Journal of Mining and Earth Sciences
Website: http://jmes.humg.edu.vn

Applying Random Forest approach in forecasting flash flood susceptibility area in Lao Cai region

Thao Phuong Thi Ngo [1,*], Long Hung Ngo [1], Khanh Quang Nguyen [1], Tinh Thanh Bui [2], Phong Van Tran [3], Ha Viet Nhu [2], Yen Hai Thi Nguyen [1]

[1] *Faculty of Information Technology, Hanoi University of Mining and Geology, Vienam*
[2] *Faculty of Geosciences and Geoengineering, Hanoi University of Mining and Geology, Vietnam*
[3] *Institute of Geological Sciences, Vietnam Academy of Science and Technology, Vietnam*

ARTICLE INFO

Article history:
Received 18th Aug. 2020
Revised 13rd Sep. 2020
Accepted 31st Oct. 2020

Keywords:
Flash Floods,
GIS,
Machine Learning,
RandomForest,
Sentinel-1A.

ABSTRACT

The main objectives of this research are to provide a new approach for flash flood prediction in Lao Cai, where frequent typhoons happen. This method is based on the Random Forest classification algorithm. The researcher applied GIS database in combination with construction machine learning model and verified the forecasting model, extracted the data based on field survey of the flash flood area of Lao Cai and GIS (Geographic Information System). The results have proved that the model can be a useful tool for flash flood forecasting model, providing more data for land planning and management for preventing and predicting flash flood for Lao Cai area.

[*]*Corresponding author*
E - mail: ngothiphuongthao@humg.edu.vn
DOI: 10.46326/JMES.2020.61(5).04

Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất
Trang điện tử: http://tapchi.humg.edu.vn

Ứng dụng phương pháp Random Forest dự báo vị trí có nguy cơ xảy ra lũ quét cho khu vực tỉnh Lào Cai

Ngô Thị Phương Thảo[1,*], Ngô Hùng Long[1], Nguyễn Quang Khánh[1], Bùi Thanh Tịnh[2], Trần Văn Phong[3], Nhữ Việt Hà[2], Nguyễn Thị Hải Yến[1]

[1] *Khoa Công nghệ thông tin, Trường Đại học Mỏ - Địa chất, Việt Nam*
[2] *Khoa Khoa học và Kỹ thuật Địa chất, Trường Đại học Mỏ - Địa chất, Việt Nam*
[3] *Viện Địa chất, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, Việt Nam*

THÔNG TIN BÀI BÁO

Quá trình:
Nhận bài 18/8/2020
Sửa xong 13/9/2020
Chấp nhận đăng 31/10/2020

Từ khóa:
Hệ thông tin địa lý,
Lũ quét,
Máy học,
Random Forest,
Sentinel-1.

TÓM TẮT

Mục tiêu chính của nghiên cứu này là cung cấp một phương pháp xây dựng mô hình dự báo vị trí có nguy cơ xảy ra lũ quét ở khu vực Lào Cai, nơi bão nhiệt đới thường xuyên xảy ra, dựa trên thuật toán phân loại Random Forest. Nghiên cứu áp dụng cơ sở dữ liệu hệ thông tin địa lý (GIS) kết hợp với mô hình máy học xây dựng và kiểm chứng mô hình dự báo, trích xuất dữ liệu dựa trên khảo sát thực địa các vùng lũ quét tại tỉnh Lào Cai và dữ liệu không gian địa lý. Kết quả cho thấy mô hình có hiệu suất cao với độ chính xác phân loại là 94,76% trên tập dữ liệu huấn luyện và khả năng dự báo là 89,29% trên tập dữ liệu kiểm tra. Kết quả đã chứng minh mô hình có thể là một công cụ hiệu quả cho mô hình dự báo vị trí có nguy cơ xảy ra lũ quét, cung cấp thêm dữ liệu cho việc quy hoạch quản lý đất sinh hoạt, phòng chống, dự báo lũ quét cho khu vực tỉnh Lào Cai.

1. Mở đầu

Lũ lụt là hiểm họa thiên nhiên thường xuyên và tàn phá lớn nhất trên toàn cầu. Không những gây thiệt hại nặng nề về tài sản mà còn ảnh hưởng tới hàng triệu người ở các đất nước khác nhau mỗi năm (Bubeck và Thieken, 2018). Theo báo cáo của các nhà nghiên cứu do tăng dân số, biến đổi khí hậu, lấn chiếm diện tích mặt nước dự báo đến năm 2050, sự phá hủy mà lũ gây ra có thể đến một nghìn tỷ USD mỗi năm (Bubeck và Thieken, 2018). Việc lập mô hình và dự báo lũ có thể làm giảm thiệt hại về kinh tế và cơ sở vật chất (Bubeck, 2012). Do đó, các nghiên cứu về xây dựng mô hình và dự báo lũ nhằm giảm thiểu những tác động xấu do lũ hiện đang là nhiệm vụ cấp bách.

Có rất nhiều phương pháp nghiên cứu và dự báo lũ quét đã được đề xuất và phát triển trên thế giới. Mô hình dự báo và đánh giá lũ lụt truyền thống thường được thiết lập trên cơ sở mô hình hóa lưu lượng dòng chảy của lưu vực tại các trạm quan trắc, từ đó dựa vào mô hình số địa hình để nội suy ra khu vực nguy có ảnh hưởng ngập lụt (Smith và Ward, 1998).

*Tác giả liên hệ
E - mail: ngothiphuongthao@humg.edu.vn
DOI: 10.46326/JMES.2020.61(5).04

Các mô hình kết hợp mô hình truyền thống với hệ thông tin địa lý và công nghệ viễn thám (Haq và nnk., 2012). Điển hình là các mô hình như HYDROTEL (Fortin và nnk., 2001), Wetspa (Liu và De Smedt, 2005) và SWAT (Jayakrishnan và nnk., 2005). Tuy nhiên, các mô hình truyền thống có nhược điểm là độ chính xác của các mô hình trong nhiều trường hợp là thấp, cần có dữ liệu quan trắc đủ dài cho mô hình hóa, cần thiết lập mạng lưới các trạm quan trắc đủ dày để cho kết quả dự báo chính xác, điều này tiêu tốn nhiều thời gian và chi phí (Sahoo và nnk., 2006; Fenicia và nnk., 2008). Có thể thấy rằng, các mô hình lũ lụt truyền thống còn nhiều hạn chế trong việc đánh giá, dự báo và phân vùng lũ cho các khu vực có địa hình phức tạp (Li và nnk., 2012). Do đó, cần thiết xây dựng một phương pháp mới để dự đoán khả năng xảy ra lũ quét và lập bản đồ dự đoán nguy cơ lũ quét hỗ trợ chính quyền địa phương và người quản lý ra quyết định trong rủi ro thiên tai.

Hiện nay, việc ứng dụng hệ thông tin địa lý (GIS), viễn thám (RS) và kỹ thuật máy học (ML) đã và đang được áp dụng phổ biến trên thế giới và có nhiều ứng dụng mang lại hiệu quả khả quan trong các lĩnh vực khoa học trái đất. Trong nghiên cứu mô hình lũ không gian, sự kết hợp GIS, RS và ML đã đem lại những thành công nhất định góp phần nâng cao hiệu quả công tác dự báo, giảm thiểu chi phí điều tra và thời gian nghiên cứu, đặc biệt với những khu vực có điều kiện địa chất phức tạp. Các công trình đã được công bố như: phân tích thứ bậc và logic mờ là các kỹ thuật định tính thường được sử dụng trong đánh giá nguy cơ lũ (Chen 2011; Tzavella và nnk., 2018; Tehrany và nnk., 2015). Mạng trí tuệ nhân tạo, máy học hỗ trợ vectơ - SVM rừng ngẫu nhiên, cây quyết định và Neural-Fuzzy là những phương pháp phổ biến nhất trong số các kỹ thuật máy học.

Trong nghiên cứu đã ứng dụng phương pháp Random Forest cho dự báo vị trí xảy ra lũ quét. Mô hình được ứng dụng thực nghiệm cho dự báo lũ quét tại hai huyện Bắc Hà và Bảo Yên thuộc tỉnh Lào Cai, Việt Nam. Đây là khu vực thường xuyên chịu ảnh hưởng nặng nề của lũ quét hàng năm (Nguyen và nnk., 2015). Kết quả nghiên cứu sẽ giúp cơ quan quản lý định hướng công tác dự báo, phòng chống khả năng xảy ra lũ quét ở khu vực nghiên cứu. Đồng thời đây cũng là dữ liệu đóng góp thêm vào lĩnh vực máy học trong nghiên cứu về các tai biến thiên nhiên.

2. Khu vực nghiên cứu

Bắc Hà và Bảo Yên bao phủ một vùng diện tích vào khoảng 1510,4 km^2, có tọa độ địa lý từ 22^{0}5′ đến 22^{0}40′ vĩ độ Bắc và từ 104^{0}10′ đến 105^{0}37′ độ kinh Đông, độ cao trải từ 38,9 m tới 1878,7 m so với mực nước biển, độ cao trung bình là 538,1 m. Các khu vực với độ dốc từ 10÷40^0, chiếm 85,4% tổng diện tích nghiên cứu, trong đó trung bình 11,5% tổng khu vực nghiên cứu có độ dốc thấp hơn 10^0 và diện tích đất có độ dốc lớn hơn 40^0 chỉ chiếm 3,1% tổng diện tích nghiên cứu. Đây là khu vực miền núi điển hình với mạng lưới sông ngòi phức tạp. Trong vùng có 2 dòng sông lớn, Sông Hồng và Sông Chảy. Sông Hồng là dòng sông lớn nhất chia đôi tỉnh Lào Cai và chảy qua vùng Bắc Hà và Bảo Yên với độ dài khoảng 28,7 km, lưu lượng dòng chảy khá lớn. Sông Chảy là dòng sông lớn chảy từ bắc sang nam với độ dài ước tính là 91,6 km, có độ dốc lớn, dòng chảy xiết, là thượng nguồn chính của thuỷ điện Thác Bà, có nhiều thác gềnh ở phía bắc.

Bắc Hà và Bảo Yên là một khu vực miền núi điển hình với khí hậu lạnh khô từ tháng mười đến tháng ba năm sau. Đáng chú ý là gió mùa nhiệt đới trong mùa mưa thường xảy ra từ tháng 4÷9. Lượng mưa hàng năm thay đổi từ 12,7 mm (tháng 12) đến 540 mm (tháng 8) và tổng lượng mưa là 1843,7 mm (được đo ở trạm Bắc Hà vào năm 2016) (GSO, 2017). Lượng mưa vào mùa mưa chiếm đến hơn 80% tổng lượng mưa một năm. Mưa tập trung chủ yếu và tháng 6, 7, 8 với tổng lượng mưa của ba tháng này chiếm tới hơn 50% lượng mưa hằng năm từ năm 2010÷2016 (GSO, 2017).

Nhiệt độ trung bình hằng năm thay đổi từ 19,27^0 C đến 23,77^0 C với nhiệt độ hàng tháng thấp nhất là 12,1^0 C vào tháng 1 (đo ở trạm Bắc Hà) và nhiệt độ hàng tháng cao nhất là 29,5^0 C vào tháng 6 (đo ở trạm Bắc Hà)(GSO, 2017).

3. Cơ sở toán học của mô hình Random Forest và phương pháp đánh giá độ chính xác

3.1. Mô hình Random Forest

Random Forest (rừng ngẫu nhiên) là phương pháp phân lớp thuộc tính được phát triển bởi Leo Breiman (Breiman, 2002; 2015) tại đại học California, Berkeley. Random Forest (RF) được xây dựng dựa trên 3 thành phần chính là: (1) CART (Classification and Regression Trees), (2)

học toàn bộ, hội đồng các chuyên gia, kết hợp các mô hình, và (3) tổng hợp bootstrap (bagging). Về bản chất RF sử dụng kỹ thuật có tên gọi là bagging. Kỹ thuật này cho phép lựa chọn một nhóm nhỏ các thuộc tính tại mỗi nút của cây phân lớp để phân chia thành các mức tiếp theo. Do đó, RF có khả năng phân chia không gian tìm kiếm rất lớn thành các không gian tìm kiếm nhỏ hơn, nhờ thế thuật toán có thể thực hiện việc phân loại một cách nhanh chóng và dễ dàng (Hình 1).

Theo Breiman 2015, thuật toán RF được mô tả gồm:

1. Chọn *T* là số lượng các cây thành phần sẽ được xây dựng.

2. Chọn *m* là số lượng các thuộc tính sẽ được dùng để phân chia tại mỗi node của cây, *m* thường nhỏ hơn p rất nhiều, p là tổng số các thuộc tính. Giá trị *m* được giữ không đổi trong suốt quá trình dựng cây.

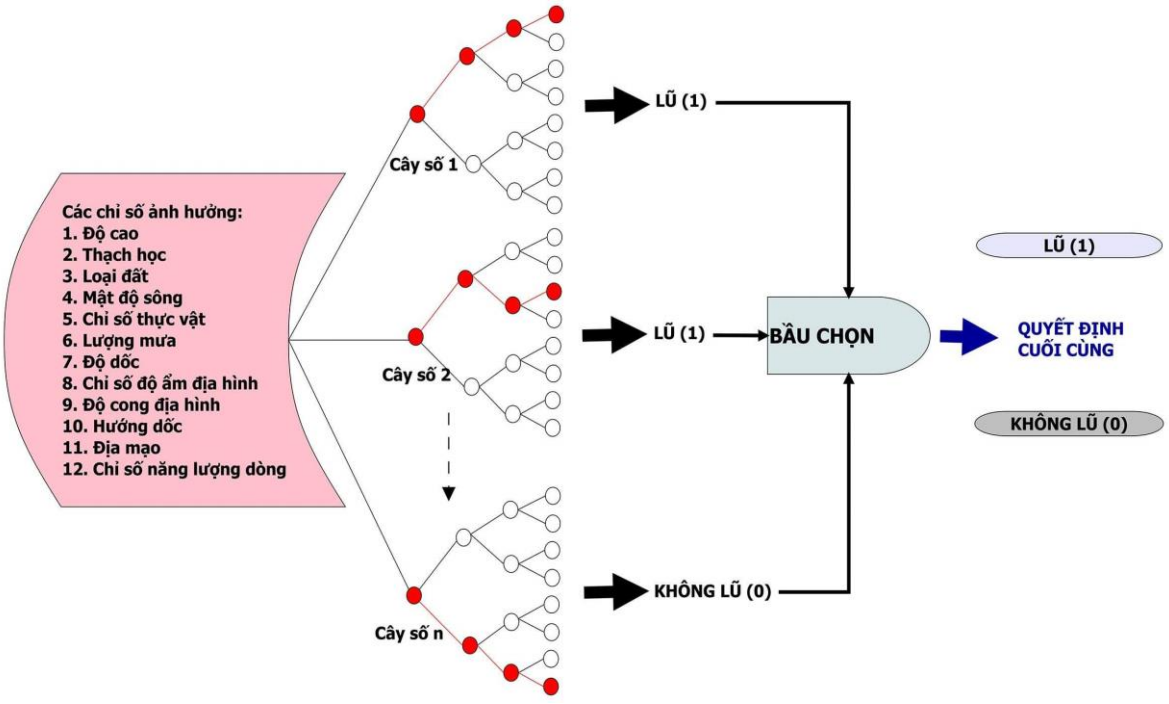
3. Dựng *T* cây quyết định. Trong đó mỗi cây được hình thành như sau: a) Xây dựng tập mẫu khởi động (bootstrap) với n mẫu, hình thành từ việc hoán vị tập các mẫu ban đầu. Mỗi cây sẽ được dựng từ tập khởi động này; b) Khi xây dựng cây, tại mỗi node sẽ chọn ra *m* thuộc tính, và sử dụng *m* thuộc tính này để tìm ra cách phân chia tốt nhất; c) Mỗi cây được phát triển lớn nhất có thể và không bị cắt xén.

4. Sau khi xây dựng được Random Forest, để phân lớp cho đối tượng *T*, thu thập kết quả phân lớp đối tượng này trên tất cả các cây quyết định và sử dụng kết quả được chọn nhiều nhất làm kết quả cuối cùng của thuật toán. Tỉ lệ lỗi của cây tổng thể phụ thuộc vào độ mạnh của từng cây quyết định thành phần và mối quan hệ qua lại giữa các cây đó.

Khi tập mẫu được rút ra từ một tập huấn luyện của một cây với sự thay thế (bagging), thì theo ước tính có khoảng 1/3 các phần tử không có nằm trong mẫu này (Breiman, 2002). Điều này có nghĩa là chỉ có khoảng 2/3 các phần tử trong tập huấn luyện tham gia vào trong các tính toán và 1/3 các phần tử này được gọi là dữ liệu out-of-bag. Dữ liệu huấn luyện bị loại ra khỏi các mẫu bootstrap được sử dụng để ước tính lỗi dự báo và tầm quan trọng của biến. Trong ước tính lỗi, các mẫu OOB được dự báo bởi các cây tương ứng và bằng cách tổng hợp các dự báo, lỗi bình phương trung bình (MSEOBB) đã được tính bằng công thức (1) (Zhang và Ma 2012):

$$MSE_{OOB} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \widehat{Yi_{OOB}}\right)^2 \qquad (1)$$

Trong đó: $\widehat{Yi_{OOB}}$ - chỉ số dự báo OOB cho việc quan sát y_i. Về tầm quan trọng của biến, các giá trị của biến dự báo cụ thể được hoán vị ngẫu nhiên



Hình 1. Mô hình Random Forest cho dự báo nguy cơ lũ quét

trong dữ liệu OOB của cây, trong khi giá trị của các yếu tố dự báo khác vẫn cố định. Dữ liệu OOB được sửa đổi đã được dự báo, sự khác biệt giữa các giá trị MSEs thu được từ dữ liệu OOB được hoán vị và dữ liệu OOB gốc đã đưa ra một thước đo về tầm quan trọng khác nhau.

3.2. Kỹ thuật thống kê đánh giá độ chính xác của mô hình

Hiệu suất dự báo nguy cơ lũ quét của mô hình được đánh giá bằng các chỉ số thống kê sau: sai số trung phương (RMSE), sai số tuyệt đối trung bình (MAE) (Mohammadzadeh và nnk., 2014). Sử dụng đường cong ROC để đánh giá hiệu suất tổng thể của mô hình. Hơn nữa, diện tích phía dưới đường cong (AUC) là chỉ số thống kê để đánh giá và so sánh định lượng hiệu suất dự báo tổng thể của mô hình (Khosravi và nnk., 2018). Giá trị AUC giao động từ 0,0 đến 1,0. Mô hình có AUC càng gần với 1,0 thì có hiệu suất dự báo lũ quét tổng thể càng cao (Bui Tien Dieu và nnk., 2016a).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - t_i)2}{n}} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - t_i| \quad (3)$$

$$r = \frac{\sum_{i=1}^{n}(y_i - \overline{y}_i)(t_i - \overline{t})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 (t_i - \overline{t})^2}} \quad (4)$$

Trong đó: y_i và $\bar{y}$ - giá trị đầu ra của của mẫu huấn luyện thứ *i* và giá trị trung bình đầu ra từ mô hình; t_i và $\overline{t}$ - giá trị gốc của mẫu huấn luyện thứ *i* và giá trị trung bình gốc của tổng số mẫu; *n* - tổng số mẫu.

Để đánh giá chi tiết chất lượng của mô hình dự báo, có các tham số thống kê gồm độ nhạy (SST), độ đặc đặc trưng (SPF), công suất dự báo dương (PPV) và công suất dự báo âm (NPV). Mức độ phù hợp của mô hình và bộ dữ liệu giá trị Kappa và độ chính xác phân loại (ACC) (Martínez-Álvarez và nnk., 2013, Bui Tien Dieu và Hoang Duc Nhat, 2017) được sử dụng theo các công thức:

$$PPV = \frac{TP}{TP + FP} \quad (5)$$

$$NPV = \frac{TN}{TN + FN} \quad (6)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$Kappa\ index\ (K) = \frac{CLA + P_{exp}}{1 - P_{exp}} \quad (8)$$

$$SST = \frac{TP}{TP + FN} \quad (9)$$

$$SPF = \frac{TN}{TN + FP} \quad (10)$$

Trong đó: *TP* - dương thực; *TN* - âm thực; *FP* - dương giả; *FN* - âm giả.

4. Phương pháp nghiên cứu

4.1. Xây dựng bản đồ thành phần

Để xây dựng mô hình dự báo và phân vùng nguy cơ lũ quét, bên cạnh bản đồ hiện trạng lũ quét, điều quan trọng là phải xác định được các bản đồ thành phần là nguyên nhân gây ra lũ quét. Cần chú ý là việc lựa chọn các bản đồ thành phần này tùy theo các đặc điểm khác nhau các khu vực nghiên cứu và dữ liệu sẵn có (Razavi Termeh và nnk., 2018). Địa hình là một thành phần chính của quá trình thủy văn, có liên quan mạnh mẽ đến sự kiện lũ quét bởi độ dốc làm tăng tốc độ dòng chảy nhanh (Destro và nnk., 2018). Do đó, các bản đồ thành phần liên quan đến địa hình như độ cao, độ dốc, độ cong địa hình, địa mạo, bề mặt, chỉ số độ ẩm địa hình (TWI) và chỉ số năng lượng dòng (SPI) được sử dụng. Trong nghiên cứu này, mô hình số độ cao (DEM) với độ phân giải không gian 10 m cho khu vực nghiên cứu được tạo ra từ bản đồ địa hình quốc gia với tỷ lệ 1: 10.000 do Bộ Tài nguyên và Môi trường Việt Nam (MONRE) thành lập. Từ mô hình DEM này, thành lập được 7 bản đồ thành phần: độ cao, độ dốc, hướng dốc, độ cong, TWI, SPI và địa mạo.

Độ cao và độ dốc được lựa chọn bởi vì dòng nước xuất hiện khi có trọng lực, di chuyển từ nơi cao xuống nơi thấp. Độ dốc có chức năng kiểm soát tốc độ dòng chảy bề mặt và thông thường những khu vực có nguy cơ lũ quét thường là khu vực bằng phẳng và thấp (Tehrany và nnk., 2013). Độ cong địa hình cũng được xem xét vì các khu vực lũ quét thường liên quan tới bản đồ thành phần hội tụ địa hình cao (Manfreda và nnk., 2014). Trong nghiên cứu này, bản đồ độ cao (Hình 2e) với 8 mức được sử dụng, trong khi đó 9 mức cho bản đồ độ

dốc (Hình 2b) và 7 mức được xây dựng cho bản đồ độ cong địa hình(Hình 2c). Các mức của ba bản đồ này được xác định dựa trên phương pháp ngắt quãng tự nhiên có sẵn trong ESRI-ArcGIS.

Bản đồ hình thái địa mạo và hướng dốc được lựa chọn vì địa mạo có thể ảnh hưởng đến sự hội tụ của dòng chảy (Santosh và nnk., 2003), trong khi đó, hướng dốc kiểm soát hướng dòng chảy mặt nước. Đối với nghiên cứu này, bản đồ hình thái địa mạo (Hình 2k) với 8 mức và bản đồ hướng dốc (Hình 2d) bao gồm 9 mức được lựa chọn. TWI và SPI là các thông số thủy văn điển hình ảnh hưởng đến cường độ dòng chảy và sự tích tụ nước (Martınez-Casasnovas, Ramos và Poesen 2004); do đó chúng đã được lựa chọn cho mô hình nguy cơ lũ quét trong nghiên cứu này. TWI (Beven và nnk., 1984) và SPI (Moore và nnk., 1991) được tính toán bằng cách sử dụng các phương trình (11), (12):

$$TWI = ln(a/\tan\beta) \quad (11)$$

$$SPI = a * \tan\beta \quad (12)$$

Trong đó: a - diện tích ngược dốc cục bộ tiêu thoát qua một điểm nhất định trên mỗi ô lưới trên DEM; β - góc dốc tính bằng radian. Trong phân tích này, bản đồ TWI (Hình 2a) và bản đồ SPI (Hình 2l) với bảy mức đã được sử dụng.

Mật độ sông suối, được tính bằng cách chia chiều dài của sông (km) trên diện tích lưu vực (km^2), là một bản đồ thành phần quan trọng ảnh hưởng đến lũ quét. Điều này là do các vùng có mật độ dòng cao hơn thường có nhiều khả năng phản ứng nhanh với mưa bão (Brody và nnk., 2007); do đó chúng dễ bị lũ quét hơn. Bản đồ mật độ sông suối với 7 mức được xem xét cho công việc hiện tại. Chỉ số thực vật NDVI là một chỉ số phản ánh mức độ thảm thực vật dày đặc và có khả năng lũ quét dễ xảy ra hơn ở những khu vực có mật độ thực vật thấp (Tehrany và nnk., 2013); do đó NDVI được lựa chọn để phân tích lũ quét.

Trong phân tích này, bản đồ NDVI được tính 8 mức sử dụng (Hình 2i) từ dữ liệu ảnh Landsat-8 (OLI) với độ phân giải là 30 m và download tại http://earthexplorer.usgs.gov theo phương trình (13) (Reed và nnk., 1994):

$$NDVI = (NIR - RED)/(NIR + RED) \quad (13)$$

Trong đó: *NIR* và *RED* - độ phản xạ bề mặt của dải cận hồng ngoại và dải màu đỏ tương ứng .

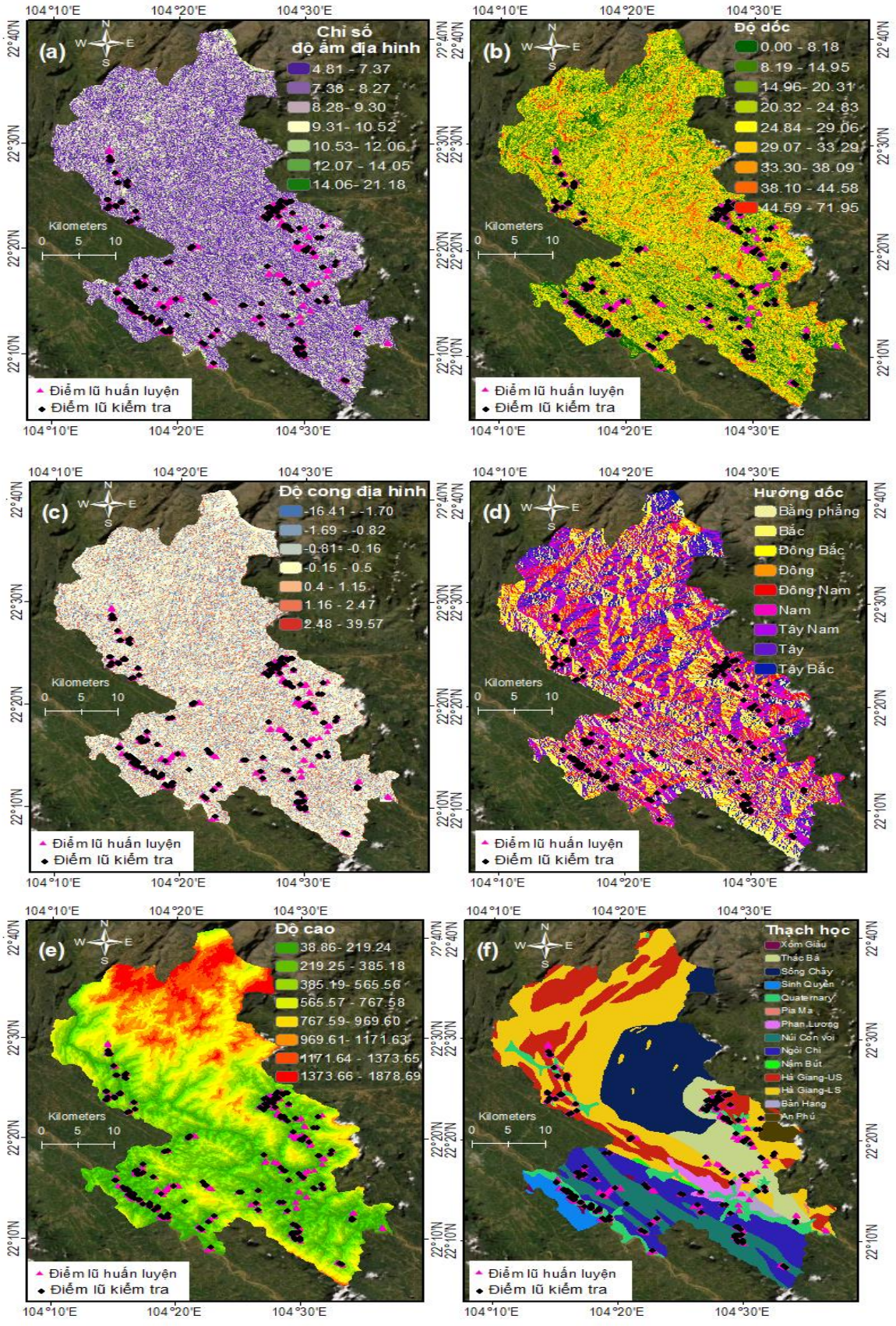
Bản đồ loại đất (Hình 2g) đã được công nhận phổ biến như là một bản đồ thành phần quan trọng ảnh hưởng đến cơ chế dòng chảy mưa, trong khi cấu trúc thạch học (Hình 2f) ảnh hưởng mạnh mẽ đến kiến trúc của mô hình thoát nước (Pizzuto 1995) liên quan đến sự phát triển của vùng đồng bằng ngập lụt. Vì lũ quét thường liên quan đến mưa bão cường độ cao và ngắn (Borga và nnk., 2011), do đó lượng mưa là bản đồ thành phần kiểm soát chính cho mô hình lũ quét. Đối với khu vực nghiên cứu này, các trận mưa lớn cường độ cao xảy ra vào ngày 10, 11 và 12 tháng 10 năm 2017 đã tạo ra lũ quét dữ dội nghiêm trọng. Ngoài ra, lượng mưa đã kéo dài trong 9 ngày trước và lượng mưa đã kết thúc sau ngày 12 tháng 10 năm 2017; do đó, tổng lượng mưa đo được từ ngày 1 đến 12 tháng 10 năm 2017 tại 16 trạm mưa trong và xung quanh khu vực nghiên cứu được sử dụng để tạo ra bản đồ lượng mưa (Hình 2j).

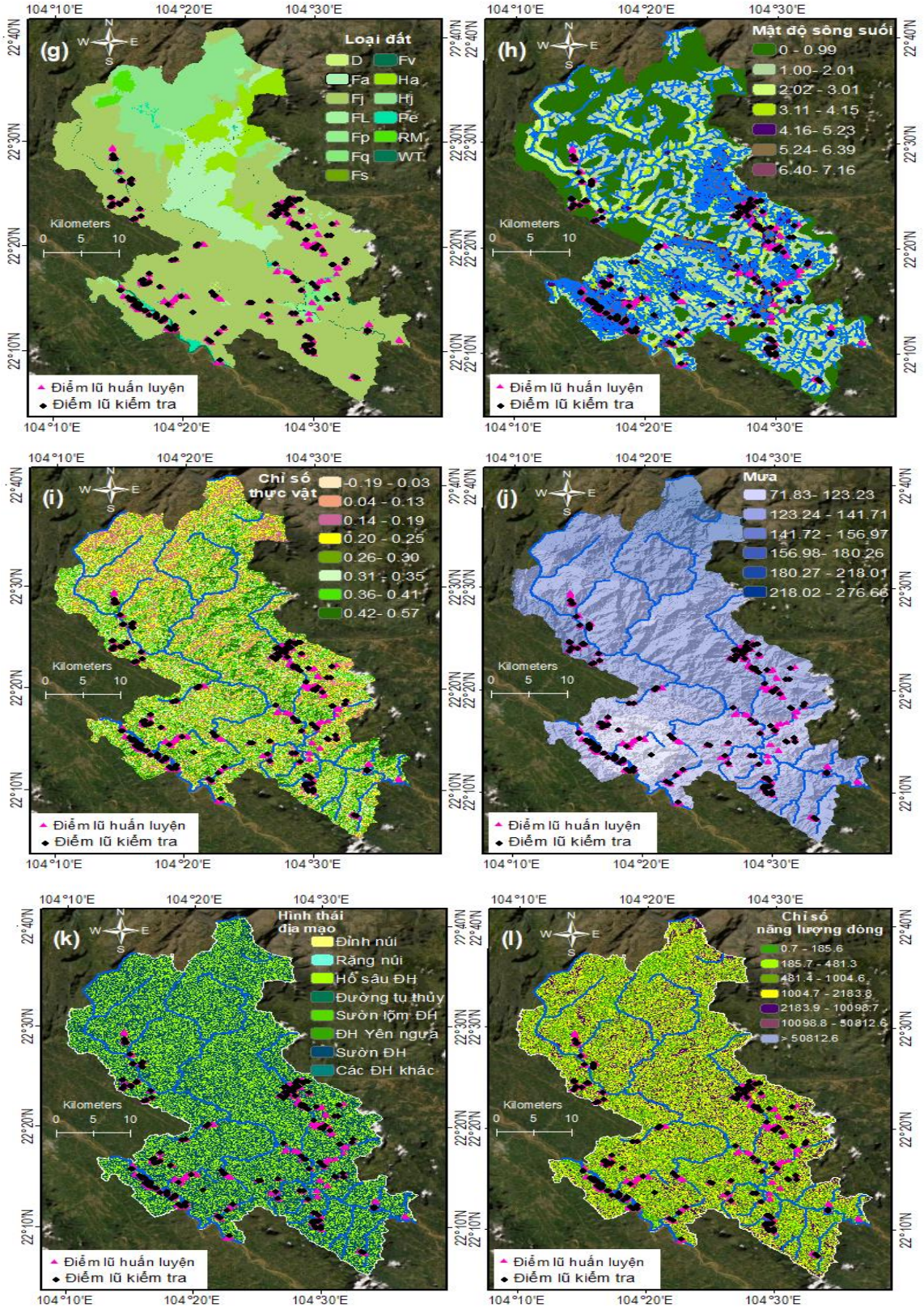
4.2. Phân tích đa cộng tuyến và lựa chọn các bản đồ thành phần

Trong bài báo này, đa cộng tuyến cho các bản đồ thành phần ảnh hưởng lũ quét đã được kiểm tra qua hệ số phóng đại phương sai VIF (Variance Inflation Factors) và dung sai TOL (Tolerances) (Dormann và nnk., 2013). Các nghiên cứu trước đây được (Bùi Tiến Diệu và nnk., 2011; Khosravi và nnk., 2018) cho thấy rằng VIF > 10 hoặc TOL < 0,1 thì vấn đề đa cộng tuyến giữa các bản đồ thành phần ảnh hưởng. Kết quả Bảng 1 cho thấy không có mối liên hệ giữa các bản đồ thành phần gây ảnh hưởng của lũ quét trong khu vực nghiên cứu.

Bảng 1. Phân tích đa cộng tuyến cho các bản đồ thành phần ảnh hưởng đến lũ quét.

TT	Bản đồ thành phần	Phân tích đa cộng tuyến	
		TOL	VIF
1	Độ cao	0,43	2,33
2	Độ dốc	0,15	6,82
3	Độ cong địa hình	0,68	1,46
4	Hình thái địa mạo	0,58	1,73
5	Hướng dốc	0,84	1,19
6	TWI	0,17	5,90
7	SPI	0,38	2,65
8	Mật độ sông suối	0,55	1,84
9	NDVI	0,64	1,57
10	Loại đất	0,79	1,26
11	Thạch học	0,80	1,24
12	Lượng mưa	0,59	1,69


(a)
Chỉ số độ ẩm địa hình
4.81 - 7.37
7.38 - 8.27
8.28- 9.30
9.31- 10.52
10.53- 12.06
12.07 - 14.05
14.06- 21.18
(b)
Độ dốc
0.00 - 8.18
8.19 - 14.95
14.96- 20.31
20.32 - 24.83
24.84 - 29.06
29.07 - 33.29
33.30- 38.09
38.10 - 44.58
44.59 - 71.95
(c)
Độ cong địa hình
-16.41 - -1.70
-1.69 - -0.82
-0.81- -0.16
-0.15 - 0.5
0.4 - 1.15
1.16 - 2.47
2.48 - 39.57
(d)
Hướng dốc
Bằng phẳng
Bắc
Đông Bắc
Đông
Đông Nam
Nam
Tây Nam
Tây
Tây Bắc
(e)
Độ cao
38.86- 219.24
219.25 - 385.18
385.19- 565.56
565.57 - 767.58
767.59- 969.60
969.61- 1171.63
1171.64 - 1373.65
1373.66 - 1878.69
(f)
Thạch học
Xóm Giấu
Thác Bà
Sông Chảy
Sinh Quyền
Quaternary
Pia Ma
Phan Lương
Núi Cốn Voi
Ngòi Chi
Nậm Bút
Hà Giang-US
Hà Giang-LS
Bản Hang
An Phú
Kilometers
0 5 10
N
S
W
E
Điểm lũ huấn luyện
Điểm lũ kiểm tra
104°10'E
104°20'E
104°30'E
22°40'N
22°30'N
22°20'N
22°10'N



Hình 2. Các bản đồ thành phần: (a) chỉ số độ ẩm địa hình, (b) độ dốc, (c) độ cong địa hình, (d) hướng dốc, (e) độ cao, (f) thạch học, (g) loại đất , (h) mật độ sông suối, (i) chỉ số thực vật, (j) lượng mưa, (k) địa mạo, (l) chỉ số năng lượng dòng.

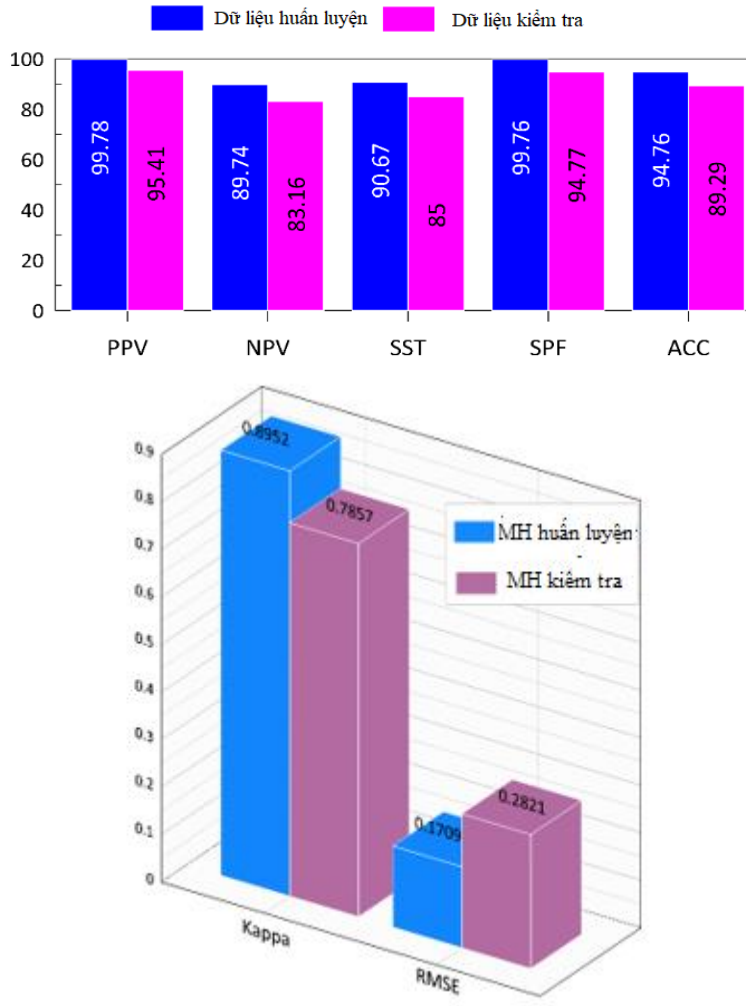
Vì vậy, các bản đồ thành phần này đã được lựa chọn cho mô hình dự báo nguy cơ lũ quét.

5. Kết quả và thảo luận

5.1. Hiệu suất của mô hình

Mô hình dự báo vị trí có nguy cơ xảy ra lũ quét được huấn luyện bằng cách sử dụng tập dữ liệu huấn luyện gồm 12 yếu tố ảnh hưởng. Từ kết quả của mô hình đánh giá (Hình 3) cho thấy mô hình đã thực hiện rất tốt với tập dữ liệu huấn luyện, mức độ chính xác của mô hình với tập dữ liệu rất cao với giá trị ACC là 94,76%. Mức độ phù hợp của mô hình và bộ dữ liệu huấn luyện là tốt ở mức 0,8952 (Kappa) với sai số trung phương thấp (RMSE) 0,1709%. Ngoài ra, tỷ lệ phần trăm của các pixel không có lũ quét được phân chia chính xác với giá trị (SPF) của mô hình là 99,76%, tỷ lệ phần trăm cho các pixel có lũ quét thấp hơn (SST) là 90,67%. Ngược lại, xác suất phân loại pixel của mô hình đối với lớp lũ quét rất cao ở mức 99,78% (PPV) và xác suất phân loại pixel của mô hình đối với lớp không lũ quét (NPV) là 89,74%.

Sau khi mô hình lũ quét được huấn luyện với tập dữ liệu huấn luyện, mô hình này được tiếp tục đánh giá với tập dữ liệu kiểm tra và kết quả trong (Hình 3) cho thấy kết quả dự báo là khá cao với 89,29% (ACC). Kappa của mô hình là 0,7857 cho thấy hiệu suất dự báo của mô hình tốt với sai số trung phương thấp (RMSE) 0,2821. Tỷ lệ phần trăm dự báo chính xác của mô hình đối với các pixel lũ quét là 95,41% (PPV) và cho các pixel không lũ quét là 83,16% (NPV). Tỷ lệ các pixel lũ quét được dự báo chính xác là 85,0% (SST) và 94,77% pixel không lũ quét được mô hình dự báo chính xác (SPF).



Hình 3. Các thông số cho mô hình đánh giá lũ quét.

5.2. Đánh giá độ chính xác

Khả năng dự báo của mô hình lũ quét được đo bằng đường cong ROC và AUC (Hình 4). Kết quả AUC của mô hình được đề xuất trong tập dữ liệu huấn luyện là 0,989 và trong tập dữ liệu kiểm tra là 0,944. Từ kết quả trên có thể kết luận rằng mô hình được đề xuất có thể dự báo chính xác các vị trí xảy ra lũ quét cho khu vực nghiên cứu này theo như phân loại chỉ số AUC của Cantor và Kattan (2000).

5.3. Xây dựng bản đồ phân vùng nguy cơ lũ quét

Mô hình dự báo các vị trí có nguy cơ xảy ra lũ quét cuối cùng đã được học bằng cách sử dụng tập dữ liệu huấn luyện để tính toán các chỉ số độ nhạy cảm xảy ra lũ quét cho khu vực nghiên cứu. Tất cả các yếu tố ảnh hưởng đã được chuyển đổi sang định dạng raster và sau đó được đưa vào mô hình Random Forest để tạo ra các chỉ số nhạy cảm được gọi là chỉ số xác suất lũ quét. Các chỉ số này được phân loại dựa trên mức độ ảnh hưởng của các yếu tố đến khả năng xảy ra lũ quét. Cuối cùng, bản đồ dự báo các vị trí có nguy cơ xảy ra lũ quét cho khu vực huyện Bắc Hà và Bảo Yên (Lào Cai) được xây dựng bằng bản đồ bởi một loạt các chỉ số xác suất lũ quét như Hình 5.

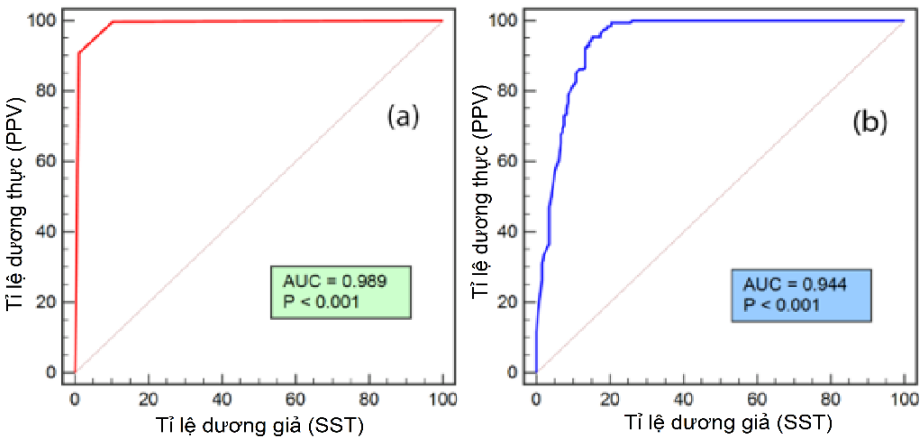
6. Kết luận và kiến nghị

Đã có nhiều nghiên cứu về việc sử dụng máy học trong các nghiên cứu lũ quét gần đây với nhiều phương pháp khác nhau. Tuy nhiên, việc xây dựng một mô hình hoàn hảo về lũ quét mà không có lỗi là gần như không thể, do đó việc xác định một mô hình với độ chính xác cao để dự báo vị trí xảy ra lũ quét ở một khu vực cụ thể là vô cùng cần thiết, điều này luôn đòi hỏi phải có những đánh giá và nghiên cứu mới để nâng cao độ chính xác trong việc sử dụng học máy trong nghiên cứu các tai biến thiên nhiên. Trong nghiên cứu này, nhóm tác giả đã ứng dụng mô hình máy học rừng ngẫu nhiên Random Forest và kết quả của nghiên cứu cho thấy độ chính xác của mô hình là tốt, với ACC là 94,76% trong tập dữ liệu huấn luyện và 89,29% trong tập dữ liệu kiểm tra. Mô hình này cũng thực hiện tốt cả dữ liệu huấn luyện và dữ liệu kiểm tra với AUC lần lượt là 0,989 và 0,944. Giá trị hiệu suất dự báo (kappa) của mô hình tốt bằng 0,8952 trong bộ dữ liệu huấn luyện và 0,7857 trong bộ dữ liệu kiểm tra.

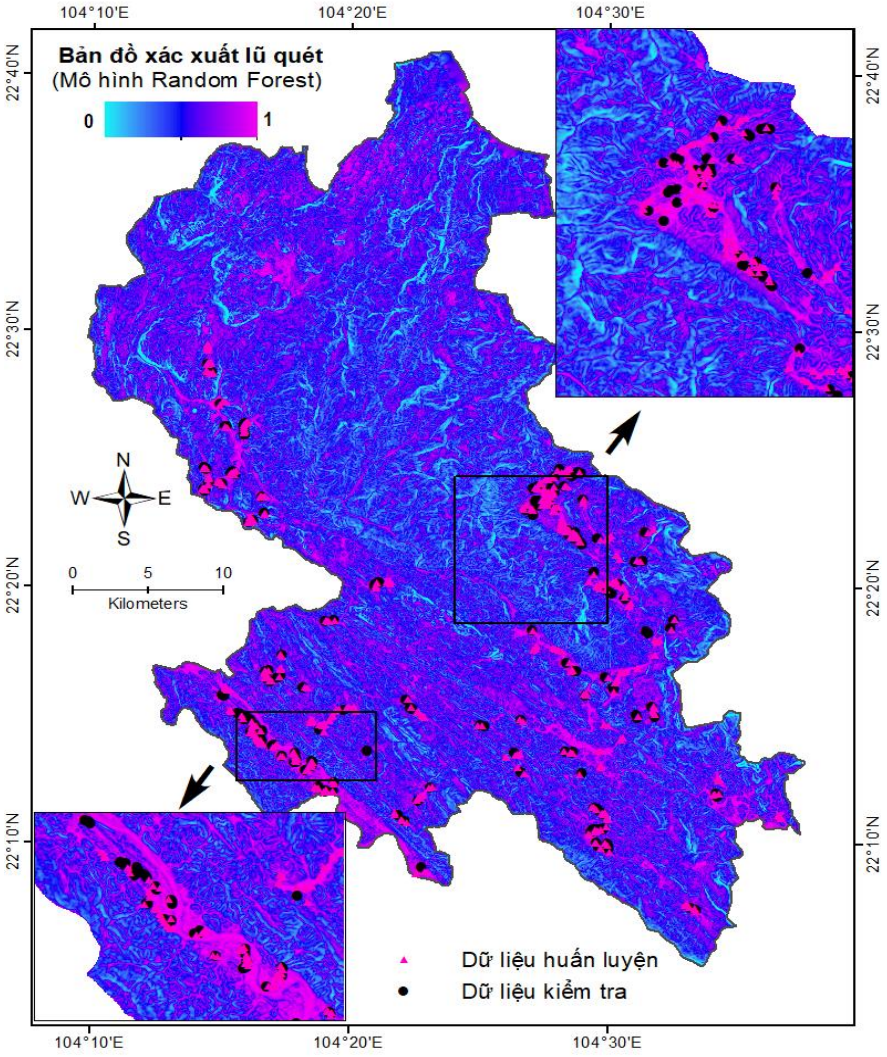
Nhìn chung, kết quả của nghiên cứu này đã minh họa hiệu quả của việc sử dụng máy học để dự báo khu vực dễ xảy ra lũ quét. Cho thấy mô hình Random Forest có tiềm năng và có thể được xem xét sử dụng để lập bản đồ dự báo vị trí xảy ra lũ quét ở các khu vực khác có cùng điều kiện môi trường địa lý. Cuối cùng, kết quả trong nghiên cứu này có thể được sử dụng để nghiên cứu thêm như lập kế hoạch cho việc phòng chống và dự báo lũ quét ở những khu vực có nguy cơ xảy ra lũ quét ở tỉnh Lào Cai.

Lời cảm ơn

Nghiên cứu này được tài trợ bởi đề tài cấp Bộ mã số B2018-MDA-18DT (Bộ Giáo dục và Đào tạo Việt Nam). Trân trọng cảm ơn Công ty cổ phần tư vấn, đầu tư xây dựng và ứng dụng công nghệ mới (Vinaconex R&D) đã giúp đỡ tác giả thu thập dữ liệu và khảo sát thực địa.



Hình 4. Phân tích ROC của mô hình: (a) tập dữ liệu huấn luyện và (b) tập dữ liệu kiểm tra.



Hình 5. Bản đồ dự báo vị trí có nguy cơ xảy ra lũ quét khu vực Lào Cai.

Tài liệu tham khảo

Beven, K., Kirkby, M., Schofield, N. & Tagg , A., (1984). Testing a physically-based flood forecasting model (TOPMODEL) for three UK catchments. *Journal of Hydrology* 69, 119-143.

Borga, M., Anagnostou, E. N. G., Blöschl & Creutin, J. D., (2011). Flash flood forecasting, warning and risk management: the HYDRATE project. *Environmental Science & Policy* 14, 834-844.

Breiman, L., (2002). Manual On Setting Up, Using, And Understanding Random Forests V3.1. Statistics Department University of California Berkeley, CA, USA,1, 58.

Breiman, L., (2015). Random forests leo breiman and adele cutler. *Random Forests-Classification Description*. Retrieved. http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm (accessed on 22 March 2016).

Brody, S. D., Zahran ,S., Maghelal , P., Grover,H. & Highfield, W. E., (2007). The rising costs of floods: Examining the impact of planning and development decisions on property damage in Florida. *Journal of the American Planning Association* 73, 330-345.

Bubeck, P. & Thieken, A. H., (2018). What helps people recover from floods? Insights from a survey among flood-affected residents in Germany. *Regional Environmental Change* 18, 287-296.

Bubeck, P., Botzen, W. J. W., Aerts, J. C. J. H., (2012). A review of risk perceptions and other factors that influence flood mitigation behavior. *Risk Anal* 32 (9), 1481-1495.

Bui Tien Dieu, Hoang Duc Nhat, (2017). A Bayesian framework based on a Gaussian mixture model and radial-basis-function Fisher discriminant analysis (BayGmmKda V1. 1) for spatial prediction of floods. *Geoscientific Model Development* 10, 3391.

Bui Tien Dieu, Owe Lofman, Inge Revhaug & Oystein Dick, (2011). Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index và logistic regression. *Natural Hazards* 59, 1413.

Cantor, S. B. & Kattan,M. W., (2000). Determining the area under the ROC curve for a binary diagnostic test. *SAGE Journals* 20, 468-470. https://doi.org/10.1177/0272989X0002000410.

Cha Zhang, Yunqian Ma, (2012). Ensemble machine learning: methods and applications. *Springer* VIII, 332.

Chen, Y., Yeh, C. H., Yu, B., (2011). Integrated application of the analytic hierarchy process và the geographic information system for flood risk assessment and flood plain management in Taiwan. *Natural Hazards* 59, 1261-1276.

Destro, E., Amponsah, W., Nikolopoulos, E. I., Marchi, L., Marra, F., Zoccatelli, D. & Borga, M., (2018). Coupled prediction of flash flood response and debris flow occurrence: Application on an alpine extreme flood event. *Journal of Hydrology* 558, 225-237.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, C., McClean, Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D. & Lautenbach, S., (2013). Collinearity: a review of methods to deal with it và a simulation study evaluating their performance. *Ecography* 36, 27-46.

Fenicia, F., Savenije, H. H., Matgen, P. & Pfister, L., (2008). Understending and catchment behavior through stepwise model concept improvement. *Water Resources Research* 44.

Fortin, J.-P., Turcotte, R., S., Massicotte, Moussa , R., Fitzback, J. & Villeneuve, J. P., (2001). Distributed watershed model compatible with remote sensing and GIS data. I: Description of model. *Journal of Hydrologic Engineering* 6, 91-99.

GSO. 2017. Lao Cai statistical year book 2016 470. Hanoi: *Statistical Publishing House*.

Haq, M., Akhtar, M., Muhammad, S., Paras, S. & Rahmatullah, J., (2012). Techniques of remote sensing and GIS for flood monitoring và damage assessment: a case study of Sindh province, Pakistan. *The Egyptian Journal of Remote Sensing and Space Science* 15, 135-141.

Jayakrishnan, R., Srinivasan, R., Santhi, C. & Arnold, J., (2005). Advances in the application of the SWAT model for water resources management. *Hydrological processes* 19, 749-762.

Katerina Tzavella, Alexander Fekete, Frank Fiedrich, (2018). Opportunities provided by geographic information systems and volunteered geographic information for a timely emergency response during flood events in Cologne, *Germany. Natural Hazards* 91, 29-57.

Khosravi, K., Binh Thai Pham, Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., Prakash, I. & Dieu Tien Bui, (2018). A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. *Science of The Total Environment* 627, 744-755.

Li, X. H., Zhang, Q., Shao, M. & Li, Y. L., (2012). A comparison of parameter estimation for distributed hydrological modelling using automatic và manual methods. *In Advanced Materials Research* 2372-2375. Trans Tech Publ.

Liu, Y. & De Smedt, F., (2005). Flood modeling for complex terrain using GIS and remote sensed information. *Water Resources Management* 19, 605-624.

Livingston, F., (2005). Implementation of Breiman's Random Forestmachine learning algorithm. *Machine Learning Journal Paper*, 1-13.

Manfreda, S., Nardi, F., Samela, C., Grimaldi, S., Taramasso, A. C., Roth,G. & Sole A., (2014). Investigation on the use of geomorphic approaches for the delineation of flood prone areas. *Journal of Hydrology* 517, 863-876.

Martínez-Álvarez, F., Reyes, J., Morales-Esteban, A. & Rubio-Escudero, C., (2013). Determining the best set of seismicity indicators to predict earthquakes. Two case studies: Chile and the Iberian Peninsula. *Knowledge-Based Systems* 50, 198-210.

Martınez-Casasnovas, Ramos, J., M. & Poesen, J., (2004). Assessment of sidewall erosion in large gullies using multi-temporal DEMs and logistic regression analysis. *Geomorphology* 58, 305-321.

Mohammadzadeh, D., Bazaz, J. B. & Alavi, A. H., (2014). An evolutionary computational approach for formulation of compression index of fine-grained soils. *Engineering Applications of Artificial Intelligence* 33, 58-68.

Moore, I. D., Grayson, R. & Ladson, A., (1991). Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrological processes* 5, 3-30.

MSN_Flood. *Water Science and Engineering*10 (3), 175-183.

Nguyen Hong Quang, Jan Degener & Martin Kappas, (2015). Flash Flood Prediction by Coupling KINEROS2 and HEC-RAS Models for Tropical Regions of Northern Vietnam. *Hydrology* 2, 242.

Nikoo, M., Ramezani, F., Hadzima-Nyarko, M., Nyarko, E. K. & Nikoo, M., (2016). Flood-routing modeling with neural network optimized by social-based algorithm. *Natural Hazards* 82, 1-24.

Pizzuto, J. E., (1995). Downstream fining in a network of gravel-bedded rivers. *Water Resources Research* 31, 753-759.

Razavi Termeh, S. V., Kornejady, A., Pourghasemi, H. R. & Keesstra, S., (2018). Flood susceptibility mapping using novel ensembles of adaptive neuro fuzzy inference system and metaheuristic algorithms. *Science of The Total Environment* 615, 438-451.

Reed, B. C., Brown, J. F., D., Lovel, T. R. & Merchant, J. W. & Ohlen, D. O., (1994). Measuring phenological variability from satellite imagery. *Journal of Vegetation Science* 5, 703-714.

Sahoo, B., Chatterjee, C., Raghuwanshi,N. S., Singh, R. & Kumar, R., (2006). Flood estimation by GIUH-based Clark and Nash models. *Journal of Hydrologic Engineering* 11, 515-525.

Santosh, K. Aryal, Russell, Mein, G., Emmett, O'Loughlin, M., (2003). The concept of effective length in hillslopes: assessing the influence of climate and topography on the contributing areas of catchments. *Hydrological Processes* 17, 131-151.

Smith, K. & Ward, R., (1998). Floods: physical processes and human impacts. *Chichester*, 382.

Flood susceptibility analysis and its verifi-cation using a novel ensemble support vector machine và frequency ratio method. *Stochastic Environmental Research and Risk Assessment* 29 (4), 1149.