

BỘ NÔNG NGHIỆP VÀ MÔI TRƯỜNG
VIỆN KHOA HỌC THỦY LỢI VIỆT NAM

BÁO CÁO SẢN PHẨM 4
QUY TRÌNH ÚNG DỤNG TRÍ TUỆ NHÂN TẠO
VÀ DỮ LIỆU ĐỊA KHÔNG GIAN ĐỂ PHÂN VÙNG
LŨ QUÉT CHO MỘT HUYỆN Ở VÙNG NÚI PHÍA BẮC

ĐỀ TÀI: NGHIÊN CỨU ÚNG DỤNG TRÍ TUỆ NHÂN TẠO
VÀ DỮ LIỆU ĐỊA KHÔNG GIAN ĐỂ PHÂN VÙNG LŨ
QUÉT QUY MÔ CẤP HUYỆN

Cơ quan chủ quản: Bộ Nông nghiệp và Môi trường
Tổ chức chủ trì: Viện Khoa học Thủy lợi Việt Nam
Chủ nhiệm: Lê Văn Thìn
Thời gian thực hiện: 01/2023÷06/2025

HÀ NỘI - 2025

MỤC LỤC

CHƯƠNG 1. DỮ LIỆU SỬ DỤNG, CHUẨN HÓA DỮ LIỆU VÀ XÂY DỰNG MÔ HÌNH TRÍ TUỆ NHÂN TẠO PHÂN VÙNG LŨ QUÉT	1
1.1. Đầu vào và cấu trúc dữ liệu	1
1.2. Xây dựng mô hình học máy	6
1.3. Xây dựng mô hình học sâu	22
1.4. Phân tích, đánh giá các mô hình trí tuệ nhân tạo trong phân vùng lũ quét....	31
CHƯƠNG 2. KẾT QUẢ PHÂN VÙNG LŨ QUÉT CHO KHU VỰC NGHIÊN CỨU	34
2.1. Kết quả phân vùng lũ quét	34
2.2. Đánh giá sự phù hợp của kết quả phân vùng lũ quét	37
2.3. Đánh giá chung	43
CHƯƠNG 3. QUY TRÌNH ÚNG DỤNG TRÍ TUỆ NHÂN TẠO VÀ VIỄN THÁM ĐỂ PHÂN VÙNG LŨ QUÉT	44
3.1. Sơ đồ quy trình.....	44
3.2. Xác định các bước thực hiện.....	44
3.2.1 Chuẩn bị dữ liệu	44
3.2.2 Xây dựng mô hình trí tuệ nhân tạo trong phân vùng lũ quét	54
3.2.3 Đánh giá sự phù hợp của mô hình	62
3.3. Xây dựng bản đồ phân vùng lũ quét theo kịch bản mưa.....	64
3.3.1 Xây dựng kịch bản mưa	64
3.3.2 Xây dựng bản đồ phân vùng nguy cơ bằng mô hình CNN.....	65
KẾT LUẬN	67

CHƯƠNG 1. DỮ LIỆU SỬ DỤNG, CHUẨN HÓA DỮ LIỆU VÀ XÂY DỰNG MÔ HÌNH TRÍ TUỆ NHÂN TẠO PHÂN VÙNG LŨ QUÉT

1.1. Đầu vào và cấu trúc dữ liệu

1.1.1.1 Chuẩn hóa dữ liệu

1. Dữ liệu input

Quá trình xây dựng mô hình bắt đầu với việc thu thập và tiền xử lý dữ liệu không gian từ các tệp raster, bao gồm các đặc trưng địa hình như độ cao, khoảng cách đến dòng chảy, độ dốc, chỉ số độ ẩm địa hình (TWI), chỉ số sức mạnh dòng chảy (SPI), và các đặc trưng khí tượng như lượng mưa tối đa trong các khung thời gian khác nhau (3 giờ, 6 giờ, 24 giờ). Dữ liệu không gian được lưu trữ dưới định dạng GeoTIFF, đòi hỏi các kỹ thuật xử lý đặc biệt để đảm bảo tính nhất quán và khả năng sử dụng trong học máy.

Một thách thức trong giai đoạn này là sự không đồng nhất về phân phối và thang đo của các đặc trưng. Các đặc trưng như độ cao (dem) hay lượng mưa (raster_max) thường có phân phối lệch, trong khi các đặc trưng như chỉ số địa hình (tpi) có thể chứa giá trị âm hoặc giá trị gần bằng không. Để giải quyết vấn đề này, các phương pháp chuẩn hóa dữ liệu được áp dụng một cách có chọn lọc. Cụ thể, các đặc trưng như độ cao và tỷ lệ thẩm nước được xử lý bằng kỹ thuật Robust Scaling, giúp giảm thiểu tác động của các giá trị ngoại lai. Trong khi đó, các đặc trưng có phân phối lệch mạnh như khoảng cách đến dòng chảy hoặc lượng mưa được chuyển đổi logarit để giảm độ lệch, sau đó được chuẩn hóa bằng MinMax Scaling để đưa về khoảng [0, 1]. Đối với các đặc trưng như độ cong mặt phẳng (planCurvature), chuẩn hóa Z-score được sử dụng để đảm bảo dữ liệu có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1. Việc áp dụng các phương pháp chuẩn hóa khác nhau cho từng loại đặc trưng không chỉ cải thiện hiệu suất mô hình mà còn phản ánh sự hiểu biết sâu sắc về đặc tính vật lý của từng biến. Christopher M. Bishop đã nói rằng việc xử lý dữ liệu đầu vào trước khi đưa vào học tập luôn luôn thuận lợi [1].

Một nguyên tắc nhỏ là các biến đầu vào nên có giá trị nhỏ, có thể nằm trong khoảng $0 \div 1$ hoặc được chuẩn hóa với giá trị trung bình là 0 và độ lệch chuẩn là 1. Tuy nhiên, nếu các giá trị của biến nhỏ (gần với 0 và 1) và phân phối dữ liệu bị hạn chế (độ lệch chuẩn lân cận 1) thì có thể không cần chia tỷ lệ dữ liệu. Điều này sẽ giúp mô hình đào tạo nhanh hơn và giảm khả năng mắc kẹt trong các tối ưu cục bộ [1].

Do vậy, toàn bộ số liệu đầu vào được chuẩn hóa theo nguyên tắc này và được thể hiện trong bảng sau:

Bảng 1. Chuẩn hóa các dữ liệu đầu vào cho mô hình

TT	Đặc trưng	Ký hiệu	Đơn vị	Phương pháp	Chuẩn hóa	Khoảng giá trị	Công thức Bước 1
1	Cao độ so với sông suối	eleStream	m	Cục bộ	Robust Scaling + MinMax	[0, 1]	$X' = \frac{X - \text{median}}{\text{IQR}}$

TT	Đặc trưng	Ký hiệu	Đơn vị	Phương pháp	Chuẩn hóa	Khoảng giá trị	Công thức Bước 1
2	Khoảng cách đến sông suối	disStream	m	Cục bộ	Log + MinMax	[0, 1]	$X' = \log(X + 1)$
3	Độ dốc	wSlope	độ	Lưu vực	MinMax	[0, 1]	$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$
4	Độ dốc lòng dẫn	stream Slope	m/m	Cục bộ	MinMax	[0, 1]	$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$
5	Chiều dài dòng chảy	flowLength	m	Cục bộ	Log + MinMax	[0, 1]	$X' = \log(X + 1)$
6	Diện tích lưu vực	area	m^2	Lưu vực	Log + MinMax	[0, 1]	$X' = \log(X + 1)$
7	Chỉ số âm địa hình	twi		Cục bộ	MinMax	[0, 1]	$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$
8	Chỉ số sức mạnh dòng chảy	spi		Cục bộ	Log + MinMax	[0, 1]	$X' = \log(X + 1)$
9	Chỉ số vị trí địa hình	tpi		Cục bộ	Z-score + MinMax	[0, 1]	$X' = \frac{X - \mu}{\sigma}$
10	Chỉ số NDVI	wNdvi		Lưu vực	MinMax	[0, 1]	$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$
11	Chỉ số CN	wCN		Lưu vực	MinMax	[0, 1]	$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$
12	Tốc độ thẩm thấu quân	wInfiRate	mm/hour	Lưu vực	Robust Scaling + MinMax	[0, 1]	$X' = \frac{X - \text{median}}{\text{IQR}}$
13	Cao độ địa hình	dem	m	Cục bộ	Robust Scaling + MinMax	[0, 1]	$X' = \frac{X - \text{median}}{\text{IQR}}$
14	Cao độ bình quân lưu vực	eleWatershed	m	Lưu vực	Robust Scaling + MinMax	[0, 1]	$X' = \frac{X - \text{median}}{\text{IQR}}$
15	Độ cong địa hình (theo hướng dốc)	profCurvature		Cục bộ	Z-score + MinMax	[0, 1]	$X' = \frac{X - \mu}{\sigma}$
16	Độ cong địa hình (phuong ngang)	planCurvature		Cục bộ	Z-score + MinMax	[0, 1]	$X' = \frac{X - \mu}{\sigma}$
17	Lượng mưa giờ lớn nhất	max_p_recip	mm	Lưu vực	Log & “÷10”	[0, 1]	$X' = \log(X + 1)$

TT	Đặc trưng	Ký hiệu	Đơn vị	Phương pháp	Chuẩn hóa	Khoảng giá trị	Công thức Bước 1
18	Lượng mưa 3 giờ lớn nhất	max_3 h_prec_ip	mm	Lưu vực	Log & “÷10”	[0, 1]	$X' = \log (X+1)$
19	Lượng mưa 6 giờ lớn nhất	max_6 h_prec_ip	mm	Lưu vực	Log & “÷10”	[0, 1]	$X' = \log (X+1)$
20	Lượng mưa 24 giờ lớn nhất	max_2 4h_precip	mm	Lưu vực	Log & “÷10”	[0, 1]	$X' = \log (X+1)$

Nếu phương pháp chuẩn hóa chỉ có 1 bước, thì thực hiện theo công thức ghi trong cột cuối, nếu có 2 bước, thì bước thứ hai là theo phương pháp MinMax (tham khảo chỉ số NDVI hoặc CN)

2. Dữ liệu dự đoán

Dữ liệu dự đoán là dữ liệu nhãn, được gán các giá trị từ 0 đến 4. Các giá trị này được đánh giá định lượng theo số (từ 0 đến 4) dựa trên đánh giá của nhóm nghiên cứu thực địa tại khu vực huyện Mù Cang Chải cho các suối chính ở một số xã điển hình cho trận lũ năm 2023.

Bảng 2. Nhãn mức độ lũ và ý nghĩa

TT	Nhãn	Giá trị nhãn	Ý nghĩa
0	Không có lũ	0	Các điểm thuộc mái dốc của núi, đỉnh núi, nơi không có tập trung dòng chảy hoặc có tập trung dòng chảy không đáng kể.
1	Lũ rất nhỏ	1	Dòng chảy trên suối không gây nguy hiểm đến các đối tượng, là dòng chảy phổ biến xuất hiện trên khu vực.
2	Lũ nhỏ	2	Dòng chảy trên suối là dòng chảy nhanh, hình thành do mưa nhưng không gây nguy hiểm đến các đối tượng.
3	Lũ trung bình	3	Dòng chảy trên sông suối là dòng chảy xiết, nằm trong lòng dẫn và an toàn để có thể đi qua các công trình cầu treo, không cuốn trôi các vật liệu lớn gây nguy hiểm cho cộng đồng sinh sống quanh khu vực
4	Lũ lớn	4	Dòng chảy trên sông suối là dòng chảy xiết, có cuốn trôi các vật liệu lớn hoặc nhỏ trong lòng dẫn, có thể tác động đến các công trình như cầu qua sông và các đối tượng cộng đồng nhà dân sinh sống xung quanh khu vực.

Dựa trên các tiêu chí này, nhóm nghiên cứu đã tiến hành thu thập thông tin tại các xã được đánh giá là có ảnh hưởng bởi lũ bao gồm các xã Khang Mao, Hồ Bón, Mò Dè, Lao Chải, Ché Tạo và Nậm Có. Chi tiết như sau:

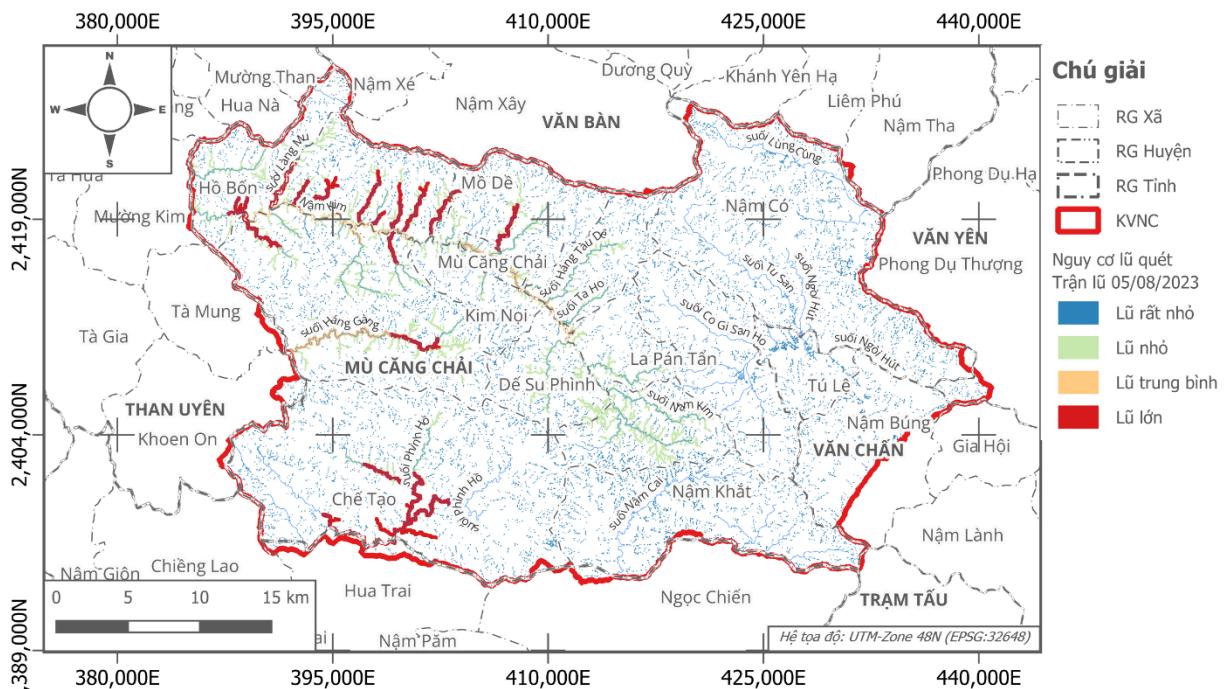
Bảng 3. Nhãn phân loại đánh giá theo tình hình mưa lũ thực tế tại địa phương

TT	Địa chỉ	Suối	Đánh giá đợt lũ 8/2023	Nhãn phân loại
1	Xã Hò Bón	Háng Nhù, Thông Gàu Bua, Làng Mu	Hầu hết các suối thuộc khu vực xã Hò Bón có dòng chảy mạnh, xiết. Người dân rất cảnh giác trong đợt mưa lũ đầu tháng 8/2023. Dòng chảy trên các suối này cuốn trôi nhiều vật liệu có đường kính lên tới 1m, phô biến là vài chục cm. Các khu vực nhánh suối đổ ra hướng suối Nậm Kim được đánh giá có nguy cơ cao, trong khi các nhánh suối đổ về phía suối Háng Đề Chu ghi nhận dòng chảy lũ bình thường (có lũ nhưng ít nguy hiểm)	Các nhánh suối chính như Hàng Nhù, Thông Gàu Bua, Làng Mu và lân cận suối được gán nhãn 4, các suối còn lại gán nhãn 3.
2	Xã Kham Mang	Suối Hàng B La Ha; Hàng B La Đề; Hàng Tàu Đề; Páo Sơ Dao; Tủa Mả Pán; Giàng Xua; Hàng Trán	Trong các nhánh suối này, khu vực Hàng B La Ha và Hàng B La Đề ghi nhận lũ lớn tại khu vực đổ vào suối chính Nậm Kim; trong khi các nhánh suối khác có lũ nhỏ hơn. Riêng suối Hàng Trán trong trận lũ này không có lũ lớn, không được coi là lũ quét.	Một số suối nhánh Hàng B La Ha và Hàng B La Đề gán nhãn 4, suối còn lại gán nhãn 3.
3	Xã Lao Chải	Suối Hàng Đề Sua, Hàng Gàng, Lao Chải	Riêng suối Hàng Đề Sua là suối được ghi nhận có lũ quét rất lớn và đổ trực tiếp vào suối Nậm Kim đợt mưa lũ này. Suối này cuốn các vật liệu lên tới 2-3m. Suối Hàng Gàng ở khu vực cuối Bản Hàng Gàng cũng xảy ra lũ rất lớn.	2 nhánh suối Hàng Đề Sua và Hàng Gàng được gán nhãn 4. Khu vực thượng nguồn bản Hàng Gàng và các suối khác đổ vào gán nhãn 3.
4	Xã Mò Đề	Suối Nà Hàng	Suối Nà Hàng thuộc xã Mò Đề là suối có lũ rất lớn trong đợt mưa lũ 8/2023. Các nhánh suối khác có ghi nhận lũ nhỏ hơn.	Suối Nà Hàng được gán nhãn 4, trong khi các nhánh suối khác đổ vào Nậm Kim gán nhãn 3.
5	Xã Chế Tạo	Suối Nậm Khắt, Nậm Khốt, Phình Hồ	Các nhánh suối đổ vào suối Phình Hồ đợt mưa này đều ghi nhận lũ lớn, đặc biệt là tại các bản Chế Tạo, Phú Vá, Tà Sung. Khu vực	Các nhánh suối Nậm Khắt, Nậm Khốt và Phình Hồ được gán nhãn 4, các nhánh suối đổ

TT	Địa chỉ	Suối	Đánh giá đợt lũ 8/2023	Nhân phân loại
			bản Nâ Háng không ghi nhận lũ lớn.	vào được gán nhãn 3.
6	Xã Nậm Cố		Các suối trên khu vực xã Nậm Cố không ghi nhận lũ lớn.	Các nhánh suối trên khu vực này được gán nhãn 2.

Ngoài ra theo mô tả, khu vực bản Mí Háng Táu (thuộc xã Púng Luông) cũng có ghi nhận lũ rất lớn, gây ảnh hưởng đến sinh hoạt của người dân. Các khu vực xã khác và các nhánh suối khác không ghi nhận lũ lớn. Nhìn chung, phần lớn các nhánh suối khu vực hạ du suối Nậm Kim đều ghi nhận lũ lên, trong khi đó, phía suối chính Ngòi Hút (đỗ về huyện Văn Yên) không có ghi nhận lũ lớn.

Ngoài các khu vực mô tả phía trên (chủ yếu ghi nhận lũ lớn), các khu vực còn lại được đánh nhăn 1 nếu nằm trên các sườn núi có độ dốc lớn (mái dốc núi), đỉnh núi và các khu vực ruộng bậc thang. Nhăn 2 được đánh cho các nhánh suối nhỏ và rất nhỏ ở thượng nguồn các khu vực không ghi nhận lũ lớn.



Hình 1. Kết quả điều tra các nhánh sông bị lũ quét trong trận lũ 05/08/2023 tại huyện Mù Cang Chải và phân loại nguy cơ lũ quét dựa trên đánh giá.

Mặc dù có dữ liệu mưa của trạm Mù Cang Chải theo giờ tại các trận lũ khác trước năm 2021, tuy nhiên, dữ liệu mưa tại một trạm không đủ đại diện cho một khu vực nhỏ bé, do đó, khó có thể đánh giá được chính xác lượng mưa sinh lũ của các trận lũ trước năm 2021. Từ năm 2021 trở đi, mật độ quan trắc mưa có thể được coi là tương đối tốt, do đó, nghiên cứu sử dụng dữ liệu năm 2023 (cho trận lũ xảy ra tại Hồ Bốn) làm cơ sở để xác định các điểm phân loại nguy cơ như đã trình bày phía trên.

1.1.1.2 Cấu trúc dữ liệu

Dữ liệu chính được tổ chức dưới dạng các tệp raster GeoTIFF, lưu trữ các đặc trưng như độ cao (dem), chỉ số ám địa hình (twi), lượng mưa (raster_max), và chỉ số NDVI (wNdvi). Mỗi tệp raster đại diện cho một đặc trưng, với các ô lưới (pixel) chứa giá trị số tương ứng với thuộc tính địa lý tại một vị trí cụ thể, được xác định bởi tọa độ (r, c). Độ phân giải không gian của các raster được đồng bộ hóa để đảm bảo tính nhất quán, với mỗi ô lưới thường đại diện cho một khu vực có kích thước cố định (trong nghiên cứu này, độ phân giải 12.5x12.5m được sử dụng). Dữ liệu nhãn, xác định các mức độ nguy cơ lũ lụt (rất thấp, thấp, trung bình, cao), cũng được lưu trữ dưới dạng raster, với các giá trị từ 1 đến 4 lớp để đơn giản hóa bài toán phân loại.

Để sử dụng trong học máy, dữ liệu raster được chuyển đổi thành bảng DataFrame của Pandas, trong đó mỗi hàng đại diện cho một điểm không gian với các cột bao gồm tọa độ (r, c), giá trị các đặc trưng (như eleStream, wSlope, raster_max), và nhãn nguy cơ lũ quét (được phân loại từ 1 đến 4). Cấu trúc này cho phép dễ dàng áp dụng các kỹ thuật tiền xử lý như chuẩn hóa (Robust Scaling, MinMax Scaling) và xử lý mất cân bằng lớp bằng SMOTE. Các đặc trưng được tổ chức thành bốn nhóm: địa hình (8 đặc trưng), thủy văn (6 đặc trưng), thực phủ (2 đặc trưng), và khí tượng (4 đặc trưng), tổng cộng 20 đặc trưng.

Cấu trúc dữ liệu này được thiết kế để tối ưu hóa cả hiệu quả tính toán và ý nghĩa vật lý. Các tệp GeoTIFF giữ được thông tin không gian, trong khi DataFrame hỗ trợ xử lý nhanh các thuật toán học máy. Hệ thống logging tích hợp ghi lại mọi bước xử lý, đảm bảo khả năng theo dõi và tái hiện. Cấu trúc dữ liệu này không chỉ đáp ứng yêu cầu của các mô hình như Random Forest hay LightGBM mà còn cho phép lưu trữ và xuất kết quả dự đoán dưới dạng raster, tích hợp dễ dàng vào các hệ thống GIS để phân tích và trực quan hóa nguy cơ lũ lụt.

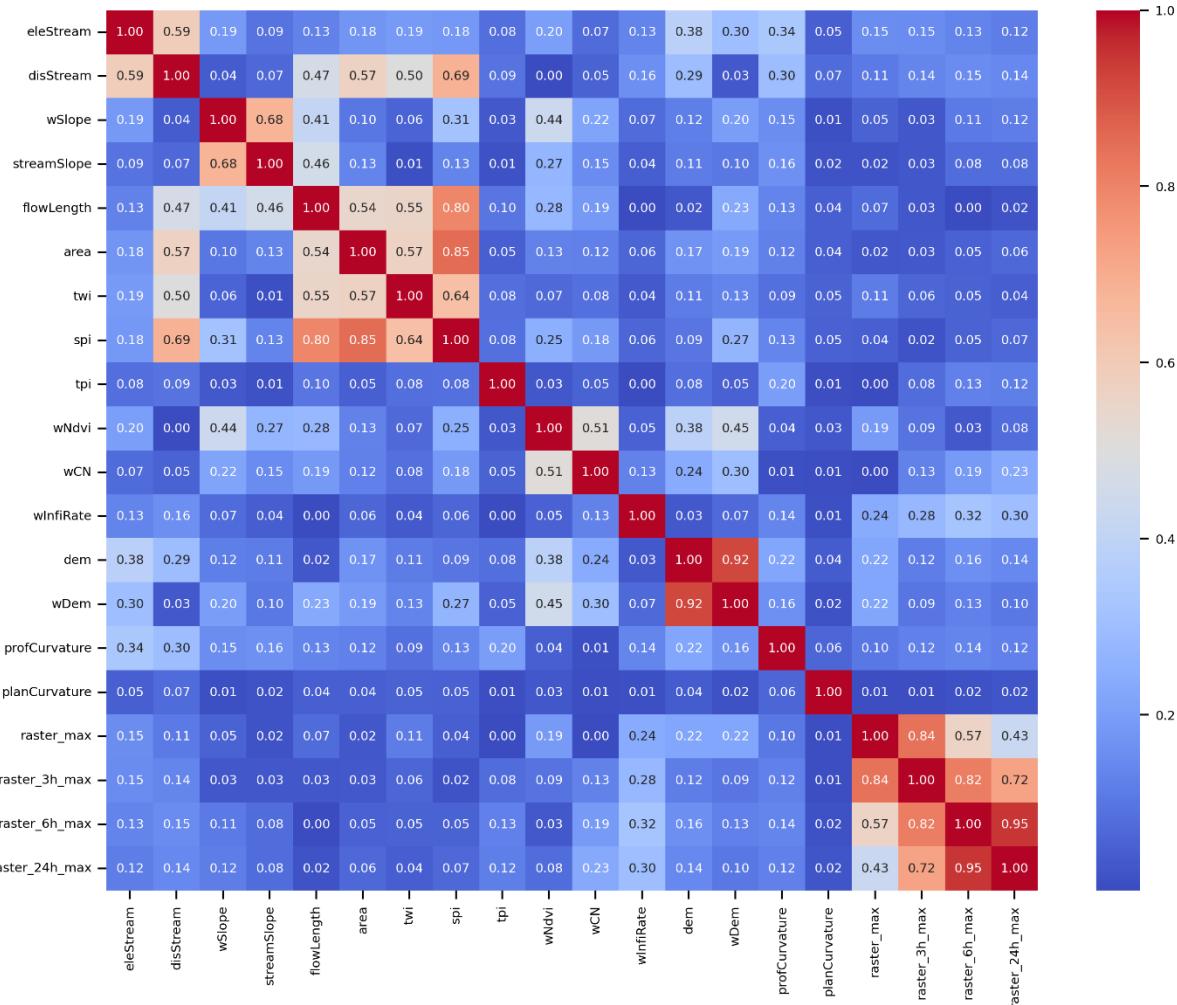
1.2. Xây dựng mô hình học máy

1. Lựa chọn đặc trưng

Lựa chọn đặc trưng là một bước quan trọng để giảm độ phức tạp của mô hình và cải thiện hiệu suất. Trong nghiên cứu này, hai kỹ thuật chính được sử dụng để giảm số lượng đặc trưng: loại bỏ các đặc trưng có tương quan cao và lựa chọn đặc trưng dựa trên điểm số thống kê.

Đầu tiên, ma trận tương quan được tính toán để xác định các đặc trưng có mức tương quan tuyệt đối lớn hơn 0,85. Các đặc trưng này bị loại bỏ để tránh hiện tượng đa cộng tuyến (multicollinearity), vốn có thể làm giảm hiệu quả của các mô hình như hồi quy logistic hoặc SVM. Việc loại bỏ các đặc trưng tương quan cao không chỉ giảm kích thước dữ liệu mà còn giúp mô hình tập trung vào các đặc trưng độc lập, mang lại thông tin phong phú hơn.

Tiếp theo, phương pháp SelectKBest với tiêu chí f_{classif} được sử dụng để chọn ra 18 đặc trưng quan trọng nhất từ tập hợp ban đầu. Kỹ thuật này đánh giá mức độ quan trọng của từng đặc trưng dựa trên mối quan hệ thống kê với biến mục tiêu, đảm bảo rằng các đặc trưng được giữ lại có khả năng phân biệt tốt giữa các lớp nguy cơ lũ quét. Danh sách các đặc trưng được chọn được lưu trữ để sử dụng trong các bước dự đoán sau này, đảm bảo tính nhất quán giữa huấn luyện và triển khai.



Hình 2. Ma trận tương quan các đặc trưng

Theo ma trận tương quan các đặc trưng của dữ liệu đưa vào mô hình học máy, cao độ bình quân lưu vực (wDem) có mức độ tương quan lớn với cao độ cửa ra (0,92), bên cạnh đó, lượng mưa lớn nhất 24 giờ cũng có tương quan rất lớn với lượng mưa lớn nhất 6 giờ. Do lũ quét thường xảy ra trong khoảng 6 giờ và thời gian tập trung dòng chảy của các lưu vực nhỏ sinh lũ quét cũng trong khoảng này, nhóm nghiên cứu loại bỏ 2/20 đặc trưng là cao độ bình quân lưu vực (wDem) và lượng mưa 24 giờ lớn nhất. Như vậy, 18/20 đặc trưng sẽ được đưa vào đánh giá và xây dựng mô hình học máy.

2. Xây dựng mô hình học máy

Năm mô hình học máy được triển khai để dự đoán nguy cơ lũ lụt: Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), LightGBM (LGBM),

và một mô hình kết hợp (ensemble) giữa RF và LGBM. Mỗi mô hình có những ưu điểm riêng, phù hợp với các đặc điểm khác nhau của bài toán.

a. Random Forest (RF)

Random Forest được cấu hình với các tham số chính để tận dụng khả năng xử lý dữ liệu không gian và chống quá khóp thông qua cơ chế tổng hợp cây quyết định. Các tham số được tối ưu hóa bao gồm:

- n_estimators: Số lượng cây quyết định trong rừng, được tìm kiếm trong khoảng từ 100 đến 200. Giá trị lớn hơn giúp cải thiện độ chính xác bằng cách giảm phương sai, nhưng tăng chi phí tính toán. Trong bài toán này, khoảng giá trị này được chọn để cân bằng giữa hiệu suất và thời gian huấn luyện.
- max_depth: Độ sâu tối đa của mỗi cây, với các giá trị được thử nghiệm là 15, 20, và 25. Giới hạn độ sâu giúp ngăn chặn quá khóp, đặc biệt khi dữ liệu không gian có tương quan cao giữa các đặc trưng như độ cao (dem) hoặc lượng mưa (raster_max).
- min_samples_split: Số lượng mẫu tối thiểu cần thiết để phân chia một nút, được tìm kiếm trong các giá trị 2 và 5. Tham số này kiểm soát độ chi tiết của cây, với giá trị lớn hơn giúp giảm nguy cơ quá khóp trên các lớp nguy cơ lũ lụt hiếm gặp.
- min_samples_leaf: Số lượng mẫu tối thiểu tại một nút lá, được thử nghiệm với các giá trị 1 và 3. Tham số này đảm bảo các lá cây không quá nhỏ, tăng tính tổng quát hóa của mô hình.

Random Forest còn được cấu hình với random_state=42 để đảm bảo tính tái lập, n_jobs=-1 để tận dụng tất cả các lõi CPU, và class_weight='balanced' để xử lý mất cân bằng lớp, đặc biệt quan trọng khi các lớp nguy cơ cao và nghiêm trọng có số lượng mẫu ít hơn. Kết quả xây dựng cho thấy bộ tham số tối ưu được xác định bao gồm: n_estimators là 120; max_depth là 25; min_samples_split là 2; min_samples_leaf là 3.

b. Support Vector Machine (SVM)

SVM được cấu hình để tận dụng khả năng phân loại phi tuyến thông qua kernel RBF, phù hợp với các bài toán có ranh giới phân loại phức tạp. Các tham số được tối ưu hóa bao gồm:

- C: Tham số điều chỉnh mức độ phạt đối với lỗi phân loại, được tìm kiếm trong phân phối đều từ 0.1 đến 20. Giá trị C lớn hơn cho phép mô hình tập trung vào việc phân loại chính xác các điểm dữ liệu, nhưng có thể dẫn đến quá khóp, đặc biệt với dữ liệu không gian có nhiễu.
- gamma: Tham số kiểm soát độ rộng của kernel RBF, được thử nghiệm với các giá trị 'scale' và 'auto'. Gamma ảnh hưởng đến mức độ ảnh hưởng của các điểm dữ liệu gần nhau, với giá trị nhỏ hơn phù hợp cho dữ liệu có phân phối không gian rộng.

- kernel: Được thử nghiệm với 'rbf' và 'linear'. Kernel RBF được ưu tiên để xử lý các mối quan hệ phi tuyến giữa các đặc trưng như chỉ số ẩm địa hình (twi) và lượng mưa.

SVM sử dụng random_state=42, probability=True để hỗ trợ dự đoán xác suất (cần thiết cho bỏ phiếu mềm trong mô hình kết hợp), và class_weight='balanced' để xử lý mất cân bằng lớp. Do tính phức tạp tính toán cao, số lần lặp trong tối ưu hóa siêu tham số được giảm xuống còn 2 (n_iter=2), giúp tiết kiệm thời gian mà vẫn đảm bảo hiệu suất. Kết quả xây dựng mô hình cho tham số tối ưu C là 7,59; gamma được lựa chọn là 'scale', kernel được xác định với 'linear'.

c. Logistic Regression (LR)

Logistic Regression được sử dụng như mô hình cơ sở, với các tham số được tối ưu hóa để đảm bảo khả năng diễn giải và hiệu quả trong phân loại đa lớp:

- C: Tham số điều chỉnh mức độ điều chuẩn hóa (regularization), được tìm kiếm trong phân phối đều từ 0.1 đến 20. Giá trị C nhỏ hơn tăng cường điều chuẩn hóa, giảm nguy cơ quá khớp trên các đặc trưng như độ dốc (wSlope) hoặc NDVI.
- penalty: Chỉ sử dụng L2 regularization để đảm bảo tính ổn định của mô hình, đặc biệt khi các đặc trưng có tương quan không gian.
- solver: Được thử nghiệm với 'lbfgs' và 'liblinear'. Solver 'lbfgs' phù hợp cho dữ liệu lớn, trong khi 'liblinear' hiệu quả cho các bài toán có số lượng đặc trưng giới hạn.

Mô hình được cấu hình với random_state=42, max_iter=1000 để đảm bảo hội tụ, n_jobs=-1 để tận dụng đa lõi, và class_weight='balanced' để xử lý mất cân bằng lớp. Các hệ số hồi quy cung cấp thông tin về độ quan trọng của đặc trưng, hữu ích trong việc diễn giải ảnh hưởng của các yếu tố như lượng mưa hoặc độ cao. Kết quả xây dựng mô hình cho tham số tối ưu C là 7,59 tương tự mô hình SVM, ngoài ra, tham số solver được xác định là 'lbfgs'.

d. LightGBM (LGBM)

LightGBM, một mô hình gradient boosting, được cấu hình để tận dụng tốc độ huấn luyện nhanh và khả năng xử lý dữ liệu lớn:

- num_leaves: Số lượng lá tối đa trong mỗi cây, được tìm kiếm trong khoảng từ 30 đến 70. Giá trị lớn hơn tăng độ phức tạp của mô hình, phù hợp với dữ liệu không gian có nhiều đặc trưng.
- learning_rate: Tốc độ học, được tìm kiếm trong phân phối đều từ 0.05 đến 0.15. Giá trị nhỏ hơn giúp mô hình học chậm và ổn định hơn, giảm nguy cơ quá khớp.
- n_estimators: Số lượng cây boosting, được tìm kiếm trong khoảng từ 100 đến 300. Giá trị này tương tự như n_estimators trong Random Forest, nhưng được tối ưu hóa cho cơ chế boosting.

- max_depth: Độ sâu tối đa của cây, được thử nghiệm với các giá trị 8, 10, và 12, giúp kiểm soát độ phức tạp và ngăn chặn quá khớp.

LightGBM sử dụng random_state=42, n_jobs=-1, và class_weight='balanced' để đảm bảo hiệu quả tính toán và xử lý mât cân bằng lớp. Mô hình này đặc biệt hiệu quả với dữ liệu không gian lớn nhờ cơ chế tối ưu hóa như histogram-based gradient boosting. Kết quả xây dựng mô hình cho tham số tối ưu num_leaves là 48; learning_rate là 0,11; n_estimators đạt 221 và max_depth đạt 12.

Việc LightGBM cho ra n_estimators lớn hơn nhiều so với RF (221 so với 120) thể hiện rõ thuật toán boosting của LightGBM, việc xây dựng thuật toán này là xây dựng tuần tự, mỗi cây phía sau sẽ sửa lỗi các cây phía trước nên cần nhiều cây nhỏ để cải thiện dàn mô hình tổng thể, trong khi đó, mô hình RF sử dụng thuật toán Bagging, các cây được huấn luyện độc lập và song song nên mỗi cây cần mạnh hơn và sâu hơn nhưng về tổng thể, số lượng cây này ít hơn mô hình LightGBM.

Độ sâu max_depth của hai mô hình cũng có sự khác biệt rõ rệt. Do RF cần mỗi cây phải đủ mạnh để phân loại một cách độc lập nên cần cây có nhiều tầng hơn (sâu hơn), do đó max_depth cũng lớn hơn (là 25), trong khi đó, LightGBM hoạt động hiệu quả với cây nông hơn (max_depth = 12) vì thông tin được tích lũy qua nhiều cây.

e. Mô hình kết hợp (Ensemble)

Mô hình kết hợp sử dụng cơ chế bỏ phiếu mềm (soft voting) để kết hợp dự đoán từ Random Forest và LightGBM, với các tham số được tối ưu hóa đồng thời cho cả hai mô hình thành phần:

- rf_n_estimators: Số lượng cây trong Random Forest, tìm kiếm trong khoảng 100 đến 200.
- rf_max_depth: Độ sâu tối đa của Random Forest, thử nghiệm với các giá trị 15 và 20.
- lgbm_num_leaves: Số lượng lá trong LightGBM, tìm kiếm trong khoảng 30 đến 50.
- lgbm_learning_rate: Tốc độ học của LightGBM, tìm kiếm trong khoảng 0.05 đến 0.15.
- lgbm_max_depth: Độ sâu tối đa của LightGBM, thử nghiệm với các giá trị 8 và 10.

Mô hình sử dụng voting='soft' để kết hợp xác suất dự đoán từ RF và LGBM, với trọng số cân bằng (0.5 cho mỗi mô hình) để đảm bảo đóng góp đồng đều. n_jobs=-1 được sử dụng để tối ưu hóa tính toán. Độ quan trọng của đặc trưng được tính bằng cách tổng hợp có trọng số từ RF và LGBM, cung cấp cái nhìn toàn diện về vai trò của các đặc trưng như lượng mưa hoặc chỉ số sức mạnh dòng chảy (spi).

Kết quả xây dựng mô hình cho tham số tối ưu rf_n_estimators đạt 174, nằm khoảng giữa mô hình RF và mô hình LGBM và nằm trên trung bình (trung bình đạt 170).

Rf_max_depth đạt 15 (cũng nằm giữa mô hình RF và LGBM độc lập). Trong khi đó, các đặc trưng mạnh của LGBM bao gồm num_leaves, learning_rate và max_depth đạt lần lượt là 48; 0,11 và 8.

3. Đánh giá mô hình học máy trong phân vùng lũ quét

a. Đánh giá hiệu suất bằng ma trận nhầm lẫn và đường cong ROC

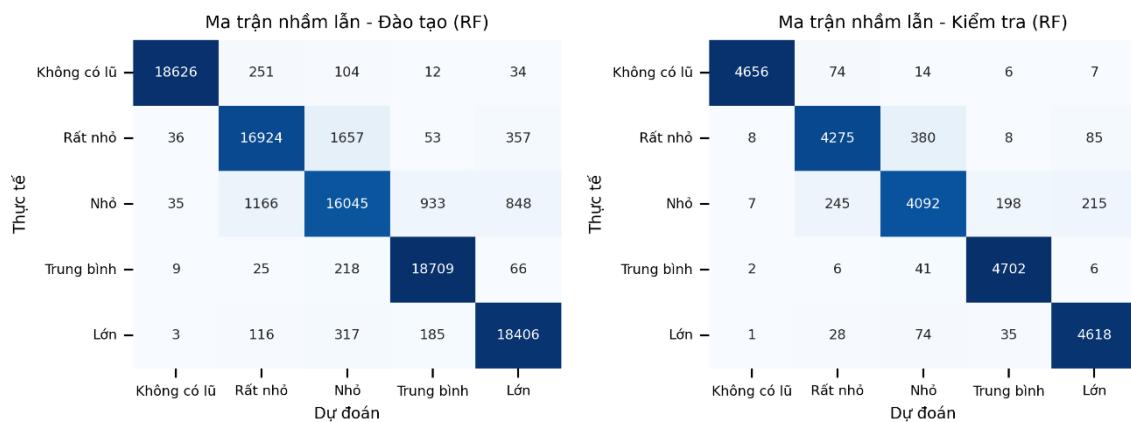
Ma trận nhầm lẫn (confusion matrix) là một công cụ quan trọng trong học máy và thống kê để đánh giá hiệu suất của mô hình phân loại. Nó hiển thị mối quan hệ giữa giá trị thực tế (actual values) và giá trị dự đoán (predicted values) của mô hình, giúp phân tích chi tiết cách mô hình hoạt động trên từng lớp (class).

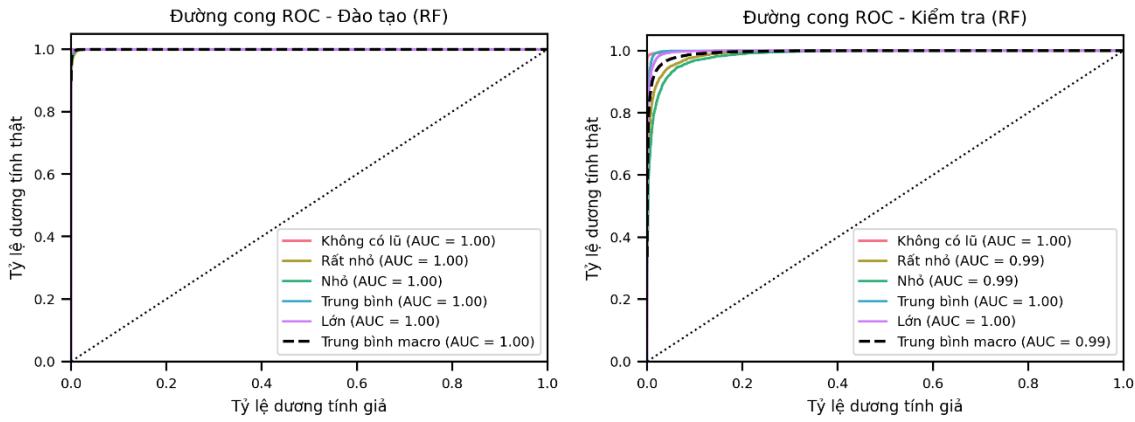
Đường cong ROC là một công cụ quan trọng để đánh giá hiệu suất của các mô hình phân loại nhị phân, với trục hoành là tỷ lệ dương tính giả (False Positive Rate - FPR) và trục tung là tỷ lệ dương tính thật (True Positive Rate - TPR). Diện tích dưới đường cong (AUC - Area Under Curve) được sử dụng để đo lường mức độ phân biệt của mô hình, với giá trị 1.0 cho thấy hiệu suất hoàn hảo và 0.5 cho thấy hiệu suất ngẫu nhiên.

Bảng 4. Bảng so sánh, đánh giá các mô hình đã xây dựng

Mô hình	Độ chính xác tổng thể (Accuracy)	Độ chính xác (Precision)	Độ nhạy (Recall)	F1 Score
rf	0,9395	0,9392	0,9395	0,9391
svm	0,6947	0,6891	0,6947	0,6908
lr	0,6897	0,6839	0,6897	0,6857
lgbm	0,9524	0,9523	0,9524	0,9522
ensemble	0,9326	0,9322	0,9326	0,9320

Mô hình RF:





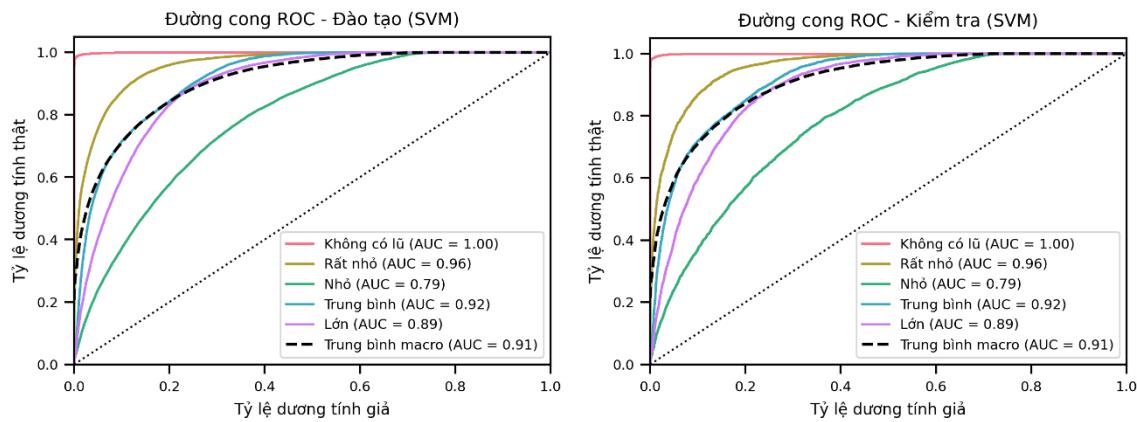
Hình 3. Ma trận nhầm lẫn và đường cong ROC mô hình RF

Random Forest (RF) đạt độ chính xác 93.95%, cho thấy khả năng phân loại tốt trên tập dữ liệu. Tuy nhiên, khi xem xét ma trận nhầm lẫn, RF gặp khó khăn trong việc phân biệt các mức độ lũ trung gian, đặc biệt là giữa "Lũ nhỏ" và "Lũ lớn".

- **Ưu điểm:**
 - Hiệu suất ổn định giữa tập đào tạo và kiểm tra, ít bị overfitting.
 - Xử lý tốt các trường hợp "Không có lũ" (97.9%) và "Lũ trung bình" (98.8%) do đặc trưng rõ ràng.
 - Phù hợp khi cần mô hình dễ giải thích (so với LGBM hoặc Ensemble).
- **Nhược điểm:**
 - Lớp "Lũ nhỏ" chỉ đạt 89.8% recall, với 4.7% bị nhầm thành "Lũ lớn" – một sai sót nguy hiểm trong cảnh báo lũ.
 - Không vượt trội ở bất kỳ lớp nào so với LGBM.

Mô hình SVM:

		Ma trận nhầm lẫn - Đào tạo (SVM)					Ma trận nhầm lẫn - Kiểm tra (SVM)					
		Không có lũ	Rất nhỏ	Nhỏ	Trung bình	Lớn	Không có lũ	Rất nhỏ	Nhỏ	Trung bình	Lớn	
Thực tế	Không có lũ	18664	207	91	34	31	Không có lũ	4664	62	20	4	7
	Rất nhỏ	58	14977	3037	239	716	Rất nhỏ	17	3752	736	45	206
Nhỏ	53	3895	7570	3887	3622	Nhỏ	11	989	1852	967	938	
Trung bình	14	59	2098	13527	3329	Trung bình	3	20	513	3398	823	
Lớn	4	497	3591	3661	11274	Lớn	0	137	878	885	2856	
Dự đoán						Dự đoán						

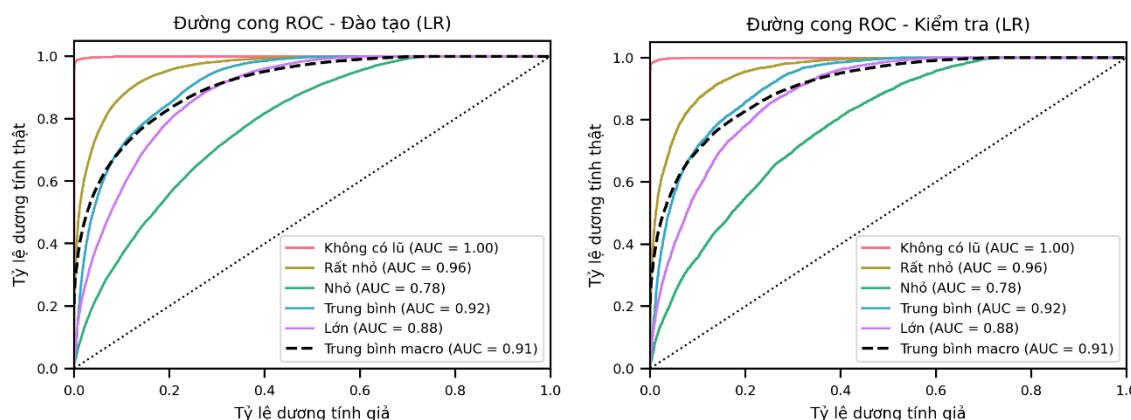


Hình 4. Ma trận nhầm lẩn và đường cong ROC mô hình SVM

Support Vector Machine (SVM) chỉ đạt 69.47% accuracy, thấp thứ 2 trong các mô hình. Ma trận nhầm lẩn cho thấy nó gần như không phân biệt được giữa các mức độ trung gian (“Lũ nhỏ”, “Lũ trung bình”, “Lũ lớn”). Nguyên nhân là bởi vì SVM là mô hình tuyến tính, trong khi ranh giới giữa các mức độ lũ có tính phi tuyến phức tạp. Lớp “Lũ nhỏ” bị nhầm lẩn nghiêm trọng: chỉ 39.5% được dự đoán đúng, phần lớn bị phân vào “Lũ lớn” (20.7%) hoặc “Lũ trung bình” (19.8%).

Mô hình LR:

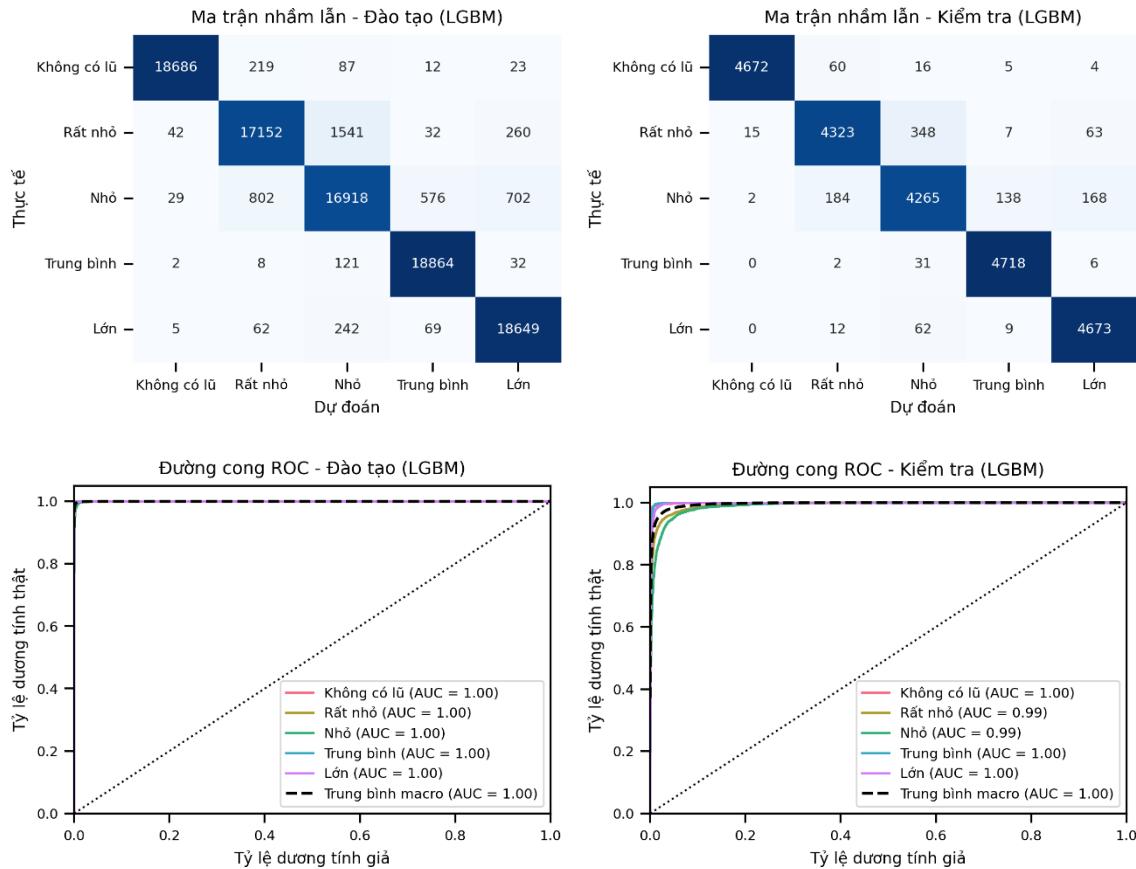
		Ma trận nhầm lẩn - Đào tạo (LR)					Ma trận nhầm lẩn - Kiểm tra (LR)					
		Không có lũ	Rất nhỏ	Nhỏ	Trung bình	Lớn	Không có lũ	Rất nhỏ	Nhỏ	Trung bình	Lớn	
Thực tế	Không có lũ	18665	223	79	25	35	4664	67	15	5	6	
	Rất nhỏ	52	14996	3007	247	725	15	3768	724	46	203	
Nhỏ	56	3888	7530	3787	3766		Nhỏ	13	997	1828	937	982
Trung bình	17	64	2286	13406	3254		Trung bình	3	23	563	3368	800
Lớn	5	488	3680	3902	10952		Lớn	0	140	891	950	2775
Dự đoán						Dự đoán						



Hình 5. Ma trận nhầm lẩn và đường cong ROC mô hình LR

Logistic Regression (LR) có hiệu suất thấp nhất và gần bằng SVM (68.97% accuracy), nhưng ma trận nhầm lẩn cho thấy nó còn tệ hơn trong phân biệt “Lũ lớn”: Chỉ 58.3% trường hợp “Lũ lớn” được dự đoán đúng, ~30% bị nhầm thành “Lũ trung bình”. Lớp “Lũ nhỏ” cũng chỉ đạt 38.6% recall. Vấn đề là mô hình LR quá đơn giản, không nắm bắt được quan hệ phi tuyến và tương quan phức tạp giữa các đặc trưng.

Mô hình LGBM:

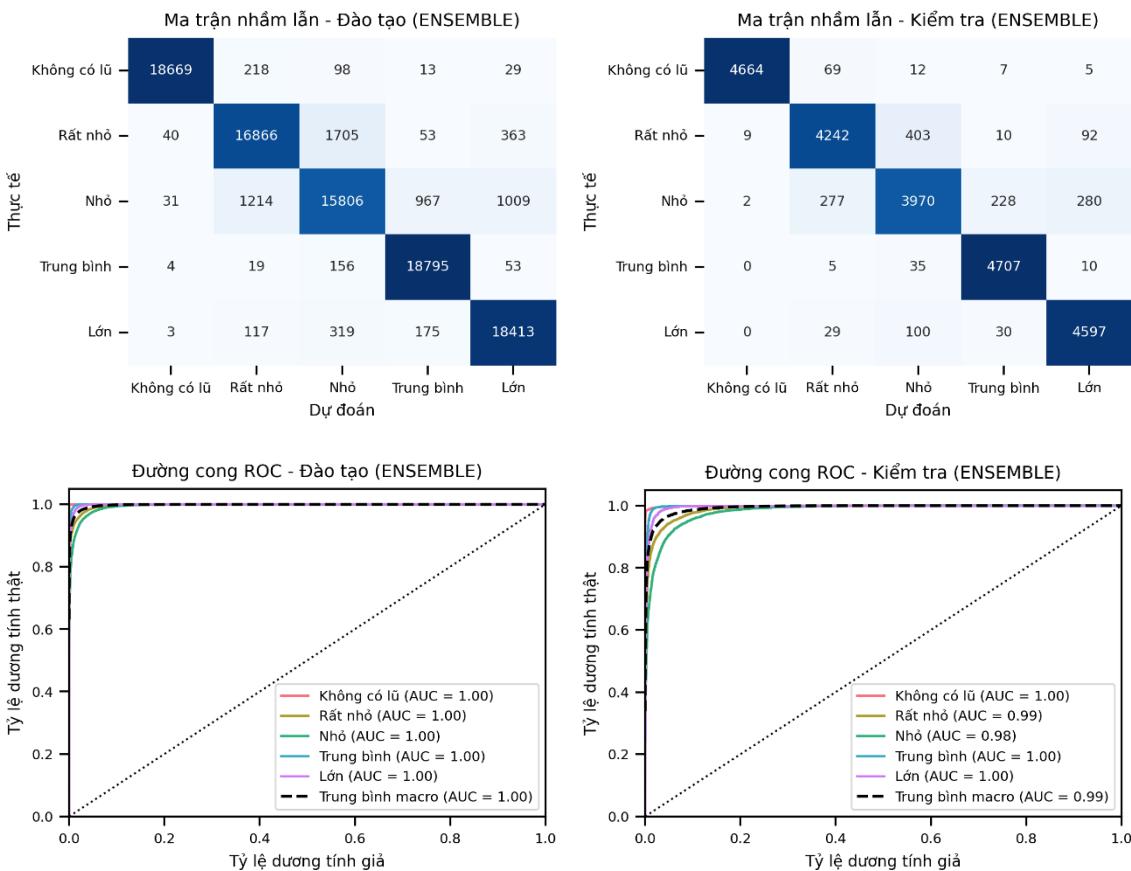


Hình 6. Ma trận nhầm lẩn và đường cong ROC mô hình LGBM

LightGBM (LGBM) là mô hình mạnh nhất, đạt 95.24% accuracy và F1-Score 95.22%.

Điểm mạnh của mô hình: mô hình đã phân loại chính xác “Lũ nhỏ” (93.6%) – tốt hơn hẳn RF (89.8%) và Ensemble (83.8%). Trong khi đó, mô hình cũng giảm thiểu nhầm lẩn như chỉ 3.7% “Lũ nhỏ” bị nhầm thành “Lũ lớn” (so với 4.7% của RF). Lý do là thuật toán Gradient Boosting giúp tối ưu hóa các trường hợp khó và xử lý tốt dữ liệu mất cân bằng.

Mô hình Ensemble:



Hình 7. Ma trận nhầm lẩn và đường cong ROC mô hình Ensemble

Ensemble đạt 93.26% accuracy, thấp hơn LGBM nhưng tốt hơn RF ở một số lớp (ví dụ: “Lũ rất nhỏ”). Ưu điểm của mô hình này là kết hợp sức mạnh của nhiều mô hình, giảm overfitting. Tuy nhiên, mô hình này không tốt bằng LGBM ở lớp “Lũ nhỏ” (83.8% recall so với 93.6% của LGBM), đồng thời nó phức tạp hơn mô hình RF trong khi mức độ cải thiện không đáng kể.

Đánh giá chung:

Dựa trên tất cả các chỉ số (Accuracy, Precision, Recall, F1 Score), các mô hình có thể được xếp thành hai nhóm chính:

- **Nhóm hiệu suất cao (Strong Performers):**

- LightGBM (LGBM): Với các chỉ số gần như hoàn hảo (Accuracy: 0.9524, Precision: 0.9523, Recall: 0.9524, F1 Score: 0.9522), LGBM vượt trội hơn hẳn. Điều này cho thấy khả năng tổng quát hóa (generalization) của mô hình này rất tốt, ít bị overfitting và có thể đưa ra dự đoán đáng tin cậy trên dữ liệu mới.
- Random Forest (RF): Hiệu suất rất đáng nể (Accuracy: 0.9395, Precision: 0.9392, Recall: 0.9395, F1 Score: 0.9391). RF là một lựa chọn mạnh mẽ và thường rất ổn định.

- Ensemble: Cũng thể hiện hiệu suất cao (Accuracy: 0.9326, Precision: 0.9322, Recall: 0.9326, F1 Score: 0.9320), chứng tỏ sức mạnh của việc kết hợp các mô hình khác nhau.
- **Nhóm hiệu suất thấp (Weak Performers):**
 - Support Vector Machine (SVM): Hiệu suất giảm mạnh (Accuracy: 0.6947, Precision: 0.6891, Recall: 0.6947, F1 Score: 0.6908).
 - Logistic Regression (LR): Thấp nhất trong số các mô hình được đánh giá (Accuracy: 0.6897, Precision: 0.6839, Recall: 0.6897, F1 Score: 0.6857).

Mô hình LGBM nổi bật với hiệu suất vượt trội, đạt độ chính xác tổng thể lên tới 95.24% và F1 Score 95.22%. Khi phân tích sâu ma trận nhầm lẫn, LGBM thể hiện khả năng phân biệt xuất sắc giữa các mức độ lũ, đặc biệt là ở các trường hợp ranh giới như giữa “Lũ rất nhỏ” và “Lũ nhỏ” hay giữa “Lũ nhỏ” và “Lũ lớn”. Điều này cho thấy kiến trúc gradient boosting của LGBM phù hợp để xử lý các mối quan hệ phi tuyến phức tạp trong dữ liệu lũ lụt.

Random Forest và mô hình Ensemble cho hiệu suất tương đối tốt nhưng kém hơn LGBM, với độ chính xác lần lượt là 93.95% và 93.26%. Các mô hình này tuy ổn định nhưng gặp khó khăn đáng kể trong việc phân biệt các mức độ lũ trung gian. Đặc biệt, lớp “Lũ nhỏ” thường bị nhầm lẫn với “Lũ lớn” với tỷ lệ khoảng 4-5%, điều này có thể gây hậu quả nghiêm trọng trong cảnh báo thực tế. Sự nhầm lẫn không đối xứng này gợi ý rằng thang phân loại hiện tại có thể chưa tối ưu hoặc cần bổ sung thêm các đặc trưng phân biệt rõ hơn giữa các mức độ.

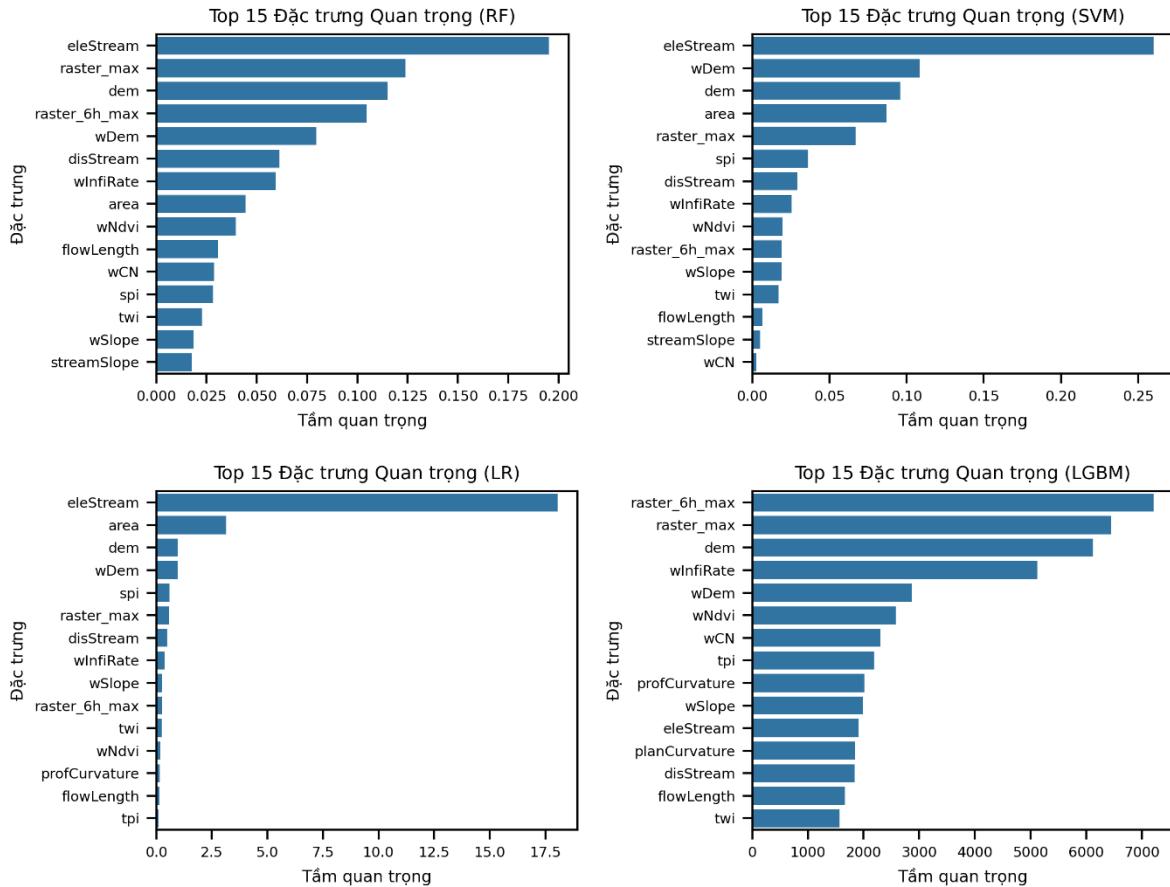
Hai mô hình SVM và Logistic Regression cho kết quả kém hơn hẳn với độ chính xác chỉ khoảng 69%. Điều này phản ánh hạn chế của các phương pháp tuyến tính khi xử lý bài toán có ranh giới phức tạp như phân loại mức độ lũ. Các mô hình này gần như không phân biệt được giữa các mức độ trung gian, đặc biệt là giữa “Lũ nhỏ”, “Lũ trung bình” và “Lũ lớn”. Hiệu suất thấp này cho thấy chúng không phù hợp để triển khai trong hệ thống cảnh báo lũ thực tế.

Bài toán phân loại lũ thể hiện rõ tính phân cấp trong độ khó - các mức độ ở hai đầu (“Không có lũ” và “Lũ trung bình”) dễ phân loại hơn hẳn (đạt >98% ở hầu hết mô hình) so với các mức trung gian. Điều này có thể do các trường hợp cực trị thường có đặc trưng rõ ràng, trong khi các mức độ trung gian có nhiều điểm tương đồng. Sự suy giảm hiệu suất ở các lớp trung gian phản ánh đúng thách thức thực tế trong việc định lượng mức độ lũ, vốn thường không có ranh giới rõ ràng.

b. Tầm quan trọng của các yếu tố/đặc trưng

Dựa trên các đặc trưng đầu vào phục vụ phân loại nguy cơ lũ quét, mỗi mô hình học máy có nhìn nhận khác biệt về tầm quan trọng của các đặc trưng đầu vào. Các hình vẽ dưới đây thể hiện 15/20 đặc trưng quan trọng nhất của mỗi mô hình. Tầm quan trọng là một thước đo định lượng mức độ mà một đặc trưng đóng góp vào việc cải thiện độ chính

xác hoặc giảm sai lệch trong dự đoán của mô hình. Nó cho biết đặc trưng nào có vai trò lớn trong việc phân biệt các lớp hoặc dự đoán giá trị đầu ra. eleStream (chênh lệch độ cao so với sông suối) và các đặc trưng mưa (3 giờ lớn nhất, 6 giờ lớn nhất và mưa giờ lớn nhất) luôn nằm trong top 5 của tất cả các mô hình, cho thấy chúng là yếu tố quyết định chính trong việc dự đoán nguy cơ. Các đặc trưng như wlnfiRate (chỉ số tốc độ thấm) và disStream (khoảng cách đến sông suối) cũng xuất hiện thường xuyên, phản ánh vai trò của môi trường tự nhiên.



Hình 8. Tầm quan trọng của các yếu tố đầu vào trong mô hình học máy

- RF và SVM: Cả hai mô hình nhấn mạnh các đặc trưng mưa và eleStream, với thang đo quan trọng cao hơn (lên đến 0.20-0.25), cho thấy chúng phản ánh đúng về nguy cơ lũ quét. Lượng mưa gây ra lũ và lũ quét ảnh hưởng đến các điểm trũng/thấp, nơi có chênh lệch độ cao so với sông suối thấp.
- LGBM: Ưu tiên cao độ địa hình (dem) và các đặc trưng mưa với thang đo lớn (lên đến 4000), phản ánh khả năng xử lý dữ liệu phức tạp tốt hơn. Tuy nhiên việc thể hiện cao độ của điểm (giá trị cao độ địa hình) có liên quan trực tiếp đến lũ quét là một trong những nhận định chưa tốt của mô hình này. Một ví dụ rất cụ thể là trong cùng một điều kiện về lưu vực và lượng mưa, nếu lưu vực đặt ở độ cao lớn hơn hay độ cao thấp hơn thì nguy cơ lũ quét tại các điểm trên lưu vực đó vẫn không thay đổi.

- LR: Có thang đo thấp hơn (lên đến 10) và tập trung vào eleStream, area, spi, cho thấy mô hình đơn giản hơn và ít nhạy với các đặc trưng phụ. Tuy nhiên, việc đánh giá thấp về lượng mưa trong việc tạo thành nguy cơ lũ quét là một điều chưa phù hợp ở mô hình này.

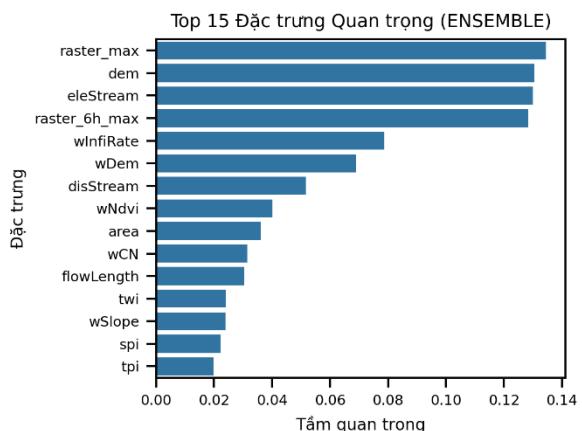
Mô hình ensemble (kết hợp của mô hình RF và mô hình LGBM) cho kết quả tổng hợp cũng tương tự 2 mô hình đối với các yếu tố chính (bao gồm lượng mưa và chênh lệch cao độ so với sông/suối gần nhất). Việc xác định lượng mưa 1 giờ lớn nhất có vai trò quan trọng nhất theo đánh giá của nhóm nghiên cứu là phù hợp để phân loại nguy cơ lũ quét.

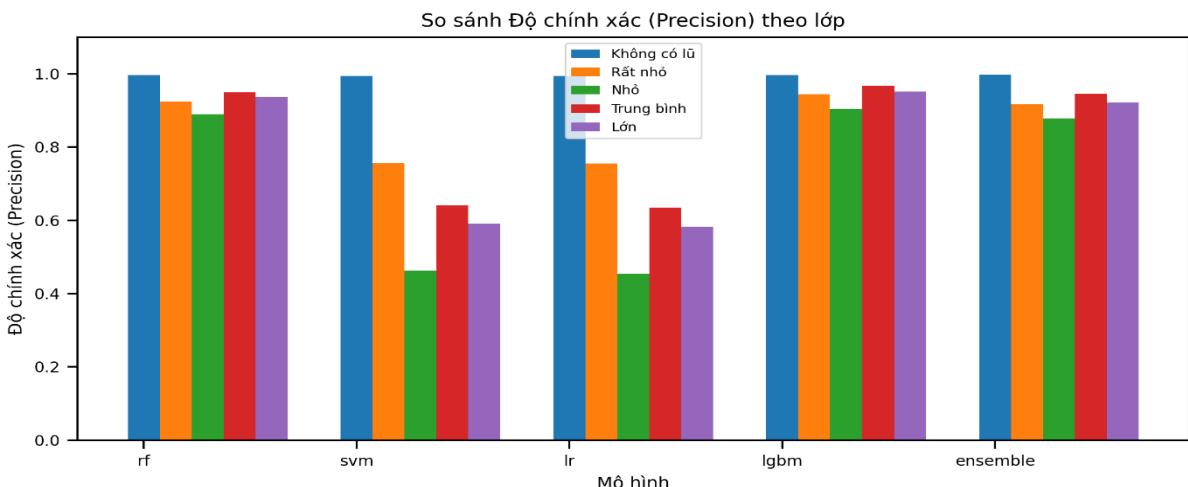
c. Các chỉ số khác

Độ chính xác (Precision), độ nhạy (Recall) và F1 Score là ba chỉ số quan trọng trong việc đánh giá hiệu suất của các mô hình phân loại, đặc biệt trong các bài toán mà dữ liệu có thể không cân bằng giữa các lớp. Độ chính xác đo lường tỷ lệ các dự đoán dương tính thực sự đúng, thể hiện khả năng mô hình tránh dự đoán sai các mẫu âm tính thành dương tính. Độ nhạy, ngược lại, đánh giá khả năng mô hình phát hiện đúng tất cả các mẫu dương tính thực sự, rất quan trọng khi việc bỏ sót các trường hợp dương tính có thể gây hậu quả nghiêm trọng. F1 Score là trung bình điều hòa của độ chính xác và độ nhạy, cung cấp một cái nhìn cân bằng về hiệu suất tổng thể, đặc biệt hữu ích khi cần ưu tiên cả hai khía cạnh.

Phân tích precision cho thấy các mô hình gặp khó khăn trong việc phân biệt chính xác giữa các mức độ lũ trung gian. Trong khi LGBM và RF duy trì được precision khá cao (0.92-0.97) cho tất cả các lớp, thì SVM và LR thể hiện sự suy giảm rõ rệt (xuống còn 0.6-0.7) ở các lớp “Lũ nhỏ” và “Lũ trung bình”. Điều này phản ánh một thực tế là ranh giới giữa các mức độ lũ này chưa thực sự rõ ràng trong dữ liệu, khiến các mô hình tuyến tính như SVM và LR khó phân biệt chính xác.

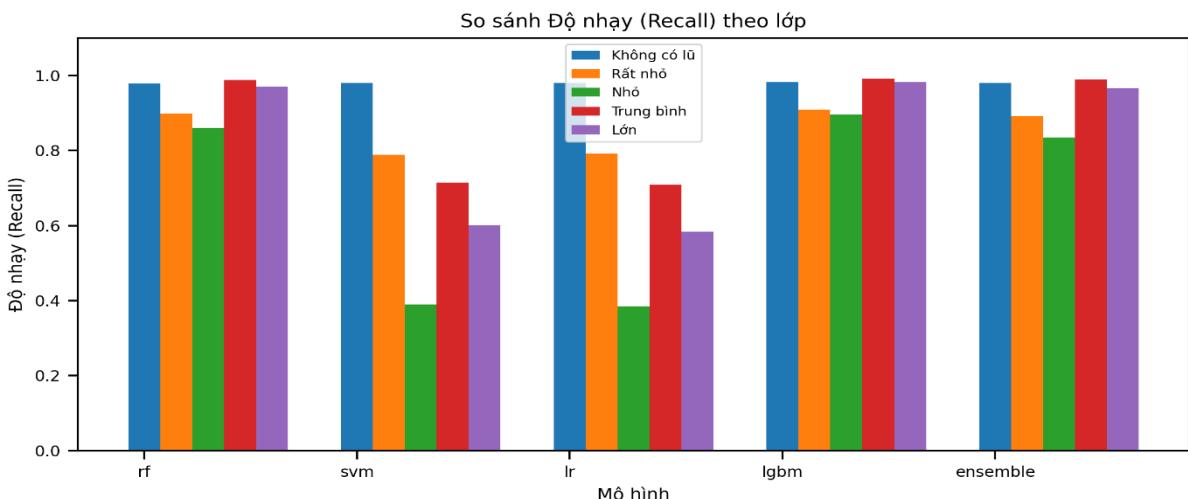
Sự khác biệt về hiệu suất giữa các mô hình cũng cho thấy tính phức tạp của bài toán. Trong khi các mô hình đơn giản như LR gần như không thể nắm bắt được sự khác biệt tinh vi giữa các lớp, thì phương pháp boosting như LGBM lại tỏ ra vượt trội nhờ khả năng học các đặc trưng phi tuyến phức tạp. Tuy nhiên, ngay cả với LGBM, precision ở lớp “Lũ nhỏ” vẫn thấp hơn so với các lớp khác, cũng có cho nhận định về sự chồng lấn đáng kể giữa các mức độ lũ liền kề trong không gian đặc trưng.





Hình 9. Độ chính xác của các mô hình học máy theo các lớp dữ liệu

Các mô hình thể hiện sự khác biệt rõ rệt về khả năng phát hiện chính xác từng mức độ lũ. LGBM tiếp tục dẫn đầu với độ nhạy recall ấn tượng trên 0.95 cho hầu hết các lớp, chứng tỏ khả năng bắt gặp gần như toàn bộ các trường hợp lũ thực tế. Đặc biệt ở lớp “Lũ trung bình”, LGBM đạt recall gần như hoàn hảo (0.99), trong khi các mô hình khác như SVM và LR chỉ đạt khoảng 0.6-0.7, cho thấy sự vượt trội trong việc nhận diện các mẫu thuộc lớp này.

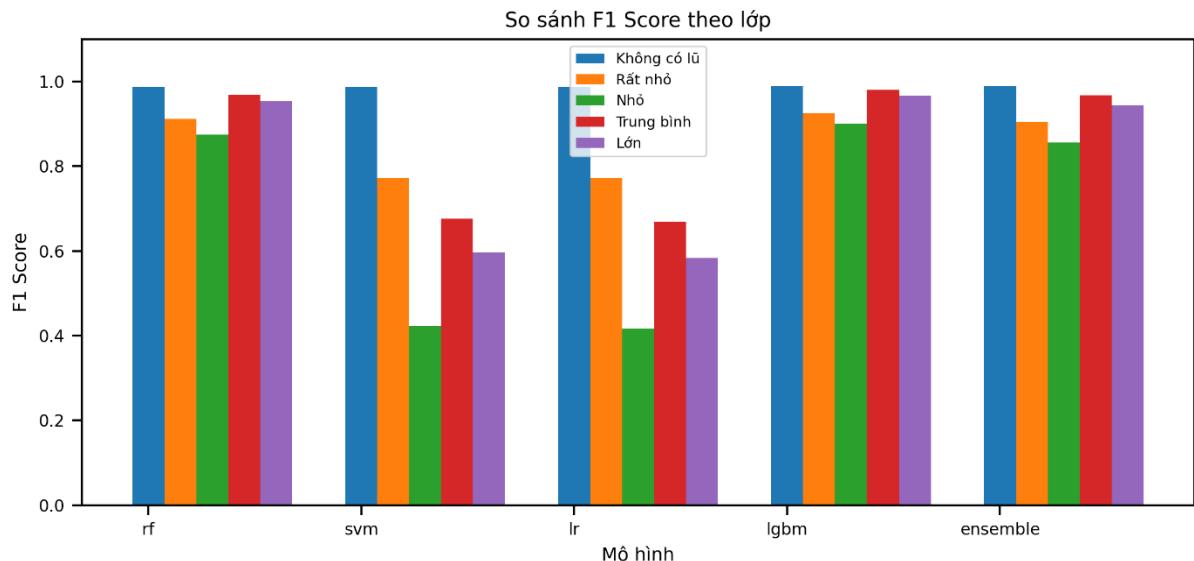


Hình 10. Độ nhạy của các mô hình học máy theo các lớp dữ liệu

Tuy nhiên, tất cả mô hình đều gặp khó khăn nhất định với lớp “Lũ nhỏ”, nơi recall dao động từ 0.83 (Ensemble) đến 0.94 (LGBM). Khoảng cách này phản ánh sự chưa rõ ràng trong định nghĩa ranh giới giữa các mức độ lũ liền kề, khiến ngay cả những mô hình mạnh nhất cũng bỏ sót một phần các trường hợp thực tế. Điều đáng chú ý là trong khi RF và Ensemble có hiệu suất tương đương nhau, thì khoảng cách giữa chúng với LGBM lại khá lớn (5-10%), nhấn mạnh ưu thế của phương pháp boosting trong việc xử lý các trường hợp khó.

Chỉ số F1 Score cho thấy bức tranh toàn diện về hiệu suất cân bằng giữa Precision và Recall của các mô hình. LGBM một lần nữa khẳng định vị thế dẫn đầu với F1 Score gần

như hoàn hảo (0.95-0.97) trên tất cả các lớp, đặc biệt ấn tượng ở lớp “Lũ trung bình” đạt 0.98. Điều này chứng tỏ LGBM không chỉ dự đoán chính xác mà còn ít bỏ sót các trường hợp thực tế.



Hình 11. F1 Score của các mô hình học máy theo các lớp dữ liệu

Các mô hình RF và Ensemble cho kết quả khá tốt với F1 Score dao động 0.91-0.94, nhưng vẫn thua kém LGBM từ 3-5%, đặc biệt ở lớp “Lũ nhỏ”. Trong khi đó, SVM và LR tiếp tục thể hiện hạn chế rõ rệt với F1 Score chỉ đạt 0.6-0.7 ở các lớp trung gian. Khoảng cách hiệu suất này càng củng cố nhận định về sự chênh lệch đáng kể giữa các mức độ lũ liền kề trong không gian đặc trưng.

d. Thời gian dự báo

Thời gian dự đoán cũng là yếu tố cần xem xét để đánh giá hiệu suất về mặt tốc độ xử lý, một yếu tố quan trọng khi triển khai các mô hình máy học trong thực tế, đặc biệt trong các ứng dụng yêu cầu phản hồi nhanh hoặc xử lý dữ liệu lớn.

Trước tiên, SVM nổi bật với thời gian dự đoán lâu nhất, lên tới 6 giờ 5 phút. Điều này phản ánh nhược điểm có hổng của SVM khi xử lý trên tập dữ liệu lớn hoặc không gian đặc trưng phức tạp. SVM thường yêu cầu tính toán khoảng cách giữa các điểm dữ liệu và tìm ranh giới phân cách tối ưu, dẫn đến độ phức tạp tính toán cao, đặc biệt nếu sử dụng kernel phi tuyến tính như RBF. Thời gian này có thể là một trở ngại lớn trong các kịch bản cần phản hồi nhanh, chẳng hạn như hệ thống phát hiện rủi ro theo thời gian thực, khiến SVM kém phù hợp nếu tốc độ là ưu tiên hàng đầu.

Ngược lại, LR và Ensemble cho thấy hiệu suất ấn tượng với thời gian dự đoán lần lượt là 8 phút. LR (Logistic Regression), với bản chất là một mô hình tuyến tính, có độ phức tạp tính toán thấp, chủ yếu dựa trên phép nhân ma trận và tối ưu hóa hàm mất mát, nên không ngạc nhiên khi nó hoạt động nhanh. Ensemble, dù thường kết hợp nhiều mô hình và có thể tồn tại nguyên hơn, vẫn đạt thời gian 8 phút, cho thấy nó đã được tối ưu hóa tốt, có thể nhờ vào việc sử dụng các mô hình đơn giản hoặc cơ chế song song hóa

trong quá trình dự đoán. Điều này khiến cả hai mô hình trở thành lựa chọn hợp lý trong các ứng dụng yêu cầu cân bằng giữa tốc độ và hiệu suất.

RF và LGBM, với thời gian dự đoán 10 phút, nằm ở mức trung bình trong nhóm. RF (Random Forest) hoạt động dựa trên tập hợp nhiều cây quyết định, và thời gian dự đoán phụ thuộc vào số lượng cây cũng như độ sâu của chúng. Mặc dù 10 phút là khá nhanh so với SVM, nó vẫn chậm hơn LR và Ensemble, có thể do RF cần tổng hợp kết quả từ nhiều cây. LGBM (Light Gradient Boosting Machine), dù được biết đến với tốc độ cao nhờ cơ chế histogram-based và xử lý dữ liệu phân loại, cũng mất 10 phút, có thể do kích thước tập dữ liệu hoặc số lượng vòng lặp boosting. Dù vậy, thời gian này vẫn cho thấy RF và LGBM là các lựa chọn khả thi khi cần tốc độ tương đối nhanh mà vẫn đảm bảo hiệu suất tốt, đặc biệt trong các bài toán phức tạp hơn.

Nhìn chung, nếu xét về mặt tốc độ, LR và Ensemble là hai mô hình hiệu quả nhất, chỉ mất 8 phút, phù hợp cho các ứng dụng cần phản hồi nhanh. RF và LGBM, với 10 phút, vẫn nằm trong ngưỡng chấp nhận được, đặc biệt khi cân nhắc rằng chúng thường mang lại độ chính xác cao hơn trong các bài toán phức tạp. Tuy nhiên, SVM, với thời gian 6 giờ 5 phút, rõ ràng không phù hợp cho các kịch bản yêu cầu tốc độ, và chỉ nên được sử dụng khi độ chính xác là ưu tiên tuyệt đối và tài nguyên tính toán không bị giới hạn. Để tối ưu hóa, có thể cân nhắc giảm kích thước dữ liệu, tinh chỉnh siêu tham số (như giảm số lượng cây trong RF hoặc sử dụng kernel đơn giản hơn cho SVM), hoặc triển khai song song hóa để cải thiện tốc độ, đặc biệt với các mô hình như SVM.

e. Đánh giá tổng hợp

Bảng 5. Bảng tổng hợp các kết quả đánh giá cho các mô hình học máy

Mô hình	Accuracy	Precision	Recall	F1 Score	Thời gian dự đoán	Đánh giá chung
RF	0.9395	0.9392	0.9395	0.9391	10 phút	Hiệu suất cao, tốc độ nhanh, đáng tin cậy.
SVM	0.6947	0.6891	0.6947	0.6908	6 giờ 5 phút	Hiệu suất thấp, tốc độ rất chậm, không khuyến nghị.
LR	0.6897	0.6839	0.6897	0.6857	8 phút	Hiệu suất thấp, tốc độ nhanh, không cạnh tranh.
LGBM	0.9524	0.9523	0.9524	0.9522	10 phút	Hiệu suất cao nhất, tốc độ hợp lý, khuyến nghị.
Ensemble	0.9326	0.9322	0.9326	0.9322	8 phút	Hiệu suất tốt, tốc độ nhanh, lựa chọn tốt.

Nhận xét:

RF đạt hiệu suất cao nhất với độ chính xác, độ nhạy, và F1 Score lần lượt là 0.9395, 0.9392, và 0.9391, trong khi thời gian dự đoán là 10 phút, khá nhanh so với các mô hình khác. LGBM và ensemble cũng thể hiện hiệu suất tốt, với LGBM đạt 0.9524 (Accuracy, Recall) và 0.9522 (F1 Score), còn ensemble đạt 0.9326 (Accuracy, Recall) và 0.9322 (F1 Score), cả hai đều có thời gian dự đoán là 10 phút và 8 phút, cho thấy sự cân bằng giữa hiệu suất và tốc độ. Ngược lại, SVM và LR có hiệu suất thấp hơn đáng kể, với SVM đạt 0.6947 (Accuracy, Recall) và 0.6908 (F1 Score), nhưng thời gian dự đoán rất

dài (6 giờ 5 phút), khiến nó không hiệu quả về mặt tốc độ. LR có các chỉ số tương tự SVM (Accuracy, Recall: 0.6897; F1 Score: 0.6857) nhưng nhanh hơn nhiều với 8 phút, dù vẫn kém về hiệu suất.

Về mặt tổng quan, LGBM nổi bật nhất khi xét cả hiệu suất và tốc độ, với các chỉ số cao nhất và thời gian dự đoán hợp lý. RF và ensemble cũng là những lựa chọn tốt, đặc biệt khi cần cân bằng giữa hiệu suất cao và thời gian chấp nhận được. SVM tỏ ra không phù hợp trong trường hợp này do thời gian dự đoán quá dài mà hiệu suất lại thấp, trong khi LR dù nhanh hơn SVM nhưng không đủ cạnh tranh về độ chính xác. Khuyến nghị sử dụng LGBM nếu ưu tiên hiệu suất cao và tốc độ hợp lý. Nếu cần một mô hình đơn giản hơn mà vẫn hiệu quả, RF là lựa chọn khả thi. Ensemble cũng đáng cân nhắc nếu muốn kết hợp ưu điểm của nhiều mô hình, dù hiệu suất không vượt trội bằng LGBM.

1.3. Xây dựng mô hình học sâu

1. Lựa chọn đặc trưng

Không giống mô hình học máy, mô hình học sâu có thể tự lựa chọn các đặc trưng theo thuật toán để đưa vào dự đoán, do đó, không cần phải loại bỏ các đặc trưng khác khi sử dụng mô hình học sâu. Các yếu tố không gian xung quanh được lựa chọn bằng việc thử dần các tham số. Nghiên cứu lựa chọn phương pháp thử dần cho các vùng lân cận từ 3x3 đến 11x11, kết quả lựa chọn được mô hình CNN được lấy vùng lân cận 5x5 và mô hình DNN được lấy đặc trưng lân cận 7x7 cho ra sự dự đoán tốt nhất (chỉ số Accuracy tốt nhất).

a. Mô hình CNN

Trong quy trình xử lý dữ liệu cho CNN, dữ liệu đầu vào được tổ chức dưới dạng các vùng lân cận (patch) kích thước 5x5, trích xuất từ các tệp raster địa lý. Mỗi tệp raster tương ứng với một tham số trong danh sách được nêu ở Bảng 1, bao gồm 16 tham số cơ bản (như độ cao địa hình, khoảng cách đến dòng chảy, chỉ số độ ẩm địa hình, chỉ số công suất dòng chảy, độ cong địa hình...) và 4 tham số lượng mưa (lượng mưa giờ lớn nhất, lượng mưa 3 giờ, 6 giờ, và 24 giờ lớn nhất). Tổng cộng, có 20 tham số, và mỗi vùng lân cận 5x5 tạo ra một mảng dữ liệu có kích thước 5x5x20 (chiều cao x chiều rộng x số kênh). Mỗi kênh (channel) đại diện cho một tham số, ví dụ, kênh đầu tiên có thể là độ cao địa hình, kênh thứ hai là khoảng cách đến dòng chảy, v.v.

Tất cả 20 tham số được giữ lại mà không loại bỏ bất kỳ tham số nào. Lý do là CNN được thiết kế để xử lý dữ liệu không gian và có khả năng tự động trích xuất các mẫu không gian phức tạp từ các vùng lân cận. Việc giữ nguyên toàn bộ tham số đảm bảo rằng mô hình có thể khai thác mọi thông tin không gian có sẵn, từ các mẫu đơn giản như độ dốc địa hình đến các mẫu phức tạp hơn như sự kết hợp giữa độ ẩm và lượng mưa. Các giá trị dữ liệu cũng đã được chuẩn hóa tương tự mô hình học máy, do đó, mô hình này mang tính kế thừa dữ liệu từ các mô hình học máy đã được xây dựng trước.

Kích thước vùng lân cận 5×5 được chọn để cân bằng giữa việc cung cấp đủ thông tin không gian và giảm chi phí tính toán. Một vùng lân cận 5×5 bao gồm 25 điểm dữ liệu cho mỗi kênh, cung cấp một cửa sổ không gian đủ lớn để mô hình nhận diện các mẫu cục bộ như độ dốc, độ cong, hoặc sự thay đổi của lượng mưa trong một khu vực nhỏ. Trong bối cảnh nghiên cứu, một vùng 5×5 đại diện cho một khu vực có diện tích hơn $300m^2$ (tương ứng độ phân giải $12,5m$), đủ để phát hiện các yếu tố như dòng chảy tập trung hoặc vùng trũng.

Nếu sử dụng kích thước lớn hơn như 9×9 (81 điểm dữ liệu) hoặc 11×11 (121 điểm dữ liệu), lượng thông tin không gian sẽ tăng lên, cho phép mô hình nắm bắt các mẫu không gian ở quy mô lớn hơn, chẳng hạn như sự phân bố lượng mưa trên một khu vực rộng hơn hoặc các đặc điểm địa hình phức tạp hơn. Tuy nhiên, điều này cũng làm tăng chi phí tính toán và có thể dẫn đến việc bao gồm các thông tin không cần thiết, đặc biệt nếu các mẫu liên quan đến lũ lụt chủ yếu xuất hiện ở quy mô cục bộ. Ngược lại, nếu sử dụng kích thước nhỏ hơn như 3×3 , mô hình có thể bỏ qua các mẫu không gian quan trọng do cửa sổ quá nhỏ. Kích thước 5×5 được chọn như một sự cân bằng hợp lý, phù hợp với dữ liệu không gian có độ phân giải vừa phải và yêu cầu tính toán hiệu quả, đặc biệt, nó cũng phù hợp với khả năng tính toán của máy tính được sử dụng trong nghiên cứu này (do mô hình CNN cần một cấu hình tương đối lớn để vận hành).

Các tham số như độ cao địa hình, khoảng cách đến dòng chảy, và chỉ số độ ẩm địa hình cung cấp thông tin về cấu trúc địa hình, trong khi các tham số lượng mưa phản ánh điều kiện thời tiết. Sự kết hợp của các tham số này trong một vùng lân cận 5×5 cho phép mô hình CNN khai thác các mối quan hệ không gian, chẳng hạn như cách lượng mưa tích lũy ở các khu vực trũng hoặc gần dòng chảy. Việc giữ nguyên 20 tham số đảm bảo rằng mô hình có thể phát hiện các mẫu phức tạp, như sự tương tác giữa độ dốc và lượng mưa, mà không cần loại bỏ thông tin trước. Chuẩn hóa từng kênh giúp mô hình tập trung vào các mẫu tương đối (relative patterns) thay vì bị chi phối bởi các giá trị tuyệt đối lớn (như độ cao hàng nghìn mét so với lượng mưa vài trăm milimet).

b. Mô hình DNN

Quy trình xử lý dữ liệu cho DNN khác biệt cơ bản so với CNN do bản chất của DNN, vốn không được thiết kế để xử lý trực tiếp dữ liệu không gian như các mảng 3D. Thay vào đó, DNN yêu cầu dữ liệu đầu vào dạng bảng, với mỗi hàng đại diện cho một điểm dữ liệu và mỗi cột là một đặc trưng. Để đáp ứng yêu cầu này, nghiên cứu đã trích xuất các đặc trưng thống kê từ các vùng lân cận kích thước 7×7 quanh mỗi điểm dữ liệu. Cụ thể, cho mỗi tham số trong danh sách 20 tham số, bốn đặc trưng thống kê được tính toán: giá trị trung bình, giá trị tối thiểu, giá trị tối đa, và độ lệch chuẩn. Ví dụ, từ tham số độ cao địa hình, năm đặc trưng được tạo ra là Giá trị trung bình, độ lệch chuẩn, giá trị nhỏ nhất, giá trị lớn nhất, trung vị của mỗi tham số. Điều này dẫn đến tối đa 100 đặc trưng (20 tham số \times 5 đặc trưng thống kê).

Lý do sử dụng đặc trưng không gian (như vùng lân cận 5x5 của CNN) là vì DNN không có khả năng tự động trích xuất các mẫu không gian từ dữ liệu mảng. Thay vào đó, DNN dựa vào các đặc trưng được tính toán trước, như các giá trị thống kê, để biểu diễn thông tin tổng quát về khu vực lân cận. Các đặc trưng này được tổ chức thành một bảng, trong đó mỗi hàng chứa tọa độ (hàng, cột) của điểm dữ liệu và các cột chứa các đặc trưng thống kê (như trung bình độ cao, tối đa lượng mưa 24 giờ).

Mặc dù lựa chọn 7x7, nhưng các giá trị đặc trưng chỉ là 1 giá trị (ví dụ như trung bình của 49 ô), khác hoàn toàn với mô hình CNN nếu nhận vùng lân cận là 7x7 thì có 49 tham số được đưa vào học tập. Đây là sự khác biệt rất lớn và cũng là lợi thế của mô hình DNN, sự khác biệt này nằm ở chỗ mô hình DNN đã đơn giản hóa được các đặc trưng đầu vào dựa vào đặc trưng không gian, làm cho dữ liệu sạch hơn và chất lượng hơn. Trong khi đó, mô hình CNN với quá nhiều tham số có thể dẫn đến việc học kém hơn nếu không đủ số lượng mẫu hoặc các nhãn đầu ra không phân định một cách định lượng hoàn toàn.

2. Xây dựng mô hình học sâu

a. Mô hình CNN

Lựa chọn vùng không gian lân cận trong mô hình CNN:

Việc thiết lập mô hình CNN trong mã nguồn được xây dựng với mục tiêu khai thác triệt để các đặc trưng không gian từ dữ liệu raster, vốn là các bản đồ địa hình và lượng mưa được biểu diễn dưới dạng patch 5x5. Kiến trúc CNN được thiết kế dựa trên các khối residual, một kỹ thuật tiên tiến giúp giảm thiểu vấn đề vanishing gradient khi huấn luyện các mạng sâu. Mỗi khối residual bao gồm ba tầng tích chập (Conv2D) với kích thước kernel lần lượt là 1x1, 3x3, và 1x1, cho phép giảm số kênh trước khi xử lý không gian và khôi phục chiều sâu kênh sau đó, tối ưu hóa tính toán mà vẫn giữ được thông tin quan trọng.

Chuẩn hóa theo lô (BatchNormalization) được áp dụng sau mỗi tầng tích chập để ổn định phân phối đầu ra, giảm sự phụ thuộc vào giá trị ban đầu của tham số và tăng tốc hội tụ. Hàm kích hoạt LeakyReLU với alpha=0.1 được chọn thay vì ReLU để tránh vấn đề nơ-ron “chết”, đặc biệt phù hợp với dữ liệu địa hình có nhiều giá trị gần 0. Cơ chế attention được tích hợp thông qua kết hợp GlobalAveragePooling2D và giảm chiều kênh, giúp mô hình tập trung vào các vùng có nguy cơ lũ cao, chẳng hạn như các khu vực trũng hoặc gần dòng chảy.

Để chống quá khớp, SpatialDropout2D với tỷ lệ 0.3 được sử dụng để ngắt kết nối không gian ngẫu nhiên, thay vì dropout thông thường, vì nó phù hợp hơn với dữ liệu hình ảnh. L2 regularization với hệ số 0.005 được áp dụng trên các tầng tích chập và dày đặc, giúp hạn chế độ lớn của trọng số, đặc biệt quan trọng khi dữ liệu có thể chứa nhiều từ các tệp raster. Dữ liệu đầu vào được chuẩn hóa bằng cách tính trung bình và độ lệch chuẩn của từng kênh trong patch, với các giá trị NaN hoặc vô cực được thay thế bằng giá trị trung bình để đảm bảo tính nhất quán.

Việc sử dụng patch 5×5 thay vì kích thước lớn hơn là một lựa chọn hợp lý để cân bằng giữa thông tin không gian và chi phí tính toán, nhưng có thể hạn chế khả năng nắm bắt các mẫu không gian ở quy mô lớn hơn. Hàm mất mát focal loss với $\text{gamma}=2.0$ và $\text{alpha}=0.25$ được chọn để giải quyết vấn đề mất cân bằng lớp, tập trung vào các lớp nguy cơ lũ cao, vốn thường hiếm gặp trong dữ liệu thực tế. Tối ưu hóa sử dụng Adam với lịch trình học CosineDecayRestarts thay vì ReduceLROnPlateau, cho phép điều chỉnh tốc độ học một cách linh hoạt, giảm nguy cơ kẹt ở điểm tối ưu cục bộ trong các vòng lặp huấn luyện dài.

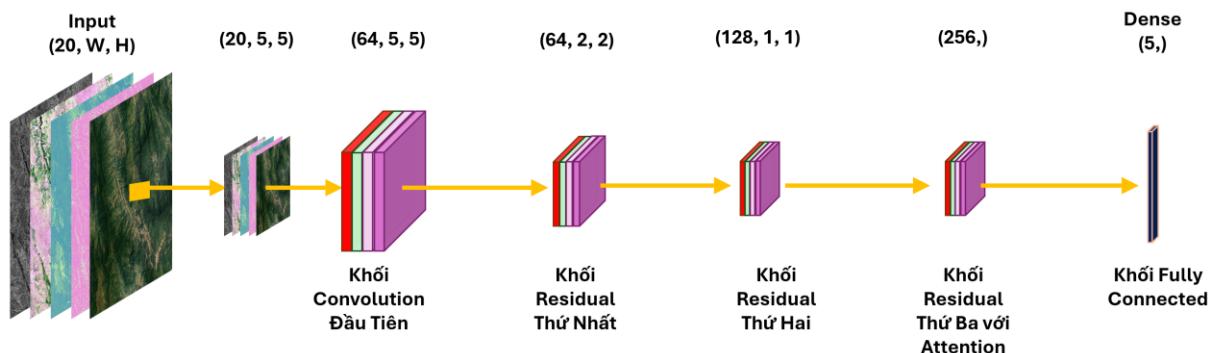
Xây dựng mô hình CNN:

Khối Convolution Đầu Tiên - Trích Xuất Đặc Trung Cơ Bản

Khối này đóng vai trò như một bộ cảm biến đầu tiên, nhận đầu vào là tensor $5 \times 5 \times 20$ chứa 16 đặc trưng địa hình cơ bản (độ cao, độ dốc, chỉ số thực vật, etc.) và 4 đặc trưng mưa tại mỗi pixel. Lớp Conv2D với 64 filters sẽ học cách kết hợp thông tin đa kênh này để tạo ra 64 feature maps khác nhau, mỗi map phản ánh một khía cạnh khác nhau của mối quan hệ giữa địa hình và lượng mưa. BatchNormalization đảm bảo quá trình học ổn định, trong khi LeakyReLU giúp mô hình có thể học được các mối quan hệ phi tuyến phức tạp giữa các yếu tố môi trường. SpatialDropout2D ngăn chặn overfitting bằng cách loại bỏ ngẫu nhiên một số feature maps hoàn chỉnh.

Khối Residual Thứ Nhất - Học Mẫu Hình Địa Phương

Hai residual blocks đầu tiên tập trung vào việc học các mẫu hình địa phương trong khung cửa sổ 5×5 . Residual connections cho phép mô hình học được cả thông tin chi tiết và thông tin tổng quát, điều quan trọng khi phân tích nguy cơ lũ vì cần xem xét cả các yếu tố địa hình nhỏ (như độ dốc cục bộ) và các yếu tố lớn hơn (như vị trí trong lưu vực). MaxPooling2D giảm kích thước từ 5×5 xuống 2×2 sẽ tạo ra một phiên bản tóm tắt của thông tin địa hình, tương tự như việc nhìn vùng nghiên cứu từ độ cao lớn hơn để nắm bắt các đặc điểm tổng quát.



Hình 12. Cấu trúc các khối thuật toán trong mô hình CNN

Khối Residual Thứ Hai - Tích Hợp Thông Tin Đa Tỷ Lệ

Khối này nâng số kênh lên 128, cho phép mô hình học được nhiều mẫu hình phức tạp hơn về mối quan hệ giữa địa hình và nguy cơ lũ. Với kích thước không gian giảm xuống 2×2 , mô hình tập trung vào việc tích hợp thông tin từ các vùng lân cận để hiểu được bối

cánh rộng hơn. Điều này đặc biệt quan trọng trong dự báo lũ vì nguy cơ lũ tại một điểm không chỉ phụ thuộc vào điều kiện tại chính điểm đó mà còn phụ thuộc vào toàn bộ vùng lưu vực xung quanh.

Khối Residual Thứ Ba với Attention - Tập Trung Vào Yếu Tố Quan Trọng

Khối cuối cùng nâng số kênh lên 256 để nắm bắt được các mẫu hình phức tạp nhất về nguy cơ lũ. Cơ chế attention đóng vai trò như một “bộ não” của mô hình, tự động xác định đâu là những yếu tố quan trọng nhất trong việc quyết định mức độ nguy hiểm của lũ. Dual pooling (average và max) cho phép attention mechanism xem xét cả giá trị trung bình (xu hướng chung) và giá trị cực đại (điểm nguy hiểm nhất) của mỗi feature map. Sigmoid gating tạo ra trọng số từ 0 đến 1 để nhấn mạnh hoặc giảm tầm quan trọng của từng đặc trưng, giống như cách chuyên gia về lũ sẽ chú ý nhiều hơn đến một số yếu tố nhất định khi đánh giá nguy cơ.

Khối Fully Connected - Quyết Định Phân Loại Cuối Cùng

GlobalAveragePooling2D chuyển đổi feature maps 2D thành vector 1D, tạo ra một “bản tóm tắt” toàn cục của tất cả thông tin đã học được. Hai lớp Dense với 128 và 64 neurons đóng vai trò như bộ não quyết định, tích hợp tất cả thông tin đã được xử lý để đưa ra quyết định về mức độ nguy hiểm của lũ. Regularization (L2, Dropout, BatchNorm) đảm bảo mô hình không bị overfit và có thể tổng quát hóa tốt trên dữ liệu mới. Lớp output cuối cùng với 5 neurons và softmax activation tương ứng với 5 mức độ nguy hiểm lũ, từ “Không có lũ” đến “Lũ lớn”, cho ra xác suất thuộc về từng lớp.

Toàn bộ kiến trúc này mô phỏng quá trình tiếp cận cho sự huấn luyện xây dựng mô hình trí tuệ nhân tạo: từ việc quan sát chi tiết các yếu tố địa hình và khí hậu cục bộ, đến việc tích hợp thông tin từ vùng rộng hơn, cuối cùng tập trung vào những yếu tố quan trọng nhất để đưa ra quyết định về mức độ nguy hiểm của lũ.

b. Mô hình DNN

Mô hình DNN được thiết lập để tận dụng các đặc trưng thống kê được trích xuất từ cửa sổ lân cận 7×7 , phù hợp với các bài toán cần phân tích đặc trưng tổng quát hơn là mẫu không gian. Kiến trúc DNN bao gồm ba tầng dày đặc với số nơ-ron giảm dần (256, 128, 64), tạo ra một cấu trúc hình tháp để giảm dần chiều không gian đặc trưng và tập trung vào các mẫu quan trọng.

Lớp Đầu Vào - Tiếp Nhận Thông Tin Tổng Hợp

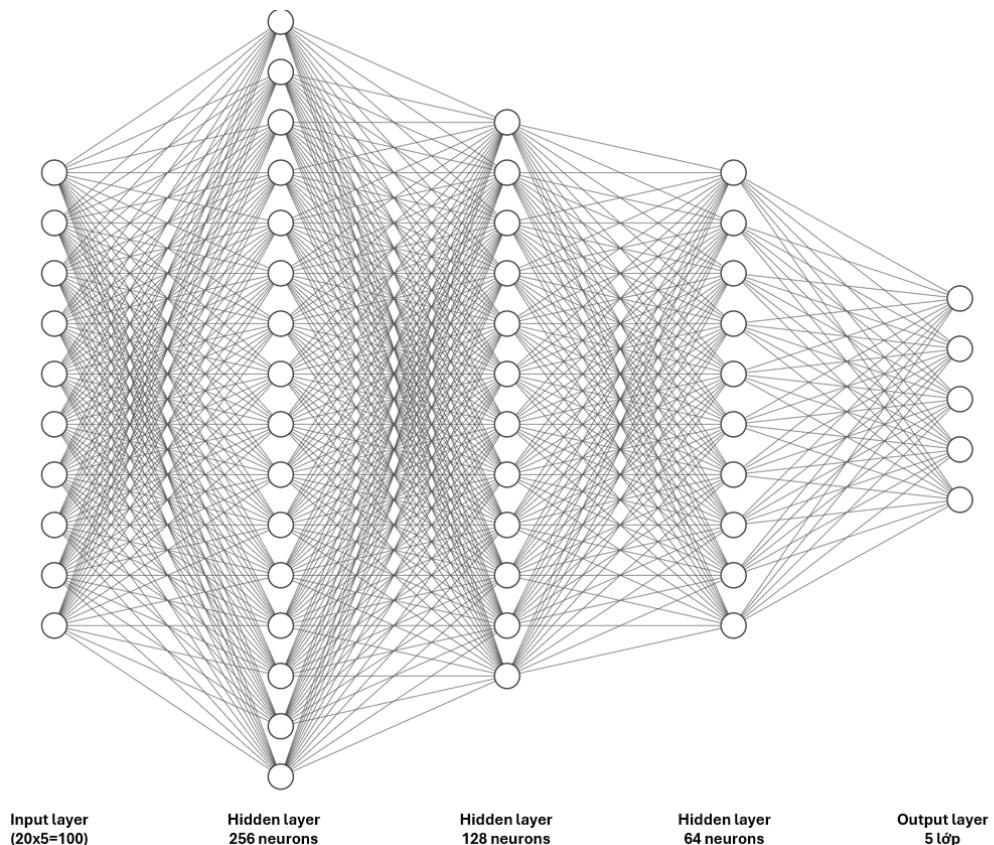
Mô hình DNN nhận đầu vào là vector một chiều với kích thước `input_dim`, chứa các đặc trưng đã được tổng hợp và trích xuất từ vùng nghiên cứu. Khác với CNN xử lý dữ liệu không gian 2D, DNN làm việc với các chỉ số thống kê tổng hợp như giá trị trung bình, độ lệch chuẩn, giá trị min/max của 16 yếu tố địa hình và 4 yếu tố mưa trong khu vực. Cách tiếp cận này phản ánh phương pháp truyền thống trong thủy văn học, nơi các chuyên gia thường sử dụng các chỉ số đặc trưng của lưu vực như độ dốc trung bình, chỉ số độ ẩm địa hình (TWI), hay lượng mưa tối đa để đánh giá nguy cơ lũ.

Lớp Dense Thứ Nhất (256 neurons) - Mở Rộng Không Gian Đặc Trung

Lớp Dense đầu tiên với 256 neurons đóng vai trò như một bộ “khuếch đại thông tin”, chuyển đổi từ không gian đặc trưng ban đầu lên không gian có 256 chiều. Điều này cho phép mô hình tạo ra nhiều tổ hợp phi tuyến khác nhau của các đặc trưng đầu vào, giống như việc một chuyên gia thủy văn xem xét không chỉ từng yếu tố riêng lẻ mà còn các mối quan hệ tương tác giữa chúng. Ví dụ, mối quan hệ giữa độ dốc và lượng mưa, hay giữa chỉ số thực vật và khả năng thẩm nước của đất. BatchNormalization đảm bảo các activation không bị bão hòa, trong khi LeakyReLU cho phép học được cả mối quan hệ âm và dương. Dropout 30% ngăn chặn overfitting bằng cách buộc mô hình không phụ thuộc quá nhiều vào một số neurons cụ thể.

Lớp Dense Thứ Hai (128 neurons) - Tinh Lọc Thông Tin Quan Trọng

Lớp 128 neurons hoạt động như một bộ lọc thông minh, giảm chiều từ 256 xuống 128 nhưng vẫn giữ lại những thông tin quan trọng nhất về nguy cơ lũ. Quá trình này tương tự như việc một chuyên gia kinh nghiệm sẽ loại bỏ những yếu tố ít ảnh hưởng và tập trung vào những chỉ số then chốt. Mô hình học cách kết hợp các đặc trưng đã được mở rộng từ lớp trước để tạo ra các “meta-features” - những đặc trưng bậc cao hơn có khả năng dự đoán mạnh mẽ về nguy cơ lũ. Regularization tiếp tục được áp dụng để đảm bảo mô hình tổng quát hóa tốt trên dữ liệu chưa thấy.



Hình 13. Cấu trúc thiết kế mô hình DNN

Lớp Dense Thứ Ba (64 neurons) - Tổng Hợp Quyết Định

Lớp 64 neurons đóng vai trò như “bộ não tổng hợp”, thu gọn thông tin xuống 64 chiều - một không gian đủ nhỏ để dễ diễn giải nhưng đủ lớn để chứa các mẫu hình phức tạp về nguy cơ lũ. Tại đây, mô hình học cách tạo ra các “signatures” đặc trưng cho từng mức độ nguy hiểm lũ. Dropout được giảm xuống 20% vì ở giai đoạn này, thông tin đã được tinh lọc cao và việc loại bỏ quá nhiều có thể làm mất đi những chi tiết quan trọng cuối cùng. Lớp này có thể được coi như giai đoạn “ra quyết định sơ bộ” trước khi đưa ra phán đoán cuối cùng.

Lớp Output - Quyết Định Phân Loại Cuối Cùng

Lớp Dense cuối cùng với 5 neurons và activation softmax thực hiện nhiệm vụ phân loại cuối cùng, chuyển đổi 64 đặc trưng đã được tinh lọc thành 5 xác suất tương ứng với các mức độ nguy hiểm lũ. Softmax đảm bảo tổng các xác suất bằng 1 và tạo ra phân phối xác suất có ý nghĩa thống kê. Việc sử dụng `dtype='float32'` đảm bảo độ chính xác số học phù hợp cho việc tính toán xác suất. Lớp này không sử dụng regularization vì đây là lớp quyết định cuối cùng và cần giữ nguyên toàn bộ thông tin để đưa ra phán đoán chính xác nhất.

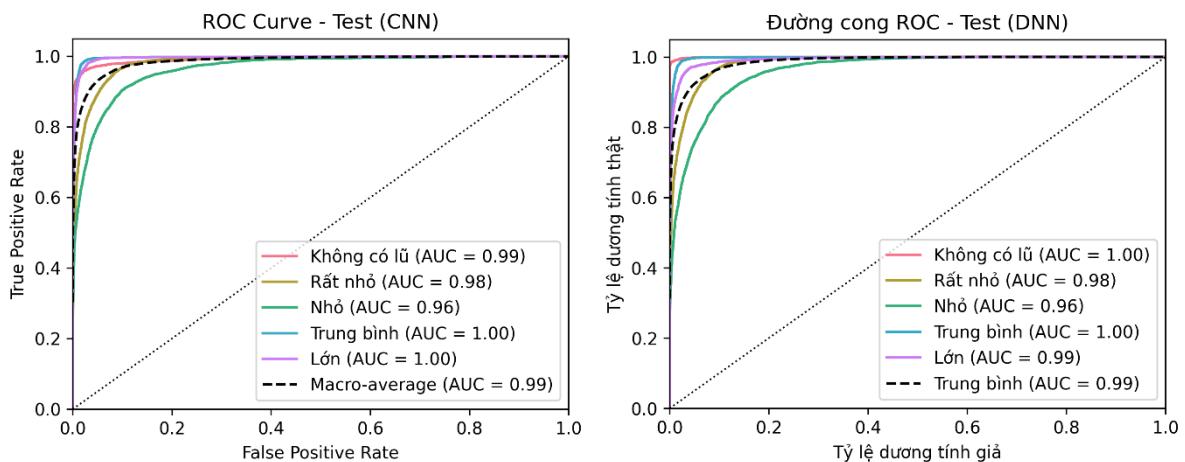
Triết Lý Thiết Kế Tổng Thể

Kiến trúc DNN này phản ánh phương pháp tiếp cận “top-down” trong phân tích nguy cơ lũ, bắt đầu từ việc mở rộng không gian đặc trưng để khám phá tất cả các mối quan hệ có thể, sau đó dần thu gọn và tinh lọc để tìm ra những mẫu hình quan trọng nhất. Quá trình giảm dần số neurons ($256 \rightarrow 128 \rightarrow 64 \rightarrow 5$) tương tự như quá trình suy nghĩ của con người: từ việc xem xét nhiều khả năng, đến việc loại bỏ những khả năng ít có khả năng, cuối cùng đưa ra quyết định dựa trên bằng chứng mạnh nhất. Mô hình này đặc biệt hiệu quả khi làm việc với dữ liệu đã được tổng hợp và các chuyên gia đã xác định được những đặc trưng quan trọng cần xem xét.

3. Đánh giá mô hình học sâu trong phân vùng lũ quét

a. Đường cong ROC

Cả hai mô hình DNN và CNN đều thể hiện hiệu suất xuất sắc với các đường cong ROC gần như lý tưởng, có AUC scores dao động từ 0.96 đến 1.00 cho các lớp khác nhau. Điều đáng chú ý là cả hai mô hình đều đạt được khả năng phân biệt hoàn hảo ($AUC = 1.00$) đối với các lớp “Không có lũ”, “Lũ trung bình” và “Lũ lớn”, cho thấy khả năng nhận diện chính xác các tình huống không có lũ cũng như các sự kiện lũ nghiêm trọng. Sự hội tụ của các đường cong về phía góc trên bên trái của biểu đồ ROC cho thấy cả hai mô hình có khả năng duy trì tỷ lệ dương tính thật cao trong khi giữ tỷ lệ dương tính giả ở mức thấp.



Hình 14. Đường cong ROC theo mô hình CNN và DNN trên tệp kiểm tra

Tuy nhiên, khi xem xét chi tiết, lớp “Lũ nhỏ” là thách thức lớn nhất đối với cả hai mô hình với AUC = 0.96, thấp hơn so với các lớp khác. Điều này có thể được giải thích bởi tính chất không rõ ràng của các sự kiện lũ nhỏ, khi các đặc trưng có thể chồng chéo với điều kiện bình thường hoặc các mức độ lũ khác. Lớp “Lũ rất nhỏ” cũng gặp khó khăn tương tự với AUC = 0.98, cho thấy việc phân loại các sự kiện lũ ở mức độ thấp đòi hỏi độ tinh tế cao hơn trong việc trích xuất đặc trưng.

Sự tương đồng cao giữa hiệu suất của DNN và CNN trong phân tích ROC cho thấy cả hai kiến trúc đều có khả năng học được các mẫu phức tạp trong dữ liệu lũ. Điều này đặc biệt quan trọng trong ứng dụng thực tế, nơi khả năng phân biệt chính xác giữa các mức độ lũ khác nhau có thể quyết định đến hiệu quả của các biện pháp cảnh báo sớm và phản ứng khẩn cấp.

b. Ma Trận nhầm lẫn

Confusion Matrix - Test (CNN)						Ma trận nhầm lẫn - Test (DNN)					
Actual	Predicted					Thực tế	Dự đoán				
	Không có lũ	Rất nhỏ	Nhỏ	Trung bình	Lớn		Không có lũ	Rất nhỏ	Nhỏ	Trung bình	Lớn
Không có lũ	4391	249	76	14	27	4653	64	25	8	7	
Rất nhỏ	69	4075	562	7	43	11	4149	552	5	39	
Nhỏ	22	456	3732	284	263	7	700	3441	241	368	
Trung bình	2	0	78	4648	29	0	2	51	4685	19	
Lớn	7	17	139	35	4558	0	26	164	108	4458	

Hình 15. Ma trận nhầm lẫn của mô hình CNN và DNN trong tệp dữ liệu kiểm tra

Ma trận nhầm lẫn của cả hai mô hình tiết lộ những thông tin chi tiết về hiệu suất phân loại trên từng lớp cụ thể. Mô hình DNN thể hiện độ chính xác đặc biệt cao trong việc nhận diện lớp “Không có lũ” với 4653 trường hợp được phân loại đúng, chỉ có số lượng nhỏ các lỗi phân loại. Tương tự, lớp “Lũ rất nhỏ” và “Lũ trung bình” cũng đạt được độ chính xác cao với 4149 và 4685 trường hợp được phân loại đúng tương ứng. Điều này

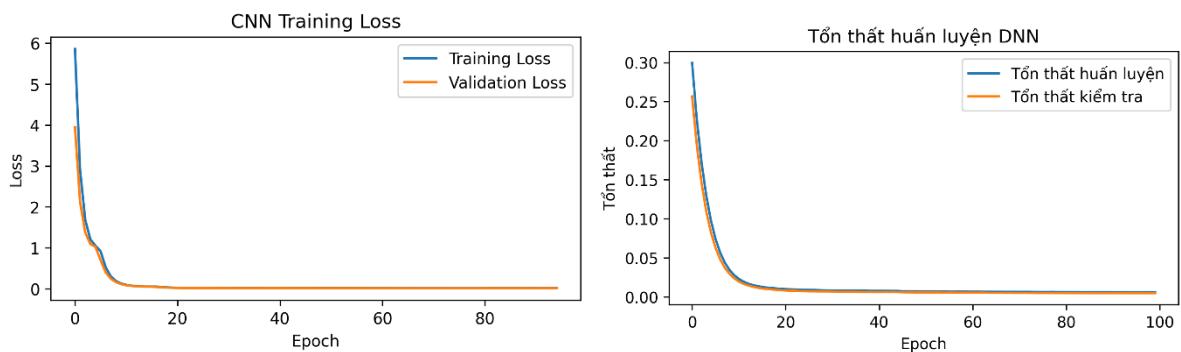
cho thấy mô hình có khả năng mạnh mẽ trong việc nhận diện các tình huống cực đoan - từ không có lũ đến lũ ở mức độ nghiêm trọng.

Mô hình CNN cho thấy cải thiện đáng kể so với DNN, đặc biệt ở lớp “Không có lũ” với 4391 dự đoán chính xác và ít lỗi phân loại hơn. Sự cải thiện này có thể được quy cho khả năng của CNN trong việc trích xuất các đặc trưng không gian và thời gian phức tạp từ dữ liệu. Tuy nhiên, cả hai mô hình đều gặp khó khăn tương đối với lớp “Lũ nhỏ”, với CNN có 3732 dự đoán chính xác so với 3441 của DNN, cho thấy CNN có ưu thế nhẹ trong việc xử lý các trường hợp biên này.

Một điểm đáng quan tâm là mô hình có xu hướng nhầm lẫn giữa các lớp liền kề nhau hơn là với các lớp cách xa. Ví dụ, lớp “Lũ nhỏ” thường bị nhầm với “Lũ rất nhỏ” hoặc “Lũ trung bình” hơn là với “Không có lũ” hoặc “Lũ lớn”. Điều này phản ánh tính chất liên tục của hiện tượng lũ trong thực tế, nơi ranh giới giữa các mức độ có thể không rõ ràng và phụ thuộc vào nhiều yếu tố môi trường phức tạp hoặc do đánh giá phân loại chủ quan ban đầu của kết quả thực địa.

c. Đường cong mất mát huấn luyện

Đường cong mất mát huấn luyện của cả hai mô hình cho thấy quá trình học tập hiệu quả và ổn định. Cả training loss (tổn thất huấn luyện) và validation loss (tổn thất kiểm tra) đều giảm mạnh trong những epoch đầu tiên, từ mức cao khoảng 6.0 xuống dưới 1.0 chỉ trong vòng 5-10 epoch đầu. Sự hội tụ nhanh chóng này cho thấy cả hai mô hình có khả năng học được các mẫu cơ bản trong dữ liệu lũ một cách hiệu quả, không gặp phải các vấn đề về gradient vanishing hay exploding thường gặp trong deep learning.



Hình 16. Đường cong mất mát huấn luyện của mô hình CNN và DNN

Điều đáng chú ý là sự đồng bộ gần như hoàn hảo giữa training loss và validation loss suốt quá trình huấn luyện, cho thấy mô hình không bị overfitting. Sự ổn định này đặc biệt quan trọng trong ứng dụng phân loại lũ, nơi mô hình cần có khả năng tổng quát hóa tốt để xử lý các tình huống mới chưa được gặp trong quá trình huấn luyện. Việc validation loss không tăng lên sau khi đạt mức thấp nhất cho thấy mô hình đã học được các đặc trưng có ý nghĩa thống kê chứ không phải chỉ ghi nhớ dữ liệu huấn luyện.

Sau epoch 20, cả hai đường cong loss đều ổn định ở mức gần 0, cho thấy mô hình đã đạt được trạng thái hội tụ tối ưu. Sự ổn định kéo dài này trong suốt 80 epoch còn lại không chỉ xác nhận tính robustness (bền vững/ổn định) của mô hình mà còn cho thấy

khả năng duy trì hiệu suất cao trong điều kiện huấn luyện mở rộng. Điều này đặc biệt có ý nghĩa trong bối cảnh ứng dụng thực tế, nơi mô hình cần phải duy trì độ chính xác cao qua thời gian và với các biến động trong dữ liệu đầu vào.

Bảng 6. Tổng hợp các tham số đánh giá mô hình học sâu

Mô hình	Accuracy	Precision	Recall	F1 Score
CNN	0.9000	0.8998	0.9000	0.8995
DNN	0.8992	0.8977	0.8992	0.8977

Một điểm đáng chú ý khác, thời gian dự đoán của mô hình CNN là 4 giờ 40 phút so với thời gian dự đoán của mô hình DNN là 5 giờ 40 phút. Các mô hình học sâu thực sự cần nhiều tài nguyên tính toán hơn rất nhiều các mô hình học máy.

1.4. Phân tích, đánh giá các mô hình trí tuệ nhân tạo trong phân vùng lũ quét

Khi đánh giá hiệu suất của các mô hình trí tuệ nhân tạo, cần xem xét cả độ chính xác và thời gian thực thi để đưa ra quyết định tối ưu. Trong nghiên cứu này, các mô hình được phân chia thành hai nhóm chính dựa trên hiệu suất tổng thể.

Nhóm mô hình truyền thống bao gồm Random Forest, SVM và Logistic Regression thể hiện sự khác biệt rõ rệt về hiệu suất. Random Forest nổi bật với độ chính xác 93.95% và thời gian dự đoán chỉ 10 phút, cho thấy sự cân bằng tốt giữa hiệu suất và tốc độ. Ngược lại, SVM mặc dù sử dụng thuật toán phức tạp nhưng lại cho kết quả thất vọng với độ chính xác chỉ 69.47% và thời gian xử lý lên tới hơn 6 giờ. Logistic Regression tuy có tốc độ nhanh (8 phút) nhưng độ chính xác thấp (68.97%) khiến nó không thể cạnh tranh với các mô hình khác.

LGBM và Ensemble đại diện cho các phương pháp cải thiện từng bước và ghép nhiều mô hình hiện đại. LGBM xuất sắc đạt được độ chính xác cao nhất 95.24% với thời gian xử lý hợp lý 10 phút, cho thấy hiệu quả của phương pháp học tăng dần theo gradient đã được tối ưu hóa. Mô hình ensemble với độ chính xác 93.26% và thời gian nhanh nhất (8 phút) trong nhóm hiệu suất cao, thể hiện lợi ích của việc kết hợp nhiều mô hình đơn giản lại với nhau.

Các mô hình deep learning gồm CNN và DNN cho thấy hiệu suất tương đối tốt nhưng đi kèm với chi phí thời gian đáng kể. CNN đạt 90.00% độ chính xác nhưng cần 4 giờ 40 phút để hoàn thành dự đoán, trong khi DNN có hiệu suất thấp hơn (89.92%) nhưng lại tốn thời gian nhiều nhất (5 giờ 40 phút). Điều này cho thấy rằng độ phức tạp của mạng neural không phải lúc nào cũng mang lại hiệu quả tương xứng.

Xét về tỷ lệ hiệu suất trên thời gian, LGBM thể hiện sự vượt trội rõ rệt khi đạt được độ chính xác cao nhất với thời gian xử lý chấp nhận được. Random Forest đứng thứ hai với sự ổn định và tin cậy cao, phù hợp cho các ứng dụng thực tế. Ensemble là lựa chọn thay thế tốt khi cần ưu tiên tốc độ mà vẫn duy trì hiệu suất cao.

Bảng 7. Kết quả tổng hợp đánh giá và khuyến nghị lựa chọn mô hình phân vùng lũ quét

Mô hình	Accuracy	Precision	Recall	F1 Score	Thời gian	Tỷ lệ Hiệu suất/Thời gian	Xếp hạng
LGBM	95.24%	95.23%	95.24%	95.22%	10 p	Xuất sắc	1
Random Forest	93.95%	93.92%	93.95%	93.91%	10 p	Rất tốt	2
Ensemble	93.26%	93.22%	93.26%	93.22%	8 p	Tốt	3
CNN	90.00%	89.98%	90.00%	89.95%	4h 40p	Trung bình	4
DNN	89.92%	89.77%	89.92%	89.77%	5h 40p	Trung bình	5
SVM	69.47%	68.91%	69.47%	69.08%	6h 5p	Rất kém	6
Logistic Regression	68.97%	68.39%	68.97%	68.57%	8 p	Kém	7

Dựa trên kết quả phân tích từ các biểu đồ ROC và bảng hiệu suất, có thể thấy rõ sự phân hóa rõ rệt giữa các nhóm thuật toán trong bài toán phân loại lũ lụt. Điều đáng chú ý nhất là sự vượt trội của các mô hình học máy truyền thống so với các mô hình học sâu, một hiện tượng tương đối bất ngờ trong bối cảnh hiện tại khi deep learning thường được kỳ vọng sẽ cho hiệu suất cao hơn.

LGBM thể hiện hiệu suất ấn tượng nhất với độ chính xác 95.24% và AUC đạt mức hoàn hảo 1.00 trên hầu hết các lớp. Sự xuất sắc này có thể được giải thích bởi bản chất của dữ liệu phân loại lũ lụt, thường bao gồm các đặc trưng có cấu trúc rõ ràng như lượng mưa và các yếu tố khí tượng thủy văn. LGBM, với khả năng xử lý hiệu quả các mối quan hệ phi tuyến tính và tương tác giữa các đặc trưng thông qua cấu trúc cây gradient boosting, đặc biệt phù hợp với loại dữ liệu này. Hơn nữa, thuật toán này có khả năng tự động xử lý các đặc trưng quan trọng và bỏ qua những đặc trưng nhiễu, điều quan trọng trong bài toán dự báo thiên tai.

Random Forest cũng cho thấy hiệu suất mạnh mẽ với độ chính xác 93.95%, chứng tỏ tính hiệu quả của phương pháp ensemble trong việc kết hợp nhiều cây quyết định. Điểm mạnh của Random Forest trong bài toán này nằm ở khả năng xử lý tốt dữ liệu có nhiều chiều và khả năng chống overfitting thông qua việc sử dụng bagging. Đặc biệt, trong lĩnh vực dự báo lũ lụt, việc có thể diễn giải được quyết định của mô hình thông qua cấu trúc cây quyết định là một lợi thế lớn, giúp các chuyên gia thủy văn hiểu được các yếu tố nào đang ảnh hưởng đến dự báo.

Mô hình Ensemble, mặc dù có hiệu suất thấp hơn các mô hình thành phần riêng lẻ với độ chính xác 93.26%, vẫn thể hiện hiệu suất ổn định. Điều này cho thấy rằng việc kết hợp các mô hình khác nhau không phải lúc nào cũng mang lại cải thiện, đặc biệt khi các mô hình thành phần đã có hiệu suất cao và có thể đã học được những pattern tương tự từ dữ liệu.

Sự thất vọng lớn nhất đến từ hiệu suất của các mô hình học sâu. CNN chỉ đạt 90.00% độ chính xác và DNN đạt 89.92%, thấp hơn đáng kể so với các mô hình học máy truyền thống. Từ biểu đồ mắt mát, có thể thấy cả CNN và DNN đều hội tụ nhanh chóng sau

khoảng 20 epoch, nhưng vẫn không thể vượt qua hiệu suất của các mô hình tree-based. Điều này có thể được giải thích bởi nhiều yếu tố quan trọng.

Thứ nhất, kích thước và tính chất của dataset có thể không phù hợp với deep learning. Các mô hình học sâu thường cần lượng dữ liệu rất lớn để học được cách thể hiện dữ liệu phức tạp, trong khi dữ liệu phân loại lũ lụt có thể có kích thước hạn chế. Hơn nữa, dữ liệu khí tượng thủy văn thường có cấu trúc độc lập với các mối quan hệ tương đối rõ ràng giữa input và output, không cần đến khả năng học các mối quan hệ quá phức tạp của deep learning. CNN xuất sắc trong phân loại thảm phủ vì có thể khai thác được tính tương quan không gian giữa các pixel lân cận và các đặc trưng thị giác liên tục. Ngược lại, phân vùng lũ quét phụ thuộc vào sự kết hợp phức tạp của nhiều yếu tố số liệu riêng lẻ mà không có cấu trúc không gian rõ ràng, khiến cho các mô hình cây quyết định như LGBM và Random Forest trở nên phù hợp hơn.

Thứ hai, bản chất của bài toán phân loại lũ lụt có thể phù hợp hơn với cách tiếp cận cây quyết định. Các quy tắc quyết định dạng “nếu lượng mưa lớn hơn X và độ dốc lưu vực lớn hơn Y thì có nguy cơ lũ cao” rất phù hợp với cấu trúc cây quyết định, trong khi các mô hình mạng thần kinh nhân tạo có thể gặp khó khăn trong việc học các quy tắc đơn giản nhưng hiệu quả này.

Thứ ba, việc tạo ra và chọn lọc các đặc điểm dữ liệu cũng rất quan trọng. Các mô hình dựa trên cây quyết định có thể tự động xử lý và chọn ra những đặc điểm quan trọng nhất, trong khi các mô hình học sâu thường cần phải chuẩn bị dữ liệu và thiết kế các đặc điểm phức tạp hơn nhiều mới có thể hoạt động hiệu quả.

Hiệu suất kém của SVM (69.47%) và Hồi quy Logistic (68.97%) có thể do dữ liệu khí tượng có nhiều mối quan hệ không tuyến tính phức tạp. Mặc dù SVM có thể xử lý được các quan hệ cong, không thẳng trong dữ liệu nhờ phương pháp biến đổi không gian, nhưng có thể việc chọn phương pháp và điều chỉnh các thông số chưa được thực hiện tốt. Còn Hồi quy Logistic chỉ có thể xử lý các quan hệ thẳng, đơn giản, nên rất bình thường khi không thể nắm bắt được những mối liên hệ phức tạp trong dữ liệu về lượng mưa.

Một điểm quan trọng khác từ biểu đồ ROC là hiệu suất phân loại không đồng đều giữa các lớp. Trong khi các lớp “Không có lũ”, “Lũ trung bình” và “Lũ lớn” thường đạt AUC gần hoàn hảo, lớp “Lũ nhỏ” thường có hiệu suất thấp hơn đáng kể. Điều này phản ánh thách thức thực tế trong dự báo lũ lụt, khi việc phân biệt giữa “không có lũ” và “lũ nhỏ” thường khó khăn hơn do ranh giới mờ nhạt giữa hai trạng thái này. Nguyên nhân có thể một phần đến từ chính việc phân loại không rõ ràng thực sự giữa 2 lớp trong quá trình chuẩn bị dữ liệu. Mặc dù vậy, hai lớp này thường không ảnh hưởng đến việc ra quyết định trong ứng phó với thiên tai do mức độ chúng mang lại.

Về mặt thời gian tính toán, các mô hình học máy truyền thống cho thấy ưu thế vượt trội. LGBM và Random Forest chỉ cần 10 phút để huấn luyện, trong khi CNN và DNN cần 4-5 giờ. Trong bối cảnh ứng dụng thực tế cho cảnh báo lũ lụt, khi tốc độ xử lý và

cập nhật mô hình là yếu tố quan trọng, sự khác biệt này càng làm nổi bật ưu thế của các mô hình cây quyết định.

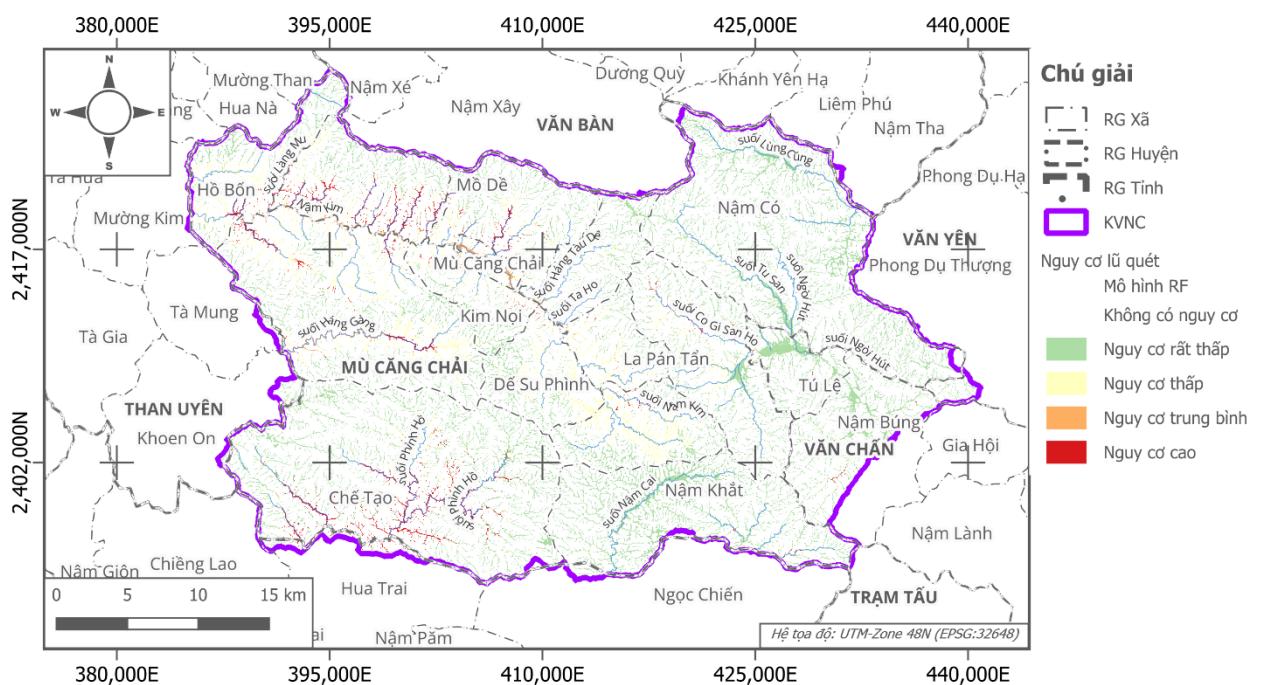
Kết quả này cho thấy tầm quan trọng của việc lựa chọn thuật toán phù hợp với bản chất dữ liệu và bài toán cụ thể, thay vì áp dụng một cách máy móc các mô hình “hiện đại” nhất. Trong lĩnh vực dự báo thiên tai, nơi tính chính xác, tốc độ và khả năng diễn giải đều quan trọng, các mô hình học máy truyền thống có thể sẽ là lựa chọn tối ưu hơn so với deep learning.

CHƯƠNG 2. KẾT QUẢ PHÂN VÙNG LŨ QUÉT CHO KHU VỰC NGHIÊN CỨU

2.1. Kết quả phân vùng lũ quét

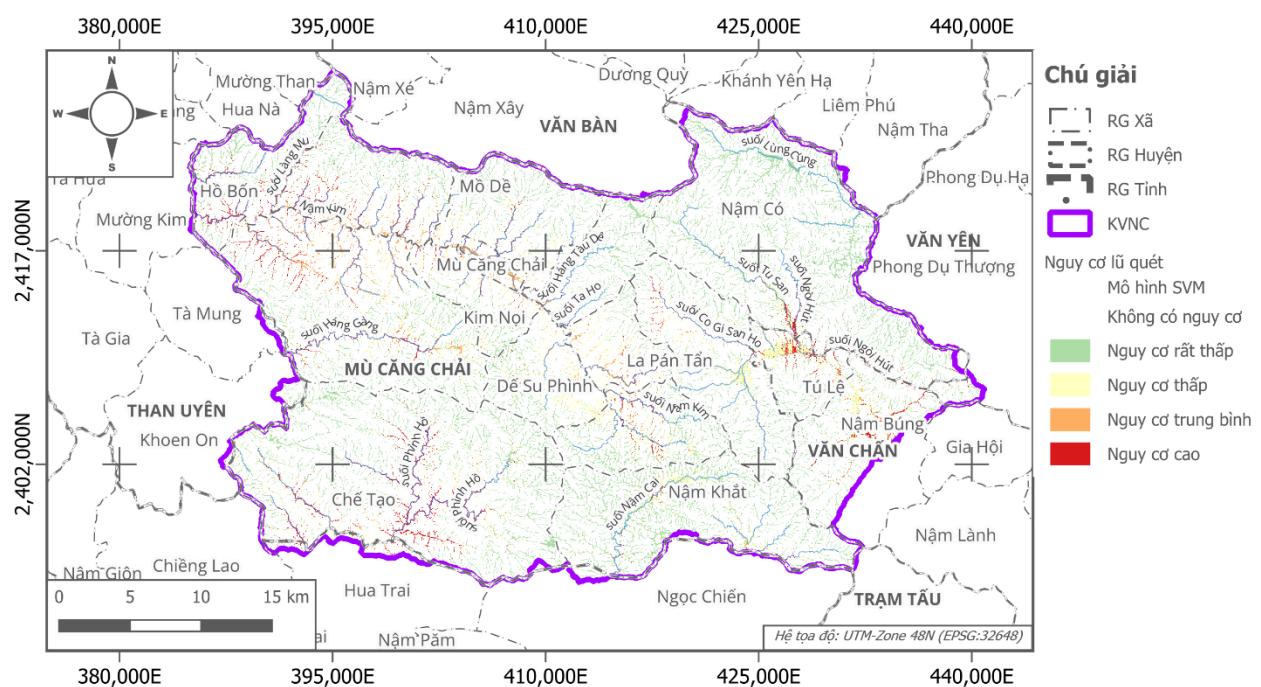
Bản đồ với độ phân giải cao được lưu trữ tại: <https://github.com/tamthat/MCC>

1. Mô hình RF (rừng ngẫu nhiên)



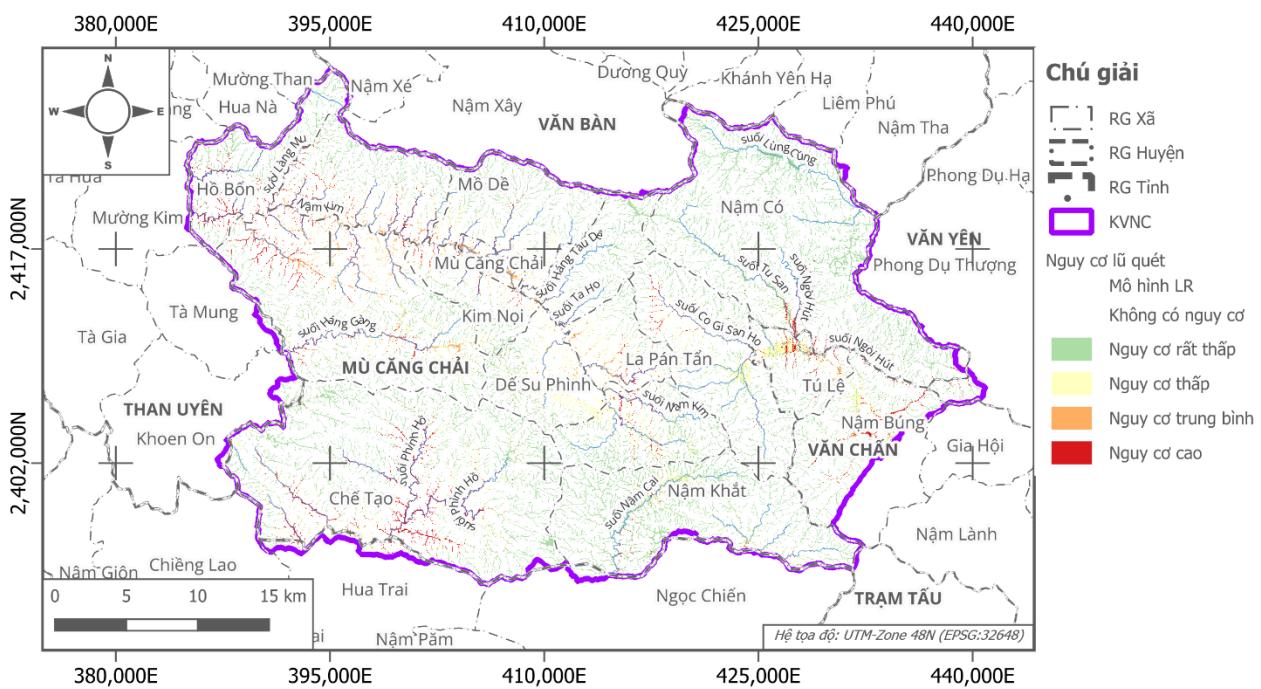
Hình 17. Kết quả xác định nguy cơ lũ quét bằng mô hình RF

2. Mô hình SVM



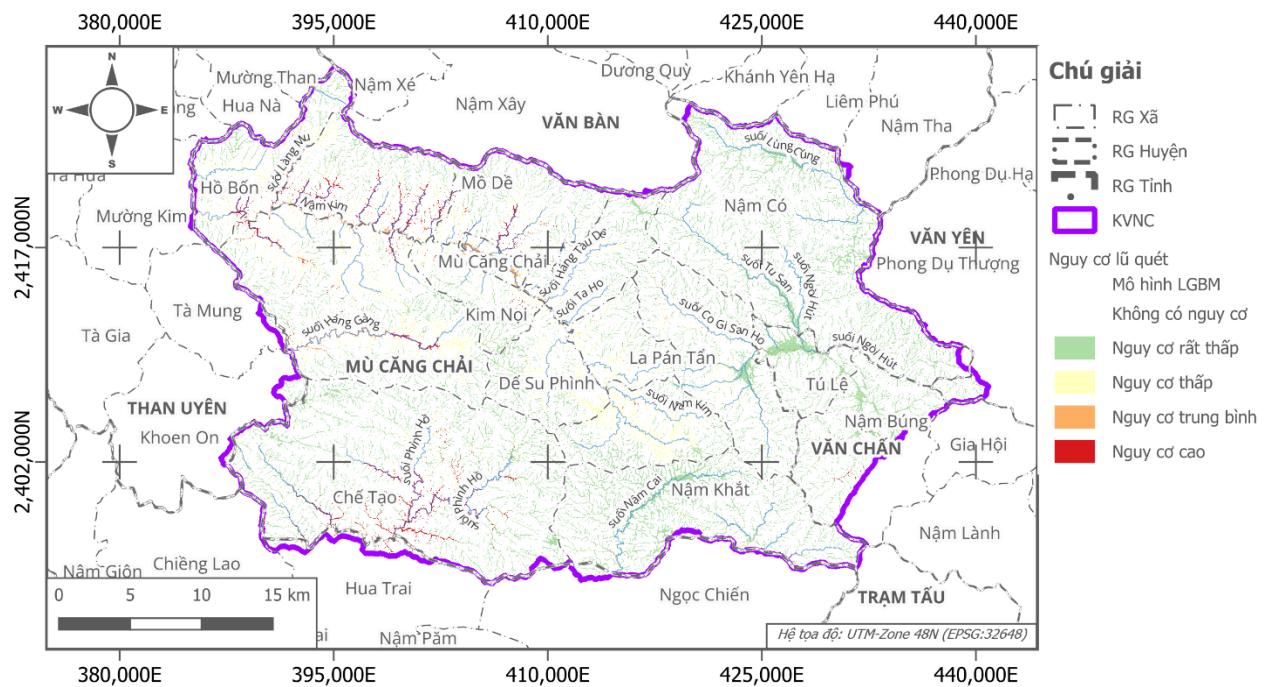
Hình 18. Kết quả xác định nguy cơ lũ quét bằng mô hình SVM

3. Mô hình hồi quy logistic (LR)



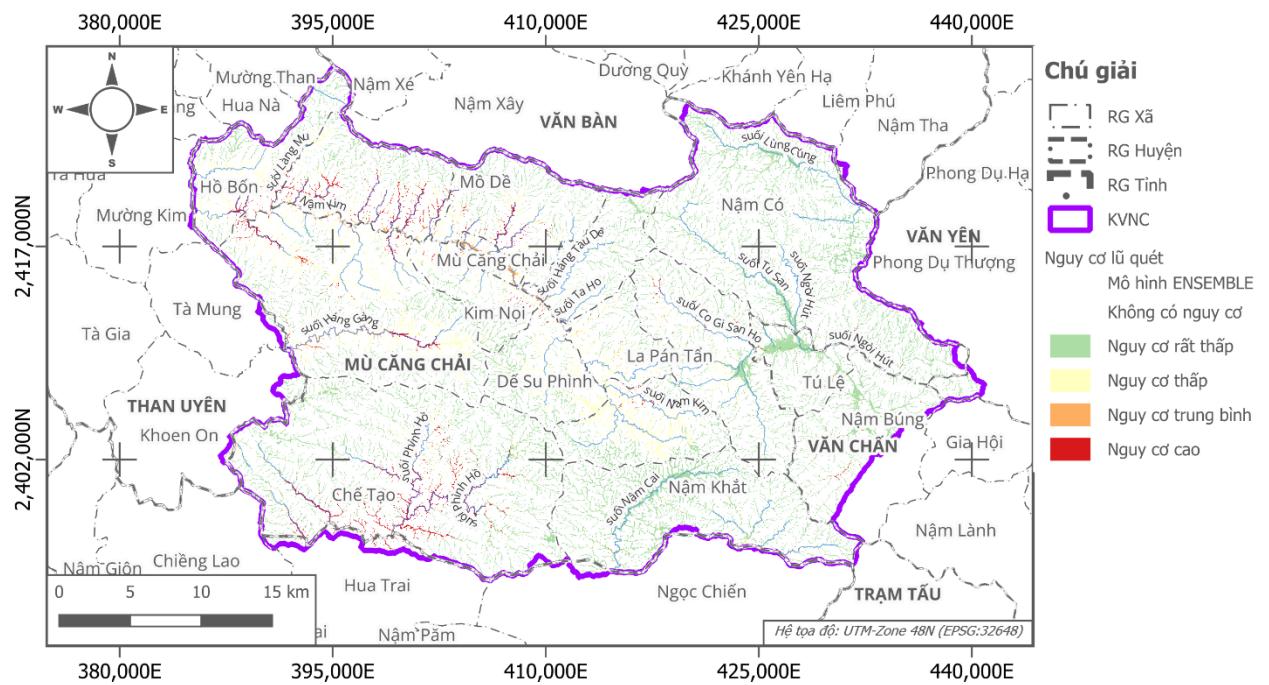
Hình 19. Kết quả xác định nguy cơ lũ quét bằng mô hình LR

4. Mô hình LGBM



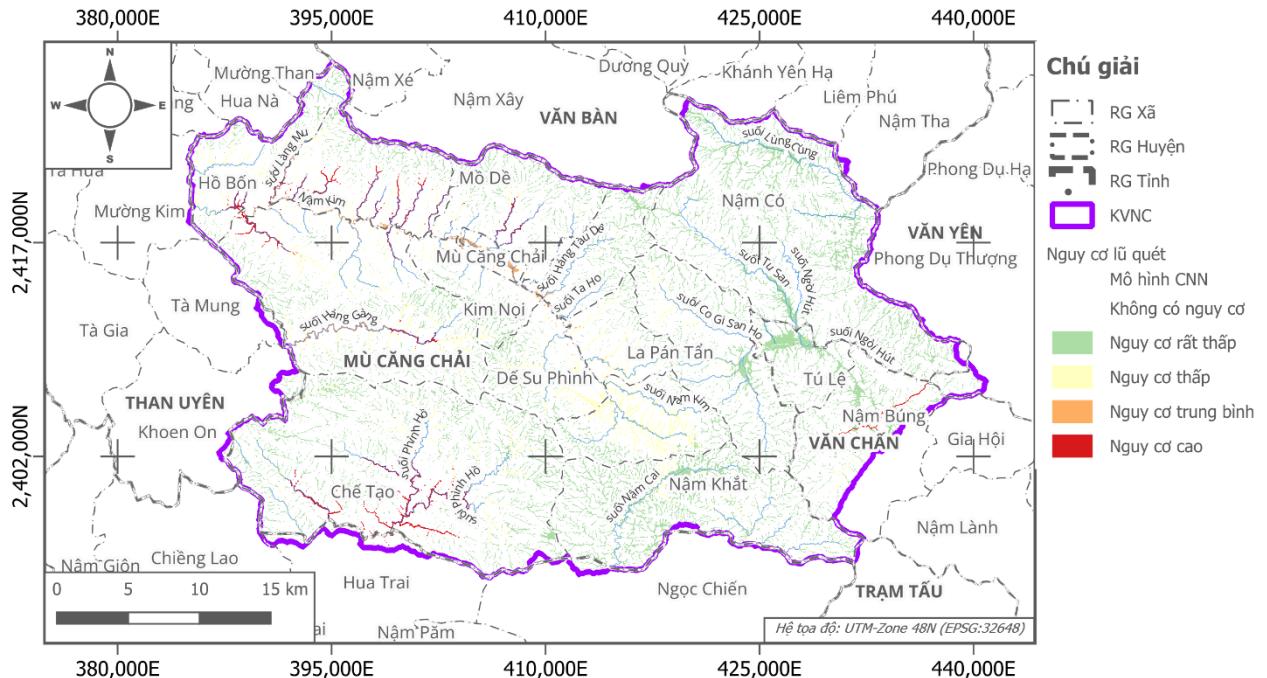
Hình 20. Kết quả xác định nguy cơ lũ quét bằng mô hình LGBM

5. Mô hình ENSEMBLE



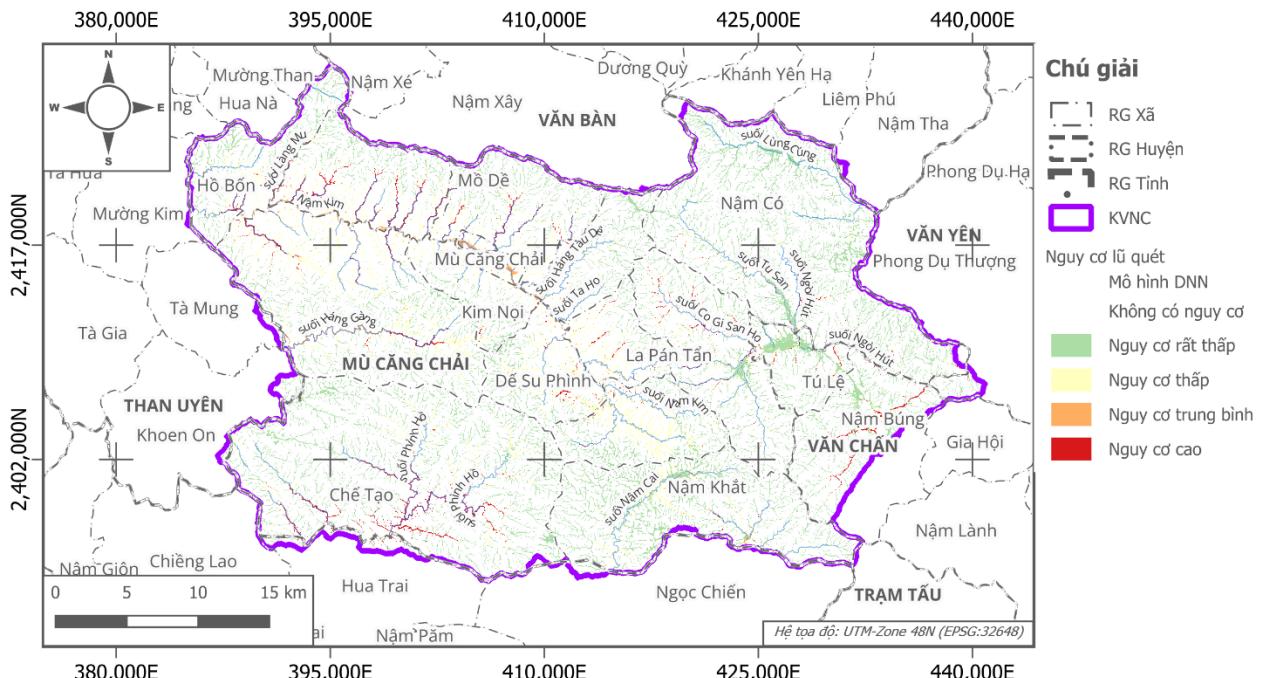
Hình 21. Kết quả xác định nguy cơ lũ quét bằng mô hình ENSEMBLE

6. Mô hình CNN



Hình 22. Kết quả xác định nguy cơ lũ quét bằng mô hình CNN

7. Mô hình DNN



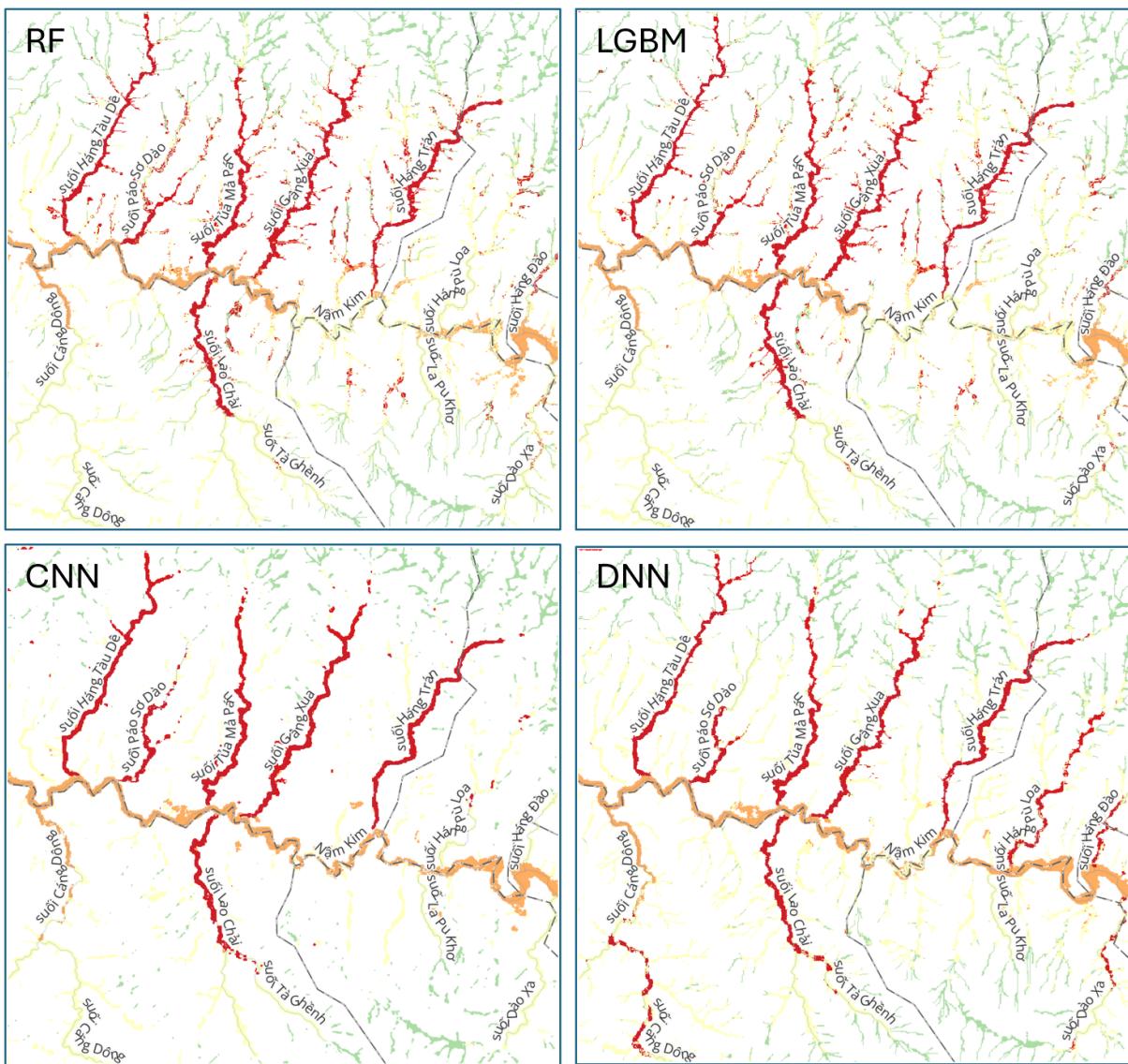
Hình 23. Kết quả xác định nguy cơ lũ quét bằng mô hình DNN

2.2. Đánh giá sự phù hợp của kết quả phân vùng lũ quét

Nội dung này sẽ đánh giá sự phù hợp của kết quả phân vùng lũ quét tại một số vị trí đã xảy ra dựa trên kết quả thu thập tại địa phương ở một số khu vực. Các khu vực được đánh giá bao gồm: (1) khu vực xã Hồ Bón; (2) khu vực xã Khao Mang; (3) khu vực xã Mồ Dề; (4) khu vực xã Lao Chải và (5) khu vực xã Chế Tạo.

Trên góc nhìn tổng thể, nhóm mô hình học sâu (bao gồm CNN và DNN) đạt được sự phù hợp rất cao và cho ra kết quả phân loại rất rõ ràng và sắc nét mặc dù có độ chính xác theo đánh giá không phải là cao nhất. Hầu hết các điểm phân loại nguy cơ cao đều nằm trên lòng dẫn và lân cận lòng dẫn, khu vực phân loại rõ ràng, không xen lẫn với các nhóm nguy cơ khác ở cùng một vị trí. Điều này cho thấy mô hình thực sự nắm bắt được các mối liên hệ không gian tại điểm lũ quét và lân cận. Tiếp đó là đến nhóm mô hình cây quyết định (bao gồm RF, LGBM và ENSEMBLE). Mặc dù nhóm mô hình cây quyết định có độ chính xác cao hơn nhưng nhiều điểm trên cùng một nhánh suối vẫn bị hiện tượng phân loại không khớp, dẫn đến sự đan xen nguy cơ ngay tại cùng một vị trí. Hạn chế này cho thấy việc thiếu đánh giá mạnh mẽ về “các điểm lân cận” có thể gây ra bản đồ phân loại bị “nhiễu”.

Trong hai mô hình học sâu là CNN và DNN, mô hình DNN tỏ ra chiếm ưu thế lớn nhờ sự phân vùng liên tục có quy luật. Các nhánh suối thượng nguồn bắt đầu từ “không có nguy cơ” đến “nguy cơ rất thấp” rồi chuyển tiếp sang “nguy cơ thấp” và “nguy cơ trung bình” rồi đến nguy cơ cao. Các khu vực nguy cơ được chuyển tiếp liền mạch sang các vùng lân cận bằng các cấp độ lân cận, điều mà mô hình CNN chưa nắm bắt tốt được.



Hình 24. Sự khác biệt phân vùng lũ quét giữa 2 nhóm cây quyết định (RF, LGBM) và nhóm học sâu (CNN, DNN)

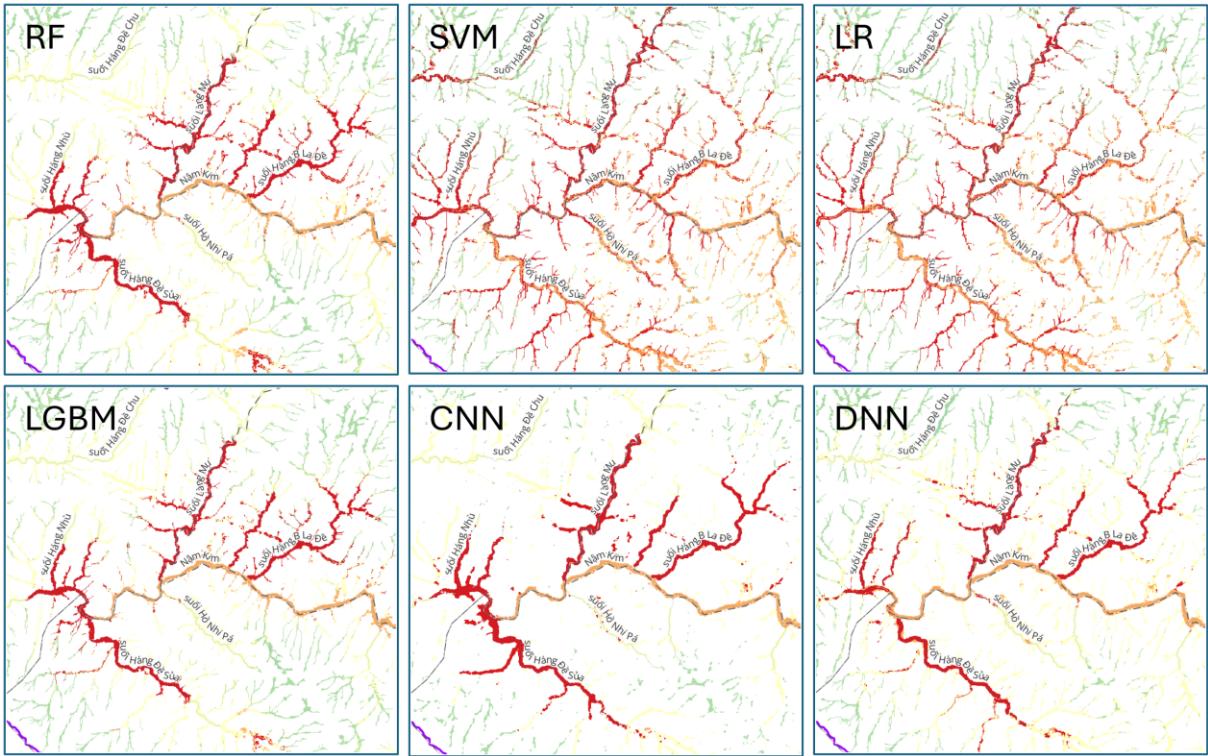
Đặc biệt, các nhánh suối đổ vào các suối nhỏ là các nhánh chỉ có diện tích lưu vực thượng nguồn vài trăm m², khó hoặc không thể có khả năng sinh lũ quét, trong khi đó ở nhóm mô hình cây quyết định, các nhánh suối này vẫn hiển thị nguy cơ cao trước nhập lưu trong khi nhóm mô hình học sâu đánh giá ở mức độ “không có nguy cơ” hoặc “nguy cơ rất thấp”. Điều này một lần nữa cho thấy nhóm mô hình học sâu đã cho kết quả phân loại phù hợp hơn tốt hơn.

1. Khu vực xã Hồ Bốn

Khu vực xã Hồ Bốn là nơi xảy ra lũ quét rất lớn tại UBND xã Hồ Bốn và trạm Y Tế xã. Nguyên nhân là lũ quét xuất hiện từ nhánh suối Háng Đè Sủa và cuốn trôi rất nhiều vật liệu trên lòng dẫn về suối Nậm Kim, do đó, toàn bộ suối Háng Đè Sủa đổ vào nhánh suối Nậm Kim và suối Nậm Kim phía sau nhập lưu bị lũ quét rất lớn.

Tại khu vực này cho thấy kết quả phân loại của 3 mô hình RF, LGBM và CNN rất phù hợp với tình hình lũ quét xảy ra trên khu vực. Trong khi đó, mô hình SVM và mô hình LR đã không chỉ ra được khu vực suối Háng Đè Sủa là khu vực có nguy cơ cao. Ngoài ra, suối Cù Di Seng và suối Xéo Dì Hồ (bên trái suối Nậm Kim, phía dưới góc phải bản đồ) không ghi nhận lũ quét nhưng lại được 2 mô hình SVM và mô hình LR chỉ ra là có nguy cơ cao trong khi các mô hình còn lại không thể hiện điều này. Riêng mô hình DNN thể hiện nguy cơ lũ quét bị ngắt quãng từ suối Háng Đè Sủa nhập lưu với suối Nậm Kim.

Một nhánh suối khác là Háng Đè Chu (góc trên bên trái bản đồ) chỉ ghi nhận lũ nhỏ, các mô hình RF, LGBM, CNN và DNN một lần nữa lại cho ra kết quả phù hợp với phân loại nguy cơ thấp trong khi hai mô hình SVM và LR cho ra nguy cơ cao, thể hiện sự phân loại chưa phù hợp.

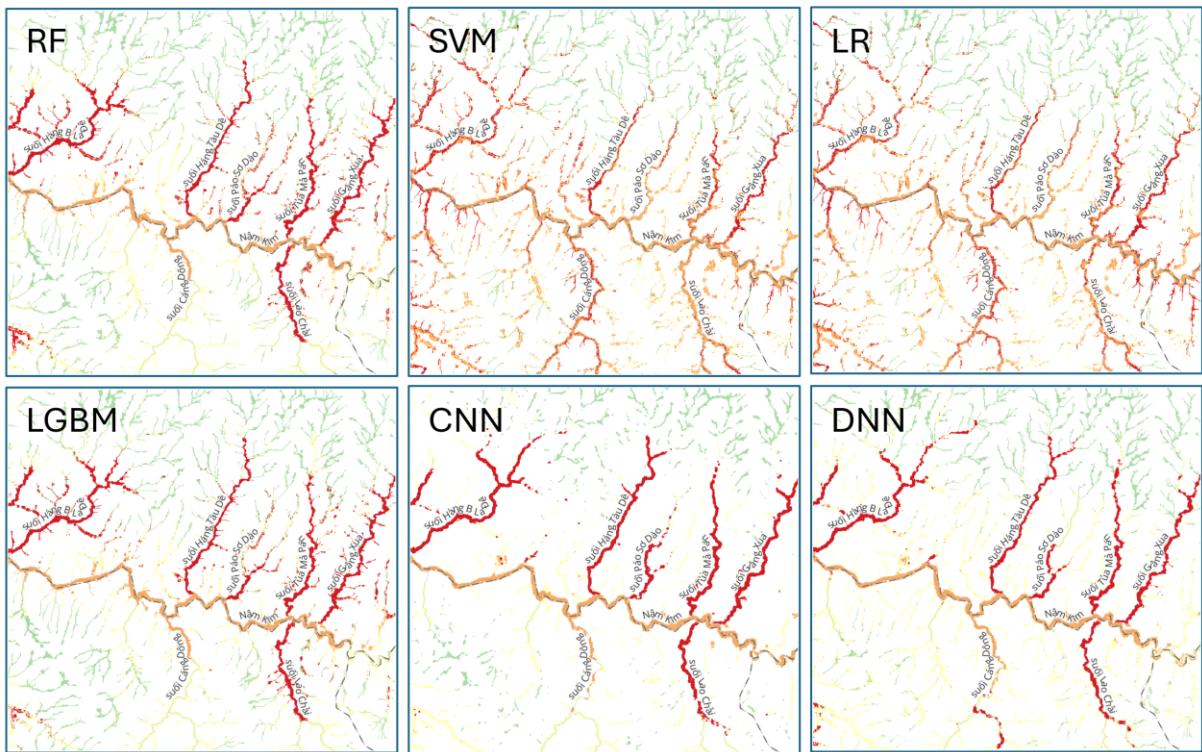


Hình 25. Phân vùng lũ quét khu vực xã Hò Bón

Như vậy, với khu vực xã Hò Bón, kết quả phân loại nguy cơ lũ quét của các mô hình xếp theo tiêu chí phù hợp với thực tế được sắp xếp theo thứ tự là CNN, LGBM, RF, CNN, DNN, SVM và LR. Mô hình CNN lần này đã thể hiện xuất sắc sự liên mạch nguyên nhân gây lũ quét bắt nguồn từ suối Háng Đề Sữa đổ vào nhánh chính Nậm Kim gây ra lũ quét dọc khu vực trong khi mô hình DNN chưa thể hiện được sự liên tục này, tuy nhiên về sự thể hiện sự chuyển tiếp giữa các hình thái nguy cơ theo cấp độ, mô hình CNN không làm tốt như mô hình LGBM và mô hình RF. Mặc dù vậy, việc gây nhiễu ở một số nhánh suối thượng nguồn khiến 2 mô hình này không được đánh giá cao bằng mô hình CNN về mặt tổng quát.

2. Khu vực xã Khao Mang

Xã Khao Mang (bên phải nhánh suối Nậm Kim) trong đợt mưa lũ năm 2023 là một trong 3 xã chịu ảnh hưởng nghiêm trọng của trận lũ. Các nhánh suối thuộc địa bàn xã hầu hết có lũ lên và đều ghi nhận lũ lớn.

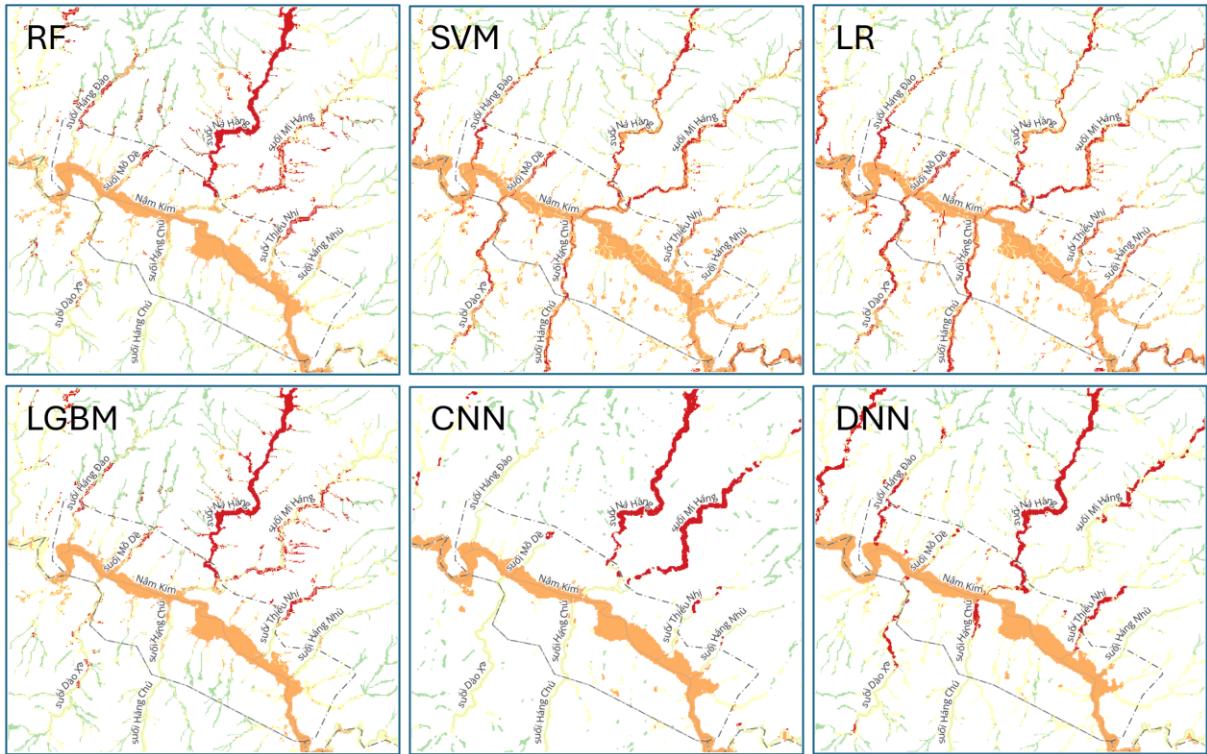


Hình 26. Phân vùng lũ quét khu vực xã Khao Mang

Kết quả phân loại một lần nữa cho thấy 2 nhóm mô hình học sâu và cây quyết định cho ra kết quả tốt hơn LR và SVM. Điều này hoàn toàn có thể lý giải được do các mô hình này khó nắm bắt được các yếu tố phi tuyến so với các mô hình còn lại. Vấn đề phân loại nhiều ở một số nhánh suối thượng nguồn vẫn làm cho các mô hình nhóm cây quyết định bị đánh giá thấp hơn so với mô hình học sâu. Cả hai mô hình CNN và DNN đã làm rất tốt sự phân vùng lũ quét trong khu vực này, tuy nhiên, việc ghi nhận thêm sự kiện lũ quét trên suối Cango Dong có thể chưa được ghi nhận hoặc xác định tại mô hình DNN, do đó các kết quả phân vùng có thể sẽ hữu ích để kiểm chứng trong tương lai khi chạy với các trận lũ mới.

3. Xã Mò Dè

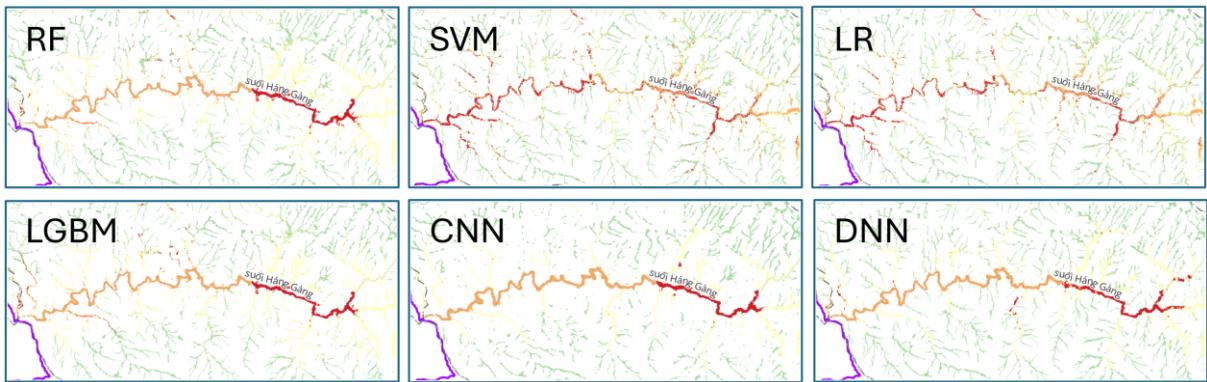
Tại khu vực xã Mò Dè và thị trấn Mù Cang Chải, nơi có xảy ra lũ lớn tại suối Nà Háng và Mý Háng. Bên cạnh đó, suối Háng Chú, nơi xảy ra lũ quét năm 2017 chỉ ghi nhận lũ trung bình. Các mô hình phân loại nguy cơ có sự biến đổi rõ rệt, mỗi mô hình cho ra một kết quả khác nhau



Hình 27. Phân vùng lũ quét khu vực xã Mò Dè và thị trấn Mù Cang Chải

Về sự phù hợp, lần này mô hình CNN đã làm rất tốt trong khi mô hình DNN đánh giá suối Háng Chủ là có nguy cơ cao, còn nhóm mô hình cây quyết định vẫn bị nhiễu ở các nhánh suối nhỏ. Đặc biệt tại suối Mí Háng có độ nhiễu rất lớn ở hầu hết các mô hình.

4. Xã Lao Chải

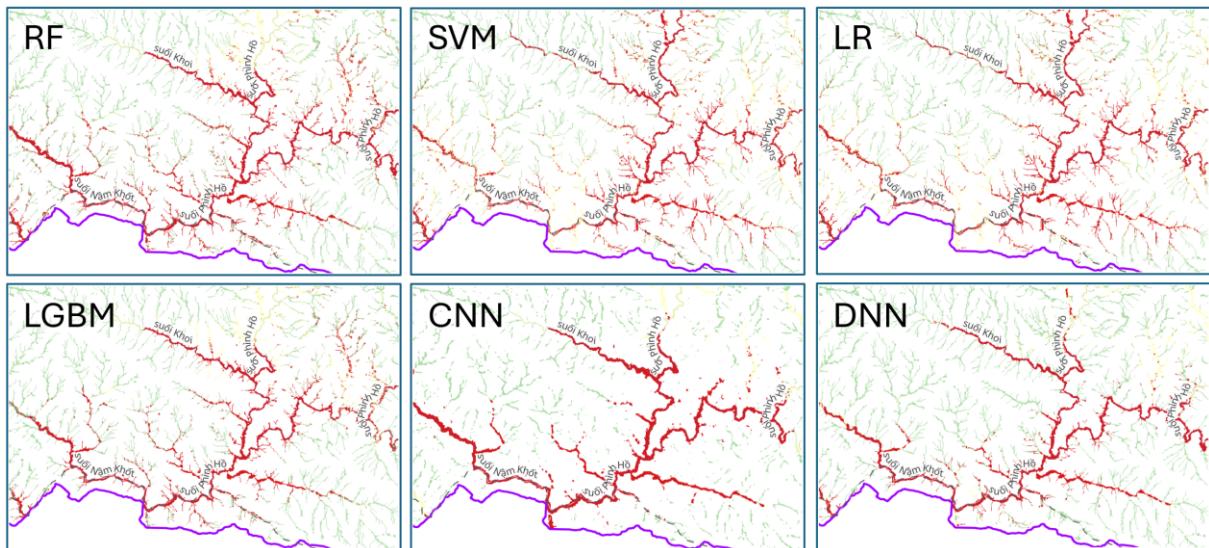


Hình 28. Phân vùng lũ quét khu vực xã Lao Chải (suối Háng Gàng)

Tại xã Lao Chải, suối Háng Gàng ở khu vực cuối Bản Háng Gàng ghi nhận xảy ra lũ lớn. Tuy nhiên, tình trạng dọc tuyến suối Háng Gàng không có đánh giá chi tiết về mức độ lũ. Do đó, các kết quả phân loại trên nhánh suối này của các mô hình đều có thể được xem là phù hợp. Mặc dù vậy, các mô hình SVM và LR đưa cả các nhánh suối nhỏ đổ vào suối Háng Gàng vào nhóm nguy cơ cao cho thấy sự khác biệt về phân loại giữa hai mô hình này với các mô hình còn lại.

5. Xã Chế Tạo

Các nhánh suối đổ vào suối Phình Hồ trong đợt mưa 5/8/2023 theo kết quả điều tra cho thấy đều có lũ lớn.



Hình 29. Phân vùng lũ quét khu vực xã Chế tạo (suối Phình Hồ)

Trên cơ sở phân loại có thể thấy mô hình CNN không bị nhiễu bởi các nhánh suối nhỏ, nơi lượng nước tập trung không đủ lớn, trong khi đó, các mô hình còn lại đều gây nhiễu nhất định, điều này cho thấy khả năng vượt trội của việc xử lý không gian trong mô hình CNN

2.3. Đánh giá chung

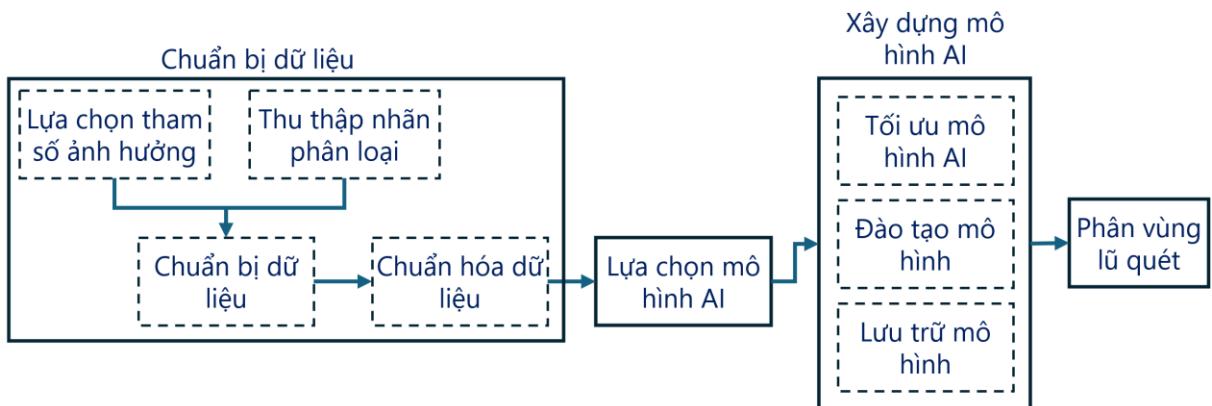
Một khía cạnh thú vị mà kết quả đường cong ROC và các chỉ số phân loại chi tiết cho thấy là CNN và DNN, mặc dù có độ chính xác tổng thể thấp hơn (90.00% và 89.92%), lại thể hiện độ tin cậy và tính nhất quán cao hơn trong các quyết định phân loại. Điều này phản ánh một đặc tính quan trọng của deep learning: khả năng cung cấp độ chắc chắn được hiệu chỉnh tốt (well-calibrated confidence) cho từng dự đoán. Trong khi các mô hình cây quyết định có thể đưa ra quyết định “cứng” (hard decisions) với điểm số tin cậy cực đoan (gần 0 hoặc gần 1), CNN và DNN thường tạo ra phân phối xác suất mượt mà và hợp lý hơn cho các trường hợp chuyển tiếp giữa các lớp.

Các mô hình học sâu thể hiện ưu điểm vượt trội trong việc xử lý các trường hợp mờ hồ và đo lường độ không chắc chắn (uncertainty quantification). Khi CNN và DNN “không chắc chắn” về một dự đoán, chúng thường thể hiện điều này thông qua điểm số đo trung gian về độ chắc chắn (ví dụ 0.6 thay vì 0.9), giúp người dùng hiểu rõ hơn về độ tin cậy của từng dự đoán. Điều này đặc biệt quan trọng trong các ứng dụng khí tượng thủy văn, nơi mà việc biết được mức độ không chắc chắn của dự báo có thể quan trọng hơn cả việc có một con số chính xác tuyệt đối. Các mô hình cây quyết định, mặc dù có hiệu suất cao hơn, có thể cho ra các điểm tin cậy thiếu chính xác đối với các mẫu khó phân loại.

Các kết quả phân loại cho thấy, mô hình học sâu có thể mang lại giá trị lớn về độ tin cậy hay sự phù hợp và khả năng xử lý dữ liệu phức tạp nhưng cần sự đánh đổi về tài nguyên. Trong bối cảnh hiệu quả nhằm áp dụng phân loại nhanh (trong các ứng dụng thực tế), mô hình LGBM vẫn là lựa chọn tối ưu. Các mô hình học sâu có thể được cân nhắc cho các ứng dụng đặc biệt quan trọng, nơi cần sự cân bằng giữa độ chính xác và sự phù hợp của kết quả phân loại.

CHƯƠNG 3. QUY TRÌNH ỦNG DỤNG TRÍ TUỆ NHÂN TẠO VÀ VIỄN THÁM ĐỂ PHÂN VÙNG LŨ QUÉT

3.1. Sơ đồ quy trình



Hình 30. Quy trình ứng dụng trí tuệ nhân tạo trong phân vùng lũ quét

3.2. Xác định các bước thực hiện

3.2.1 Chuẩn bị dữ liệu

1. Lựa chọn tham số

a. Các tham số đầu vào

Phân vùng lũ quét bằng trí tuệ nhân tạo (AI) đang trở thành một phương pháp tiên tiến trong quản lý rủi ro thiên tai, tuy nhiên hiệu quả của các mô hình AI phụ thuộc hoàn toàn vào chất lượng và tính đại diện của dữ liệu đầu vào. Khác với các hiện tượng thiên tai khác, lũ quét có đặc thùy đặc biệt phức tạp do tính chất động lực học không gian-thời gian và sự tương tác đa tầng giữa các yếu tố thủy văn, địa hình và khí tượng. Việc chuẩn bị dữ liệu đầu vào không chỉ đơn thuần là thu thập thông tin mà còn đòi hỏi sự hiểu biết sâu sắc về bản chất vật lý của hiện tượng lũ quét và khả năng biểu diễn các quá trình này thông qua các tham số có ý nghĩa khoa học.

Lũ quét là hiện tượng thủy văn cực đoan được đặc trưng bởi thời gian hình thành ngắn (thường dưới 6 giờ), lưu lượng đỉnh cao và khả năng phá hủy lớn. Khác với lũ lụt thông thường, lũ quét được hình thành chủ yếu bởi các quá trình diễn ra ở quy mô lưu vực nhỏ đến trung bình, nơi mà thời gian tập trung ngắn và khả năng phản ứng nhanh chóng của hệ thống thủy văn đối với các sự kiện mưa cực đoan. Điều này tạo ra những thách thức đặc biệt trong việc lựa chọn và chuẩn bị dữ liệu đầu vào cho các mô hình AI.

Thứ nhất, tính chất phi tuyến và ngưỡng của các quá trình lũ quét đòi hỏi dữ liệu phải có khả năng nắm bắt được các điểm chuyển đổi quan trọng trong hệ thống. Ví dụ, khả năng thấm của đất có thể thay đổi đột ngột khi độ ẩm đạt đến mức bão hòa, dẫn đến sự gia tăng lớn của dòng chảy bề mặt (tạo những đường quá trình lưu lượng lũ đột biến). Dữ liệu đầu vào cần phải có độ phân giải không gian và thời gian đủ cao để có thể phát hiện và mô hình hóa những thay đổi này.

Thứ hai, tính không đồng nhất không gian cao của các yếu tố điều khiển lũ quét đòi hỏi dữ liệu phải có khả năng biểu diễn sự biến thiên không gian phức tạp. Địa hình, thổ nhưỡng, thực phủ và sử dụng đất có thể thay đổi đáng kể trong phạm vi một lưu vực nhỏ, tạo ra các vùng có đặc tính thủy văn hoàn toàn khác nhau. Việc tổng hợp hóa quá mức hoặc sử dụng dữ liệu có độ phân giải thấp có thể dẫn đến việc mất mát thông tin quan trọng về sự không đồng nhất này. Phần lớn các nghiên cứu phân vùng lũ quét hiện tại đều dựa trên phương pháp tiếp cận “pixel-based” hoặc “point-based”, trong đó mỗi điểm trong không gian được đặc trưng bởi một tập hợp các thuộc tính nội tại như độ dốc, độ cao, loại đất, lượng mưa cục bộ, và chỉ số thực phủ. Mặc dù phương pháp này có ưu điểm về tính đơn giản trong việc thu thập và xử lý dữ liệu, nhưng nó bỏ qua hoàn toàn bản chất hệ thống của hiện tượng lũ quét.

Hạn chế căn bản của phương pháp này nằm ở việc không tính đến “yếu tố lưu vực” - tức là ảnh hưởng của toàn bộ lưu vực thượng nguồn đối với một điểm cụ thể. Trong thực tế, nguy cơ lũ quét tại một vị trí không chỉ phụ thuộc vào các đặc điểm cục bộ mà còn phụ thuộc rất lớn vào khả năng tích lũy dòng chảy từ toàn bộ khu vực thượng nguồn. Một điểm có địa hình tương đối bằng phẳng và thổ nhưỡng có khả năng thấm tốt vẫn có thể bị lũ quét nghiêm trọng nếu nó nằm ở vị trí hội tụ của nhiều dòng chảy từ các khu vực có độ dốc lớn và khả năng thấm kém ở phía thượng nguồn.

Hơn nữa, việc sử dụng dữ liệu nội tại điểm còn bỏ qua các quá trình động lực học quan trọng như truyền lũ và sự suy giảm của dòng chảy. Dòng chảy sinh ra từ một sự kiện mưa không đơn giản là tổng các đóng góp từ các điểm riêng lẻ mà còn trải qua quá trình biến đổi phức tạp khi di chuyển qua mạng lưới thủy văn. Thời gian chảy truyền, khả năng tích trữ tạm thời trong các vùng trũng, và sự gia tăng đột biến về đỉnh lũ do mặt đệm bão hòa đều là những yếu tố quan trọng quyết định đến cường độ và thời điểm xuất hiện của lũ quét tại một vị trí cụ thể.

Do đó, để khắc phục những hạn chế của phương pháp tiếp cận dữ liệu nội tại điểm, cần chuyển đổi từ việc mô tả “điều kiện” sang việc lượng hóa “nguyên nhân” của lũ quét. Điều này đòi hỏi việc phát triển các chỉ số và biến số có khả năng nắm bắt được các quá trình vật lý cơ bản điều khiển sự hình thành và phát triển của lũ quét.

Thứ nhất, cần tích hợp các chỉ số liên quan đến quá trình hình thành dòng chảy. Trong thủy văn, khi một lượng mưa rơi xuống sẽ xét đến quá trình thẩm thấu vào mặt đất, lượng nước dư thừa mới tạo nên dòng chảy. Dòng chảy này được tích lũy dần vào các nhánh suối chảy về hạ lưu. Các yếu tố này thường được xác định thông qua mô hình

thủy văn để mô phỏng quá trình này. Do đó, các tham số thủy văn mang tính lưu vực cần được xét đến một cách tương đối đầy đủ như diện tích lưu vực, độ dốc bình quân lưu vực, chiều dài dòng chảy, khả năng hấp thụ nước, thời gian tập trung dòng chảy... Trong nghiên cứu này, nhóm nghiên cứu đã sử dụng một số tham số liên quan trực tiếp bao gồm: chỉ số CN bình quân lưu vực (liên quan đến tần suất dòng chảy), chiều dài dòng chảy (liên quan đến thời gian tập trung dòng chảy), độ dốc lòng dẫn (liên quan đến năng lượng dòng chảy), độ dốc bình quân lưu vực (liên quan đến thời gian tập trung dòng chảy)... Các tham số này không phải là những tham số riêng lẻ (cho từng điểm) mà là các tham số đại diện cho một lưu vực thượng nguồn chảy ra điểm đó. Cách tiếp cận này thực sự đã đưa những đặc điểm thủy văn làm dữ liệu đầu vào cho mô hình trí tuệ nhân tạo, từ đó giúp cho những nguyên lý thủy văn được tích hợp vào mô hình trí tuệ nhân tạo.

Thứ hai là các chỉ số tương tác không gian như cao độ so với sông suối gần nhất, khoảng cách đến sông suối là những chỉ số quan trọng. Một điểm bị lũ quét sẽ kéo theo các điểm lân cận bị lũ quét. Do đó, các cơ sở hạ tầng như nhà cửa hay các công trình trên sông thường bị tác động mạnh mẽ bởi lũ quét. Các tham số này nên được xem xét một cách cẩn thận nhằm làm rõ bản chất của quá trình hình thành lũ quét.

Thứ ba là lượng mưa, nhiều nghiên cứu đã sử dụng lượng mưa bình quân năm để đánh giá nguy cơ lũ quét. Điều này có phần chưa phản ánh được nguyên nhân sinh lũ quét, bởi vì lũ quét thường chỉ diễn ra trong khoảng 6 giờ, điều này có nghĩa thời gian diễn ra lũ quét chính là thời gian đạt đỉnh (time to peak) – một tham số trong mô hình thủy văn. Lượng mưa nếu có bước thời gian lớn hơn thời gian đạt đỉnh sẽ không thể nào phản ánh được nguy cơ sinh lũ. Do đó, lượng mưa trong khoảng từ 1÷6 giờ thường được xem xét như nguyên nhân sinh lũ. Bên cạnh đó, lượng mưa tích lũy thời đoạn trước cũng cần được xem xét do có tác động lớn đến trạng thái bề mặt của lưu vực. Do đó, các lượng mưa lớn hơn 6 giờ sẽ khái quát hóa được sự hình thành lũ một cách chính xác hơn.

Cuối cùng là các tham số nội tại, nếu chỉ xét đến các tham số thủy văn, việc xây dựng mô hình trí tuệ nhân tạo không mang ý nghĩa lớn về sự đột phá trong phương pháp. Khác với các công cụ khác, trí tuệ nhân tạo có thể nhận diện được các mối quan hệ phi tuyến phức tạp và tương tác đa chiều giữa các yếu tố môi trường. Do đó, việc bổ sung các tham số nội tại như địa hình chi tiết (độ dốc cục bộ, độ cong bề mặt...), đặc tính thổ nhưỡng (thành phần cơ giới, độ xốp, khả năng thấm), và đặc điểm sử dụng đất (mật độ che phủ thực vật, tỷ lệ bề mặt không thấm) vẫn có giá trị quan trọng. Những tham số này hoạt động như các “điều kiện biên” cục bộ, ảnh hưởng đến cách thức mà các quá trình thủy văn cấp lưu vực được biểu hiện tại từng vị trí cụ thể. Ví dụ, hai điểm có cùng điều kiện lưu vực thượng nguồn nhưng khác biệt về độ thấm của đất sẽ có mức độ nguy hiểm lũ quét khác nhau do khả năng thoát nước cục bộ khác biệt. Hơn nữa, các thuật toán AI hiện đại như Random Forest hay Neural Networks có khả năng tự động phát

hiện và khai thác các tương tác giữa tham số lưu vực và tham số nội tại, tạo ra những phát hiện mới về cơ chế hình thành lũ quét mà các phương pháp truyền thống khó có thể nhận diện được. Điều quan trọng là cần duy trì cân bằng hợp lý giữa tham số thủy văn (chiếm tỷ trọng chính) và tham số nội tại (đóng vai trò bổ sung và tinh chỉnh), đảm bảo mô hình vừa năm bắt được bản chất vật lý của hiện tượng vừa tận dụng được sức mạnh tính toán của AI trong việc khám phá các mẫu hình phức tạp.

Tuy nhiên, số lượng các yếu tố hay việc lựa chọn các yếu tố đầu vào không có một chuẩn mực nào bắt buộc, mà phần lớn phụ thuộc vào kinh nghiệm và phán đoán chuyên môn của người xây dựng mô hình. Điều này tạo ra một thách thức lớn trong việc chuẩn hóa quy trình phát triển mô hình AI cho phân vùng lũ quét. Khác với các lĩnh vực khác như nhận dạng hình ảnh hay xử lý ngôn ngữ tự nhiên có đầu vào đã được chuẩn hóa thông qua các định dạng cố định, việc lựa chọn tham số cho mô hình lũ quét đòi hỏi sự hiểu biết về cả khoa học thủy văn lẫn đặc điểm địa phương của khu vực nghiên cứu. Quyết định lựa chọn hay loại bỏ một tham số cụ thể có thể ảnh hưởng trực tiếp đến độ chính xác và khả năng tổng quát hóa của mô hình, nhưng lại không có quy tắc định lượng rõ ràng nào để hướng dẫn vì đôi khi nó phụ thuộc vào sự sẵn có của dữ liệu. Do đó, quá trình lựa chọn đặc trưng trong mô hình AI lũ quét không chỉ đơn thuần là bài toán tối ưu hóa kỹ thuật mà còn đòi hỏi sự kết hợp giữa kiến thức chuyên ngành và khoa học dữ liệu, tuy nhiên cần đảm bảo những tham số được chọn vừa có cơ sở vật lý vững chắc vừa mang giá trị thông tin cao cho quá trình học máy.

Trên cơ sở đó, nhóm nghiên cứu đề xuất 4 nhóm dữ liệu đầu vào cho mô hình bao gồm: (1) Nhóm dữ liệu về địa hình; (2) Nhóm dữ liệu về thủy văn; (3) Nhóm dữ liệu về thực phủ; và (4) Nhóm dữ liệu về khí tượng. Trong đó, các dữ liệu cụ thể của từng nhóm có thể được liệt kê trong bảng sau đây:

Bảng 8. Các dữ liệu khuyến nghị trong phân vùng lũ quét

Nhóm dữ liệu	Tham số	Ý nghĩa trong phân vùng/xác định lũ quét	Mức độ ưu tiên (1-10)	Yếu tố lưu vực
1. Dữ liệu Địa hình	Độ dốc (Slope)	Ảnh hưởng trực tiếp đến tốc độ dòng chảy và năng lượng xói mòn. Độ dốc lớn tăng nguy cơ lũ quét	10	✓
	Độ cao (Elevation)	Xác định vị trí tương đối trong lưu vực, ảnh hưởng đến hướng dòng chảy và tích tụ nước	7	
	Hướng dốc (Aspect)	Ảnh hưởng đến điều kiện khí hậu cục bộ, bay hơi và độ ẩm đất	5	
	Độ cong địa hình (theo hướng dốc)	Ảnh hưởng đến sự tập trung hay phân tán dòng chảy	6	
	Độ cong địa hình (phương ngang)	Xác định khả năng tập trung nước trên bề mặt	6	
	Chỉ số vị trí địa hình (TPI)	Mô tả vị trí tương đối của điểm so với địa hình xung quanh	6	

Nhóm dữ liệu	Tham số	Ý nghĩa trong phân vùng/xác định lũ quét	Mức độ ưu tiên (1-10)	Yếu tố lưu vực
	Cao độ so với sông suối	Xác định khả năng tích tụ nước và nguy cơ ngập úng	10	
2. Dữ liệu Thủy văn	Khoảng cách đến sông	Ảnh hưởng đến thời gian tập trung dòng chảy và khả năng tiêu thoát	9	
	Chỉ số ẩm địa hình (TWI)	Dự đoán các khu vực có khả năng tích tụ nước cao	8	
	Chỉ số sức mạnh dòng chảy (SPI)	Đánh giá năng lượng xói mòn và vận chuyển của dòng chảy	8	
	Mật độ sông (Stream Density)	Phản ánh khả năng tiêu thoát nước của lưu vực	7	
	Chiều dài dòng chảy	Liên quan đến thời gian tập trung dòng chảy	9	
	Diện tích lưu vực	Xác định lượng nước tập trung tại điểm xét	9	
	Độ dốc lòng dẫn	Ảnh hưởng đến tốc độ và năng lượng dòng chảy	9	
	Chỉ số CN	Đánh giá khả năng sinh dòng chảy của lưu vực	10	✓
3. Dữ liệu Thực phủ	Sử dụng đất/Lớp phủ (LULC)	Ảnh hưởng đến khả năng thấm và sinh dòng chảy	8	✓
	Độ che phủ thực vật	Giảm tốc độ dòng chảy và tăng khả năng thấm	7	✓
	Chỉ số NDVI	Đánh giá mật độ thực vật, ảnh hưởng đến khả năng giữ nước	8	✓
	Chỉ số NDBI	Xác định mức độ đô thị hóa, ảnh hưởng đến bì mặt không thấm	6	✓
	Loại đất (Soil Type)	Quyết định khả năng thấm và giữ nước của đất	8	✓
	Tốc độ thấm bình quân	Ảnh hưởng trực tiếp đến lượng nước thấm vào đất	7	✓
	Độ ẩm đất (Soil Moisture)	Xác định khả năng hấp thụ nước bổ sung của đất	8	✓
	Thạch học (Lithology)	Ảnh hưởng đến tính chất thấm của nền địa chất	6	
	Mật độ rãnh xói (Gully Density)	Phản ánh mức độ xói mòn và khả năng tập trung dòng chảy	5	
4. Dữ liệu Khí tượng	Lượng mưa giờ lớn nhất	Nguyên nhân trực tiếp gây lũ quét, cường độ mưa ngắn hạn	10	✓
	Lượng mưa 3 giờ lớn nhất	Phản ánh cường độ mưa trong thời gian ngắn gây lũ quét	10	✓
	Lượng mưa 6 giờ lớn nhất	Tương ứng với thời gian đạt đỉnh của lũ quét	10	✓
	Lượng mưa 24 giờ lớn nhất	Ảnh hưởng đến trạng thái bão hòa của lưu vực	9	✓
	Nhiệt độ (Temperature)	Ảnh hưởng đến bay hơi và trạng thái độ ẩm đất	5	
	Ước lượng mưa	Dự báo nguy cơ lũ quét trong thời gian thực	8	✓

Trên cơ sở đó, tùy vào điều kiện dữ liệu và kinh nghiệm nghiên cứu, các nhà khoa học có thể linh hoạt lựa chọn hoặc bổ sung các dữ liệu phù hợp cho từng khu vực nghiên cứu nhằm phản ánh được quá trình phân vùng lũ quét một cách hiệu quả.

b. Nhận dự đoán/phân loại

Dữ liệu nhãn (label data) đóng vai trò then chốt trong việc huấn luyện các mô hình học có giám sát cho bài toán phân vùng lũ quét. Tuy nhiên, việc xây dựng tập dữ liệu nhãn phân loại cho hiện tượng lũ quét gấp phải những thách thức lớn về mặt khoa học do tính chất phức tạp và không xác định của hiện tượng này. Khác với các bài toán phân loại truyền thống có ranh giới rõ ràng, việc định nghĩa lũ quét thiếu vắng các tiêu chí định lượng thống nhất.

Cụ thể, các vấn đề chưa được giải quyết bao gồm: (1) thiếu vắng ngưỡng định lượng rõ ràng để phân biệt giữa lũ thông thường và lũ quét dựa trên các thông số thủy lực như lưu lượng đỉnh, tốc độ gia tăng mực nước, hoặc thời gian đạt đỉnh; (2) chưa có phương pháp chuẩn để xác định ranh giới không gian của vùng ảnh hưởng lũ quét, dẫn đến sự không nhất quán trong việc gán nhãn các điểm thuộc lớp 'có lũ quét' hay 'không có lũ quét'; và (3) tính nhạy cảm trong quá trình gán nhãn dựa trên đánh giá chuyên gia hoặc báo cáo thiệt hại, có thể dẫn đến sai lệch trong tập dữ liệu huấn luyện. Những hạn chế này không chỉ ảnh hưởng đến chất lượng mô hình mà còn gây khó khăn cho việc so sánh và đánh giá hiệu suất giữa các nghiên cứu khác nhau.

Như trong nghiên cứu này, việc xác định một suối có lũ lớn hay không là có thể khảo sát được, nhưng nếu vẫn trận lũ đó mà gây thiệt hại về người thì thông thường sẽ được quy về lũ quét. Việc định tính lũ có lớn hay không hoàn toàn phụ thuộc vào khảo sát và câu trả lời từ địa phương cũng như kinh nghiệm của người thu thập số liệu. Do đó rất khó để phân loại một cách chính xác nếu không có những tiêu chí cụ thể.

Có 2 cách để xác định nhãn trong phân loại lũ quét: (1) phân loại nhị phân: gán có và không tương ứng với 1 và 0 cho những vị trí đã xảy ra/chưa xảy ra lũ quét; và (2) phân loại đa lớp: gán giá trị phân theo cấp độ lũ quét cho trường hợp cụ thể (trận lũ cụ thể) trong lịch sử.

Phân loại nhị phân (có – không - Binary Classification)

Việc áp dụng phương pháp phân loại nhị phân cho bài toán phân vùng lũ quét đòi hỏi quy trình thu thập và xác thực dữ liệu nhãn nghiêm ngặt. Mỗi điểm quan sát cần được gán nhãn dựa trên bằng chứng lịch sử về sự xuất hiện/không xuất hiện của lũ quét, kết hợp với việc phân tích các điều kiện về lượng mưa tương ứng tại thời điểm sự kiện. Đặc biệt, dữ liệu lượng mưa cần được đồng bộ hóa với các biến đầu vào khác (như đã trình bày trong các phần trước) để đảm bảo tính nhất quán về mặt thời gian và không gian.

Tuy nhiên, độ tin cậy của tập dữ liệu nhãn phụ thuộc chặt chẽ vào chất lượng và độ chi tiết của thông tin lịch sử có sẵn. Trong trường hợp thiếu vắng dữ liệu quan trước đầy đủ hoặc không có khả năng xác thực chéo thông tin từ nhiều nguồn độc lập, các điểm

dữ liệu sẽ không đáp ứng tiêu chuẩn chất lượng cần thiết cho quá trình huấn luyện mô hình, dẫn đến nguy cơ dự đoán sai bởi mô hình trí tuệ nhân tạo.

Trong nghiên cứu này, các dữ liệu trước năm 2021 không có đủ độ chi tiết để xác định lượng mưa sinh lũ quét tại các khu vực đã xảy ra (do nằm ở xa khu vực đã xảy ra và có quá ít trạm để có thể tạo phân bố không gian hợp lý), do đó nghiên cứu chỉ lựa chọn phân lại cho một trận lũ cụ thể năm 2023 dựa trên việc thu thập một cách đầy đủ về lượng mưa thay vì đưa toàn bộ các sự kiện lũ quét trong quá khứ với độ tin cậy về lượng mưa bị suy giảm gây ra chất lượng mô hình không đảm bảo.

Phương pháp phân loại đa lớp (Multi-class Classification)

Phương pháp phân loại đa lớp cung cấp cách tiếp cận chi tiết hơn trong việc đánh giá mức độ nguy hiểm lũ quét thông qua việc phân chia thành các cấp độ rủi ro khác nhau (ví dụ: không có lũ, lũ rất nhỏ, lũ nhỏ, lũ trung bình, lũ lớn, lũ rất lớn). Tuy nhiên, việc triển khai phương pháp này đòi hỏi hệ thống tiêu chí phân cấp định lượng chặt chẽ và thống nhất.

Các thách thức chính trong phân loại đa lớp bao gồm: Thiết lập ngưỡng phân cấp hoặc khảo sát định tính để phân cấp. Cả hai phương pháp này đều có những khoảng trống học thuật nhất định do chưa có những nghiên cứu chuyên sâu trong phân loại. Trong thực tế, tần suất xuất hiện các sự kiện/khu vực xuất hiện lũ lớn thường thấp hơn rất nhiều so với các sự kiện/khu vực lũ nhỏ hoặc không có lũ, tạo ra tập dữ liệu không cân bằng ảnh hưởng đến hiệu suất mô hình. Do đó cần tạo thêm những dữ liệu bổ sung một cách khoa học và xử lý cân bằng lớp để có nhãn phân loại chất lượng cho mô hình trí tuệ nhân tạo. Các khu vực sông/suối có điểm lũ lớn cần xem xét thêm các điểm lân cận, bởi chính các điểm lân cận sẽ có đặc điểm lũ lân cận với điểm đang xét. Ví dụ điểm đang xét là lũ lớn thì các điểm lân cận nằm trên lòng dãy thường sẽ được gán nhãn là “rất lớn” hoặc “trung bình” – các nhãn lân cận nhãn lũ lớn vì tính liên tục của dòng chảy. Yếu tố này rất quan trọng và cần kinh nghiệm trong nghiên cứu lũ và dòng chảy để có thể định tính một cách phù hợp với quy luật tự nhiên, từ đó nâng cao chất lượng phân vùng lũ.

Dù sử dụng phương pháp phân loại nào (nhi phân hay đa lớp) thì dữ liệu biến đổi (lượng mưa) cần phải xác định một cách thận trọng và đảm bảo độ tin cậy, đặc biệt là các dữ liệu mưa ngắn hạn, là nguyên nhân trực tiếp hình thành trận lũ và có tầm quan trọng đặc biệt quyết định hiệu quả của mô hình trí tuệ nhân tạo.

2. Chuẩn hóa dữ liệu

Tùy từng loại dữ liệu mà sử dụng phương pháp chuẩn hóa khác nhau. Hầu hết các dữ liệu cần được chuẩn hóa về khoảng từ $0 \rightarrow 1$ hoặc lân cận nhằm đảm bảo tính nhất quán. Nghiên cứu này khuyến nghị sử dụng khoảng giá trị từ $0 \rightarrow 1$ sau khi được chuẩn hóa cho tất cả các dữ liệu đầu vào. Việc chuẩn hóa này sử dụng phương pháp chuẩn hóa MinMax, với công thức:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

MinMax scaling hoạt động dựa trên nguyên lý đơn giản là chuyển đổi tuyến tính toàn bộ dữ liệu vào khoảng [0,1] bằng cách sử dụng giá trị nhỏ nhất và lớn nhất trong tập dữ liệu. Tuy nhiên, phương pháp này có một điểm yếu cơ bản: nó hoàn toàn phụ thuộc vào hai giá trị cực trị này. Khi dữ liệu chứa các điểm ngoại lai (điểm đột biến - outliers), những giá trị cực trị này sẽ bị méo mó, dẫn đến việc phân lón dữ liệu bình thường bị nén vào một khoảng rất nhỏ gần 0, trong khi outliers chiếm phần lớn không gian [0,1]. Điều này làm mất đi thông tin quan trọng về sự phân bố thực sự của dữ liệu.

Vì vậy, việc sử dụng các phương pháp chuẩn hóa khác trước khi áp dụng MinMax scaling thực chất là một chiến lược đa tầng để xử lý những thách thức mà MinMax scaling gặp phải khi làm việc với dữ liệu thô. Khi áp dụng một số các phương pháp khác như Robust Scaling hay Log trước MinMax sẽ giúp dữ liệu có phân phối cân bằng hơn trước khi MinMax scaling được áp dụng. Kết quả là sau khi MinMax scaling, dữ liệu sẽ được phân bổ đều hơn trong khoảng [0,1] thay vì bị tập trung ở một đầu (tạo ra phân phối không đồng đều).

Sự kết hợp giữa các phương pháp chuẩn hóa không đơn thuần chỉ nhằm khắc phục những hạn chế mà còn phát huy tối đa thế mạnh đặc trưng của từng kỹ thuật. Mỗi phương pháp chuẩn hóa được phát triển với mục đích giải quyết một vấn đề riêng biệt trong đặc tính dữ liệu, và khi dữ liệu được áp dụng chuẩn hóa theo một trình tự phù hợp, nó sẽ trở thành một quy trình xử lý dữ liệu vững chắc và có hiệu suất cao hơn nhiều so với việc chỉ dựa vào một phương pháp chuẩn hóa đơn lẻ.

Việc lựa chọn phương pháp chuẩn hóa phù hợp phụ thuộc vào đặc tính cơ bản của từng loại dữ liệu và những thách thức cụ thể mà dữ liệu tạo ra trong quá trình phân tích. Mỗi tổ hợp phương pháp chuẩn hóa được thiết kế để giải quyết một nhóm vấn đề riêng biệt, tạo nên một hệ thống xử lý dữ liệu có tính thống nhất và hiệu quả.

Đối với dữ liệu có phân phối lệch phải như lượng mưa, khoảng cách, diện tích và mật độ, tổ hợp Log transformation kết hợp MinMax scaling trở thành lựa chọn được ưu tiên. Những loại dữ liệu này thường có đặc điểm chung là chứa nhiều giá trị nhỏ và ít giá trị lớn, tạo ra một “đuôi dài” về phía bên phải của phân phối. Khi áp dụng MinMax scaling trực tiếp, phần lớn dữ liệu sẽ bị nén vào khoảng [0, 0.2] chẳng hạn, trong khi những giá trị lớn hiếm hoi chiếm phần còn lại của thang đo. Log transformation hoạt động như một “bộ cân bằng” bằng cách nén những giá trị lớn và mở rộng những giá trị nhỏ, biến đổi phân phối lệch thành phân phối gần như chuẩn. Sau đó, MinMax scaling có thể phân bổ dữ liệu một cách đồng đều trong khoảng [0,1], đảm bảo rằng mọi khoảng giá trị đều được đại diện công bằng.

Robust Scaling kết hợp MinMax scaling được thiết kế đặc biệt cho những loại dữ liệu dễ chứa các điểm ngoại lai như cao độ, tốc độ thẩm, và mực nước. Những biến này thường có phân phối tương đối bình thường nhưng bị ảnh hưởng bởi các giá trị cực trị

do điều kiện địa lý hoặc môi trường đặc biệt. Robust Scaling sử dụng trung vị thay vì trung bình và phạm vi liên tú phân vị thay vì phân phối chuẩn, do đó không bị ảnh hưởng bởi những giá trị cực trị này. Phương pháp này giữ nguyên cấu trúc phân phối cơ bản của dữ liệu trong khi giảm thiểu tác động của các giá trị ngoại lai. Khi sau đó áp dụng MinMax scaling, dữ liệu đã được “làm sạch” sẽ được phân bố đều trong khoảng [0,1] mà không bị méo mó bởi những giá trị ngoại lai.

Z-score normalization kết hợp MinMax scaling là giải pháp lý tưởng cho những biến có thể nhận giá trị âm và dương như độ cong địa hình, TPI (Topographic Position Index), và nhiệt độ. Những biến này có đặc điểm là phân phối xung quanh một giá trị trung tâm với sự biến thiên theo cả hai hướng. Z-score transformation đưa dữ liệu về phân phối chuẩn với trung bình bằng 0 và độ lệch chuẩn bằng 1, đảm bảo rằng cả giá trị âm và dương đều được xử lý một cách công bằng. Điều này đặc biệt quan trọng vì MinMax scaling yêu cầu tất cả giá trị phải không âm để có thể ánh xạ về khoảng [0,1]. Z-score transformation thực chất “dịch chuyển” toàn bộ phân phối để đảm bảo tính đối xứng, sau đó MinMax scaling có thể áp dụng một cách hiệu quả.

Cuối cùng, MinMax scaling độc lập được dành riêng cho những loại dữ liệu đã “sạch” về mặt thống kê. Đây là những biến đã có phạm vi giá trị cố định, không chứa outliers nghiêm trọng, và có phân phối tương đối đồng đều. Các chỉ số đã được tính toán sẵn, dữ liệu categorical được encode, hoặc những biến đã qua xử lý từ các bước trước đều thuộc nhóm này. Việc áp dụng MinMax scaling trực tiếp trong trường hợp này không chỉ đơn giản và hiệu quả mà còn tránh được việc over-processing - một hiện tượng có thể làm mất thông tin quan trọng hoặc tạo ra những biến đổi không cần thiết trong dữ liệu.

Nguyên tắc chuẩn hóa thống nhất này tạo ra một framework có thể áp dụng rộng rãi, giúp đảm bảo rằng mỗi loại dữ liệu được xử lý theo cách phù hợp nhất với đặc tính riêng của nó, đồng thời vẫn đạt được mục tiêu chung là đưa tất cả về cùng một thang đo [0,1] để thuận tiện cho việc phân tích và mô hình hóa.

Trên cơ sở đó, nhóm nghiên cứu đề xuất sử dụng các phương pháp chuẩn hóa sau đây cho từng loại dữ liệu trước khi đưa dữ liệu vào mô hình trí tuệ nhân tạo để xây dựng mô hình:

Bảng 9. Các phương pháp chuẩn hóa dữ liệu được khuyến nghị theo từng loại dữ liệu

Nhóm Dữ Liệu	Tham Số	Phương Pháp Đề Xuất	Lý Do
Dữ liệu Địa hình	Độ dốc (Slope)	MinMax	Thống nhất với các tham số địa hình khác
	Độ cao (Elevation)	Robust Scaling + MinMax	Có outliers ở vùng núi cao, cần robust method
	Hướng dốc (Aspect)	MinMax	Dữ liệu tuần hoàn (0-360°), cần giữ tỷ lệ
	Độ cong (Curvature)	Z-score + MinMax	Có giá trị âm/dương, cần Z-score trước

Nhóm Dữ Liệu	Tham Số	Phương Pháp Đề Xuất	Lý Do
	Profile Curvature	Z-score + MinMax	Có giá trị âm/dương, cần Z-score trước
	Plan Curvature	Z-score + MinMax	Có giá trị âm/dương, cần Z-score trước
	Chỉ số vị trí địa hình (TPI)	Z-score + MinMax	Có giá trị âm/dương, cần Z-score
	Cao độ địa hình	Robust Scaling + MinMax	Có outliers, cần robust method
	Cao độ bình quân lưu vực	Robust Scaling + MinMax	Tương tự cao độ địa hình
	Độ cong địa hình (theo hướng dốc)	Z-score + MinMax	Có giá trị âm/dương, cần Z-score
	Độ cong địa hình (phương ngang)	Z-score + MinMax	Có giá trị âm/dương, cần Z-score
	Cao độ so với sông suối	Robust Scaling + MinMax	Có outliers, cần robust method và đưa về [0,1]
Dữ liệu Thủy văn	Khoảng cách đến sông (Distance to River)	Log + MinMax	Phân phối lệch phải, giá trị tập trung gần 0
	Chỉ số ẩm địa hình (TWI)	MinMax	Thống nhất với các chỉ số khác
	Chỉ số sức mạnh dòng chảy (SPI)	Log + MinMax	Phân phối lệch mạnh, cần log transform
	Mật độ sông (Stream Density)	Log + MinMax	Phân phối lệch phải
	Chiều dài dòng chảy	Log + MinMax	Phân phối lệch phải
	Diện tích lưu vực (Basin Area)	Log + MinMax	Sự khác biệt lớn, phân phối lệch
	Độ dốc lòng dẫn	MinMax	Thống nhất với độ dốc
	Chiều dài sông (River Length)	Log + MinMax	Phân phối lệch phải
	Khoảng cách đến sông suối	Log + MinMax	Phân phối lệch phải, giá trị tập trung gần 0
Dữ liệu Thực phủ	Sử dụng đất/Lớp phủ (LULC)	MinMax	Dữ liệu phân nhóm đã mã hóa
	Độ phủ thực vật (Vegetation Cover)	MinMax	Phạm vi 0-100%, phân phối tương đối đều
	Chỉ số NDVI	MinMax	Có phạm vi cố định (-1 đến 1)
	Chỉ số NDBI	MinMax	Có phạm vi cố định (-1 đến 1)
	Loại đất (Soil Type)	MinMax	Dữ liệu categorical đã encode
	Tốc độ thấm bình quân	Robust Scaling + MinMax	Có outliers, cần robust method
	Độ ẩm đất (Soil Moisture)	MinMax	Phạm vi tương đối cố định
	Thạch học (Lithology)	MinMax	Dữ liệu phân nhóm đã mã hóa
	Mật độ rãnh xói (Gully Density)	Log + MinMax	Phân phối lệch phải
	Chỉ số CN	MinMax	Có phạm vi cố định (30-100)
Dữ liệu Khí tượng	Lượng mưa (Rainfall)	Log + MinMax	Phân phối lệch phải rất mạnh

Nhóm Dữ Liệu	Tham Số	Phương Pháp Đề Xuất	Lý Do
	Nhiệt độ (Temperature)	Z-score + MinMax	Có thể có giá trị âm, phân phối gần chuẩn
	Ước lượng mưa (Precipitation Estimates)	Log + MinMax	Tương tự lượng mưa
	Lượng mưa giờ lớn nhất	Log + MinMax	Phân phối lệch phai rất mạnh
	Lượng mưa 3 giờ lớn nhất	Log + MinMax	Phân phối lệch phai rất mạnh
	Lượng mưa 6 giờ lớn nhất	Log + MinMax	Phân phối lệch phai rất mạnh
	Lượng mưa 24 giờ lớn nhất	Log + MinMax	Phân phối lệch phai rất mạnh

3.2.2 Xây dựng mô hình trí tuệ nhân tạo trong phân vùng lũ quét

1. Lựa chọn mô hình trí tuệ nhân tạo

Bài toán phân vùng lũ quét mang tính phức tạp cao do đặc thù của dữ liệu đầu vào có nhiều chiều và đa dạng về nguồn gốc. Dữ liệu nghiên cứu gồm ba nhóm chính: dữ liệu địa không gian như địa hình, cách sử dụng đất, loại đất; dữ liệu viễn thám bao gồm ảnh vệ tinh và mô hình số độ cao; cùng với dữ liệu lượng mưa theo nhiều khoảng thời gian khác nhau. Sự kết hợp này tạo ra bộ dữ liệu có nhiều dạng thức, vừa mang tính chất bảng số liệu, vừa có tính chất không gian và thời gian. Điều này đòi hỏi các mô hình trí tuệ nhân tạo phải có khả năng xử lý đồng thời nhiều loại dữ liệu khác nhau và nắm bắt được mối quan hệ phi tuyến phức tạp giữa các yếu tố môi trường và khả năng xảy ra lũ quét.

Đặc điểm nổi bật của dữ liệu lũ quét là tính chất mất cân bằng, do các sự kiện lũ quét xảy ra với tần suất thấp nhưng mức độ tác động rất lớn. Hơn thế nữa, mối quan hệ giữa các yếu tố đầu vào và kết quả đầu ra có tính phi tuyến cao, với nhiều ngưỡng và tương tác phức tạp giữa các biến số. Chẳng hạn, cùng một lượng mưa có thể gây ra lũ quét ở vùng có địa hình dốc và đất không thấm nước, nhưng lại không gây nguy hiểm ở vùng đất bằng với khả năng thoát nước tốt.

Trong bài toán phân vùng lũ quét từ dữ liệu địa không gian, viễn thám và lượng mưa đa thời đoạn, mỗi loại mô hình trí tuệ nhân tạo thể hiện những đặc điểm và khả năng khác biệt rõ rệt. Các mô hình học máy truyền thống như LightGBM và Random Forest nổi bật với khả năng xử lý hiệu quả dữ liệu dạng bảng phức tạp, có thể tự động học được tầm quan trọng của từng biến và xử lý tốt các mối quan hệ phi tuyến giữa các yếu tố địa hình, khí tượng mà không cần quá nhiều tiền xử lý dữ liệu. Random Forest thể hiện tính ổn định cao và khả năng chống nhiễu tốt, đặc biệt phù hợp khi dữ liệu có nhiều dữ liệu thiếu hoặc ngoại lai, đồng thời cung cấp khả năng giải thích rõ ràng về vai trò của từng biến trong việc dự báo nguy cơ lũ quét.

Bên cạnh đó, các mô hình học sâu lại thể hiện những ưu thế khác biệt. Mạng nơ-ron sâu (DNN) có thể giải quyết trong việc học các mẫu phức tạp và tương tác phi tuyến giữa nhiều biến, có khả năng tự động phát hiện các mối quan hệ ẩn giữa lượng mưa các

thời đoạn khác nhau với đặc điểm địa hình, tạo ra các biểu diễn đặc trưng phong phú cho bài toán phân loại. Mạng nơ-ron tích chập (CNN) thể hiện sức mạnh vượt trội trong việc xử lý dữ liệu không gian, có thể tự động trích xuất các đặc trưng địa hình từ mô hình số độ cao (DEM) và ảnh viễn thám mà không cần trích xuất đặc trưng thủ công, đặc biệt hiệu quả trong việc nhận dạng các mẫu không gian như lưu vực, độ dốc phức tạp hay mạng lưới thủy văn. Mạng nơ-ron hồi tiếp (LSTM, GRU) lại chuyên biệt trong việc xử lý dữ liệu chuỗi thời gian, có khả năng ghi nhớ và học từ các mẫu lượng mưa trong quá khứ để dự báo xu hướng và nguy cơ lũ quét trong tương lai.

Trong khi Support Vector Machine thể hiện tính robust cao với dữ liệu nhiều chiều và khả năng tạo ra các ranh giới quyết định rõ ràng giữa các vùng có và không có nguy cơ lũ quét, thì Decision Tree lại nổi bật với tính diễn giải cao, có thể tạo ra các quy tắc đơn giản và dễ hiểu cho việc đánh giá nguy cơ lũ quét dựa trên các ngưỡng cụ thể của lượng mưa và đặc điểm địa hình.

a. Nhóm mô hình học máy

Trong nhóm mô hình học máy truyền thống, các thuật toán tăng cường độ dốc như LightGBM, XGBoost và CatBoost nổi lên như những ứng cử viên hàng đầu với mức độ phù hợp rất cao (Bảng 10). Những mô hình này thể hiện khả năng xuất sắc trong việc xử lý dữ liệu dạng bảng kết hợp, với ưu điểm vượt trội trong việc xử lý dữ liệu thiếu và cung cấp thông tin về tầm quan trọng của từng đặc trưng một cách có ý nghĩa. Đặc biệt, những mô hình này rất hiệu quả trong việc nắm bắt các mối quan hệ phi tuyến phức tạp giữa các yếu tố địa lý, khí tượng và thủy văn.

Rừng ngẫu nhiên (RF) và các biến thể của nó cũng thể hiện hiệu suất cao với mức độ phù hợp cao. Điểm mạnh của rừng ngẫu nhiên nằm ở khả năng xử lý ổn định với các giá trị bất thường và tính năng lựa chọn đặc trưng tự động, điều này đặc biệt quan trọng khi làm việc với dữ liệu đa nguồn có thể chứa nhiều nhiễu và bất thường. Khả năng giải thích tầm quan trọng đặc trưng của rừng ngẫu nhiên cũng tạo thuận lợi cho việc hiểu rõ yếu tố nào đóng vai trò quan trọng nhất trong việc hình thành lũ quét, từ đó hỗ trợ các nhà quản lý trong việc đưa ra chính sách phòng chống thiên tai phù hợp.

Các mô hình đơn giản hơn như máy véc-tơ hỗ trợ và cây quyết định có mức độ phù hợp trung bình, thích hợp cho các tình huống cần tính giải thích cao hoặc khi lượng dữ liệu còn hạn chế. Tuy nhiên, độ chính xác thấp hơn và khả năng mở rộng hạn chế khiến mô hình này không được ưa chuộng trong các ứng dụng thực tế quy mô lớn. Hồi quy logistic và k-láng giềng gần nhất có mức độ phù hợp thấp do không thể nắm bắt được độ phức tạp của dữ liệu lũ quét, mặc dù có thể được sử dụng như các mô hình cơ sở để so sánh hiệu suất.

Căn cứ trên những phân tích, nhóm nghiên cứu khuyến nghị sử dụng nhóm mô hình Gradient Boosting và Rừng ngẫu nhiên cho phân vùng lũ quét. Ưu điểm lớn nhất chính là khả năng nắm bắt được các mối quan hệ phi tuyến phức tạp của các loại mô hình này, điều mà có thể chỉ được nhận biết bằng một số các đặc điểm.

b. Nhóm mô hình học sâu

Nhóm mô hình học sâu mang lại những khả năng đặc biệt phù hợp với tính chất đa dạng của dữ liệu lũ quét. Mạng nơ-ron tích chập (CNN) với các kiến trúc hiện đại như ResNet, EfficientNet và DenseNet thể hiện khả năng cao trong việc xử lý dữ liệu không gian. Điểm mạnh nổi bật của mạng tích chập là khả năng tự động trích xuất đặc trưng từ dữ liệu viễn thám và mô hình số độ cao, giúp phát hiện các mẫu địa hình phức tạp mà con người có thể bỏ qua. Khả năng này đặc biệt quan trọng trong phân tích lũ quét, nơi các đặc trưng địa hình nhỏ như độ dốc cục bộ, hướng dốc, và mức độ gồ ghề có thể ảnh hưởng lớn đến khả năng thoát nước.

Mạng nơ-ron hồi quy và các biến thể như bộ nhớ dài ngắn hạn (RNN), cỗng hồi quy mang lại khả năng xử lý chuỗi thời gian với mức độ phù hợp trung bình cao. Những mô hình này xuất sắc trong việc phân tích các mẫu thời gian của dữ liệu lượng mưa, có thể phát hiện các xu hướng và chu kỳ trong dữ liệu lượng mưa mà có thể báo hiệu nguy cơ lũ quét. Tuy nhiên, hiệu quả của mạng hồi quy phụ thuộc nhiều vào độ dài chuỗi thời gian và chất lượng dữ liệu thời gian, do đó cần có chiến lược tiền xử lý dữ liệu phù hợp.

Mạng nơ-ron sâu (DNN) nhiều lớp ẩn cũng thể hiện hiệu suất ẩn tượng với mức độ phù hợp cao. Các kiến trúc như mạng nơ-ron sâu đa lớp và TabNet được thiết kế đặc biệt để xử lý dữ liệu dạng bảng, có khả năng xử lý tốt các đặc trưng hỗn hợp từ nhiều nguồn khác nhau. Điểm mạnh của nhóm mô hình này nằm ở khả năng học các mối quan hệ phi tuyến phức tạp thông qua nhiều lớp ẩn, cho phép mô hình nắm bắt được những tương tác tinh tế giữa các yếu tố địa lý, khí tượng và thủy văn mà các mô hình đơn giản khó có thể phát hiện. Trong nhóm này, nhóm nghiên cứu khuyến nghị sử dụng mạng CNN và DNN trong phân vùng lũ quét.

Bảng 10. Các mô hình trí tuệ nhân tạo và đánh giá sự phù hợp của các mô hình trong phân vùng lũ quét

Nhóm mô hình	Mô hình cụ thể	Đặc điểm với dữ liệu của bạn	Ứng dụng cho lũ quét	Mức độ phù hợp	Lý do đánh giá
Học máy					
Gradient Boosting	- LightGBM - XGBoost - CatBoost - Gradient Boosting Regressor	Xuất sắc với dữ liệu dạng bảng kết hợp, xử lý tốt dữ liệu thiếu, đánh giá các đặc trưng quan trọng	- Phân vùng nguy cơ lũ quét - Dự báo cường độ lũ quét - Xếp hạng mức độ nguy hiểm	RẤT CAO	Tối ưu cho dữ liệu kết hợp không thời gian, xử lý tốt các mối quan hệ phi tuyến
Random Forest	- Random Forest Classifier - Extra Trees - Balanced Random Forest	Xử lý hiệu quả dữ liệu nhiều chiều, bền vững trước các giá trị ngoại lai, tự động lựa chọn đặc trưng.	- Phân loại vùng nguy cơ cao/thấp - Xác định yếu tố quan trọng nhất	CAO	Ôn định với dữ liệu đa nguồn, dễ giải thích các yếu tố quan trọng.
Support Vector Machine	- SVM với RBF kernel - Linear SVM - Nu-SVM	Hiệu quả với dữ liệu có chiều cao, bền vững trước các giá trị ngoại lai.	- Phân loại nhị phân nguy cơ lũ quét - Phát hiện ranh giới vùng nguy hiểm	TRUNG BÌNH	Tốt với dữ liệu ít, nhưng khó triển khai với quy mô lớn
Decision Tree	- CART - C4.5 - ID3	Dễ diễn giải, xử lý được các loại dữ liệu hỗn hợp	- Tạo các quy tắc quyết định đơn giản - Phân tích nguyên nhân lũ quét	TRUNG BÌNH	Dễ hiểu nhưng dễ quá khớp với dữ liệu phức tạp
Logistic Regression	- Logistic Regression - Ridge/Lasso Logistic	Đơn giản, nhanh, cho kết quả đầu ra xác suất	- Dự báo xác suất xảy ra lũ quét - Mô hình cơ sở	THẤP	Quá đơn giản cho dữ liệu phức tạp, không bắt được mối quan hệ phi tuyến
K-Nearest Neighbors	- KNN Classifier - Weighted KNN	Dựa trên độ tương đồng của các điểm lân cận	- Nội suy không gian tốt - Dự báo cục bộ	THẤP	Không phù hợp với dữ liệu đa chiều, chậm
Ensemble Methods	- Voting Classifier - Stacking - Blending	Kết hợp nhiều mô hình	- Tăng độ chính xác tổng thể - Giảm hiện tượng quá khớp	RẤT CAO	Kết hợp ưu điểm của nhiều models

Nhóm mô hình	Mô hình cụ thể	Đặc điểm với dữ liệu của bạn	Ứng dụng cho lũ quét	Mức độ phù hợp	Lý do đánh giá
Học sâu					
Convolutional Neural Network	- ResNet50/101 - EfficientNet - DenseNet - VGG16/19	Xuất sắc với dữ liệu spatial/ảnh viễn thám, tự động trích xuất đặc trưng	- Phân tích địa hình từ DEM - Trích xuất đặc trưng từ ảnh vệ tinh - Nhận diện được các mẫu	CAO	Tối ưu cho dữ liệu không gian, có thể kết hợp với dữ liệu dạng bảng
Recurrent Neural Network	- LSTM - GRU - BiLSTM - Attention-LSTM	Xử lý dữ liệu tuần tự kiểu time-serial (mưa – mực nước – lưu lượng...)	- Dự báo từ chuỗi lượng mưa - Phân tích mẫu thời gian - Dự báo ngắn hạn	TRUNG BÌNH CAO	Tốt cho dữ liệu thời gian nhưng cần nhiều bước thời gian
Hybrid CNN-RNN	- CNN+LSTM - ConvLSTM - CNN+GRU	Kết hợp không – thời gian	- Phân tích đồng thời không gian và thời gian - Dự báo không gian – thời gian	RẤT CAO	Tốt cho dữ liệu không gian kết hợp với dữ liệu thời gian liên tục.
Deep Neural Network (DNN)	- Deep MLP - Feedforward DNN - Wide & Deep - Deep & Cross Network - TabNet - NODE (Neural ODEs)	Kiến trúc đa lớp kết nối toàn phần, tối ưu cho dữ liệu dạng bảng phức tạp	- Phân vùng lũ quét từ các đặc trưng hỗn hợp - Nhận dạng mẫu phi tuyến - Học tương tác đặc trưng	RẤT CAO	Xuất sắc cho dữ liệu dạng bảng với các mối quan hệ phức tạp, kiến trúc linh hoạt
Transformer	- Vision Transformer - Swin Transformer - TabTransformer	Cơ chế chú ý, xử lý dữ liệu đa phương thức	- Kết hợp đa phương thức - Phụ thuộc tâm xa	TRUNG BÌNH	Cần nhiều dữ liệu, phức tạp cho quy mô bài toán
Graph Neural Networks	- GraphSAGE - GCN - GAT	Xử lý các mối quan hệ không gian	- Mô hình phụ thuộc không gian - Kết nối lưu vực	TRUNG BÌNH	Phù hợp nếu có cấu trúc đồ thị rõ ràng

2. Lựa chọn phương pháp tối ưu cho mô hình trí tuệ nhân tạo

a. Tối ưu siêu tham số

Tối ưu siêu tham số (Hyperparameter Optimization) là quá trình tìm kiếm các giá trị tối ưu cho những tham số không được học tự động trong quá trình huấn luyện mô hình. Đây là bước quan trọng để nâng cao hiệu suất dự đoán của mô hình học máy và học sâu. Trong bài toán phân vùng lũ quét, việc tối ưu siêu tham số đặc biệt quan trọng do tính phức tạp của dữ liệu địa không gian và sự tương tác phức tạp giữa các yếu tố địa hình, thuỷ văn và khí tượng.

Đối với bài toán phân vùng lũ quét từ dữ liệu địa không gian và lượng mưa, việc lựa chọn mô hình phù hợp và tối ưu siêu tham số sẽ quyết định khả năng dự đoán chính xác các khu vực có nguy cơ lũ quét. Dữ liệu đầu vào phong phú với nhiều đặc trưng bao gồm các yếu tố địa hình, thuỷ văn, thổ nhưỡng, thực vật và khí tượng tạo ra không gian đặc trưng đa chiều, đòi hỏi việc điều chỉnh cẩn thận các siêu tham số để mô hình có thể học được các mối quan hệ phi tuyến phức tạp.

Quá trình tối ưu siêu tham số thường được thực hiện thông qua các phương pháp như Grid Search, Random Search, Bayesian Optimization hoặc các thuật toán tối ưu tiến hóa. Trong bối cảnh phân vùng lũ quét, việc sử dụng Cross-validation với chia dữ liệu theo không gian sẽ đảm bảo tính tổng quát hóa tốt hơn do tính chất tự tương quan không gian của dữ liệu địa lý.

Bảng 11. Khuyến nghị tối ưu siêu tham số cho nhóm mô hình Gradient Boosting

Siêu Tham Số	Mức Độ Ưu Tiên	Khoảng Tối Ưu Khuyến Nghị	Giải thích và Ảnh hưởng
n_estimators	RẤT CẦN THIẾT	100-1500	Ảnh hưởng trực tiếp đến hiệu suất.
learning_rate	RẤT CẦN THIẾT	0.01-0.2	Quyết định tốc độ học và khả năng hội tụ.
max_depth	CẦN THIẾT	3-15	Kiểm soát overfitting.
min_child_weight	CẦN THIẾT	1-15	Quan trọng với dữ liệu không cân bằng.
subsample	KHUYẾN NGHỊ	0.7-1.0	Giảm overfitting, tăng tốc training.
colsample_bytree	KHUYẾN NGHỊ	0.7-1.0	Với ít features, có thể để 0.8-1.0.
reg_alpha	TÙY CHỌN	0-5	L1 regularization. Chỉ thử nếu overfitting
reg_lambda	TÙY CHỌN	0-5	L2 regularization.

Bảng 12. Khuyến nghị tối ưu siêu tham số cho mô hình Random Forest

Siêu Tham Số	Mức Độ Ưu Tiên	Khoảng Tối Ưu Khuyến Nghị	Giải thích và Ảnh hưởng
n_estimators	RẤT CẦN THIẾT	100-500	Default (100) thường quá thấp. Cải thiện đáng kể hiệu suất. Khuyến nghị: 200-300

Siêu Tham Số	Mức Độ Ưu Tiên	Khoảng Tối Ưu Khuyến Nghị	Giải Thích và Ảnh Hưởng
max_depth	RÂT CẦN THIẾT	5-25	Default (None) dễ overfitting. Với dữ liệu địa không gian: 10-15 tối ưu
min_samples_split	CẦN THIẾT	2-15	Default (2) có thể quá nhỏ. Với dữ liệu lũ quét: 5-10 giúp tránh overfitting
max_features	CẦN THIẾT	'sqrt', 4-8, 0.4-0.7	Default ('sqrt') thường tốt. Với 15-20 features có thể thử 0.5-0.6
min_samples_leaf	KHUYẾN NGHỊ	1-8	Default (1) có thể để nguyên. Thủ 2-4 nếu overfitting
bootstrap	KHUYẾN NGHỊ	True/False	Default (True) thường tối ưu. Hiếm khi cần thay đổi
class_weight	KHUYẾN NGHỊ	'balanced'	Bạn đã đặt đúng cho dữ liệu không cân bằng

Bảng 13. Khuyến nghị tối ưu siêu tham số cho mô hình CNN

Siêu Tham Số	Mức Độ Ưu Tiên	Khoảng Tối Ưu Khuyến Nghị	Giải Thích và Ảnh Hưởng
Số lớp conv	RÂT CẦN THIẾT	2-5 lớp	Quyết định khả năng học pattern. Với ít features: 3-4 lớp thường tối ưu
Số filters	RÂT CẦN THIẾT	16-256 mỗi lớp	Ảnh hưởng lớn đến bộ nhớ. Bắt đầu: $32 \rightarrow 64 \rightarrow 128$
Learning rate	RÂT CẦN THIẾT	$1e-4 \div 1e-3$	Quan trọng nhất trong deep learning. Thủ: $1e-4, 5e-4, 1e-3$
Batch size	CẦN THIẾT	32-128	Ảnh hưởng đến gradient stability. Thủ: 32, 64, 128
Dropout rate	CẦN THIẾT	0.2-0.5	Quan trọng để tránh overfitting. Thủ: 0.3, 0.4, 0.5
Kernel size	KHUYẾN NGHỊ	3x3, 5x5, 7x7	3x3 thường tốt. Chỉ thử 5x5 nếu cần pattern lớn hơn
Padding	TÙY CHỌN	'same'	'same' thường tối ưu cho dữ liệu địa không gian

Bảng 14. Khuyến nghị tối ưu siêu tham số cho mô hình DNN

D Tham Số	Mức Độ Ưu Tiên	Khoảng Tối Ưu Khuyến Nghị	Giải Thích và Ảnh Hưởng
Số hidden layers	RÂT CẦN THIẾT	2-6 layers	Quyết định model complexity. Với 15-20 features: 3-4 layers

D Tham Số	Mức Độ Ưu Tiên	Khoảng Tối Ưu Khuyên Nghị	Giải Thích và Ảnh Hưởng
Số neurons per layer	RẤT CẦN THIẾT	32-256	Ảnh hưởng lớn đến capacity. Thử: 64→128→64 hoặc 128→64→32
Learning rate	RẤT CẦN THIẾT	1e-4 to 1e-3	Quan trọng nhất. Thử: 1e-4, 2e-4, 5e-4, 1e-3
Dropout rate	CẦN THIẾT	0.1-0.5	Tránh overfitting. Tăng dần qua layers: 0.2→0.3→0.4
Batch size	CẦN THIẾT	32-256	Ảnh hưởng training stability. Thử: 64, 128
Activation function	KHUYẾN NGHỊ	ReLU, Leaky ReLU	ReLU thường tốt. Chỉ thử Leaky ReLU nếu dying ReLU
L2 regularization	KHUYẾN NGHỊ	1e-4 to 1e-3	Giúp tránh overfitting. Thử: 1e-4, 5e-4, 1e-3
Batch normalization	TÙY CHỌN	True/False	True thường tốt hơn nhưng tăng complexity
Optimizer	TÙY CHỌN	Adam	Adam với default params thường tối ưu

Mức độ “rất cần thiết” là mức độ quan trọng, cần phải tối ưu nhằm đạt hiệu suất tốt hơn việc sử dụng mặc định. Đối với các mô hình cây quyết định, số lượng cây nên được tăng lên lớn hơn 100 cây để mang lại kết quả tốt hơn, bên cạnh đó, độ sâu cây tùy thuộc vào số lượng tham số đầu vào, nếu nhiều tham số (hơn 20 tham số đầu vào) có thể sử dụng giới hạn khoảng từ 15-20 tầng, nếu ít tham số hơn có thể thử với giới hạn 10-15 tầng.

Có một số quy tắc có thể ước tính được độ sâu cây (số tầng) dựa trên các tham số đầu vào. Nếu số lượng tham số càng lớn và phức tạp thì độ sâu cây càng lớn nhưng không quá số lượng tham số. Các quy tắc bao gồm quy tắc Logarithm, Square Root và Linear. Với quy tắc Logarithm và Square Root, một hệ số bổ sung thường từ 3-5 sẽ được thêm vào để xác định tùy vào độ phức tạp của dữ liệu (mang tính phi tuyến cao hay thấp). Ví dụ nếu có 16 tham số với quy tắc Square Root thì số tầng có thể xác định là $\sqrt{\text{số_features}} + \text{constant} = 4 + 5 = 9$. Trong đó 4 là căn bậc 2 của 16 và 5 là hệ số bổ sung (với dữ liệu có tính phi tuyến cao). Tương tự với quy tắc Logarithm sẽ có số tầng là $\log_2(\text{số_features}) + \text{constant} = 9$. Riêng đối với quy tắc Linear, số tầng sẽ có quy tắc là từ $\text{số_features}/3$ đến $\text{số_features}/2$. Trong trường hợp là 16 tham số đầu vào thì số tầng từ 5 – 8 tầng.

Đối với mô hình học sâu, điều chỉnh số lớp và tốc độ học là điều quan trọng, đặc biệt, tốc độ học nên được cài đặt tùy chỉnh một cách tự động giảm dần khi độ chính xác không được cải thiện sau nhiều bước đào tạo.

b. Tối ưu dữ liệu

Tối ưu dữ liệu là một bước tối ưu vô cùng quan trọng để có thể đạt được hiệu suất cao trong xây dựng mô hình trí tuệ nhân tạo. Công tác này bao gồm: (1) tăng cường dữ liệu; (2) làm sạch và tiền xử lý dữ liệu; (3) tạo thêm đặc trưng dữ liệu từ dữ liệu gốc; và (4) xác định được các đặc trưng quan trọng nhất.

Tăng cường dữ liệu là việc bổ sung thêm các nhãn (đầu ra) hoặc tạo thêm dữ liệu huấn luyện có chất lượng đảm bảo. Đối với mô hình trí tuệ nhân tạo, càng nhiều dữ liệu có chất lượng thì mô hình càng đảm bảo độ tin cậy và khả năng dự đoán đúng. Trong nghiên cứu này, nhóm nghiên cứu đã tạo thêm các dữ liệu nhãn đầu ra cho mô hình bằng việc xác định các điểm lân cận trên dòng chính (với khoảng cách vài pixels) và gán giá trị nguy cơ tương ứng. Điều này phù hợp với nguyên tắc ảnh hưởng của lũ quét, không làm mất đi bản chất của việc dự đoán nguy cơ lũ. Đối với việc dự đoán phân loại hình ảnh (như bản đồ sử dụng đất dựa trên ảnh vệ tinh), các phương pháp biến đổi như xoay, lật, hay thay đổi độ sáng cũng là một trong những cách giúp tăng cường dữ liệu hiệu quả.

Làm sạch và tiền xử lý dữ liệu bao gồm các công tác chuẩn hóa hay xử lý dữ liệu bị mất. Phương pháp chuẩn hóa rất quan trọng để làm sạch dữ liệu, giúp dữ liệu biểu thị tốt hơn và phân bố đều hơn. Điều này còn giúp cho mô hình dễ nhận diện phân loại và hội tụ tốt hơn.

Tạo thêm đặc trưng từ dữ liệu gốc và xác định các đặc trưng quan trọng nhất như các dữ liệu độ dốc, độ cong... từ DEM địa hình hay các dữ liệu về chiều dài dòng chảy, độ dốc lòng dẫn bằng phương pháp thủy văn... Trên thực tế, việc tạo thêm đặc trưng dữ liệu từ dữ liệu gốc thể hiện sự hiểu biết sâu sắc của người xây dựng mô hình trí tuệ nhân tạo, đưa các vấn đề chuyên môn, gốc rễ vào xây dựng mô hình chứ không đơn thuần chỉ là “ném dữ liệu vào và chạy”. Điều này có ý nghĩa vô cùng quan trọng, đặc biệt là trong các chuyên ngành hẹp, nơi cần chuyên gia thực sự để cải tiến thay vì các chuyên gia công nghệ.

c. Tối ưu kiến trúc và hiệu suất mô hình

Kiến trúc mô hình có thể được tối ưu nhằm nâng cao độ chính xác. Một số phương pháp có thể kể đến bao gồm kết hợp nhiều mô hình để tăng hiệu suất hay tự động tìm kiếm kiến trúc mô hình tối ưu, điều chỉnh tốc độ học theo thời gian, dừng huấn luyện khi mô hình bắt đầu overfitting, lưu trữ phiên bản mô hình có độ chính xác cao nhất, sử dụng GPU, TPU để tăng tốc huấn luyện... Vấn đề tối ưu kiến trúc và hiệu suất phần lớn dựa vào kinh nghiệm của người lập mô hình trí tuệ nhân tạo mà không bị giới hạn bởi bất cứ ràng buộc tiêu chuẩn nào.

3.2.3 Đánh giá sự phù hợp của mô hình

Một trong những thách thức lớn nhất trong ứng dụng trí tuệ nhân tạo cho dự báo thiên tai là sự mâu thuẫn tiềm ẩn giữa độ chính xác thống kê và tính phù hợp thực tiễn. Nhiều mô hình có thể đạt được độ chính xác cao trong các phép đo kiểm định truyền thống như accuracy, precision, recall, nhưng lại tạo ra những kết quả không hợp lý về mặt vật lý hoặc không phù hợp với thực tế địa lý. Điều này đặc biệt nghiêm trọng trong bài toán phân vùng lũ quét, nơi mà các dự báo không chỉ cần chính xác mà còn phải tuân thủ các quy luật tự nhiên và thể hiện tính liên tục không gian hợp lý.

Vấn đề cốt lõi nằm ở chỗ các mô hình học máy truyền thống thường xử lý từng điểm dữ liệu một cách độc lập, không quan tâm đến mối quan hệ không gian giữa các vùng lân cận. Kết quả là có thể xuất hiện những “đảo” nguy cơ cao được bao quanh bởi các vùng nguy cơ thấp một cách không hợp lý, hoặc ngược lại, những vùng nguy cơ thấp bất thường nằm giữa các khu vực có nguy cơ cao. Điều này không chỉ làm giảm độ tin cậy của mô hình mà còn có thể dẫn đến những quyết định sai lầm trong công tác phòng chống thiên tai, gây ra hậu quả nghiêm trọng cho cộng đồng.

Trong lĩnh vực thủy văn, nguyên lý cơ bản là nước luôn chảy từ nơi cao xuống nơi thấp theo các tuyến thoát nước tự nhiên, tạo thành một hệ thống mạng lưới sông suối có tính liên tục và phân cấp rõ ràng. Lũ quét không phải là hiện tượng xảy ra ngẫu nhiên tại một điểm cô lập, mà là kết quả của quá trình tích tụ và vận chuyển nước mưa trong toàn bộ lưu vực. Do đó, nguy cơ lũ quét tại một vị trí bất kỳ phải phù hợp với điều kiện thủy văn của các vùng thượng lưu và hạ lưu, tạo thành một vùng biến đổi nguy cơ có tính logic và liên tục.

Tuy nhiên, nhiều mô hình trí tuệ nhân tạo hiện tại, đặc biệt là các mô hình học máy truyền thống như Random Forest, SVM, hoặc thậm chí một số mô hình học sâu được thiết kế không phù hợp, có xu hướng tạo ra các dự báo “rời rạc” không tuân thủ nguyên lý liên tục thủy văn này. Chẳng hạn, một mô hình có thể dự báo nguy cơ lũ quét cao tại một điểm nằm giữa lưu vực nhưng lại dự báo nguy cơ thấp tại các điểm ngay lân cận phía trên hoặc phía dưới thuộc lòng dẫn. Điều này không chỉ vi phạm các quy luật vật lý cơ bản mà còn làm mất niềm tin của các chuyên gia thủy văn và người sử dụng vào khả năng ứng dụng thực tế của mô hình.

Bài toán phân vùng lũ quét thường được tiếp cận như một bài toán phân loại, trong đó mỗi vùng được gán nhãn thuộc một trong các mức độ nguy cơ khác nhau như “rất thấp”, “thấp”, “trung bình”, “cao”... Cách tiếp cận này, mặc dù đơn giản và dễ hiểu, lại tạo ra những thách thức đặc biệt về tính phù hợp. Trong thực tế, nguy cơ lũ quét là một đại lượng liên tục, thay đổi một cách mềm mại theo không gian và thời gian. Việc chia cắt thành các mức độ rời rạc có thể tạo ra những “đường biên” cứng nhắc giữa các vùng có mức độ nguy cơ khác nhau, dẫn đến hiện tượng hai vùng liền kề có thể được phân loại vào hai mức độ nguy cơ hoàn toàn khác nhau mặc dù điều kiện địa lý và khí tượng của chúng rất tương tự nhau.

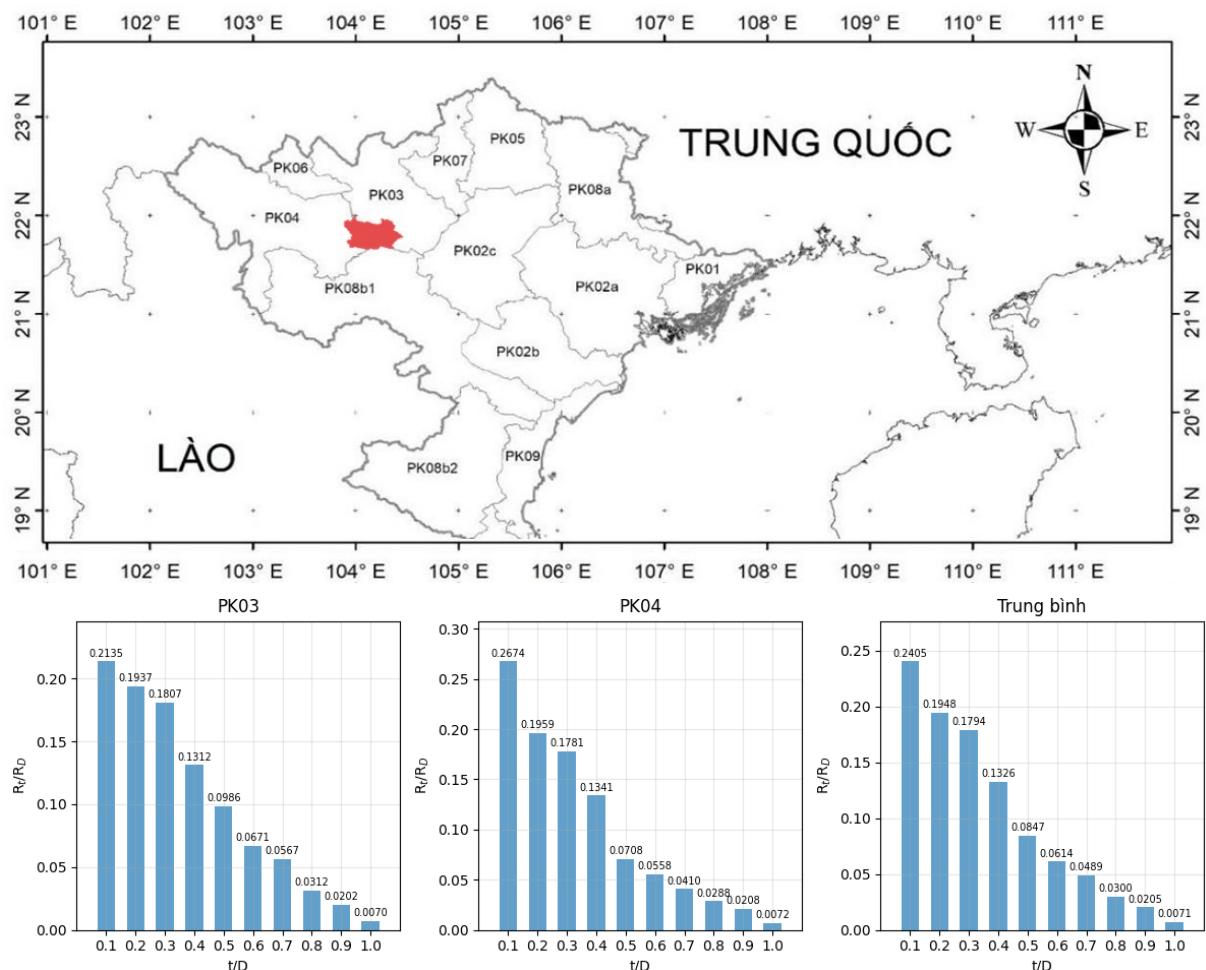
Vấn đề này được khuếch đại khi các mô hình phân loại tập trung vào việc tối ưu hóa độ chính xác tổng thể mà không quan tâm đến tính hợp lý của các quyết định phân loại tại biên giới giữa các lớp. Một mô hình có thể đạt được độ chính xác 95% nhưng lại tạo ra những “vùng đồng tâm” kỳ lạ với nguy cơ cao được bao quanh bởi nguy cơ thấp, hoặc những “hành lang” nguy cơ thấp cắt ngang qua vùng địa hình đồng nhất. Những bất thường này không chỉ làm giảm độ tin cậy của mô hình mà còn gây khó khăn cho việc giải thích và truyền đạt kết quả đến cộng đồng.

Để giải quyết những thách thức về tính phù hợp, xu hướng phát triển hiện tại đang hướng tới các “physics-informed models” - những mô hình có ý thức về các quy luật vật lý. Những mô hình này không chỉ học từ dữ liệu mà còn được thiết kế để tuân thủ các nguyên lý cơ bản của thủy văn, khí tượng và địa chất. Chẳng hạn, các mô hình này có thể được thiết kế để đảm bảo rằng nguy cơ lũ quét tại các điểm hạ lưu phải phù hợp với điều kiện tại thượng lưu, hoặc nguy cơ tại các vùng có cùng đặc điểm địa hình và khí hậu phải có mức độ tương tự nhau. Và điều quan trọng nhất là kết quả dự đoán/phân loại cần được đánh giá lại bởi chính các chuyên gia trong lĩnh vực.

3.3. Xây dựng bản đồ phân vùng lũ quét theo kịch bản mưa

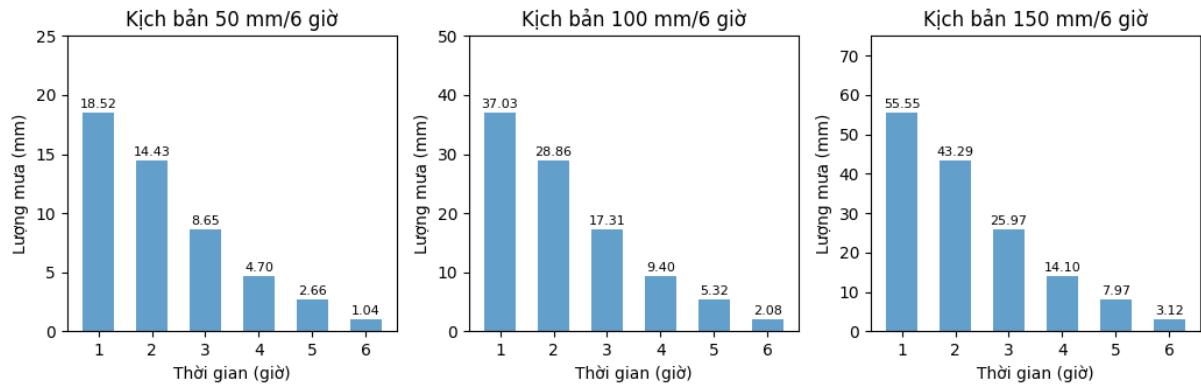
3.3.1 Xây dựng kịch bản mưa

Kết quả nghiên cứu phía trên là kết quả phân vùng lũ quét cho khu vực Mù Cang Chải trận lũ 8/2023. Trên cơ sở đó, nghiên cứu tiếp tục triển khai xây dựng bản đồ phân vùng lũ quét cho các kịch bản mưa giả định. Do có rất nhiều kịch bản mưa, nghiên cứu này sử dụng kịch bản mưa bất lợi nhất theo TCVN 13615:2022 làm cơ sở xác định biểu đồ phân bố tương ứng với lượng mưa thời đoạn 6 giờ cho các dự báo với tổng lượng mưa lần lượt là 50mm; 100mm; và 150mm. Mô hình CNN được áp dụng.



Hình 31. Phân vùng mưa rào theo TCVN 13615:2022 và khu vực nghiên cứu

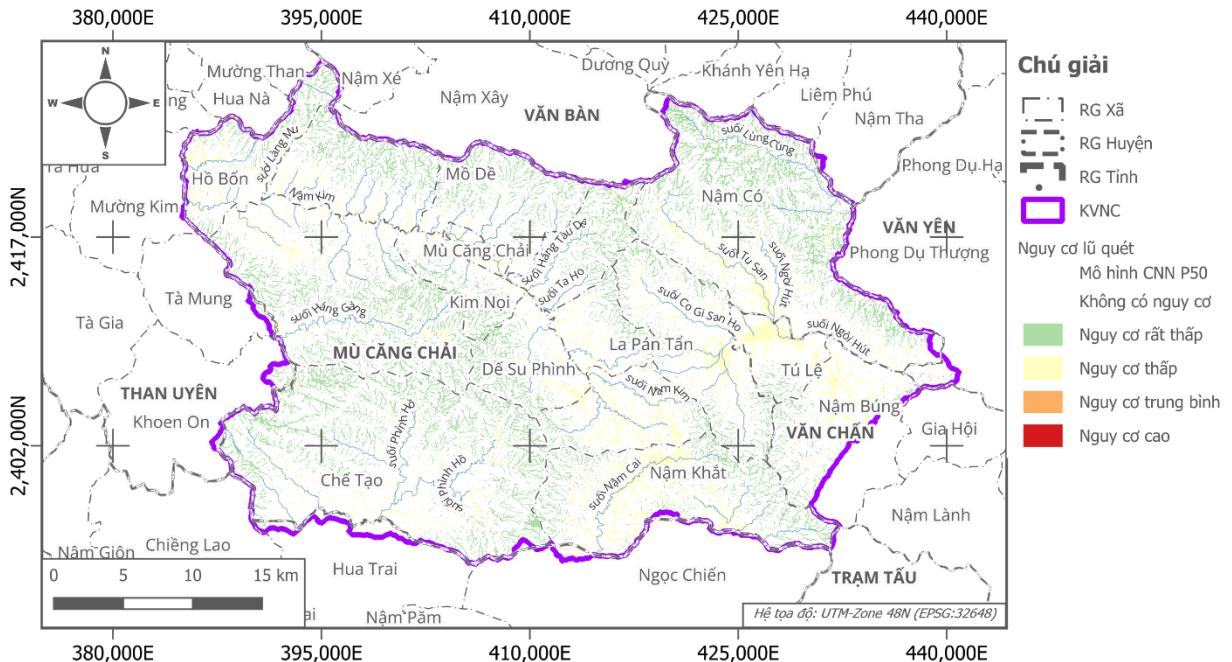
Khu vực nghiên cứu nằm trong vùng PK04 và PK03. Do đó, nghiên cứu này sẽ lấy bình quân phân vùng mưa rào bất lợi nhất của cả hai vùng làm cơ sở để xây dựng biểu đồ phân bố mưa cho các kịch bản thời tiết hiện như sau:



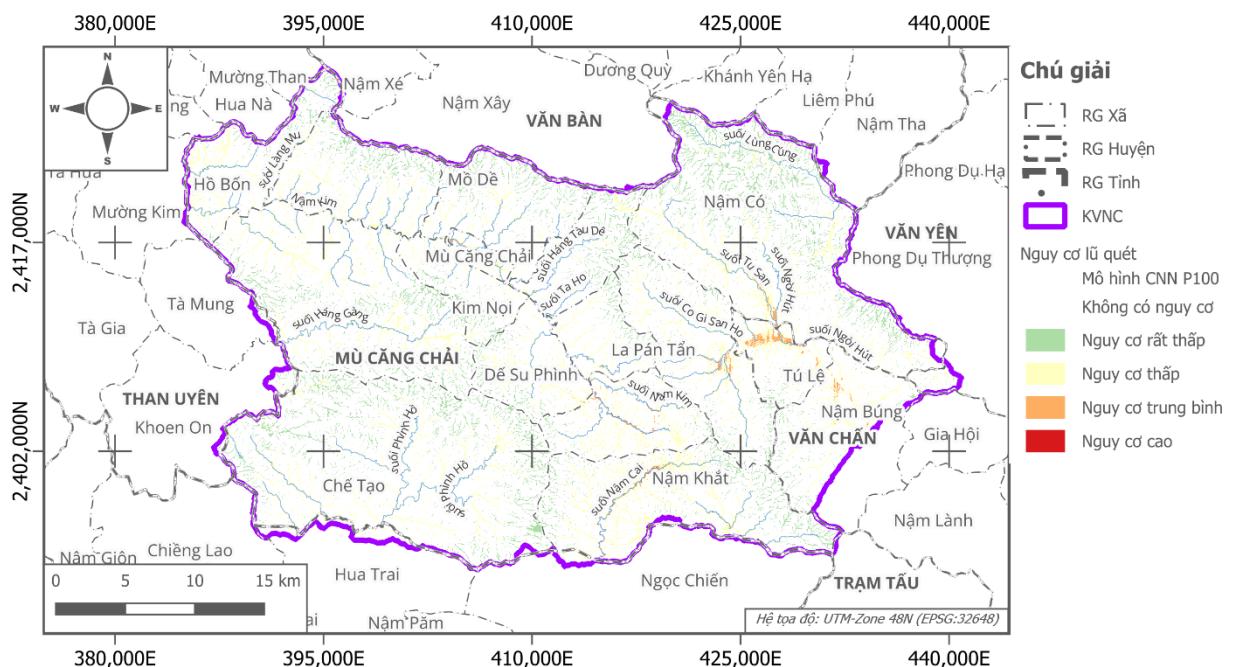
Hình 32. Phân bố mưa tương ứng với mô hình mưa bất lợi nhất cho KVNC

Tham số	Kịch bản mưa trong 6 giờ (mm)		
	50 mm	100 mm	150 mm
1 giờ max	18.52	37.03	55.55
3 giờ max	41.6	83.2	124.81
6 giờ max	50	100	150
24 giờ max	102.39	204.79	307.18

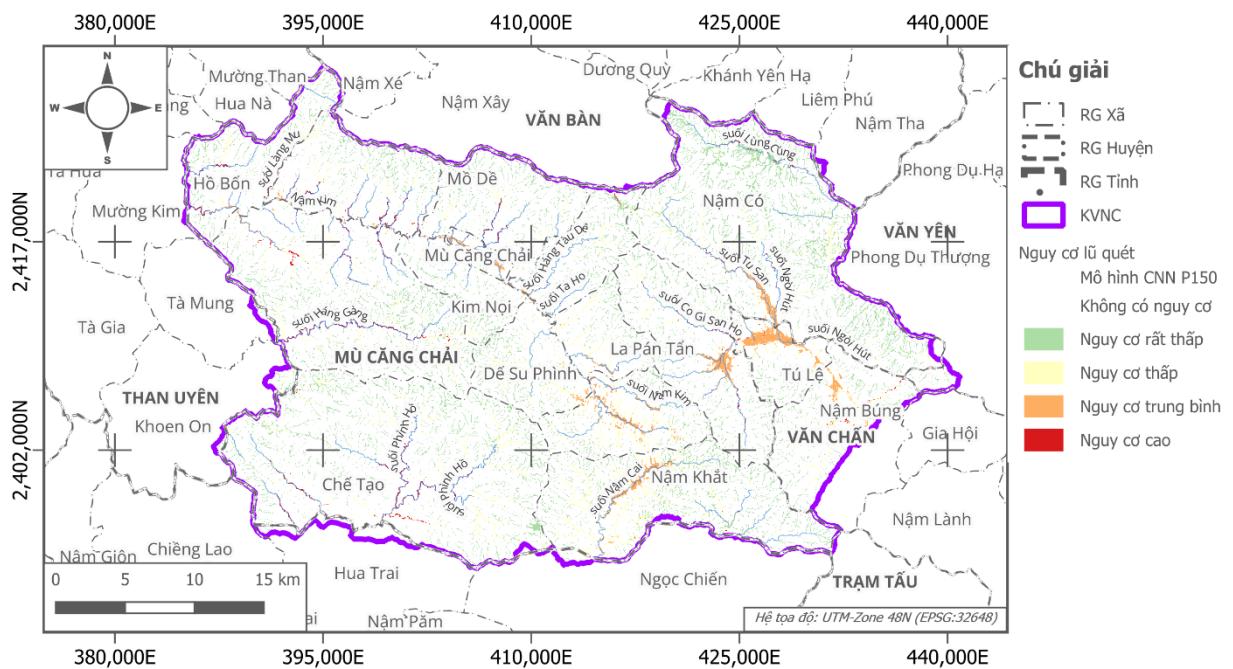
3.3.2 Xây dựng bản đồ phân vùng nguy cơ bằng mô hình CNN



Hình 33. Kết quả xây dựng bản đồ phân vùng nguy cơ kịch bản mưa 50mm/6 giờ



Hình 34. Kết quả xây dựng bản đồ phân vùng nguy cơ kịch bản mưa 100mm/6 giờ



Hình 35. Kết quả xây dựng bản đồ phân vùng nguy cơ kịch bản mưa 150mm/6 giờ

KẾT LUẬN

Kết quả của nghiên cứu đã đạt được những thành tựu đột phá trong việc ứng dụng trí tuệ nhân tạo cho lĩnh vực phòng chống thiên tai, thể hiện qua việc xây dựng thành công hệ thống phân vùng lũ quét với độ chính xác vượt trội. Mô hình LGBM (Light Gradient Boosting Machine) đã khẳng định vị thế dẫn đầu với độ chính xác lên tới 95.24%, kèm theo các chỉ số Precision, Recall và F1-Score đều duy trì ở mức cao trên 95%. Điều đáng chú ý là thời gian huấn luyện mô hình chỉ trong vòng 10 phút, thể hiện tính hiệu quả vượt trội trong việc xử lý dữ liệu địa hình và khí tượng phức tạp, mở ra khả năng triển khai ứng dụng thực tế với chi phí tính toán hợp lý.

Việc tích hợp thành công 20 tham số đầu vào đa dạng, từ các đặc trưng địa hình như độ cao so với sông suối (eleStream), độ dốc lưu vực (wSlope), chỉ số độ ẩm địa hình (TWI) đến các yếu tố khí tượng như lượng mưa tối đa trong các khoảng thời gian khác nhau, đã tạo nên một hệ thống đánh giá toàn diện và khoa học. Đặc biệt, việc sử dụng kết hợp cả các tham số điểm và tham số trung bình lưu vực thể hiện sự hiểu biết sâu sắc về tính chất đa tỷ lệ không gian của hiện tượng lũ quét dưới vai trò thủy văn học, từ đó nâng cao độ chính xác của mô hình dự báo.

Thành tựu quan trọng khác là việc chứng minh thành công khả năng áp dụng đa dạng các thuật toán AI, từ các phương pháp truyền thống như Logistic Regression, SVM đến các mô hình tiên tiến như ensemble learning và deep learning. Mô hình Random Forest với độ chính xác 93.95% đã chứng minh khả năng xử lý tốt các biến đầu vào có tính phi tuyến cao, trong khi các mô hình ensemble đạt 93.26% cho thấy tiềm năng kết hợp sức mạnh của nhiều thuật toán. Điều này không chỉ tạo ra sự linh hoạt trong lựa chọn mô hình phù hợp với từng điều kiện cụ thể mà còn mở ra hướng nghiên cứu tối ưu hóa kết hợp các phương pháp để đạt hiệu quả cao nhất.

TÀI LIỆU THAM KHẢO

- [1] Christopher M. Bishop, Neural Networks for Pattern Recognition, Birmingham: Oxford, 1995.
- [2] Kuz'min, K. K., "The catastrophic flash flood of 1973 and the medeo dam," *Hydrotechnical Construction*, vol. 8, no. 3, p. 203–206, 1974. DOI:10.1007/bf02403378.
- [3] *Quyết định số 18/2021/QĐ-TTg ngày 22 tháng 4 năm 2021 quy định về dự báo, cảnh báo, truyền tin thiên tai và cấp độ rủi ro thiên tai*, 2021. Địa chỉ: vanban.chinhphu.vn/default.aspx?pageid=27160&docid=203152.
- [4] Romdani, R. P, Tamamdin, M, Susandi, A, Pratama, A and Wijaya, A. R, "Development of Flash Flood Hazard Map in Bima City (NTB) using Analytical Hierarchy Process," *IOP Conference Series: Earth and Environmental Science*, vol. 166, p. 012035, 2018. DOI:10.1088/1755-1315/166/1/012035.
- [5] S. Talha, M. Maanan, H. Atika and H. Rhinane, "Prediction of flash flood susceptibility using Fuzzy Analytical Hierarchy Process (FAHP) algorithms and gis: A study case of Guelmim region in southwestern of Morocco," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vols. XLII-4/W19, pp. 407-414, 2019. DOI:10.5194/isprs-archives-xlii-4-w19-407-2019.
- [6] Chen Cao, Peihua Xu, Yihong Wang and et al., "Flash Flood Hazard Susceptibility Mapping Using Frequency Ratio and Statistical Index Methods in Coalmine Subsidence Areas," *Sustainability*, vol. 8, no. 9, p. 948, 2016. DOI:10.3390/su8090948.
- [7] A Saleh, N Sabtu and M R Bunmi, "Flash Flood Susceptibility Mapping in Sungai Pinang catchment using Weight of Evidence," *IOP Conference Series: Earth and Environmental Science*, vol. 1091, no. 1, p. 012017, 2022. DOI:10.26226/m.61a82ed54a84e7b4701d8bdb.
- [8] Costache, Romulus, Pham, Quoc Bao, Sharifi, Ehsan and et al., "Flash-Flood Susceptibility Assessment Using Multi-Criteria Decision Making and Machine Learning Supported by Remote Sensing and GIS Techniques," *Remote Sensing*, vol. 12, p. 106, 2019. DOI:10.3390/rs12010106.
- [9] Zhao, Gang, Liu, Ronghua, Yang, Mingxiang, Tu, Tongbi, Ma, Meihong, Hong, Yang and Wang, Xiekang, "Large-scale flash flood warning in China using deep learning," *Journal of Hydrology*, vol. 604, p. 127222, 2022. DOI:10.1016/j.jhydrol.2021.127222.

- [10] Carpenter, T.M., Sperfslage, J.A., Georgakakos, K.P., Sweeney, T. and Fread, D.L., “National threshold runoff estimation utilizing GIS in support of operational flash flood warning systems,” *Journal of Hydrology*, vol. 224, no. 1-2, p. 21–44, 1999. DOI:10.1016/S0022-1694(99)00115-8.
- [11] Forest Service, “CAUSES OF FLASHY FLOODS AND MUD FLOWS IN UTAH,” *Monthly Weather Review*, vol. 59, p. 122–122, 1931. DOI:10.1175/1520-0493(1931)59<122a:coffam>2.0.co;2.
- [12] Hales, John E., “The Kansas City Flash Flood of 12 September 1977,” *Bulletin of the American Meteorological Society*, vol. 59, no. 6, p. 706–710, 1978. DOI:10.1175/1520-0477-59.6.706.
- [13] NOAA, National Weather Service forecasting handbook (No. 1 - 1979), University of Michigan Library (January 1, 1979), 1979. Địa chỉ: books.google.com.vn/books?id=8jRRAAAAMAAJ&printsec=frontcover&hl=v i#v=onepage&q&f=false.
- [14] Georgakakos, Konstantine P., “A generalized stochastic hydrometeorological model for flood and flash-flood forecasting: 1. Formulation,” *Water Resources Research*, vol. 22, no. 13, p. 2083–2095, 1986. DOI:10.1029/WR022i013p02096.
- [15] Sweeney, Timothy L., “Modernized Areal Flash Flood Guidance,” NOAA Technical Memorandum NWS HYDRO 44, 1992. Địa chỉ: blob:<https://repository.library.noaa.gov/17df97fc-73c5-44db-8f38-5064333c9f3d>.
- [16] William W. Emmett, “The Channels and Waters of the Upper Salmon River Area, Idaho,” United States Government Printing Office, Washington, 1975. DOI:10.3133/pp870a.
- [17] Chang, Tiao J. and Sun, Hong Y., “Study of Potential Flash Floods by Kriging Method,” *Journal of Hydrologic Engineering*, vol. 2, no. 3, p. 104–108, 1997. DOI:10.1061/(asce)1084-0699(1997)2:3(104).
- [18] Sharada, D., Devi, D. Kaveri, Prasad, S. and Kumar, Seelan Santhosh, “Modelling flash flood hazard to a railway line: A GIS approach,” *Geocarto International*, vol. 12, no. 3, p. 77–82, 1997. DOI:10.1080/10106049709354600.
- [19] Marco Borga, Markus Stoffel, Lorenzo Marchi, Francesco Marra and Matthias Jakob, “Hydrogeomorphic response to extreme rainfall in headwater systems: Flash floods and debris flows,” *Journal of Hydrology*, vol. 518, pp. 194-205, 2014. DOI:10.1016/j.jhydrol.2014.05.022.

- [20] Shahin Khosh Bin Ghomash, Daniel Bachmann, Daniel Caviedes-Voulli`eme and Christoph Hinz, “Impact of Rainfall Movement on Flash Flood Response: A Synthetic Study of a Semi-Arid Mountainous Catchment,” *Water*, vol. 14, no. 12, p. 1844, 2022. DOI:10.3390/w14121844.
- [21] Lorenzo Alfieri, Marc Berenguer, Valentin Knechtl and et al., “Handbook of Hydrometeorological Ensemble Forecasting,” in *Flash Flood Forecasting Based on Rainfall Thresholds*, Berlin, Springer Berlin Heidelberg, 2015, pp. 1-38.DOI:10.1007/978-3-642-40457-3_49-1.
- [22] L. Alfieri, D. Velasco and J. Thielen, “Flash flood detection through a multi-stage probabilistic warning system for heavy precipitation events,” *Advances in Geosciences*, vol. 29, pp. 69-75, 2011. DOI:10.5194/adgeo-29-69-2011.
- [23] Nel Caine, “The Rainfall Intensity: Duration Control of Shallow Landslides and Debris Flows,” *Geografiska Annaler: Series A, Physical Geography*, vol. 62, no. 1/2, p. 23, 1980. DOI:10.2307/520449.
- [24] Fausto Guzzetti, Silvia Peruccacci, Mauro Rossi and Colin P. Stark, “The rainfall intensity-duration control of shallow landslides and debris flows: an update,” *Landslides*, vol. 5, no. 1, pp. 3-17, 2007. DOI:10.1007/s10346-007-0112-1.
- [25] Nejc Bezak, Mojca vSraj and Matjaz Mikovs, “Copula-based IDF curves and empirical rainfall thresholds for flash floods and rainfall-induced landslides,” *Journal of Hydrology*, vol. 541, pp. 272--284, 2016. DOI:10.1016/j.jhydrol.2016.02.058.
- [26] T. Turkington, J. Ettema, C. J. van Westen and K. Breinl, “Empirical atmospheric thresholds for debris flows and flash floods in the southern French Alps,” *Natural Hazards and Earth System Sciences*, vol. 14, no. 6, pp. 1517-1530, 2014. DOI:10.5194/nhess-14-1517-2014.
- [27] Geraldo Moura Ramos Filho, Victor Hugo Rabelo Coelho, Emerson da Silva Freitas and et al., “An improved rainfall-threshold approach for robust prediction and warning of flood and flash flood hazards,” *Natural Hazards*, vol. 105, no. 3, pp. 2409-2429, 2020. DOI:10.1007/s11069-020-04405-x.
- [28] Xiaoyan Zhai, Liang Guo, Ronghua Liu and Yongyong Zhang, “Rainfall threshold determination for flash flood warning in mountainous catchments with consideration of antecedent soil moisture and rainfall pattern,” *Natural Hazards*, vol. 94, no. 2, pp. 605-625, 2018. DOI:10.1007/s11069-018-3404-y.

- [29] Wenlin Yuan, Xinyu Tu, Chengguo Su and et al., “Research on the Critical Rainfall of Flash Floods in Small Watersheds Based on the Design of Characteristic Rainfall Patterns,” *Water Resources Management*, vol. 35, no. 10, pp. 3297-3319, 2021. DOI:10.1007/s11269-021-02893-5.
- [30] Mohamed Saber and Koray Yilmaz, “Evaluation and Bias Correction of Satellite-Based Rainfall Estimates for Modelling Flash Floods over the Mediterranean region: Application to Karpuz River Basin, Turkey,” *Water*, vol. 10, no. 5, p. 657, 2018. DOI:10.3390/w10050657.
- [31] I. K. Westerberg, J.-L. Guerrero, P. M. Younger and et al., “Calibration of hydrological models using flow-duration curves,” *Hydrology and Earth System Sciences*, vol. 15, no. 7, pp. 2205-2227, 2011. DOI:10.5194/hess-15-2205-2011.
- [32] F. Silvestro, N. Rebora, G. Cummings and L. Ferraris, “Experiences of dealing with flash floods using an ensemble hydrological nowcasting chain: implications of communication, accessibility and distribution of the results,” *Journal of Flood Risk Management*, vol. 10, no. 4, pp. 446-462, 2015. DOI:10.1111/jfr3.12161.
- [33] Seann Reed, John Schaake and Ziya Zhang, “A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations,” *Journal of Hydrology*, vol. 337, no. 3-4, pp. 402-420, 2007. DOI:10.1016/j.jhydrol.2007.02.015.
- [34] Jonathan D Phillips, “Geomorphic impacts of flash flooding in a forested headwater basin,” *Journal of Hydrology*, vol. 269, no. 3-4, pp. 236-250, 2002. DOI:10.1016/s0022-1694(02)00280-9.
- [35] Thomas L Saaty, “A scaling method for priorities in hierarchical structures,” *Journal of Mathematical Psychology*, vol. 15, no. 3, pp. 234-281, 1977. DOI:10.1016/0022-2496(77)90033-5.
- [36] Aqil Tariq, Jianguo Yan, Bushra Ghaffar and et al., “Flash Flood Susceptibility Assessment and Zonation by Integrating Analytic Hierarchy Process and Frequency Ratio Model with Diverse Spatial Data,” *Water*, vol. 14, no. 19, p. 3069, 2022. DOI:10.3390/w14193069.
- [37] Kairong Lin, Haiyan Chen, Chong-Yu Xu and et al., “Assessment of flash flood risk based on improved analytic hierarchy process method and integrated maximum likelihood clustering algorithm,” *Journal of Hydrology*, vol. 584, p. 124696, 2020. DOI:10.1016/j.jhydrol.2020.124696.
- [38] Mohammed Sadek and Xuxiang Li, “Low-Cost Solution for Assessment of Urban Flash Flood Impacts Using Sentinel-2 Satellite Images and Fuzzy Analytic

Hierarchy Process: A Case Study of Ras Ghareb City, Egypt," *Advances in Civil Engineering*, vol. 2019, pp. 1-15, 2019. DOI:10.1155/2019/2561215.

- [39] Romulus Costache and Liliana Zaharia, "Flash-flood potential assessment and mapping by integrating the weights-of-evidence and frequency ratio statistical methods in GIS environment - case study: Basca Chiojdului River catchment (Romania)," *Journal of Earth System Science*, vol. 126, no. 4, pp. 1-19, 2017. DOI:10.1007/s12040-017-0828-9.
- [40] Amoroch, J. and Brandstetter, A., "Determination of Nonlinear Functional Response Functions in Rainfall-Runoff Processes," *Water Resources Research*, vol. 7, no. 5, pp. 1087-1101, 1971. DOI:10.1029/wr007i005p01087.
- [41] Hsu, Kuo-lin, Gupta, Hoshin Vijai and Sorooshian, Soroosh, "Artificial Neural Network Modeling of the Rainfall-Runoff Process," *Water Resources Research*, vol. 31, no. 10, pp. 2517-2530, 1995. DOI:10.1029/95wr01955.
- [42] Sahoo, G.B., Ray, C. and De Carlo, E.H., "Use of neural network to predict flash flood and attendant water qualities of a mountainous stream on Oahu, Hawaii," *Journal of Hydrology*, vol. 327, no. 3-4, pp. 525-538, 2006. DOI:10.1016/j.jhydrol.2005.11.059.
- [43] Janál, Petr and Starý, Miloš, "Fuzzy model use for prediction of the state of emergency of river basin in the case of flash flood," *Journal of Hydrology and Hydromechanics*, vol. 57, no. 3, 2009. DOI:10.2478/v10098-009-0013-1.
- [44] Toukourou, Mohamed Samir, Johannet, Anne and Dreyfus, Gérard, "Flash Flood Forecasting by Statistical Learning in the Absence of Rainfall Forecast: A Case Study," in *Communications in Computer and Information Science*, Springer Berlin Heidelberg, 2009, pp. 98-107. DOI:10.1007/978-3-642-03969-0_10.
- [45] Lamovec, Peter, Veljanovski, Tatjana, Mikoš, Matjaž and Oštir, Krištof, "Detecting flooded areas with machine learning techniques: case study of the Selška Sora river flash flood in September 2007," *Journal of Applied Remote Sensing*, vol. 7, no. 1, p. 073564, 2013. DOI:10.1117/1.jrs.7.073564.
- [46] Piotrowski, A., Napiórkowski, J. J. and Rowiński, P.M., "Flash-flood forecasting by means of neural networks and nearest neighbour approach – a comparative study," *Nonlinear Processes in Geophysics*, vol. 13, no. 4, pp. 443-448, 2006. DOI:10.5194/npg-13-443-2006.
- [47] Kong A Siou, L., Johannet, A., Pistre, S. and Borrell, V., "Flash Floods Forecasting in a Karstic Basin Using Neural Networks: the Case of the Lez Basin

(South of France)," in *Environmental Earth Sciences*, Springer Berlin Heidelberg, 2010, pp. 215-221. DOI:10.1007/978-3-642-12486-0_33.

- [48] Artigue, G., Johannet, A., Borrell, V. and Pistre, S., "Flash floods forecasting without rainfalls forecasts by recurrent neural networks. Case study on the Mialet basin (Southern France)," in *2011 Third World Congress on Nature and Biologically Inspired Computing*, 2011. DOI:10.1109/nabic.2011.6089612.
- [49] Izyan 'Izzati Abdul Rahman and Nik Mohd Asrol Alias, "Rainfall forecasting using an artificial neural network model to prevent flash floods," in *8th International Conference on High-capacity Optical Networks and Emerging Technologies*, 2011. DOI:10.1109/honet.2011.6149841.
- [50] Boukharouba, Khaled, Roussel, Pierre, Dreyfus, Gerard and Johannet, Anne, "Flash flood forecasting using Support Vector Regression: An event clustering based approach," in *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2013. DOI:10.1109/mlsp.2013.6661958.
- [51] Cao Đăng Dư and Lương Tuấn Anh, "Phân vùng khả năng xuất hiện lũ quét," *Tạp chí Khí tượng thủy văn*, p. 1÷8, 1995. Địa chỉ: tapchikttv.vn/data/article/2126/2.pdf.
- [52] Phạm Thị Hương Lan and Vũ Minh Cát, "Một số kết quả nghiên cứu xây dựng bản đồ tiềm năng lũ quét phục vụ công tác cảnh báo lũ quét vùng núi đồng bắc Việt Nam," *Tạp chí Khí tượng Thủy văn*, vol. 556, no. 2, pp. 11-16, 2008. Địa chỉ: http://tapchikttv.vn/data/article/1401/thang%202%202008_02.pdf.
- [53] Lã Thanh Hà, "Nghiên cứu xây dựng bản đồ phân vùng nguy cơ lũ quét phục vụ công tác phòng tránh lũ quét cho tỉnh Yên Bái," *Tạp chí Khí tượng Thủy văn*, vol. 578, no. 2, pp. 11-15, 2009. Địa chỉ: <http://tapchikttv.vn/data/article/362/B%C3%A0i%203.pdf>.
- [54] Dương Thị Lợi and Đặng Phuong Lan, "Ứng dụng mô hình đa chi tiêu nhằm đánh giá nguy cơ lũ quét trong bối cảnh biến đổi khí hậu toàn cầu. Trường hợp nghiên cứu cụ thể: miền núi Tây Bắc - Việt Nam," *Tạp chí Khí tượng thủy văn*, vol. 721, p. 31÷45, 2021. Địa chỉ: <http://tapchikttv.vn/data/article/1388/4.pdf>.
- [55] Minh Đức, Đào, Cao Minh, Vũ, Hải Yến, Hoàng, Quang Anh, Phạm and Kinh Bắc, Đặng, "Đánh giá nguy cơ hình thành lũ quét trên suối Nghĩa Đô, huyện Bảo Yên, tỉnh Lào Cai bằng phương pháp phân tích thống kê," *Vietnam Journal of Hydrometeorology*, vol. EME4, no. 1, pp. 341-354, 2022. DOI:10.36335/vnjhm.2022(eme4).341-354.

- [56] Thị Huyền, Nguyễn, Quốc Khánh, Nguyễn, huy Dương, Nguyễn, Hoàng Ninh, Nguyễn and Đức Hà, Nguyễn, “Kết quả khoanh định các khu vực nhạy cảm về trượt lở, lũ quét khu vực Thành phố Đà Nẵng,” *Vietnam Journal of Hydrometeorology*, vol. 1, no. 745, pp. 21-33, 2023. DOI:10.36335/vnjhm.2023(745).21-33.
- [57] Thị Phương Thảo, Ngô, Hùng Long, Ngô, Anh Tuấn, Trần and Minh Hằng, Lê, “Sử dụng ảnh Sentinel-1A đa thời gian để phát hiện lũ quét, thử nghiệm tại tỉnh Lào Cai,” *Journal of Hydro-meteorology*, vol. 8, no. 764, pp. 29-37, 2024. DOI:10.36335/vnjhm.2024(764).29-37.
- [58] Ngô Thị Phương Thảo, “Luận án: Nghiên cứu phát triển mô hình trí tuệ nhân tạo trong phân vùng nguy cơ lũ quét ở Việt Nam,” Trường đại học Mỏ - Địa chất, 2024. Địa chỉ: <https://humg.edu.vn/content/tintuc/Lists/News/Attachments/9000/Luan%20an%20tien%20si-Ngo%20Thi%20Phuong%20Thao.pdf>.
- [59] Ha, Hang, Bui, Quynh Duy, Khuc, Thanh Dong, Tran, Dinh Trong, Pham, Binh Thai, Mai, Sy Hung, Nguyen, Lam Phuong and Luu, Chinh, “A machine learning approach in spatial predicting of landslides and flash flood susceptible zones for a road network,” *Modeling Earth Systems and Environment*, vol. 8, no. 4, p. 4341÷4357, 2022. DOI:10.1007/s40808-022-01384-9.
- [60] Nguyễn Viết Nghĩa and Nguyễn Cao Cường, “Ứng dụng mạng nơ-ron nhân tạo đa lớp trong thành lập mô hình phân vùng lũ quét khu vực miền núi Tây Bắc, thực nghiệm tại tỉnh Yên Bái,” *Tạp chí khoa học Đo đạc và Bản đồ*, vol. 44, no. 6, p. 56÷64, 2020. Địa chỉ: <https://jgac.vn/index.php/journal/article/view/304/291>.
- [61] Hoang, Duc-Vinh and Liou, Yuei-An, “Assessing the influence of human activities on flash flood susceptibility in mountainous regions of Vietnam,” *Ecological Indicators*, vol. 158, no. , p. 111417, 2024. DOI:10.1016/j.ecolind.2023.111417.
- [62] He, Fei, Liu, Suxia, Mo, Xingguo and Wang, Zhonggen, “Interpretable flash flood susceptibility mapping in Yarlung Tsangpo River Basin using H2O Auto-ML,” *Scientific Reports*, vol. 15, no. 1, p. , 2025. DOI:10.1038/s41598-024-84655-y.
- [63] Alarifi, Saad S., Abdelkareem, Mohamed, Abdalla, Fathy and Alotaibi, Mislat, “Flash Flood Hazard Mapping Using Remote Sensing and GIS Techniques in Southwestern Saudi Arabia,” *Sustainability*, vol. 14, no. 21, p. 14145, 2022. DOI:10.3390/su142114145.

- [64] MDE, “Method for Designing Infiltration Structures,” [Trực tuyến]. Địa chỉ: https://mde.maryland.gov/programs/water/StormwaterManagementProgram/Documents/www.mde.state.md.us/assets/document/sedimentstormwater/Appnd_D13.pdf. [Truy cập 15/6/2023].
- [65] NRCS, “Soil Infiltration,” [Trực tuyến]. Địa chỉ: https://web.archive.org/web/20240301064123/https://cropwatch.unl.edu/documents/USDA_NRCS_infiltration_guide6-4-14.pdf. [Truy cập 11/6/2023].
- [66] Rahmati, Mehdi, Weihermüller, Lutz, Vanderborgh, Jan, Pachepsky, Yakov A., Mao, Lili, Sadeghi, Seyed Hamidreza, Moosavi, Niloofar, Kheirfam, Hossein, Montzka, Carsten, Van Looy, Kris, Toth, Brigitta, Hazbavi, Zeinab, Al Yamani, Wafa, Albalasmeh, Ammar A., Alghzawi, Ma'in Z., Angulo-Jaramillo, Rafael, Antonino, Antônio Celso Dantas, Arampatzis, George, Armindo, Robson André, Asadi, Hossein, Bamutaze, Yazidhi, Batlle-Aguilar, Jordi, Béchet, Béatrice, Becker, Fabian, Blöschl, Günter, Bohne, Klaus, Braud, Isabelle, Castellano, Clara, Cerdà, Artemi, Chalhoub, Maha, Cichota, Rogerio, Císlarová, Milena, Clothier, Brent, Coquet, Yves, Cornelis, Wim, Corradini, Corrado, Coutinho, Artur Paiva, de Oliveira, Muriel Bastista, de Macedo, José Ronaldo, Durães, Matheus Fonseca, Emami, Hojat, Eskandari, Iraj, Farajnia, Asghar, Flammini, Alessia, Fodor, Nándor, Gharaibeh, Mamoun, Ghavimipanah, Mohamad Hossein, Ghezzehei, Teamrat A., Giertz, Simone, Hatzigiannakis, Evangelos G., Horn, Rainer, Jiménez, Juan José, Jacques, Diederik, Keesstra, Saskia Deborah, Kelishadi, Hamid, Kiani-Harchegani, Mahboobeh, Kouselou, Mehdi, Kumar Jha, Madan, Lassabatere, Laurent, Li, Xiaoyan, Liebig, Mark A., Lichner, Lubomír, López, María Victoria, Machiwal, Deepesh, Mallants, Dirk, Mallmann, Micael Stolben, de Oliveira Marques, Jean Dalmo, Marshall, Miles R., Mertens, Jan, Meunier, Félicien, Mohammadi, Mohammad Hossein, Mohanty, Binayak P., Pulido-Moncada, Mansonia, Montenegro, Suzana, Morbidelli, Renato, Moret-Fernández, David, Moosavi, Ali Akbar, Mosaddeghi, Mohammad Reza, Mousavi, Seyed Bahman, Mozaffari, Hasan, Nabiollahi, Kamal, Neyshabouri, Mohammad Reza, Ottoni, Marta Vasconcelos, Ottoni Filho, Theophilo Benedicto, Pahlavan-Rad, Mohammad Reza, Panagopoulos, Andreas, Peth, Stephan, Peyneau, Pierre-Emmanuel, Picciafuoco, Tommaso, Poesen, Jean, Pulido, Manuel, Reinert, Dalvan José, Reinsch, Sabine, Rezaei, Meisam, Roberts, Francis Parry, Robinson, David, Rodrigo-Comino, Jesús, Rotunno Filho, Otto Corrêa, Saito, Tadaomi, Suganuma, Hideki, Saltalippi, Carla, Sándor, Renáta, Schütt, Brigitta, Seeger, Manuel, Sepehrnia, Nasrollah, Sharifi Moghaddam, Ehsan, Shukla, Manoj, Shutaro, Shiraki, Sorando, Ricardo, Stanley, Ajayi Asishana, Strauss, Peter, Su, Zhongbo, Taghizadeh-Mehrjardi, Ruhollah, Taguas,

Encarnación, Teixeira, Wenceslau Geraldes, Vaezi, Ali Reza, Vafakhah, Mehdi, Vogel, Tomas, Vogeler, Iris, Votrubova, Jana, Werner, Steffen, Winarski, Thierry, Yilmaz, Deniz, Young, Michael H., Zacharias, Steffen, Zeng, Yijian, Zhao, Ying, Zhao, Hong and Vereecken, Harry, “Development and analysis of the Soil Water Infiltration Global database,” *Earth System Science Data*, vol. 10, no. 3, pp. 1237-1263, 2018. DOI:10.5194/essd-10-1237-2018.

- [67] Thùy Thanh, “Yên Bai: Chủ động ứng phó với mưa lớn diện rộng từ chiều tối ngày 14 đến 16/10,” Báo Yên Bai, 13/10/2020. [Trực tuyến]. Địa chỉ: <https://baoyenbai.com.vn/PrintPreview/198871/>. [Truy cập 15/6/2023].
- [68] Lal, Aleena B, M, Megha, P, Muhammed Sufiyan M, Thekkal, David Joseph, V, Aswathy M and Meckamalil, Rotney Roy, “Flash Flood Detection and Alert System Using Machine Learning,” *International Journal for Research in Applied Science and Engineering Technology*, vol. 12, no. 6, p. 45÷51, 2024. DOI:10.22214/ijraset.2024.62971.
- [69] Sellami, El Mehdi and Rhinane, Hassan, “A modern method for building damage evaluation using deep learning approach - Case study: Flash flooding in Derna, Libya,” *E3S Web of Conferences*, vol. 502, no. , p. 03010, 2024. DOI:10.1051/e3sconf/202450203010.
- [70] Lugt, Dorien, van Hoek, Mattijn, Meirink, Jan Fokke and van der Kooij, Eva, “Nowcasting for urban flash floods in Africa: a machine-learning and satellite-observation based model,” , vol. , no. , p. , 2021. DOI:10.5194/egusphere-egu21-16002.
- [71] Band, Shahab S., Janizadeh, Saeid, Chandra Pal, Subodh, Saha, Asish, Chakrabortty, Rabin, Melesse, Assefa M. and Mosavi, Amirhosein, “Flash Flood Susceptibility Modeling Using New Approaches of Hybrid and Ensemble Tree-Based Machine Learning Algorithms,” *Remote Sensing*, vol. 12, no. 21, p. 3568, 2020. DOI:10.3390/rs12213568.
- [72] Costache, Romulus, Arabameri, Alireza, Blaschke, Thomas, Pham, Quoc Bao, Pham, Binh Thai, Pandey, Manish, Arora, Aman, Linh, Nguyen Thi Thuy and Costache, Iulia, “Flash-Flood Potential Mapping Using Deep Learning, Alternating Decision Trees and Data Provided by Remote Sensing Sensors,” *Sensors*, vol. 21, no. 1, p. 280, 2021. DOI:10.3390/s21010280.
- [73] Ilia, Ioanna, Tsangaratos, Paraskevas, Tzampoglou, Ploutarchos, Chen, Wei and Hong, Haoyuan, “Flash flood susceptibility mapping using stacking ensemble machine learning models,” *Geocarto International*, vol. 37, no. 27, p. 15010÷15036, 2022. DOI:10.1080/10106049.2022.2093990.

- [74] SELLAMI, EL Mehdi and Rhinane, Hassan, “Google Earth Engine and Machine Learning for Flash Flood Exposure Mapping”Case Study: Tetouan, Morocco,” *Geosciences*, vol. 14, no. 6, p. 152, 2024. DOI:10.3390/geosciences14060152.
- [75] Razavi-Termeh, Seyed Vahid, Seo, MyoungBae, Sadeghi-Niaraki, Abolghasem and Choi, Soo-Mi, “Flash flood detection and susceptibility mapping in the Monsoon period by integration of optical and radar satellite imagery using an improvement of a sequential ensemble algorithm,” *Weather and Climate Extremes*, vol. 41, no. , p. 100595, 2023. DOI:10.1016/j.wace.2023.100595.
- [76] C. ROSS, L. PRIHODKO, J. ANCHANG, S. KUMAR, W. JI and N. HANAN, *Global Hydrologic Soil Groups (HYSOGs250m) for Curve Number-Based Runoff Modeling*, ORNL Distributed Active Archive Center, 2018. DOI:10.3334/ORNLDAAAC/1566. Địa chỉ: https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1566.
- [77] Theo Vietnamplus, “Yên Bai: Bảy người mất tích do lũ ống, lũ quét ở Mù Cang Chải,” [Trực tuyến]. Địa chỉ: <https://daidoanket.vn/yen-bai-bay-nguo-mat-tich-do-lu-ong-lu-quet-o-mu-cang-chai-10078324.html>. [Truy cập 17/4/2025].
- [78] Thanh Thủy, “Lũ quét ở Mù Cang Chải: Uớc thiệt hại khoảng 150 tỷ đồng,” [Trực tuyến]. Địa chỉ: https://baoyenbai.com.vn/12/151798/Luquet_o_Mu_Cang_Chai_Uoc_thiet_hai_khoang_150_ty_dong.htm. [Truy cập 17/4/2025].
- [79] Báo Nông Nghiệp Và Môi Trường and Thanh Ngà - Trần Nam, “Mưa lũ ở Mù Cang Chải, Yên Bai: Nhiều thôn bản của xã Hồ Bốn vẫn chưa thể tiếp cận,” 2023. [Trực tuyến]. Địa chỉ: <https://nongnghiepmoitruong.vn/mua-lu-o-mu-cang-chai-yen-bai-nhieu-thon-ban-cua-xa-ho-bon-van-chua-the-tiep-can-i717002.html>. [Truy cập 17/4/2025].
- [80] Định Sơn, “Mưa lũ làm 3 người chết và 11 nạn nhân mất tích ở Yên Bai,” 2018. [Trực tuyến]. Địa chỉ: <https://znews.vn/mua-lu-lam-3-nguo-chet-va-11-nan-nhan-mat-tich-o-yen-bai-post861935.html>. [Truy cập 17/4/2025].
- [81] Mạnh Cường, “Thiệt hại do mưa, lũ gây ra trên địa bàn huyện Mù Cang Chải ước khoảng 920 triệu đồng,” Trang Thông tin điện tử Mù Cang Chải, 22/7/2019. [Trực tuyến]. Địa chỉ: <https://mucangchai.yenbai.gov.vn/news/tin-moi/?UserKey=Thiet-hai-do-mua-lu-gay-ra-tren-dia-ban-huyen-Mu-Cang-Chai-uoc-khoang-920-trieu-dong&PageIndex=12>. [Truy cập 17/4/2025].

- [82] Vũ Bá Thao and Bùi Xuân Việt, “Phân tích ngưỡng mưa phát sinh một số trận lũ quét, lũ bùn đá thuộc các tỉnh Lai Châu, Điện Biên, Yên Bai, Sơn La,” *Tạp chí Khí tượng thủy văn*, vol. 749, pp. 96-110, 2023. DOI:10.36335/VNJHM.2023(749).96-110.
- [83] JICA, “Khảo sát thu thập dữ liệu về các giải pháp phòng chống lũ quét và sạt lở đất tại khu vực miền núi phía Bắc của Việt Nam,” Cơ quan hợp tác quốc tế Nhật Bản (JICA), Hà Nội, 2021. Địa chỉ: https://openjicareport.jica.go.jp/pdf/12357885_01.pdf.
- [84] “Yên Bai: Hiệu quả từ mô hình phòng chống, giảm nhẹ thiên tai tại trường THPT Mù Cang Chải,” 2011. [Trực tuyến]. Địa chỉ: <https://baochinhphu.vn/yen-bai-hieuqua-tu-mo-hinh-phong-chong-giam-nhe-thien-tai-tai-truong-thpt-mu-cang-chai-102104536.htm>. [Truy cập 17/4/2025].
- [85] “Ứng dụng khoa học công nghệ trong phòng, chống thiên tai trên địa bàn tỉnh Yên Bai còn nhiều hạn chế,” [Trực tuyến]. Địa chỉ: <https://phongchongthientai.mard.gov.vn/Pages/Ung-dung-khoa-hoc-cong-nghe-trong-phong-chong-thie-5821976939.aspx>. [Truy cập 17/4/2025].
- [86] I. D. Moore, R. B. Grayson and A. R. Ladson, “Digital terrain modelling: A review of hydrological, geomorphological, and biological applications,” *Hydrological Processes*, vol. 5, no. 1, pp. 3-30, jan/1991. DOI:10.1002/hyp.3360050103.
- [87] Han J., Kamber M. and Pei J., Data Mining: Concepts and Techniques, Elsevier, 2012. DOI:10.1016/c2009-0-61819-5.
- [88] T. Hastie, R. Tibshirani and J. Friedman, “Model Assessment and Selection,” in *The Elements of Statistical Learning*, Springer New York, 2008, pp. 219-259. DOI:10.1007/978-0-387-84858-7_7.
- [89] Pedregosa, Fabian, “Scikit-learn: Machine Learning in Python,” 2012.
- [90] Box G. E. P. and Cox D. R., “An Analysis of Transformations,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 26, no. 2, pp. 211-252, 1964. Địa chỉ: <https://www.ime.usp.br/~abe/lista/pdfQWaCMboK68.pdf>.
- [91] In-Kwon Yeo and Richard A. Johnson, “A New Family of Power Transformations to Improve Normality or Symmetry,” *Biometrika*, vol. 87, no. 4, pp. 954-959, 2000. Địa chỉ: <https://www.jstor.org/stable/2673623>.
- [92] Petr Sercl, Martin Pecha, Petr Novak, Hana Kyznarova, Ondrej Ledvinka, Vojtech Svoboda and Jan Danhelka, “Flash Flood Indicator,” Czech

Hydrometeorological Institute, Prague, 2023. Địa chỉ:
<https://www.chmi.cz/files/portal/docs/reditel/SIS/nakladatelstvi/assets/ffi.pdf>.

- [93] Bá Thao, Vũ, “Phương pháp xác định khu vực rủi ro lũ bùn đá dựa vào bản đồ địa hình,” *Vietnam Journal of Hydrometeorology*, vol. 713, no. 5, pp. 37-46, 2020. DOI:10.36335/vnjhm.2020(713).37-46.
- [94] NRCS, “Part 630 - Hydrology,” 3/2020. [Trực tuyến]. Địa chỉ:
<https://directives.nrcc.usda.gov/sites/default/files2/1712930634/Part%20630%20-%20Hydrology.pdf>. [Truy cập 11/12/2023].