

Mẫu B23. BCTK-BNN

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

BÁO CÁO TỔNG KẾT  
ĐỀ TÀI TIỀM NĂNG CẤP BỘ

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

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

**BÁO CÁO TỔNG KẾT  
ĐỀ TÀI TIỀM NĂNG CẤP BỘ**

**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**

## **DANH SÁCH NHỮNG NGƯỜI THAM GIA THỰC HIỆN ĐỀ TÀI**

TT	Họ và tên	Cơ quan/tổ chức
1	ThS. Lê Văn Thìn	Viện Khoa học Thủy lợi Việt Nam
2	TS. Nguyễn Đăng Giáp	Viện Khoa học Thủy lợi Việt Nam
3	ThS. Đào Anh Tuấn	Viện Khoa học Thủy lợi Việt Nam
4	ThS. Lê Thế Cường	Viện Khoa học Thủy lợi Việt Nam
5	ThS. Mai Thị Ngân Anh	Viện Khoa học Thủy lợi Việt Nam
6	ThS. Nguyễn Đức Diện	Viện Khoa học Thủy lợi Việt Nam
7	KS. Nguyễn Hương Trà	Viện Khoa học Thủy lợi Việt Nam
8	KS. Phạm Thị Tuyết	Viện Khoa học Thủy lợi Việt Nam
9	KS. Nguyễn Đức Hoan	Viện Khoa học Thủy lợi Việt Nam
10	ThS. Lê Xuân Cầu	Viện Khoa học Thủy lợi Việt Nam

### **Thông tin đề tài:**

- Tên đề tài: Nghiên cứu ứng dụng trí tuệ nhân tạo và dữ liệu địa không gian để phân vùng lũ quét quy mô cấp huyện
- Mục tiêu: Đánh giá khả năng ứng dụng công nghệ trí tuệ nhân tạo và dữ liệu địa không gian để nâng cao độ tin cậy trong phân vùng lũ quét quy mô cấp huyện
- Thời gian thực hiện: Từ tháng 01/2023 đến tháng 06/2025.
- Kinh phí: 500.000.000 đồng (Năm trăm triệu đồng chẵn)
- Cơ quan chủ trì: Viện Khoa học Thủy lợi Việt Nam
- Chủ nhiệm: Lê Văn Thìn

## MỤC LỤC

CHƯƠNG 1. TỔNG QUAN TÌNH HÌNH NGHIÊN CỨU XÁC ĐỊNH LŨ QUÉT VÀ ỨNG DỤNG TRÍ TUỆ NHÂN TẠO TRONG PHÂN VÙNG LŨ QUÉT .....	3
1.1. Sơ lược về trí tuệ nhân tạo và ứng dụng trí tuệ nhân tạo trong nghiên cứu lũ quét .....	3
1.1.1 Trí tuệ nhân tạo .....	3
1.1.2 Ứng dụng trí tuệ nhân tạo trong nghiên cứu lũ quét .....	4
1.2. Tổng quan các loại dữ liệu địa không gian và ứng dụng trong nghiên cứu lũ quét. ....	5
1.1.1. Giới thiệu về dữ liệu địa không gian .....	5
1.1.2. Quá trình phát triển ảnh viễn thám theo lịch sử.....	7
1.1.3. Các loại ảnh viễn thám và ứng dụng.....	13
1.1.4. Các dữ liệu địa không gian khác .....	16
1.3. Các nghiên cứu về phân vùng lũ quét.....	18
1.3.1 Các nghiên cứu trên thế giới .....	19
1.3.2 Các nghiên cứu ở Việt Nam.....	36
1.4. Các sản phẩm chính từ kết quả phân vùng lũ quét sử dụng trí tuệ nhân tạo và dữ liệu địa không gian. ....	41
1.4.1 Bản đồ phát hiện lũ quét .....	42
1.4.2 Bản đồ nhạy cảm với lũ quét .....	43
1.4.3 Bản đồ nguy cơ lũ quét .....	43
1.5. Đánh giá các phương pháp xác định lũ quét và dữ liệu sử dụng .....	45
1.5.1 Các phương pháp phổ biến xác định lũ quét .....	45
1.5.2 Dữ liệu sử dụng.....	53
1.5.3 Sự khác biệt dữ liệu giữa học máy và học sâu.....	54
1.6. Đánh giá những hạn chế còn tồn tại và định hướng nghiên cứu .....	56
CHƯƠNG 2. PHƯƠNG PHÁP PHÂN VÙNG LŨ QUÉT ỨNG DỤNG TRÍ TUỆ NHÂN TẠO VÀ DỮ LIỆU ĐỊA KHÔNG GIAN .....	58
2.1. Sơ đồ tiếp cận và phương pháp nghiên cứu.....	58
2.1.1 Phương pháp thu thập dữ liệu.....	59
2.1.2 Phương pháp nghiên cứu GIS.....	59
2.1.3 Phương pháp học máy .....	59
2.1.4 Phương pháp học sâu .....	64
2.1.5 Một số kỹ thuật tối ưu hóa mô hình trí tuệ nhân tạo .....	66
2.2. Dữ liệu sử dụng.....	67
2.2.1 Dữ liệu thu thập .....	67

2.2.2 Chuẩn bị dữ liệu.....	75
2.3. Quy trình ứng dụng trí tuệ nhân tạo và dữ liệu địa không gian để phân vùng lũ quét .....	82
2.3.1 Sơ đồ quy trình .....	82
2.3.2 Xác định các bước thực hiện .....	83
2.3.3 Chuẩn bị dữ liệu.....	83
2.3.4 Xây dựng mô hình trí tuệ nhân tạo trong phân vùng lũ quét.....	93
2.3.5 Đánh giá sự phù hợp của mô hình .....	102
<b>CHƯƠNG 3. XÂY DỰNG MÔ HÌNH PHÂN VÙNG LŨ QUÉT CHO KHU VỰC MÙ CANG CHẢI.....</b>	<b>104</b>
3.1. Lựa chọn khu vực nghiên cứu .....	104
3.2. Đánh giá các dữ liệu địa không gian trong nghiên cứu lũ quét .....	105
3.2.1 Nghiên cứu xây dựng dữ liệu thảm phủ từ ảnh viễn thám .....	105
3.2.2 Xây dựng bộ cơ sở dữ liệu không gian khác phục vụ phân vùng lũ quét .....	115
3.3. Thiết lập mô hình phân vùng lũ quét cho khu vực nghiên cứu .....	129
3.3.1 Đầu vào và cấu trúc dữ liệu .....	129
3.3.2 Xây dựng mô hình học máy.....	138
3.3.3 Xây dựng mô hình học sâu .....	155
3.3.4 Phân tích, đánh giá các mô hình trí tuệ nhân tạo trong phân vùng lũ quét .....	164
3.4. Kết quả phân vùng lũ quét cho khu vực nghiên cứu .....	167
3.4.1 Kết quả phân vùng lũ quét .....	167
3.4.2 Đánh giá sự phù hợp của kết quả phân vùng lũ quét.....	170
3.4.3 Đánh giá chung .....	176
3.5. Xây dựng bản đồ phân vùng lũ quét theo kịch bản mưa .....	177
3.5.1 Xây dựng kịch bản mưa.....	177
3.5.2 Xây dựng bản đồ phân vùng nguy cơ bằng mô hình CNN .....	178
<b>KẾT LUẬN, KIẾN NGHỊ .....</b>	<b>180</b>
Kết luận .....	180
Những hạn chế còn tồn tại .....	181
Kiến nghị .....	181

## I. MỞ ĐẦU

Lũ quét thường có những đặc điểm là một dòng chảy mạnh, bất ngờ và có tốc độ cao trên lòng sông hoặc khu vực đất thấp. Lũ quét tự nhiên thường xảy ra sau một cơn mưa lớn hoặc vỡ đập do nghẽn dòng. Đây là một hiện tượng nguy hiểm, gây thiệt hại lớn về người và tài sản nếu không được cảnh báo và quản lý kịp thời. Vì vậy, phân vùng lũ quét (flash-flood zoning) là một nhiệm vụ quan trọng trong công tác quản lý lũ quét, đặc biệt ở các khu vực miền núi ở Việt Nam.

Trong những năm gần đây, các công nghệ trí tuệ nhân tạo (AI) đã bước vào giai đoạn phát triển mạnh mẽ và được ứng dụng trong nhiều lĩnh vực khác nhau, trong đó có lĩnh vực quản lý rủi ro thiên tai. Với khả năng xử lý dữ liệu lớn, học máy và thị giác máy tính, AI đóng vai trò quan trọng trong việc phân tích và dự báo các hiện tượng thời tiết cực đoan, cũng như hỗ trợ công tác phân vùng lũ quét.

Phương pháp truyền thống trong phân vùng lũ quét thường dựa vào các mô hình thủy lực và thủy văn, cùng với các số liệu khảo sát và đo đạc trực tiếp. Tuy nhiên, việc thu thập dữ liệu đầu vào cho các mô hình này thường tốn kém và mất nhiều thời gian. Hơn nữa, các mô hình này cũng có những hạn chế nhất định trong việc dự đoán chính xác diễn biến của lũ quét, đặc biệt là trong các tình huống phức tạp và thay đổi nhanh chóng. Ngoài ra, việc xác định lũ quét dựa vào mô hình thủy lực cũng chưa có sự phù hợp do tính chất dòng chảy xiết không ổn định.

Cùng với sự phát triển về công nghệ, AI có thể đóng vai trò hỗ trợ đáng kể trong việc phân tích và xử lý dữ liệu từ nhiều nguồn khác nhau, bao gồm ảnh vệ tinh, dữ liệu radar, số liệu quan trắc môi trường, cũng như dữ liệu từ các nguồn thông tin xã hội. Các thuật toán học máy có khả năng nhận dạng và xác định các đặc trưng quan trọng liên quan đến lũ quét, như độ cao mực nước, vận tốc dòng chảy, khu vực ngập lụt, v.v. Điều này giúp cải thiện độ chính xác của các mô hình dự báo và phân vùng.

Hơn nữa, các kỹ thuật trí tuệ nhân tạo như học sâu (deep learning) và mạng nơron (neural networks) có khả năng xử lý các chuỗi dữ liệu dài và phức tạp, cho phép dự đoán diễn biến của lũ quét trong tương lai dựa trên các dữ liệu lịch sử. Điều này đóng vai trò quan trọng trong việc chuẩn bị và ứng phó kịp thời với các tình huống khẩn cấp.

Do đó, nghiên cứu “Ứng dụng trí tuệ nhân tạo và dữ liệu địa không gian để phân vùng lũ quét quy mô cấp huyện” với mục tiêu “Đánh giá khả năng ứng dụng công nghệ trí tuệ nhân tạo và dữ liệu địa không gian để nâng cao độ tin cậy trong phân vùng lũ quét quy mô cấp huyện” được thực hiện nhằm tìm hiểu công nghệ trí tuệ nhân tạo và ứng dụng công nghệ để đánh giá tiềm năng trong việc sử dụng trí tuệ nhân tạo để phân vùng lũ quét. Đặc điểm của các mô hình và khả năng phân vùng của các mô hình phổ biến.

## **II. MỤC TIÊU CỦA ĐỀ TÀI**

Mục tiêu của đề tài là đánh giá khả năng ứng dụng công nghệ trí tuệ nhân tạo và dữ liệu địa không gian để nâng cao độ tin cậy trong phân vùng lũ quét quy mô cấp huyện.

## **III. CÁCH TIẾP CẬN**

Nghiên cứu đề xuất một cách tiếp cận tích hợp nhằm ứng dụng trí tuệ nhân tạo (AI) và dữ liệu địa không gian để phân vùng lũ quét tại huyện Mù Cang Chải, tỉnh Yên Bái, với mục tiêu nâng cao độ tin cậy trong phân vùng rủi ro lũ quét ở quy mô cấp huyện. Cách tiếp cận bao gồm các bước chính sau:

**1. Thu thập và chuẩn bị dữ liệu:** Dữ liệu sẽ được thu thập từ nhiều nguồn khác nhau, bao gồm dữ liệu địa không gian (địa hình, sử dụng đất, thảm phủ, loại đất) và khí tượng (lượng mưa). Các nhóm dữ liệu chính bao gồm: (1) Địa hình: Độ dốc, chỉ số ẩm địa hình (TWI), độ cong địa hình, cao độ bình quân lưu vực; (2) Thủy văn: Khoảng cách/chênh lệch độ cao đến sông/suối, diện tích lưu vực, độ dốc lòng dân; (3) Thảm phủ: Chỉ số NDVI, chỉ số CN; (4) Khí tượng: Lượng mưa lớn nhất trong 1 giờ, 3 giờ, 6 giờ và 24 giờ. Dữ liệu sẽ được làm sạch, chuẩn hóa bằng các phương pháp như Z-score Normalization hoặc MinMax để đảm bảo tính nhất quán và loại bỏ nhiễu, đặc biệt với các chỉ số như NDVI và lượng mưa có phân bố lệch.

**2. Xây dựng mô hình học máy và học sâu:** Nghiên cứu sẽ triển khai các mô hình học máy như Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), LightGBM (LGBM), và các mô hình học sâu như Deep Neural Network (DNN) và Convolutional Neural Network (CNN). Các mô hình sẽ được cấu hình với các tham số tối ưu, ví dụ: RF với số cây quyết định và độ sâu tối đa được điều chỉnh, SVM với kernel tuyến tính hoặc phi tuyến, và CNN với số lượng đặc trưng tăng dần qua các lớp tích chập. Dữ liệu huấn luyện sẽ dựa trên nhãn phân loại nguy cơ lũ quét từ 0 (không có lũ) đến 4 (nguy cơ rất cao), được xây dựng từ điều tra thực địa và dữ liệu lịch sử.

**3. Phân tích GIS:** Dữ liệu địa không gian sẽ được xử lý thông qua công cụ GIS để tạo bản đồ thảm phủ và các đặc trưng địa hình như độ dốc, TWI, chỉ số vị trí địa hình (TPI), và cao độ bình quân lưu vực. Ảnh viễn thám từ Sentinel-2 và Sentinel-1C sẽ được sử dụng để xây dựng bản đồ thảm phủ với các tham số đầu vào, hỗ trợ xác định các khu vực có nguy cơ lũ quét cao dựa trên đặc điểm địa hình và thủy văn.

**4. Phân vùng lũ quét:** Kết quả dự kiến là xây dựng bản đồ phân vùng lũ quét, xác định các khu vực có nguy cơ cao, trung bình và thấp tại huyện Mù Cang Chải. Các mô hình sẽ được đánh giá và so sánh dựa trên các chỉ số như độ chính xác, độ nhạy, và khả năng phân biệt các mức nguy cơ. Kết quả sẽ được tích hợp với bản đồ hành chính để hỗ trợ quản lý rủi ro thiên tai.

# **CHƯƠNG 1. TỔNG QUAN TÌNH HÌNH NGHIÊN CỨU XÁC ĐỊNH LŨ QUÉT VÀ ÚNG DỤNG TRÍ TUỆ NHÂN TẠO TRONG PHÂN VÙNG LŨ QUÉT**

## **1.1. Sơ lược về trí tuệ nhân tạo và ứng dụng trí tuệ nhân tạo trong nghiên cứu lũ quét**

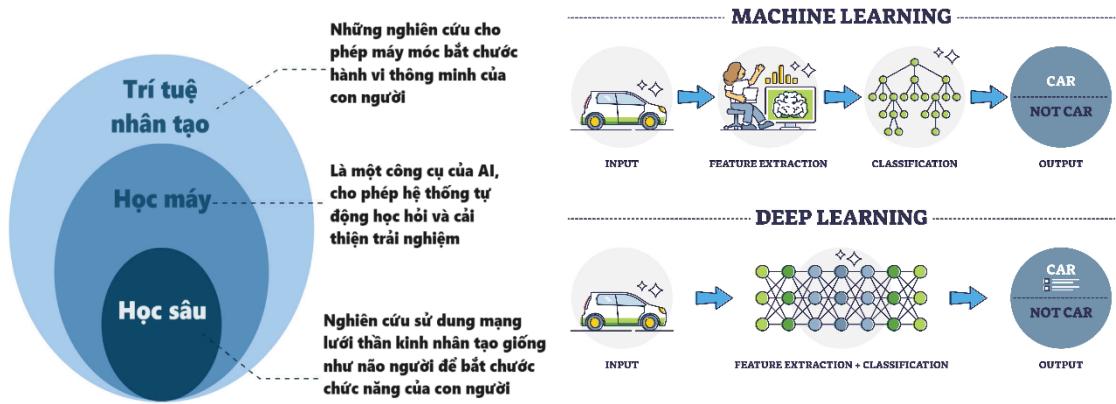
### **1.1.1 Trí tuệ nhân tạo**

Trí tuệ nhân tạo (AI) là lĩnh vực khoa học máy tính chuyên giải quyết các vấn đề nhận thức thường liên quan đến trí tuệ con người, chẳng hạn như học tập, sáng tạo và nhận diện hình ảnh. Các tổ chức hiện đại thu thập vô số dữ liệu từ nhiều nguồn khác nhau như cảm biến thông minh, nội dung do con người tạo, công cụ giám sát và nhật ký hệ thống. Mục tiêu của AI là tạo ra các hệ thống tự học có thể tìm ra ý nghĩa của dữ liệu. Sau đó, AI áp dụng kiến thức thu được để giải quyết các vấn đề mới theo cách giống như con người.

Có thể hiểu một cách đơn giản, trí tuệ nhân tạo chỉ đơn giản là quá trình học tập, từ đó đưa ra phán đoán với những thông tin tương tự những thông tin đã được học tập. Ví dụ như để mô hình trí tuệ nhân tạo có thể đoán được giá nhà tại một khu vực, chỉ với điều kiện dữ liệu đầu vào bao gồm: (1) vị trí; (2) tên chủ đầu tư; (3) vị trí số tầng; (4) diện tích; và (5) tiện ích. Mô hình cần có đủ dữ liệu (có thể lên tới hàng ngàn dữ liệu) các căn hộ tương tự để học, sau đó dự đoán giá nhà cho một nhà bất kỳ với đủ 5 yếu tố trên. Việc này hoàn toàn tương tự như việc cung cấp tư vấn dịch vụ bất động sản của một nhân viên sale bất động sản. Các nhân viên càng lâu năm, dữ liệu được học càng nhiều thì việc dự đoán giá nhà mới càng tiệm cận giá thực tế (tương tự như mô hình trí tuệ nhân tạo, với dữ liệu học càng lớn, mô hình được huấn luyện càng tốt hơn).

Đây là một ví dụ điển hình để có thể hiểu được cách hoạt động của trí tuệ nhân tạo. Trên thực tế, khi càng có nhiều dữ liệu, giá nhà càng có sự biến động mạnh làm cho ngay cả những người làm việc lâu năm cũng khó có thể dựa ra giá tiệm cận do độ phức tạp của dữ liệu. Điều này dẫn đến các thuật ngữ như Overfitting hay Underfitting... Do đó, một lượng dữ liệu vừa đủ và “đáng tin cậy” sẽ tốt hơn rất nhiều việc đưa quá nhiều các “dữ liệu bẩn” vào để đào tạo mô hình làm suy giảm chất lượng mô hình trí tuệ nhân tạo xây dựng. Người xây dựng mô hình trí tuệ nhân tạo cần hiểu rõ bản chất của các yếu tố phụ thuộc dẫn đến kết quả dự đoán thay vì “tra tấn dữ liệu” để xây dựng mô hình.

Trí tuệ nhân tạo bao gồm các thuật toán học máy (machine learning) và học sâu (deep learning) đều được hoạt động theo nguyên tắc sử dụng dữ liệu đầu vào → thông qua thuật toán → dữ liệu đầu ra. Học máy cũng tương tự như “mô hình hộp trắng” và học sâu có những nét tương tự như “mô hình hộp đen”. Do đó về bản chất, học máy có thể được giải thích thông qua các thuật toán một cách tương đối rõ ràng (có thể hiện sự logic bên trong và có thể kiểm tra được), còn học sâu thì mang tính trừu tượng hơn và có những tham số ẩn.



Hình 1-1. Sự khác biệt giữa học máy và học sâu

Phương pháp học máy đòi hỏi người sử dụng có những hiểu biết nhất định và định hướng thuật toán (như dự đoán dạng hàm số, kiểu phân bố dữ liệu), học máy sử dụng nhiều loại thuật toán khác nhau, trong khi đó học sâu lại tập trung vào mạng nơ-ron nhân tạo. Để sử dụng được mô hình học máy, người dùng cần trích xuất những đặc trưng theo hiểu biết (ví dụ như độ dốc từ địa hình...) và đưa vào mô hình, nhưng học sâu có thể tự động học những đặc trưng phức tạp từ dữ liệu đầu vào. Do đó, học sâu cũng yêu cầu nhiều dữ liệu hơn và tài nguyên tính toán mạnh mẽ hơn. Học máy cũng yêu cầu dạng dữ liệu đầu vào là vector (dạng số/matrice...) đã được chuẩn hóa/xử lý, tuy nhiên học sâu có thể nhận dữ liệu thô với sự đa dạng hơn (âm thanh, hình ảnh, số liệu...)

### 1.1.2 *Ứng dụng trí tuệ nhân tạo trong nghiên cứu lũ quét*

AI có tiềm năng to lớn trong nhiều lĩnh vực, bao gồm y tế, tài chính, giáo dục, và môi trường. Một trong những ứng dụng đáng chú ý của AI là trong nghiên cứu và dự đoán lũ quét. Lũ quét là lũ xảy ra bất ngờ trên sườn dốc và trên các sông suối nhỏ miền núi, dòng chảy xiết, thường kèm theo bùn đá, lũ lên nhanh, xuống nhanh, có sức tàn phá lớn (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). Đây là một loại hình thiên tai khó dự báo và gây nhiều thiệt hại cho khu vực miền núi nước ta.

Nói như vậy có nghĩa rằng lũ quét là một dạng lũ, do đó, nó là kết quả phản ứng thủy văn của một lưu vực (được xác định từ mưa → dòng chảy). Điều này dẫn đến quá trình xây dựng mô hình trí tuệ nhân tạo để dự đoán lũ quét tại một vị trí thì vị trí đó cần phải có các thông số đại diện cho lưu vực mà cửa ra chính là điểm dự đoán. Nếu đưa các thông số chỉ đại diện cho các điểm (thông số nội tại), việc dự đoán lũ quét có thể gây ra sai lệch rất lớn (ví dụ một điểm trên mái nhà có các yếu tố bất lợi như độ dốc lớn, không có thảm phủ...). Bên cạnh đó, các thông số được đưa vào cũng cần phải là các thông số ảnh hưởng trực tiếp đến quá trình hình thành dòng chảy và mức độ sẵn có của dữ liệu.

Bên cạnh ứng dụng dự đoán (tính nhạy cảm/nguy cơ) lũ quét, mô hình trí tuệ nhân tạo cũng được sử dụng để phát hiện lũ quét (nhân dạng lũ) từ ảnh viễn thám. Về bản chất, đây là quá trình phát hiện “nước” trên bề mặt giữa hai thời kỳ trước và sau lũ để so sánh và phát hiện khu vực bị ngập lũ. Ứng dụng này ban đầu sử dụng cho phát hiện lũ và ngập lũ, sau đó triển khai linh hoạt để ứng dụng trong phát hiện lũ quét.

Nghiên cứu này tin rằng với sự kết hợp của các tham số địa không gian (từ ảnh vệ tinh và quan sát bề mặt) và lượng mưa, mô hình trí tuệ nhân tạo có thể phát hiện ra được nguy cơ lũ quét một cách đáng tin cậy.

## **1.2. Tổng quan các loại dữ liệu địa không gian và ứng dụng trong nghiên cứu lũ quét.**

### ***1.1.1. Giới thiệu về dữ liệu địa không gian***

Dữ liệu địa không gian là dữ liệu không lồ do các trung tâm nghiên cứu và tổ chức tạo ra, đòi hỏi phải truy cập và truy xuất nhanh chóng để đưa ra các quyết định quan trọng liên quan đến khủng hoảng như thiên tai, hỏa hoạn... Dữ liệu này bao gồm việc sử dụng hệ thống máy chủ không lồ để cung cấp tài nguyên tính toán và lưu trữ để truy cập và tìm kiếm dữ liệu này, điều này đặt ra những thách thức do khối lượng dữ liệu không lồ, vị trí ảnh chụp đa dạng và mang tính toàn cầu, đặc biệt là tính không đồng nhất về định dạng và cấu trúc giữa các loại dữ liệu (Bakri Bashir, Mohammed, et al., 2016).

Dữ liệu địa không gian là dữ liệu về các đối tượng, sự kiện hoặc hiện tượng có vị trí trên bề mặt Trái Đất (Stock, Kristin & Guesgen, Hans, 2016). Vị trí này có thể tĩnh trong ngắn hạn (ví dụ: vị trí của một con đường, một trận động đất, trẻ em sống trong nghèo đói) hoặc động (ví dụ: một chiếc xe đang di chuyển hoặc một người đi bộ, sự lây lan của một bệnh truyền nhiễm). Dữ liệu địa không gian kết hợp thông tin vị trí (thường là tọa độ trên Trái Đất), thông tin thuộc tính (các đặc điểm của đối tượng, sự kiện hoặc hiện tượng đang xét) và thường cả thông tin thời gian (thời gian hoặc khoảng thời gian mà vị trí và thuộc tính tồn tại).

Nhiều dữ liệu địa không gian có mối quan tâm chung đối với nhiều người dùng. Ví dụ, đường xá, địa phương, vùng nước và tiện ích công cộng hữu ích như thông tin tham khảo cho nhiều mục đích. Vì lý do này, cho dù được thu thập bởi tổ chức công cộng hay tư nhân, một lượng lớn dữ liệu địa không gian có sẵn dưới dạng dữ liệu mở. Điều này có nghĩa là nó có thể được truy cập tự do bởi người dùng và được cung cấp thông qua các tiêu chuẩn mở. Việc phát triển và sử dụng các tiêu chuẩn mở trong cộng đồng địa không gian đã được hỗ trợ mạnh mẽ do phạm vi rộng lớn của các ứng dụng mà dữ liệu địa không gian có thể được áp dụng và do số lượng lớn các cơ quan trên toàn cầu và địa phương tham gia vào việc thu thập dữ liệu như vậy. Hệ thống Google Earth Engine (GEE) là một trong những hệ thống lớn nhất thế giới về dữ liệu địa không gian

toàn cầu, cho phép người dùng có thể truy cập, sử dụng dữ liệu địa không gian trên toàn thế giới miễn phí.

Theo truyền thống, dữ liệu địa không gian chủ yếu được thu thập bởi các bộ, ngành chính phủ, thường liên quan đến nhiều bộ, ngành trong bất kỳ khu vực quản lý nào. Ví dụ, một bộ, ngành có thể chịu trách nhiệm thu thập dữ liệu ranh giới đất đai (địa chính), một bộ, ngành khác chịu trách nhiệm thu thập dữ liệu đường xá và giao thông, một bộ, ngành khác chịu trách nhiệm thu thập dữ liệu môi trường và một bộ, ngành khác chịu trách nhiệm thu thập dữ liệu y tế, v.v., theo từng lĩnh vực chuyên môn của họ. Việc tích hợp dữ liệu từ các bộ, ngành khác nhau thường khó khăn do sử dụng các định dạng, mô hình dữ liệu và ngữ nghĩa khác nhau. Ví dụ, bộ, ngành chịu trách nhiệm về ranh giới đất đai có thể đã thu thập dữ liệu về đường xá theo nghĩa là khu vực đã được xác nhận là đường xá hợp pháp, trong khi bộ, ngành chịu trách nhiệm bảo trì đường xá có thể đã thu thập dữ liệu về con đường được xây dựng: vật liệu, bề mặt và diện tích vật lý của chính con đường đó, và bộ, ngành liên quan đến bảo tồn có thể đã thu thập dữ liệu về các điểm giao cắt đường xá của động vật hoang dã. Trong mỗi trường hợp, "đường xá" có nghĩa khác nhau (có ngữ nghĩa khác nhau), có khả năng có các thuộc tính khác nhau, cấu trúc dữ liệu khác nhau và cơ chế nhận dạng khác nhau. Hơn nữa, tình huống này lặp lại ở mỗi khu vực quản lý khác nhau, do đó các bộ, ngành duy trì cùng một loại dữ liệu (ví dụ: ranh giới đất đai) ở các khu vực quản lý khác nhau sử dụng các định dạng, cấu trúc và ngữ nghĩa khác nhau cho dữ liệu của riêng họ. Trong những trường hợp cần chia sẻ dữ liệu giữa các khu vực quản lý, điều này gây ra các vấn đề. Ví dụ, ở Úc, mỗi tiểu bang và lãnh thổ sử dụng hệ thống, định dạng và cấu trúc riêng của họ cho mỗi bộ dữ liệu địa không gian khác nhau, khiến việc tạo ra dữ liệu thống nhất trên toàn tiểu bang và lãnh thổ hoặc tạo ra một bộ dữ liệu quốc gia cho các vấn đề toàn quốc trở nên rất khó khăn. Ở châu Âu, những thách thức tương tự đã xảy ra với sự bổ sung của sự khác biệt ngôn ngữ. Ví dụ, dữ liệu về các địa điểm được bảo vệ thường yêu cầu một cách tiếp cận xuyên biên giới vì môi trường sống của các loài không dừng lại ở biên giới quốc gia, nhưng mỗi quốc gia duy trì bộ dữ liệu riêng của mình với các định dạng và cấu trúc khác nhau (Stock, Kristin & Guesgen, Hans, 2016).

Do những thách thức tích hợp dữ liệu này, dữ liệu địa không gian mở đã là một mục tiêu quan trọng trong cộng đồng địa không gian trong một số năm. Hiệp hội Địa không gian Mở (OGC) đã phát triển một số tiêu chuẩn để cho phép chia sẻ dữ liệu mở, trong đó không thể không kể đến Đặc tả Dịch vụ Tính năng Web (WFS), định nghĩa một yêu cầu dịch vụ web và định dạng phản hồi để cho phép các nhà cung cấp dữ liệu cung cấp dữ liệu của họ cho người dùng dữ liệu. Tiêu chuẩn này và các tiêu chuẩn OGC khác chủ yếu tập trung vào định dạng dữ liệu và không giải quyết mô hình dữ liệu hoặc ngữ nghĩa của nội dung dữ liệu.

Dữ liệu địa không gian và viễn thám có mối quan hệ đồng điệu chặt chẽ trong nghiên cứu và ứng dụng thực tiễn. Về bản chất, dữ liệu viễn thám chính là một dạng dữ liệu địa không gian, mang thông tin về vị trí địa lý và thuộc tính của đối tượng. Ảnh vệ tinh, dữ liệu radar đều được địa chiêu và tích hợp vào hệ thống thông tin địa lý.

Trong xử lý và phân tích, hai loại dữ liệu này bổ trợ cho nhau. Dữ liệu viễn thám cung cấp thông tin bề mặt Trái Đất theo thời gian thực, trong khi dữ liệu địa không gian khác (DEM, bản đồ địa chất, thổ nhưỡng...) cung cấp thông tin nền tảng. Sự kết hợp này tạo ra khả năng phân tích toàn diện.

Về công nghệ, các phần mềm GIS hiện đại đều tích hợp công cụ xử lý ảnh viễn thám. Ngược lại, các phần mềm viễn thám cũng có khả năng phân tích không gian. Sự hội tụ này phản ánh mối liên kết không thể tách rời giữa hai lĩnh vực. Trong xu hướng phát triển, công nghệ địa không gian và viễn thám ngày càng hòa quyện, tạo nên các giải pháp tổng thể trong nhiều ứng dụng từ quản lý tài nguyên đến ứng phó thiên tai.

### ***1.1.2. Quá trình phát triển ảnh viễn thám theo lịch sử***

Viễn thám là khoa học thu thập thông tin về bề mặt trái đất mà không cần tiếp xúc vật lý với nó. Kỹ thuật này bao gồm việc sử dụng nhiều thiết bị khác nhau để thu thập thông tin về bề mặt trái đất, chẳng hạn như các đặc điểm, tính chất vật lý và điều kiện của nó. Viễn thám có lịch sử lâu đời bắt đầu từ đầu thế kỷ 19.

Trường hợp đầu tiên được ghi nhận về cảm biến từ xa có thể bắt nguồn từ năm 1858 khi nhiếp ảnh gia người Pháp Gaspard-Félix Tournachon chụp ảnh trên không Paris từ khinh khí cầu (Wikipedia, n.d.). Những bức ảnh của ông cho thấy thành phố từ góc nhìn của một chú chim và cung cấp một góc nhìn mới về bố cục và thiết kế của thành phố.

Việc sử dụng sớm nhất của cảm biến từ xa là cho mục đích quân sự, cụ thể là để trinh sát trong Thế chiến thứ nhất. Máy bay được sử dụng để chụp ảnh trên không của lãnh thổ đối phương, sau đó được phân tích để phục vụ cho tình báo quân sự. Trong Thế chiến thứ hai, công nghệ cảm biến từ xa đã tiến bộ đáng kể với sự phát triển của máy bay trinh sát ảnh, chẳng hạn như Lockheed P-38 Lightning và North American P-51 Mustang. Những máy bay này được trang bị máy ảnh có thể chụp ảnh mặt đất có độ phân giải cao từ độ cao lớn.

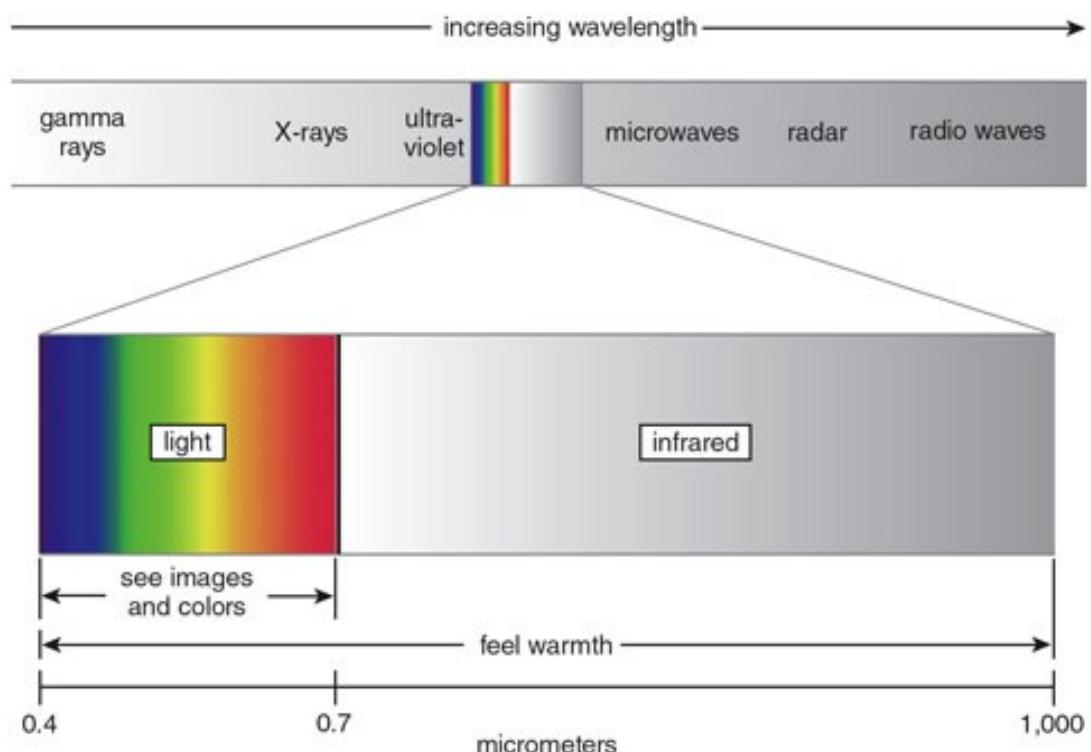
#### ***1.1.2.1. Phát hiện ra tia hồng ngoại (1800)***

Việc phát hiện ra tia hồng ngoại của Sir William Herschel vào năm 1800 đã đánh dấu sự khởi đầu của việc hiểu biết về quang phổ điện từ (White, Jack, 2012). Các thí nghiệm của Herschel bao gồm việc truyền ánh sáng mặt trời qua lăng kính và đo nhiệt độ thay đổi trong các màu ánh sáng khác nhau. Ông phát hiện ra rằng nhiệt độ tăng lên vượt quá ánh sáng đỏ có thể nhìn thấy, dẫn đến việc xác định bức xạ hồng ngoại. Khám

phá này đã đặt nền tảng cho các cuộc khám phá trong tương lai về sóng điện từ và các ứng dụng của chúng trong công nghệ cảm biến từ xa, cho phép các nhà khoa học hiểu và sử dụng các bước sóng vượt quá quang phổ có thể nhìn thấy cho nhiều mục đích quan sát khác nhau.

Ranh giới giữa ánh sáng và hồng ngoại được xác định bởi giới hạn bước sóng dài của phản ứng của mắt người (một dải bước sóng hẹp từ 0,4 đến 0,7 micromet). Kinh nghiệm hàng ngày sẽ không khiến mọi người tin rằng ánh sáng và hồng ngoại là cùng một loại năng lượng. Thật vậy, hai bằng chứng thuyết phục cho thấy, về mặt logic, rằng chúng không liên quan.

Đầu tiên, chúng ta trải nghiệm ánh sáng và hồng ngoại khác nhau bằng các giác quan khác nhau. Mọi người nhìn thấy ánh sáng, cảm nhận các bước sóng khác nhau thành các màu khác nhau, nhưng mọi người chỉ cảm thấy hồng ngoại như hơi ấm. Thứ hai, ánh sáng và hồng ngoại không phải lúc nào cũng được tìm thấy cùng nhau. Hầu hết các nguồn sáng cũng phát ra hồng ngoại, nhưng hồng ngoại thường được tìm thấy riêng lẻ. Một ví dụ quen thuộc là một chiếc lò nướng điện không đủ nóng để phát sáng: Nếu căn phòng hoàn toàn tối, mọi người vẫn có thể cảm nhận được hơi ấm từ lò nướng nhưng không thể nhìn thấy nó. Tia hồng ngoại trải dài trong phạm vi bước sóng giữa ánh sáng và vi sóng, lên đến khoảng 1.000 micromet.

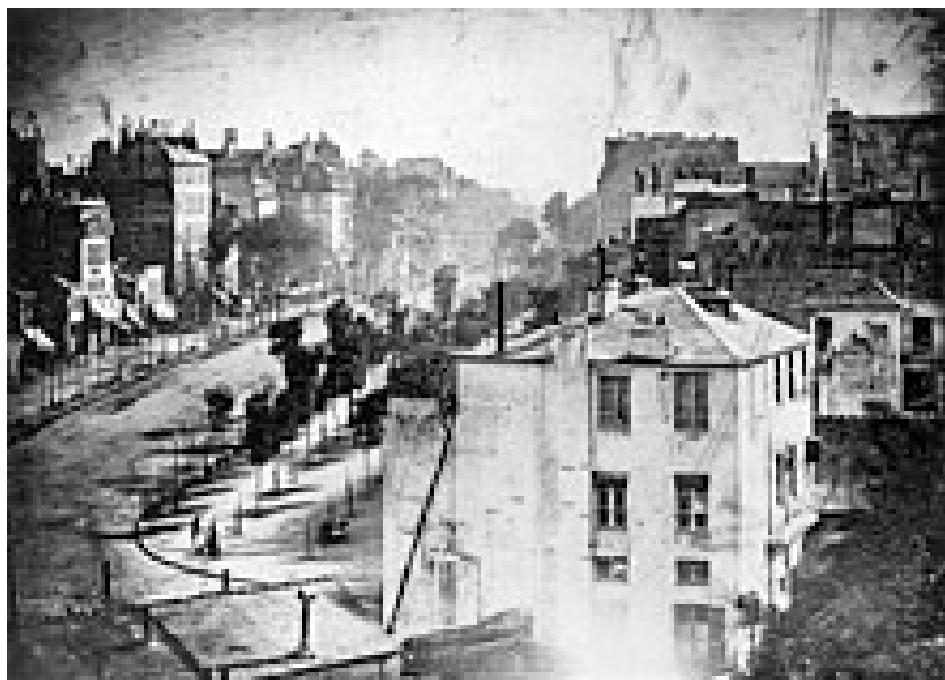


Hình 1. Bước sóng ánh sáng và bước sóng hồng ngoại

### 1.1.2.2. Sự khởi đầu của kỹ thuật nhiếp ảnh đen trắng (1839)

Trong khi phát hiện của Herschel về tia hồng ngoại cung cấp nền tảng lý thuyết về ánh sáng ngoài tầm nhìn, phát minh của Daguerre đã giúp chụp ảnh bằng các nguyên lý đó trong các ứng dụng thực tế. Sự kết hợp giữa hiểu biết về các bước sóng khác nhau (nhờ Herschel) và có phương pháp ghi lại hình ảnh (nhờ Daguerre) đã tạo tiền đề cho những tiến bộ trong tương lai về công nghệ cảm biến từ xa.

Năm 1839, Louis Daguerre giới thiệu quy trình daguerreotype (Davidson, Michael W., 2010), đây là phương pháp chụp ảnh đầu tiên được công khai. Kỹ thuật này bao gồm việc phơi sáng một tấm đồng mạ bạc, tạo ra những hình ảnh có độ chi tiết cao, thu hút công chúng. Khả năng chụp các chi tiết tinh tế của daguerreotype đã đánh dấu một cột mốc quan trọng trong nhiếp ảnh, cho phép ghi lại các sự kiện lịch sử và chân dung cá nhân. Sự phổ biến của quy trình này đã thúc đẩy những tiến bộ nhanh chóng trong các kỹ thuật và vật liệu chụp ảnh, cuối cùng dẫn đến sự quan tâm rộng rãi của công chúng đối với nhiếp ảnh như một hình thức nghệ thuật và phương tiện truyền thông.



Hình 2. Một trong những bức ảnh đầu tiên được thực hiện vào năm 1837 hoặc 1938 do Louis Daguerre sử dụng quy trình daguerreotype

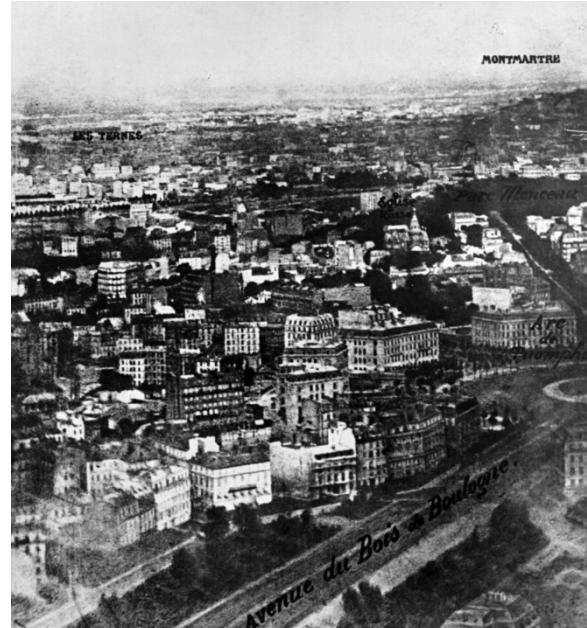
Cụm từ sự ra đời của nhiếp ảnh đã được nhiều tác giả khác nhau sử dụng để chỉ những điều khác nhau – hoặc là việc công khai quá trình này (vào năm 1839) như một phép ẩn dụ để chỉ ra rằng trước đó, quá trình chụp ảnh daguerreotype đã được giữ bí mật hoặc là ngày chụp bức ảnh đầu tiên bằng máy ảnh hoặc máy ảnh (sử dụng quy trình nhựa đường hoặc heliography)

### 1.1.2.3. Phát hiện ra cả phổ hồng ngoại và phổ nhìn thấy (1847)

Năm 1847, các nhà khoa học đã mở rộng hiểu biết của mình để bao gồm cả phổ hồng ngoại và phổ nhìn thấy (được hiểu là có chung đặc tính như phản xạ và khúc xạ) (White, Jack, 2012). Khám phá này nhấn mạnh rằng các bước sóng khác nhau có thể được sử dụng cho nhiều ứng dụng khác nhau ngoài việc chỉ biểu diễn trực quan. Khả năng chụp ảnh bằng ánh sáng hồng ngoại trở nên quan trọng đối với các ứng dụng cảm biến từ xa sau này, đặc biệt là trong việc đánh giá các điều kiện môi trường như sức khỏe thảm thực vật và thay đổi sử dụng đất.

### 1.1.2.4. Chụp ảnh từ khinh khí cầu (1850-1860)

Giai đoạn từ năm 1850 đến năm 1860 chứng kiến những người tiên phong thử nghiệm chụp ảnh trên không bằng khinh khí cầu. Gaspard-Félix Tournachon (Nadar) (Philip McCouat, 2016) đã chụp thành công hình ảnh từ các vị trí cao, cung cấp góc nhìn mới về cảnh quan và môi trường đô thị. Sự đổi mới này chịu ảnh hưởng trực tiếp từ những tiến bộ trước đó trong nghiệp ảnh và chứng minh cách nhìn từ trên không có thể nâng cao nỗ lực lập bản đồ và ghi chép địa lý (hình bên).



### 1.1.2.5. Xây dựng lý thuyết về phổ điện từ (1873)

Năm 1873, James Clerk Maxwell đã phát triển lý thuyết về sóng điện từ, cung cấp nền tảng khoa học để hiểu cách ánh sáng hoạt động như một sóng và một hạt (Maxwell, James Clerk, 2010). Khung lý thuyết này rất quan trọng để hiểu cách các bước sóng khác nhau có thể được khai thác cho mục đích giao tiếp và quan sát. Công trình của Maxwell dựa trên những khám phá trước đó về tia hồng ngoại và ánh sáng nhìn thấy, mở đường cho những tiến bộ trong tương lai về công nghệ cảm biến từ xa dựa trên bức xạ điện từ.

### 1.1.2.6. Chụp ảnh từ máy bay (1906)

Những bức ảnh chụp trên không thành công đầu tiên được chụp từ máy bay diễn ra vào năm 1906 khi phi công George Lawrence chụp ảnh trên bầu trời San Francisco bằng máy ảnh khổ lớn treo trên máy bay hai tầng (Christopher Turner, n.d.). Sự đổi mới này cải thiện đáng kể độ phân giải không gian so với chụp ảnh bằng khinh khí cầu, cho phép lập bản đồ chi tiết hơn và các nỗ lực trinh sát. Các kỹ thuật được phát triển thông

qua chụp ảnh bằng khinh khí cầu trước đó đã đặt nền tảng cho sự tiến bộ này, nâng cao khả năng trinh sát quân sự và các sáng kiến quy hoạch đô thị.



Hình 3. Hình ảnh được chụp từ máy bay năm 1906 thực hiện bởi Lawrence

#### 1.1.2.7. Sự phát triển của công nghệ radar (1930-1940)

Sự phát triển của công nghệ radar trong những năm 1930 đến 1940 đánh dấu bước tiến đáng kể trong khả năng cảm biến từ xa bổ sung cho các kỹ thuật hình ảnh quang học. Các quốc gia như Đức, Hoa Kỳ và Anh đã đầu tư mạnh vào nghiên cứu radar trong Thế chiến II, nhận ra tiềm năng phát hiện các vật thể ở khoảng cách xa bất kể điều kiện tầm nhìn. Bước nhảy vọt về công nghệ này dựa trên các phương pháp hình ảnh trước đó bằng cách cung cấp một phương tiện quan sát thay thế không bị ảnh hưởng bởi thời tiết hoặc điều kiện ánh sáng.

Công việc phát triển nghiêm túc về radar bắt đầu vào những năm 1930, nhưng ý tưởng cơ bản về radar có nguồn gốc từ các thí nghiệm cổ điển về bức xạ điện từ do nhà vật lý người Đức thực hiện Heinrich Hertz vào cuối những năm 1880 (Kock, Winston E., 1978). Hertz bắt đầu xác minh bằng thực nghiệm công trình lý thuyết trước đó của nhà vật lý người Scotland James Clerk Maxwell. Maxwell đã xây dựng các phương trình tổng quát của trường điện từ, xác định rằng cả ánh sáng và sóng vô tuyến đều là ví dụ về sóng điện từ chịu sự chi phối của cùng một định luật cơ bản nhưng có tần số rất khác nhau. Công trình của Maxwell dẫn đến kết luận rằng sóng vô tuyến có thể phản xạ từ các vật thể kim loại và khúc xạ bởi môi trường điện môi, giống như sóng ánh sáng. Hertz đã chứng minh các

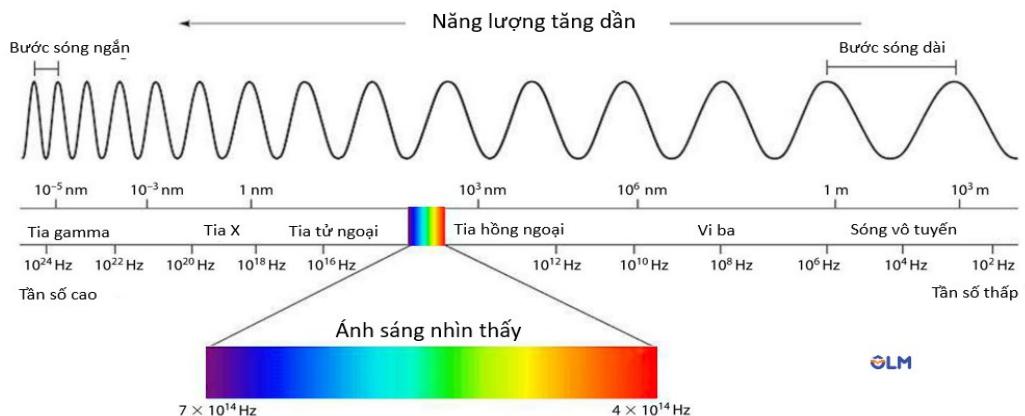


Early military radar system

đặc tính này vào năm 1888, sử dụng sóng vô tuyến ở bước sóng 66 cm (tương ứng với tần số khoảng 455 MHz).

#### 1.1.2.8. Xác định phạm vi quang phổ từ khả kiến đến vô hình (1950)

Vào đầu những năm 1950, các nhà nghiên cứu bắt đầu xác định toàn bộ dải quang phổ bao gồm cả bước sóng nhìn thấy và không nhìn thấy (như hồng ngoại) (Sliney, David H., et al., 2012). Sự hiểu biết toàn diện này cho phép các nhà khoa học phát triển các cảm biến có khả năng thu thập dữ liệu trên nhiều dải quang phổ, nâng cao khả năng giám sát các thay đổi về môi trường như sức khỏe thảm thực vật và chất lượng nước—một sự tiến hóa bắt nguồn từ những khám phá trước đó về bức xạ hồng ngoại. Đây cũng chính là thời điểm bắt đầu nghiên cứu về các ảnh đa phổ.



Hình 4. Dải bước sóng điện từ

#### 1.1.2.9. Khám phá không gian (1961)

Vào ngày 12 tháng 4 năm 1961, Yuri Gagarin trở thành người đầu tiên bay quanh Trái Đất trên tàu Vostok 1, đánh dấu một cột mốc quan trọng trong khám phá không gian, mở rộng khả năng quan sát vượt ra ngoài các phương pháp máy bay truyền thống (Heppener, Marc, 2008). Sự kiện này mở ra những con đường mới để chụp ảnh Trái Đất từ không gian bằng công nghệ vệ tinh, dựa trên những tiến bộ trước đây trong chụp ảnh trên không đồng thời giới thiệu những khả năng mới cho việc giám sát toàn cầu.

#### 1.1.2.10. Lần đầu tiên sử dụng thuật ngữ Viễn thám (1960-1970)

Thuật ngữ "viễn thám" lần đầu tiên được đặt ra trong giai đoạn này khi các nhà khoa học tìm cách định nghĩa hoạt động thu thập thông tin về một vật thể hoặc khu vực mà không cần tiếp xúc trực tiếp thông qua các cảm biến trên vệ tinh hoặc máy bay bởi Evelyn Lord Pruitt (Fussell, J., Rundquist & D., & Harrington, J. A., 1986). Khung khái niệm này đã cung cấp cho viễn thám như một lĩnh vực riêng biệt trong khoa học môi trường và địa lý, dẫn đến



Evelyn Lord Pruitt, 1918–2000

việc tăng cường tài trợ nghiên cứu và những tiến bộ công nghệ dựa trên các kỹ thuật hình ảnh trước đó.

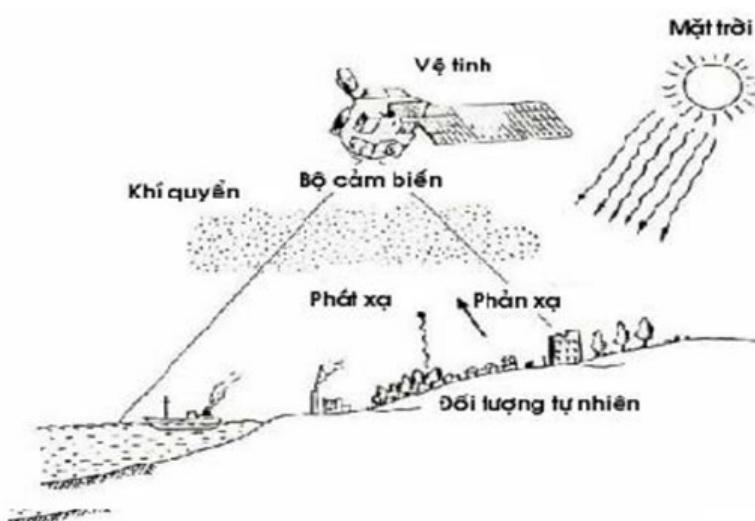
#### 1.1.2.11. Phóng vệ tinh Landsat-1 (1972)

Vào ngày 23 tháng 7 năm 1972, NASA đã phóng Landsat-1 - vệ tinh đầu tiên được thiết kế riêng cho mục đích quan sát Trái đất - đánh dấu một thời điểm quan trọng trong lịch sử cảm biến từ xa (Landsat Science, n.d.). Được trang bị máy quét đa phổ (MSS), Landsat-1 cung cấp phạm vi phủ sóng liên tục của bề mặt Trái đất theo thời gian, cho phép giám sát có hệ thống các thay đổi về sử dụng đất, hoạt động nông nghiệp, tỷ lệ phá rừng và mở rộng đô thị - những tiến bộ có thể thực hiện được nhờ nhiều thập kỷ nghiên cứu về công nghệ hình ảnh.

Từ năm 1972 cho đến nay, rất nhiều loại vệ tinh khác với mục đích quan sát trái đất đã được phóng lên quỹ đạo nhằm nghiên cứu các hiện tượng tự nhiên, dự báo thời tiết và các hoạt động khác phục vụ con người. Đây là nguồn tài nguyên vô cùng quý giá trong việc hiểu rõ hơn về các hiện tượng và cơ chế hình thành, đồng thời có thể nhận biết kịp thời những thay đổi của trái đất và đưa ra những dự đoán trong tương lai.

#### 1.1.3. Các loại ảnh viễn thám và ứng dụng

Nguyên lý cơ bản của viễn thám đó là đặc trưng phản xạ hay bức xạ của các đối tượng tự nhiên tương ứng với từng giải phổ khác nhau. Kết quả của việc giải đoán các lớp thông tin phụ thuộc rất nhiều vào sự hiểu biết về mối quan giữa đặc trưng phản xạ phổ với bản chất, trạng thái của các đối tượng tự nhiên. Những thông tin về đặc trưng phản xạ phổ của các đối tượng tự nhiên sẽ cho phép các nhà chuyên môn chọn các kênh ảnh tối ưu, chứa nhiều thông tin nhất về đối tượng nghiên cứu, đồng thời đây cũng là cơ sở để phân tích nghiên cứu các tính chất của đối tượng, tiến tới phân loại chúng.



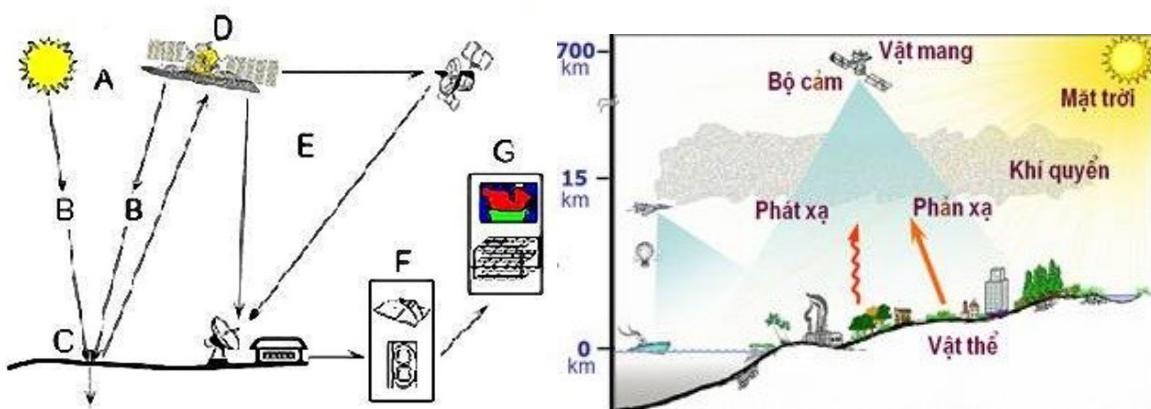
Hình 5. Mô tả về nguyên lý thu thập dữ liệu ảnh bằng công nghệ viễn thám.

Bộ cảm biến là thiết bị được sử dụng để thu thập sóng điện từ được phản xạ / bức xạ từ vật thể, cảm biến đó có thể là máy chụp hoặc máy quét. Những phương tiện dùng để mang các cảm biến này được gọi là vật mang, chúng có thể là máy bay, tàu con thoi, vệ tinh hoặc khinh khí cầu...

Nguồn năng lượng chính được sử dụng để thu thập dữ liệu trong công nghệ viễn thám là bức xạ mặt trời, bộ cảm biến được đặt trên vật mang thu nhận năng lượng của sóng điện từ do các vật thể phản xạ hay bức xạ lại. Thông tin về năng lượng mà các vật thể phản xạ được thu thập lại bởi ảnh viễn thám, thông qua quá trình xử lý tự động trên máy tính hoặc giải đoán trực tiếp hình ảnh theo kinh nghiệm của chuyên gia viễn thám để cho ra kết quả cuối cùng.

Toàn bộ quá trình thu nhận và xử lý ảnh bằng công nghệ viễn thám bao gồm 5 thành phần cơ bản sau:

- Nguồn cung cấp năng lượng.
- Sự tương tác của nguồn năng lượng đó với khí quyển.
- Sự tương tác của năng lượng với các vật thể trên bề mặt Trái Đất.
- Chuyển đổi năng lượng phản xạ từ vật thể thành dữ liệu ảnh số thông qua bộ cảm biến.
- Hiển thị ảnh số, là đầu vào cho việc giải đoán và xử lý dữ liệu.



Hình 6. Sơ đồ về quá trình thu nhận, xử lý và ứng dụng của viễn thám.

Trong đó:

- A: Năng lượng sóng điện từ được bức xạ từ nguồn cấp.
- B: Năng lượng từ A tương tác với các yếu tố có trong khí quyển.
- C: Năng lượng từ B tương tác với các thành phần trên bề mặt Trái Đất.
- D: Năng lượng C phản xạ lại và được ghi nhận bởi bộ cảm biến.
- E: Năng lượng sau khi thu nhận bởi cảm biến được truyền về trạm thu để xử lý.
- F: Giải đoán và phân tích ảnh viễn thám thu nhận được.
- G: Ứng dụng dữ liệu sau khi giải đoán vào các lĩnh vực khác nhau.

Quá trình thu thập dữ liệu thông qua bộ cảm biến chịu tác động mạnh mẽ bởi khả năng lan truyền sóng điện từ trong khí quyển. Mỗi dải phổ điện từ sẽ có đặc điểm và một tốc độ lan truyền sóng khác nhau, ví dụ như:

- **Khả kiến (hay còn gọi là bước sóng nhìn thấy):** Có bước sóng từ 0,4 – 0,76 $\mu\text{m}$ , rất ít hấp thu bởi oxy, hơi nước. Năng lượng phản xạ cực đại khi bước sóng đạt 0,5 $\mu\text{m}$  trong khí quyển. Năng lượng do dải phổ điện từ này cung cấp giữ vai trò quan trọng trong viễn thám.
- **Sau Khả kiến (cận hồng ngoại – hồng ngoại sóng ngắn):** Hồng ngoại gần trung bình, có bước sóng từ 0,77-1,34 & 1,55-2,4, năng lượng phản xạ mạnh với các bước sóng hồng ngoại gần từ 0,77 – 0,9, sử dụng trong chụp ảnh hồng ngoại theo dõi biến đổi thực vật từ 1,55-2,4
- **Hồng ngoại nhiệt:** Có bước sóng từ 3 – 22 $\mu\text{m}$ , có một số vùng bị hấp thụ mạnh bởi hơi nước, dải sóng điện từ này giữ vai trò quan trọng trong việc phát hiện cháy rừng và hoạt động của núi lửa (từ 3,5 – 5 $\mu\text{m}$ ). Bức xạ nhiệt của Trái Đất có năng lượng cao nhất được phát hiện qua dải sóng hồng ngoại nhiệt tại bước sóng 10 $\mu\text{m}$ .
- **Vô tuyến (Radar):** Có bước sóng từ 1mm – 30cm, dải sóng điện từ này không được hấp thụ mạnh từ khí quyển, cho phép cảm biến thu nhận dữ liệu 24/24 mà không bị ảnh hưởng bởi mây, sương mù hay mưa.

Bảng 1-1. Tổng hợp các ảnh viễn thám phân loại theo cơ chế hoạt động

Tiêu chí	Ảnh viễn thám quang học (Optical)	Ảnh hồng ngoại nhiệt (Thermal IR)	Ảnh radar chủ động (Active Radar)	Ảnh viễn thám vi ba thụ động (Passive Microwave)
Phân nhóm	Ảnh thụ động (Passive)	Ảnh thụ động (Passive)	Ảnh chủ động (Active)	Ảnh thụ động (Passive)
Dải phổ	0.4-2.5 $\mu\text{m}$ (Visible & Near-IR)	3-14 $\mu\text{m}$ (Thermal IR)	1 mm - 1 m (Microwave)	1 mm - 1 m (Microwave)
Mô tả/Giới thiệu	Thu nhận bức xạ phản xạ từ mặt trời trong dải sóng nhìn thấy và cận hồng ngoại	Thu nhận bức xạ nhiệt tự nhiên từ vật thể trên bề mặt Trái đất	Phát và thu sóng radar chủ động (bao gồm cả SAR)	Thu nhận bức xạ tự nhiên trong dải vi ba
Ưu điểm	<ul style="list-style-type: none"> <li>- Độ phân giải không gian cao</li> <li>- Dễ giải đoán</li> <li>- Chi phí thấp</li> <li>- Có thông tin màu sắc thực</li> </ul>	<ul style="list-style-type: none"> <li>- Hoạt động cả ngày và đêm</li> <li>- Phát hiện nhiệt độ</li> <li>- Theo dõi hoạt động nhiệt</li> </ul>	<ul style="list-style-type: none"> <li>- Hoạt động mọi thời tiết</li> <li>- Xuyên qua mây</li> <li>- Độ phân giải cao (SAR)</li> </ul>	<ul style="list-style-type: none"> <li>- Không cần nguồn phát</li> <li>- Đo được độ ẩm</li> <li>- Xuyên qua mây nhẹ</li> </ul>

Tiêu chí	Ảnh viễn thám quang học (Optical)	Ảnh hồng ngoại nhiệt (Thermal IR)	Ảnh radar chủ động (Active Radar)	Ảnh viễn thám vi ba thụ động (Passive Microwave)
Nhược điểm	<ul style="list-style-type: none"> <li>- Phụ thuộc thời tiết</li> <li>- Không xuyên mây</li> <li>- Chỉ hoạt động ban ngày</li> </ul>	<ul style="list-style-type: none"> <li>- Độ phân giải trung bình</li> <li>- Ảnh hưởng bởi điều kiện khí quyển</li> <li>- Chi phí thiết bị cao</li> </ul>	<ul style="list-style-type: none"> <li>- Chi phí rất cao</li> <li>- Xử lý phức tạp</li> <li>- Nhiều tín hiệu</li> </ul>	<ul style="list-style-type: none"> <li>- Độ phân giải thấp</li> <li>- Tín hiệu yếu</li> <li>- Dễ nhiễu</li> </ul>
Ứng dụng chung	<ul style="list-style-type: none"> <li>- Bản đồ thực vật</li> <li>- Quy hoạch đô thị</li> <li>- Giám sát môi trường</li> </ul>	<ul style="list-style-type: none"> <li>- Phát hiện cháy rừng</li> <li>- Nghiên cứu đô thị nhiệt</li> <li>- Theo dõi hoạt động núi lửa</li> </ul>	<ul style="list-style-type: none"> <li>- Bản đồ 3D</li> <li>- Quan trắc biển</li> <li>- Giám sát biến dạng</li> </ul>	<ul style="list-style-type: none"> <li>- Dự báo thời tiết</li> <li>- Đo độ ẩm đất</li> <li>- Nghiên cứu đại dương</li> </ul>
Ứng dụng cho lũ quét	<ul style="list-style-type: none"> <li>- Đánh giá thiệt hại</li> <li>- Lập bản đồ rủi ro</li> <li>- Theo dõi phục hồi</li> </ul>	<ul style="list-style-type: none"> <li>- Phát hiện vùng ngập</li> <li>- Đánh giá nhiệt độ nước</li> <li>- Theo dõi dòng chảy</li> </ul>	<ul style="list-style-type: none"> <li>- Bản đồ ngập chi tiết</li> <li>- Giám sát đê điều</li> <li>- Cảnh báo sớm</li> </ul>	<ul style="list-style-type: none"> <li>- Đánh giá độ ẩm đất</li> <li>- Dự báo mưa</li> <li>- Theo dõi tích nước</li> </ul>
Danh sách ảnh	Landsat, SPOT, Sentinel-2...	ASTER, Thermal, MODIS, Landsat TIR...	TerraSAR-X, Sentinel-1, RADARSAT...	SMOS, AMSR-E, SSM/I...
Cơ chế hoạt động	Thu bức xạ phản xạ ánh sáng mặt trời	Thu bức xạ nhiệt từ vật thể	Phát và thu sóng radar phản xạ	Thu bức xạ tự nhiên dài vi ba

Ngoài các loại ảnh viễn thám trên, có một sự kết hợp giữa ảnh radar chủ động và vi ba thụ động và được gọi là vi ba kết hợp. Ảnh này có độ phân giải thấp hơn các ảnh radar chủ động và tập trung vào môi trường chuyên biệt như độ ẩm đất. Các ảnh này thường sử dụng băng tần L-band để quan trắc yếu tố chuyên biệt. Một ví dụ về ảnh này là SMAP, CIMR...

#### 1.1.4. Các dữ liệu địa không gian khác

Ngoài ảnh viễn thám, các loại dữ liệu địa không gian khác là các dữ liệu thực tế thông qua công tác đo đạc trực tiếp tại thực địa, điều tra khảo sát thực tế, tổng hợp từ các nguồn thống kê hay số hóa từ bản đồ giấy... Bên cạnh đó, các sản phẩm gián tiếp từ các dữ liệu này cũng được xem xét

#### 2. Địa hình và các sản phẩm từ địa hình

Dữ liệu địa hình là nền tảng quan trọng trong nghiên cứu lũ quét, cung cấp thông tin về cấu trúc bề mặt Trái Đất, bao gồm độ cao, độ dốc, hướng dốc, và các đặc điểm địa

hình như sông suối, thung lũng, hoặc đồi núi. Dữ liệu này thường được thu thập thông qua các phương pháp đo đạc thực địa sử dụng các thiết bị như máy toàn đạc điện tử, GPS độ chính xác cao, hoặc công nghệ LiDAR (không sử dụng ảnh viễn thám). Ngoài ra, dữ liệu địa hình có thể được số hóa từ các bản đồ địa hình giấy truyền thống, chẳng hạn như bản đồ tỷ lệ 1/10.000 hoặc 1/25.000, do các cơ quan đo đạc bản đồ cung cấp.

Các sản phẩm từ dữ liệu địa hình bao gồm mô hình địa hình số (Digital Elevation Model - DEM), bản đồ độ dốc, bản đồ hướng dốc, và bản đồ lưu vực nước, tất cả đều đặc biệt quan trọng trong nghiên cứu lũ quét. Chẳng hạn, bản đồ địa hình tỷ lệ 1/10.000 cung cấp chi tiết về các đường đồng mức, giúp xác định các khu vực trũng thấp dễ bị ngập lụt hoặc các sườn dốc có nguy cơ lũ quét cao. Bản đồ độ dốc, được tạo ra từ DEM, giúp xác định các khu vực có độ dốc lớn (>15-20 độ), nơi nước mưa chảy nhanh, làm tăng nguy cơ lũ quét. Bản đồ lưu vực nước hỗ trợ xác định các khu vực tập trung dòng chảy, đặc biệt là các thung lũng hẹp hoặc vùng hạ lưu.

Trong nghiên cứu lũ quét, dữ liệu địa hình được sử dụng để mô phỏng dòng chảy bề mặt và đánh giá nguy cơ. Ví dụ, kết hợp độ dốc và hướng dốc giúp xác định các khu vực dễ xảy ra lũ quét do nước mưa tập trung nhanh. Dữ liệu này thường được thu thập bởi các cơ quan chuyên môn như Tổng cục Địa chính hoặc các đơn vị khảo sát địa phương, đảm bảo độ chính xác cao để phục vụ các mô hình thủy văn như HEC-HMS hoặc SWAT. Các sản phẩm từ địa hình không ch

### 3. Sử dụng đất

Dữ liệu sử dụng đất cung cấp thông tin về cách các khu vực đất đai được khai thác, chẳng hạn như đất nông nghiệp, đất rừng, đất đô thị, đất công nghiệp, hoặc đất chưa sử dụng. Dữ liệu này được thu thập trực tiếp từ các cơ quan địa phương như Sở Tài nguyên và Môi trường, các đơn vị hành chính cấp tỉnh hoặc huyện, thông qua khảo sát thực địa, phỏng vấn cộng đồng, hoặc tổng hợp từ các báo cáo thống kê như báo cáo sử dụng đất hàng năm. Ngoài ra, dữ liệu sử dụng đất có thể được số hóa từ các bản đồ giấy hoặc tài liệu quy hoạch đất đai lưu trữ tại địa phương.

Trong nghiên cứu lũ quét, dữ liệu sử dụng đất giúp đánh giá tác động của các hoạt động con người đến nguy cơ lũ quét. Ví dụ, các khu vực đất rừng (rừng tự nhiên hoặc rừng trồng) thường có lớp phủ thực vật dày, giúp giảm tốc độ dòng chảy bề mặt và hạn chế nguy cơ lũ quét. Ngược lại, đất nông nghiệp hoặc đất trồng, đặc biệt là những khu vực bị canh tác không bền vững, dễ bị xói mòn và làm tăng nguy cơ lũ quét do thiếu thảm thực vật. Đất đô thị, với bề mặt bị bê tông hóa, làm tăng tốc độ dòng chảy, dẫn đến nguy cơ lũ quét cao hơn ở các khu vực hạ lưu.

Các sản phẩm từ dữ liệu sử dụng đất bao gồm bản đồ sử dụng đất, bản đồ thay đổi sử dụng đất theo thời gian, và các báo cáo phân tích xu hướng. Những sản phẩm này hỗ trợ xác định các khu vực cần bảo vệ hoặc cải tạo, chẳng hạn như trồng rừng để tăng độ

che phủ hoặc áp dụng các biện pháp canh tác bền vững để giảm xói mòn. Việc thu thập dữ liệu sử dụng đất đòi hỏi sự phối hợp chặt chẽ giữa các cơ quan địa phương và các nhóm khảo sát thực địa, đảm bảo thông tin phản ánh đúng thực trạng và được cập nhật thường xuyên để phục vụ nghiên cứu lũ quét.

#### 4. Loại đất, thổ nhưỡng

Dữ liệu loại đất cung cấp thông tin về đặc tính vật lý, hóa học, và sinh học của đất, bao gồm thành phần đất (đất sét, đất cát, đất phù sa), độ pH, độ phì nhiêu, và khả năng thấm nước. Dữ liệu này thường được thu thập bởi các cơ quan nghiên cứu địa chất, như Viện Địa chất Việt Nam, thông qua khảo sát thực địa, lấy mẫu đất, và phân tích trong phòng thí nghiệm. Ngoài ra, dữ liệu loại đất có thể được tổng hợp từ các nghiên cứu trước đây hoặc số hóa từ các bản đồ đất giấy, đảm bảo không sử dụng các nguồn viễn thám.

Trong nghiên cứu lũ quét, dữ liệu loại đất đóng vai trò quan trọng trong việc đánh giá khả năng thấm nước và nguy cơ xói mòn. Ví dụ, đất cát hoặc đất đá có khả năng thấm nước thấp, dẫn đến dòng chảy bề mặt nhanh, làm tăng nguy cơ lũ quét. Ngược lại, đất sét hoặc đất phù sa có thể giữ nước tốt hơn, nhưng nếu bị nén chặt do hoạt động nông nghiệp hoặc xây dựng, cũng làm tăng nguy cơ dòng chảy bề mặt. Bản đồ nguy cơ xói mòn, được tạo ra từ dữ liệu loại đất kết hợp với độ dốc, giúp xác định các khu vực dễ bị rửa trôi đất, góp phần làm tăng lượng phù sa trong dòng lũ quét.

Các sản phẩm từ dữ liệu loại đất bao gồm bản đồ phân loại đất, bản đồ độ phì nhiêu, và bản đồ nguy cơ xói mòn. Những sản phẩm này hỗ trợ để xuất các biện pháp giảm thiểu lũ quét, như trồng cây che phủ, xây dựng các công trình chống xói mòn (rãnh thoát nước, kè đá), hoặc cải tạo đất để tăng khả năng thấm nước. Việc thu thập dữ liệu loại đất đòi hỏi sự chuyên môn hóa cao, thường được thực hiện bởi các nhà địa chất học hoặc chuyên gia về khoa học đất, nhằm đảm bảo độ chính xác và tính ứng dụng trong nghiên cứu lũ quét.

### 1.3. Các nghiên cứu về phân vùng lũ quét

Phân vùng lũ quét đã được nhiều nhà nghiên cứu thực hiện nhằm tìm ra các khu vực chịu tác động bởi lũ quét, về cơ bản, phân vùng lũ quét được chia thành hai loại: phân vùng theo điểm và phân vùng theo lưu vực.

**Phân vùng lũ quét theo điểm** là lập một bản đồ không gian các vị trí chịu tác động bởi lũ quét dựa trên đặc điểm không gian của điểm đó. Thông thường, quá trình phân vùng theo điểm được xây dựng bằng phương pháp chồng chập đa nhân tố như: Phương pháp phân tích thứ bậc (AHP) (Romdani, R. P, et al., 2018; S. Talha, et al., 2019); phương pháp tỷ số tần suất (FR) (Chen Cao, et al., 2016); phương pháp trọng số dẫn chứng (WOE) (A Saleh, et al., 2022); phương pháp học máy (ML) (Costache, Romulus, et al., 2019); phương pháp học sâu (DL) (Zhao, Gang, et al., 2022)...

**Phân vùng lũ quét theo lưu vực** thường được tiếp cận dưới góc độ thủy văn, là quá trình xác định các yếu tố đặc trưng của dòng chảy lũ (từ mô phỏng mưa – dòng chảy) như lưu lượng đỉnh lũ, vận tốc dòng chảy, thời gian tập trung dòng chảy... để xác định lũ quét tại cửa ra của một lưu vực. Khi dòng chảy thuộc lưu vực có khả năng là lũ quét (vượt ngưỡng quy định nào đó) thì lưu vực đó được cảnh báo có nguy cơ lũ quét. Một trong những phương pháp điển hình của loại hình phân vùng lũ quét theo lưu vực là phương pháp FFG (Flash flood guidance) trong hệ thống FFGS (Flash flood guidance system) (Carpenter, T.M., et al., 1999).

### 1.3.1 Các nghiên cứu trên thế giới

Mặc dù các nghiên cứu về lũ quét có từ rất sớm, các nghiên cứu về phân vùng lũ quét lại khởi đầu muộn hơn. Trước năm 2000, hầu hết các nghiên cứu về lũ quét tập trung vào đánh giá các trận lũ quét (Forest Service, 1931; Kuz'min, K. K., 1974; Hales, John E., 1978) và tính toán thủy văn cho các vị trí có khả năng xảy ra lũ quét (NOAA, 1979; Georgakakos, Konstantine P., 1986). Đến năm 1992, Sweeney viết hướng dẫn hiện đại hóa lũ quét lưu vực (Modernized Areal Flash Flood Guidance) và đăng tải trên bản ghi nhớ hướng dẫn kỹ thuật NWS HYDRO 44 (Sweeney, Timothy L., 1992). Trong đó mô tả đầy đủ các khái niệm cũng như hướng dẫn các bước tính toán xác định ngưỡng mưa sinh lũ. Khái niệm FFG (Flash Flood Guidance) cũng được hình thành (mặc dù lưu lượng tràn bờ đã được Emmett nghiên cứu trước đó gần 20 năm (William W. Emmett, 1975)), mô tả về một lượng mưa bình quân đủ để sinh ra lũ lụt trên các sông suối nhỏ. Cơ sở của FFG là tính toán các giá trị ngưỡng mưa hiệu quả trong một khoảng thời gian cần thiết để gây ra lũ, lượng mưa hiệu quả này được hiểu là lượng mưa dư sau tốn thát do thẩm thấu, giam giữ và bốc hơi trên lưu vực để hình thành dòng chảy tràn trên bề mặt. Sau đó, các ngưỡng FFG được Trung tâm dự báo xác định cụ thể cho từng khu vực.

Một số phân tích khác sử dụng bản đồ lượng mưa tần suất để dự báo khả năng xuất hiện lũ quét (Chang, Tiao J. & Sun, Hong Y., 1997) với giả thiết lượng mưa hàng ngày lớn hơn giá trị xác định. Phương pháp này giúp đưa ra khu vực có tiềm năng lũ quét hoàn toàn dựa vào lượng mưa, tuy nhiên chưa xét đến các yếu tố bề mặt. Ngoài ra, sử dụng phân bố mưa có thể gây ra cảnh báo không do vị trí xuất hiện lũ quét có thể nằm ở hạ lưu khu vực mưa lớn – nơi không có lượng mưa phản ánh đúng thực trạng lũ quét.

Năm 1997, Sharada trình bày một phương pháp xác định ảnh hưởng của nguy cơ lũ quét tới tuyến đường sắt tại quận Warangal, Ấn Độ (Sharada, D., et al., 1997) bằng việc kết hợp nhiều yếu tố địa chất không gian bao gồm (1) Mặt nước; (2) đặc điểm bờ kè; (3) số lượng hồ trong khu vực; (4) mật độ sông/suối; (5) đặc điểm thoát nước; (6) khoảng cách tới đường sắt; (7) sử dụng đất; (8) độ lún; và (9) độ dốc giữa suối và đường sắt. Phương pháp này gần giống với các nghiên cứu chồng chập đa nhân tố hiện nay và đưa

ra được bản đồ tác động được phân cấp nguy cơ (từ rất thấp đến rất cao) bằng việc sử dụng công cụ GIS.

Từ năm 2000 trở lại đây, các nghiên cứu phân vùng lũ quét đã trở nên dần phổ biến dưới sự hỗ trợ của các công cụ GIS và máy tính hiệu năng cao và đi về hai hướng: (1) phân vùng lũ quét dựa trên nguyên lý thủy văn và các yếu tố liên quan đến thủy văn (bao gồm mưa, kết quả tính toán thủy văn, thủy lực) và (2) phân vùng lũ quét dựa trên các tham số địa không gian. Các nghiên cứu dựa trên dữ liệu địa không gian được phân tích, đánh giá trên hàng loạt các yếu tố được raster hóa cho khu vực nghiên cứu, từ đó phân vùng lũ quét từ nhiều tham số. Trong nghiên cứu này, tác giả phân loại nghiên cứu phân vùng lũ quét theo dữ liệu địa không gian làm 2 loại: (1) sử dụng phương pháp trọng số và (2) sử dụng phương pháp trí tuệ nhân tạo.

#### *1.3.1.1 Phân vùng lũ quét dựa trên lượng mưa*

Mưa lớn là một trong những nguyên nhân chính trong quá trình hình thành lũ quét tự nhiên, do vậy, yếu tố lượng mưa được xét đến trong các nghiên cứu về lũ nói chung và lũ quét nói riêng là một hướng đi đúng đắn. Mặc dù vậy, dự báo được lượng mưa một cách chính xác không phải là một vấn đề dễ dàng trong các nghiên cứu.

Borga và cộng sự đã có một rà soát các nghiên cứu ở châu Âu nói riêng và công bố quốc tế nói chung đến năm 2014 về hệ thống cảnh báo lũ quét và lũ bùn đá về các khía cạnh: kỹ thuật sử dụng, khoảng trống nghiên cứu và đề xuất các hướng nghiên cứu tiếp theo (Marco Borga, et al., 2014). Trong đó, đề xuất tiếp theo bao gồm: xây dựng hệ thống giám sát và dự báo và xác định ngưỡng mưa cho lũ quét. Ghomash và cộng sự đã nghiên cứu tác động của diễn biến lượng mưa đến quá trình hình thành lũ quét cho các khu vực miền núi bán khô cằn ở lưu vực Kan (Iran) (Shahin Khosh Bin Ghomash, et al., 2022). Các hình thái mưa khác nhau cho ra kết quả khác nhau về đỉnh lũ, đường quá trình lũ và các yếu tố thủy văn.

Alfieri và cộng sự đã trình bày hướng dẫn dự báo lũ quét dựa trên ngưỡng về lượng mưa trong cuốn sổ tay dự báo khí tượng thủy văn vào năm 2015 (Lorenzo Alfieri, et al., 2015). Alfieri cho biết, chỉ sử dụng dữ liệu mưa có thể ước tính được thiệt hại do lũ lụt gây ra. Một trong những phương pháp phổ biến là phương pháp chỉ dẫn lũ quét (FFG) trong hệ thống chỉ dẫn lũ quét (FFGS) đang được sử dụng rộng rãi trên toàn thế giới.

Một nghiên cứu khác của Alfieri năm 2011 đã đề xuất hệ thống cảnh báo sớm lũ quét dựa trên các sản phẩm dự báo khí tượng như dự báo thời tiết số trị (NWP), kết hợp NWP-Radar và truyền sóng radar cho các khu vực miền núi ở Catalonia (thuộc Tây Ban Nha) (L. Alfieri, et al., 2011). Nhược điểm của phương pháp này là bỏ qua các điều kiện ban đầu của bề mặt và hiệu chỉnh của lượng mưa.

Phương pháp phổ biến nhất là xây dựng ngưỡng mưa dựa trên quan hệ cường độ - thời gian (I-D) được Caine đề xuất năm 1980 (Nel Caine, 1980) với công thức  $I = \alpha D^\beta$

(công thức kinh nghiệm). Trong đó,  $\alpha$  và  $\beta$  là các hệ số và  $D$  là thời gian mưa (theo giờ),  $I$  là cường độ mưa. Trên cơ sở thu thập các trận mưa trong quá khứ, Caine đã tìm ra được ngưỡng mưa  $I = 14.82D^{-0.39}$  (với thời đoạn mưa nhỏ hơn 500 giờ) là ngưỡng mưa có khả năng gây ra lũ bùn đá và sạt lở đất.

Năm 2007, tiếp tục nghiên cứu dựa theo Caine, Guzzetti và cộng sự (Fausto Guzzetti, et al., 2007) đã tìm ra được ngưỡng mưa  $I = 2.28D^{-0.2}$  cho các trận mưa có thời đoạn dưới 48 giờ và  $I = 0.48D^{-0.11}$  cho các trận mưa từ 48 giờ đến dưới 1000 giờ. Nghiên cứu này sử dụng cơ sở dữ liệu toàn cầu từ 2.626 trận mưa và 19.800 trạm quan trắc mưa trên toàn thế giới. Con số này thấp hơn khá đáng kể so với nghiên cứu của Caine và làm tăng ngưỡng an toàn về mưa.

Bezak và cộng sự đã xây dựng đặc điểm các đợt mưa cực lớn gây ra sạt lở đất và lũ quét ở Slovenia trong 25 năm (Nejc Bezak, et al., 2016). Kết quả phân tích ở các trạm cho thấy công thức của Caine phù hợp nhất trong việc xây dựng ngưỡng mưa cảnh báo.

Turkington và cộng sự đã đưa ra ngưỡng cảnh báo về lượng mưa cho lũ bùn đá và lũ quét ở khu vực miền núi phía Nam nước Pháp (thung lũng Ubaye) là lượng mưa lớn hơn 20mm/ngày hoặc 22mm/4 ngày (T. Turkington, et al., 2014). Mặc dù trong nghiên cứu có nhiều ngưỡng phân loại khác nhau độc lập (cho các yếu tố khí tượng khác), đây là một ngưỡng mưa quá an toàn để đưa ra cảnh báo và có khả năng cảnh báo không cao trong khi đáp ứng đầy đủ về tiêu chí phân loại cho các trận mưa sinh lũ ở khu vực nghiên cứu.

Ramos Filho và cộng sự đã trình bày một phương pháp cải tiến trong việc sử dụng ngưỡng cường độ mưa cực đại để đánh giá và cảnh báo lũ quét ở bang São Paulo (Brazil) (Geraldo Moura Ramos Filho, et al., 2020) bằng mô hình số mũ có mối quan hệ với cường độ mưa và chỉ số lượng mưa kỳ trước (API) bằng công thức  $I = ae^{b \cdot API} + c$ . Trong đó:  $a$ ,  $b$ ,  $c$  là các hằng số cần xác định. Nghiên cứu cũng chỉ ra giá trị của  $a$  nằm trong khoảng từ cường độ mưa tối thiểu đến ba lần cường độ mưa tối đa (của các trận lũ quét đã xảy ra), giá trị  $b$  thay đổi trong khoảng từ  $-1.00 \div -0.01$  và giá trị  $c$  nằm trong khoảng từ cường độ mưa tối thiểu đến cường độ mưa trung bình (của các trận lũ quét đã xảy ra). Kết quả nghiên cứu cho thấy xác suất phát hiện các trận lũ quét trên ngưỡng trên tăng trung bình 14%.

Zhai và cộng sự đã nghiên cứu ngưỡng mưa sinh lũ quét tại ba khu vực miền núi ở miền Nam Trung Quốc (Zhong, Balisi và Yu) dựa trên mô hình thủy văn lũ quét Trung Quốc (Xiaoyan Zhai, et al., 2018). Quá trình xác định bao gồm mực nước và lưu lượng sinh lũ quét tại các khu vực, từ đó tính ngược lượng mưa sinh lũ quét. Kết quả nghiên cứu chỉ ra: (1) tại Zhong, lưu lượng sinh lũ quét là  $356,2\text{m}^3/\text{s}$  và ngưỡng mưa dao động từ  $93 \div 344\text{mm/trận}$ ; (2) tại Balisi, lưu lượng sinh lũ quét là  $544,69\text{m}^3/\text{s}$  và ngưỡng mưa

dao động từ 77÷246mm/trận; (3) tại Yu, lưu lượng sinh lũ quét là 335,2m<sup>3</sup>/s và ngưỡng mưa dao động từ 111÷420mm/trận.

Yuan và cộng sự cho ra kết quả về ngưỡng mưa tương ứng với điều kiện độ ẩm kỳ trước (ASMC) và sự phân bố mưa (Wenlin Yuan, et al., 2021). Lượng mưa được tính toán cho các độ ẩm kỳ trước là 0.2 (khô); 0.5 (trung bình); 0.8 (ướt) và ngưỡng mưa tương ứng từ 95÷140mm; 79÷125mm; và 73÷117mm/6 giờ cho lưu vực Xinxian, huyện Tín Dương, tỉnh Hà Nam, Trung Quốc.

### 1.3.1.2 Phương pháp phân vùng lũ quét dựa vào thủy văn, thủy lực

Saber và cộng sự đã sử dụng lượng mưa vệ tinh GSMP để lập mô hình dự báo lũ quét cho lưu vực sông Karpuz (Thổ Nhĩ Kỳ) (Mohamed Saber & Koray Yilmaz, 2018), mô hình thủy văn Hydro-BEAM được sử dụng để xác định dòng chảy dựa trên dữ liệu GSMP đã được hiệu chỉnh. Mặc dù đã có sự hiệu chỉnh độ lệch nhưng không có sự đảm bảo rằng dữ liệu mưa vệ tinh có đủ tin cậy để dự báo dòng chảy lũ. Thậm chí, Westerberg và cộng sự (I. K. Westerberg, et al., 2011) trong một nghiên cứu khác đã chỉ ra rằng việc hiệu chỉnh lại lượng mưa vệ tinh không phải lúc nào cũng là một phương pháp đáng tin cậy vì các lỗi vệ tinh có thể dẫn đến những suy luận sai lệch trong quá trình hiệu chỉnh mô hình thủy văn cho các trận lũ tiếp theo.

Silvestro và cộng sự đã sử dụng mô hình thủy văn bán phân bố DriFt kết hợp với lượng mưa thực đo và lượng mưa dự báo trước 2 giờ để dự báo nguy cơ lũ quét cho lưu vực sông Entella ở phía đông vùng Liguria, Ý (F. Silvestro, et al., 2015). Số liệu mưa dự báo được lấy từ các nguồn ECMWF; COSMO-LAMI; BOLAM; MOLOCH. Kết quả tính toán cho thấy, sai lệch % đỉnh lũ dao động từ -15÷43% với khoảng thời gian dự báo trước từ 1:40p ÷ 4:40p cho 7 trận lũ quét đã xảy ra trong quá khứ. Nghiên cứu không đánh giá khả năng ứng dụng trong tương lai và có đưa ra một số khuyến nghị khi áp dụng mô hình.

Việc sử dụng các tham số thủy lực để tính toán nguy cơ lũ quét (mực nước, vận tốc...) đã được áp dụng, tuy nhiên về bản chất là các mô hình toán một chiều trên sông. Để có thể nói là mô phỏng lũ quét (với đặc điểm diễn biến nhanh và không chậm biến đổi đều) là chưa phù hợp, tuy nhiên, hầu hết các mô hình này sử dụng chỉ với mục đích xác định dòng chảy lũ trên sông suối, khi đạt ngưỡng về mực nước, vận tốc... sẽ được xem xét cảnh báo nguy cơ lũ nói chung và lũ quét nói riêng.

Ngưỡng sinh dòng chảy được sử dụng để đánh giá dòng chảy trên sông khi vượt qua ngưỡng càn cảnh báo. Đây là kết quả của việc sử dụng mô hình thủy lực để tính toán các yếu tố về mực nước, vận tốc... từ mưa và các mô hình thủy văn. Các kết quả tính toán sẽ được so sánh trực tiếp với giá trị ngưỡng dòng chảy. Ngưỡng dòng chảy có thể là mực nước lớn nhất đã xảy ra tương ứng với trận lũ quét trong quá khứ, cũng có

thể là một giá trị tương ứng với tần suất khi có quan trắc trong nhiều năm về mực nước (phương pháp thống kê) (Seann Reed, et al., 2007).

Trong hệ thống FFG, việc xác định lưu lượng tràn bờ được sử dụng bằng phương pháp thủy lực mặt cắt, từ đó đưa ra ngưỡng tràn bờ tại các vị trí cửa ra của lưu vực tính toán. Một số các nghiên cứu khác có thể sử dụng ngưỡng tràn bờ bằng công thức kinh nghiệm (Jonathan D Phillips, 2002).

#### 1.3.1.3 Phân vùng lũ quét theo trọng số dữ liệu địa không gian

Phần lớn các kết quả nghiên cứu về lũ quét được công bố có liên quan đến đánh giá tính nhạy cảm với lũ quét. Mặc dù là phương pháp đơn giản, các nghiên cứu này lại có lượng dữ liệu lớn và đa dạng. Các loại dữ liệu được sử dụng thông thường bao gồm: (1) Dữ liệu liên quan đến yếu tố địa hình như: độ dốc, độ cong, cao độ...; (2) Dữ liệu liên quan đến yếu tố thủy văn như: khoảng cách đến sông suối, mật độ sông suối, chiều dài dòng chảy...; (3) Dữ liệu liên quan đến thảm phủ bề mặt như: sử dụng đất, chỉ số NDVI...; (4) dữ liệu liên quan đến loại đất như: nhóm đất thủy văn, nhóm đất, kết cấu đất...

Để kết hợp được rất nhiều các yếu tố, các nhà nghiên cứu sử dụng phương pháp chồng chập đa nhân tố để đưa ra bản đồ cuối cùng về nguy cơ. Mỗi loại dữ liệu được xác định cho một hệ số tác động đến kết quả hình thành lũ quét. Các phương pháp sau đây là phổ biến trong việc đánh giá tính nhạy cảm với lũ quét:

##### 1. Phương pháp phân tích thứ bậc

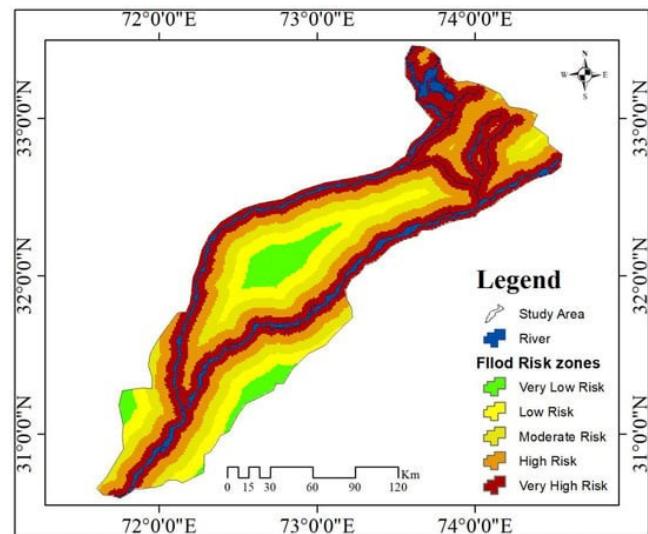
Phương pháp phân tích thứ bậc (AHP) được biết đến là một phương pháp đưa ra phương án quyết định dựa trên đa tiêu chí. Phương pháp này được phát triển bởi Thomas L. Saaty năm 1977 (Thomas L Saaty, 1977). Bài báo này có số lượt trích dẫn lên tới hơn 5.000 lượt theo thống kê của Crossref (tính đến tháng 9/2023).

Nội dung của phương pháp AHP bao gồm 5 bước: (1) Xác định các yếu tố liên quan và phương án quyết định; (2) Xây dựng ma trận tương quan để đánh giá mức độ quan trọng của các yếu tố liên quan. Trong ma trận này, mỗi ô thể hiện mức độ quan trọng của yếu tố dòng (row) đối với yếu tố cột (column); (3) Xác định trọng số của các yếu tố thể hiện mức độ quan trọng của các yếu tố đó. Trọng số của các yếu tố được tính toán từ ma trận tương quan; (4) Xây dựng ma trận quyết định để đánh giá mức độ ưu tiên của các phương án quyết định đối với các yếu tố liên quan. Trong ma trận này, mỗi ô thể hiện mức độ ưu tiên của phương án dòng (row) đối với yếu tố cột (column); và (5) Tính toán kết quả là thứ tự ưu tiên của các phương án quyết định. Thứ tự ưu tiên này được tính toán từ ma trận quyết định và trọng số của các yếu tố.

Với ưu điểm là một phương pháp tổng quát, có thể áp dụng cho nhiều loại vấn đề khác nhau, phương pháp này đã được áp dụng trong việc xác định nguy cơ lũ quét với

các yếu tố đã được trình bày ở trên. Do sử dụng dữ liệu lớn với đa phần là dữ liệu về không gian, phương pháp này đang dần phổ biến trong những năm gần đây.

Aquil Tariq và cộng sự đã sử dụng phương pháp AHP trong việc xác định nguy cơ lũ quét cho 2 lưu vực sông Jhelum và Chenab ở Ấn Độ (Aquil Tariq, et al., 2022) dựa trên 8 yếu tố bao gồm (1) độ dốc; (2) độ cao; (3) mật độ sông suối; (4) Lượng mưa; (5) thảm phủ; (6) khoảng cách đến sông suối; (7) địa chất; và (8) loại đất. Kết quả xây dựng bản đồ và phân tích trọng số được thể hiện ở hình bên.



Bảng 1-2. Bản đồ ma trận và trọng số quyết định theo AHP

Các yếu tố	SL	E	DD	R	LULC	DS	G	So	SFWV
Độ dốc (SL)	1.00	0.20	5.00	4.00	2.00	1.00	4.00	0.30	0.0152
Độ cao (E)	1.00	0.23	0.90	5.00	1.00	4.00	2.00	0.30	0.0254
Mật độ sông suối (DD)	2.00	5.00	4.00	0.20	1.00	5.00	3.00	1.00	0.2569
Lượng mưa (R)	0.25	0.12	1.00	0.23	0.50	1.00	1.00	0.30	0.0452
Thảm phủ (LULC)	0.10	0.12	0.17	0.45	0.50	0.13	0.25	1.00	0.1400
KC đến sông suối (DS)	1.00	1.00	3.00	5.00	3.00	5.00	4.00	1.00	0.2592
Địa chất (G)	0.34	0.23	0.25	1.00	0.25	0.18	2.00	0.35	0.1190
Loại đất (So)	0.25	4.00	0.25	1.00	2.00	4.00	3.00	1.00	0.1800

Yếu tố lượng mưa trong nghiên cứu này sử dụng là sự chênh lệch lượng mưa so với lượng mưa trung bình nhiều năm tại các trạm (cụ thể trong nghiên cứu này là lượng mưa trung bình từ năm 2010 đến 2015 của 5 trạm quan trắc trong khu vực nghiên cứu). Giá trị phương sai của lượng mưa sẽ được tính toán theo các trạm và được nội suy thành raster mưa dựa trên thuật toán trọng số nghịch đảo IDW.

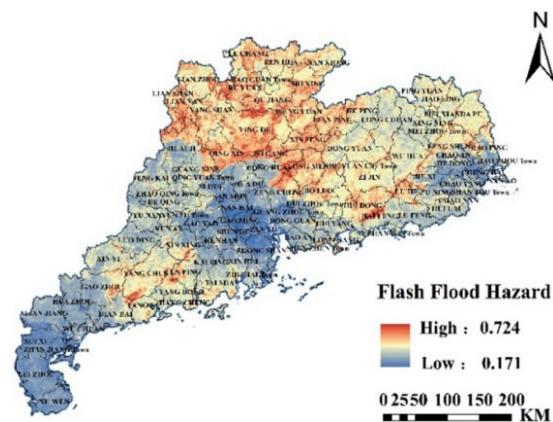
Kairong Lin và cộng sự cũng sử dụng phương pháp AHP để xác định nguy cơ lũ quét cho tỉnh Quảng Đông ở Trung Quốc (Kairong Lin, et al., 2020), Lin chỉ sử dụng 4 yếu tố nguy cơ đầu vào bao gồm (1) cao độ địa hình; (2) độ dốc; (3) lượng mưa; và (4) mật độ sông suối. Kết quả được thể hiện như hình bên. Độ dốc và lượng mưa được xét với trọng số như nhau (khoảng 34%), trong khi đó, cao độ địa hình là gần 20% và mật độ sông suối chiếm gần 12%.

Một biến thể khác của AHP là FAHP (Fuzzy Analytic Hierarchy Process) (S. Talha, et al., 2019; Mohammed Sadek & Xuxiang Li, 2019) khi không phân loại các yếu tố thành 9 ngưỡng (từ 1 đến 9) như phương pháp gốc mà thay vì đó là quy đổi về phạm vi từ 0 đến 1. Các tác giả cho rằng FAHP có lợi thế hơn vì nó có thể đáp ứng được sự không chắc chắn trong dữ liệu và phân tích và phù hợp để xác định khu vực dễ bị lũ lụt ở các thành phố, đặc biệt là do sự trùng khớp với các khu vực bị tàn phá nhiều nhất được xác định bằng cách phát hiện sự thay đổi giữa hai hình ảnh Sentinel-2.

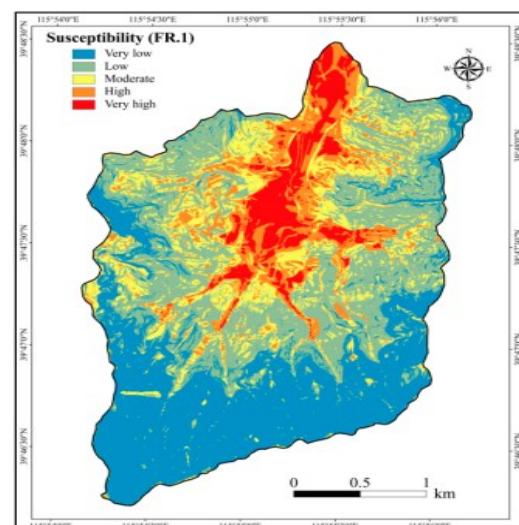
## 2. Phương pháp tỷ lệ tàn suất

Phương pháp tỷ lệ tàn suất - Frequency Ratio (FR) là một trong những phương pháp phổ biến đo lường tàn suất xuất hiện của lũ quét trong một khu vực nhất định. Các khu vực có chỉ số FR cao hơn thể hiện nguy cơ lũ quét cao hơn. Chỉ số FR là thương của tỷ lệ xuất hiện trên tỷ lệ diện tích. Tương tự như AHP, phương pháp này sử dụng các yếu tố đa dạng về địa không gian.

Chen Cao đã sử dụng phương pháp FR để xây dựng bản đồ nguy cơ lũ quét cho khu vực Xiqu Gully (thuộc Bắc Kinh – Trung Quốc) dựa trên 85 điểm đã xảy ra lũ quét (Chen Cao, et al., 2016) và 10 yếu tố bao gồm: độ cao; độ dốc; độ cong; sử dụng đất; địa chất; kết cấu đất; khu vực có nguy cơ sạt lún; chỉ số năng lượng dòng chảy; chỉ số ám địa hình và mưa lớn ngắn hạn. Kết quả xây dựng bản đồ cho thấy

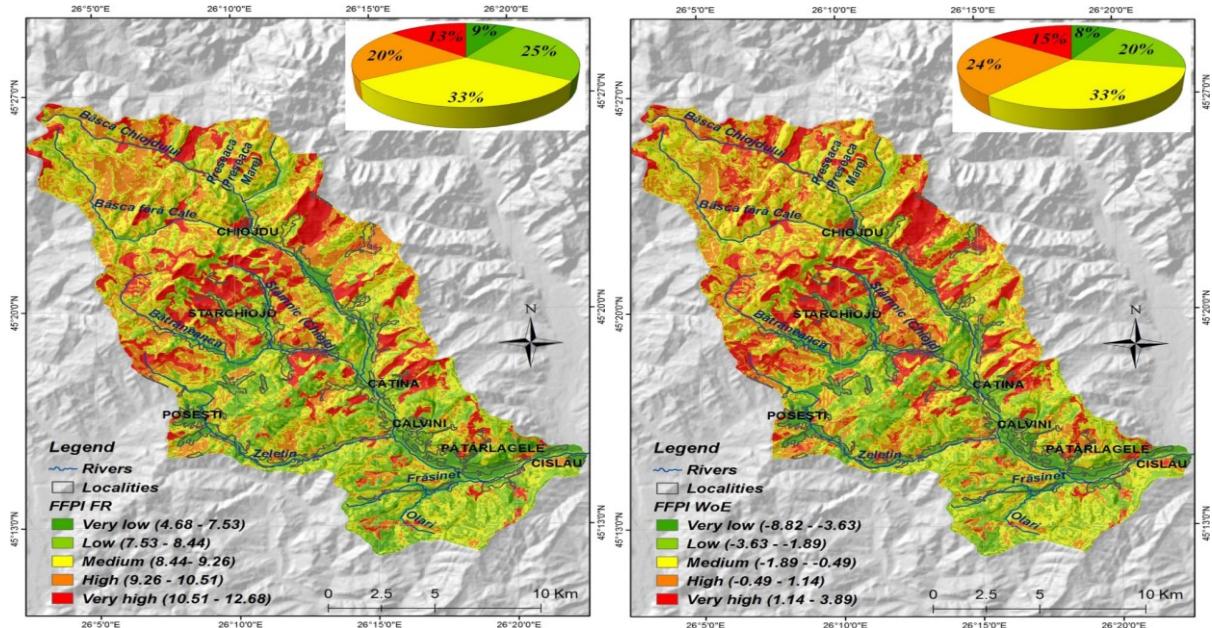


Hình 1-2. Bản đồ nguy cơ lũ quét tỉnh Quảng Đông – Trung Quốc



các vị trí có nguy cơ rất cao về lũ quét nằm ở khu vực trũng thấp của địa hình.

Ở Romania, Romulus Costache đã xây dựng bản đồ tiềm năng lũ quét bằng phương pháp FR kết hợp với phương pháp trọng số dẫn chứng (WOE) với 9 yếu tố cho lưu vực sông Bâscă Chiojdului có diện tích khoảng 340km<sup>2</sup> (Romulus Costache & Liliana Zaharia, 2017).



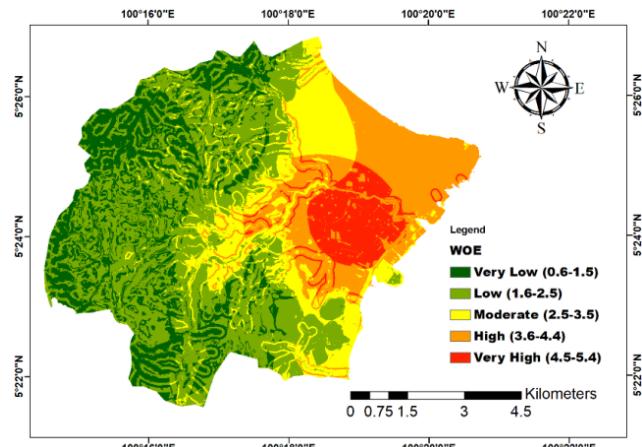
Hình 1-4. Kết quả xây dựng bản đồ tiềm năng lũ quét lưu vực Bâscă Chiojdului bằng phương pháp FR (bên trái) và WOE (bên phải)

### 3. Phương pháp trọng số dẫn chứng

Phương pháp trọng số dẫn chứng - Weights Of Evidence (WOE) được sử dụng để đo lường mức độ liên quan của một biến độc lập với một biến phụ thuộc. WOE được tính bằng cách sử dụng hàm logistic để chuyển đổi các giá trị của biến độc lập thành các giá trị số.  $WOE = \ln(p/(1 - p))$ , trong đó p là xác suất biến phụ thuộc có giá trị là 1 và  $(1-p)$  là xác suất biến phụ thuộc có giá trị là 0. Để tính toán WOE, cần chia dữ liệu thành 2 nhóm, một nhóm có biến phụ thuộc có giá trị là 1 và một nhóm có biến phụ thuộc có giá trị là 0. Sau đó, tính xác suất của biến phụ thuộc có giá trị là 1 cho từng nhóm. Cuối cùng, sử dụng công thức trên để tính WOE cho từng giá trị của biến độc lập.

Hình 1-3. Bản đồ nguy cơ lũ quét khu vực Xiqu Gully

A Saleh đã sử dụng phương pháp WOE để xây dựng bản đồ nhạy cảm với lũ quét cho lưu vực Sungai Pinang ở Malaysia dựa trên 6 yếu tố bao gồm: độ cao, độ dốc, lượng mưa, độ che phủ đất, khoảng cách từ sông và thạch học (A Saleh, et al., 2022). Kết quả xây dựng bản đồ cho độ tin cậy lên tới 0,839.



Hình 1-5. Bản đồ nhạy cảm lũ quét lưu vực Sungai Pinang

#### 1.3.1.4 Phân vùng lũ quét ứng dụng trí tuệ nhân tạo

Trí tuệ nhân tạo được ứng dụng rộng rãi trong phân vùng lũ quét những năm gần đây. Nghiên cứu này phân tích hơn 300 nghiên cứu về lũ quét trong những năm gần đây về vấn đề xây dựng các bản đồ nhạy cảm và bản đồ nguy cơ lũ quét trên cơ sở dữ liệu của CrossRef. Bảng sau đây tổng hợp một số mô hình trí tuệ nhân tạo (bao gồm học máy và học sâu) thường được sử dụng trong phân vùng lũ quét.

Bảng 1-3. Danh sách các mô hình phổ biến được sử dụng trong phân vùng lũ quét

TT	Tên	Đặc điểm
<b>I Mô hình học máy (machine learning)</b>		
1	Hồi quy tuyến tính (Linear Regression)	<b>Đặc điểm:</b> Mô hình hóa mối quan hệ tuyến tính giữa biến đầu vào và đầu ra. <b>Ưu điểm:</b> Đơn giản, dễ hiểu, nhanh chóng. <b>Nhược điểm:</b> Chỉ phù hợp với quan hệ tuyến tính, nhạy cảm với các điểm đột biến. <b>Ứng dụng trong lũ quét:</b> Dự đoán mức nước lũ dựa trên lượng mưa.
2	Hồi quy logistic (Logistic Regression)	<b>Đặc điểm:</b> Dùng cho phân loại nhị phân, ước tính xác suất của một sự kiện. <b>Ưu điểm:</b> Hiệu quả với dữ liệu phân loại, dễ giải thích. <b>Nhược điểm:</b> Giả định quan hệ tuyến tính giữa biến đầu vào và log-odds của đầu ra. <b>Ứng dụng trong lũ quét:</b> Dự đoán khả năng xảy ra lũ quét (có/không).
3	Cây quyết định (Decision Trees)	<b>Đặc điểm:</b> Mô hình dựa trên quy tắc quyết định dạng cây. <b>Ưu điểm:</b> Dễ hiểu và diễn giải, xử lý tốt dữ liệu categorical và số. <b>Nhược điểm:</b> Có thể overfitting, không ổn định với thay đổi nhỏ trong dữ liệu.

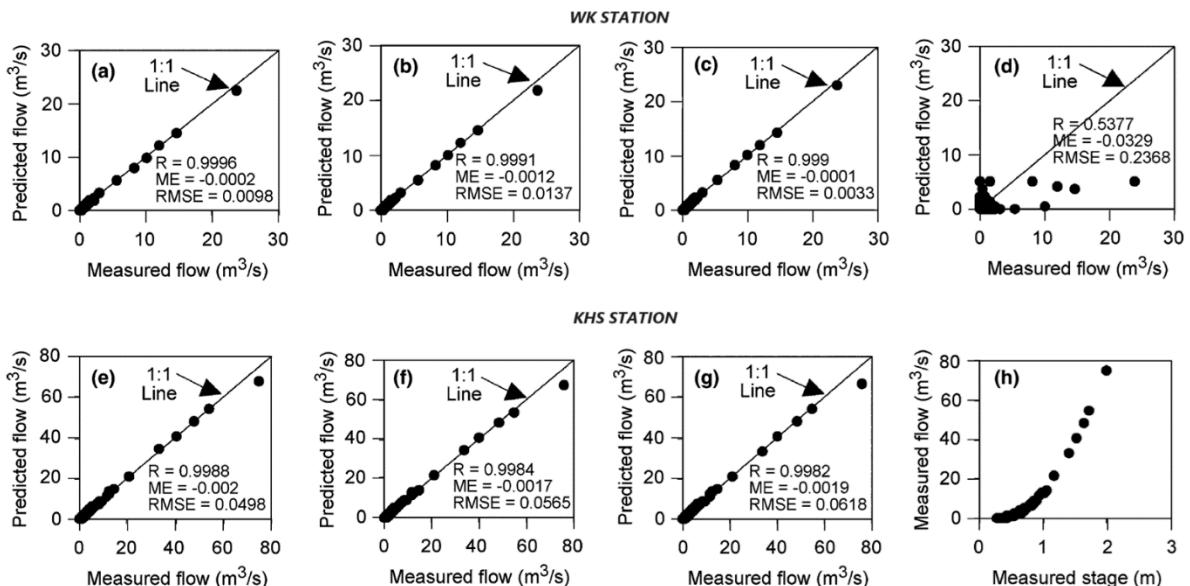
TT	Tên	Đặc điểm
		<b>Ứng dụng trong lũ quét:</b> Phân loại mức độ nguy hiểm của lũ quét.
4	Máy véc-tơ hỗ trợ (Support Vector Machines - SVM)	<b>Đặc điểm:</b> Tìm siêu phẳng tối ưu để phân loại dữ liệu. <b>Ưu điểm:</b> Hiệu quả trong không gian nhiều chiều, linh hoạt với nhiều kernel. <b>Nhược điểm:</b> Khó chọn kernel phù hợp, kém hiệu quả với dữ liệu lớn. <b>Ứng dụng trong lũ quét:</b> Phân loại các khu vực có nguy cơ lũ quét.
5	Mô hình phân lớp Naive Bayes	<b>Đặc điểm:</b> Dựa trên định lý Bayes với giả định độc lập giữa các đặc trưng. <b>Ưu điểm:</b> Đơn giản, nhanh, hiệu quả với dữ liệu văn bản và phân loại. <b>Nhược điểm:</b> Giả định độc lập giữa các đặc trưng thường không đúng trong thực tế. <b>Ứng dụng trong lũ quét:</b> Phân loại các loại cảnh báo lũ quét dựa trên nhiều yếu tố.
6	K láng giềng gần nhất (K-Nearest Neighbors KNN)	<b>Đặc điểm:</b> Phân loại dựa trên khoảng cách với K điểm dữ liệu gần nhất. <b>Ưu điểm:</b> Đơn giản, hiệu quả với dữ liệu có ranh giới quyết định phức tạp. <b>Nhược điểm:</b> Chậm với dữ liệu lớn, nhạy cảm với tỷ lệ các đặc trưng. <b>Ứng dụng trong lũ quét:</b> Dự đoán mức độ lũ quét dựa trên các sự kiện tương tự trong quá khứ.
7	Phân cụm K-means (K-Means Clustering)	<b>Đặc điểm:</b> Phân nhóm dữ liệu thành K cụm dựa trên khoảng cách. <b>Ưu điểm:</b> Đơn giản, hiệu quả với dữ liệu lớn, dễ hiểu. <b>Nhược điểm:</b> Cần xác định số cụm trước, nhạy cảm với outliers và điểm khởi tạo. <b>Ứng dụng trong lũ quét:</b> Phân nhóm các khu vực có đặc điểm lũ quét tương tự.
<b>II Mô hình học sâu (deep learning)</b>		
1	Mạng nơ-ron tích chập (Convolutional Neural Networks CNN)	<b>Đặc điểm:</b> Được thiết kế để xử lý dữ liệu có cấu trúc lưới, đặc biệt là hình ảnh. <b>Cấu trúc:</b> Gồm các lớp tích chập, lớp gộp (pooling) và lớp kết nối đầy đủ. <b>Ưu điểm:</b> <ul style="list-style-type: none"> <li>- Hiệu quả trong việc trích xuất đặc trưng từ dữ liệu không gian.</li> <li>- Giảm số lượng tham số cần học so với mạng nơ-ron thông thường.</li> <li>- Có khả năng học tự động các đặc trưng phức tạp.</li> </ul> <b>Nhược điểm:</b> <ul style="list-style-type: none"> <li>- Cần lượng dữ liệu lớn để huấn luyện.</li> <li>- Tính toán nặng, đòi hỏi phần cứng mạnh.</li> </ul>

TT	Tên	Đặc điểm
		<p><b>Ứng dụng trong nghiên cứu lũ quét:</b> Phân tích ảnh vệ tinh hoặc radar để đánh giá nguy cơ lũ quét.</p>
2	Mạng nơ-ron truy hồi (Recurrent Neural Networks - RNN)	<p><b>Đặc điểm:</b> Được thiết kế để xử lý dữ liệu chuỗi và dữ liệu thời gian.</p> <p><b>Cấu trúc:</b> Có các kết nối truy hồi, cho phép thông tin được truyền từ bước thời gian trước đến bước hiện tại.</p> <p><b>Ưu điểm:</b></p> <ul style="list-style-type: none"> <li>- Hiệu quả trong việc xử lý dữ liệu chuỗi thời gian.</li> <li>- Có khả năng học các phụ thuộc thời gian.</li> </ul> <p><b>Nhược điểm:</b></p> <ul style="list-style-type: none"> <li>- Khó huấn luyện do vấn đề gradient biến mất/bùng nổ.</li> <li>- Khó xử lý các phụ thuộc dài hạn.</li> </ul> <p><b>Ứng dụng trong nghiên cứu lũ quét:</b> Dự báo lũ quét dựa trên dữ liệu thời gian thực về lượng mưa, mực nước sông.</p>
3	Bộ nhớ dài-ngắn hạn (Long Short-Term Memory)	<p><b>Đặc điểm:</b> Một loại đặc biệt của RNN, được thiết kế để khắc phục vấn đề của RNN thông thường.</p> <p><b>Cấu trúc:</b> Sử dụng các cổng (gates) để kiểm soát luồng thông tin, bao gồm cổng quên, cổng đầu vào và cổng đầu ra.</p> <p><b>Ưu điểm:</b></p> <ul style="list-style-type: none"> <li>- Có khả năng học các phụ thuộc dài hạn tốt hơn RNN thông thường.</li> <li>- Giải quyết vấn đề gradient biến mất.</li> <li>- Hiệu quả trong việc xử lý chuỗi dài.</li> </ul> <p><b>Nhược điểm:</b></p> <ul style="list-style-type: none"> <li>- Phức tạp hơn RNN, có nhiều tham số hơn cần học.</li> <li>- Vẫn có thể gặp khó khăn với chuỗi rất dài.</li> </ul> <p><b>Ứng dụng trong nghiên cứu lũ quét:</b> Dự báo lũ quét dài hạn, phân tích xu hướng lũ quét theo thời gian.</p>
4	Mạng đối nghịch tạo sinh (Generative adversarial networks)	<p><b>Đặc điểm:</b> Bao gồm hai mạng neural cạnh tranh nhau: mạng sinh (generator) và mạng phân biệt (discriminator).</p> <p><b>Cấu trúc:</b> Mạng sinh tạo ra dữ liệu giả, mạng phân biệt cố gắng phân biệt dữ liệu thật và giả.</p> <p><b>Ưu điểm:</b></p> <ul style="list-style-type: none"> <li>- Có khả năng tạo ra dữ liệu mới rất giống với dữ liệu thật.</li> <li>- Học được các phân phối phức tạp của dữ liệu.</li> </ul> <p><b>Nhược điểm:</b></p> <ul style="list-style-type: none"> <li>- Khó huấn luyện và điều chỉnh.</li> <li>- Có thể không ổn định trong quá trình huấn luyện.</li> </ul> <p><b>Ứng dụng trong nghiên cứu lũ quét:</b> Tạo ra các kịch bản lũ quét giả định để đánh giá rủi ro, tăng cường dữ liệu cho các mô hình dự báo.</p>

Bản chất của mô hình trí tuệ nhân tạo là quá trình học hỏi những đặc trưng của các dữ liệu đầu vào để dự đoán kết quả đầu ra từ các dữ liệu được đào tạo. Mặc dù các mô hình tuyến tính và phi tuyến đã được áp dụng trong mô hình mưa – dòng chảy từ rất

sớm, các nghiên cứu trước đây thường đưa ra các phát hiện dưới dạng công thức thay vì “học” (Amoroch, J. & Brandstetter, A., 1971). Năm 1995, Hsu đã sử dụng mô hình mạng thần kinh nhân tạo (ANN) để mô phỏng mưa – dòng chảy ở Mississippi (Hsu, Kuo-lin, et al., 1995), đây là một trong những nghiên cứu điển hình ứng dụng trí tuệ nhân tạo trong tính toán thủy văn.

Năm 2006, Sahoo đã ứng dụng mô hình ANN để mô phỏng lũ quét cho lưu vực suối ở Oahu, Hawaii (Sahoo, G.B., et al., 2006) thông qua quá trình mô phỏng mưa dòng chảy. ANN được kết hợp với thuật toán lan truyền ngược và gọi là thuật toán BPNN với 2 lớp ẩn. Đầu vào của mô hình bao gồm (1) lượng mưa tại thời điểm  $t$ ; (2) bốc hơi tại thời điểm  $t$ ; (3) mực nước tại thời điểm  $t$ ; và (4) mực nước tại thời điểm  $t - 1$ . Sahoo đã mô phỏng với các trường hợp bao gồm: (a) với tất cả dữ liệu đầu vào; (b) tất cả dữ liệu ngoại trừ bốc hơi; (c) chỉ sử dụng mực nước; (d) tất cả ngoại trừ mực nước. Kết quả mô phỏng cho 2 trạm WK và KHS cho thấy, hệ số tương quan  $R$  trong tất cả các trường hợp đều lớn hơn 0,99.



Hình 1-6. Kết quả mô phỏng mưa – dòng chảy bằng mô hình BPNN của Sahoo

Mặc dù đạt với độ tương quan rất cao, tác giả chưa trình bày quá trình dự đoán cho một thời đoạn khác, do vậy chưa thể đánh giá được khả năng dự đoán dòng chảy từ lượng mưa (do chuỗi dữ liệu trong tương lai chưa có mực nước). Khởi đầu này cho thấy tiềm năng của dự báo dòng chảy bằng phương pháp ANN trong tính toán thủy văn.

Năm 2009, Janál đã sử dụng mô hình tập mờ (Fuzzy) để dự báo tình trạng khẩn cấp lưu vực sông trong tình huống lũ quét cho lưu vực Morva và Odra (cộng hòa Séc) trên MATLAB (Janál, Petr & Starý, Miloš, 2009). Với các tham số đầu vào bao gồm: (1) cường độ mưa; (2) thời lượng mưa; (3) chỉ số mưa tích lũy; (4) diện tích lưu vực; (5) hệ số hình dạng lưu vực; (6) độ dốc lưu vực; và (7) độ che phủ lưu vực. Đầu ra của mô hình là giá trị dòng chảy cực đại (lưu lượng). Các đại lượng đầu vào được chuẩn hóa về

từ 0÷1, kết quả đầu ra được phân thành 4 cấp độ lũ lụt, trong đó cấp 3 và cấp 4 được xem là thảm khốc.

Toukourou đã sử dụng phương pháp học thống kê để dự báo dòng chảy lũ dựa vào lượng mưa và mực nước, áp dụng cho lưu vực Gardon (Pháp) (Toukourou, Mohamed Samir, et al., 2009). Mô hình ANN được sử dụng với một lớp ẩn. Kết quả cho thấy hệ số NAS dự đoán lưu lượng lũ giảm dần theo thời đoạn dự báo. Hệ số lớn nhất đạt 0,98 cho thời đoạn dự báo sau 30 phút, sau đó giảm dần đến 0,84 cho thời đoạn dự báo sau 4 giờ và thấp nhất là 0,58 cho thời đoạn dự báo lớn hơn 5 giờ.

Bảng 1-4. Kết quả dự báo theo thời đoạn của mô hình ANN cho lưu vực sông Gardon

Forecasting horizon ( $f$ )	0.5 h	1 h	2 h	3 h	4 h	5 h	mean
$N_C$	2	2	5	3	3	3	3
Persistency criterion	0.45	0.65	0.32	0.28	0.23	0.59	0.42
$R^2$ (Nash-Sutcliffe criterion)	0.98	0.93	0.87	0.93	0.84	0.58	0.85
Estimated/Observed peak values	0.90	0.84	0.73	0.82	0.79	0.60	0.78

Số lượng lớp ẩn và tham số đầu vào sẽ quyết định các kết quả đầu ra của mô hình dự báo. Các tham số này rất khó điều chỉnh/quyết định để có thể lựa chọn được mô hình hiệu quả. Thông thường, đối với dữ liệu đầu vào, các tham số cần đạt một số tiêu chí bao gồm:

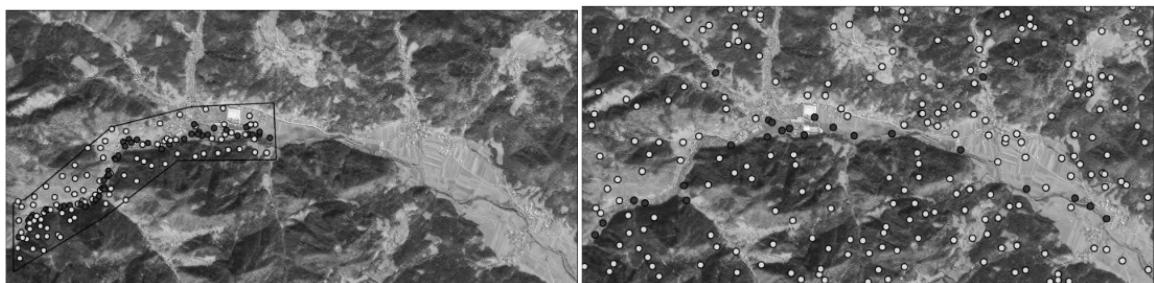
- Tính tương quan với các tham số khác: các tham số độc lập sẽ cho ra kết quả tốt hơn các tham số phụ thuộc. Ví dụ nếu cùng đưa chỉ số tập trung dòng chảy và diện tích lưu vực làm thông số đầu vào, mô hình sẽ hoạt động kém hiệu quả hơn (do tham số diện tích phụ thuộc vào tham số tập trung dòng chảy).
- Tác động đến kết quả đầu ra: các yếu tố đầu vào phải là các yếu tố có tác động đến kết quả đầu ra về mặt lý thuyết. Mô hình có các yếu tố đầu vào càng tác động mạnh đến kết quả đầu ra thì sẽ hoạt động hiệu quả hơn các mô hình đưa các tham số đầu vào ít có liên quan đến kết quả đầu ra.

Khi đã thỏa mãn các điều kiện trên, một mô hình hiệu quả cần phải xác định các tham số hiệu quả (through qua các thuật toán lựa chọn/số lớp ẩn....). Việc này khó xác định và cần một bước thử dần/tối ưu dựa trên các thuật toán đánh giá kết quả đầu ra. Một mô hình có kết quả đầu ra tốt sẽ có bộ tham số tốt.

Các tiếp cận trên hướng đến việc dự báo dòng chảy lũ cho lưu vực sông, tương tự như cách hoạt động của mô hình thủy văn/thủy lực. Các yếu tố tác động đến quá trình hình thành dòng chảy được chú trọng làm đầu vào cho mô hình dự báo. Có thể nói, đây là một hướng đi đúng đắn và phù hợp thực tế do lũ quét là một dạng lũ nên kết quả phản ứng thủy văn của một lưu vực sẽ giúp xác định được dòng chảy lũ quét. Tuy nhiên, kết

quả này không thể phân vùng lũ quét mà chỉ có thể xác định dòng chảy tại một vị trí cụ thể và phân loại ngưỡng nhằm cảnh báo lũ/lũ quét.

Năm 2013, Lamovec đã sử dụng mô hình học máy để phát hiện vùng ngập lũ cho trận lũ quét tại sông Selška Sora (Slovenia) xảy ra vào tháng 9 năm 2007 (Lamovec, Peter, et al., 2013). Hàng loạt các mô hình được sử dụng bao gồm mô hình Bayes (NavieBayes; BayesNet), mô hình cây quyết định (J48; Random Tree; Random Forest); mô hình Rules (JRip) và một số thuật toán bao gồm AdaBoostM1, LogitBoost, Bagging dựa trên bộ dữ liệu ảnh SPOT (2,5m), DTM ở độ phân giải 12,5m, mạng sông. Các tham số đầu vào bao gồm: (1) cao độ; (2) độ dốc; (3) hướng; (4) NDVI; (5) NDBI; (6) NBI cho 145 điểm thuộc một phạm vi nhỏ được trích xuất phục vụ đào tạo cho các thuộc tính có độ phân giải cao nhất và 255 điểm thuộc phạm vi lớn hơn cho độ phân giải thấp hơn.

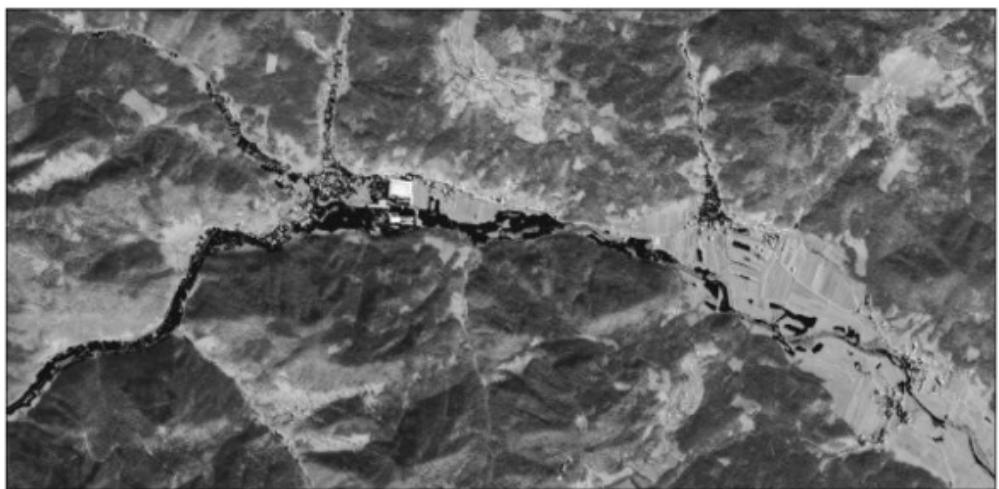


Hình 1-7. Các điểm trích xuất cho dữ liệu đào tạo mô hình

Toàn bộ các điểm (400 điểm) được sử dụng cho đào tạo mô hình (trainning), trong khi đó, 145 điểm ở độ phân giải cao sẽ được sử dụng trích xuất thành 145 điểm ở độ phân giải thấp phục vụ kiểm tra (testing). Kết quả về độ chính xác dự đoán được thể hiện ở bảng sau:

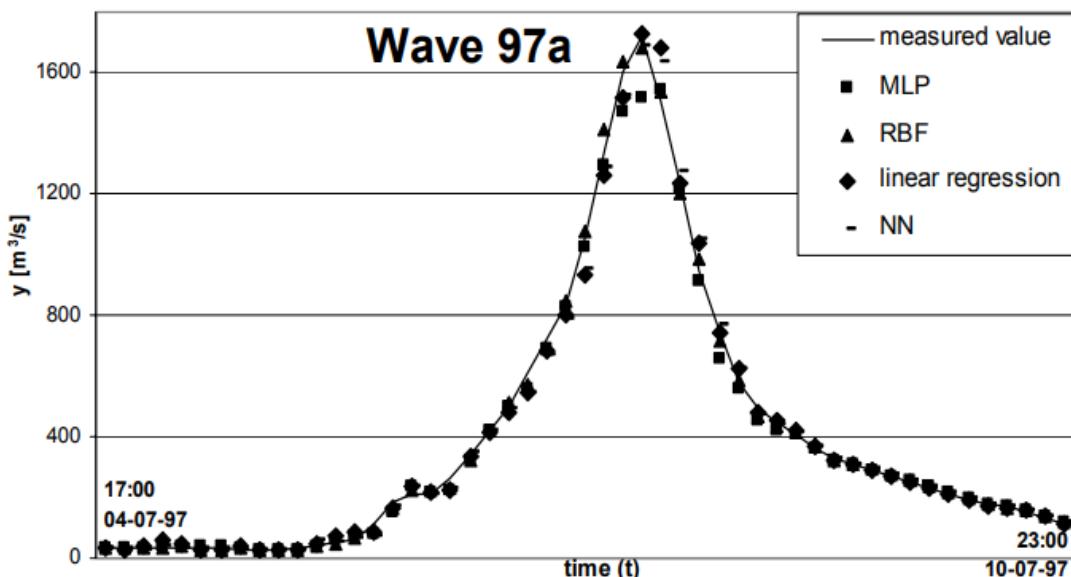
Bảng 1-5. Độ chính xác của phát hiện ngập lũ theo các thuật toán

Phương pháp	Độ chính xác	
	Đào tạo	Tập huấn
J48	88%	95%
Jrip	86%	85%
Bagging J48 (10 cây)	90%	92%
Rừng ngẫu nhiên (10 cây)	89%	92%



Hình 1-8. Kết quả phát hiện ngập lũ theo thuật toán J48

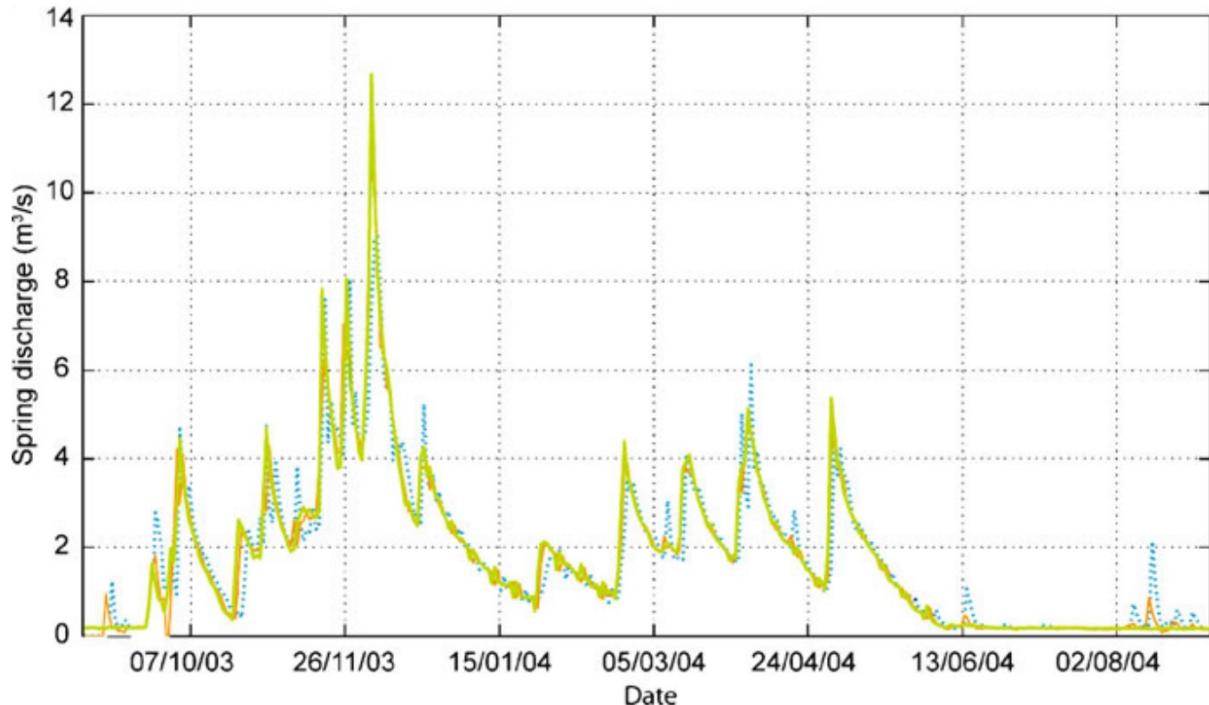
Piotrowski đã trình bày một nghiên cứu đánh giá tính hiệu quả của các kỹ thuật học máy khác nhau, đặc biệt là mạng lưới thần kinh (Neural Network) trong việc dự báo các sự kiện mưa – dòng chảy lũ nhằm so sánh hiệu quả của các mô hình khác nhau trong việc dự báo lũ quét, từ đó nâng cao khả năng dự báo lũ (Piotrowski, A., et al., 2006). Nghiên cứu này sử dụng các mô hình bao gồm Multi-Layer Perceptron (MLP), Mạng chức năng cơ sở xuyên tâm (RBF), Phương pháp tiếp cận láng giềng gần nhất (NN), Hồi quy tuyến tính cho khu vực sông Nysa Kłodzka với dữ liệu bao gồm lượng mưa và lưu lượng (mỗi 3 giờ) trong thời gian từ năm 1965 đến năm 2000. Các trận lũ được phân tích phải thỏa mãn độ trễ tối đa 24 giờ. Kết quả nghiên cứu cho thấy, mạng RBF vượt trội đáng kể so với các kỹ thuật dự báo khác, bên cạnh đó, kết quả dự báo cho thời đoạn 3 giờ có độ tin cậy cao.



Hình 1-9. Kết quả dự báo thời đoạn 3 giờ của các mô hình trong nghiên cứu của Piotrowski

Kong A Siou đã phát triển mô hình dự báo lũ quét ở lưu vực Lez (miền nam nước Pháp) sử dụng mạng lưới thần kinh nhân tạo nhằm mục đích nâng cao sự hiểu biết về đặc tính thủy động lực của lưu vực Lez và cải thiện độ chính xác của dự báo lũ (Kong A Siou, L., et al., 2010). Tác giả đã sử dụng mạng thần kinh nhân tạo (ANN) để dự báo dòng chảy ra dựa trên dữ liệu lượng mưa và lưu lượng lịch sử. Quá trình đào tạo bao gồm các kỹ thuật như dừng sớm và xác thực chéo để ngăn chặn việc trang bị quá mức và đảm bảo khả năng khái quát hóa của mô hình. Tập dữ liệu bao gồm dữ liệu lịch sử của 16 năm, bao gồm dữ liệu lượng mưa từ ba trạm quan trắc và trạm đo lưu lượng nằm trên suối Lez. Dữ liệu được sử dụng để huấn luyện mạng lưới thần kinh, tập trung vào việc lựa chọn các tham số đầu vào tối ưu thông qua phân tích tương quan chéo và kiểm soát độ phức tạp của mô hình. Kết quả chỉ ra rằng mô hình mạng thần kinh đã mô phỏng và dự báo thành công các dòng chảy ra với tiêu chí Nash xấp xỉ 0,95 cho mô phỏng và 0,84 cho dự báo trong hai ngày. Mô hình này thể hiện hiệu suất thỏa đáng trong việc đồng bộ hóa dòng chảy đỉnh với dữ liệu quan sát được, cho thấy độ tin cậy của nó trong việc dự báo lũ ngắn hạn.

Forecasting horizon $f$	0	1 day	2 days	3 days	Selected architecture
Valflaunès	0.95	0.90	0.84	0.73	8 inputs, 5 hidden neurons
Prades	0.94	0.90	0.84	0.73	9 inputs, 3 hidden neurons
St Martin	0.94	0.90	0.87	0.75	9 inputs, 3 hidden neurons

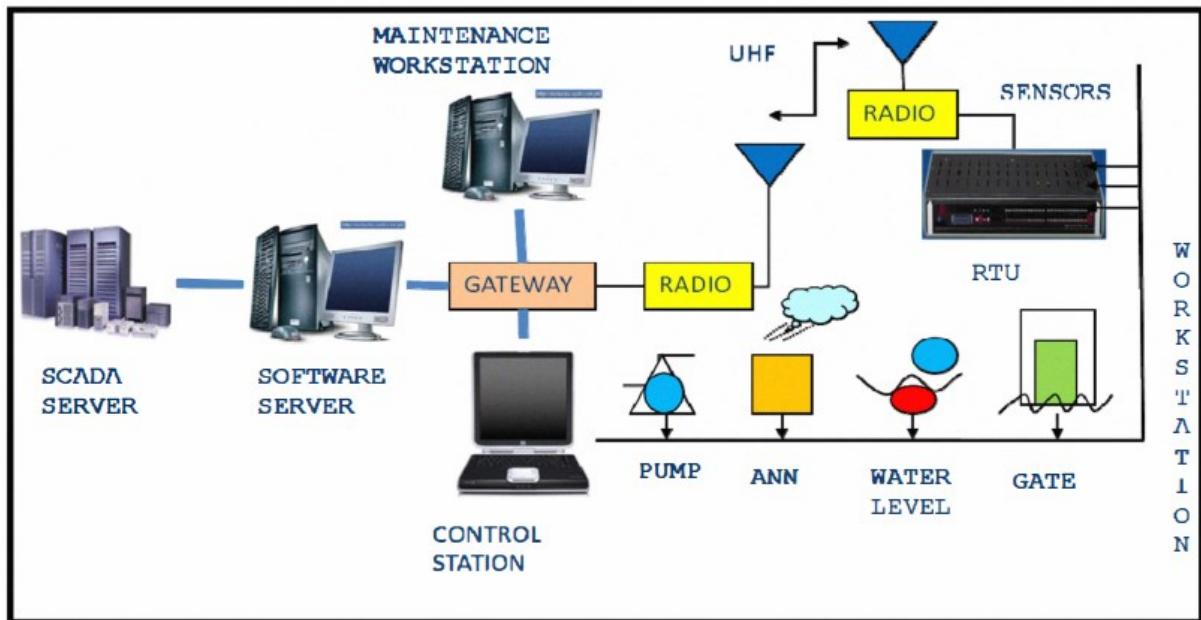


Hình 1-10. Kết quả mô phỏng dự báo lưu lượng dựa trên lượng mưa (màu xanh nõn chuối: thực đo; màu cam: mô phỏng; màu xanh chấm: dự báo trước 2 ngày)

Năm 2011, Artigue đã nghiên cứu phát triển mô hình dự báo lũ quét hiệu quả bằng cách sử dụng mạng thần kinh, được thiết kế đặc biệt để vận hành mà không dựa vào dữ

báo lượng mưa trong tương lai hoặc dữ liệu lưu lượng quan sát được trước đó. Nghiên cứu này nhằm mục đích tăng cường hệ thống cảnh báo sớm lũ quét ở các lưu vực không có trạm đo (Artigue, G., et al., 2011). Khu vực áp dụng là lưu vực Gardon de Mialet, một tiểu lưu vực của vùng Anduze ở miền Nam nước Pháp. Khu vực này có đặc điểm là độ dốc lớn, đất đá mỏng và dễ bị ảnh hưởng bởi lượng mưa lớn cục bộ, khiến nơi đây trở thành địa điểm quan trọng trong việc dự báo lũ quét. Nghiên cứu sử dụng mô hình mạng nơ-ron perceptron đa lớp (MLP). Mô hình này được chọn vì các đặc tính gần đúng và phân tích phổ quát, cho phép mô hình hóa một cách hiệu quả mối quan hệ phức tạp giữa lượng mưa và lưu lượng trong lưu vực. Kiến trúc bao gồm cả các thành phần tuyến tính và phi tuyến tính để nắm bắt động lực của phản ứng thủy văn.

Nghiên cứu sử dụng dữ liệu từ ba máy đo mưa đặt tại Barre-des-Cévennes, Mialet và Saint-Roman-de-Tousque, cùng với các phép đo lưu lượng từ máy đo tại Mialet. Bộ dữ liệu bao gồm các bản ghi lượng mưa và lưu lượng, được xử lý để chọn các biến có liên quan cho việc huấn luyện mạng lưới thần kinh. Kết quả chỉ ra rằng mô hình mạng lưới thần kinh cung cấp chất lượng dự báo tuyệt vời, đạt các giá trị tiêu chí Nash trên 0,8, được coi là chuẩn mực cho hiệu suất tốt trong lĩnh vực thủy văn. Đáng chú ý, mô hình này duy trì độ chính xác dự đoán ngay cả khi phạm vi dự báo tăng lên, thể hiện tính chắc chắn trong dự đoán của nó. Tuy nhiên, thời điểm dự đoán lưu lượng đỉnh điểm kém khả quan hơn, cho thấy còn nhiều cơ hội để cải thiện.



Hình 1-11. Hệ thống cảnh báo lũ quét khu vực Selangor, Malaysia

Izyan đã sử dụng mô hình ANN kết hợp với hệ thống SCADA bằng cách sử dụng dữ liệu khí tượng hàng ngày được thu thập từ ba trạm chính ở Selangor trong những năm 2007 đến 2010. Các thông số đầu vào bao gồm nhiệt độ bầu ướt, độ ẩm tương đối, tốc độ gió, độ mây và áp suất không khí, trong khi biến mục tiêu là lượng mưa đo bằng

máy đo mưa tại các trạm đã chọn để dự báo nguy cơ lũ quét cho khu vực Selangor, Malaysia (Izyan 'Izzati Abdul Rahman & Nik Mohd Asrol Alias, 2011). Kết quả chỉ ra rằng tuy hiệu suất dự báo hàng ngày của mô hình không cao như mong đợi nhưng nó vẫn cung cấp những dự đoán hữu ích cho các ứng dụng thực tế trong dự báo lượng mưa và quản lý lũ lụt. Mô hình ANN đã chứng tỏ tính hiệu quả trong việc xử lý dữ liệu thời tiết ồn ào và không ổn định, đặc trưng trong các tình huống như vậy. Việc tích hợp các kết quả dự báo lượng mưa vào hệ thống SCADA được coi là một bước quan trọng hướng tới cải thiện các chiến lược phòng chống lũ lụt.

Boukharouba năm 2013 đã sử dụng một phương pháp liên quan đến việc phân cụm các sự kiện lũ lụt thành các nhóm riêng biệt dựa trên đặc điểm của chúng, cho phép tạo các mô hình Hồi quy vectơ hỗ trợ (SVR) cụ thể cho từng cụm (Boukharouba, Khaled, et al., 2013). Cách tiếp cận này trái ngược với các phương pháp truyền thống sử dụng một mô hình toàn cầu duy nhất. Nghiên cứu sử dụng cơ sở dữ liệu toàn diện về lượng mưa và dữ liệu mực nước được thu thập từ năm 1993 đến năm 2008, bao gồm 23 trận lũ lớn, với thời gian lấy mẫu là 30 phút.

Việc dự báo lũ quét dựa trên lượng mưa của mô hình trí tuệ nhân tạo bẩn chất vẫn hướng về dự báo mưa – dòng chảy và đây là một cách tiếp cận đúng bởi mưa là yếu tố tự nhiên chính gây ra lũ quét (các trường hợp sự cố/nghẽn dòng là các yếu tố bất định). Mặc dù vậy, nhiều nghiên cứu hướng đến mục tiêu xác định các khu vực dễ ảnh hưởng bởi lũ quét dựa trên đặc điểm địa hình, địa mạo (susceptibility). Cách tiếp cận này có thể chỉ ra những khu vực dễ bị ảnh hưởng bởi lũ quét, nhưng không chỉ ra được khi nào thì sẽ xảy ra lũ quét. Tuy nhiên, kết quả nghiên cứu của các dạng nghiên cứu này lại có ý nghĩa lớn đối với quy hoạch sử dụng đất hoặc bố trí dân cư/tái định cư.

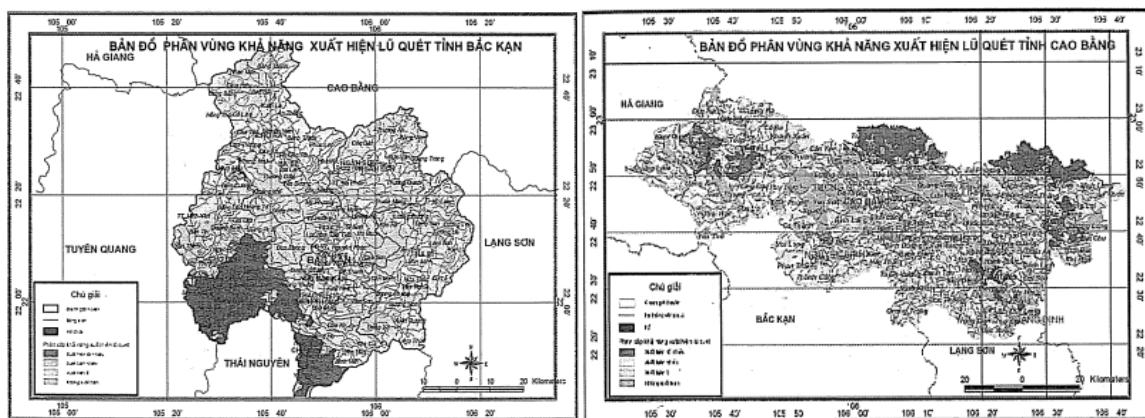
### **1.3.2 Các nghiên cứu ở Việt Nam**

#### 1.3.2.1 Phân vùng lùi quét theo trọng số dữ liệu địa không gian

Năm 1995, Cao Đăng Dư đã phân vùng khả năng xuất hiện lũ quét cho vùng Tây Bắc bằng phương pháp phân tích nhân tố (Cao Đăng Dư & Lương Tuấn Anh, 1995), các nhân tố được đưa vào đánh giá bao gồm: (1) lượng mưa một ngày lớn nhất ứng với tần suất 1% và 5%; (2) độ dốc lưu vực; (3) độ dốc lòng sông; (4) lớp phủ; (5) loại đất; và (6) mô đun dòng chảy đỉnh lũ tần suất 1% và 5%. Bản đồ được xây dựng ở tỷ lệ 1:500.000 và kết quả được phân thành 4 cấp độ: (1) khả năng xuất hiện cao; (2) khả năng xuất hiện khá; (3) khả năng xuất hiện trung bình; (4) ít có khả năng xuất hiện.



Phạm Thị Hương Lan và Vũ Minh Cát đã xây dựng bản đồ tiềm năng lũ quét phục vụ công tác cảnh báo lũ quét vùng núi Đông Bắc Việt Nam (Phạm Thị Hương Lan & Vũ Minh Cát, 2008) sử dụng 4 yếu tố bao gồm: (1) độ dốc bề mặt; (2) lượng mưa một ngày lớn nhất; (3) thảm thực vật; (4) khả năng thấm của đất. Mỗi yếu tố được phân thành 20 cấp, có 4.845 tổ hợp được xem xét, trong đó 97 tổ hợp cho thấy có khả năng xuất hiện lũ quét cao và phù hợp với điều kiện tự nhiên ở Việt Nam. Kết quả cho thấy tác giả đã xác định được ngưỡng mưa một ngày cho từng trạm bao gồm: Ở Thái Nguyên – trạm Kỳ Phú 220mm; Phú Bình 167mm; Cao Bằng – trạm Trùng Khánh 286,3mm, Hà Quảng 212mm; Bắc Kạn – trạm Bắc Kạn 205,7mm, Thác Riềng 106,3mm, Chợ Mới 143mm.

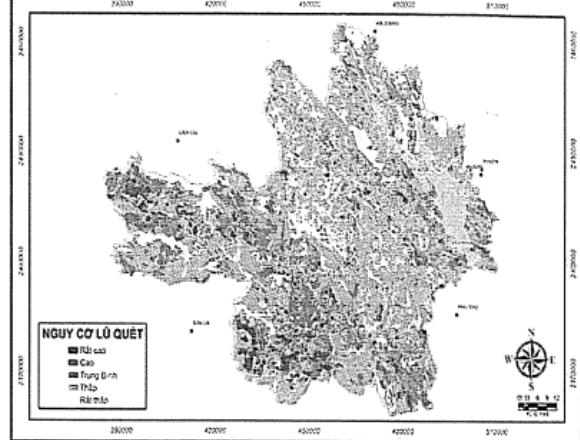


Hình 1-12. Bản đồ phân vùng khả năng xuất hiện lũ quét các tỉnh Đông Bắc

Lã Thanh Hà (2009) đã xây dựng bản đồ phân vùng nguy cơ lũ quét cho tỉnh Yên Bái (Lã Thanh Hà, 2009) dựa trên phương pháp trọng số thủ dàn với số liệu kiểm chứng là 58 trạm lũ quét đã xảy ra trên địa bàn tỉnh. Nghiên cứu sử dụng các số liệu đầu vào bao gồm: (1) lượng mưa ngày lớn nhất ứng với tần suất 50%; (2) nguy cơ xói mòn đất; (3) độ dốc lưu vực; và (4) khả năng phòng hộ của rừng.

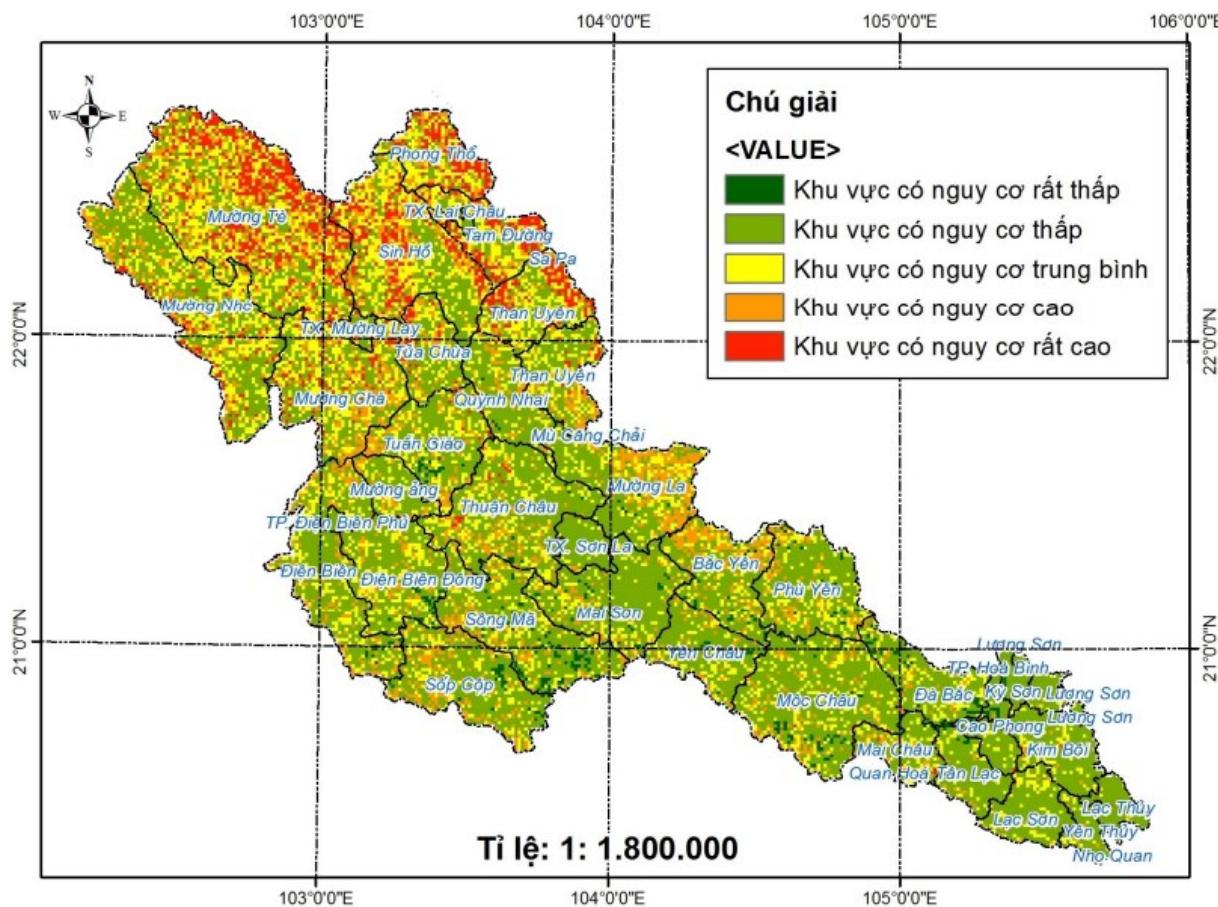
Kết quả phân vùng được xây dựng trên bản đồ tỷ lệ 1:100.000

Dương Thị Lợi (Dương Thị Lợi & Đặng Phương Lan, 2021) đã sử dụng phương pháp phân tích thứ bậc (AHP) để phân tích đa chỉ tiêu các yếu tố bao gồm: (1) độ dốc; (2) thành phần cơ giới của đất; (3) sử dụng đất; (4) lượng mưa một ngày lớn nhất. Kết quả ma trận trọng số thể hiện như sau:



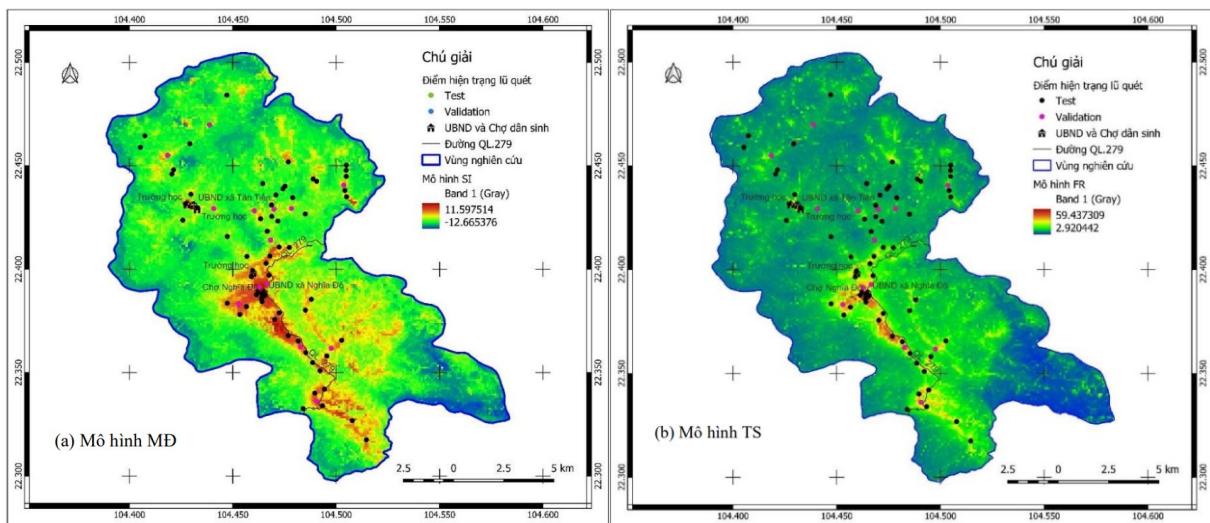
Bảng 1-6. Bảng đánh giá trọng số theo phương pháp AHP

	<b>Độ dốc</b>	<b>LM</b>	<b>HTSDD</b>	<b>TPCGĐ</b>	<b>Tổng</b>
Độ dốc	1,00	3,00	5,00	7,00	16,00
LM	0,33	1,00	3,00	5,00	9,33
HTSDD	0,2	0,33	1,00	3,00	4,53
TPCGĐ	0,14	0,20	0,33	1,00	1,68
Total	2	5	9	16	



Hình 1-13. Bản đồ nguy cơ lũ quét khu vực Tây Bắc

Đào Minh Đức và cộng sự đã xây dựng bản đồ nguy cơ lũ quét cho suối Nghĩa Đô, huyện Bảo Yên, tỉnh Lào Cai bằng việc sử dụng kết hợp kết quả chồng chập đa nhân tố và lượng mưa ngày để đưa ra bản đồ theo thời gian (Minh Đức, Đào, et al., 2022).

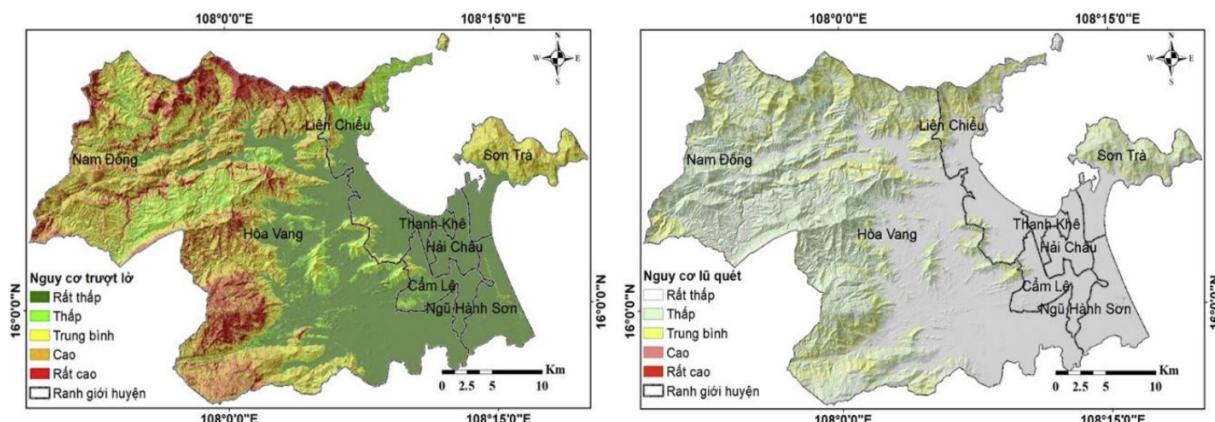


Hình 1-14. Kết quả xây dựng bản đồ nguy cơ lũ quét theo 2 phương pháp chòng chập đa nhân tố cho suối Nghĩa Đô

Có 9 yếu tố đưa vào đánh giá chòng chập đa nhân tố bằng phương pháp tỷ lệ tần suất và phương pháp thống kê mật độ bao gồm: (1) cao độ; (2) độ dốc; (3) độ cong; (4) phân cắt sâu; (5) phân cắt ngang; (6) năng lượng dòng chảy; (7) ẩm địa hình; (8) NDVI; và (9) thạch học. Kết quả xây dựng bản đồ nguy cơ ở phía trên một lần nữa lại được kết hợp với bản đồ lượng mưa trong 24 giờ để tạo ra bản đồ nguy cơ lũ quét theo thời gian. Trong đó, cả hai loại bản đồ đều được phân thành 5 cấp độ nguy cơ và kết quả cuối cùng được phân thành 7 cấp nguy cơ lũ quét

	<10mm	10-25mm	25-45mm	45-70mm	>70mm
1	I	I	II	II	II
2	I	III	III	III	III
3	II	III	IV	IV	IV
4	III	IV	V	V	V
5	III	IV	V	VI	VII

Tiềm năng hình thành lũ quét trên lưu vực suối Nghĩa Đô



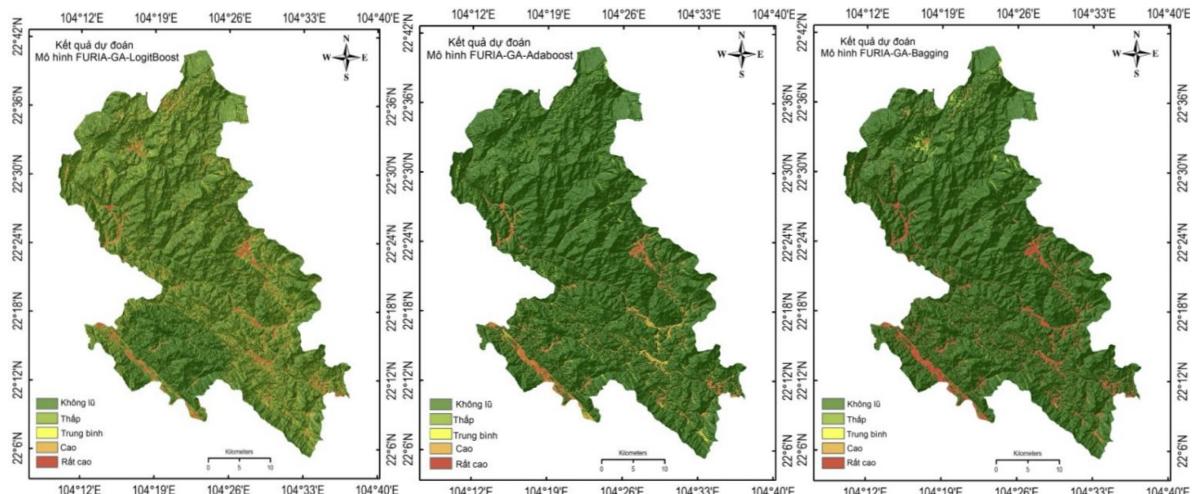
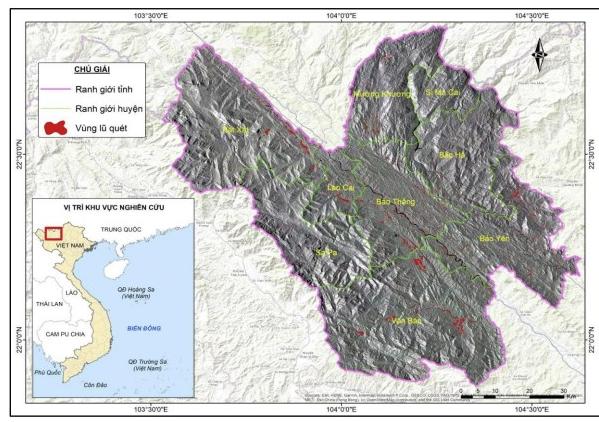
Hình 1-15. Bản đồ nguy cơ trượt lở và lũ quét cho khu vực thành phố Đà Nẵng

Nguyễn Thị Huyền và cộng sự (Thị Huyền, Nguyễn, et al., 2023) đã khoanh định các khu vực nhạy cảm về trượt lở, lũ quét cho khu vực thành phố Đà Nẵng bằng phương pháp đánh giá đa tiêu chí SMCE với 9 yếu tố. Mỗi yếu tố được chia thành 5 cấp độ, sau

đó kết hợp lại với bộ tiêu chí thành phần để xây dựng bản đồ. Kết quả đã xây dựng được 2 bản đồ cho trượt lở và lũ quét tờ tỷ lệ 1:50.000 cho khu vực thành phố Đà Nẵng.

### 1.3.2.2 Phân vùng lũ quét ứng dụng trí tuệ nhân tạo

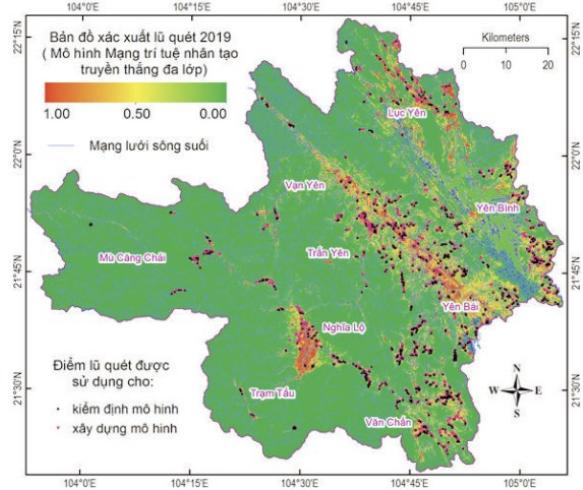
Trí tuệ nhân tạo và dữ liệu địa không gian trong những năm gần đây cũng được các nhà khoa học ở Việt Nam quan tâm và ứng dụng trong nghiên cứu về lũ quét. Ngô Thị Phương Thảo (Thị Phương Thảo, Ngô, et al., 2024) đã sử dụng ảnh Sentinel-1A để phát hiện lũ quét và xây dựng bản đồ hiện trạng lũ quét Lào Cai năm 2017. Dữ liệu lựa chọn sử dụng là băng tần VV cho các thời điểm trước và sau khi xảy ra lũ. Ngoài ra, trong một nghiên cứu khác (Ngô Thị Phương Thảo, 2024), tác giả cũng đã xây dựng mô hình FA-LM-ANN dựa trên nền tảng thuật toán ANN để tự động cập nhật và tối ưu các trọng số của mô hình dự báo lũ quét, đồng thời sử dụng mô hình tổ hợp học máy với các thuật toán di truyền GA, luật mò FURIA và cây quyết định DT để dự đoán nguy cơ lũ quét.



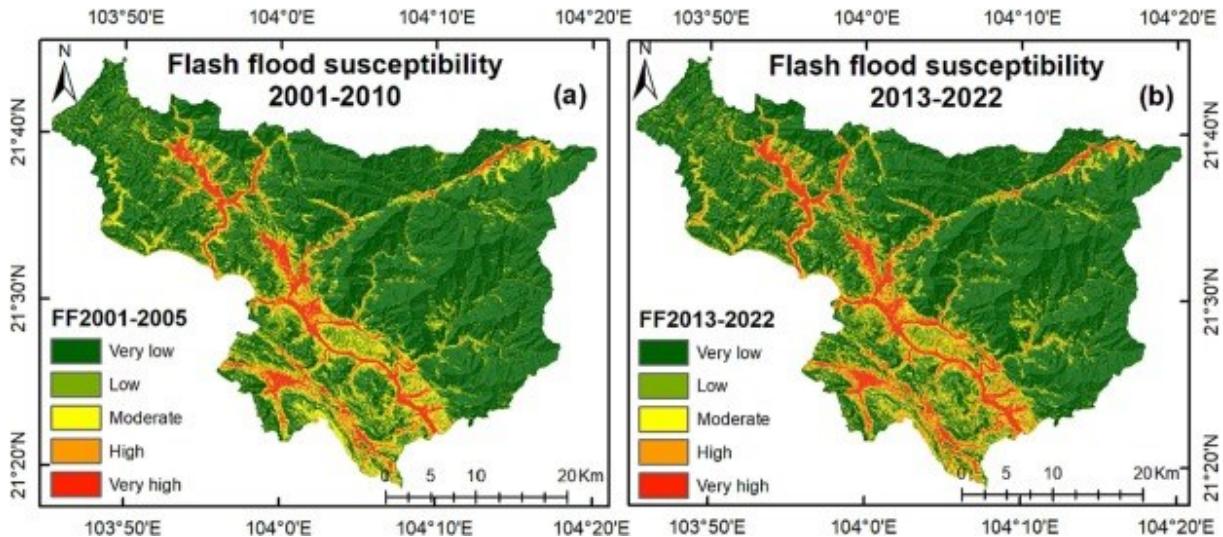
Hình 1-16. Bản đồ phân vùng nguy cơ lũ quét bởi các mô hình học máy khác nhau (Ngô Thị Phương Thảo, 2024)

Hà Thị Hằng (Ha, Hang, et al., 2022) đã phát triển mô hình học máy để dự báo lũ quét và sạt lở đất trên các tuyến đường bộ ở khu vực miền núi với dữ liệu thu thập là 235 điểm sạt lở và 88 khu vực lũ quét trên quốc lộ 6 (địa phận tỉnh Hòa Bình, Việt Nam). Các mô hình được phát triển bao gồm hình học máy kết hợp nâng cao Decorate-SYS, Rotation Forest-SYS, và Vote-SYS với bộ phân lớp cơ bản là thuật toán Systematically Developed Forest of Multiple Decision Trees (SYS).

Nguyễn Việt Nghĩa (Nguyễn Việt Nghĩa & Nguyễn Cao Cường, 2020) đã ứng dụng mạng nơ-ron nhân tạo đa lớp trong thành lập mô hình phân vùng lũ quét khu vực miền núi Tây Bắc và xây dựng bản đồ cho tỉnh Yên Bái. Mô hình sử dụng 7 yếu tố chính bao gồm độ dốc, độ cao, địa chất, lượng mưa ngày lớn nhất, thảm phủ, mật độ dòng chảy và loại đất để làm đầu vào. Kết quả đầu ra là bản đồ xác suất lũ quét cho tỉnh Yên Bái năm 2019.



Hoàng Đức Vinh (Hoang, Duc-Vinh & Liou, Yuei-An, 2024) đã đánh giá định lượng tác động của con người đến khả năng xảy ra lũ quét ở vùng núi Việt Nam bằng việc sử dụng 6 mô hình Multi-layer Perceptron, Gaussian Naïve Bayes, Support Vector Machines, K-Nearest Neighbors, XGBoost, and Random Forest để xây dựng bản đồ nhạy cảm với lũ quét cho khu vực huyện Mường La, tỉnh Sơn La. Kết quả lựa chọn mô hình RF với độ tin cậy cao nhất để lập bản đồ. Kết quả dự đoán được phân loại thành 5 cấp độ dựa trên phương pháp Natural Break.



Hình 1-17. Bản đồ nhạy cảm với lũ quét cho 2 giai đoạn tại huyện Mường La

#### 1.4. Các sản phẩm chính từ kết quả phân vùng lũ quét sử dụng trí tuệ nhân tạo và dữ liệu địa không gian.

Sự kết hợp giữa trí tuệ nhân tạo (AI), bao gồm học máy (ML) và học sâu (DL), cùng với dữ liệu địa không gian đã tạo ra những bước tiến đáng kể trong việc phân vùng và quản lý lũ quét. Các sản phẩm chính từ quá trình này bao gồm bản đồ phát hiện lũ quét, bản đồ nhạy cảm với lũ quét (susceptibility maps), và bản đồ nguy cơ lũ quét (hazard maps). Những sản phẩm này không chỉ hỗ trợ trong việc dự báo và cảnh báo

sớm mà còn đóng vai trò quan trọng trong việc lập kế hoạch giảm thiểu rủi ro và quản lý tài nguyên nước.

#### **1.4.1 Bản đồ phát hiện lũ quét**

Bản đồ phát hiện lũ quét là sản phẩm trực tiếp từ các mô hình AI nhằm xác định và phân loại các khu vực đang hoặc đã trải qua lũ quét dựa trên dữ liệu thời gian thực hoặc lịch sử. Đây là công cụ quan trọng trong việc phát hiện nhanh các sự kiện lũ quét, đặc biệt ở những khu vực có nguy cơ cao và thiếu trạm quan trắc (ungauged basins). Các nghiên cứu như “Flash Flood Detection and Alert System Using Machine Learning” (2024) (Lal, Aleena B, et al., 2024) đã sử dụng MobileNet CNN để phân loại hình ảnh vệ tinh, đạt độ chính xác lên đến 90%, từ đó tạo ra bản đồ phát hiện lũ quét với khả năng cảnh báo kịp thời qua tin nhắn hoặc mạng xã hội.

##### **1. Quy trình phát triển:**

Quá trình xây dựng bản đồ phát hiện lũ quét thường bắt đầu bằng việc thu thập dữ liệu địa không gian, bao gồm hình ảnh vệ tinh (Sentinel-1, Landsat), dữ liệu mưa từ các sản phẩm như IMERG hoặc PERSIANN, và thông tin thủy văn như mực nước hoặc lưu lượng. Các mô hình DL như Convolutional Neural Network (CNN) hoặc U-Net được sử dụng để phân đoạn (segmentation) các khu vực ngập lụt từ hình ảnh vệ tinh, trong khi các mô hình ML như Random Forest (RF) hoặc XGBoost được áp dụng để phân loại các sự kiện lũ dựa trên dữ liệu đa nguồn. Ví dụ, nghiên cứu “A modern method for building damage evaluation using deep learning approach - Case study: Flash flooding in Derna, Libya” (2024) (Sellami, El Mehdi & Rhinane, Hassan, 2024) đã khai thác dữ liệu SAR từ Sentinel-1 để xác định phạm vi ngập lụt, kết hợp với U-Net để trích xuất các khu vực bị ảnh hưởng, từ đó tạo bản đồ chi tiết về mức độ thiệt hại.

##### **2. Ứng dụng:**

Bản đồ phát hiện lũ quét có giá trị lớn trong các hệ thống cảnh báo sớm (Early Warning Systems - EWS). Chúng cung cấp thông tin thời gian thực về vị trí và phạm vi lũ quét, giúp các nhà quản lý triển khai các biện pháp ứng phó khẩn cấp như sơ tán dân cư hoặc bảo vệ cơ sở hạ tầng. Ví dụ, trong nghiên cứu “Nowcasting for urban flash floods in Africa” (2021) (Lugt, Dorien, et al., 2021), mô hình TrajGRU kết hợp với dữ liệu MSG-SEVIRI đã tạo ra bản đồ dự báo ngắn hạn (0-2 giờ) cho Ghana, hỗ trợ cảnh báo lũ quét đô thị với độ phân giải 3 km.

##### **3. Ưu điểm và hạn chế:**

Ưu điểm của bản đồ phát hiện lũ quét là khả năng xử lý dữ liệu lớn và cung cấp kết quả nhanh chóng, đặc biệt khi kết hợp AI với dữ liệu vệ tinh. Tuy nhiên, hạn chế nằm ở độ chính xác phụ thuộc vào chất lượng dữ liệu đầu vào (ví dụ: độ phân giải của hình ảnh vệ tinh) và khả năng phân biệt lũ quét với các loại ngập lụt khác. Ngoài ra, các

mô hình DL như CNN yêu cầu dữ liệu huấn luyện phong phú, điều này có thể là thách thức ở các khu vực thiếu dữ liệu lịch sử.

#### **1.4.2 Bản đồ nhạy cảm với lũ quét**

Bản đồ nhạy cảm với lũ quét (flash flood susceptibility maps) là sản phẩm dự đoán các khu vực có khả năng xảy ra lũ quét dựa trên các yếu tố điều kiện địa lý, thủy văn và khí tượng. Đây là công cụ quan trọng trong việc lập kế hoạch dài hạn và quản lý rủi ro, giúp xác định các vùng dễ bị tổn thương để ưu tiên các biện pháp giảm thiểu. Các nghiên cứu như “Flash-Flood Susceptibility Modeling Using New Approaches of Hybrid and Ensemble Tree-Based Machine Learning Algorithms” (2020) (Band, Shahab S., et al., 2020) đã sử dụng RF và Extremely Randomized Trees (ERT) để lập bản đồ độ nhạy lũ quét ở lưu vực Kalvan, Iran, đạt AUC lên đến 0.82.

##### **1. Quy trình phát triển:**

Để xây dựng bản đồ này, các nhà nghiên cứu thu thập dữ liệu địa không gian như độ dốc, độ cao, khoảng cách đến sông, lượng mưa, và sử dụng đất từ các nguồn như DEM (Digital Elevation Model), Landsat, hoặc dữ liệu khí tượng vệ tinh. Các yếu tố này được đưa vào các mô hình ML/DL như RF, SVM, hoặc DNN để tính toán chỉ số độ nhạy lũ quét (Flash-Flood Potential Index - FFPI). Nghiên cứu “Flash-Flood Potential Mapping Using Deep Learning, Alternating Decision Trees and Data Provided by Remote Sensing Sensors” (2021) (Costache, Romulus, et al., 2021) đã sử dụng DNN-WOE với AUC 0.96, kết hợp dữ liệu từ Google Earth và các yếu tố địa hình để lập bản đồ ở lưu vực Bâsca Chiojdului, Romania, cho thấy 59.38% khu vực có độ nhạy cao.

##### **2. Ứng dụng:**

Bản đồ nhạy cảm với lũ quét được sử dụng để xác định các khu vực cần ưu tiên trong quy hoạch đô thị, xây dựng hệ thống thoát nước, hoặc bảo vệ đất nông nghiệp. Ví dụ, nghiên cứu “Flash flood susceptibility mapping using stacking ensemble machine learning models” (2022) (Ilia, Ioanna, et al., 2022) đã tạo bản đồ cho một khu vực ở Hy Lạp, hỗ trợ các nhà quản lý trong việc giảm thiểu thiệt hại kinh tế và môi trường. Chúng cũng là cơ sở để đánh giá tác động của biến đổi khí hậu lên tần suất lũ quét.

##### **3. Ưu điểm và hạn chế:**

Ưu điểm của bản đồ này là khả năng tích hợp nhiều yếu tố điều kiện để đưa ra dự đoán không gian chính xác, đặc biệt khi sử dụng các mô hình ensemble như RF hoặc XGBoost. Tuy nhiên, hạn chế bao gồm việc phụ thuộc vào dữ liệu lịch sử lũ quét để huấn luyện mô hình, điều này có thể không khả thi ở các khu vực chưa từng ghi nhận sự kiện lũ. Ngoài ra, việc lựa chọn các yếu tố đầu vào đôi khi thiếu cơ sở lý thuyết rõ ràng, dẫn đến sự không nhất quán giữa các nghiên cứu.

#### **1.4.3 Bản đồ nguy cơ lũ quét**

Bản đồ nguy cơ lũ quét (flash flood hazard maps) là sản phẩm phân tích không gian nhằm xác định các khu vực có khả năng xảy ra lũ quét trong những điều kiện cụ thể, chẳng hạn như lượng mưa cực đại, đặc điểm địa hình, hoặc đặc tính thủy văn. Khác với bản đồ rủi ro, bản đồ nguy cơ tập trung hoàn toàn vào xác suất và phạm vi xảy ra của hiện tượng lũ quét, không xem xét đến các tác động tiềm tàng đối với con người, cơ sở hạ tầng, hay các yếu tố kinh tế-xã hội. Đây là công cụ nền tảng trong việc dự báo và lập kế hoạch giảm thiểu lũ quét, cung cấp thông tin quan trọng để các nhà quản lý hiểu rõ những khu vực nào dễ bị ảnh hưởng bởi sự kiện lũ trong các kịch bản thời tiết nhất định. Các nghiên cứu như “Google Earth Engine and Machine Learning for Flash Flood Exposure Mapping—Case Study: Tetouan, Morocco” (2024) (SELLAMI, EL Mehdi & Rhinane, Hassan, 2024) đã sử dụng Random Forest (RF) với độ chính xác 96% để lập bản đồ nguy cơ, tập trung vào các yếu tố tự nhiên như lượng mưa và độ dốc.

## 1. Quy trình phát triển:

Quá trình xây dựng bản đồ nguy cơ lũ quét bắt đầu bằng việc thu thập dữ liệu địa không gian, bao gồm các yếu tố như độ cao (elevation), độ dốc (slope), khoảng cách đến sông (distance to river), lượng mưa (rainfall), và chỉ số độ ẩm địa hình (Topographic Wetness Index - TWI). Dữ liệu này thường được lấy từ Digital Elevation Model (DEM), hình ảnh vệ tinh (Sentinel-1, Landsat), hoặc sản phẩm mưa vệ tinh như IMERG. Các mô hình trí tuệ nhân tạo (AI) như RF, Support Vector Machine (SVM), hoặc Deep Neural Network (DNN) được áp dụng để phân tích mối quan hệ giữa các yếu tố đầu vào và khả năng xảy ra lũ quét. Ví dụ, nghiên cứu “Flash-Flood Potential Mapping Using Deep Learning, Alternating Decision Trees and Data Provided by Remote Sensing Sensors” (2021) (Costache, Romulus, et al., 2021) đã sử dụng DNN kết hợp với Weights of Evidence (WOE) để tạo bản đồ nguy cơ cho lưu vực Bâscă Chiojdului, Romania, đạt AUC 0.96, xác định 59.38% khu vực có nguy cơ cao dựa trên các điều kiện địa hình và lượng mưa.

Các mô hình AI thường được huấn luyện với dữ liệu lịch sử lũ quét để dự đoán xác suất xảy ra trong các điều kiện cụ thể, chẳng hạn như mưa lớn kéo dài 6 giờ với cường độ 50 mm/giờ. Một số nghiên cứu khác, như “Flash flood susceptibility mapping using stacking ensemble machine learning models” (2022) (Ilia, Ioanna, et al., 2022), đã sử dụng mô hình ensemble (RF kết hợp XGBoost) để lập bản đồ nguy cơ với độ chính xác cao, tập trung vào các yếu tố tự nhiên mà không đưa vào dữ liệu về dân số hay cơ sở hạ tầng.

## 2. Ứng dụng:

Bản đồ nguy cơ lũ quét đóng vai trò quan trọng trong việc dự báo và lập kế hoạch phòng chống thiên tai ở cấp độ khu vực. Chúng giúp xác định các khu vực cần theo dõi sát sao trong mùa mưa hoặc khi có dự báo thời tiết cực đoan, từ đó hỗ trợ triển khai các

biện pháp phòng ngừa như xây dựng hệ thống thoát nước hoặc cảnh báo sớm. Ví dụ, nghiên cứu “Flash Flood Detection and Susceptibility Mapping in the Monsoon Period” (2022) (Razavi-Termeh, Seyed Vahid, et al., 2023) đã tạo bản đồ nguy cơ cho một khu vực ở Ấn Độ, cung cấp thông tin về các vùng có nguy cơ cao để điều chỉnh chiến lược quản lý nước trong mùa gió mùa. Bản đồ này cũng là nền tảng để phát triển các bản đồ rủi ro khi kết hợp với dữ liệu về tác động và thiệt hại.

### 3. Ưu điểm và hạn chế:

Ưu điểm của bản đồ nguy cơ lũ quét là khả năng cung cấp thông tin không gian chính xác về các khu vực tiềm ẩn lũ quét dựa trên dữ liệu tự nhiên, với sự hỗ trợ của các mô hình AI như RF hoặc DNN mang lại độ tin cậy cao (AUC thường  $>0.90$ ). Chúng đơn giản hơn bản đồ rủi ro vì không yêu cầu dữ liệu phức tạp về kinh tế-xã hội, do đó dễ triển khai ở các khu vực thiếu thông tin chi tiết. Tuy nhiên, hạn chế nằm ở chỗ bản đồ này không phản ánh đầy đủ mức độ nghiêm trọng của lũ quét đối với cộng đồng hoặc cơ sở hạ tầng, mà chỉ dừng lại ở việc chỉ ra khả năng xảy ra. Ngoài ra, độ chính xác phụ thuộc vào chất lượng dữ liệu đầu vào (ví dụ: độ phân giải của DEM) và khả năng mô phỏng các điều kiện thời tiết cụ thể, vốn có thể thay đổi do biến đổi khí hậu.

Bản đồ nguy cơ lũ quét là sản phẩm quan trọng trong việc xác định các khu vực có khả năng xảy ra lũ quét dưới các điều kiện cụ thể, dựa trên dữ liệu địa không gian và các mô hình AI như RF, SVM, hoặc DNN. Với vai trò tập trung vào yếu tố tự nhiên, nó cung cấp nền tảng cho dự báo và phòng ngừa lũ quét, đồng thời là bước đệm để phát triển bản đồ rủi ro khi cần đánh giá tác động rộng hơn. Mặc dù đơn giản và hiệu quả, bản đồ nguy cơ cần được cập nhật thường xuyên để phản ánh các thay đổi trong điều kiện khí hậu và địa hình, đảm bảo tính ứng dụng trong quản lý thiên tai dài hạn.

## 1.5. Đánh giá các phương pháp xác định lũ quét và dữ liệu sử dụng

Để đánh giá bao quát hơn về các phương pháp/mô hình được sử dụng trong lũ quét và dữ liệu sử dụng, nghiên cứu đã phân tích, đánh giá gần 500 nghiên cứu trên toàn thế giới dựa trên cơ sở dữ liệu CrossRef với các từ khóa bao gồm: flash flood; susceptibility; hazard; machine learning; deep learning. Kết quả xác định như sau:

### 1.5.1 Các phương pháp phổ biến xác định lũ quét

#### 1.5.1.1 Các mô hình trí tuệ nhân tạo phổ biến sử dụng trong nghiên cứu lũ quét

Các mô hình thành ba nhóm: Học máy truyền thống (ML), Học sâu (DL), và Khác (Other) (bao gồm các phương pháp thống kê hoặc lai ghép không thuộc ML/DL thuận túy). Dưới đây là bảng tổng hợp các mô hình được sử dụng, mức độ phổ biến, và hiệu suất:

Bảng 1-7. Tổng hợp các mô hình được sử dụng để xác định lũ quét

Loại mô hình	Mô hình cụ thể	Số lần xuất hiện	Hiệu suất nổi bật	Nhận xét
ML	Random Forest (RF)	68	AUC: 0.86–0.98, Accuracy: 81–96%	Phổ biến nhất, hiệu quả cao trong lập bản đồ độ nhạy lũ và dự báo.
	Support Vector Machine (SVM)	46	AUC: 0.75–0.96, Accuracy: ~90%	Hiệu quả trong phân loại, nhưng cần tối ưu hóa tham số.
	Logistic Regression (LR)	38	AUC: 0.72–0.93, Accuracy: ~75–90%	Đơn giản, phù hợp cho dữ liệu tuyến tính, hiệu suất trung bình.
	Decision Tree (DT)	24	AUC: ~0.93	Dễ hiểu, nhưng dễ bị overfitting nếu không được tối ưu.
	Classification and Regression Trees (CART)	20	AUC: ~0.91	Phù hợp cho dữ liệu phi tuyến, hiệu suất khá.
	XGBoost	18	AUC: 0.85–0.98	Hiệu suất cao, đặc biệt trong các mô hình ensemble.
	K-Nearest Neighbors (kNN)	16	AUC: 0.75–0.91	Đơn giản, hiệu quả với dữ liệu nhỏ, nhạy cảm với nhiễu.
	Naïve Bayes (NB)	14	AUC: 0.76–0.83	Phù hợp cho phân loại, hiệu suất trung bình.
	Alternating Decision Tree (ADT)	10	AUC: ~0.95	Ít phổ biến, nhưng hiệu quả trong một số nghiên cứu.
	Boosted Regression Tree (BRT)	10	AUC: ~0.82	Hiệu quả trong độ nhạy lũ, nhưng ít được sử dụng hơn RF.
DL	LightGBM	6	KGE: >0.9	Hiệu suất cao trong dự báo dòng chảy ngắn hạn.
	Extreme Learning Machine (ELM)	5	AUC: ~0.90	Nhanh, nhưng ít phổ biến trong lũ quét.
	Gradient Boosting Machine (GBM)	4	AUC: ~0.90	Hiệu quả, nhưng ít được sử dụng hơn XGBoost.
	GLMnet	3	AUC: ~0.94	Ít phổ biến, dùng trong hồi quy tuyến tính tổng quát.
	TreeBag	3	AUC: ~0.94	Ít được sử dụng, hiệu suất khá.
	Functional Tree (FT)	2	AUC: ~0.95	Ít phổ biến, hiệu quả trong một số trường hợp cụ thể.
	Kernel Logistic Regression (KLR)	2	AUC: ~0.95	Biến thể của LR, ít được sử dụng.
	Quadratic Discriminant Analysis (QDA)	2	AUC: ~0.95	Ít phổ biến, dùng trong phân loại.
	Deep Neural Network (DNN)	28	AUC: 0.92–0.96, Accuracy: >90%	Phổ biến trong các nghiên cứu gần đây, hiệu suất cao với dữ liệu phức tạp.

Loại mô hình	Mô hình cụ thể	Số lần xuất hiện	Hiệu suất nổi bật	Nhận xét
	Long Short-Term Memory (LSTM)	22	NSE: 0.80–0.87, AUC: ~0.90	Hiệu quả trong dự báo chuỗi thời gian (time-series forecasting).
	Convolutional Neural Network (CNN)	16	AUC: ~0.935, Dice: ~79.75%	Hiệu quả trong phân tích ảnh vệ tinh và lập bản đồ lũ.
	Deep Belief Network (DBN)	6	AUC: ~0.90	Ít phổ biến, nhưng hiệu quả với dữ liệu phức tạp.
	U-Net	5	Dice: ~79.75%	Chủ yếu dùng cho phân đoạn ảnh vệ tinh (segmentation).
	Deep 1D-CNN	4	AUC: 0.969, Accuracy: 90.2%	Hiệu suất rất cao trong dự báo không gian.
	PSO-BP Neural Network	3	MAE: 2.51%, RMSE: 3.74%	Kết hợp tối ưu hóa, hiệu quả trong dự báo lưu lượng đỉnh.
	MobileNet CNN	2	Accuracy: ~90%	Dùng cho phân loại hình ảnh lũ, hiệu quả cao.
	TrajGRU	1	-	Dùng trong dự báo ngắn hạn (nowcasting), ít phổ biến.
Other	Frequency Ratio (FR)	32	AUC: ~0.72–0.93	Phương pháp thống kê, thường kết hợp với ML/DL để tăng hiệu suất.
	Analytical Hierarchy Process (AHP)	26	AUC: ~0.80–0.845	Phương pháp ra quyết định đa tiêu chí, thường kết hợp với ML.
	Weights of Evidence (WOE)	12	AUC: ~0.96	Thống kê, hiệu quả khi kết hợp với DL (như DLNN-WOE).
	Fuzzy Logic	10	AUC: ~0.90	Xử lý dữ liệu không chắc chắn, thường kết hợp với ML/DL.
	Monte Carlo	5	-	Dùng trong mô phỏng xác suất, ít phổ biến trong lũ quét.
	Reinforcement Learning	3	-	Mới xuất hiện, tiềm năng trong cảnh báo thời gian thực.
	Bayesian Belief Network (BBN)	3	AUC: ~0.90	Ít phổ biến, dùng trong mô hình xác suất.
	Hydrological Models (HEC-HMS, GR4H)	15	NSE: 0.56–0.94	Mô hình thủy văn truyền thống, thường kết hợp với ML/DL.
	Statistical Methods (khác)	10	-	Các phương pháp thống kê chung, ít được chỉ rõ.

Các mô hình học máy phổ biến nhất: Random Forest (RF) là mô hình được sử dụng nhiều nhất (68 lần), nhờ tính linh hoạt, khả năng xử lý dữ liệu phi tuyến, và hiệu suất cao (AUC thường  $>0.90$ ). SVM (46 lần) và Logistic Regression (38 lần) cũng rất phổ biến, đặc biệt trong các nghiên cứu về lập bản đồ độ nhạy lũ (susceptibility mapping). XGBoost (18 lần) và LightGBM (6 lần) đang nổi lên như các lựa chọn thay thế mạnh mẽ cho RF, với hiệu suất cạnh tranh.

Các mô hình học sâu phổ biến nhất: Deep Neural Network (DNN) (28 lần) và Long Short-Term Memory (LSTM) (22 lần) là các mô hình DL được sử dụng nhiều nhất, đặc biệt trong các nghiên cứu từ 2020 trở đi. Convolutional Neural Network (CNN) (16 lần) và Deep 1D-CNN (4 lần) cho thấy hiệu quả vượt trội trong xử lý dữ liệu hình ảnh vệ tinh và dự báo không gian (AUC lên đến 0.969).

Các mô hình/phương pháp khác: Frequency Ratio (FR) (32 lần) và Analytical Hierarchy Process (AHP) (26 lần) là các phương pháp thống kê phổ biến, thường được dùng để xác định trọng số của các yếu tố đầu vào hoặc kết hợp với ML/DL. Hydrological Models như HEC-HMS và GR4H (15 lần) vẫn được sử dụng, đặc biệt trong các nghiên cứu kết hợp với ML/DL để cải thiện dự báo.

Hiệu suất của các phương pháp/mô hình: Các mô hình ensemble (như RF, XGBoost, DNN-FR) thường đạt AUC >0.90, vượt trội so với các mô hình đơn lẻ như LR hoặc NB. Trong DL, Deep 1D-CNN ghi nhận AUC cao nhất (0.969), nhưng chỉ xuất hiện trong một số nghiên cứu cụ thể. LSTM cho thấy hiệu suất tốt trong dự báo chuỗi thời gian (NSE: 0.80–0.87).

Xu hướng: Các mô hình DL (LSTM, CNN) và ensemble (RF, XGBoost) ngày càng được ưa chuộng, đặc biệt trong các khu vực thiếu dữ liệu (ungauged basins) hoặc cần xử lý dữ liệu phức tạp (hình ảnh vệ tinh, chuỗi thời gian). RF, SVM, LR, DNN, và LSTM là những mô hình chủ đạo, với RF dẫn đầu về số lần xuất hiện và hiệu suất ổn định.

#### *1.5.1.2 Phân tích, lựa chọn mô hình trí tuệ nhân tạo trong phân vùng lũ quét*

Dựa trên các phân tích tại chương 2, có thể thấy bốn mô hình học máy gồm rừng ngẫu nhiên (RF), máy hỗ trợ vectơ (SVM), hồi quy logistic (LR) và LightGBM (LGBM) cùng với hai mô hình học sâu là mạng nơ-ron sâu (DNN), mạng nơ-ron tích chập (CNN) và mạng nơ-ron hồi tiếp dài ngắn hạn (LSTM) là các phương pháp nổi bật hiện nay trong lĩnh vực dự đoán nguy cơ lũ quét. Tuy nhiên, không phải mô hình nào cũng phù hợp cho mọi loại dữ liệu hoặc mục đích sử dụng, đặc biệt là trong bối cảnh dữ liệu địa không gian có tính chất phức tạp và không đồng nhất.

Trong nghiên cứu này, bộ dữ liệu bao gồm các raster đặc trưng, mỗi raster đại diện cho một yếu tố ảnh hưởng đến nguy cơ lũ quét như địa hình, thảm phủ, lớp phủ đất, lượng mưa, chỉ số CN, v.v. Dữ liệu đầu ra (label) là bản đồ nguy cơ lũ quét dạng raster, trong đó mỗi pixel mang một nhãn (ví dụ: nguy cơ cao/thấp hoặc các cấp nguy cơ). Một điểm quan trọng là bản chất không gian của dữ liệu: mỗi điểm dữ liệu (pixel) không chỉ phụ thuộc vào đặc trưng của chính nó mà còn bị ảnh hưởng bởi khu vực xung quanh.

Do đó, việc lựa chọn mô hình cần xem xét kỹ khả năng xử lý dữ liệu raster, khả năng học các đặc trưng không gian, và cả tính toán hiệu quả với dữ liệu lớn.

1. Đánh giá các mô hình trí tuệ nhân tạo và đề xuất lựa chọn mô hình.

a. Nhóm mô hình học máy

### **Rừng ngẫu nhiên (Random Forest – RF)**

RF là một thuật toán học máy phổ biến với ưu điểm:

- Có khả năng xử lý dữ liệu có nhiều đặc trưng mà không cần giả định phân phối.
- Khả năng chống overfitting tốt nhờ trung bình hóa nhiều cây quyết định.
- Dữ liệu đầu vào: RF yêu cầu dữ liệu dạng bảng (tabular), tức là mỗi pixel được biểu diễn dưới dạng một vector đặc trưng gồm 20 giá trị từ 20 raster tương ứng.

Hạn chế: RF không tự động nắm bắt thông tin không gian, tức là không xem xét mối liên hệ giữa pixel và các pixel lân cận. Để bổ sung yếu tố không gian, cần thủ công mở rộng đặc trưng (ví dụ: tính toán giá trị trung bình hoặc phương sai của từng raster trong vùng  $3 \times 3$  hoặc  $5 \times 5$  quanh điểm trung tâm).

### **Máy hỗ trợ vectơ (Support Vector Machine – SVM)**

SVM là thuật toán mạnh về phân loại nhị phân:

- Đặc biệt hiệu quả khi dữ liệu không tuyến tính (nhờ kernel trick).
- Ổn định với dữ liệu nhỏ và trung bình.
- Dữ liệu đầu vào: Giống RF, SVM nhận đầu vào là vector đặc trưng tại từng pixel.

Hạn chế: SVM cũng không tự khai thác đặc trưng không gian. Ngoài ra, với tập dữ liệu lớn (nhiều pixel), chi phí tính toán của SVM có thể rất cao.

### **Hồi quy logistic (Logistic Regression – LR)**

LR là mô hình thống kê cổ điển:

- Dễ triển khai, dễ diễn giải.
- Thích hợp cho những bài toán phân loại tuyến tính.
- Dữ liệu đầu vào: Mỗi pixel là một vector đặc trưng gồm 20 đặc trưng.

Hạn chế: Khả năng học phi tuyến kém hơn RF và SVM, và hoàn toàn không nắm bắt được ngữ cảnh không gian trừ khi có đặc trưng bổ sung.

### **Light Gradient Boosting Machine (LGBM)**

LGBM là thuật toán gradient boosting hiện đại với nhiều cải tiến vượt trội:

- Tốc độ huấn luyện nhanh nhờ kỹ thuật leaf-wise tree growth thay vì level-wise như các thuật toán boosting truyền thống.
- Hiệu quả xử lý dữ liệu lớn với yêu cầu bộ nhớ thấp nhờ tối ưu hóa lưu trữ histogram.

- Khả năng học mối quan hệ phi tuyến phức tạp giữa các đặc trưng thông qua việc kết hợp tuần tự nhiều weak learner.
- Tích hợp sẵn các kỹ thuật regularization để giảm overfitting như dropout cho boosting và constraining tree complexity.
- Dữ liệu đầu vào: LGBM xử lý dữ liệu dạng bảng, với mỗi pixel được biểu diễn bởi vector 20 đặc trưng từ các raster tương ứng.

Hạn chế: Tương tự RF và SVM, LGBM không tự động khai thác thông tin không gian giữa các pixel lân cận. Mô hình chỉ xem xét từng pixel như một điểm dữ liệu độc lập, không nắm bắt được cấu trúc không gian của ảnh viễn thám. Để tận dụng ngũ cảnh không gian, cần bổ sung thủ công các đặc trưng texture hoặc spatial features như local variance, spatial autocorrelation trong cửa sổ lân cận.

## b. Nhóm mô hình học sâu

### **Mạng nơ-ron sâu (Deep Neural Network – DNN)**

DNN là phiên bản mở rộng của mạng perceptron đa lớp:

- Có khả năng học phi tuyến mạnh mẽ.
- Linh hoạt và dễ tùy biến về kiến trúc.
- Dữ liệu đầu vào: DNN nhận vector đặc trưng giống như RF/SVM, có thể là 20 đặc trưng gốc hoặc mở rộng (ví dụ 20 đặc trưng + 20 đặc trưng từ vùng lân cận).
- Tự động học các mối quan hệ phi tuyến phức tạp.

Hạn chế: Dù mạnh về học phi tuyến, DNN không tự nắm bắt spatial pattern nếu dữ liệu chỉ là vector.

### **Mạng nơ-ron tích chập (Convolutional Neural Network – CNN)**

CNN nổi bật nhờ khả năng xử lý dữ liệu dạng ảnh hoặc raster:

- Có khả năng học đặc trưng không gian (spatial pattern) thông qua các lớp convolution.
- Phù hợp tự nhiên cho bài toán bản đồ và dữ liệu raster.
- Dữ liệu đầu vào: Một patch raster (ví dụ:  $15 \times 15$  pixel) với 20 kênh tương ứng với 20 raster đặc trưng. Pixel trung tâm của patch là điểm cần dự đoán.
- Học được mối quan hệ không gian giữa pixel và khu vực xung quanh.
- Tận dụng tốt cấu trúc dữ liệu raster mà không cần thủ công mở rộng đặc trưng.

Lưu ý: Kích thước patch cần hợp lý để vừa đảm bảo có đủ thông tin không gian, vừa tối ưu tài nguyên tính toán.

### **Mạng nơ-ron hồi tiếp dài ngắn hạn (Long Short-Term Memory – LSTM)**

LSTM là một kiến trúc nổi bật cho chuỗi thời gian:

- Có khả năng học các phụ thuộc dài hạn trong dữ liệu tuần tự.
- Rất mạnh nếu dữ liệu mang tính chất thời gian.
- Dữ liệu đầu vào: Chuỗi đặc trưng theo thời gian tại từng pixel (ví dụ: mưa tích lũy theo giờ/ngày).

**Hạn chế:** Trong bài toán hiện tại, dữ liệu chính là các raster tĩnh (không có chuỗi thời gian rõ ràng), do đó LSTM không thực sự phù hợp nếu chỉ có dữ liệu tĩnh. Tuy nhiên, nếu mở rộng thêm dữ liệu khí tượng theo thời gian, LSTM hoặc hybrid CNN-LSTM có thể phát huy sức mạnh.

## 2. Đánh giá tổng hợp và lựa chọn mô hình xác định lũ quét

Bảng 1-8. Đặc điểm tổng hợp các mô hình trí tuệ nhân tạo trong nghiên cứu lũ quét

Mô hình	Định dạng dữ liệu	Khả năng học phi tuyến	Khả năng khai thác không gian	Độ phức tạp tính toán	Phù hợp bài toán hiện tại
RF	Vector	Tốt	Không (cần thủ công)	Trung bình	Tốt
LGB M	Vector	Rất tốt	Không (cần thủ công)	Thấp	Rất tốt
SVM	Vector	Trung bình	Không (cần thủ công)	Cao	Khá
LR	Vector	Yếu	Không	Thấp	Trung bình
DNN	Vector	Rất tốt	Không (cần thủ công)	Cao	Tốt
CNN	Ảnh raster (patch)	Rất tốt	Tự động học các đặc trưng không gian	Cao	Rất tốt
LSTM	Chuỗi thời gian	Rất tốt	Không	Cao	Hạn chế (trừ khi dữ liệu tuân theo chuỗi thời gian liên tục)

Trong quá trình lựa chọn mô hình dự đoán nguy cơ lũ quét, việc cân nhắc giữa các phương pháp học máy và học sâu phải dựa trên cả đặc điểm dữ liệu và năng lực của từng mô hình. Bộ dữ liệu sử dụng trong nghiên cứu bao gồm 20 raster đặc trưng, đại diện cho các yếu tố địa lý, khí tượng và thủy văn khác nhau. Đây là dữ liệu dạng không gian (spatial data), mà mỗi pixel trong bản đồ nguy cơ lũ quét không chỉ chịu tác động của các yếu tố tại chính nó mà còn phụ thuộc rất lớn vào điều kiện của khu vực xung quanh. Chính vì thế, việc lựa chọn mô hình cần xem xét khả năng khai thác thông tin không gian một cách hiệu quả.

Mạng nơ-ron tích chập (CNN) nổi bật là một lựa chọn phù hợp nhờ kiến trúc được thiết kế chuyên biệt để xử lý dữ liệu ảnh hoặc raster. Khác với các mô hình học máy truyền thống vốn chỉ xem xét từng điểm dữ liệu như một vector đặc trưng rời rạc, CNN có khả năng tiếp nhận đầu vào dạng patch raster (ví dụ: một vùng  $15 \times 15$  pixel với 20 kênh đặc trưng), trong đó pixel trung tâm là đối tượng dự đoán. Nhờ cơ chế convolution,

CNN tự động học được mối quan hệ không gian giữa pixel và vùng lân cận, cho phép mô hình nhận diện các pattern phức tạp như đường rãnh, sườn dốc hay khu vực tập trung nước mưa – những yếu tố quan trọng trong nguy cơ lũ quét. Đây chính là lợi thế vượt trội mà các mô hình như Logistic Regression, SVM hay thậm chí DNN không thể tự động làm được nếu chỉ sử dụng vector đặc trưng đơn lẻ. Việc xây dựng dữ liệu đầu vào dưới dạng patch cũng giúp CNN linh hoạt mở rộng quy mô mà không làm mất đi tính liên kết không gian.

Tuy nhiên, CNN yêu cầu tài nguyên tính toán lớn và thời gian huấn luyện dài hơn so với các phương pháp truyền thống. Trong bối cảnh đó, Rừng ngẫu nhiên (RF) trở thành một lựa chọn thay thế hiệu quả, nhất là khi cần một mô hình dễ triển khai và đánh giá nhanh. RF xử lý tốt dữ liệu có nhiều đặc trưng và không yêu cầu giả định phức tạp về phân phối dữ liệu. Dù không có khả năng học thông tin không gian một cách trực tiếp như CNN, RF cho phép người nghiên cứu mở rộng bộ đặc trưng bằng cách tích hợp thêm các thống kê không gian thủ công, ví dụ như giá trị trung bình hoặc phương sai trong vùng lân cận  $5 \times 5$  hoặc  $7 \times 7$  pixel quanh điểm trung tâm. Cách tiếp cận này tuy không tối ưu bằng CNN nhưng lại là một giải pháp trung gian đáng tin cậy, giúp tận dụng sức mạnh của RF mà vẫn bổ sung yếu tố không gian ở mức độ nhất định.

Mặc dù vậy, khi triển khai thực tế trên tập dữ liệu raster gồm 20 lớp đặc trưng không gian với số lượng ô raster rất lớn, mô hình RF bộc lộ một hạn chế rõ rệt: thời gian dự đoán chậm do phải thực hiện voting trên hàng trăm cây quyết định, mỗi cây lại duyệt qua toàn bộ tập dữ liệu.

Trước thực tế đó, Light Gradient Boosting Machine (LightGBM) mới nổi lên trong những năm gần đây như một giải pháp thay thế tối ưu hơn (do đó chưa có nhiều nghiên cứu về lũ quét sử dụng). LightGBM về bản chất vẫn thuộc nhóm các mô hình cây quyết định – tương tự RF – nhưng cải tiến bằng kỹ thuật gradient boosting. Điều này giúp LightGBM khắc phục hai vấn đề chính mà RF gặp phải: (1) tối ưu hóa chất lượng dự đoán nhờ việc xây dựng các cây liên tiếp để liên tục sửa lỗi dự đoán, và (2) tốc độ huấn luyện cũng như dự đoán nhanh hơn nhờ thuật toán leaf-wise và kỹ thuật chia nhỏ theo histogram, tối ưu cho dữ liệu lớn.

Ngoài ra, LightGBM vẫn duy trì được các ưu điểm cốt lõi vốn là lý do lựa chọn RF: khả năng xử lý tốt dữ liệu không chuẩn hóa, đánh giá tầm quan trọng đặc trưng rõ ràng, và khả năng phân tích cấu trúc cây quyết định để hỗ trợ diễn giải mô hình. Điểm mạnh này giúp LightGBM không chỉ đáp ứng được yêu cầu về hiệu quả tính toán mà còn giữ nguyên tính minh bạch và dễ giải thích trong ứng dụng khoa học và quản lý tài nguyên thiên nhiên.

Từ các phân tích trên có thể kết luận rằng, trong điều kiện lý tưởng về tài nguyên tính toán và mục tiêu tối ưu hóa chất lượng mô hình, CNN là lựa chọn ưu tiên nhờ khả

năng khai thác triệt để cấu trúc không gian của dữ liệu raster. Trong khi đó, LightGBM đóng vai trò như một mô hình nền tảng, vừa đơn giản triển khai vừa cho kết quả khá tốt khi biết cách mở rộng đặc trưng hợp lý. Việc kết hợp cả hai mô hình này trong nghiên cứu không chỉ giúp so sánh hiệu quả giữa học sâu và học máy truyền thống mà còn tăng độ tin cậy và tính khách quan của kết quả cuối cùng.

Trong nghiên cứu này, nhóm nghiên cứu sẽ phát triển tất cả các mô hình để có sự đánh giá tốt hơn bao gồm 6 mô hình bao gồm: RF, SVM, LR, LGBM, CNN và DNN. Riêng mô hình LSTM được đánh giá là chưa phù hợp trong nghiên cứu này đối với dữ liệu không phải là dạng chuỗi theo thời gian.

### **1.5.2 Dữ liệu sử dụng**

Các yếu tố đầu vào (input factors) được sử dụng trong các nghiên cứu rất đa dạng, bao gồm các đặc trưng địa lý, thủy văn, khí tượng, và nhân tạo. Dưới đây là danh sách tổng hợp các yếu tố phổ biến:

Bảng 1-9. Tổng hợp các dữ liệu phổ biến sử dụng trong nghiên cứu lũ quét

Dữ liệu	Mô tả	Loại dữ liệu
Slope (Độ dốc)	Góc nghiêng của bề mặt địa hình, ảnh hưởng đến tốc độ dòng chảy.	Địa hình (Topographic)
Elevation (Độ cao)	Độ cao so với mực nước biển, ảnh hưởng đến phân bố mưa và dòng chảy.	Địa hình
Aspect (Hướng dốc)	Hướng của độ dốc, ảnh hưởng đến tiếp xúc ánh sáng và độ ẩm đất.	Địa hình
Distance to River (Khoảng cách đến sông)	Khoảng cách từ điểm nghiên cứu đến sông gần nhất, liên quan đến nguy cơ lũ.	Thủy văn (Hydrological)
Topographic Wetness Index (TWI)	Chỉ số độ ẩm địa hình, biểu thị khả năng tích tụ nước.	Thủy văn
Stream Power Index (SPI)	Chỉ số sức mạnh dòng chảy, liên quan đến năng lượng dòng chảy.	Thủy văn
Topographic Position Index (TPI)	Chỉ số vị trí địa hình, xác định vị trí tương đối (đỉnh, sườn, thung lũng).	Địa hình
Rainfall (Lượng mưa)	Lượng mưa tích lũy hoặc cường độ mưa, yếu tố chính gây lũ quét.	Khí tượng (Meteorologica l)
Land Use/Land Cover (LULC)	Loại sử dụng đất (nông nghiệp, đô thị, rừng), ảnh hưởng đến khả năng thấm nước.	Nhân tạo (Anthropogeni c)
Soil Type (Loại đất)	Đặc tính đất (đất sét, cát, v.v.), ảnh hưởng đến khả năng thấm và dòng chảy.	Địa chất (Geological)
Lithology (Thạch học)	Lớp bề mặt, ảnh hưởng đến khả năng thấm nước và xói mòn.	Địa chất
Stream Density (Mật độ sông)	Mật độ mạng lưới sông trong khu vực, liên quan đến thoát nước.	Thủy văn
Curvature (Độ cong)	Độ cong của địa hình (lồi, lõm), ảnh hưởng đến dòng chảy và tích tụ nước.	Địa hình
Profile Curvature	Độ cong theo hướng dốc, ảnh hưởng đến tốc độ dòng chảy.	Địa hình

Dữ liệu	Mô tả	Loại dữ liệu
Plan Curvature	Độ cong ngang, ảnh hưởng đến hướng dòng chảy.	Địa hình
Distance to Road (Khoảng cách đến đường)	Khoảng cách đến đường, liên quan đến tác động nhân tạo và khả năng tiếp cận.	Nhân tạo
Population Density (Mật độ dân số)	Mật độ dân cư, ảnh hưởng đến mức độ tổn thương và thiệt hại.	Nhân tạo
River Length (Chiều dài sông)	Tổng chiều dài sông trong lưu vực, liên quan đến dòng chảy.	Thủy văn
Basin Area (Diện tích lưu vực)	Diện tích lưu vực, ảnh hưởng đến lượng nước tập trung.	Thủy văn
Soil Moisture (Độ ẩm đất)	Độ ẩm của đất, ảnh hưởng đến khả năng thấm nước và dòng chảy bề mặt.	Thủy văn
Gully Density (Mật độ rãnh xói mòn)	Mật độ các rãnh xói mòn, liên quan đến nguy cơ lũ quét.	Địa hình
Normalized Difference Built-up Index (NDBI)	Chỉ số xây dựng, biểu thị mức độ đô thị hóa.	Nhân tạo
Temperature (Nhiệt độ)	Nhiệt độ môi trường, ảnh hưởng đến bốc hơi và độ ẩm đất.	Khí tượng
Vegetation Cover (Độ phủ thực vật)	Mức độ che phủ thực vật, ảnh hưởng đến xói mòn và thấm nước.	Nhân tạo
Precipitation Estimates (Uớc lượng mưa)	Dữ liệu mưa từ vệ tinh (IMERG, PERSIANN), dùng khi thiếu trạm đo.	Khí tượng
Discharge (Lưu lượng)	Lưu lượng nước trong sông, yếu tố trực tiếp liên quan đến lũ.	Thủy văn
Water Level (Mực nước)	Mực nước sông hoặc hồ, biểu thị tình trạng thủy văn.	Thủy văn

Các yếu tố độ dốc (slope), độ cao (elevation), khoảng cách đến sông (distance to river), lượng mưa (rainfall) và sử dụng đất (LULC) là những yếu tố được sử dụng phổ biến nhất, xuất hiện trong hầu hết các nghiên cứu, do chúng có tác động trực tiếp đến dòng chảy bề mặt và nguy cơ lũ quét.

Dữ liệu khí tượng (như lượng mưa) và địa hình (như độ dốc, độ cao) đóng vai trò quan trọng nhất trong việc dự báo lũ quét. Sử dụng đất (LULC) và Loại đất (Soil Type) ngày càng được chú trọng, đặc biệt trong các nghiên cứu về độ nhạy lũ, vì chúng ảnh hưởng đến khả năng thấm nước và xói mòn.

Dữ liệu từ vệ tinh (IMERG, PERSIANN, Sentinel) ngày càng được sử dụng phổ biến để bổ sung cho các khu vực thiếu trạm đo (ungauged basins). Một số nghiên cứu gần đây đã bắt đầu sử dụng các yếu tố nhân tạo như mật độ dân số và khoảng cách đến đường để đánh giá mức độ tổn thương (vulnerability).

### 1.5.3 Sự khác biệt dữ liệu giữa học máy và học sâu

Dựa trên các tài liệu tìm hiểu, mô hình học máy và học sâu trong AI có sự khác biệt đáng kể. Chi tiết thể hiện như sau:

Bảng 1-10. Sự khác biệt về dữ liệu giữa hai mô hình trí tuệ nhân tạo

Tiêu chí	Học máy (Machine Learning)	Học sâu (Deep Learning)
Loại dữ liệu đầu vào	<ul style="list-style-type: none"> <li>Dữ liệu có cấu trúc: bảng số liệu (lượng mưa, mực nước, độ dốc địa hình, lưu lượng sông).</li> <li>Đặc trưng được trích xuất thủ công (ví dụ: tổng mưa theo giờ, độ ẩm đất).</li> </ul>	<ul style="list-style-type: none"> <li>Dữ liệu thô hoặc ít xử lý: hình ảnh vệ tinh, radar thời tiết, video dòng chảy, bản đồ địa hình.</li> <li>Tự động trích xuất đặc trưng từ dữ liệu như hình ảnh hoặc chuỗi thời gian.</li> </ul>
Yêu cầu chuẩn hóa dữ liệu	<ul style="list-style-type: none"> <li>Bắt buộc chuẩn hóa dữ liệu số (lượng mưa, mực nước) về cùng thang đo (ví dụ: [0, 1]).</li> <li>Phương pháp: Min-Max Scaling, Z-score...</li> <li>Dữ liệu địa hình cần đồng nhất định dạng.</li> </ul>	<ul style="list-style-type: none"> <li>Chuẩn hóa ít nghiêm ngặt hơn, nhưng cần cho hình ảnh (pixel về [0, 1]) hoặc chuỗi thời gian (chuẩn hóa biên độ).</li> <li>Dữ liệu radar/hình ảnh thường tự động chuẩn hóa trong mô hình.</li> </ul>
Phạm vi giá trị	<ul style="list-style-type: none"> <li>Dữ liệu số trong phạm vi chuẩn hóa (ví dụ: lượng mưa 0-100 mm, mực nước 0-10 m).</li> <li>Outliers (mưa cực lớn) cần xử lý để tránh sai lệch mô hình.</li> </ul>	<ul style="list-style-type: none"> <li>Hình ảnh: pixel [0, 255] hoặc [0, 1].</li> <li>Chuỗi thời gian (mưa, mực nước): chuẩn hóa về [-1, 1] hoặc [0, 1].</li> <li>Outliers ít ảnh hưởng nhờ mạng sâu.</li> </ul>
Cần gán giá trị (xử lý dữ liệu thiếu)	<ul style="list-style-type: none"> <li>Cần gán giá trị cho dữ liệu thiếu (ví dụ: lượng mưa thiếu tại trạm đo).</li> <li>Phương pháp: trung bình, trung vị, hoặc mô hình dự đoán (KNN, Random Forest).</li> </ul>	<ul style="list-style-type: none"> <li>Ít cần gán giá trị thủ công.</li> <li>Mô hình sâu có thể xử lý dữ liệu thiếu qua các tầng mạng hoặc sử dụng dữ liệu thô (hình ảnh, radar) để bù đắp.</li> </ul>
Dữ liệu có phải là số không?	<ul style="list-style-type: none"> <li>Bắt buộc là số.</li> <li>Dữ liệu phi số (loại đất, vùng địa lý) cần mã hóa (One-Hot Encoding, Label Encoding).</li> </ul>	<ul style="list-style-type: none"> <li>Không bắt buộc là số.</li> <li>Hình ảnh vệ tinh, radar, hoặc video dòng chảy được xử lý trực tiếp và chuyển thành số trong quá trình huấn luyện.</li> </ul>
Xử lý dữ liệu tạp nham (noise)	<ul style="list-style-type: none"> <li>Nhạy cảm với nhiễu (ví dụ: dữ liệu mưa sai lệch từ cảm biến).</li> <li>Cần làm sạch dữ liệu (loại bỏ outliers, lọc nhiễu) trước khi huấn luyện.</li> </ul>	<ul style="list-style-type: none"> <li>Chịu nhiễu tốt hơn nhờ kiến trúc mạng sâu.</li> <li>Có thể học từ dữ liệu radar/hình ảnh có nhiễu, nhưng nhiễu quá lớn vẫn cần làm sạch cơ bản.</li> </ul>
Kích thước dữ liệu	<ul style="list-style-type: none"> <li>Hoạt động tốt với dữ liệu nhỏ/trung bình (ví dụ: vài nghìn bản ghi từ trạm đo mưa, mực nước).</li> <li>Hiệu quả phụ thuộc vào chất lượng đặc trưng thủ công.</li> </ul>	<ul style="list-style-type: none"> <li>Yêu cầu dữ liệu lớn (hàng chục nghìn hình ảnh vệ tinh, chuỗi thời gian dài).</li> <li>Dữ liệu nhỏ dễ gây overfitting, đặc biệt với mạng nơ-ron sâu.</li> </ul>
Loại bài toán chính	<ul style="list-style-type: none"> <li>Dự báo lũ quét (hồi quy: dự đoán mực nước, phân loại: nguy cơ lũ cao/thấp).</li> <li>Phản phẩm chính: dự báo mưa, phân tích rủi ro lũ.</li> <li>Phân tích cụm (clustering) vùng nguy cơ lũ.</li> </ul>	<ul style="list-style-type: none"> <li>Dự báo lũ quét phức tạp (dự đoán thời gian, vị trí lũ).</li> <li>Nhận diện vùng ngập từ hình ảnh vệ tinh.</li> <li>Phân tích chuỗi thời gian mưa/mực nước.</li> <li>Mô phỏng dòng chảy lũ bằng mạng nơ-ron.</li> </ul>
Công cụ xử lý dữ liệu	<ul style="list-style-type: none"> <li>Thư viện: Scikit-learn, Pandas, NumPy.</li> <li>Công cụ trích xuất đặc trưng: thống kê (tổng mưa, trung bình mực nước), GIS.</li> </ul>	<ul style="list-style-type: none"> <li>Thư viện: TensorFlow, PyTorch, Keras.</li> <li>Công cụ tiền xử lý: OpenCV (hình ảnh vệ tinh), xarray (dữ liệu radar), Hugging Face (NLP cho báo cáo lũ).</li> </ul>

Tiêu chí	Học máy (Machine Learning)	Học sâu (Deep Learning)
	- Làm sạch: lọc nhiễu, quy tắc kinh doanh.	- Tiền xử lý tập trung vào định dạng dữ liệu thô.
Ví dụ ứng dụng dữ liệu	- Dự báo nguy cơ lũ quét dựa trên dữ liệu bảng (lượng mưa, mực nước, độ dốc). - Phân loại vùng nguy cơ lũ (cao/thấp) dựa trên đặc trưng địa hình, đất đai.	- Dự báo lũ quét từ hình ảnh vệ tinh và radar thời tiết. - Phân tích video dòng chảy để phát hiện lũ quét. - Dự đoán thời gian lũ dựa trên chuỗi thời gian mưa/mực nước.

## 1.6. Đánh giá những hạn chế còn tồn tại và định hướng nghiên cứu

Mô hình trí tuệ nhân tạo và dữ liệu địa không gian là một sự kết hợp hoàn hảo để tận dụng được khả năng phán đoán nguy cơ lũ quét với dữ liệu lớn, mặc dù đã được triển khai trong nhiều năm gần đây, tuy nhiên các nghiên cứu vẫn còn tồn tại nhiều vấn đề chưa được giải quyết, dẫn đến khả năng ứng dụng trong thực tế còn gặp nhiều khó khăn.

Khó khăn thứ nhất là bản chất của dòng chảy lũ quét chưa được tiếp cận một cách đầy đủ. Lũ quét không thể sinh ra từ nội tại của một vị trí, mà nó được hình thành từ việc tập trung dòng chảy từ thượng nguồn, do đó, các nghiên cứu sử dụng giá trị nội tại của điểm làm đầu vào cho các mô hình để dự đoán nguy cơ lũ quét là chưa diễn đạt được bản chất của dòng chảy lũ quét. Do đó việc áp dụng vào để cảnh báo nguy cơ lũ quét có thể gây ra cảnh báo không do dữ liệu phản ánh chưa chính xác.

Khó khăn thứ hai là dữ liệu mưa được sử dụng. Theo đặc điểm của lũ quét là thời gian ngắn và có sự đột biến về lưu lượng, dữ liệu mưa giờ (hoặc 10 phút) nên được sử dụng để xây dựng bản đồ nguy cơ lũ quét thay vì dữ liệu mưa có độ phân giải cao hơn. Các dữ liệu mưa ngày chưa thể đưa ra được kết quả dự báo tốt cho loại hình thiên tai lũ quét. Bên cạnh đó, phân bố mưa gây ảnh hưởng lớn đến cường độ của lũ quét.

Khó khăn thứ ba là trạng thái độ ẩm thời đoạn trước chưa được xem xét. Rất ít nghiên cứu đưa dữ liệu độ ẩm đất theo thời gian vào để dự đoán nguy cơ lũ quét. Trong điều kiện độ ẩm bề mặt cao, khả năng sinh lũ quét cao hơn rất nhiều so với các điều kiện khác.

Các khó khăn trên cũng là các thách thức trong nghiên cứu lũ quét, nghiên cứu này với mục tiêu đánh giá khả năng ứng dụng công nghệ trí tuệ nhân tạo và dữ liệu địa không gian để nâng cao độ tin cậy trong phân vùng lũ quét quy mô cấp huyện. Cụ thể:

- Đánh giá được các loại dữ liệu địa không gian trong xác định nguy cơ lũ quét về: sự sẵn có của dữ liệu, độ chi tiết của dữ liệu, tiềm năng áp dụng.
- Đánh giá được sự phù hợp của các mô hình trí tuệ nhân tạo trong xác định nguy cơ lũ quét, ưu điểm, nhược điểm của các mô hình và đề xuất lựa chọn mô hình.
- Giải pháp khắc phục một số hạn chế còn tồn tại trong xác định nguy cơ lũ quét sử dụng mô hình trí tuệ nhân tạo và dữ liệu địa không gian nhằm tăng cường

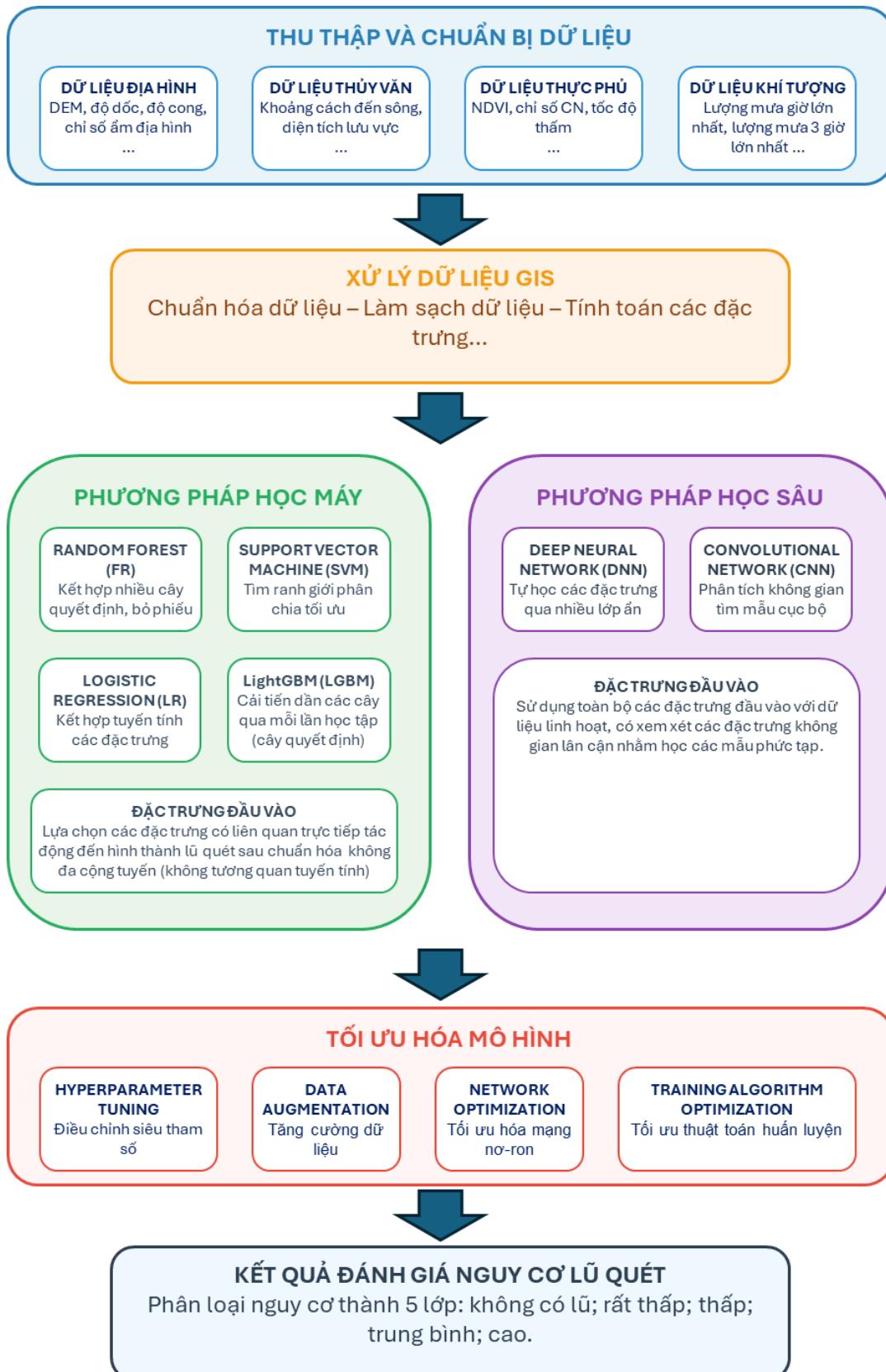
chất lượng phân vùng nguy cơ lũ quét, tiến tới xây dựng hệ thống cảnh báo nguy cơ lũ quét theo thời gian bằng mô hình trí tuệ nhân tạo.

### Nghiên cứu này hướng tới phạm vi quy mô cấp huyện do nhiều yếu tố:

- Phạm vi nghiên cứu lũ quét: lũ quét thường xảy ra ở các lưu vực nhỏ có diện tích trong khoảng từ 10 đến 100km<sup>2</sup>, ở quy mô cấp huyện có thể có vài chục lưu vực hợp thành và tập trung ở sông chính. Số lượng này đủ để mô hình trí tuệ nhân tạo có thể học được sự đa dạng về đặc trưng các lưu vực sinh lũ quét, từ lưu vực nhỏ (với đặc điểm thung lũng hẹp) đến trung bình (lòng chảo rộng hơn) trong cùng một phạm vi nghiên cứu. Điều mà phạm vi xã khó có thể đủ về mặt quy mô để cung cấp đầy đủ sự đa dạng của các lưu vực.
- Năng lực xử lý tính toán: Nghiên cứu tiềm năng không được đầu tư hệ thống máy chủ hiện đại với chi phí lớn, do đó việc xử lý dữ liệu không gian cho mô hình học máy, học sâu không thể thực hiện được đối với khu vực lớn ở quy mô cấp tỉnh hay quốc gia. Điều này ảnh hưởng trực tiếp đến giới hạn tính toán. Ở quy mô cấp tỉnh hay quốc gia, năng lực xử lý tính toán ở độ phân giải 30m có thể cần đến hệ thống máy chủ lớn và chuyên nghiệp.
- Mật độ quan trắc: ở quy mô cấp huyện, số lượng các trạm quan trắc mưa hiện nay có thể đủ để phản ánh được sự phân bố mưa, trong khi đó ở quy mô cấp xã hay nhỏ hơn, chỉ có thể một vài trạm quan trắc mưa được ghi nhận, dẫn đến khó nắm bắt được sự phân bố hay tác động hình thành đến lũ quét.
- Mục tiêu nghiên cứu tập trung vào việc đánh giá khả năng ứng dụng công nghệ AI và dữ liệu địa không gian trong phân vùng lũ quét, điều này không đòi hỏi phải thực hiện ở quy mô quá lớn như cấp tỉnh hay vùng. Bản chất của việc "đánh giá khả năng ứng dụng" là chứng minh tính khả thi, hiệu quả và độ tin cậy của phương pháp, chứ không phải đo lường quy mô triển khai tối đa. Tại quy mô cấp huyện, chúng ta có thể kiểm chứng đầy đủ các khía cạnh then chốt: khả năng xử lý dữ liệu đa nguồn, hiệu quả của các thuật toán machine learning/deep learning, độ chính xác của kết quả phân vùng, và tính ứng dụng thực tế của sản phẩm. Việc mở rộng lên quy mô lớn hơn không làm tăng giá trị khoa học của đánh giá mà có thể làm phức tạp hóa quá trình phân tích, gây khó khăn trong việc xác định chính xác nguồn gốc của sai sót và hạn chế. Hơn nữa, với quy mô huyện, chúng ta có thể thực hiện được đánh giá toàn diện từ khâu thu thập dữ liệu, xây dựng mô hình, đến kiểm chứng - tất cả trong phạm vi ngân sách và thời gian cho phép. Do đó, quy mô cấp huyện là lựa chọn hợp lý để hoàn thành mục tiêu đánh giá khả năng ứng dụng một cách khoa học và thuyết phục.

## CHƯƠNG 2. PHƯƠNG PHÁP PHÂN VÙNG LŨ QUÉT ỨNG DỤNG TRÍ TUỆ NHÂN TẠO VÀ DỮ LIỆU ĐỊA KHÔNG GIAN

### 2.1. Sơ đồ tiếp cận và phương pháp nghiên cứu



Hình 2-18. Sơ đồ tiếp cận trong nghiên cứu

Hình 2-18 thể hiện sơ đồ tiếp cận trong việc ứng dụng trí tuệ nhân tạo và dữ liệu địa không gian để phân vùng lũ quét. Bốn khía cạnh công việc được thể hiện bao gồm:

### **2.1.1 Phương pháp thu thập dữ liệu**

Phương pháp thu thập dữ liệu giúp nghiên cứu có đầy đủ bộ dữ liệu để phục vụ đánh giá và phân vùng lũ quét cho khu vực thí điểm (Mù Cang Chải), bên cạnh đó, công tác thu thập dữ liệu nghiên cứu trước đây sẽ góp phần làm tăng cường nhận thức và nâng cao hiểu biết về lũ quét, trí tuệ nhân tạo trong nghiên cứu lũ quét và cách thức triển khai thực hiện. Các dữ liệu thu thập dữ liệu được thể hiện chi tiết trong mục 2.2.1, và công tác chuẩn bị dữ liệu được thể hiện chi tiết trong mục 2.2.2.

### **2.1.2 Phương pháp nghiên cứu GIS**

Phương pháp nghiên cứu GIS là một tập hợp các phương pháp diễn toán các dữ liệu địa không gian nhằm xây dựng các bản đồ có nghĩa trong một lĩnh vực hay một phạm vi nào đó. Ví dụ đơn giản nhất là dữ liệu DEM địa hình (có thể được thu thập bằng khảo sát hoặc từ ảnh vệ tinh) là đầu vào để xác định các dữ liệu có liên quan bao gồm độ dốc, độ cong địa hình hoặc địa mạo...

Trong nghiên cứu về lũ quét, các loại dữ liệu địa không gian cơ bản bao gồm: (1) dữ liệu địa hình; (2) các dữ liệu ảnh vệ tinh; và (3) dữ liệu khí tượng thủy văn.

### **2.1.3 Phương pháp học máy**

#### **1. Lựa chọn mô hình, dữ liệu và phương pháp chung để chuẩn hóa dữ liệu**

Nghiên cứu sử dụng bốn mô hình học máy để đánh giá nguy cơ lũ quét dưới dạng bài toán phân loại: Rừng ngẫu nhiên (Random Forest - RF), Máy hỗ trợ vectơ (Support Vector Machine - SVM), Hồi quy Logistic (Logistic Regression - LR) và LightGBM - LGBM. Các mô hình này được chọn dựa trên khả năng xử lý dữ liệu phức tạp và phù hợp với đầu ra là một lớp về phân loại lũ.

#### **Các nhóm dữ liệu đầu vào bao gồm 4 nhóm chính:**

- Đặc trưng địa hình: bao gồm bản đồ mô hình số độ cao và các sản phẩm từ mô hình số độ cao như độ dốc, chỉ số ám địa hình, độ cong địa hình...
- Đặc trưng thủy văn: bao gồm khoảng cách/chênh lệch độ cao so với sông/suối, độ dốc lòng dẫn, chiều dài dòng chảy, diện tích lưu vực...
- Đặc trưng thực phủ: bao gồm các chỉ số liên quan đến bề mặt như NDVI (liên quan đến thảm phủ); chỉ số CN (liên quan đến thảm phủ và đặc trưng địa chất).
- Đặc trưng khí tượng: bao gồm các dữ liệu đặc trưng mưa như lượng mưa giờ lớn nhất, tổng lượng mưa 3, 6 giờ lớn nhất...

Chuẩn hóa đặc trưng: Thay vì sử dụng các đặc trưng định tính như loại đất hoặc loại hình sử dụng đất, nghiên cứu chọn các chỉ số định lượng như tốc độ thấm, NDVI,

và CN để đại diện cho đất và thảm phủ. Điều này giúp giảm thiểu việc mã hóa phức tạp và tăng tính chính xác.

Phương pháp chuẩn hóa: Với các đặc trưng định tính (nếu có), các phương pháp như trọng số dẫn chứng (WOE) hoặc tỷ lệ tàn suất (FR) có thể được sử dụng. Tuy nhiên, các phương pháp này yêu cầu dữ liệu lớn và đa dạng để đảm bảo độ tin cậy.

Làm sạch dữ liệu: Dữ liệu nhiễu (như các điểm không đại diện cho lũ quét) được loại bỏ trong quá trình tiền xử lý, đảm bảo chất lượng đầu vào.

Nghiên cứu xem lũ quét như phản ứng thủy văn của một lưu vực, tập trung vào các điểm ở cửa ra lưu vực thay vì các vị trí không đặc trưng (như mái nhà, đỉnh núi). Điều này phản ánh thực tế rằng lũ quét thường tích tụ ở thượng nguồn và gây hậu quả ở hạ du, giúp mô hình dự đoán chính xác hơn.

Không giống SVM và LR, RF và LGBM có khả năng xử lý dữ liệu định tính (nếu cần) mà không yêu cầu mã hóa phức tạp. Các đặc trưng định tính có thể được mã hóa bằng One-Hot Encoding hoặc mã hóa ngẫu nhiên. RF hoạt động dựa trên việc phân chia dữ liệu theo các đặc trưng tối ưu, tận dụng ý tưởng từ thống kê (như xác suất mẫu) nhưng được nâng cấp thông qua kỹ thuật lấy mẫu ngẫu nhiên và kết hợp nhiều cây quyết định.

Bằng việc kết hợp kiến thức thủy văn và học máy, nghiên cứu xây dựng một mô hình hồi quy đánh giá nguy cơ lũ quét phù hợp với các khu vực miền núi. Các đặc trưng được chọn đảm bảo tính khoa học, trong khi quá trình xử lý dữ liệu tăng cường độ tin cậy của mô hình.

## 2. Các mô hình sử dụng và nguyên tắc xác định lũ quét

### a. Rừng ngẫu nhiên (Random Forest - RF)

Rừng ngẫu nhiên là một mô hình kết hợp nhiều cây quyết định, mỗi cây đưa ra một dự đoán riêng, sau đó lấy trung bình để có kết quả cuối cùng. Mỗi cây quyết định hoạt động như một chuỗi câu hỏi dựa trên các đặc trưng dữ liệu.

Ví dụ: Một cây quyết định có thể hỏi:

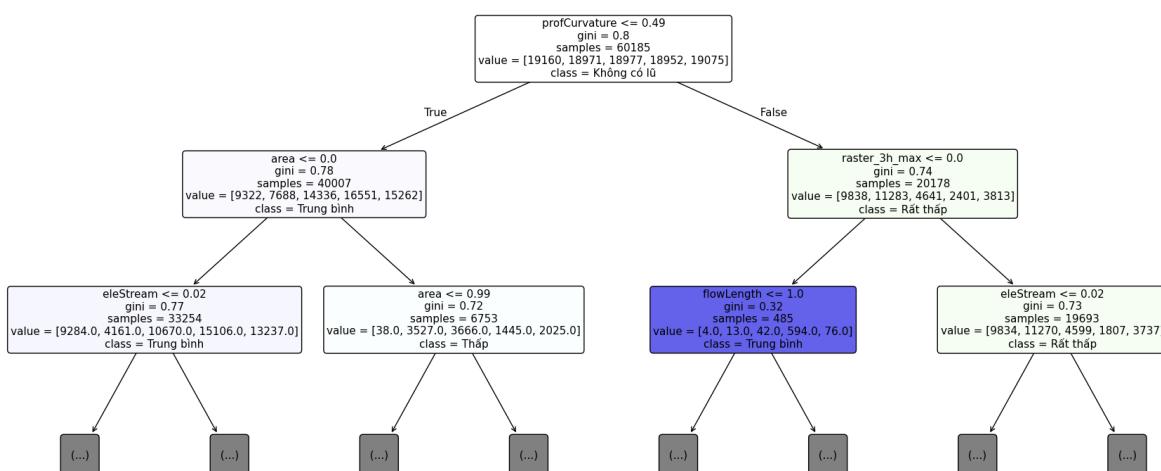
- Lượng mưa ngày lớn nhất có vượt quá 100 mm không?
- Nếu có, độ dốc bình quân lưu vực có lớn hơn 20 độ không?
- Nếu cả hai đều đúng, dự đoán nguy cơ lũ quét là 80/100.

Rừng ngẫu nhiên tạo ra hàng trăm cây như vậy, mỗi cây sử dụng một phần dữ liệu ngẫu nhiên (như lượng mưa, NDVI, hoặc tốc độ thẩm thấu), rồi kết hợp kết quả để tăng độ chính xác.

- **Nguyên tắc dự đoán nguy cơ lũ quét**

- Phân tích nhiễu đặc trưng: Mô hình xem xét đồng thời toàn bộ đặc trưng (lượng mưa, độ dốc, NDVI, v.v.) và tự động xác định đặc trưng nào quan trọng nhất (ví dụ: lượng mưa lớn thường làm tăng nguy cơ lũ).
- Kết hợp dự đoán: Bằng cách lấy trung bình từ nhiều cây quyết định, mô hình giảm thiểu sai sót và đưa ra dự đoán đáng tin cậy hơn.
- Đầu ra số: Kết quả là một giá trị từ 0 đến 100, biểu thị mức độ nguy cơ lũ quét tại cửa ra cửa lưu vực.

Rừng ngẫu nhiên rất hiệu quả khi xử lý dữ liệu phức tạp, như các yếu tố thủy văn (mưa, độ dốc, thảm phủ). Nó có thể phát hiện các mối quan hệ không đơn giản, ví dụ: nguy cơ lũ tăng mạnh khi lượng mưa lớn kết hợp với tốc độ thẩm thấu thấp. Tuy nhiên, mô hình này khó giải thích chi tiết lý do đưa ra một con số cụ thể, vì nó dựa trên nhiều cây quyết định.



Hình 2-19. Hình mô tả sự phân loại cho 5 lớp (cây có độ sâu là 2)

Ví dụ trong một cây được thể hiện như hình trên (với dữ liệu đã được chuẩn hóa về khoảng từ -1 đến 1), tại nút gốc (trên cùng), cây này sẽ xem xét độ cong theo phương dốc của địa hình và so sánh với ngưỡng 0,49. Khi đó, số lượng mẫu phân loại cho 5 lớp được thể hiện trong value (có hơn 19 nghìn mẫu ở lớp 0 – không có lũ; gần 19 nghìn mẫu ở lớp 1 – nguy cơ rất thấp; gần 19 nghìn mẫu ở lớp 2 – nguy cơ thấp....). Số lượng mẫu ở lớp 0 – không có lũ là lớn nhất, do đó tại nút này sẽ phân loại là không có lũ. Tuy nhiên, phân loại này bị phiến diện do các lớp khác cũng có số lượng tương tự, không quá chênh lệch. Do đó, tiếp tục xét đến tầng thứ hai, tại tầng này, nếu độ cong địa hình theo phương dốc nhỏ hơn 0,49, mô hình sẽ xét tiếp diện tích lưu vực và nếu độ cong lớn hơn 0,49, mô hình sẽ xét tổng lượng mưa 3 giờ lớn nhất. Cứ như vậy cho đến khi kết thúc. Một mô hình thường có rất nhiều cây (nên gọi là rừng), số lượng cây trong mô hình được thể hiện qua tham số n\_estimators, trong khi đó, số tầng (hay còn gọi là độ sâu cây) được thể hiện qua tham số max\_depth. Các cây khác nhau có thuật toán phân

loại khác nhau. Mỗi bộ dữ liệu đầu vào sẽ được từng cây đánh giá, sau đó cho ra kết quả dưới dạng bô phiếu. Tỷ lệ các cây bô phiếu cho một lớp nào đó càng cao thì dữ liệu tương ứng sẽ được phân vào lớp đó.

#### b. Máy hỗ trợ vectơ (Support Vector Machine - SVM)

Máy hỗ trợ vectơ tìm cách phân biệt các mức độ nguy cơ lũ quét bằng cách tạo ra một ranh giới trong không gian dữ liệu. Không gian này bao gồm các đặc trưng như lượng mưa, độ dốc, và chỉ số CN.

Ví dụ: SVM có thể xác định rằng các lưu vực với lượng mưa ngày lớn hơn 120 mm và độ dốc trên 25 độ có nguy cơ lũ cao (ví dụ: 85/100), trong khi các lưu vực có mưa dưới 50 mm và độ dốc thấp có nguy cơ thấp (ví dụ: 20/100). Mô hình sẽ tìm ranh giới tối ưu để phân chia các mức nguy cơ khác nhau và dự đoán một giá trị số.

- **Nguyên tắc dự đoán nguy cơ lũ quét**
- Xác định ranh giới: SVM tạo ra một ranh giới (hay mặt phẳng) trong không gian dữ liệu, sao cho các lưu vực có nguy cơ lũ cao và thấp được tách biệt rõ ràng nhất.
- Dự đoán mức độ nguy cơ: Dựa trên vị trí của một lưu vực (được xác định bởi các đặc trưng), SVM tính toán một giá trị nguy cơ từ 0 đến 100.
- Xử lý mẫu phức tạp: SVM có thể nhận diện các mối quan hệ phức tạp giữa các đặc trưng (như lượng mưa và NDVI) bằng cách sử dụng kỹ thuật biến đổi dữ liệu.

SVM có sự phù hợp với bài toán lũ quét vì nó có thể xử lý nhiều đặc trưng cùng lúc và tìm ra các mẫu nguy cơ dựa trên dữ liệu thủy văn. Mô hình này đặc biệt hiệu quả khi dữ liệu được chuẩn hóa tốt (ví dụ: lượng mưa và độ dốc được đưa về cùng thang đo). Tuy nhiên, SVM có thể chậm khi xử lý lượng dữ liệu lớn và yêu cầu điều chỉnh tham số cẩn thận, bên cạnh đó vấn đề phi tuyến trong dự đoán lũ quét có thể sẽ làm khó cho mô hình này mặc dù mô hình có thể đưa ra các mặt phẳng cong trong ranh giới phân chia.

#### c. Hồi quy logistic (Logistic Regression - LR)

Hồi quy Logistic là một mô hình đơn giản, sử dụng một công thức để kết hợp các đặc trưng và dự đoán nguy cơ lũ quét. Mỗi đặc trưng (như lượng mưa, độ dốc, NDVI) được gán một trọng số, thể hiện mức độ ảnh hưởng của nó.

Ví dụ: Nếu lượng mưa ngày lớn hơn 100 mm có trọng số +4, độ dốc trên 20 độ có trọng số +2, và chỉ số NDVI thấp có trọng số +1, mô hình sẽ tính tổng các trọng số này và đưa ra một con số nguy cơ (ví dụ: 70/100).

- **Nguyên tắc dự đoán nguy cơ lũ quét**

- Kết hợp đặc trưng: Mô hình gán trọng số cho từng đặc trưng (ví dụ: mưa lớn có trọng số cao hơn diện tích lưu vực) và tính toán tổng để dự đoán nguy cơ.
- Dự đoán số liệu: Kết quả là một giá trị từ 0 đến 100, được điều chỉnh để dễ hiểu, thể hiện mức độ nguy cơ lũ quét.
- Giả định tuyến tính: Mô hình cho rằng các đặc trưng ảnh hưởng đến nguy cơ lũ theo cách đơn giản (ví dụ: mưa càng lớn, nguy cơ càng cao).

Hồi quy Logistic dễ triển khai, nhanh, và dễ giải thích, giúp xác định đặc trưng nào (như lượng mưa hay độ dốc) ảnh hưởng mạnh nhất đến lũ quét. Tuy nhiên, mô hình này có thể không hiệu quả nếu các đặc trưng có mối quan hệ phức tạp, ví dụ: nguy cơ lũ chỉ tăng khi lượng mưa lớn kết hợp với tốc độ thẩm thấu thấp, đặc biệt là các quan hệ phi tuyến của các dữ liệu.

#### d. LightGBM (Light Gradient Boosting Machine - LGBM)

LightGBM (Light Gradient Boosting Machine) là một mô hình học máy tiên tiến dựa trên gradient boosting, được thiết kế để xử lý dữ liệu lớn và phức tạp với tốc độ cao và hiệu quả. Mô hình này xây dựng nhiều cây quyết định liên tiếp, trong đó mỗi cây cố gắng sửa lỗi của các cây trước đó, từ đó cải thiện dự đoán nguy cơ lũ quét.

Ví dụ LightGBM có thể phân tích các đặc trưng như lượng mưa ngày lớn nhất, độ dốc lưu vực, chỉ số NDVI, và tốc độ thẩm đất. Nếu lượng mưa vượt quá 100 mm và độ dốc lớn hơn 20 độ, mô hình có thể dự đoán nguy cơ lũ quét cao (ví dụ: 85/100). Mỗi cây trong LightGBM học từ sai số của cây trước, giúp dự đoán chính xác hơn.

- **Nguyên tắc dự đoán nguy cơ lũ quét**

- Xây dựng tuần tự: LightGBM tạo ra các cây quyết định tuần tự, trong đó mỗi cây tập trung vào việc sửa lỗi dự đoán của các cây trước đó bằng cách sử dụng gradient của hàm mất mát.
- Tối ưu hóa hiệu quả: Mô hình sử dụng kỹ thuật "histogram-based" để nhóm dữ liệu thành các bin, giảm thời gian tính toán và tăng tốc độ xử lý, đặc biệt với dữ liệu lớn.
- Dự đoán số liệu: Kết quả là một giá trị từ 0 đến 100, biểu thị mức độ nguy cơ lũ quét, được tính dựa trên sự kết hợp của các cây trong mô hình.
- Xử lý đặc trưng quan trọng: LightGBM tự động xác định và ưu tiên các đặc trưng quan trọng (ví dụ: lượng mưa lớn hoặc độ dốc cao) để cải thiện độ chính xác.

LightGBM đặc biệt hiệu quả trong bài toán dự đoán nguy cơ lũ quét nhờ khả năng xử lý nhanh dữ liệu thủy văn phức tạp và phát hiện các mối quan hệ phi tuyến giữa các đặc trưng, chẳng hạn như sự kết hợp giữa lượng mưa cao và tốc độ thẩm thấu thấp. Mô hình này cũng hỗ trợ dữ liệu lớn và có thể được điều chỉnh thông qua các tham số như

số lượng cây (n\_estimators), độ sâu cây (max\_depth), và tỷ lệ học (learning\_rate). Tuy nhiên, LightGBM có thể yêu cầu điều chỉnh tham số cẩn thận để tránh hiện tượng quá khớp (overfitting) và đôi khi khó giải thích chi tiết lý do đưa ra dự đoán cụ thể.

#### 2.1.4 Phương pháp học sâu

##### 1. Lựa chọn mô hình, dữ liệu và phương pháp chuẩn hóa dữ liệu

Đối với phương pháp học sâu, nghiên cứu sử dụng 3 mô hình bao gồm Mạng nơ-ron sâu (Deep Neural Network - DNN), Mạng nơ-ron tích chập (Convolutional Neural Network - CNN), và Mạng nơ-ron hồi tiếp dài ngắn hạn (Long Short-Term Memory - LSTM) để đánh giá lũ quét cho khu vực nghiên cứu. Khác với các mô hình học máy, các dữ liệu đầu vào của mô hình học sâu đa dạng và linh hoạt hơn. Các yếu tố dữ liệu như cao độ (không thể hiện xu hướng lũ quét) cũng có thể được sử dụng để làm đầu vào cho mô hình.

##### 2. Các mô hình sử dụng và nguyên tắc xác định lũ quét

Nghiên cứu sử dụng ba mô hình học sâu để dự đoán nguy cơ lũ quét dưới dạng bài toán phân loại: DNN, CNN và LSTM. Mỗi mô hình dựa vào các tham số đầu vào sẽ dự đoán nhãn phân loại lũ ở đầu ra dựa trên dữ liệu đào tạo và kiểm tra. Các mô hình tận dụng các đặc trưng định lượng, bao gồm lượng mưa, độ dốc lưu vực, chỉ số NDVI, và các yếu tố thủy văn khác, để phân tích và dự đoán. Dưới đây là mô tả chi tiết về từng mô hình, cách chúng hoạt động, và lý do chúng phù hợp với bài toán lũ quét.

###### a. Mạng nơ-ron tích chập (Convolutional Neural Network - CNN)

Mạng nơ-ron tích chập được thiết kế để phân tích dữ liệu có cấu trúc không gian, như hình ảnh, nhưng cũng có thể áp dụng cho dữ liệu số nếu được tổ chức phù hợp. Trong bài toán lũ quét, CNN xem các đặc trưng như lượng mưa, độ dốc, và NDVI như một "bản đồ" của lưu vực, tìm ra các mẫu liên quan đến nguy cơ lũ.

Ví dụ: Nếu tổ chức dữ liệu thành một bảng, với các cột là lượng mưa, độ dốc, và NDVI tại các điểm trong lưu vực, CNN sẽ phân tích bảng này để tìm các khu vực có nguy cơ cao (như nơi mưa lớn và độ dốc cao xuất hiện cùng nhau).

- **Nguyên tắc dự đoán nguy cơ lũ quét**
- Phân tích cục bộ: CNN xem xét các đặc trưng (như lượng mưa, độ dốc, chỉ số CN) trong các nhóm nhỏ, ví dụ: kiểm tra xem mưa lớn có đi kèm độ dốc cao ở một khu vực cụ thể không.
- Tìm mẫu phức tạp: Qua các lớp tích chập, CNN nhận diện các mẫu lớn hơn, như sự kết hợp của lượng mưa giờ lớn nhất và tốc độ thẩm thấu thấp, để đánh giá nguy cơ lũ.
- Dự đoán số liệu: Kết quả là một giá trị từ 0 đến 100, thể hiện nguy cơ lũ quét, ví dụ: 90/100 cho lưu vực có nhiều khu vực với mưa lớn và địa hình bất lợi.

CNN phù hợp khi dữ liệu lũ quét được tổ chức thành các cấu trúc không gian, ví dụ: các giá trị lượng mưa và NDVI tại nhiều điểm trong lưu vực. Mô hình tìm ra các mẫu cục bộ, như khu vực có mưa lớn gần sông suối, giúp dự đoán chính xác hơn. Tuy nhiên, CNN cần dữ liệu lớn và định dạng phù hợp, đồng thời tốn thời gian huấn luyện hơn so với các mô hình đơn giản.

#### b. Mạng nơ-ron sâu (Deep Neural Network - DNN)

Mạng nơ-ron sâu giống như một hệ thống xử lý thông tin, lấy dữ liệu đầu vào (như lượng mưa, độ dốc) và biến đổi chúng qua nhiều lớp (layers) để đưa ra dự đoán. Mỗi lớp giống như một bước xử lý, tìm ra các mối quan hệ giữa các đặc trưng.

Ví dụ: Khi nhận dữ liệu về lượng mưa ngày (150 mm), độ dốc lưu vực (25 độ), và chỉ số NDVI (0.3), DNN phân tích các con số này qua các lớp để xác định mức độ nguy cơ lũ. Nó tự động học cách kết hợp các đặc trưng, chẳng hạn nhận ra rằng mưa lớn và độ dốc cao thường dẫn đến nguy cơ lũ cao.

- **Nguyên tắc dự đoán nguy cơ lũ quét**
- Xử lý đồng thời các đặc trưng: DNN nhận tất cả các đặc trưng (lượng mưa, độ dốc, NDVI, v.v.) và biến đổi chúng qua nhiều lớp. Mỗi lớp tìm ra các mẫu phức tạp, ví dụ: lượng mưa lớn kết hợp với tốc độ thẩm thấu thấp làm tăng nguy cơ lũ.
- Tự động học đặc trưng: Không cần con người chọn yếu tố nào quan trọng, DNN tự tìm ra cách các đặc trưng ảnh hưởng đến lũ quét qua quá trình huấn luyện.
- Dự đoán số liệu: Kết quả là một giá trị từ 0 đến 100, thể hiện nguy cơ lũ quét tại cửa ra của lưu vực, ví dụ: 85/100 cho khu vực có mưa lớn và địa hình dốc.

DNN phù hợp với bài toán lũ quét vì nó có thể xử lý nhiều đặc trưng cùng lúc và tìm ra các mối quan hệ phức tạp mà các mô hình đơn giản có thể bỏ sót. Ví dụ, nó có thể nhận ra rằng nguy cơ lũ tăng không chỉ do mưa lớn mà còn do sự kết hợp với chỉ số CN cao và độ cao tương đối thấp. Tuy nhiên, DNN cần nhiều dữ liệu để học hiệu quả và có thể khó giải thích chi tiết tại sao đưa ra một con số cụ thể.

#### c. Mạng nơ-ron hồi tiếp dài ngắn hạn (Long Short-Term Memory - LSTM)

Mạng LSTM là một loại mạng nơ-ron chuyên xử lý dữ liệu theo thời gian, như các chuỗi số thay đổi qua từng giờ hoặc ngày. Trong bài toán lũ quét, LSTM phân tích các đặc trưng như lượng mưa theo thời gian để dự đoán nguy cơ lũ.

Ví dụ: Nếu lượng mưa giờ lớn nhất tăng từ 10 mm lên 50 mm trong 3 giờ, đồng thời chỉ số NDVI thấp (thảm phủ kém), LSTM sẽ xem xét xu hướng này để dự đoán nguy cơ lũ quét tăng cao.

- **Nguyên tắc dự đoán nguy cơ lũ quét**

- Phân tích theo thời gian: LSTM xem xét các đặc trưng như lượng mưa giờ hoặc ngày theo thứ tự thời gian, nhận ra xu hướng dần đến lũ quét, ví dụ: mưa lớn kéo dài làm đất bão hòa.
- Ghi nhớ dài hạn: Mô hình "nhớ" các sự kiện trước đó, như lượng mưa tích lũy trong 24 giờ, để đánh giá tác động của chúng đến nguy cơ lũ quét hiện tại.
- Dự đoán số liệu: Kết quả là một giá trị từ 0 đến 100, thể hiện nguy cơ lũ quét, ví dụ: 80/100 nếu mưa lớn kéo dài kết hợp với độ dốc lưu vực cao.

LSTM rất hiệu quả khi nguy cơ lũ quét phụ thuộc vào dữ liệu thời gian, như lượng mưa tăng dần hoặc sự thay đổi của tốc độ thẩm đất qua các trận mưa. Mô hình có thể dự đoán chính xác hơn bằng cách xem xét lịch sử dữ liệu, ví dụ: một lưu vực nhận mưa lớn liên tục trong 24 giờ có nguy cơ lũ cao hơn so với mưa lớn trong 1 giờ. Tuy nhiên, LSTM phức tạp và yêu cầu dữ liệu thời gian chi tiết, cùng với thời gian huấn luyện dài.

### **2.1.5 Một số kỹ thuật tối ưu hóa mô hình trí tuệ nhân tạo**

Về cơ bản, kỹ thuật tối ưu nhằm tìm kiếm các tham số/thuật toán để giảm thiểu các chi phí đào tạo mô hình và tăng cường chất lượng dự đoán.

#### *2.1.5.1 Tinh chỉnh siêu tham số (Hyperparameter tuning)*

Siêu tham số là các tham số không được học trong quá trình huấn luyện, như tỷ lệ học (learning rate), số lượng lớp ẩn (hidden layer size), số lượng epochs, và batch size. Tinh chỉnh siêu tham số là quá trình tìm kiếm và chọn giá trị tốt nhất cho các siêu tham số này để cải thiện hiệu suất mô hình.

#### *2.1.5.2 Tăng cường dữ liệu (Data augmentation)*

Tăng cường dữ liệu là quá trình tạo ra thêm dữ liệu huấn luyện bằng cách áp dụng các biến đổi nhỏ lên dữ liệu hiện có, như xoay, thu phóng, cắt tia, lật ngang, hay thay đổi ánh sáng. Việc tăng cường dữ liệu giúp mô hình học được các biến thể khác nhau của dữ liệu và giảm hiện tượng quá khớp (overfitting).

#### *2.1.5.3 Tối ưu hóa mạng nơ-ron (Network optimization)*

Cải thiện cấu trúc mạng nơ-ron bằng cách thay đổi số lượng lớp, kích thước lớp, số lượng tham số, và các thay đổi khác để cải thiện hiệu suất. Có thể sử dụng các kiến trúc mạng nơ-ron sâu (deep neural networks) như Convolutional Neural Networks (CNN) hay Recurrent Neural Networks (RNN) để nâng cao hiệu suất mô hình.

#### *2.1.5.4 Tối ưu hóa thuật toán huấn luyện (Training algorithm optimization)*

Cải tiến thuật toán huấn luyện bằng cách sử dụng các phương pháp tối ưu hóa như Adam, RMSprop, hoặc Stochastic Gradient Descent (SGD) với các biến thể như Momentum hay Nesterov. Cũng có thể sử dụng kỹ thuật regularization như L1 regularization hoặc L2 regularization để giảm hiện tượng quá khớp.

## 2.2. Dữ liệu sử dụng

### 2.2.1 Dữ liệu thu thập

Có 05 nhóm dữ liệu ở dữ liệu thô được thu thập bao gồm: (1) Địa hình; (2) sử dụng đất – thảm phủ; (3) loại đất; (4) dữ liệu vệ tinh quan trắc bề mặt; và (5) dữ liệu mưa. Mỗi nhóm dữ liệu này có thể tạo ra các loại dữ liệu khác nhau phục vụ phân tích, đánh giá (VD như dữ liệu địa hình có thể tạo bản đồ độ dốc, độ cong địa hình...; dữ liệu vệ tinh có thể xác định chỉ số NDVI và các chỉ số khác hay dữ liệu mưa vệ tinh ...).

Bên cạnh 05 nhóm dữ liệu cơ bản, dữ liệu về sự kiện lũ quét cũng được thu thập nhằm đánh giá, kiểm chứng phương pháp đánh giá lũ quét. Toàn bộ dữ liệu thu thập sẽ được tổng hợp, phân tích và lựa chọn các loại dữ liệu đầu vào cho mô hình trí tuệ nhân tạo trong xác định lũ quét.

Tổng hợp của 05 nhóm dữ liệu này sẽ thông qua quá trình sàng lọc, đánh giá và lựa chọn để đưa ra 04 nhóm dữ liệu dạng số có liên quan trực tiếp tác động đến phân vùng lũ quét bao gồm: (1) Địa hình; (2) Thủy văn; (3) Thực phủ; và (4) Khí tượng.

#### 1. Địa hình

##### a. Ý nghĩa của dữ liệu địa hình trong nghiên cứu lũ quét

Địa hình (hay còn gọi là mô hình số độ cao) là dữ liệu cơ bản trong phân tích thủy văn GIS. Lượng mưa rơi xuống bề mặt sẽ đi theo hướng dòng chảy để đến sông, suối và chảy về hạ du. Do đó, địa hình là một trong những dữ liệu quan trọng trong việc xác định nguy cơ lũ quét. Trong nghiên cứu này, mô hình số độ cao sẽ là dữ liệu cơ bản để xác định: (1) Độ dốc địa hình; (2) Độ dốc lòng dẫn; (3) Chiều dài lòng dẫn; (4) Diện tích lưu vực; (5) Khoảng cách đến sông suối; (6) Độ cao tương đối đến sông suối gần nhất; và (7) Chỉ số ẩm địa hình. Toàn bộ các yếu tố trên đây được xác định trực tiếp từ địa hình hoặc các sản phẩm từ bản đồ địa hình.

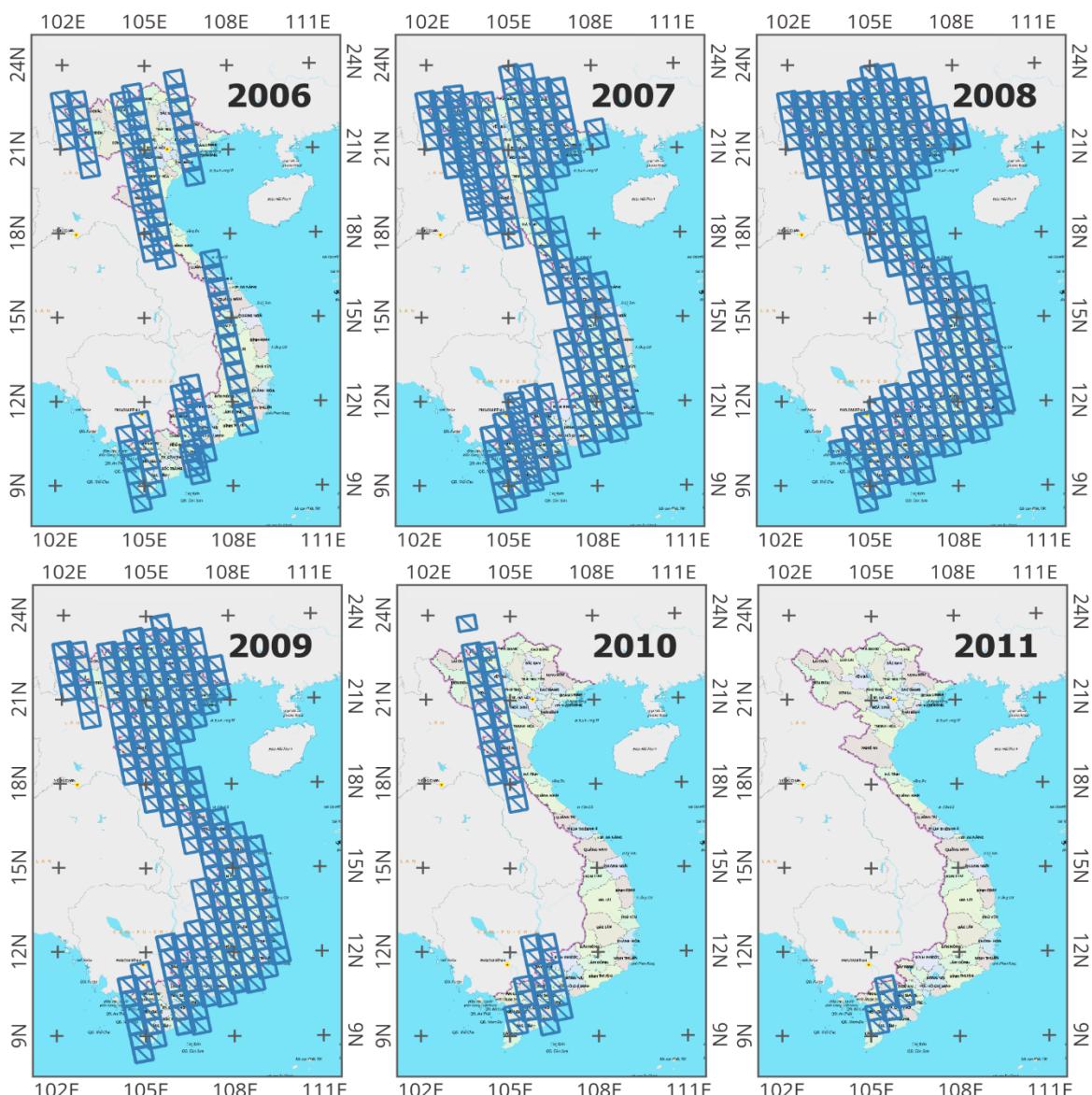
##### b. Các nguồn dữ liệu sẵn có và lựa chọn dữ liệu sử dụng

Hiện nay các nguồn dữ liệu về địa hình chủ yếu được thu thập bởi các nguồn sau:

- Khảo sát: kết quả khảo sát địa hình là một kết quả đáng tin cậy và tốn kém, đây được coi là cách tốt nhất để có được dữ liệu địa hình (với số liệu mới nhất) phục vụ đa mục tiêu.
- Mua bản đồ số độ cao: Việc mua bản đồ số độ cao cũng đáng tin cậy như công tác khảo sát và có độ tốn kém ít hơn. Tuy nhiên, số liệu bản đồ số độ cao ở Việt Nam do cục đo đạc và bản đồ cấp phần lớn có thời gian từ năm 2010. Ở Mù Cang Chải, dữ liệu này cũng sẵn có năm 2010.
- Dữ liệu DEM vệ tinh miễn phí: bao gồm 2 nguồn phổ biến là nguồn ảnh ALOS (band L) và Sentinel (band 1C). ALOS sử dụng band L để xác định cao độ bề mặt, band L có khả năng xuyên mây rất tốt và ít bị tác động bởi khí quyển, cây

cối, do đó có độ tin cậy cao hơn band C (xuyên mây kém hơn và bị chịu tác động bởi các tầng phủ dày). Ở khu vực Mù Cang Chải, dữ liệu DEM do ALOS cung cấp miễn phí có ở 2 độ phân giải là 30m và 12,5m (năm 2010). Band 1C của sentinel là dữ liệu miễn phí thường xuyên được cung cấp, do đó có thể có được địa hình mới nhất từ nguồn dữ liệu này (khoảng 15 ngày có một chu kỳ ảnh), trong khi đó dữ liệu ALOS có chu kỳ ảnh là khoảng 45 ngày (dữ liệu trả phí).

Trong nghiên cứu này, nhóm nghiên cứu sử dụng DEM vệ tinh miễn phí nguồn ảnh ALOS ở độ phân giải 12,5m chụp năm 2006-2011 để phục vụ công tác nghiên cứu do là nguồn dữ liệu đáng tin cậy, được sử dụng rộng rãi trong các nghiên cứu trong nước và quốc tế. Bên cạnh đó, yếu tố địa hình với độ phân giải 12,5m được đánh giá khá chi tiết cho công tác nghiên cứu lũ quét, thông thường các nghiên cứu công bố trên thế giới sử dụng bản đồ ở độ phân giải 30m cho các nghiên cứu về lũ quét.

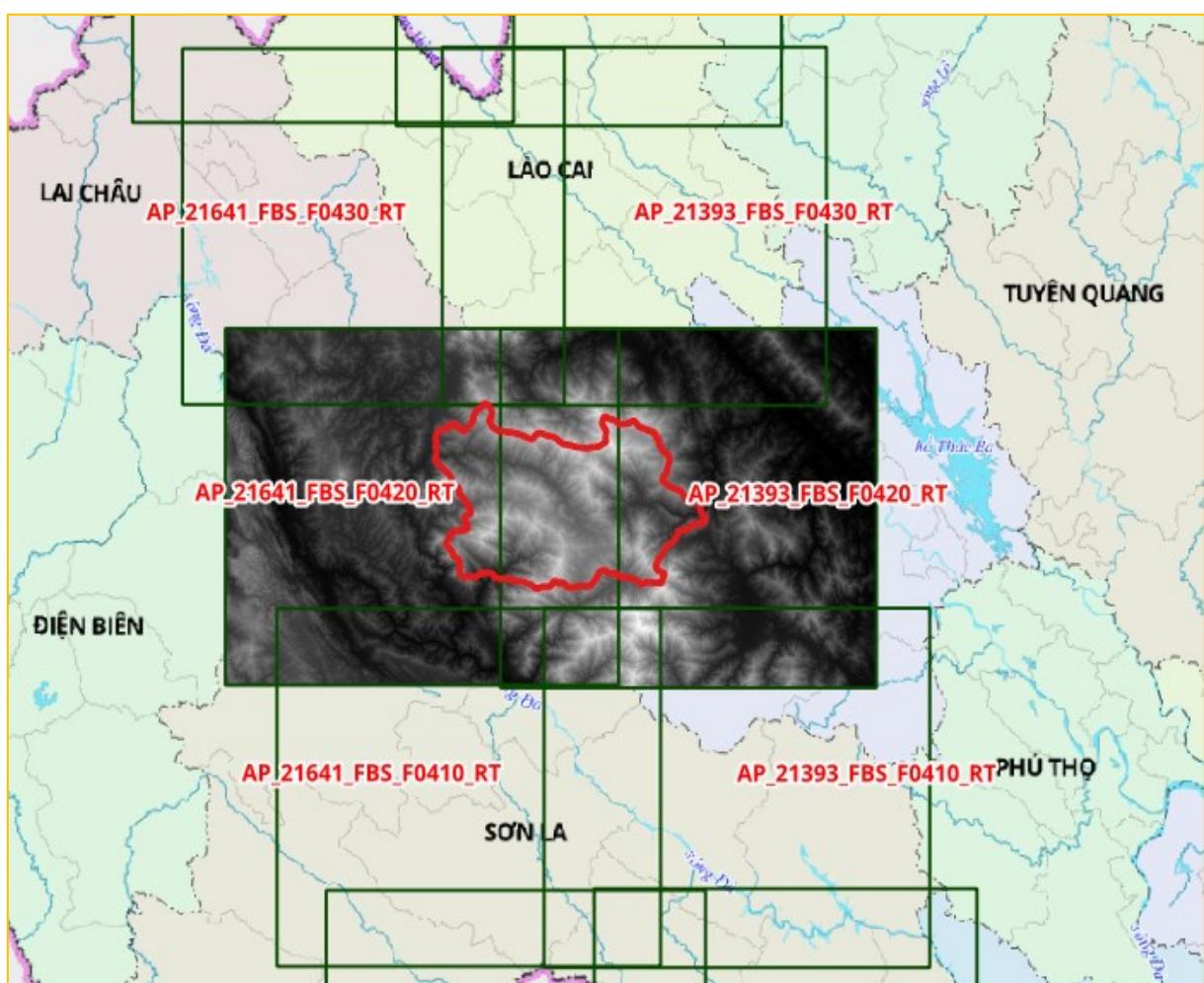


Hình 2-20. Sơ đồ phân bố các mảnh bản đồ địa hình độ phân giải 12,5m của ALOS

Trên lãnh thổ Việt Nam, DEM ở độ phân giải 12,5m bao trùm toàn bộ lãnh thổ không đồng nhất về mặt thời gian. Bắt đầu từ năm 2006, sứ mệnh vệ tinh ALOS đã cung cấp các dữ liệu địa hình toàn cầu, tuy nhiên, các dữ liệu ảnh chỉ được công bố theo các mảnh ở dạng miễn phí. Riêng năm 2008, toàn bộ Việt Nam được công bố ảnh miễn phí trên toàn lãnh thổ.

Việc lựa chọn dữ liệu địa hình ALOS có độ bao phủ rộng không chỉ giúp tăng cường sự đồng nhất của dữ liệu địa hình, mà còn nâng cao khả năng mở rộng mô hình dự báo trong tương lai cho các khu vực ngoài Mù Cang Chải. Toàn bộ dữ liệu ALOS tại Việt Nam cung cấp miễn phí ở phân cực đơn HH. Đây là phân cực có độ tin cậy cao trong việc phát hiện các bề mặt phẳng hoặc cấu trúc nhân tạo (đô thị, đường xá). Với băng tần L, HH có khả năng xuyên qua thảm thực vật mỏng để phản xạ từ mặt đất, đặc biệt hữu ích trong xây dựng DEM địa hình.

Huyện Mù Cang Chải có dữ liệu địa hình theo ALOS đầy đủ từ năm 2007-2010 với dữ liệu gần nhất là năm 2010 thuộc các mảnh ghép có số hiệu là AP\_21641\_FBS\_F0420\_RT1 và AP\_21393\_FBS\_F0420\_RT1. Hai mảnh ghép này mỗi mảnh bao phủ một phần huyện Mù Cang Chải.



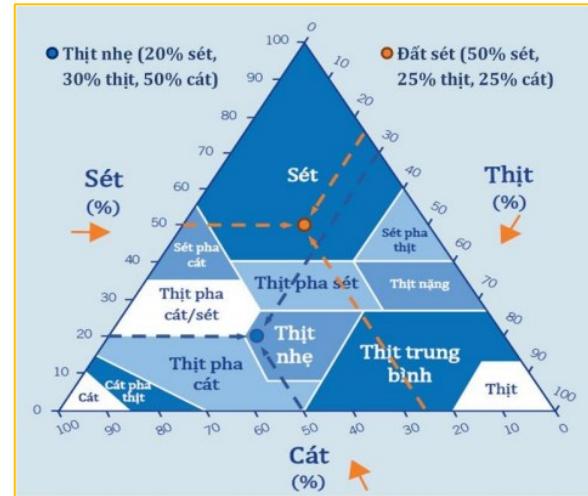
Hình 2-21. Bản đồ DEM địa hình 12,5x12,5m từ ALOS

Dựa trên thông tin thuộc tính hình ảnh, ảnh bên trái được thu nhận thời điểm ngày 15/02/2010 và ảnh bên phải được thu nhận ngày 29/01/2010 ở phân cực HH sử dụng băng tần L-Band. Nghiên cứu cũng thu thập bản đồ địa hình tỷ lệ 1/10.000 tại khu vực huyện Mù Cang Chải, tuy nhiên để đồng nhất dữ liệu địa hình và tăng cường khả năng mở rộng mô hình dự đoán nguy cơ lũ quét trong tương lai, nghiên cứu lựa chọn dữ liệu địa hình của ALOS như đã trình bày ở trên.

## 2. Đất và thuộc tính

Đất trong nghiên cứu lũ quét là một trong những dữ liệu quan trọng, trọng tâm của lũ quét ngoài ở lượng mưa còn nằm ở khả năng hấp thụ nước của đất. Khi cường độ mưa vượt quá tốc độ thấm, lượng nước dư thừa sẽ chuyển thành dòng chảy mặt. Mỗi loại đất khác nhau có tốc độ thấm khác nhau dẫn đến phản ứng với cùng một lượng mưa rơi xuống là khác nhau.

Thành phần của đất quyết định khả năng hấp thụ nước và trữ nước. Toàn bộ các loại đất được cấu tạo chủ yếu bởi 3 thành phần là Sét (clay), Thịt (silt) và Cát (sand). Sét có đặc điểm là tốc độ thấm chậm và khả năng giữ nước cao, và cát thì ngược lại, tốc độ thấm nhanh trong khi giữ nước thấp. Thịt có đặc điểm cân bằng trong thấm và giữ nước. Bộ Nông nghiệp Hoa Kỳ (USDA) cung cấp bản đồ cấu trúc đất ở độ phân giải 250m toàn cầu ở 6 độ sâu đất (0, 10, 30, 60, 100 và 200 cm) với 12 mã, chi tiết thể hiện như sau:



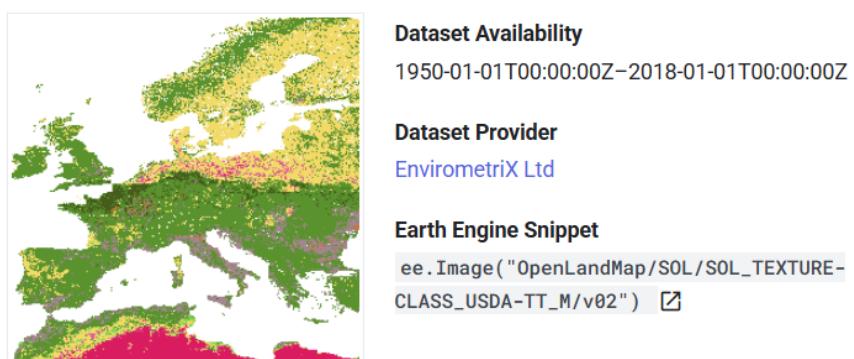
Bảng 2-11. Thuộc tính các loại đất trong bản đồ kết cấu đất (MDE, n.d.; NRCS, n.d.; Rahmati, Mehdi, et al., 2018)

TT	Mã	Loại đất	Định tính		Định lượng	
			Tốc độ thấm nước	Khả năng trữ nước	Tốc độ thấm (mm/giờ)	Mức giữ nước (mm/m)
1	Cl	Sét	Rất chậm	Rất cao	0,1–0,3	180–250
2	SiCl	Sét pha thịt	Rất chậm	Rất cao	0,2–0,5	180–250
3	SaCl	Sét pha cát	Rất chậm	Trung bình	0,3–0,8	150–200
4	ClLo	Thịt pha sét	Rất chậm	Cao	0,8–1,5	160–220
5	SiClLo	Thịt nặng	Rất chậm	Rất cao	0,5–1,0	170–230
6	SaClLo	Thịt pha cát/sét	Rất chậm đến Chậm	Trung bình	1,0–2,5	140–180

TT	Mã	Loại đất	Định tính		Định lượng	
			Tốc độ thấm nước	Khả năng trữ nước	Tốc độ thấm (mm/giờ)	Mức giữ nước (mm/m)
7	Lo	Thịt nhẹ	Trung bình	Trung bình	5,1–10,2	150–200
8	SiLo	Thịt trung bình	Chậm	Cao	2,5–6,4	160–220
9	SaLo	Thịt pha cát	Nhanh	Thấp	10,2–20,3	100–150
10	Si	Thịt	Chậm	Rất cao	1,3–3,8	170–230
11	LoSa	Cát pha thịt	Rất nhanh	Thấp	13,2–25,4	70–120
12	Sa	Cát	Rất nhanh	Rất thấp	38,1–203,2	50–100

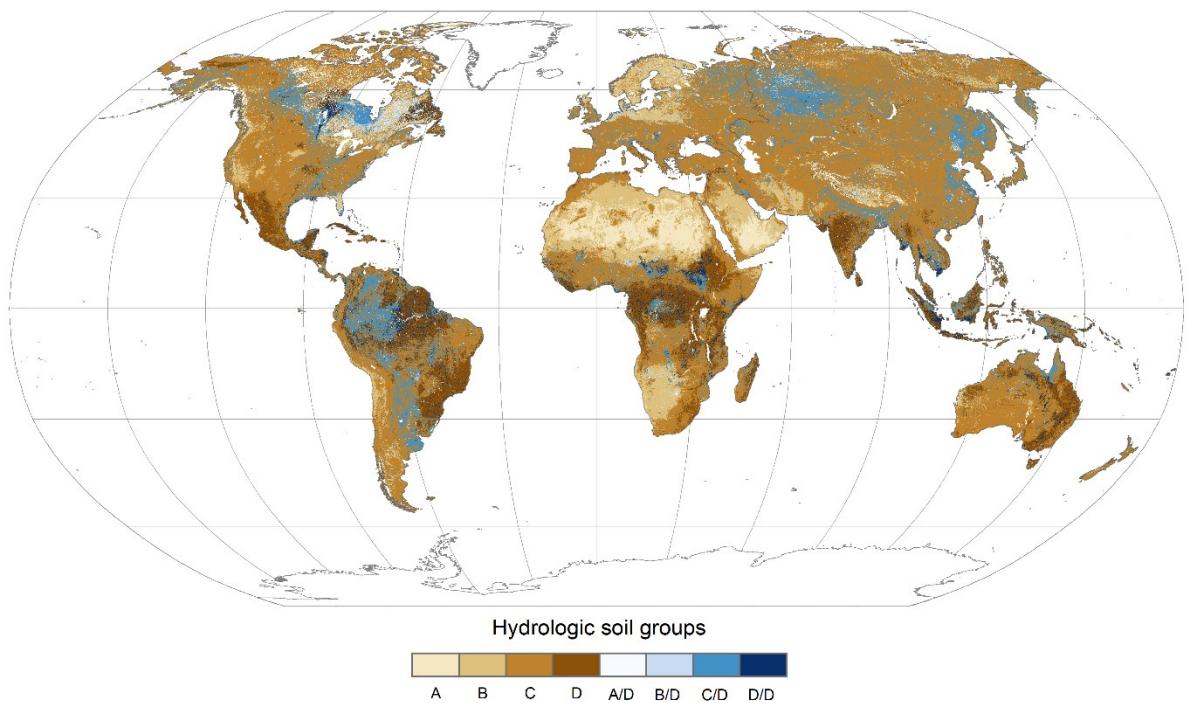
Tốc độ thấm và khả năng giữ nước của các loại đất là đối lập, trong nghiên cứu lũ lụt, tốc độ thấm của bờ mặt càng lớn thì lưu lượng đỉnh lũ sẽ càng giảm. Mức giữ nước thường được nghiên cứu rộng rãi hơn trong nông nghiệp (cho sự phát triển của cây trồng) và trong sạt lở đất (về ứng suất cắt).

#### OpenLandMap Soil Texture Class (USDA System)



Hình 2-22. Dữ liệu cấu trúc đất toàn cầu của USDA có sẵn trong Google Earth Engine

Nghiên cứu này sử dụng giá trị tốc độ thấm bình quân và xây dựng bản đồ tốc độ thấm bình quân để nghiên cứu sự ảnh hưởng của đất đến nguy cơ lũ quét. Đây là loại dữ liệu riêng biệt, là kết quả của phân tích ảnh viễn thám (không phải là ảnh quan sát bờ mặt trực tiếp).



Hình 2-23. Dữ liệu nhóm đất thủy văn được sử dụng trong nghiên cứu

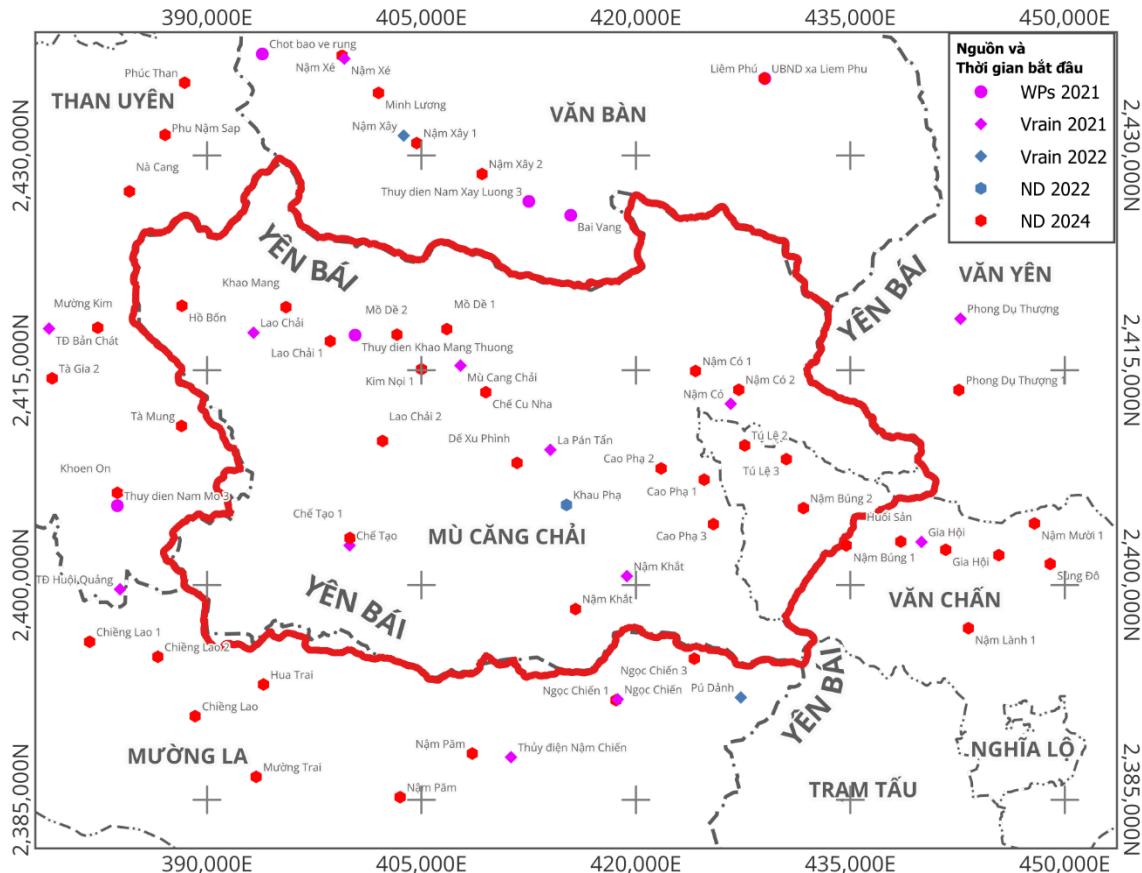
Ngoài ra, một loại dữ liệu đất nữa cũng được sử dụng là dữ liệu nhóm đất thủy văn (hydrologic soil groups). Nhóm đất thủy văn được chia thành 4 loại: A, B, C, và D theo xu hướng tăng dần về mức độ không thấm, phục vụ xác định chỉ số CN khi kết hợp với đặc điểm thảm phủ. Đây là dữ liệu then chốt cho việc ước tính lượng dòng chảy bì mặt trong thủy văn. Dữ liệu này được lấy từ nguồn ORNL DAAC (ROSS, et al., 2018) ở độ phân giải 250m.

Nhóm đất thủy văn là một hệ thống phân loại đất dựa trên khả năng thấm nước tối thiểu của chúng khi được bão hòa hoàn toàn và không có thảm phủ thực vật. Thông thường, đất được chia thành bốn nhóm chính: A, B, C và D. Đất nhóm A có khả năng thấm nước cao nhất (ví dụ: cát, sỏi), dẫn đến lượng dòng chảy bì mặt thấp. Đất nhóm B có khả năng thấm nước vừa phải (ví dụ: đất thịt pha cát). Đất nhóm C có khả năng thấm nước thấp hơn (ví dụ: đất thịt pha sét). Cuối cùng, đất nhóm D có khả năng thấm nước thấp nhất (ví dụ: đất sét), tạo ra lượng dòng chảy bì mặt cao nhất. Việc xác định đúng nhóm đất thủy văn là bước quan trọng đầu tiên trong việc ước tính dòng chảy.

### 3. Dữ liệu vệ tinh quan trắc bì mặt

Dữ liệu vệ tinh quan trắc bì mặt bao gồm các dữ liệu Sentinel, Landsat và Modis sẽ được sử dụng trong nghiên cứu này. Các dữ liệu quang học (Sentinel và landsat) hỗ trợ giải đoán thảm phủ, trong khi Modis cung cấp dữ liệu độ ẩm bì mặt. Dữ liệu Sentinel-1C (SAR) cũng được sử dụng kết hợp để hỗ trợ công tác giải đoán thảm phủ. Sử dụng dữ liệu vệ tinh quan trắc bì mặt để tính toán các chỉ số: (1) NDVI; (2) MNDVI; và (3) CN (khi kết hợp với nhóm đất thủy văn).

#### 4. Dữ liệu mưa



Hình 2-24. Các trạm quan trắc mưa và thời gian bắt đầu đo khu vực Mù Cang Chải

Dữ liệu mưa được sử dụng là dữ liệu mưa quan trắc tại các trạm và dữ liệu mưa vệ tinh nhằm thay thế và tăng cường dữ liệu trong quá khứ trong điều kiện hạn chế về quan trắc mưa trạm. Các dữ liệu này đều được xây dựng bằng việc nội suy thành raster (đối với dữ liệu mưa trạm) làm đầu vào cho công tác dự đoán nguy cơ lũ quét.

Dữ liệu mưa giờ tại các trạm được thu thập từ năm 2021–2024 từ Vrain và các nguồn khác (Thủy Thanh, 2020). Khu vực nghiên cứu và phụ cận năm 2021 có 19 trạm được lắp đặt, trong đó Vrain là 13 trạm và WPs là 6 trạm. Năm 2022 có 3 trạm được ghi nhận thêm dữ liệu và năm 2024 có 48 trạm ghi nhận dữ liệu. Như vậy, năm 2024 vừa qua là năm có số lượng trạm quan trắc nhiều nhất trên khu vực.

#### 5. Sự kiện lũ quét

Sự kiện lũ quét được thu thập tại địa phương thông qua thực địa, bên cạnh đó, một số các nguồn tài liệu từ internet và các đề tài/dự án trước đây cũng được khai thác bổ sung.

Các sự kiện lũ quét được ghi nhận là không đầy đủ, lý do của việc này là đôi khi lũ quét xảy ra ở các khu vực hẻo lánh và không để lại thiệt hại về người và tài sản thì không được ghi nhận. Do đó, bản chất của lũ quét vẫn chưa thể được phản ánh rõ ràng. Nếu trong cùng một điều kiện mưa, có tới 3 khu vực đều xảy ra lũ quét mà chỉ có 1 khu

vực gây thiệt hại được ghi nhận, các khu vực còn lại không được ghi nhận sẽ không phản ánh đúng được tình hình mưa lũ trên khu vực. Điều này không chỉ gây ra khó khăn trong việc đảm bảo độ tin cậy của dữ liệu, mà còn gây khó khăn trong cả quá trình đánh giá các mô hình dự báo.

Tổng hợp các sự kiện lũ quét thu thập tại huyện Mù Cang Chải trong 10 năm trở lại đây thể hiện như sau:

Bảng 2-12. Sự kiện lũ quét tổng hợp tại Mù Cang Chải

<b>Ngày</b>	<b>Thời điểm</b>	<b>Vị trí</b>	<b>Thiệt hại</b>
03/08/2017 (Theo Vietnamplus, n.d.; Vũ Bá Thao & Bùi Xuân Việt, 2023; JICA, 2021)	Khoảng 05:30	- Suối Háng Chú - Dốc suối Nậm Kim (TT. Mù Cang Chải) - Suối Háng Gàng	15 người chết và mất tích, 8 người bị thương. 46 nhà bị thiệt hại (32 nhà bị cuốn trôi hoàn toàn, 14 nhà bị sập do sạt lở đất). Hư hỏng nhiều công trình công cộng, như Trường mầm non Hoa Lan, Trường Tiểu học và THCS Thị trấn, Trung tâm Bồi dưỡng chính trị huyện. (Thanh Thủy, n.d.)
20/07/2018 (Đinh Sơn, 2018; Vũ Bá Thao & Bùi Xuân Việt, 2023; JICA, 2021)	Khoảng 07:00	- Suối Ngòi Hút	- 3 người bị lũ cuốn trôi. - 79 nhà bị sập, cuốn trôi
20/7/2019 (Mạnh Cường, 2019)	Đêm 19/7 đến trưa 20/7	- Suối Mí Háng	- Một người bị thương - 12 hộ bị thiệt hại về nhà cửa
06/08/2023 (Báo Nông Nghiệp VÀ Môi Trường & Thanh Niên - Trần Nam, 2023)	Đêm 05/8 đến rạng sáng 06/8	- Xã Hồ Bón - Xã Kim Nọi, Lao Chải, Khao Mang	- 248 ngôi nhà bị sập, trôi hoàn toàn hoặc hư hỏng nặng. - Hư hỏng cơ sở hạ tầng như điện, đường, trường, trạm.

Ngoài các trận lũ trên, tại suối Nậm Khắt (8 giờ sáng ngày 23/6/2011) cũng xảy ra một trận lũ quét làm 5 người bị cuốn trôi (gây hậu quả 4 người chết). Tuy nhiên, trận lũ này ít có thông tin và chỉ được mô tả là “một trận lũ quét nghiêm trọng” (Anon., 2011; Anon., n.d.).

## 2.2.2 Chuẩn bị dữ liệu

### 2.2.2.1 Nhóm dữ liệu địa không gian

Độ dốc địa hình: là một sản phẩm của dữ liệu địa hình, thể hiện vai trò của sự kiểm soát vận tốc dòng chảy và khả năng thấm của bờ mặt. Có thể nói rằng độ dốc là một trong những dữ liệu quan trọng nhất đối với loại hình thiên tai lũ quét. Đây cũng là lý do mà lũ quét thường xuyên xảy ra ở các khu vực đồi núi, nơi có độ dốc cao. Có nghiên cứu đã chỉ ra rằng, độ dốc lớn hơn  $20^\circ$  làm giảm khả năng thấm của bờ mặt lên tới  $40\text{--}60\%$  và làm tăng vận tốc dòng chảy (He, Fei, et al., 2025). Công cụ Slope có sẵn trong ArcGIS và QGIS là công cụ chính được sử dụng trong nghiên cứu này.

Độ dốc lòng dẫn: Độ dốc lòng dẫn (channel slope) biểu thị độ nghiêng dọc theo kinh dòng chảy chính, ảnh hưởng đến vận tốc và thời gian tập trung nước. Độ dốc lòng dẫn cao làm tăng nguy cơ lũ quét do giảm thời gian dòng chảy. Khác với độ dốc bờ mặt, độ dốc lòng dẫn được tính bằng cách trích xuất hồ sơ dọc (longitudinal profile) của kinh từ DEM, sử dụng công cụ Flow Path Profiling trong ArcGIS hoặc Profile Tool trong QGIS. Độ dốc lòng dẫn được xác định bằng tỷ số giữa chênh lệch độ cao giữa hai điểm trên kinh và chiều dài kinh.

Chiều dài lòng dẫn: Chiều dài lòng dẫn là khoảng cách dọc theo kinh từ điểm cao nhất của lưu vực đến cửa ra, ảnh hưởng đến thời gian tập trung nước. Lòng dẫn ngắn làm tăng nguy cơ lũ quét do nước chảy nhanh. Chiều dài lòng dẫn được tính từ DEM bằng công cụ Flow Length trong ArcGIS hoặc QGIS, sử dụng flow direction để xác định đường đi của dòng chảy.

Diện tích lưu vực: Diện tích lưu vực quyết định lượng nước mưa tích tụ, ảnh hưởng đến quy mô lũ quét. Lưu vực nhỏ với độ dốc cao thường gây lũ quét nhanh. Diện tích lưu vực được tính từ DEM bằng công cụ Watershed trong ArcGIS hoặc QGIS, dựa trên flow direction và flow accumulation.

Khoảng cách đến sông suối: Khoảng cách từ một điểm đến sông suối gần nhất ảnh hưởng đến nguy cơ ngập lũ. Các điểm gần sông ( $<100$  m) dễ bị lũ quét do nhận nước trực tiếp. Khoảng cách được tính bằng công cụ Euclidean Distance trong ArcGIS hoặc QGIS, sử dụng bản đồ dòng chảy chính

Độ cao tương đối đến sông/suối gần nhất: Độ cao tương đối so với sông/suối gần nhất thể hiện khả năng ngập lũ của một điểm. Điểm có độ cao thấp dễ bị lũ quét hơn.

Độ cao tương đối được tính bằng cách trừ độ cao DEM của điểm cho độ cao của sông/suối gần nhất, sử dụng ArcGIS hoặc QGIS

Chỉ số ẩm địa hình (TWI) với vai trò là phản ánh độ bão hòa đất tự nhiên và tiềm năng tạo ra dòng chảy (Alarifi, Saad S., et al., 2022). Với sự liên hệ mật thiết với độ dốc và diện tích lưu vực thượng nguồn ( $TWI = \ln(As / \tan(Slope))$ ), chỉ số TWI càng lớn càng có khả năng xảy ra lũ quét.

Tốc độ thấm bình quân của đất: Tốc độ thấm của đất quyết định lượng nước mưa được hấp thụ hoặc tạo dòng chảy bề mặt. Đất thấm kém (như đất sét) làm tăng nguy cơ lũ quét. Tốc độ thấm bình quân được trích xuất từ bản đồ đất USDA được mô tả trong mục 2.2.1.2, tính trung bình trên lưu vực bằng ArcGIS hoặc QGIS

Chỉ số NDVI bình quân: Chỉ số NDVI (Normalized Difference Vegetation Index) phản ánh mật độ thảm phủ thực vật, ảnh hưởng đến giữ nước và dòng chảy bề mặt. NDVI thấp làm tăng nguy cơ lũ quét. NDVI bình quân được tính từ ảnh vệ tinh (Landsat, Sentinel-2) bằng công cụ Raster Calculator trong ArcGIS hoặc QGIS

Chỉ số CN bình quân: Chỉ số Curve Number (CN) biểu thị tiềm năng tạo dòng chảy bề mặt dựa trên loại đất và sử dụng đất. CN cao (đất tro, đô thị) làm tăng nguy cơ lũ quét. CN bình quân được trích xuất từ bản đồ đất và nhóm đất thủy văn, tính trung bình trên lưu vực bằng ArcGIS hoặc QGIS.

#### 2.2.2.2 Nhóm dữ liệu khí tượng thủy văn

Nhóm dữ liệu mưa sẽ được xác định trong phạm vi 24 giờ đến thời điểm xảy ra lũ quét. Trong đó trích xuất các đặc trưng bao gồm: lượng mưa giờ lớn nhất; lượng mưa 3 giờ lớn nhất; và lượng mưa 24 giờ lớn nhất.

#### 2.2.2.3 Dữ liệu lũ quét lịch sử

Dữ liệu lũ quét lịch sử là thành phần cốt lõi để huấn luyện các mô hình trí tuệ nhân tạo (AI) dự đoán nguy cơ lũ quét. Dữ liệu được chia thành ba loại: dữ liệu thực tế, dữ liệu tăng cường, và dữ liệu không phải lũ quét, với các điểm được gán nhãn khác nhau (lũ quét) hoặc 0 (không lũ quét). Quá trình chuẩn bị đảm bảo tính phù hợp thủy văn và tối ưu hóa cho các mô hình như RF, SVM, LR, LGBM, CNN, DNN, và LSTM.

##### 1. Dữ liệu thực tế

Dữ liệu thực tế bao gồm các vị trí đã xảy ra lũ quét, thu thập từ báo cáo thiên tai hoặc bản đồ lũ lịch sử (Bảng 2-12). Vì lũ quét thường được ghi nhận dưới dạng vùng (polygon), dữ liệu được chuyển thành dạng điểm tương ứng với độ phân giải raster (12.5 m) bằng công cụ Polygon to Point trong ArcGIS hoặc QGIS. Mỗi điểm trong khu vực lũ quét ban đầu được gán nhãn lũ quét, đại diện cho nguy cơ lũ quét cao (nhãn 4). Quá trình chuyển đổi đảm bảo giữ nguyên các điểm đại diện cho khu vực chịu ảnh hưởng trực tiếp của lũ quét.

## 2. Dữ liệu tăng cường

Dữ liệu tăng cường được tạo để mở rộng tập dữ liệu lũ quét dựa trên tình hình thực tế nhằm cải thiện khả năng khai quát hóa của mô hình AI. Lũ quét không chỉ xảy ra trên dòng chảy chính mà còn ảnh hưởng đến các khu vực lân cận như bãi bồi, sườn dốc gần suối, hoặc vùng ngập lũ hạ du. Do đó, một vùng đệm xung quanh các dòng chảy chính tại vị trí lũ quét lịch sử được tạo bằng công cụ Buffer trong ArcGIS hoặc QGIS. Các điểm trong vùng đệm có độ cao tương đối thấp so với lòng suối ( $\leq 5$  m, tính bằng DEM trừ độ cao kênh) được gán nhãn có nguy cơ cao (nhãn 4) vì trên thực tế đây là các đối tượng trực tiếp bị ảnh hưởng.

Nếu ghi nhận lũ lớn trên các nhánh suối mà chưa phân định được là lũ quét hay không phải là lũ quét, các nhánh suối này được gán giá trị nhãn là 3, thể hiện mức độ nguy cơ lũ quét ở mức trung bình. Do đó, cần khảo sát thực tế để nắm bắt được tình hình mưa lũ trên khu vực nghiên cứu. Nhãn 3 được gán cho các điểm thuộc lòng dẫn và lân cận lòng dẫn thuộc khu vực có lũ lớn trong khu vực.

Các nhánh suối khu vực lân cận (thượng và hạ lưu) các nhánh suối gán nhãn 3 được gán nhãn 2 tương ứng với nguy cơ thấp cho các điểm thuộc lòng dẫn, bên cạnh đó các nhánh suối có mưa và tạo ra các dòng chảy thông thường cũng được gán nhãn 2, các nhánh suối này thường được mô tả là các nhánh suối có mức nước lũ thông thường trong mùa mưa lũ, không có ghi nhận về dòng chảy bất thường.

Các đoạn suối thượng nguồn, nơi có diện tích lưu vực nhỏ nằm trên sườn dốc núi không ghi nhận dòng chảy lũ bất thường được gán nhãn 1 (nguy cơ rất thấp). Lý do của việc này là tại các khu vực thượng nguồn, dòng chảy tập trung nhỏ, chưa hình thành đủ năng lượng để có thể hình thành lũ quét trong bất cứ điều kiện nào (không xét đến sạt lở, nghẽn dòng).

Dữ liệu không phải lũ quét (nhãn 0) đóng vai trò quan trọng trong việc huấn luyện mô hình AI, vì gán nhãn sai có thể dẫn đến đánh giá thấp nguy cơ lũ quét. Do dữ liệu lũ quét lịch sử thường không đầy đủ (một số sự kiện có thể không được ghi nhận), việc chọn các điểm không lũ quét đòi hỏi tiêu chí thủy văn chặt chẽ. Các điểm có xác suất lũ quét thấp được chọn dựa trên hai tiêu chí:

- Khoảng cách xa sông/suối: Các điểm cách sông/suối  $> 200$  m, nằm ở vị trí cao (gần đỉnh núi, sườn núi) hoặc có flow accumulation thấp ( $< 100$  ô lưới) được coi là an toàn, vì thiếu lượng nước tập trung cần thiết cho lũ quét. Khoảng cách được tính bằng công cụ Euclidean Distance trong ArcGIS hoặc QGIS
- Lượng mưa thấp: Các điểm có lượng mưa 1 giờ hoặc 6 giờ  $< 10$  mm, trích xuất từ trạm đo, được gán nhãn 0. Nguồn này dựa trên nghiên cứu cho thấy mưa  $< 10$  mm khó kích hoạt lũ quét, ngay cả ở khu vực gần sông/suối (Petr Sercl, et al., 2023).

#### 2.2.2.4 Chuẩn hóa dữ liệu

Việc chuẩn hóa dữ liệu (data normalization/standardization) là bước quan trọng trong các bài toán học máy và học sâu, đặc biệt khi các đặc trưng đầu vào có thang đo khác nhau (ví dụ: lượng mưa từ 0-200 mm, độ dốc từ 0-45 độ, NDVI từ 0-1 trong bài toán lũ quét). Chuẩn hóa giúp đưa các đặc trưng về cùng thang đo, đảm bảo mô hình (như RF, SVM, LR, LGBM, DNN, CNN, LSTM) không bị thiên lệch bởi các đặc trưng có giá trị lớn và cải thiện hiệu suất huấn luyện. Các lý do cần phải chuẩn hóa bao gồm:

- Thang đo khác nhau: Các đặc trưng như lượng mưa (mm), độ dốc (độ), flow accumulation (số tế bào), hay NDVI (0-1) có đơn vị và phạm vi giá trị khác nhau. Nếu không chuẩn hóa, các đặc trưng có giá trị lớn (như flow accumulation) có thể áp đảo các đặc trưng nhỏ (như NDVI) trong các mô hình nhạy cảm với thang đo, như SVM, LR, DNN, CNN, LSTM.
- Tối ưu hóa mô hình: Chuẩn hóa giúp các thuật toán dựa trên gradient (như DNN, CNN, LSTM) hội tụ nhanh hơn và cải thiện độ chính xác.
- Tính tương thích: Một số mô hình (như SVM) yêu cầu dữ liệu trong cùng phạm vi (ví dụ: [0, 1] hoặc [-1, 1]) để hoạt động hiệu quả.

Các phương pháp chuẩn hóa dữ liệu cơ bản:

##### 1. Min-Max Scaling (Chuẩn hóa tuyến tính)

Min-Max Scaling là một kỹ thuật cơ bản trong xử lý dữ liệu, phát triển từ các phương pháp thống kê và khoa học máy tính từ những năm 1970. Nó được sử dụng rộng rãi trong các thuật toán học máy sớm như mạng nơ-ron và SVM, không gắn với một cá nhân cụ thể mà là sản phẩm của cộng đồng nghiên cứu (Han J., et al., 2012).

###### a. Các đặc điểm của chuẩn hóa tuyến tính:

- Biến đổi dữ liệu về một phạm vi cố định, thường là [0, 1], bằng cách chia tỷ lệ giá trị dựa trên giá trị nhỏ nhất và lớn nhất của đặc trưng.
- Phù hợp khi mô hình yêu cầu dữ liệu trong phạm vi cố định (như DNN, CNN) và dữ liệu không có nhiều giá trị ngoại lai.

###### b. Công thức

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

###### c. Ưu điểm

- Đơn giản, dễ triển khai, giữ nguyên phân bố tương đối của dữ liệu.
- Phù hợp với các mô hình yêu cầu dữ liệu trong [0, 1], như DNN, CNN, SVM.
- Hiệu quả khi dữ liệu có phân bố đều và ít giá trị ngoại lai.

#### d. Nhược điểm

- Nhạy cảm với giá trị ngoại lai (outliers). Nếu  $X_{\max}$  hoặc  $X_{\min}$  là ngoại lai, phạm vi chuẩn hóa sẽ bị méo mó.
- Không phù hợp với dữ liệu lệch mạnh (skewed), như flow accumulation hoặc lượng mưa trong lũ quét.

#### e. Ứng dụng trong lũ quét:

- Phù hợp cho các đặc trưng có phạm vi giới hạn và ít outliers, như NDVI (0-1), chỉ số CN (0-100), hoặc độ dốc (0-90 độ).
- Cần làm sạch outliers (ví dụ: loại bỏ giá trị lượng mưa bất thường) trước khi áp dụng.

### 2. Standardization (Z-score Normalization)

Standardization dựa trên khái niệm Z-score trong thống kê, phát triển từ các nghiên cứu về phân phối chuẩn của Carl Friedrich Gauss (thế kỷ 19). Trong học máy, nó được chuẩn hóa từ những năm 1980-1990, đặc biệt trong các thuật toán như SVM và hồi quy tuyến tính (Hastie, et al., 2008).

#### a. Đặc điểm của Standardization

- Biến đổi dữ liệu thành phân phối có trung bình bằng 0 và độ lệch chuẩn bằng 1, dựa trên trung bình và độ lệch chuẩn của đặc trưng.
- Phù hợp với các mô hình giả định dữ liệu gần phân phối chuẩn, như SVM, LR, DNN, và dữ liệu có phân bố lệch hoặc outliers.

#### b. Công thức

$$X_{\text{std}} = \frac{X - \mu}{\sigma}$$

Trong đó:

$X_{\text{std}}$ : Giá trị đã được chuẩn hóa (Z-score) của một điểm dữ liệu.

$X$ : Giá trị gốc của điểm dữ liệu trong tập đặc trưng (ví dụ: lượng mưa, độ dốc, flow accumulation).

$\mu$ : Giá trị trung bình (mean) của toàn bộ tập đặc trưng  $X$ .

$\sigma$ : Độ lệch chuẩn (standard deviation) của toàn bộ tập đặc trưng  $X$ .

#### c. Ưu điểm

- Ít nhạy cảm với outliers hơn Min-Max Scaling, vì dựa trên trung bình và độ lệch chuẩn.
- Phù hợp với dữ liệu lệch hoặc có outliers, như lượng mưa, flow accumulation.
- Tốt cho các mô hình dựa trên gradient (DNN, CNN, LSTM) và SVM, LR.

#### d. Nhược điểm

- Không giới hạn dữ liệu trong phạm vi cố định (giá trị có thể âm hoặc lớn hơn 1), có thể không phù hợp với một số CNN yêu cầu [0, 1].
- Nếu dữ liệu rất lệch, cần biến đổi trước (như log transform).

#### e. Ứng dụng trong lũ quét

- Lý tưởng cho các đặc trưng có phân bố lệch hoặc outliers, như lượng mưa (1 giờ, 3 giờ, 6 giờ, 24 giờ), flow accumulation, TWI.
- Phù hợp với SVM, LR, DNN, CNN, LSTM, vì chúng nhạy cảm với thang đo.

### 3. Robust Scaling

Robust Scaling dựa trên thống kê mạnh (robust statistics), phát triển từ những năm 1970-1980 để xử lý dữ liệu có outliers. Nó được tích hợp vào các thư viện học máy như Scikit-learn và phổ biến từ những năm 2000 trong các ứng dụng thực tế (Pedregosa, Fabian, 2012).

#### a. Đặc điểm của Robust Scaling

- Chuẩn hóa dữ liệu dựa trên trung vị và khoảng tứ phân vị (IQR), thay vì trung bình và độ lệch chuẩn, để giảm ảnh hưởng của giá trị ngoại lai.
- Phù hợp với dữ liệu có nhiều outliers, như flow accumulation ở các điểm trên dòng chảy chính.

#### b. Công thức

$$X_{\text{robust}} = \frac{X - Q_2}{Q_3 - Q_1}$$

Trong đó Q là các phân phân vị.

#### c. Ưu điểm

- Rất mạnh với outliers, vì sử dụng trung vị và IQR.
- Phù hợp với dữ liệu có giá trị ngoại lai nghiêm trọng, như flow accumulation, TWI.
- Tốt cho SVM, LR, DNN khi dữ liệu không sạch.

#### d. Nhược điểm

- Không giới hạn phạm vi, có thể không phù hợp với CNN yêu cầu [0, 1].
- Mất thông tin nếu dữ liệu có phân bố phức tạp hoặc ít outliers.

#### e. Ứng dụng trong lũ quét

- Lý tưởng cho các đặc trưng có nhiều outliers, như flow accumulation (giá trị lớn trên dòng chảy chính), TWI, hoặc lượng mưa.
- Phù hợp khi dữ liệu chưa được làm sạch hoàn toàn.

## 4. Log Transformation

Log Transformation bắt nguồn từ thống kê, được sử dụng từ thế kỷ 19 để xử lý dữ liệu lệch. Một biến thể là Box-Cox Transformation, được George E. P. Box và David R. Cox đề xuất năm 1964. Yeo-Johnson Transformation (2000) là phiên bản mở rộng cho cả giá trị âm (Box G. E. P. & Cox D. R., 1964; In-Kwon Yeo & Richard A. Johnson, 2000).

### a. Đặc điểm của Log Transformation

- Biến đổi dữ liệu bằng hàm logarithm để giảm độ lệch (skewness), đặc biệt với dữ liệu lệch phải (right-skewed). Sau đó, có thể áp dụng Standardization hoặc Min-Max Scaling.
- Phù hợp với các đặc trưng có giá trị dương lớn, như flow accumulation, lượng mưa.

### b. Công thức

$$X_{\log} = \log(X + c)$$

Với  $c$  là hằng số chuyển đổi

### c. Ưu điểm

- Hiệu quả với dữ liệu lệch mạnh (flow accumulation, lượng mưa).
- Giảm tác động của outliers bằng cách nén các giá trị lớn.
- Tăng hiệu quả khi kết hợp với Standardization.

### d. Nhược điểm

- Không áp dụng trực tiếp cho giá trị âm hoặc 0 (cần thêm hằng số  $c$ ).
- Có thể làm mất thông tin nếu dữ liệu không lệch mạnh.
- Thường cần bước chuẩn hóa tiếp theo.

### e. Ứng dụng trong lũ quét

- Rất phù hợp cho các đặc trưng lệch mạnh, như flow accumulation, lượng mưa (1 giờ, 3 giờ, 6 giờ, 24 giờ), diện tích lưu vực thượng nguồn.
- Nên kết hợp với Standardization để chuẩn hóa thêm

## 5. Lựa chọn phương pháp chuẩn hóa trong nghiên cứu

Bảng 2-13. Tổng hợp một số phương pháp chuẩn hóa dữ liệu phổ biến

Phương pháp	Công thức	Ưu điểm	Nhược điểm	Ứng dụng lũ quét
Min-Max Scaling	$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$	Đơn giản, phạm vi $[0, 1]$	Nhạy cảm với outliers	NDVI, CN, không tốt cho mưa
Standardization	$X_{\text{std}} = \frac{X - \mu}{\sigma}$	Ít nhạy cảm với outliers, tốt cho dữ liệu lệch	Không giới hạn phạm vi	Mưa, TPI

Phương pháp	Công thức	Ưu điểm	Nhược điểm	Ứng dụng lũ quét
Robust Scaling	$X_{robust} = \frac{X - Q_2}{Q_3 - Q_1}$	Rất mạnh với outliers	Không giới hạn phạm vi	TWI có outliers
Log Transformation	$X_{log} = \log(X + c)$	Giảm độ lệch, nén outliers	Không áp dụng cho âm/0	Kết hợp với Standardization cho mưa

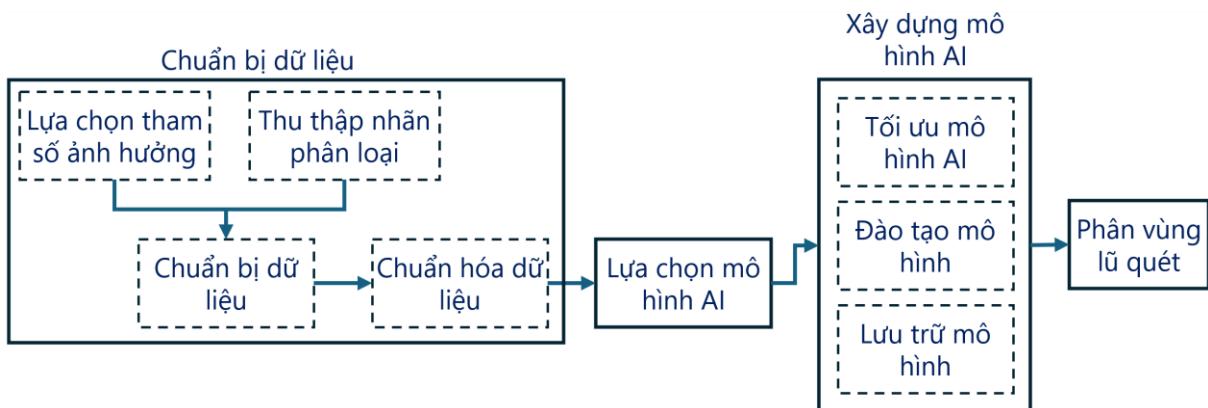
Các phương pháp chuẩn hóa dữ liệu cho từng loại dữ liệu được thống kê ở bảng sau:

Bảng 2-14. Chuẩn hóa dữ liệu cho từng loại dữ liệu

Phương pháp	Đặc trưng
Min-Max Scaling	- Các đặc trưng: Độ dốc, độ dốc lòng dẫn, chỉ số ẩm địa hình, NDVI, CN. - Chuẩn hóa lần cuối (sau các phương pháp khác) cho các đặc trưng còn lại. Ngoại trừ mưa
Standardization	Chỉ số vị trí địa hình; độ cong địa hình.
Robust Scaling	Cao độ so với sông suối; tốc độ thẩm bình quân; cao độ địa hình; cao độ bình quân lưu vực.
Log Transformation	Khoảng cách đến sông suối; chiều dài dòng chảy; diện tích lưu vực; chỉ số sức mạnh dòng chảy; lượng mưa.

## 2.3. Quy trình ứng dụng trí tuệ nhân tạo và dữ liệu địa không gian để phân vùng lũ quét

### 2.3.1 Sơ đồ quy trình



Hình 2-25. Quy trình ứng dụng trí tuệ nhân tạo trong phân vùng lũ quét

### **2.3.2 Xác định các bước thực hiện**

#### **2.3.3 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 thẩm 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 thấp như vùng có khả năng thẩm 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 thẩm 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 tổn thất 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 thương nguồn nhưng khác biệt về độ thấm của đất sẽ có mức độ nguy hiểm lũ quét khác nhau do khả năng tiêu thoát 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 nắm 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 2-15. 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	
	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	✓

Nhóm dữ liệu	Tham số	Ý nghĩa trong phân vùng/xác định lũ quét	Mức độ ưu tiên (1-10)	Yếu tố lưu vực
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 bề 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 gặp 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 cù thê) 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 đây 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 cù thê 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 (nhị 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→1 hoặc lân cận nhằm đảm bảo tính nhất quán. Nghiên cứu này khuyến nghị sử dụng khoảng giá trị từ 0→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 phần lớn 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 bố 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 bố 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 phôi xung quanh một giá trị trung tâm với sự biến thiên theo cả hai hướng. Z-score transformation đưa dữ liệu về phân phôi 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 phôi 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 phôi tương 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 2-16. 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
	Profile Curvature	Z-score + MinMax	Có giá trị âm/dương, cần Z-score trước

Nhóm Dữ Liệu	Tham Số	Phương Pháp Đề Xuất	Lý Do
	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	Thông 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	Thông 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 thám 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
	Nhiệt độ (Temperature)	Z-score + MinMax	Có thể có giá trị âm, phân phối gần chuẩn

Nhóm Dữ Liệu	Tham Số	Phương Pháp Đề Xuất	Lý Do
	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

### 2.3.4 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 mảng 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 2-17). 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 ấn 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 khuyên nghị sử dụng mạng CNN và DNN trong phân vùng lũ quét.

Bảng 2-17. 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
<b>Học máy</b>					
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
<b>Học sâu</b>					
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 2-18. Khuyến nghị tối ưu siêu tham số cho nhóm mô hình Gradient Boosting

<b>Siêu Tham Số</b>	<b>Mức Độ Ưu Tiên</b>	<b>Khoảng Tối Ưu Khuyến Nghị</b>	<b>Giải Thích và Ảnh Hưởng</b>
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 2-19. Khuyến nghị tối ưu siêu tham số cho mô hình Random Forest

<b>Siêu Tham Số</b>	<b>Mức Độ Ưu Tiên</b>	<b>Khoảng Tối Ưu Khuyến Nghị</b>	<b>Giải Thích và Ảnh Hưởng</b>
n_estimators	RẤT CẦN THIẾT	100-500	Default (100) thường quá thấp. Cải thiện đáng kể hiệu suất. Khuyến nghị: 200-300
max_depth	RẤT 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. Thử 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 2-20. Khuyến nghị tối ưu siêu tham số cho mô hình CNN

<b>Siêu Tham Số</b>	<b>Mức Độ Ưu Tiên</b>	<b>Khoảng Tối Ưu Khuyến Nghị</b>	<b>Giải Thích và Ảnh Hưởng</b>
Số lớp conv	RẤT CẦN THIẾT	2-5 lớp	Quyết định khả năng học pattern. Với ít features: 3-4 lớp thường tối ưu
Số filters	RẤT CẦN THIẾT	16-256 mỗi lớp	Ảnh hưởng lớn đến bộ nhớ. Bắt đầu: $32 \rightarrow 64 \rightarrow 128$
Learning rate	RẤT CẦN THIẾT	$1e-4 \div 1e-3$	Quan trọng nhất trong deep learning. Thử: $1e-4, 5e-4, 1e-3$
Batch size	CẦN THIẾT	32-128	Ảnh hưởng đến gradient stability. Thử: 32, 64, 128
Dropout rate	CẦN THIẾT	0.2-0.5	Quan trọng để tránh overfitting. Thử: 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 2-21. 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	RẤT CẦN THIẾT	2-6 layers	Quyết định model complexity. Với 15-20 features: 3-4 layers
Số neurons per layer	RẤT CẦN THIẾT	32-256	Ảnh hưởng lớn đến capacity. Thủ: $64 \rightarrow 128 \rightarrow 64$ hoặc $128 \rightarrow 64 \rightarrow 32$
Learning rate	RẤT CẦN THIẾT	1e-4 to 1e-3	Quan trọng nhất. Thủ: 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. Thủ: 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. Thủ: 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 số\_features/3 đến 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.

### 2.3.5 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.

## **CHƯƠNG 3. XÂY DỰNG MÔ HÌNH PHÂN VÙNG LŨ QUÉT CHO KHU VỰC MÙ CANG CHẢI**

### **3.1. Lựa chọn khu vực nghiên cứu**

Mù Cang Chải có lịch sử chịu thiệt hại nặng nề từ các trận lũ quét. Một ví dụ điển hình là trận lũ quét ngày 5 tháng 8 năm 2023 tại xã Hồ Bón (Báo Nông Nghiệp Và Môi Trường & Thanh Niên - Trần Nam, 2023), gây thiệt hại nghiêm trọng. Trận lũ này không chỉ làm rõ tính bất ngờ và sức tàn phá của lũ quét mà còn nhấn mạnh sự cần thiết của các công cụ dự báo và quản lý rủi ro.



Hình 3-26. Trận lũ xảy ra tại Hồ Bón năm 2023

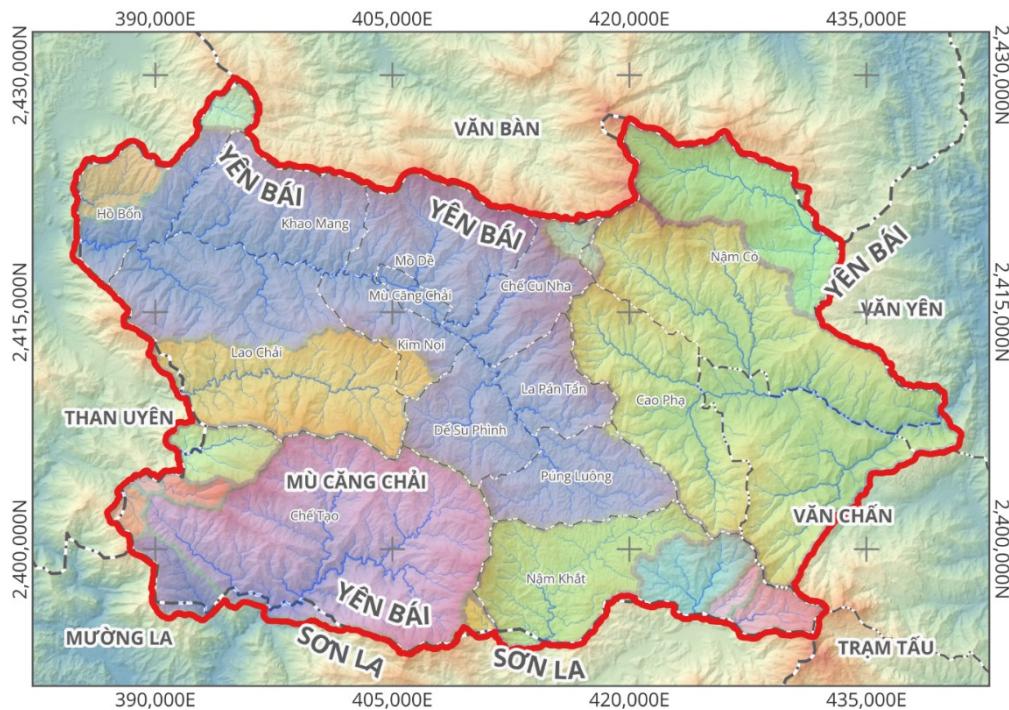
Các sự kiện lũ quét trước đó, như trận lũ năm 2017 tại trung tâm thị trấn Mù Cang Chải, cũng gây thiệt hại lớn về người và tài sản (Theo Vietnamplus, n.d.; Vũ Bá Thảo & Bùi Xuân Việt, 2023; JICA, 2021). Những sự kiện này cho thấy Mù Cang Chải là một trong những khu vực có nguy cơ lũ quét cao nhất ở miền Bắc Việt Nam, với các yếu tố như độ dốc lớn, mật độ dân cư cao ở các khu vực ven sông suối và thiếu hệ thống cảnh báo sớm.



Hình 3-27. Lũ quét tại suối Hàng Chú năm 2017 (Bá Thảo, Vũ, 2020)

Mù Cang Chải được chọn vì nó đại diện cho các khu vực miền núi Việt Nam, nơi mà các yếu tố như địa hình dốc, mưa lớn và thay đổi sử dụng đất tạo ra nguy cơ lũ quét cao. Việc nghiên cứu khu vực này cho phép các nhà khoa học phát triển các mô hình có thể áp dụng rộng rãi, cung cấp thông tin chi tiết về cách giảm thiểu rủi ro lũ quét trong bối cảnh biến đổi khí hậu và đô thị hóa. Phân tích địa hình thủy văn cho khu vực Mù

Cang Chải cho thấy khu vực nghiên cứu là thượng nguồn của các lưu vực sông như lưu vực Nậm Kim; lưu vực ngòi Hút; lưu vực Nậm Tha; và lưu vực Nậm Chang.



Hình 3-28. Huyện Mù Cang Chải là thượng nguồn của nhiều lưu vực, không chịu sự tác động lũ quét từ bên ngoài mà chỉ xảy ra ở nội tại huyện

### 3.2. Đánh giá các dữ liệu địa không gian trong nghiên cứu lũ quét

#### 3.2.1 Nghiên cứu xây dựng dữ liệu thảm phủ từ ảnh viễn thám

Dữ liệu thảm phủ là dữ liệu quan trọng mô tả bề mặt và có tác động đến dòng chảy trong thủy văn. Một số phân loại thảm phủ chính thường được sử dụng như của Esri và Jaxa do các đơn vị này liên tục phân loại thảm phủ hàng năm.

Hệ thống phân loại của Esri bao gồm 9 lớp chính như Mặt nước, Rừng cây, Thực vật ngập nước, Cây trồng, Khu vực xây dựng, Đất trống, Tuyết/Băng, Mây che phủ, và Đồng cỏ/Thảo nguyên. Hệ thống này có xu hướng tập trung vào đặc điểm tự nhiên và nhân tạo, với sự nhấn mạnh vào các khu vực ngập nước và thực vật (như rừng ngập mặn, ruộng lúa), phù hợp với phân tích đa dạng sinh học và quản lý tài nguyên ở các khu vực nhiệt đới như Việt Nam.

Hệ thống của JAXA bao gồm 12 lớp như Nước, Khu vực đô thị, Ruộng lúa, Cây trồng khác, Cỏ/cây bụi, Cây gỗ/vườn cây, Đất trống, Rừng thường xanh, Rừng rụng lá, Rừng trồng, Rừng ngập mặn, và Nuôi trồng thủy sản. Hệ thống này chi tiết hơn với sự phân biệt rõ ràng giữa các loại rừng (thường xanh, rụng lá, trồng) và bổ sung lớp Nuôi trồng thủy sản, phản ánh sự quan tâm đến các hoạt động kinh tế ven biển.

Mặc dù các hệ thống phân loại này là hiện hữu, việc phân loại lại thảm phủ cho một khu vực cụ thể là cần thiết khi mục tiêu chính là tối ưu hóa phân tích thủy văn, đặc

biệt trong bối cảnh quản lý tài nguyên nước và dự đoán dòng chảy. Thảm phủ bề mặt ảnh hưởng trực tiếp đến các quá trình thủy văn như thấm nước, dòng chảy bề mặt, và bốc hơi, do đó một hệ thống phân loại không phù hợp hoặc thiếu chi tiết có thể dẫn đến sai lệch trong các mô hình thủy văn. Tại một khu vực cụ thể, các đặc điểm tự nhiên và nhân tạo như rừng, đồng cỏ, đất canh tác, hoặc khu vực xây dựng có thể thay đổi theo thời gian do tác động của con người hoặc biến đổi khí hậu, khiến các phân loại có sẵn như của Esri hoặc JAXA không còn phản ánh chính xác thực tế địa phương.

Mục tiêu của việc phân loại lại thảm phủ để phù hợp với thủy văn là xây dựng một cơ sở dữ liệu chi tiết, phản ánh chính xác điều kiện thực địa, từ đó cải thiện độ tin cậy của các mô hình tính toán dòng chảy như phương pháp hệ số đường cong (curve number) của NRCS. Một phân loại mới có thể tập trung vào các yếu tố như mật độ thực vật, mức độ thấm nước của đất, và cách quản lý đất (chẳng hạn canh tác theo đường đồng mức hay chăn thả), vốn ảnh hưởng lớn đến lượng nước thẩm vào lòng đất hoặc chảy tràn ra sông ngòi. Ngoài ra, việc này cho phép đánh giá tác động của các thay đổi sử dụng đất, như chuyển từ rừng sang khu vực đô thị hóa, lên tài nguyên nước, đặc biệt trong mùa mưa hoặc lũ lụt. Tại khu vực cụ thể, phân loại lại cũng giúp xác định các vùng dễ bị ngập úng hoặc khô hạn, từ đó hỗ trợ lập kế hoạch thủy lợi và bảo vệ môi trường một cách hiệu quả hơn. Vì vậy, quá trình này không chỉ là cập nhật dữ liệu mà còn là bước quan trọng để đảm bảo các quyết định quản lý dựa trên cơ sở khoa học, phù hợp với nhu cầu thực tiễn của khu vực nghiên cứu.

Nghiên cứu này tập trung vào lũ quét, điều này đồng nghĩa với việc đi tìm những đặc điểm về sử dụng đất có liên quan đến yếu tố dòng chảy. Nghiên cứu lũ quét đòi hỏi một cách tiếp cận chi tiết về thảm phủ để đánh giá chính xác tiềm năng dòng chảy và ngập úng, đặc biệt trong các khu vực dễ bị tổn thương. Hệ thống phân loại thảm phủ của NRCS, được trình bày trong Chương 8 và Chương 9 của National Engineering Handbook (NEH) (NRCS, 2020), cung cấp một khuôn khổ khoa học để phân tích các yếu tố thủy văn như thấm nước, dòng chảy bề mặt, và khả năng giữ nước. Theo NRCS, thảm phủ được chia thành các lớp sử dụng đất và xử lý đất (land use and treatment classes), bao gồm đất canh tác, đồng cỏ, rừng, và khu vực đô thị, với các điều kiện thủy văn được đánh giá qua các cấp độ như tốt, trung bình, và kém. Đối với lũ quét, việc áp dụng phân loại này giúp xác định các khu vực có hệ số đường cong runoff (CN) cao, vốn phản ánh khả năng dòng chảy nhanh và mạnh, đặc biệt trên đất trống, đất canh tác không che phủ, hoặc khu vực đô thị hóa với nhiều bề mặt không thấm nước. Ví dụ, đất canh tác để trống (fallow) hoặc cây hàng (row crop) trong điều kiện kém có thể đạt CN từ 76 đến 89 tùy theo nhóm đất, cho thấy nguy cơ lũ quét cao do giảm thấm nước.

Phân loại theo NRCS cũng nhấn mạnh vai trò của các biện pháp xử lý đất như trồng cây theo đường đồng mức (contouring) hoặc làm bậc thang (terracing), vốn làm chậm dòng chảy và tăng khả năng thấm, từ đó giảm rủi ro lũ quét. Trong khu vực nghiên

cứu, việc kết hợp dữ liệu thực địa về mật độ thực vật, lượng rác hữu cơ, và mức độ chấn thả với bảng CN (như Bảng 9-1) (NRCS, 2020) cho phép xây dựng mô hình thủy văn chính xác hơn. Đặc biệt, đối với các khu vực rừng hoặc đồng cỏ, NRCS phân loại theo điều kiện thủy văn (poor, fair, good) dựa trên tỷ lệ che phủ và quản lý, giúp đánh giá tác động của quản lý đất đai đến dòng chảy cực đoan. Ngoài ra, việc áp dụng phương trình tính CN tổng hợp (equation 9-1 hoặc 9-2) từ Chương 9 (NRCS, 2020) cho các khu vực đô thị hoặc hỗn hợp sử dụng đất giúp điều chỉnh rủi ro lũ quét khi có sự thay đổi trong tỷ lệ bề mặt không thấm. Vì vậy, việc sử dụng hệ thống NRCS không chỉ hỗ trợ phân tích hiện trạng mà còn định hướng chiến lược quản lý thảm phủ để giảm thiểu tác động của lũ quét, đặc biệt trong các khu vực có địa hình dốc hoặc mưa lớn như ở Việt Nam.

Nghiên cứu này thể hiện một nỗ lực hệ thống hóa các đặc điểm tự nhiên và nhân tạo của khu vực Mù Cang Chải nhằm phục vụ mục tiêu thủy văn, đặc biệt trong quản lý dòng chảy và lũ quét. Nhóm 1 (Mặt nước: Hồ, sông, đầm lầy) đại diện cho các khu vực có tiềm năng dòng chảy bề mặt rất thấp do tính chất tự nhiên thấm nước, với hệ số đường cong runoff (CN) được NRCS xác định là 98, phản ánh khả năng giữ nước tối đa. Nhóm 2 (Rừng rậm: Rừng dày, điều kiện Good) mô tả các khu vực thực vật dày đặc, có khả năng thấm nước cao nhờ lớp thực bì và rác hữu cơ, với CN dao động từ 30-55 tùy thuộc vào nhóm đất, giúp giảm đáng kể nguy cơ lũ quét.

Nhóm 3 (Đất trống: Fallow hoặc đất không có thực vật) chỉ ra các khu vực nông nghiệp để trống hoặc không có che phủ, với CN cao từ 67-89 tùy nhóm đất, cho thấy nguy cơ dòng chảy bề mặt lớn do thiếu thực vật giữ nước. Nhóm 4 (Khu xây dựng: Bề mặt không thấm nước) bao gồm các khu vực như nhà cửa và đường sá, có CN là 98, dẫn đến dòng chảy nhanh và tăng rủi ro lũ quét do không có khả năng thấm. Nhóm 5 (Rừng thưa: Rừng thưa, điều kiện Fair/Poor) đại diện cho vùng thực vật thưa thớt, với CN trung bình từ 55-79, phản ánh giảm khả năng thấm so với rừng rậm và tiềm năng dòng chảy cao hơn.

Nhóm 6 (Ruộng bậc thang: Đất canh tác với hệ thống terracing) nhấn mạnh kỹ thuật canh tác tạo bậc thang, giúp lưu trữ nước và giảm tốc độ dòng chảy, với CN từ 60-80 tùy điều kiện quản lý, mang lại hiệu quả trong kiểm soát lũ. Nhóm 7 (Đất trồng lúa: Đất canh tác ngập nước) bao gồm ruộng lúa, nơi dòng chảy phụ thuộc vào quản lý nước, với CN từ 70-85 tùy mức độ ngập, đòi hỏi đánh giá kỹ lưỡng về hệ thống thủy lợi. Logic của phân loại này nằm ở việc kết hợp các yếu tố tự nhiên (thảm thực vật, nước) và nhân tạo (quản lý đất, xây dựng), phù hợp với phương pháp NRCS, giúp phân tích chi tiết tác động của từng loại thảm phủ lên dòng chảy, hỗ trợ lập kế hoạch giảm thiểu lũ quét tại các khu vực như Mù Cang Chải. Chi tiết được thể hiện trong bảng sau đây:

Bảng 3-22. Nhóm phân loại thảm phủ sử dụng trong nghiên cứu

Lớp	Mô tả	Điều kiện thủy văn	Ảnh hưởng đến dòng chảy
1	Mặt nước (Hồ, sông, đầm lầy)	Bè mặt nước	Tạo dòng chảy trực tiếp, không thâm thấu
2	Rừng rậm (Rừng dày, điều kiện Good)	Good	Tăng thâm thấu, giảm dòng chảy bè mặt
3	Đất trống (Fallow hoặc đất không có thực vật)	Poor	Tăng dòng chảy bè mặt, giảm thâm thấu
4	Khu xây dựng (Bè mặt không thấm nước)	Impervious	Tạo dòng chảy nhanh, không thâm thấu
5	Rừng thưa (Rừng thưa, điều kiện Fair/Poor)	Fair/Poor	Thâm thấu trung bình, dòng chảy trung bình
6	Ruộng bậc thang (Đất canh tác với hệ thống terracing)	Contoured/terraced	Giảm dòng chảy nhờ lưu trữ nước trong ruộng
7	Đất trống lúa (Đất canh tác ngập nước)	Wet meadow	Giữ nước lâu dài, dòng chảy chậm

### 1. Chuẩn bị dữ liệu

Nghiên cứu lựa chọn 16 chỉ số đầu vào (NDVI, EVI, GNDVI, GRVI, NDWI1, NDWI2, GSI, NDBI, NDWI, BSI, Slope, Red, Green, Blue, VV, VH) để phân loại thảm phủ theo 7 nhóm (Mặt nước, Rừng rậm, Đất trống, Khu xây dựng, Rừng thưa, Ruộng bậc thang, Đất trống lúa) nhằm tận dụng tối đa thông tin từ dữ liệu đa phổ và radar để phân biệt các đặc trưng quang phổ, cấu trúc, và địa hình của từng loại thảm phủ. Trước tiên, các chỉ số thực vật như NDVI (Normalized Difference Vegetation Index), EVI (Enhanced Vegetation Index), GNDVI (Green Normalized Difference Vegetation Index), và GRVI (Green-Red Vegetation Index) được chọn để đánh giá mật độ thực vật và sức khỏe cây trồng, rất hữu ích trong việc phân biệt Rừng rậm (nhóm 2) và Rừng thưa (nhóm 5) dựa trên sự khác biệt về độ che phủ và sinh khối. NDVI và EVI, với khả năng giảm nhiễu từ khí quyển và đất nền, đặc biệt hiệu quả trong việc xác định các khu vực có thực vật dày đặc (CN thấp, 30-55) so với khu vực thưa thớt (CN 55-79). GNDVI và GRVI bổ sung thông tin về phản xạ ở kênh xanh, hỗ trợ phân loại Đất trống lúa (nhóm 7) nhờ khả năng phát hiện sự thay đổi quang phổ trong môi trường ngập nước.

Các chỉ số liên quan đến nước như NDWI1, NDWI2, NDWI (Normalized Difference Water Index), và GSI (Green-SWIR Index) đóng vai trò quan trọng trong việc nhận diện Nhóm 1 (Mặt nước) và Nhóm 7 (Đất trống lúa). NDWI, sử dụng các dải phổ hồng ngoại gần (NIR) và xanh lá (Green), giúp phát hiện các khu vực có nước nhờ đặc tính hấp thụ mạnh ở dải NIR của nước. NDWI1 và NDWI2, với sự kết hợp các dải hồng ngoại ngắn (SWIR), tăng cường khả năng phân biệt giữa mặt nước tự nhiên và đất ngập nước canh tác, hỗ trợ xác định các khu vực có CN cao (98) như mặt nước hoặc CN dao động (70-85) như ruộng lúa. GSI bổ sung thông tin về độ ẩm đất, giúp phân loại chính xác hơn giữa đất trống (nhóm 3) và đất canh tác ngập nước. Trong khi đó, NDBI

(Normalized Difference Built-up Index) và BSI (Bare Soil Index) được sử dụng để nhận diện Nhóm 4 (Khu xây dựng) và Nhóm 3 (Đất trống). NDBI, dựa trên sự khác biệt giữa hồng ngoại ngắn và hồng ngoại gần, nổi bật trong việc phát hiện bề mặt không thấm nước như nhà cửa (CN 98), trong khi BSI giúp xác định đất trống hoặc đất đê hoang (CN 67-89) nhờ phản xạ đặc trưng của đất tràn ở dải SWIR.

Các dải quang phổ cơ bản Red, Green, Blue cung cấp thông tin trực tiếp về màu sắc và phản xạ bề mặt, hỗ trợ phân loại trực quan và bổ sung cho các chỉ số tổng hợp, đặc biệt trong việc phân biệt Đất trống và Khu xây dựng. Dữ liệu radar Sentinel-1 (VV và VH) mang lại giá trị lớn trong việc phân loại các khu vực có cấu trúc bề mặt phức tạp, chẳng hạn như Ruộng bậc thang (nhóm 6). Độ phân cực VV và VH phản ánh độ nhám bề mặt và cấu trúc thực vật, giúp phát hiện các hệ thống terracing (CN 60-80) nhờ sự khác biệt trong tín hiệu tán xạ giữa đất canh tác có cấu trúc và đất trống. Cuối cùng, Slope (độ dốc) là yếu tố địa hình quan trọng, ảnh hưởng trực tiếp đến dòng chảy bề mặt và khả năng thấm nước, đặc biệt trong các nhóm như Ruộng bậc thang và Rừng rậm, nơi địa hình dốc có thể làm tăng nguy cơ lũ quét nếu không được quản lý tốt.

Bảng 3-23. Mô tả các tham số đầu vào trong phân loại thảm phủ

T T	Tham số	Mô tả	Ý nghĩa	Lớp tác động chính
1	NDVI	Chỉ số chênh lệch thực vật chuẩn hóa (NIR - Red)/(NIR + Red)	Dánh giá mật độ và sức khỏe thực vật dựa trên phản xạ hồng ngoại gần (NIR) và đỏ (Red), rất hữu ích trong việc phân biệt Rừng rậm và Rừng thưa dựa trên sự khác biệt về độ che phủ và sinh khối.	Rừng rậm (2), Rừng thưa (5)
2	EVI	Chỉ số thực vật cải tiến $(2.5*(NIR - Red)/(NIR + 6*Red - 7.5*Blue + 1))$	Cải thiện NDVI, giảm nhiễu khí quyển và đất nền, đặc biệt hiệu quả trong việc xác định các khu vực có thực vật dày đặc so với khu vực thưa thớt.	Rừng rậm (2), Rừng thưa (5)
3	GND VI	Chỉ số chênh lệch thực vật xanh (NIR - Green)/(NIR + Green)	Dánh giá sức khỏe thực vật qua phản xạ kẽm xanh, hỗ trợ phân loại Đất trống lúa nhờ khả năng phát hiện sự thay đổi quang phổ trong môi trường ngập nước.	Rừng rậm (2), Đất trống lúa (7)
4	GRVI	Chỉ số thực vật xanh-đỏ (Green - Red)/(Green + Red)	Phân biệt thực vật qua sự chênh lệch giữa kẽm xanh và đỏ, hỗ trợ phát hiện Đất trống nhờ phản xạ đặc trưng của đất không che phủ.	Rừng thưa (5), Đất trống (3)
5	NDWI 1	Chỉ số nước (Green - NIR)/(Green + NIR)	Phát hiện nước qua sự hấp thụ mạnh ở dải NIR, giúp nhận diện Mặt nước và Đất trống lúa nhờ đặc tính quang phổ của nước bề mặt.	Mặt nước (1), Đất trống lúa (7)
6	NDWI 2	Chỉ số nước cải tiến (Green - SWIR)/(Green + SWIR)	Tăng cường phát hiện nước qua dải hồng ngoại ngắn (SWIR), giảm nhiễu từ thực vật, hỗ trợ phân biệt giữa mặt nước tự nhiên và đất ngập nước canh tác.	Mặt nước (1), Đất trống lúa (7)

T T	Tham số	Mô tả	Ý nghĩa	Lớp tác động chính
7	GSI	Chỉ số xanh-SWIR (Green - SWIR)/(Green + SWIR)	Đánh giá độ ẩm đất và nước trên bề mặt, hỗ trợ phân loại Đất trống lúa và Đất trống nhờ khả năng phát hiện sự thay đổi độ ẩm.	Đất trống lúa (7), Đất trống (3)
8	NDBI	Chỉ số xây dựng (SWIR - NIR)/(SWIR + NIR)	Phát hiện khu vực xây dựng qua phản xạ SWIR cao, nổi bật trong việc nhận diện Khu xây dựng nhờ đặc trưng bề mặt không thấm nước.	Khu xây dựng (4)
9	NDWI	Chỉ số nước tổng quát (NIR - SWIR)/(NIR + SWIR)	Nhận diện nước và vùng ngập qua sự chênh lệch NIR-SWIR, ứng dụng rộng rãi trong phân tích thủy văn cho Mặt nước và Đất trống lúa.	Mặt nước (1), Đất trống lúa (7)
10	BSI	Chỉ số đất trần ((SWIR + Red) - (NIR + Blue))/((SWIR + Red) + (NIR + Blue))	Phát hiện đất trần qua phản xạ đặc trưng của đất ở dải SWIR, hỗ trợ phân loại Đất trống nhờ khả năng nhận diện đất không che phủ.	Đất trống (3)
11	Slope	Độ dốc địa hình (tính bằng độ hoặc phần trăm)	Đánh giá độ nghiêng của bề mặt, ảnh hưởng trực tiếp đến dòng chảy và khả năng thấm nước, quan trọng trong phân loại Ruộng bậc thang và Rừng rậm.	Ruộng bậc thang (6), Rừng rậm (2)
12	Red	Dải đỏ từ dữ liệu đa phỏ	Phản xạ ánh sáng đỏ, hỗ trợ phân biệt bề mặt như Đất trống và Khu xây dựng, cung cấp cơ sở cho các chỉ số tổng hợp khác.	Đất trống (3), Khu xây dựng (4)
13	Green	Dải xanh từ dữ liệu đa phỏ	Phản xạ ánh sáng xanh, hỗ trợ phát hiện nước và thực vật, đặc biệt hiệu quả trong nhận diện Mặt nước và Đất trống lúa.	Mặt nước (1), Đất trống lúa (7)
14	Blue	Dải xanh lam từ dữ liệu đa phỏ	Phản xạ ánh sáng xanh lam, hỗ trợ phân biệt đất và nước, giảm nhiễu khí quyển trong phân loại Đất trống và Mặt nước.	Đất trống (3), Mặt nước (1)
15	VV	Độ phân cực thẳng đứng (Vertical-Vertical) từ radar Sentinel-1	Phản ánh độ nhám bề mặt và cấu trúc thực vật, hiệu quả trong phân loại Ruộng bậc thang nhờ tín hiệu tán xạ đặc trưng.	Ruộng bậc thang (6), Đất trống (3)
16	VH	Độ phân cực chéo (Vertical-Horizontal) từ radar Sentinel-1	Phân tích cấu trúc thực vật và bề mặt phức tạp, nhạy với sinh khối và hỗ trợ phân loại Rừng rậm và Ruộng bậc thang.	Rừng rậm (2), Ruộng bậc thang (6)

## 2. Mô hình CNN

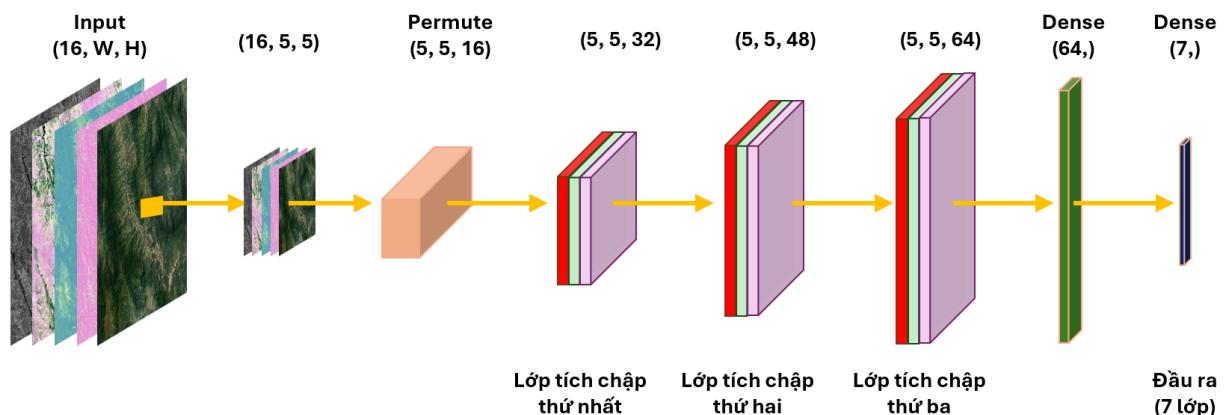
Mô hình CNN là mô hình học sâu được thiết kế để tận dụng mối quan hệ không gian giữa các pixel lân cận, một yếu tố quan trọng trong phân loại LULC do sự khác biệt về cấu trúc bề mặt giữa các lớp. Mô hình bao gồm ba lớp tích chập chính:

**Lớp tích chập đầu tiên:** Sử dụng kernel  $5 \times 5$  để nắm bắt các đặc trưng không gian rộng hơn, chẳng hạn như ranh giới giữa rừng và đất trống. Hàm kích hoạt LeakyReLU (negative\_slope=0.1) được sử dụng để tránh vấn đề "dying ReLU", trong khi bình thường hóa hàng loạt (batch normalization) giúp ổn định quá trình huấn luyện.

**Lớp tích chập giãn nở:** Áp dụng dilation rate  $2 \times 2$  để mở rộng trường tiếp nhận mà không tăng số lượng tham số, cho phép mô hình phát hiện các mẫu không gian lớn hơn như các vùng nước hoặc khu xây dựng. Lớp này sử dụng kernel  $3 \times 3$  và cũng được theo sau bởi LeakyReLU và batch normalization.

**Lớp tích chập thứ ba:** Sử dụng kernel  $3 \times 3$  để xử lý các đặc trưng kết hợp từ các lớp trước, tập trung vào các chi tiết nhỏ hơn như ranh giới giữa ruộng bậc thang và đất trống lúa.

Sau các lớp tích chập, một lớp tổng hợp toàn cục (global average pooling) được sử dụng để giảm số chiều, tiếp theo là hai lớp dày đặc (dense) với 64 và 7 nơ-ron, tương ứng với số lớp đầu ra. Lớp dropout (0.2) được thêm vào để ngăn chặn quá khớp. Mô hình được tối ưu hóa bằng thuật toán Adam với tốc độ học ban đầu 0.00005 và clipnorm=1.0 để kiểm soát gradient lớn. Hàm mất mát sparse categorical crossentropy được sử dụng do nhãn là các giá trị nguyên. Cấu trúc mô hình được minh họa như sau:



Hình 3-29. Cấu trúc mạng CNN

Trong thiết kế mô hình CNN, số lượng đặc trưng (feature maps) được tăng dần qua các lớp tích chập, từ 32 ở lớp đầu tiên lên 48 ở lớp giãn nở và 64 ở lớp cuối cùng. Chiến lược này nhằm mục đích học các đặc trưng phức tạp hơn ở các tầng sâu, tận dụng mối quan hệ không gian giữa các pixel lân cận để phân biệt các lớp LULC có cấu trúc bề mặt đa dạng, chẳng hạn như ranh giới giữa rừng rậm và đất trống hoặc các mẫu không gian lớn của khu xây dựng. So với các phương pháp thu hẹp dần (ví dụ: sử dụng các lớp max pooling hoặc stride lớn để giảm kích thước không gian), cách tiếp cận tăng dần đặc trưng giữ nguyên độ phân giải không gian qua các lớp tích chập, cho phép mô hình duy trì thông tin chi tiết về ranh giới và kết cấu bề mặt. Điều này đặc biệt quan trọng trong phân loại LULC, nơi các đặc trưng không gian như ranh giới giữa ruộng bậc thang và

đất trồng lúa có thể ảnh hưởng đến kết quả thủy văn. Ngoài ra, việc sử dụng lớp tổng hợp toàn cục (global average pooling) thay vì các lớp thu hẹp dần ở giai đoạn sau giúp giảm số chiều mà không làm mất thông tin tổng quát, đồng thời giảm nguy cơ quá khóp so với các lớp dày đặc truyền thống. Tuy nhiên, nhược điểm của chiến lược này là yêu cầu tính toán cao hơn do kích thước không gian được giữ nguyên qua các lớp, có thể làm tăng thời gian huấn luyện và tiêu tốn tài nguyên trên các tập dữ liệu lớn. Hơn nữa, việc tăng số đặc trưng có thể dẫn đến nguy cơ học các đặc trưng dư thừa nếu không được kiểm soát chặt chẽ bởi các kỹ thuật như dropout và bình thường hóa hàng loạt.

Bảng 3-24. Mô tả mô hình CNN

TT	Lớp	Loại	Kích thước	Số filter/units	Đặc điểm đặc biệt	Kích thước đầu ra
0	Input	Input	-	-	16 bands, patch $5 \times 5$	(16, 5, 5)
1	Permute	Reshape	-	-	Chuyển (16,5,5) $\rightarrow$ (5,5,16)	(5, 5, 16)
2	Conv2D	Tích chập	$5 \times 5$	32	padding='same', nǎm bắt lân cận	(5, 5, 32)
3	LeakyReLU	Kích hoạt	-	-	negative_slope=0.1	(5, 5, 32)
4	BatchNorm	Chuẩn hóa	-	-	Chuẩn hóa theo batch	(5, 5, 32)
5	Conv2D	Tích chập giãn nở	$3 \times 3$	48	dilation_rate=(2,2), mở rộng trường tiếp nhận	(5, 5, 48)
6	LeakyReLU	Kích hoạt	-	-	negative_slope=0.1	(5, 5, 48)
7	BatchNorm	Chuẩn hóa	-	-	Chuẩn hóa theo batch	(5, 5, 48)
8	Conv2D	Tích chập	$3 \times 3$	64	padding='same', xử lý đặc trưng kết hợp	(5, 5, 64)
9	LeakyReLU	Kích hoạt	-	-	negative_slope=0.1	(5, 5, 64)
10	BatchNorm	Chuẩn hóa	-	-	Chuẩn hóa theo batch	(5, 5, 64)
11	GlobalAvgPool	Pooling	-	-	Trung bình toàn cục	(64)
12	Dense	Kết nối đầy đủ	-	64	Tầng ẩn	(64)
13	LeakyReLU	Kích hoạt	-	-	negative_slope=0.1	(64)
14	BatchNorm	Chuẩn hóa	-	-	Chuẩn hóa theo batch	(64)
15	Dropout	Regularization	-	-	rate=0.2, giảm overfitting	(64)
16	Dense	Kết nối đầy đủ	-	7	activation='softmax', phân loại 7 lớp	(7)

Tập dữ liệu bao gồm 100,000 mẫu (trên tổng số 122,393 mẫu dựa trên thuật toán không ché só lượng mẫu tối đa) được chia thành tập huấn luyện (80%) và tập xác thực (20%) bằng phương pháp phân chia ngẫu nhiên có phân tầng để đảm bảo phân bố nhẫn đồng đều. Toàn bộ mẫu được lấy ở dạng vùng và chuyển về dạng điểm thông qua công cụ rasterize trong QGIS. Mô hình được huấn luyện trong 50 epoch với kích thước batch là 32. Các callback được sử dụng bao gồm:

**EarlyStopping:** Dừng huấn luyện nếu hàm mất mát xác thực không cải thiện sau 5 epoch, đồng thời khôi phục trọng số tốt nhất.

**ReduceLROnPlateau:** Giảm tốc độ học xuống 50% nếu hàm mất mát xác thực không cải thiện sau 3 epoch, với tốc độ học tối thiểu là 1e-6.

Phân bố nhãn trong tập dữ liệu được trình bày trong Bảng 3, phản ánh sự đa dạng của các lớp LULC trong khu vực nghiên cứu. Bài báo không trình bày bản đồ tập dữ liệu huấn luyện và kiểm tra của mô hình do các mẫu được lấy ở dạng vùng, sau đó chuyển về dạng điểm với độ phân giải là 12,5m (quá nhỏ và gần nhau so với kích thước khu vực nghiên cứu) gây ra hiệu ứng chồng lấn và thể hiện không rõ ràng trong các bản đồ.

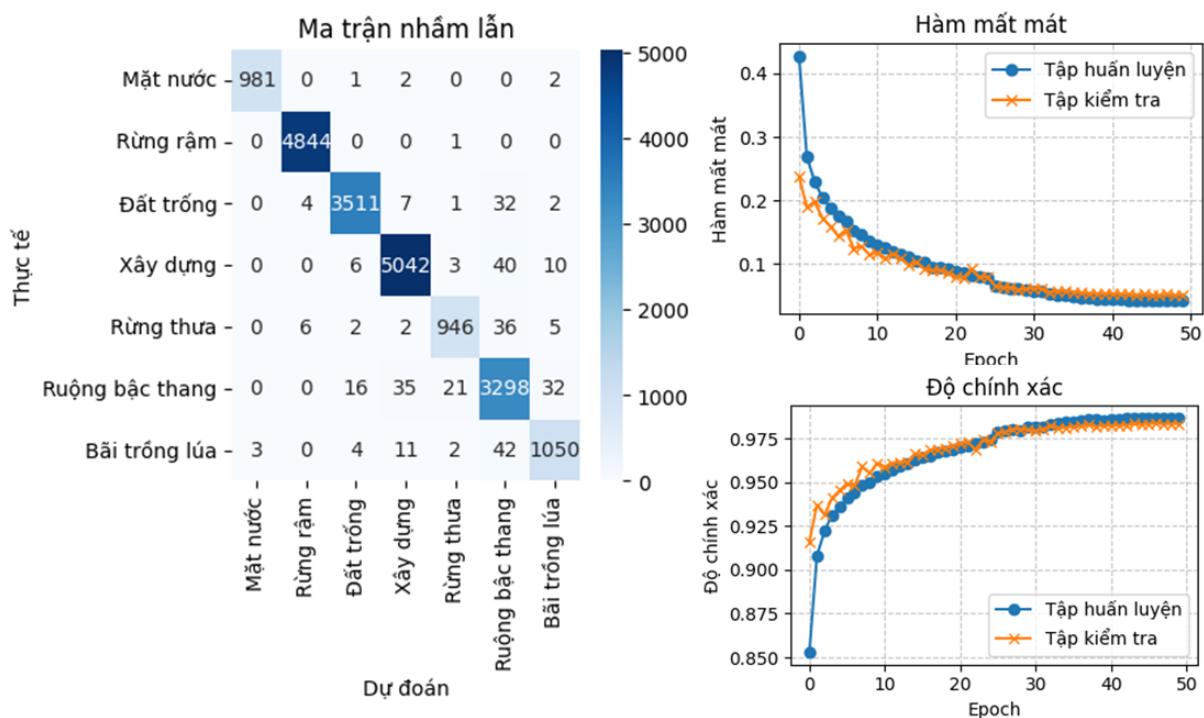
Bảng 3-25. Phân bố nhãn trong tập dữ liệu huấn luyện

Lớp	Tên	Số mẫu	Tỷ lệ
1	Mặt nước	4,929	4.93%
2	Rừng rậm	24,223	24.22%
3	Đất trống	17,784	17.78%
4	Khu xây dựng	25,505	25.51%
5	Rừng thưa	4,986	4.99%
6	Ruộng bậc thang	17,012	17.01%
7	Đất trống lúa	5,561	5.56%

Tập dữ liệu huấn luyện thể hiện sự mất cân bằng đáng kể giữa các lớp, với các lớp như "Khu xây dựng" (25.51%) và "Rừng rậm" (24.22%) chiếm tỷ lệ lớn, trong khi các lớp như "Mặt nước" (4.93%) và "Rừng thưa" (4.99%) có số mẫu ít hơn. Tuy nhiên, mô hình CNN vẫn đạt độ chính xác xác thực cao (98.32%) nhờ vào các chiến lược thiết kế hiệu quả. Thứ nhất, việc sử dụng phân chia ngẫu nhiên có phân tầng (stratified random splitting) đảm bảo rằng tỷ lệ các lớp được duy trì trong cả tập huấn luyện và xác thực, giúp mô hình học được đặc trưng của các lớp thiểu số. Thứ hai, các lớp tích chập giãn nở và kernel lớn cho phép mô hình nắm bắt các mẫu không gian phức tạp, đặc biệt là ở các lớp hiếm như "Mặt nước", vốn có đặc trưng quang học rõ rệt (NDWI, NDWI1). Ngoài ra, hàm mất mát sparse categorical crossentropy và kỹ thuật dropout (0.2) giúp giảm thiểu tình trạng quá khớp đối với các lớp chiếm ưu thế, đảm bảo mô hình tổng quát hóa tốt trên tất cả các lớp. Những yếu tố này đã giảm thiểu tác động của dữ liệu mất cân bằng, dẫn đến hiệu suất phân loại đồng đều, như được thể hiện trong ma trận nhầm lẫn.

### 3. Kết quả xây dựng bản đồ thảm phủ

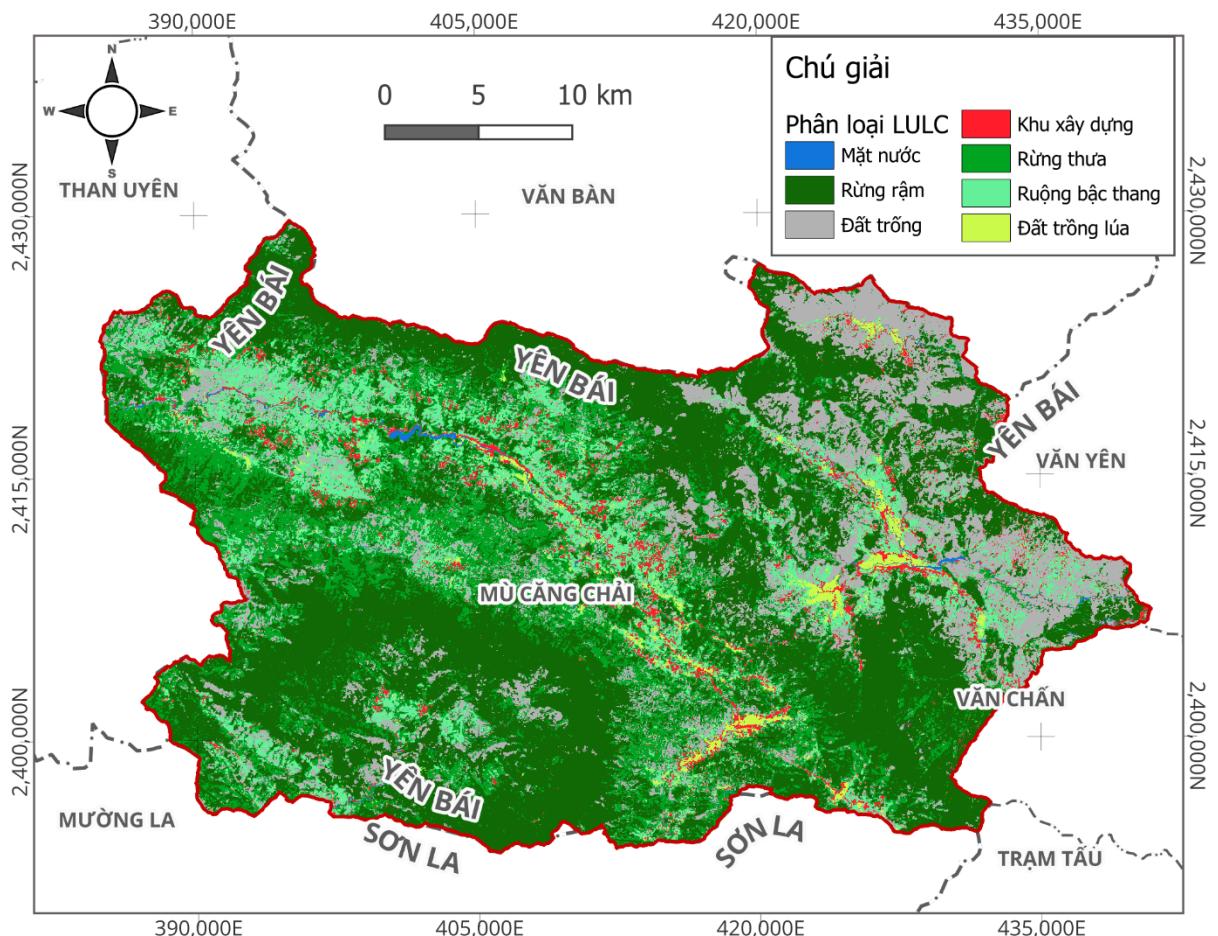
Mô hình CNN đạt độ chính xác huấn luyện là 98.68% và độ chính xác xác thực là 98.32% sau 50 epoch. Quá trình hội tụ được thể hiện trong Hình 3, cho thấy hàm mất mát giảm đều và ổn định, đặc biệt sau khi tốc độ học được điều chỉnh tự động. Ma trận nhầm lẫn (Hình 4) cho thấy mô hình phân loại chính xác hầu hết các lớp, với một số nhầm lẫn nhỏ giữa rừng rậm (lớp 2) và rừng thưa (lớp 5) do sự tương đồng về mật độ thực vật. Độ chính xác và độ nhạy của từng lớp được trình bày trong Bảng 3-26.



Hình 3-30. Ma trận nhầm lẫn, đồ thị hàm mất mát và độ chính xác trong quá trình xây dựng mô hình CNN trong phân loại thảm phủ tại Mù Cang Chải

Bảng 3-26. Độ chính xác và độ nhạy của từng lớp phân loại thảm phủ

Lớp	Tên	Độ chính xác (%)	Độ nhạy (%)	Phương pháp nghiên cứu	ESRI Land Cover	JAXA ALOS
1	Mặt nước	99.7	99.49	Có	Có	Có
2	Rừng rậm	99.79	99.98	Có (Good)	Có (Trees)	Có (chưa phân loại)
3	Đất trống	99.18	98.71	Có (Poor)	Có (Bare ground)	Có (Barren)
4	Khu xây dựng	98.88	98.84	Có (Impervious)	Có (Built-up)	Có (Urban)
5	Rừng thưa	97.13	94.88	Có (Fair/Poor)	Không	Có (chưa phân loại)
6	Ruộng bậc thang	95.65	96.94	Có (Contoured)	Không	Không
7	Đất trồng lúa	95.37	94.42	Có (Wet meadow)	Không	Có (rice)



Hình 3-31. Bản đồ LULC được tạo ra từ mô hình CNN

Bằng cách sử dụng 16 tham số đầu vào từ Sentinel-1, Sentinel-2 và ALOS, mô hình CNN đã phân loại thành công bảy lớp LULC, mỗi lớp có mối liên hệ rõ ràng với các quá trình thủy văn như thẩm thấu và dòng chảy. So với các sản phẩm LULC hiện có như ESRI Land Cover và JAXA ALOS, phương pháp này cung cấp các lớp phân loại chuyên biệt hơn, đặc biệt là ruộng bậc thang, đất trồng lúa và chất lượng rừng ( thông qua CSDL rừng), vốn quan trọng trong mô phỏng dòng chảy ở các khu vực nông nghiệp miền núi phía Bắc Việt Nam.

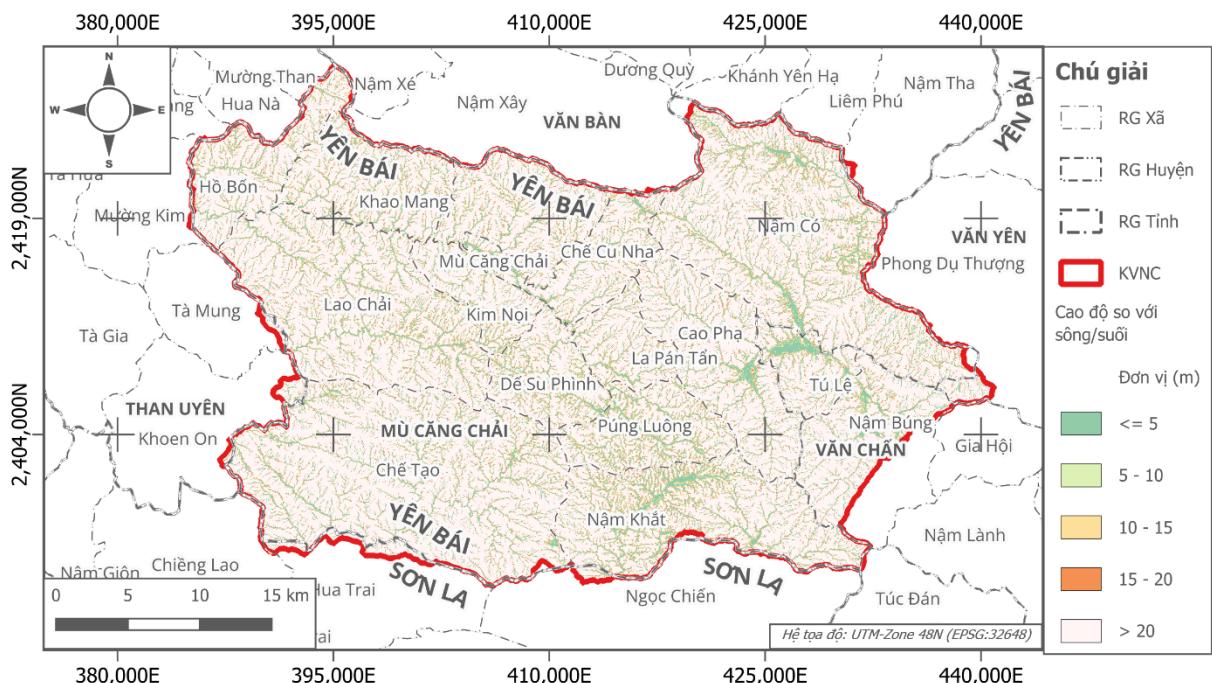
### 3.2.2 Xây dựng bộ cơ sở dữ liệu không gian khác phục vụ phân vùng lũ quét

Trong bối cảnh dự đoán nguy cơ lũ quét, việc lựa chọn và hiểu rõ các đặc trưng dữ liệu là yếu tố cốt lõi để xây dựng các mô hình học máy hiệu quả. Các đặc trưng được sử dụng trong nghiên cứu này thuộc bốn nhóm chính: địa hình, thủy văn, thực phủ và khí tượng. Mỗi nhóm đặc trưng phản ánh một khía cạnh cụ thể của môi trường tự nhiên và đóng vai trò quan trọng trong việc xác định mức độ nguy cơ lũ quét. Trong nội dung này, nhóm nghiên cứu sẽ phân tích chi tiết các đặc trưng trong đó nhấn mạnh ý nghĩa vật lý, cách thu thập, và vai trò của chúng trong việc mô hình hóa nguy cơ lũ quét.

## 1. Đặc trưng địa hình:

### a. Cao độ so với sông suối

Cao độ so với sông suối là chênh lệch độ cao giữa một điểm trên địa hình và cao độ tại sông hoặc suối gần nhất theo hướng dòng chảy, được đo bằng mét. Đặc trưng này phản ánh mức độ dễ bị ngập lụt của một khu vực: các điểm có độ cao thấp so với sông suối thường dễ bị ngập hơn do gần với mực nước. Trong GIS, đặc trưng này được tính toán bằng cách trừ độ cao của điểm địa hình (từ DEM) cho cao độ tham chiếu của dòng chảy gần nhất, thường dựa trên phân tích không gian như thuật toán D8 hoặc D-infinity. Cao độ so với sông suối đặc biệt quan trọng trong các khu vực đồng bằng, nơi sự chênh lệch nhỏ về độ cao có thể dẫn đến khác biệt lớn về nguy cơ lũ lụt.



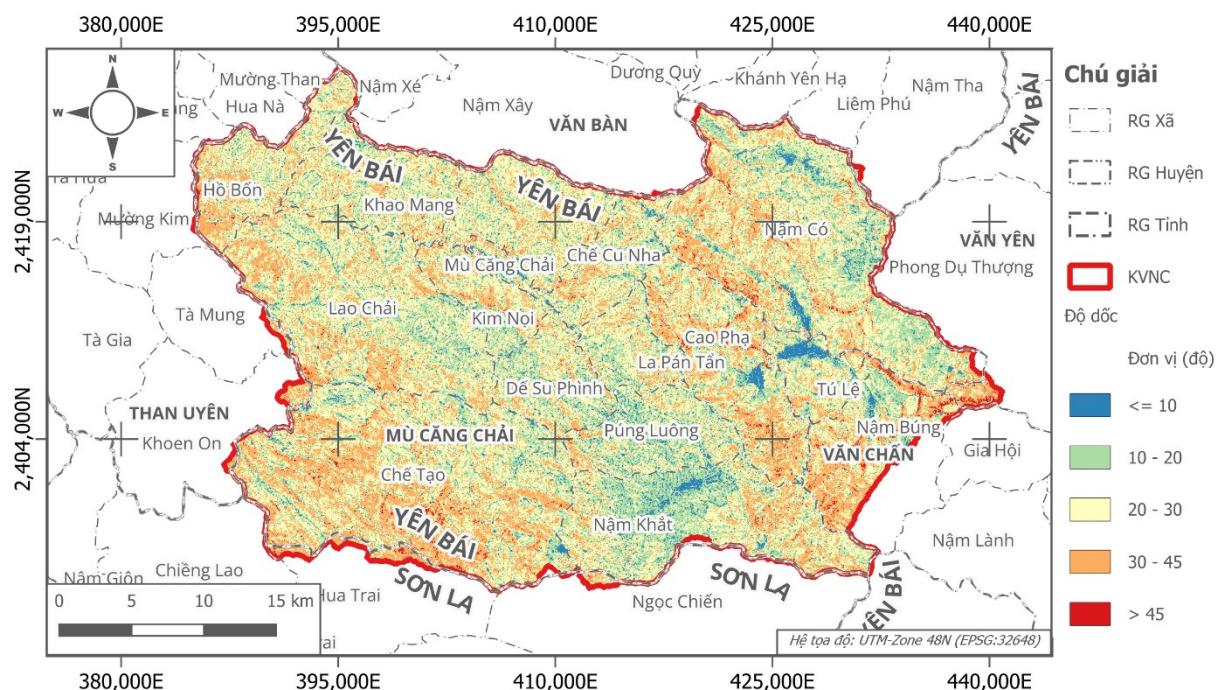
Hình 3-32. Cao độ các điểm đến sông/suối gần nhất (m)

Nghiên cứu này cho rằng các điểm có chênh lệch độ cao so với sông suối gần nhất nhỏ sẽ dễ bị dòng lũ quét chính ở trên sông tác động, đặc biệt là đối với các cơ sở hạ tầng xung quanh lòng dẫn. Do đó, các điểm có chênh lệch nhỏ sẽ bị gán các giá trị nguy cơ tương đương với các điểm thuộc lòng dẫn dựa trên thuật toán hướng dòng chảy trong phân tích thủy văn. Thuật toán xác định chênh lệch cao độ địa hình được nhóm nghiên cứu xây dựng và phát triển bằng các mã với ngôn ngữ Python. Kết quả được thể hiện trong Hình 3-32.

### b. Độ dốc bình quân lưu vực

Độ dốc địa hình biểu thị mức độ nghiêng của bề mặt địa hình, được tính bằng phần trăm hoặc độ, dựa trên sự thay đổi độ cao giữa các ô lưới trong DEM. Độ dốc ảnh hưởng trực tiếp đến tốc độ dòng chảy bề mặt: địa hình dốc làm tăng tốc độ dòng chảy, giảm khả năng tích tụ nước, trong khi địa hình bằng phẳng dễ gây ngập lụt do nước chảy

chậm. Trong mô hình hóa, độ dốc được tính toán bằng các thuật toán GIS như phương pháp Horn hoặc Zevenbergen-Thorne. Đặc trưng này rất quan trọng trong việc xác định các khu vực dễ bị ngập lụt do tích tụ nước hoặc xói mòn do dòng chảy mạnh.

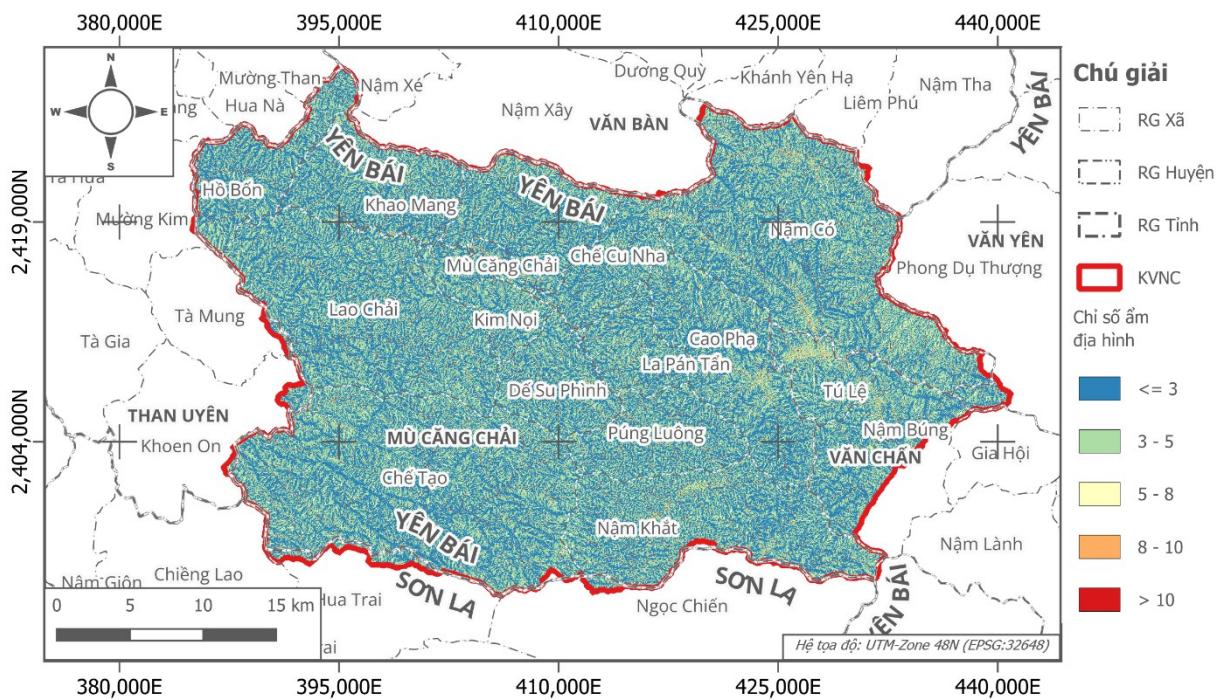


Hình 3-33. Độ dốc bình quân lưu vực địa hình khu vực nghiên cứu (độ)

Nghiên cứu này xem mỗi điểm là cửa ra của một lưu vực, do đó, dựa trên thuật toán hướng dòng chảy, các ô lưới thuộc mỗi lưu vực sẽ được xác định. Giá trị bình quân độ dốc của tất cả các ô lưới trong một lưu vực sẽ được gán là giá trị độ dốc bình quân lưu vực tại điểm tính. Kết quả thể hiện trong Hình 3-33. Độ dốc này khác với độ dốc điểm (bề mặt địa hình), nơi thể hiện độ dốc nội tại của từng điểm trên bề mặt.

### c. Chỉ số ẩm địa hình

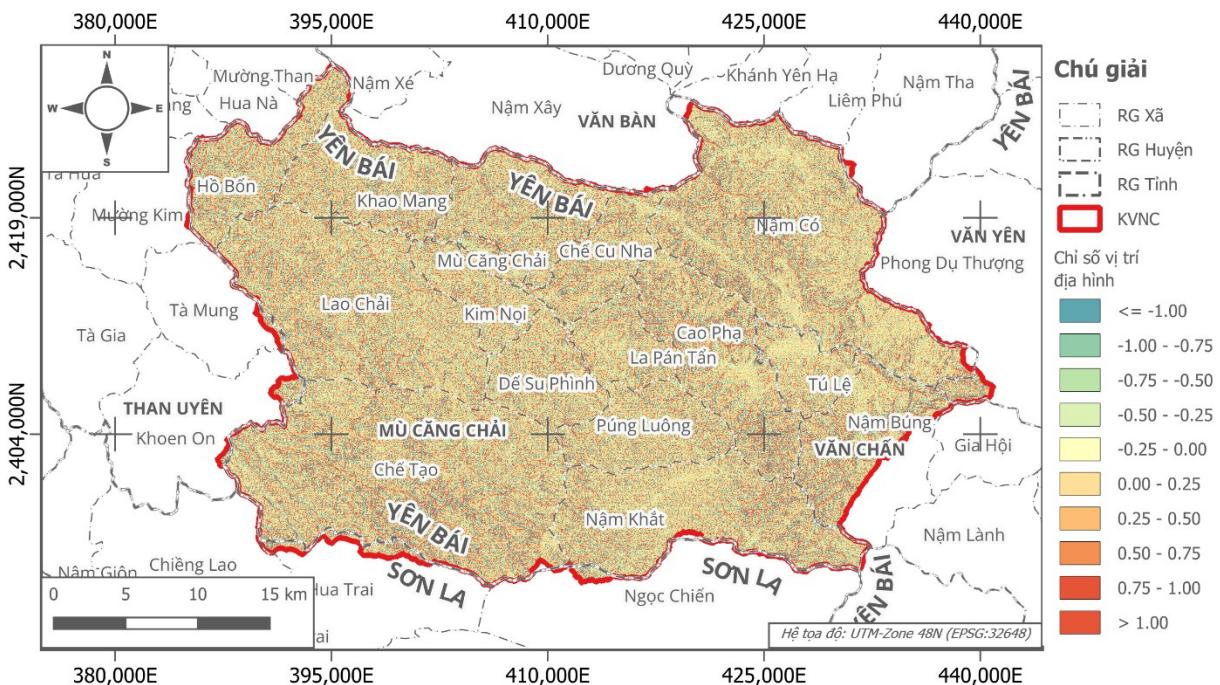
Chỉ số ẩm địa hình (Topographic Wetness Index - TWI) đo lường mức độ tích tụ nước tại một điểm dựa trên độ dốc và diện tích đóng góp dòng chảy. Giá trị TWI cao biểu thị các khu vực dễ tích tụ nước, chẳng hạn như vùng trũng hoặc thung lũng, trong khi giá trị thấp thường xuất hiện ở các khu vực cao hoặc dốc. Đặc trưng này rất hữu ích trong việc xác định các khu vực có nguy cơ ngập lụt cao do khả năng giữ nước lâu dài.



Hình 3-34. Chỉ số ẩm địa hình các điểm nằm trong khu vực nghiên cứu

#### d. Chỉ số vị trí địa hình

Chỉ số vị trí địa hình (Topographic Position Index - TPI) so sánh độ cao của một điểm với độ cao trung bình của các điểm xung quanh trong một bán kính nhất định. TPI dương biểu thị các điểm cao hơn khu vực xung quanh (như đỉnh đồi), trong khi TPI âm chỉ các điểm thấp hơn (như thung lũng).



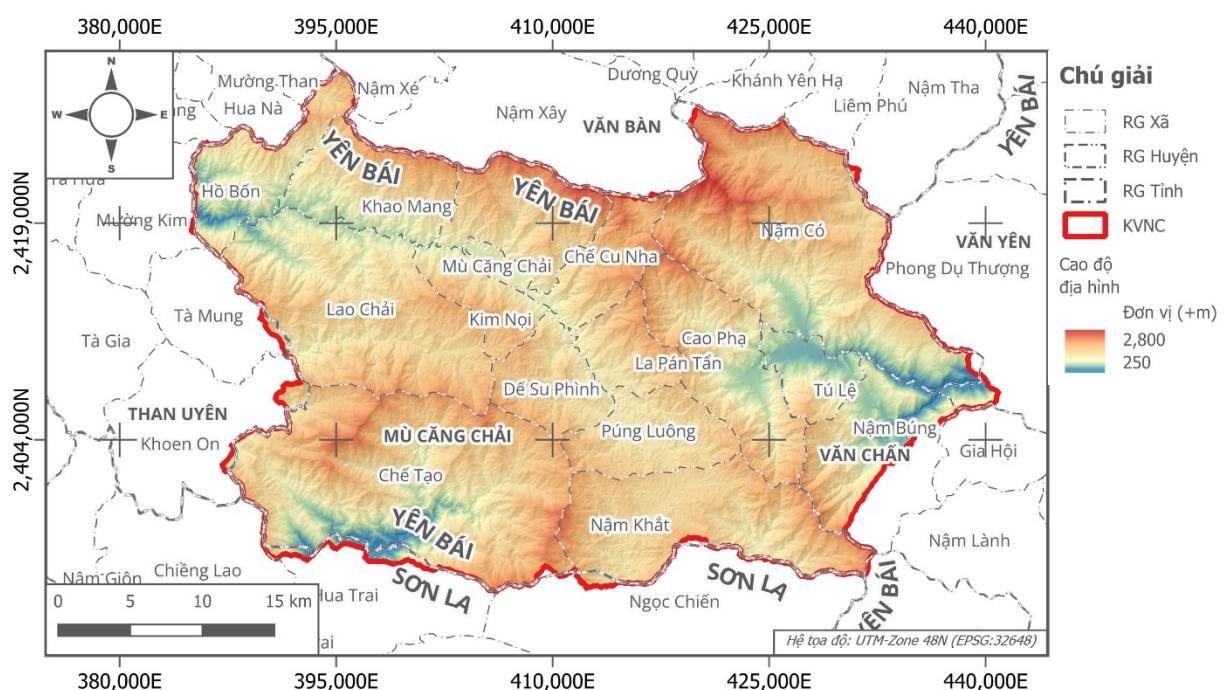
Hình 3-35. Chỉ số vị trí địa hình trong khu vực nghiên cứu

Đặc trưng chỉ số vị trí địa hình giúp xác định các khu vực dễ bị ngập lụt, đặc biệt là các vùng trũng có TPI âm. Trong GIS, TPI được tính toán bằng cách sử dụng các bộ

lọc không gian trên dữ liệu DEM, với bán kính lân cận được chọn dựa trên quy mô địa hình. Kết quả được thể hiện trong Hình 3-35.

#### e. Cao độ địa hình

Cao độ địa hình, hay độ cao tuyệt đối so với mực nước biển, được trích xuất trực tiếp từ DEM. Đặc trưng này cung cấp thông tin cơ bản về vị trí của khu vực trong không gian, ảnh hưởng đến khả năng thoát nước và mức độ ngập lụt. Các khu vực có độ cao thấp hơn thường có nguy cơ ngập lụt cao hơn, đặc biệt trong các sự kiện mưa lớn.

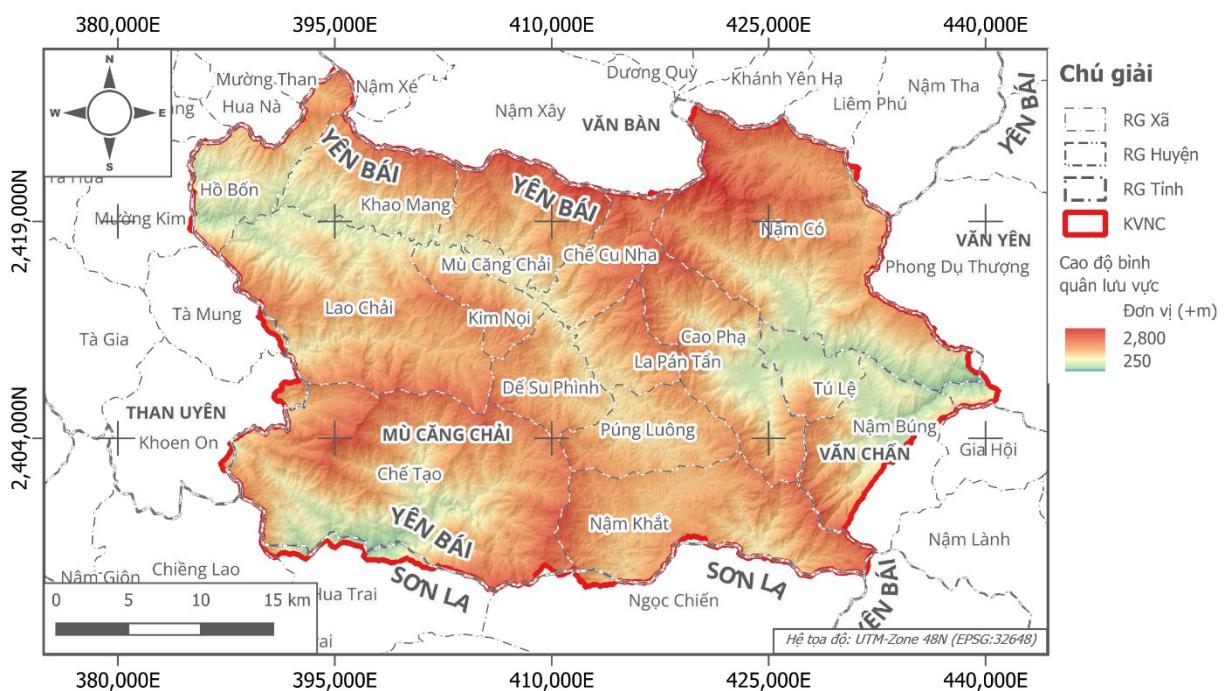


Hình 3-36. Bản đồ cao độ địa hình khu vực nghiên cứu

#### f. Cao độ bình quân lưu vực

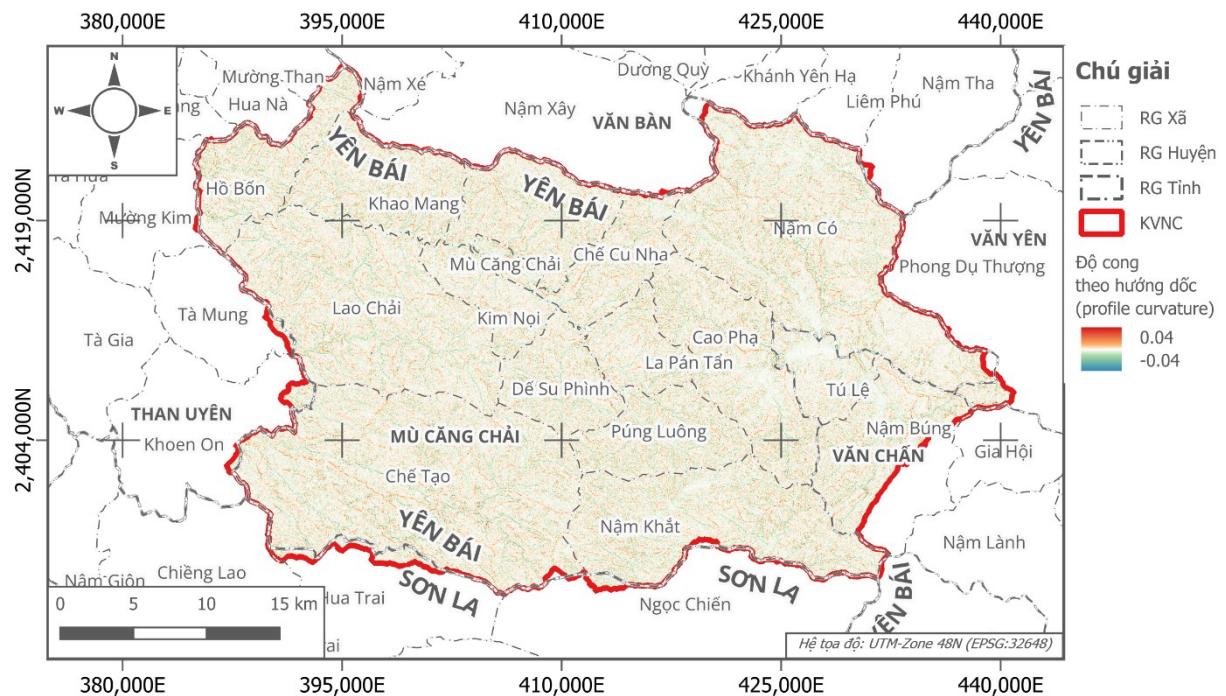
Cao độ bình quân lưu vực là giá trị trung bình của độ cao trong một lưu vực cụ thể, phản ánh đặc tính chung của địa hình trong khu vực thoát nước. Đặc trưng này hữu ích trong việc đánh giá khả năng thoát nước của toàn bộ lưu vực, các lưu vực có độ cao trung bình lớn hơn thường sẽ có khả năng thoát nước tốt hơn. Trong GIS, cao độ bình quân được tính bằng cách lấy trung bình các giá trị DEM trong phạm vi lưu vực, được xác định thông qua phân tích thủy văn.

Khác với cao độ địa hình phía trên, mỗi điểm trong cao độ bình quân lưu vực được xem là cửa ra lưu vực theo nguyên tắc ô lưới, do đó, toàn bộ các điểm thuộc lưu vực sẽ được tổng hợp bởi giá trị bình quân và đưa ra giá trị cao độ đại diện cho điểm cửa ra. Kết quả được thể hiện trong Hình 3-37, giá trị độ cao trong hình này cao hơn giá trị độ cao cục bộ do giá trị độ cao cục bộ là giá trị tại đúng điểm cửa ra, trong khi giá trị độ cao này là giá trị bình quân của lưu vực với điểm cửa ra là thấp nhất.



Hình 3-37. Bản đồ cao độ bình quân lưu vực khu vực nghiên cứu  
g. Độ cong địa hình (theo hướng dốc)

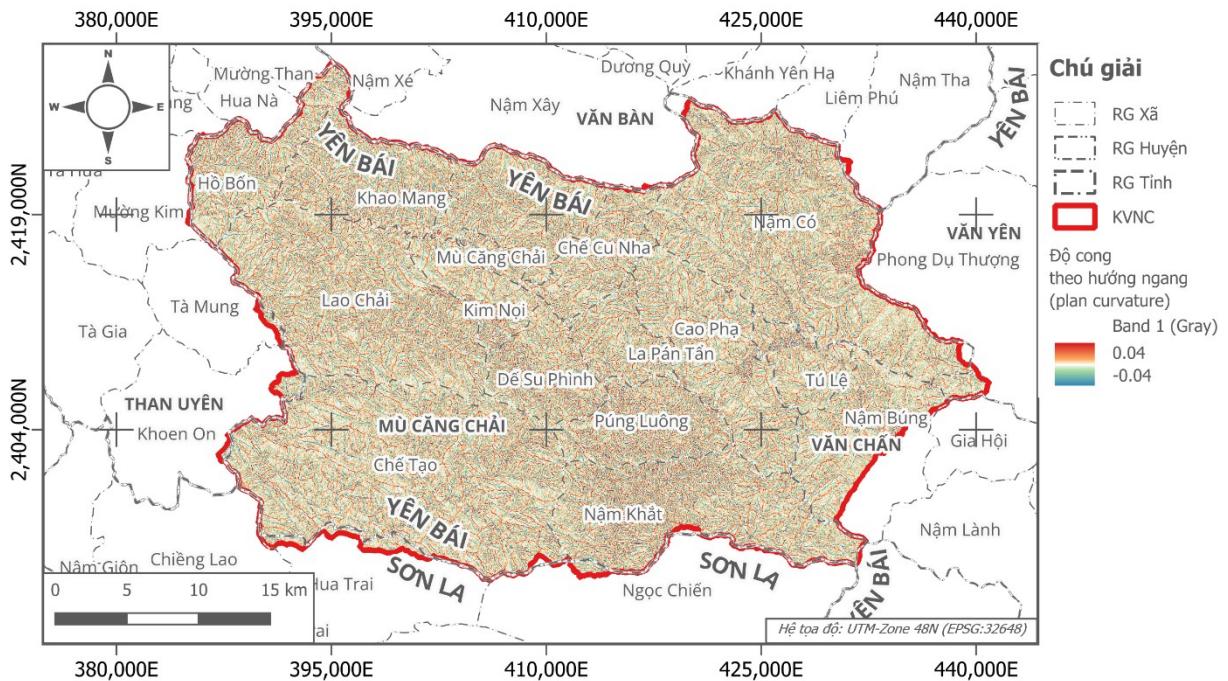
Độ cong theo hướng dốc (profile curvature) đo lường sự thay đổi độ dốc theo hướng dòng chảy, ảnh hưởng đến tốc độ và hướng của dòng chảy bề mặt. Độ cong dương (lồi) làm tăng tốc độ dòng chảy, trong khi độ cong âm (lõm) làm chậm dòng chảy, dẫn đến tích tụ nước. Đặc trưng này được tính toán từ DEM bằng cách sử dụng các đạo hàm bậc hai, thường thông qua các công cụ GIS như ArcGIS Spatial Analyst.



Hình 3-38. Độ cong địa hình theo hướng dốc khu vực nghiên cứu

## h. Độ cong địa hình (phương ngang)

Độ cong phương ngang (plan curvature) đo lường sự thay đổi độ dốc theo phương vuông góc với hướng dòng chảy, ảnh hưởng đến sự phân tán hoặc tập trung của dòng chảy. Độ cong âm biểu thị các khu vực tập trung dòng chảy (như thung lũng), trong khi độ cong dương biểu thị các khu vực phân tán (như sườn đồi). Đặc trưng này giúp xác định các khu vực dễ bị ngập lụt do sự tập trung dòng chảy.



Hình 3-39. Độ cong địa hình theo phương ngang khu vực nghiên cứu

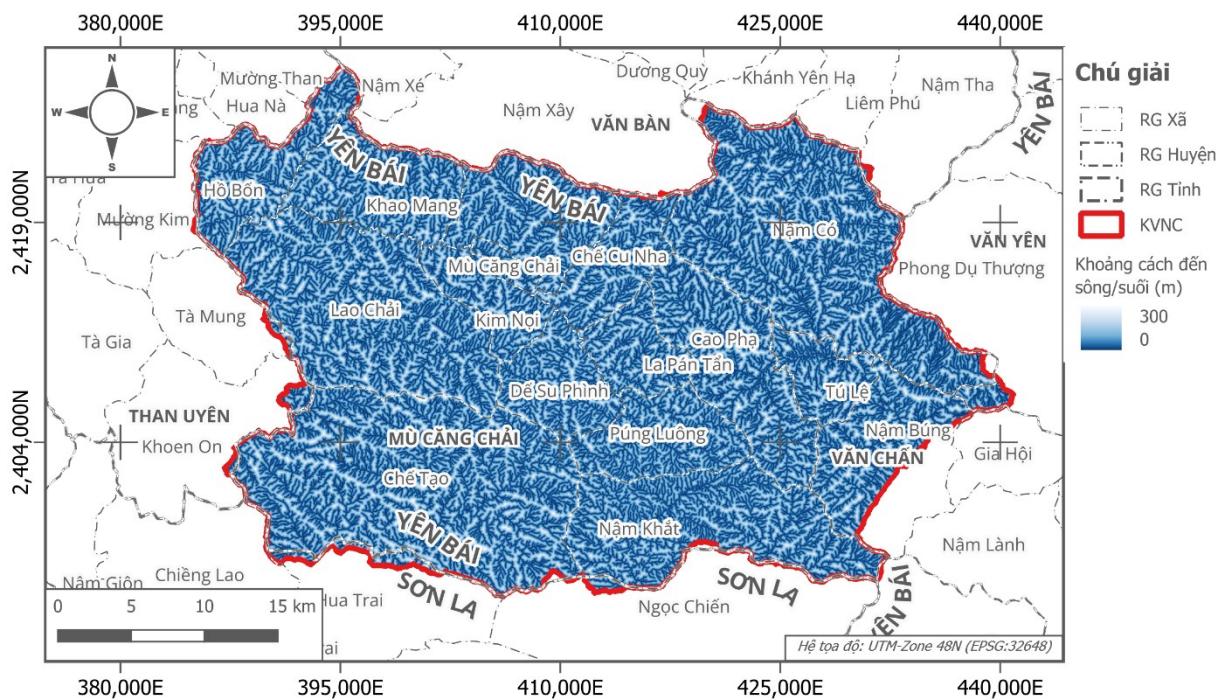
## 2. Đặc trưng thủy văn:

Các đặc trưng thủy văn mô tả các yếu tố liên quan đến dòng chảy và khả năng thoát nước của hệ thống thủy văn, đóng vai trò quan trọng trong việc dự đoán nguy cơ lũ lụt.

### a. Khoảng cách đến sông suối

Khoảng cách đến sông suối là khoảng cách Euclidean từ một điểm địa hình đến dòng chảy gần nhất, được đo bằng mét. Các khu vực gần sông suối thường có nguy cơ ngập lụt cao hơn do dễ bị ảnh hưởng bởi mực nước dâng cao. Trong GIS, đặc trưng này được tính toán bằng cách sử dụng các thuật toán phân tích không gian, cụ thể là proximity analysis, dựa trên mạng lưới sông suối được trích xuất từ DEM.

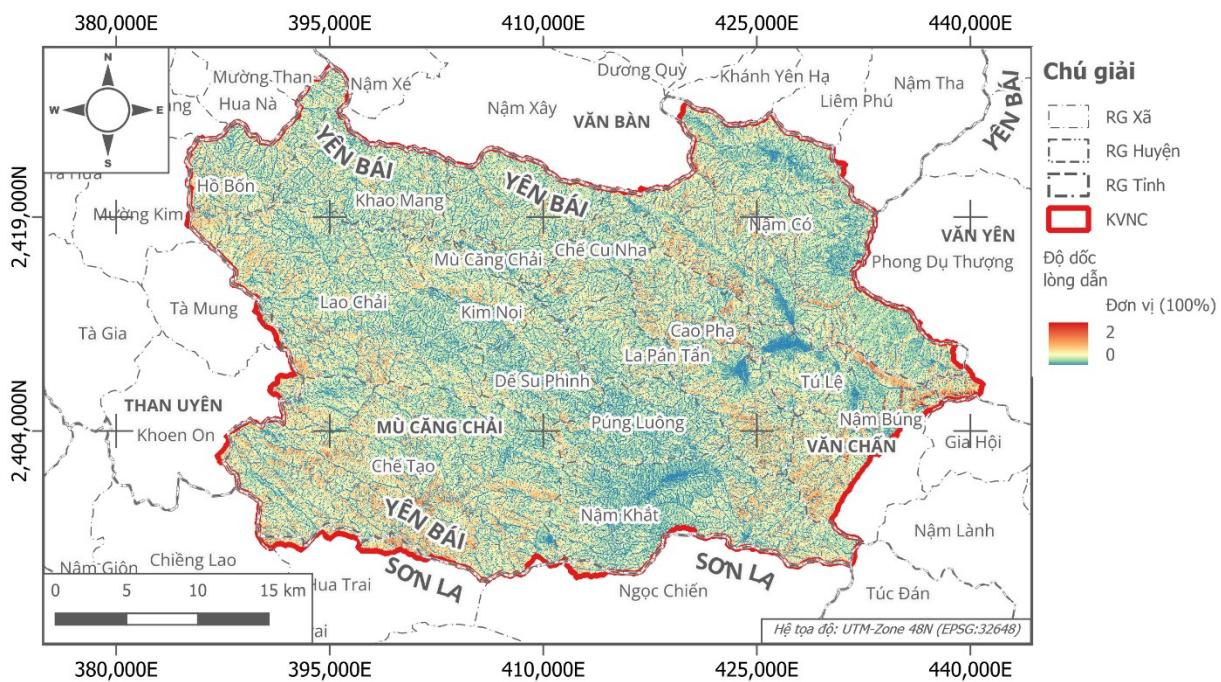
Mạng lưới sông, suối được phân tích từ địa hình theo nguyên lý thủy văn, theo độ phân giải  $12,5 \times 12,5$ m trong nghiên cứu này, mạng lưới sông suối được trích xuất có tổng số ô lưới tích lũy (flow accumulation) là 100 ô lưới. Điều này đồng nghĩa là dòng suối bắt đầu từ vị trí mà có diện tích lưu vực thương nguồn đạt tối thiểu là  $15.625\text{m}^2$  hay  $0,015625\text{km}^2$ . Kết quả xây dựng bản đồ khoảng cách đến sông, suối của từng vị trí được thể hiện trong Hình 3-40.



Hình 3-40. Khoảng cách đến sông, suối trong khu vực nghiên cứu

### b. Độ dốc lòng dẫn

Độ dốc lòng dẫn là độ dốc của dòng chảy tại một điểm trên sông hoặc suối, ảnh hưởng đến tốc độ dòng chảy và khả năng thoát nước. Các dòng chảy có độ dốc thấp thường dẫn đến tích tụ nước, làm tăng nguy cơ ngập lụt. Đặc trưng này được tính toán từ DEM bằng cách phân tích mạng lưới dòng chảy và xác định độ dốc tại các điểm dọc theo sông suối.

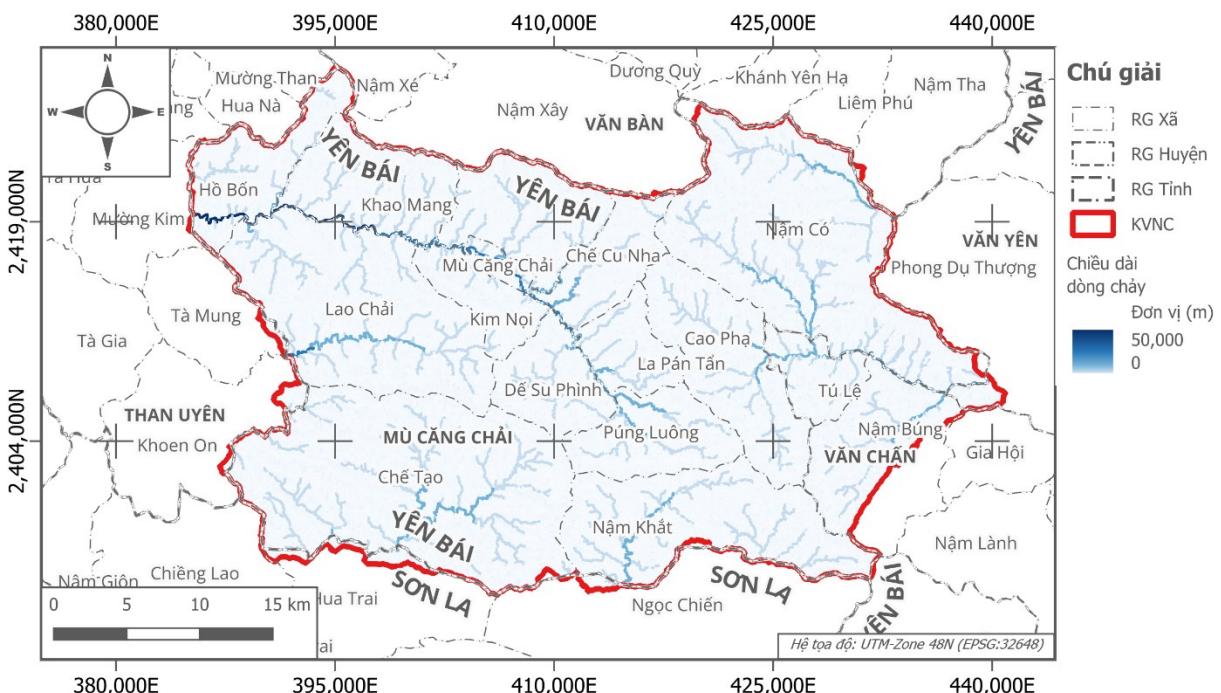


Hình 3-41. Độ dốc lòng dẫn khu vực nghiên cứu

Nghiên cứu này sử dụng độ dốc lòng dẫn tại một điểm bằng việc xác định hướng dòng chảy ngược về thượng nguồn và lấy trung bình độ dốc các điểm thuộc tuyến lòng dẫn để làm cơ sở xác định độ dốc lòng dẫn. Các suối có độ dốc lòng dẫn lớn dễ gây ra lũ quét hơn các suối có độ dốc nhỏ.

#### c. Chiều dài dòng chảy

Chiều dài dòng chảy là khoảng cách mà nước chảy từ một điểm nằm xa nhất trên đường phân thủy theo hướng dòng chảy đến điểm cần xác định. Nghiên cứu sử dụng thuật toán GIS kết hợp với D8 (hướng dòng chảy) để xác định chiều dài dòng chảy.



Hình 3-42. Chiều dài dòng chảy khu vực nghiên cứu

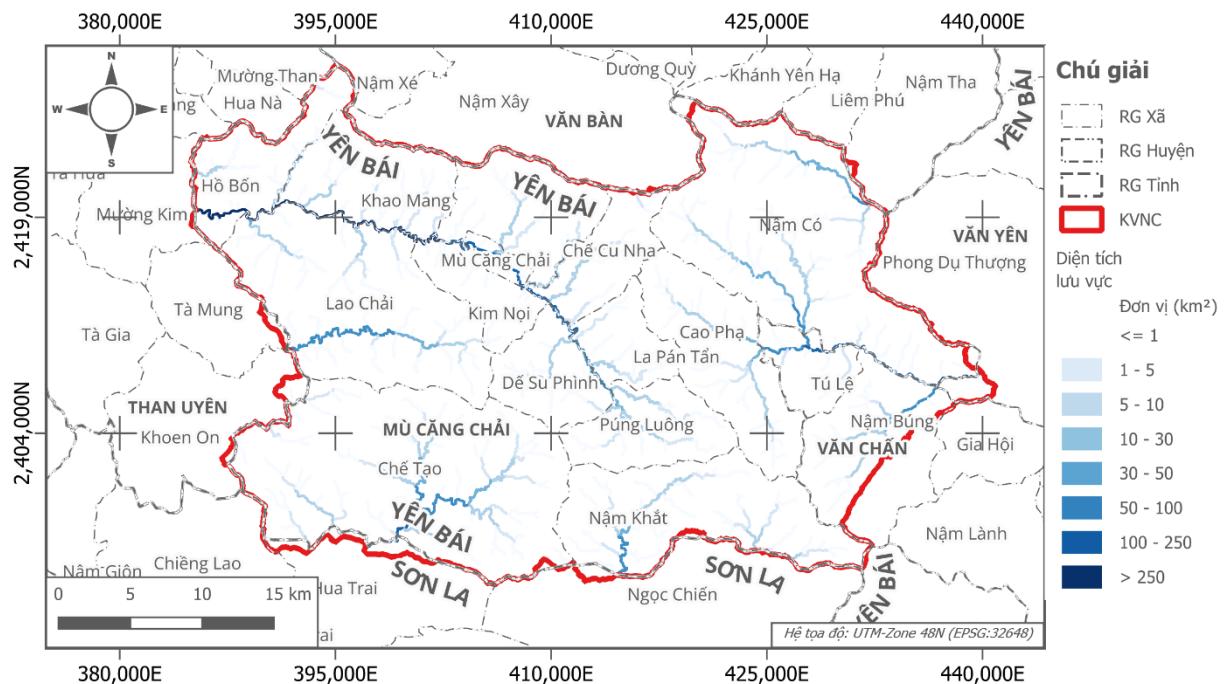
Theo thủy văn, chiều dài dòng chảy có tác động trực tiếp đến thời gian tập trung dòng chảy, đặc biệt trong nghiên cứu lũ quét, chiều dài dòng chảy là một trong các tham số quan trọng. Khi kết hợp với độ dốc lòng dẫn, việc ước tính thời gian tập trung dòng chảy dựa vào lượng mưa thường được thể hiện trong các tính toán thủy văn.

#### d. Diện tích lưu vực

Diện tích lưu vực là tổng diện tích của khu vực đóng góp dòng chảy vào một điểm cụ thể, được đo bằng mét vuông hoặc kilômét vuông. Lưu vực lớn hơn thường thu nhận lượng nước lớn hơn trong các sự kiện mưa, làm tăng nguy cơ lũ lụt. Đặc trưng này được trích xuất từ DEM bằng cách sử dụng các công cụ phân tích thủy văn như flow accumulation và kích thước pixel tính toán.

Flow accumulation là một số biểu thị tổng số lượng pixel có hướng dòng chảy đi qua nó. Do đó, giá trị này cũng biểu thị số lượng pixel trong một lưu vực mà tại đó là cửa ra của lưu vực. Khi đó, mỗi pixel sẽ có một diện tích (ở độ phân giải 12,5x12,5, diện

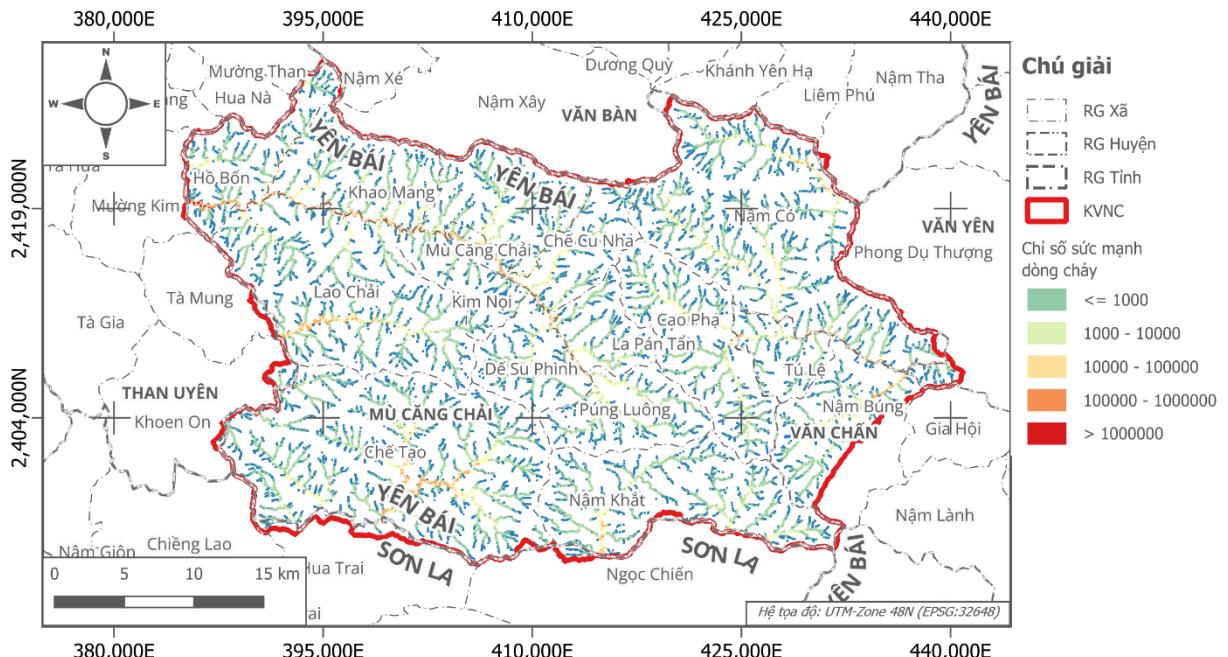
tích một pixel là 156,25m<sup>2</sup>). Tích số Flow accumulation và diện tích một pixel chính là bản đồ diện tích lưu vực cho khu vực nghiên cứu.



Hình 3-43. Bản đồ diện tích lưu vực khu vực nghiên cứu

#### e. Chỉ số sức mạnh dòng chảy

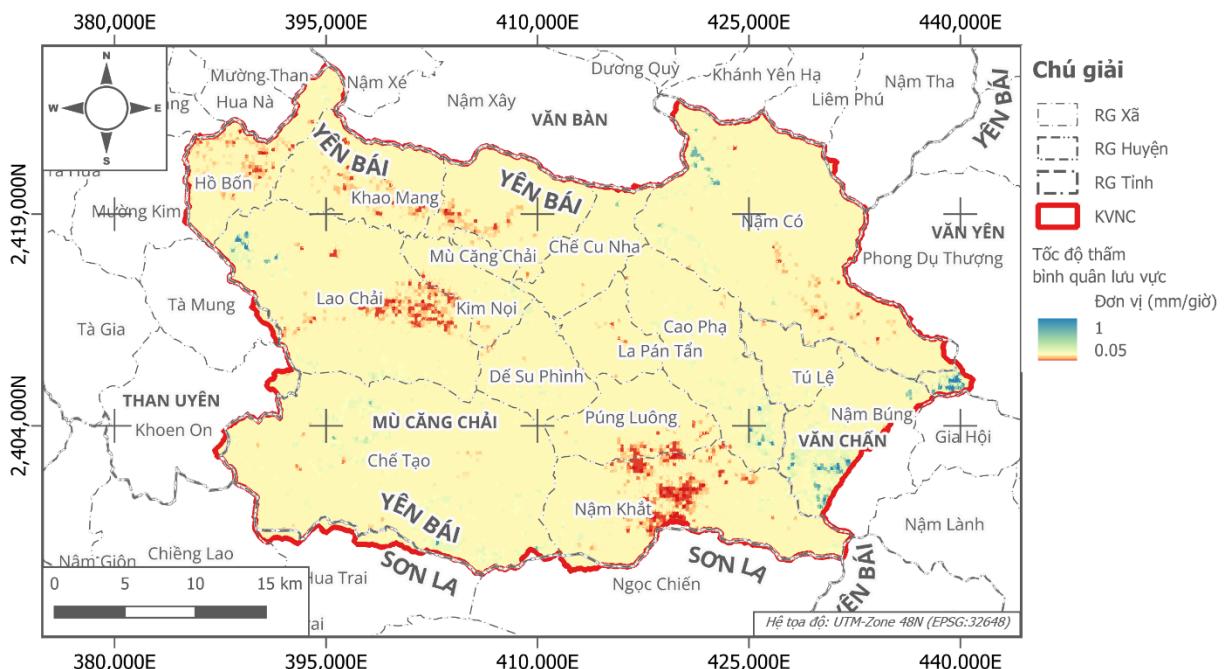
Chỉ số sức mạnh dòng chảy (Stream Power Index - SPI) đo lường khả năng xói mòn hoặc vận chuyển của dòng chảy. SPI cao biểu thị các khu vực có dòng chảy mạnh, có thể dẫn đến xói mòn hoặc lũ quét, trong khi SPI thấp liên quan đến tích tụ nước và ngập lụt.



Hình 3-44. Chỉ số sức mạnh dòng chảy SPI trên khu vực nghiên cứu

## f. Tốc độ thấm bình quân lưu vực

Tốc độ thấm bình quân đo lường khả năng của đất trong việc hấp thụ nước, được xác định dựa trên loại đất, cấu trúc địa chất, và thực phủ. Các khu vực có tốc độ thấm thấp (như đất sét) dễ bị ngập lụt hơn do nước không thể thấm xuống đất nhanh chóng. Đặc trưng này thường được trích xuất từ bản đồ đất, sau đó được ánh xạ lên lưới raster và cuối cùng là tổng hợp dữ liệu theo lưu vực. Như vậy, lưu vực nào có tốc độ thấm cao sẽ có nguy cơ sinh lũ thấp hơn các lưu vực có độ thấm thấp.



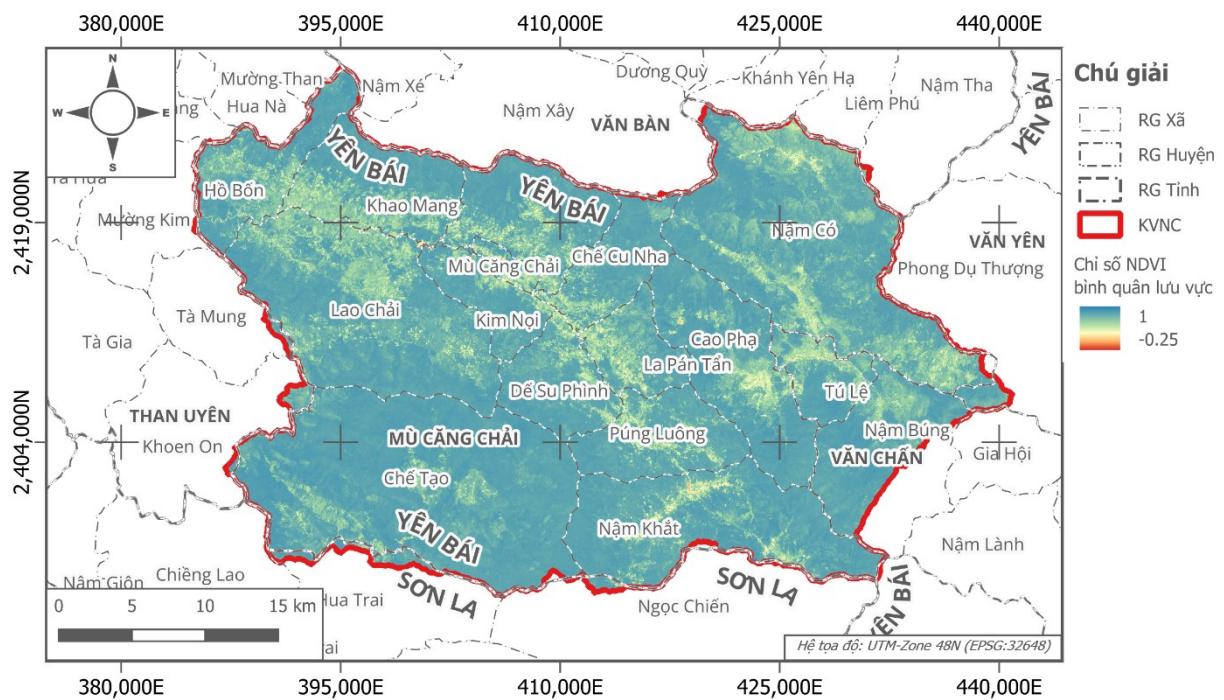
Hình 3-45. Tốc độ thấm bình quân lưu vực

## 3. Đặc trưng thực phủ:

Các đặc trưng thực phủ phản ánh mức độ che phủ của thảm thực vật và các đặc tính bề mặt liên quan đến khả năng giữ nước hoặc thoát nước.

### a. Chỉ số NDVI bình quân lưu vực

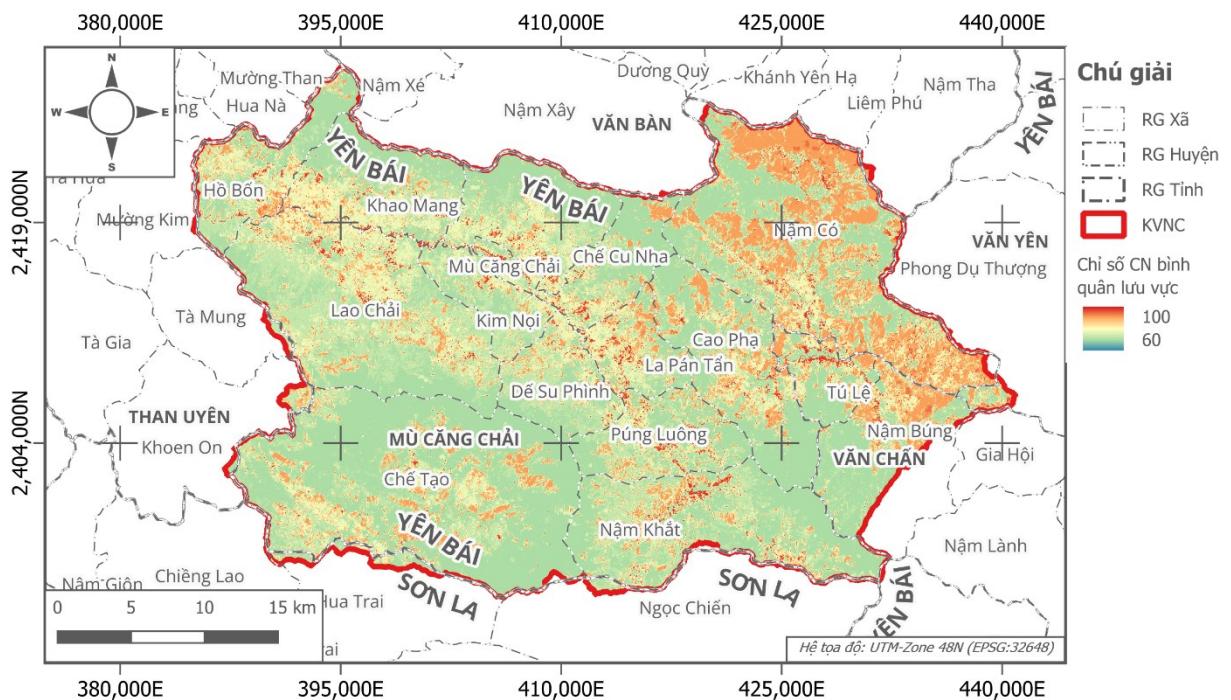
Chỉ số NDVI (Normalized Difference Vegetation Index) đo lường mức độ che phủ của thảm thực vật dựa trên sự khác biệt giữa phản xạ ánh sáng ở dải hồng ngoại gần và dải đỏ. NDVI cao biểu thị thảm thực vật dày đặc, giúp giảm nguy cơ lũ lụt do khả năng giữ nước và giảm dòng chảy bề mặt. Dữ liệu NDVI thường được thu thập từ hình ảnh vệ tinh như Landsat hoặc Sentinel-2, với độ phân giải không gian phù hợp để phân tích khu vực. Nghiên cứu này sử dụng dữ liệu ảnh Sentinel-2 với độ phân giải 10m để xác định chỉ số NDVI



Hình 3-46. Chỉ số NDVI bình quân lưu vực khu vực nghiên cứu

### b. Chỉ số CN bình quân lưu vực

Chỉ số CN (Curve Number) là một tham số thủy văn biểu thị khả năng giữ nước của bề mặt đất, dựa trên loại đất, thực phủ, và điều kiện sử dụng đất. CN có giá trị từ 0 đến 100, với giá trị cao hơn biểu thị khả năng giữ nước thấp (như khu vực đô thị hóa) và nguy cơ lũ lụt cao hơn. CN thường được lấy từ các bảng tra cứu của USDA hoặc tính toán dựa trên bản đồ đất và thực phủ.

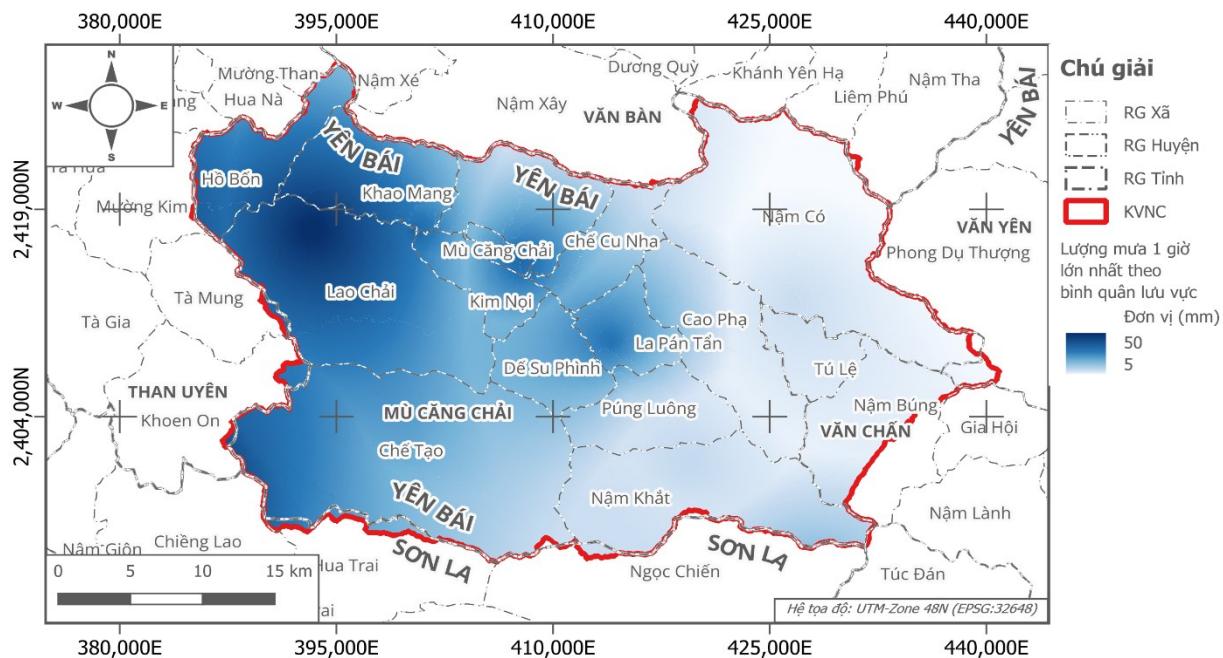


Hình 3-47. Chỉ số CN bình quân lưu vực

#### 4. Đặc trưng khí tượng:

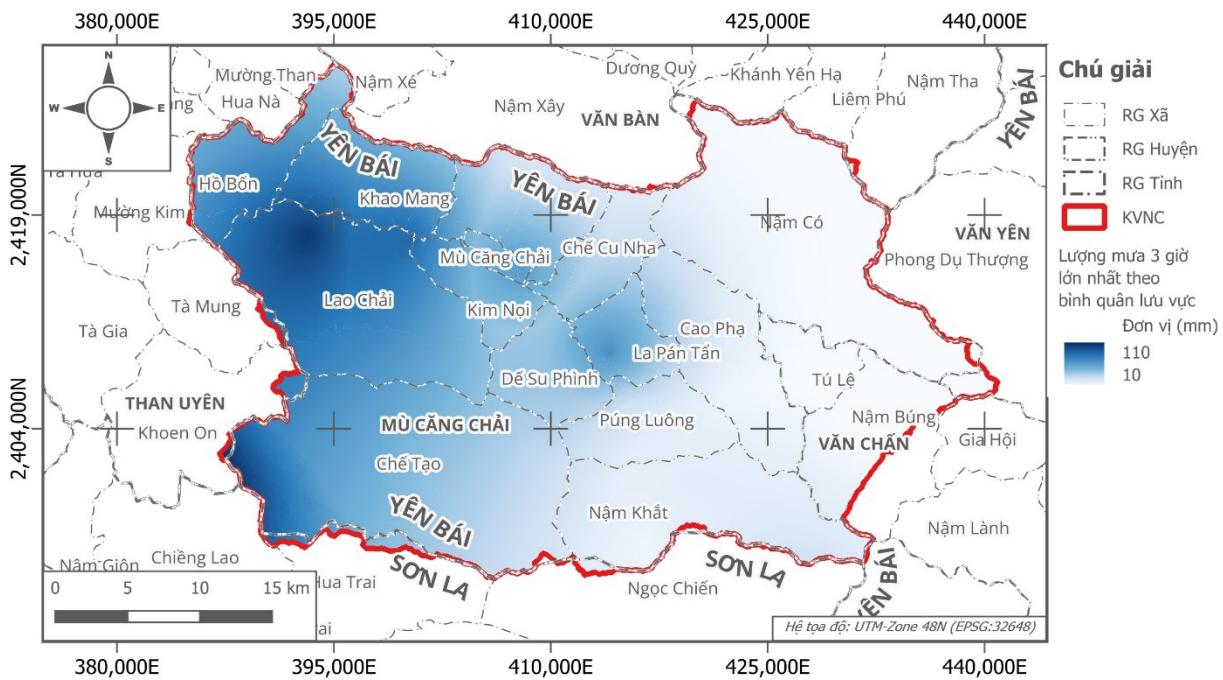
##### a. Lượng mưa giờ lớn nhất

Lượng mưa giờ lớn nhất là lượng mưa lớn nhất ghi nhận trong một giờ tại một điểm, được đo bằng milimet. Đặc trưng này phản ánh cường độ mưa cực đại, có thể gây ra lũ quét hoặc ngập lụt đô thị. Trong nghiên cứu này, nhóm nghiên cứu sử dụng lượng mưa giờ lớn nhất trước thời điểm xảy ra lũ quét năm 2023 nằm trong khoảng thời gian tập trung dòng chảy của lưu vực.



Hình 3-48. Lượng mưa 1 giờ lớn nhất trận lũ 05/08/2023

##### b. Lượng mưa 3 giờ lớn nhất

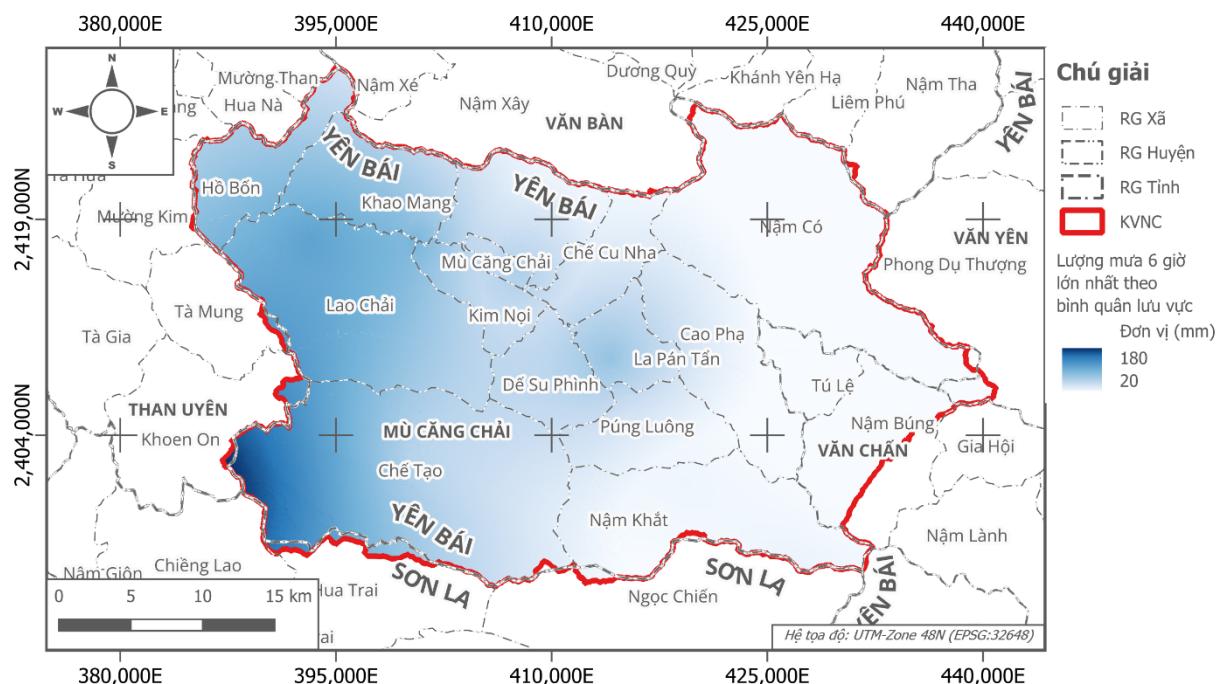


Hình 3-49. Tổng lượng mưa 3 giờ lớn nhất trận lũ 05/08/2023

Lượng mưa 3 giờ lớn nhất đo lường tổng lượng mưa trong khung thời gian 3 giờ liên tục lớn nhất. Đặc trưng này quan trọng trong việc đánh giá các sự kiện mưa ngắn hạn nhưng có cường độ cao, thường liên quan đến lũ quét.

#### c. Lượng mưa 6 giờ lớn nhất

Tương tự, lượng mưa 6 giờ lớn nhất đo lường lượng mưa trong khung thời gian 6 giờ. Đặc trưng này hữu ích trong việc đánh giá các sự kiện mưa kéo dài hơn, có thể gây ngập lụt ở các khu vực đồng bằng hoặc lưu vực lớn.

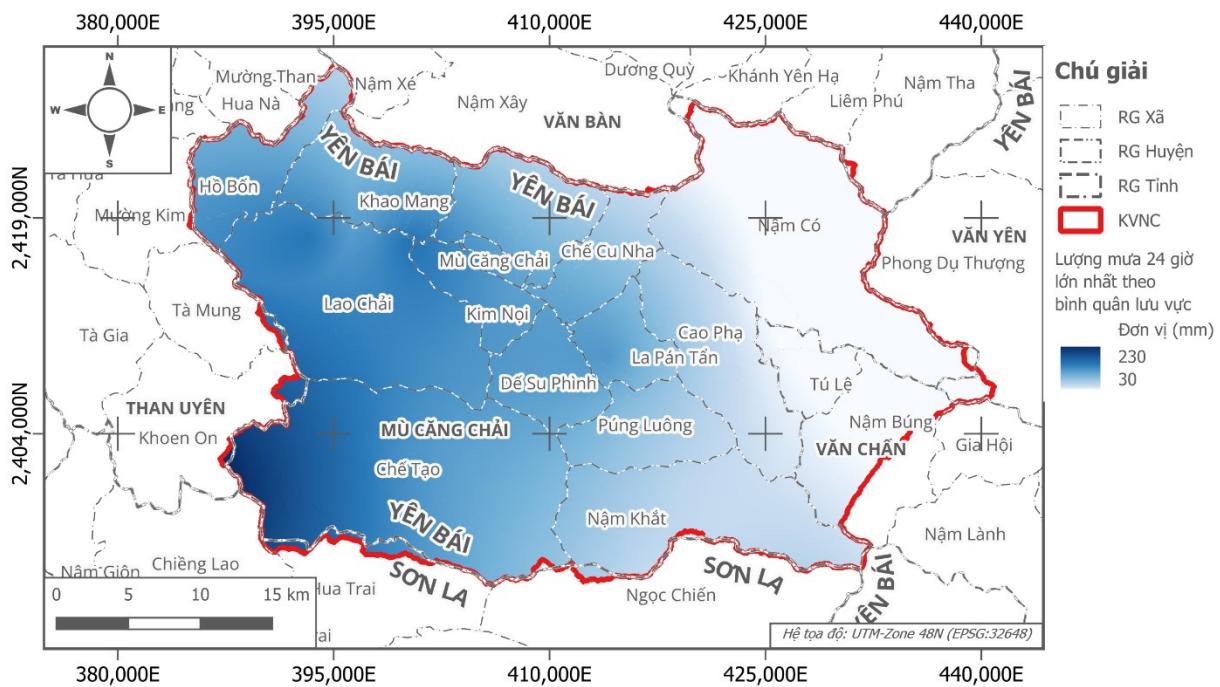


Hình 3-50. Tổng lượng mưa 6 giờ lớn nhất trận lũ 05/08/2023

Như vậy, trong vòng 6 giờ, lượng mưa tích lũy tại khu vực phía Đông giáp tỉnh Sơn La lên tới hơn 180mm. Đây là lượng mưa rất lớn gây ảnh hưởng trong khu vực, không chỉ vậy, lượng mưa lớn nhất 3 giờ đạt 110mm và mưa giờ lớn nhất là hơn 50mm tại khu vực xã Lao Chải và xã Hồ Bón có thể là nguyên nhân chính gây ra trận lũ quét tại xã Hồ Bón năm 2023.

#### d. Lượng mưa 24 giờ lớn nhất

Lượng mưa 24 giờ lớn nhất phản ánh tổng lượng mưa trong một ngày, là yếu tố chính trong các sự kiện lũ lụt quy mô lớn. Đặc trưng này đặc biệt quan trọng trong việc dự đoán lũ sông hoặc ngập lụt trên diện rộng.



Hình 3-51. Tổng lượng mưa 24 giờ lớn nhất trận lũ 05/08/2023

### 3.3. Thiết lập mô hình phân vùng lũ quét cho khu vực nghiên cứu

#### 3.3.1 Đầu vào và cấu trúc dữ liệu

##### 3.3.1.1 Phân tích lựa chọn dữ liệu đầu vào

Hầu hết các dữ liệu phổ biến đều được sử dụng trong nghiên cứu bao gồm cao độ, độ dốc, các đặc trưng về thực phủ và đặc trưng lượng mưa. Do sử dụng 2 loại mô hình trí tuệ nhân tạo bao gồm học máy và học sâu, các dữ liệu được lựa chọn đều có liên quan mật thiết trực tiếp đến khả năng sinh lũ quét. Những yếu tố trừu tượng (như loại đất nào tác động đến lũ quét chưa được làm rõ) thì sẽ được sử dụng bằng các chỉ số thay thế (như tốc độ thẩm bình quân của loại đất là bao nhiêu) bằng định lượng.

Bảng sau đây tổng hợp các yếu tố lựa chọn và lý do lựa chọn các yếu tố trong phân vùng lũ quét cho khu vực Mu Cang Chải.

Bảng 3-27. Phân tích lựa chọn các yếu tố dữ liệu đầu vào.

TT	Đặc trưng	Ý nghĩa	Phân tích lý do lựa chọn	Các yếu tố liên quan và lý do loại bỏ
1	Cao độ so với sông suối	Chênh lệch độ cao giữa điểm địa hình với sông, suối gần nhất theo hướng dòng chảy	Các điểm có chênh lệch cao độ so với sông, suối thấp có nguy cơ lũ quét cao hơn	Không có
2	Khoảng cách đến sông suối	Khoảng cách giữa điểm địa hình với sông, suối gần nhất theo hướng dòng chảy	Các điểm có khoảng cách gần sông, suối có nguy cơ lũ quét cao hơn	Không có

TT	Đặc trưng	Ý nghĩa	Phân tích lý do lựa chọn	Các yếu tố liên quan và lý do loại bỏ
3	Độ dốc	Độ dốc địa hình (độ dốc cục bộ điểm)	Có tác động đến lũ quét: độ dốc lớn làm cho nước di chuyển nhanh hơn.	Không có
4	Độ dốc lòng dẫn	Độ dốc của lòng dẫn (theo hướng dòng chảy)	Độ dốc lòng dẫn lớn → năng lượng dòng chảy lớn hơn trong cùng điều kiện về lưu lượng. Tác động đến thời gian tập trung dòng chảy và các yếu tố thủy văn khác	Không có
5	Chiều dài dòng chảy	Khoảng cách từ điểm xa nhất của lưu vực đến điểm tính toán	Ảnh hưởng đến thời gian tập trung dòng chảy, chiều dài ngắn thì thời gian tập trung nhanh hơn	Không có
6	Diện tích lưu vực	[Nguyên tắc lưu vực] – Diện tích từng lưu vực. Mỗi điểm được coi là cửa ra một lưu vực	Diện tích lớn → lưu lượng lớn hơn trong cùng lượng mưa, là tham số ảnh hưởng trực tiếp đến kết quả tính toán thủy văn	Không có
7	Chỉ số âm địa hình	Phản ánh khả năng tích tụ nước tại một điểm.	TWI cao mô tả các khu vực thấp, trũng (chân đồi, lòng dẫn...) → nguy cơ ngập lụt và lũ quét cao hơn.	Không có
8	Chỉ số sức mạnh dòng chảy	Mô tả năng lượng tiềm năng của dòng chảy bề mặt trong đánh giá xói mòn và vận chuyển trầm tích	SPI cao → năng lượng lớn → khả năng xói mòn mạnh → tăng nguy cơ lũ quét	Không có
9	Chỉ số vị trí địa hình	Chênh lệch cao độ của một điểm so với bình quân các điểm xung quanh	TWI thấp ( $<0$ ) là khu vực tích tụ, bị tập trung nước. TWI cao ( $>0$ ) là khu vực thoát nước	Không có
10	Chỉ số NDVI	[Nguyên tắc lưu vực] Mô tả mật độ và sức khỏe của thực vật bình quân trong một lưu vực.	NDVI cao → thảm thực vật dày → giữ nước tốt → giảm nguy cơ lũ quét	Yếu tố liên quan trực tiếp là thảm phủ và sử dụng đất. Hai yếu tố này là định tính đối với lũ quét, do đó không lựa chọn trong nghiên cứu này. (Thay bằng chỉ số NDVI, CN)
11	Chỉ số CN	[Nguyên tắc lưu vực] Phản ánh khả năng thấm nước,	CN cao → trữ nước kém và tạo dòng chảy	Lý do: chưa có các nghiên cứu rõ ràng về đất

TT	Đặc trưng	Ý nghĩa	Phân tích lý do lựa chọn	Các yếu tố liên quan và lý do loại bỏ
		trữ nước, sử dụng đất và loại đất	mặt lớn → tăng nguy cơ lũ quét	trồng lúa và hoa màu thì loại nào có tác động bất lợi đến lũ quét...
12	Tốc độ thám bình quân	[Nguyên tắc lưu vực] Mô tả tốc độ thám của nước mưa xuống bề mặt	Tốc độ thám càng lớn → giảm dòng chảy mặt → giảm nguy cơ lũ quét	Yếu tố liên quan trực tiếp là loại đất (phân loại theo nhóm), các loại đất có tác động đến lũ quét khác nhau nhưng chưa có chỉ tiêu định lượng rõ ràng, do đó loại đất không được lựa chọn trong nghiên cứu này (thay bằng tốc độ thám bình quân của đất).
13	Cao độ địa hình	Giá trị cao độ tại một vị trí	Ảnh hưởng trực tiếp hoặc gián tiếp đến lũ quét. Cao độ địa hình thấp có nguy cơ bị lũ quét hơn.	Không có
14	Cao độ bình quân lưu vực	[Nguyên tắc lưu vực] Giá trị cao độ bình quân lưu vực	Lưu vực có cao độ bình quân cao là các lưu vực ở vùng đồi, núi. Có ảnh hưởng trực tiếp hoặc gián tiếp đến lũ quét.	Không có
15	Độ cong địa hình (theo hướng dốc)	Mô tả bề mặt của địa hình theo hướng dốc	Nếu âm (lõm) → tích tụ dòng chảy (chân dốc, thung lũng) → tăng nguy cơ lũ quét	Không có
16	Độ cong địa hình (phuong ngang)	Mô tả bề mặt của địa hình theo hướng ngang	Nếu âm (lõm) → hội tụ dòng chảy → tăng nguy cơ lũ quét	Không có
17	Lượng mưa giờ lớn nhất	Lượng mưa giờ lớn nhất trong một thời đoạn	Lượng mưa thời đoạn ngắn lớn có khả năng sinh lũ quét lớn. Lũ quét trong tự nhiên được hình thành do mưa lớn trong thời đoạn ngắn.	Không có
18	Lượng mưa 3 giờ lớn nhất	Lượng mưa 3 giờ lớn nhất trong một thời đoạn		
19	Lượng mưa 6 giờ lớn nhất	Lượng mưa 6 giờ lớn nhất trong một thời đoạn	Thời gian thường dưới 6 giờ.	
20	Lượng mưa 24 giờ lớn nhất	Lượng mưa 24 giờ lớn nhất trong một thời đoạn		Yếu tố liên quan trực tiếp là độ ẩm kỳ trước (bằng các bản đồ như SMAP...). Nghiên cứu này sử dụng lượng mưa

TT	Đặc trưng	Ý nghĩa	Phân tích lý do lựa chọn	Các yếu tố liên quan và lý do loại bỏ
				24 giờ lớn nhất thay thế cho độ ẩm kỳ trước (bởi các bản đồ vệ tinh) do chu kỳ ảnh không liên tục.

### 3.3.1.2 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 thuận lợi (Christopher M. Bishop, 1995).

Một nguyên tắc nhỏ là các biến đầu vào nên có giá trị nhỏ, có thể nằm trong khoảng 0÷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 (Christopher M. Bishop, 1995).

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

Bảng 3-28. 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]	$X' = \frac{X - \text{median}}{\text{IQR}}$
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	$\text{m}^2$	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	wInfiltrate	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]	$X' = \frac{X - \mu}{\sigma}$

TT	Đặc trưng	Ký hiệu	Đơn vị	Phương pháp	Chuẩn hóa	Khoảng giá trị	Công thức Bước 1
17	Lượng mưa giờ lớn nhất	max_p_recip	mm	Lưu vực	Log & “÷10”	[0, 1]	$X' = \log(X + 1)$
18	Lượng mưa 3 giờ lớn nhất	max_3h_prec_ip	mm	Lưu vực	Log & “÷10”	[0, 1]	$X' = \log (X+1)$
19	Lượng mưa 6 giờ lớn nhất	max_6h_prec_ip	mm	Lưu vực	Log & “÷10”	[0, 1]	$X' = \log (X+1)$
20	Lượng mưa 24 giờ lớn nhất	max_24h_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), ngoại trừ lượng mưa.

## 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 3-29. 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, Hò Bón, Mồ Dề, Lao Chải, Ché Tạo và Nậm Cố. Chi tiết như sau:

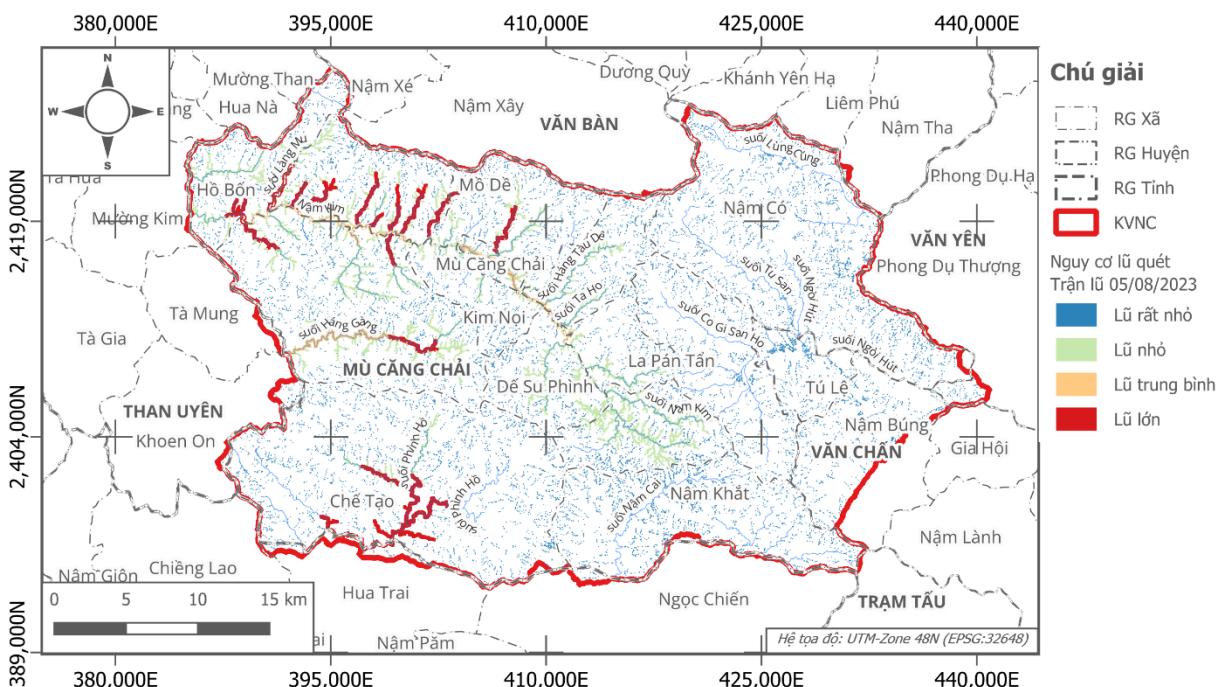
Bảng 3-30. Nhãn phân loại đánh giá theo tình hình mưa lũ thực tế tại địa phương

<b>TT</b>	<b>Địa chỉ</b>	<b>Suối</b>	<b>Đánh giá đợt lũ 8/2023</b>	<b>Nhãn phân loại</b>
1	Xã Hò Bón	Háng Nhù, Thống 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 Đè Chu ghi nhận dòng chảy lũ bình thường (có lũ nhưng ít nguy hiểm)	Các nhánh suối chính như Hàng Nhù, Thống Gàu Bua, Làng Mu và lân cận suối được gán nhãn 4, các suối còn lại gán nhãn 3.
2	Xã Khao Mang	Suối Hàng B La Ha; Hàng B La Đê; Hàng Tàu Đê; 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 Đê 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 Đê gán nhãn 4, suối còn lại gán nhãn 3.
3	Xã Lao Chải	Suối Hàng Đè Sửa, Hàng Gàng, Lao Chải	Riêng suối Hàng Đè Sửa 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 Đè Sửa 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 Mồ Dề	Suối Nà Hàng	Suối Nà Hàng thuộc xã Mồ Dề 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.

TT	Địa chỉ	Suối	Đánh giá đợt lũ 8/2023	Nhãn phân loại
5	Xã Chẽ 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 Chẽ Tạo, Phú Vá, Tà Sung. Khu vực bản Nả Háng không ghi nhận lũ lớn.	Các nhánh suối Nậm Khắt, Nậm Khốt và Phình Hồ được gán nhãn 4, các nhánh suối đổ vào được gán nhãn 3.
6	Xã Nậm Có		Các suối trên khu vực xã Nậm Có không ghi nhận lũ lớn.	Các nhánh suối trên khu vực này được gán nhãn 2.

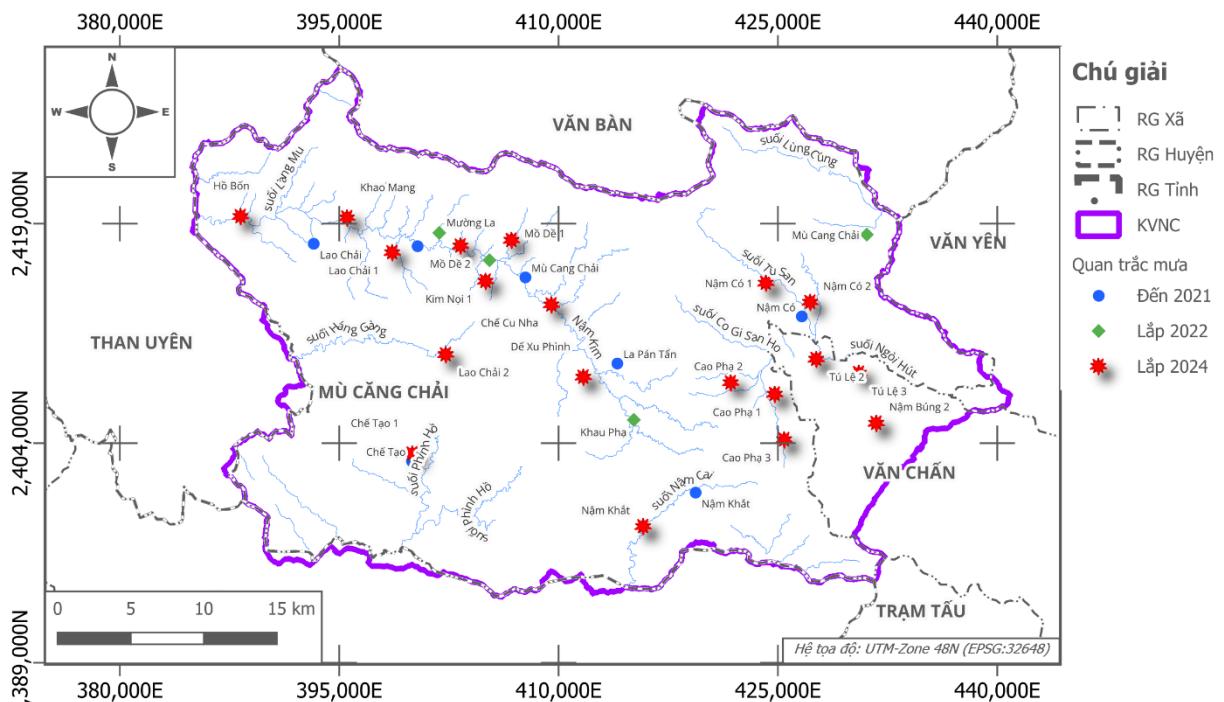
Ngoài ra theo mô tả, khu vực bản Mí Háng Táu (thuộc xã Púng Luông) cũng có ghi nhận lũ rất lớn, gây ảnh hưởng đến sinh hoạt của người dân. Các khu vực xã khác và các nhánh suối khác không ghi nhận lũ lớn. Nhìn chung, phần lớn các nhánh suối khu vực hạ du suối Nậm Kim đều ghi nhận lũ lên, trong khi đó, 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 nhãn 1 nếu nằm trên các sườn núi có độ dốc lớn (mái dốc núi), đỉnh núi và các khu vực ruộng bậc thang. Nhãn 2 được đánh cho các nhánh suối nhỏ và rất nhỏ ở thượng nguồn các khu vực không ghi nhận lũ lớn.



Hình 3-52. 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 Bón) làm cơ sở để xác định các điểm phân loại nguy cơ như đã trình bày phía trên. Bên cạnh đó, lượng mưa trạm đo Mù Cang Chải trận lũ năm 2017 tại suối Háng Chú cũng được sử dụng để tăng cường dữ liệu do có khoảng cách khá gần đối với vị trí xảy ra lũ quét.



Hình 3-53. Phân bố các trạm quan trắc mưa khu vực Mù Cang Chải

### 3.3.1.3 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 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.

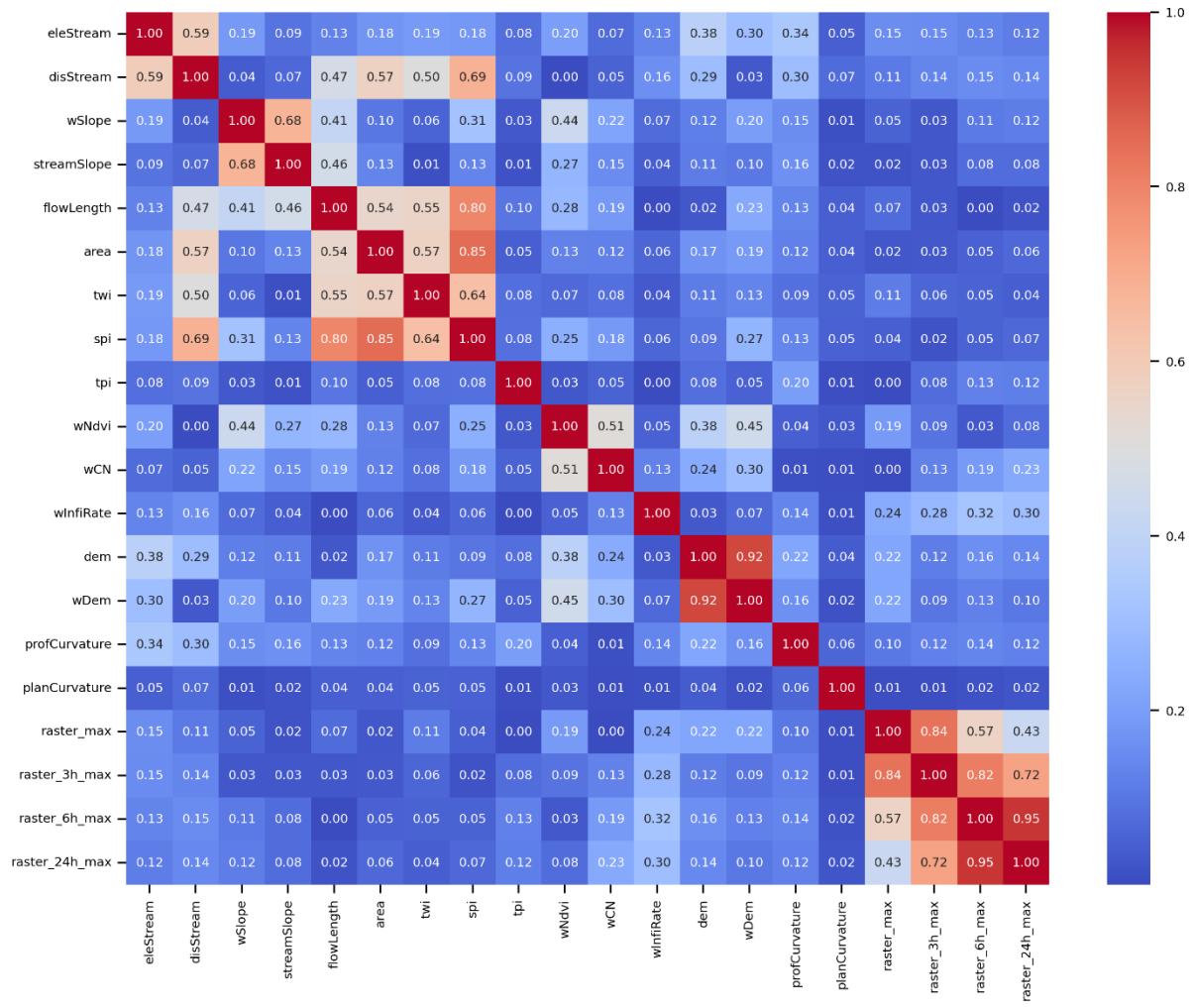
Cấu trúc dữ liệu này được thiết kế để tối ưu hóa cả hiệu quả tính toán và ý nghĩa vật lý. Các tệp GeoTIFF giữ được thông tin không gian, trong khi DataFrame hỗ trợ xử lý nhanh các thuật toán học máy. Hệ thống logging tích hợp ghi lại mọi bước xử lý, đảm bảo khả năng theo dõi và tái hiện. Cấu trúc dữ 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.

### **3.3.2 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.



Hình 3-54. Ma trận tương quan các đặc trưng

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.

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 khớp 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 khớp, 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 khớp 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 khớp, 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 mảng 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 dần 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 và giá trị dự đoán 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 và trục tung là tỷ lệ dương tính thật. 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.

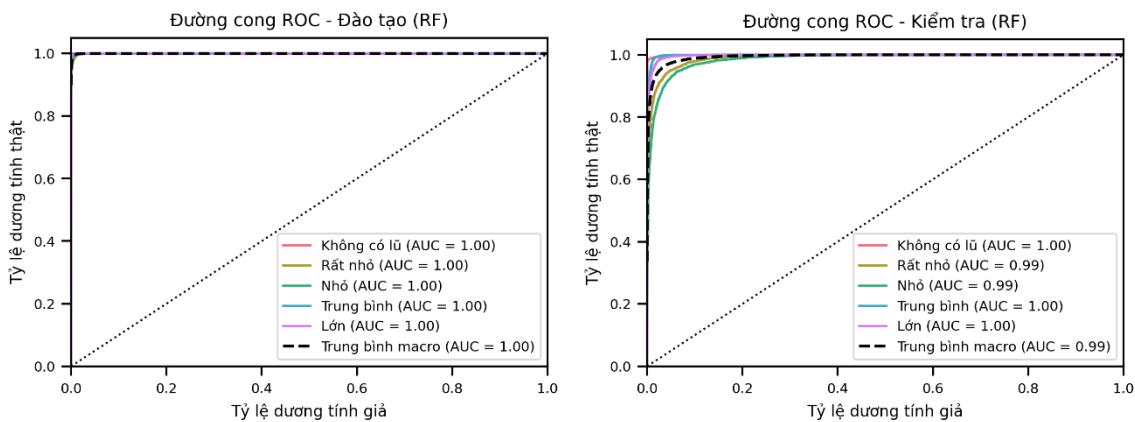
Bảng 3-31. Bảng so sánh, đánh giá các mô hình đã xây dựng

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:

Ma trận nhầm lẩn - Đào tạo (RF)						Ma trận nhầm lẩn - Kiểm tra (RF)						
Thực tế	Không có lũ	251	104	12	34	Thực tế	Không có lũ	74	14	6	7	
	Rất nhỏ	36	16924	1657	53		Rất nhỏ	8	4275	380	8	85
	Nhỏ	35	1166	16045	933		Nhỏ	7	245	4092	198	215
	Trung bình	9	25	218	18709		Trung bình	2	6	41	4702	6
	Lớn	3	116	317	185		Lớn	1	28	74	35	4618
	Dự đoán	Không có lũ	Rất nhỏ	Nhỏ	Trung bình	Dự đoán	Không có lũ	Rất nhỏ	Nhỏ	Trung bình	Lớn	

Hình 3-55. Ma trận nhầm lẩn mô hình RF



Hình 3-56. Đường cong ROC mô hình RF

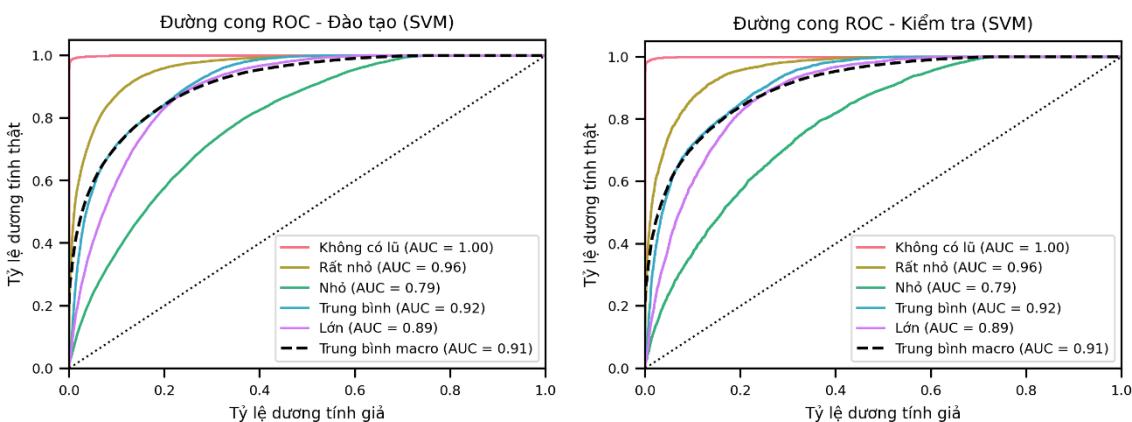
Random Forest (RF) đạt độ chính xác 93.95%, cho thấy khả năng phân loại tốt trên tập dữ liệu. Tuy nhiên, khi xem xét ma trận nhầm lẩn, RF gặp khó khăn trong việc phân biệt các mức độ lũ trung gian, đặc biệt là giữa "Lũ nhỏ" và "Lũ lớn".

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

Hình 3-57. Ma trận nhầm lẩn mô hình SVM

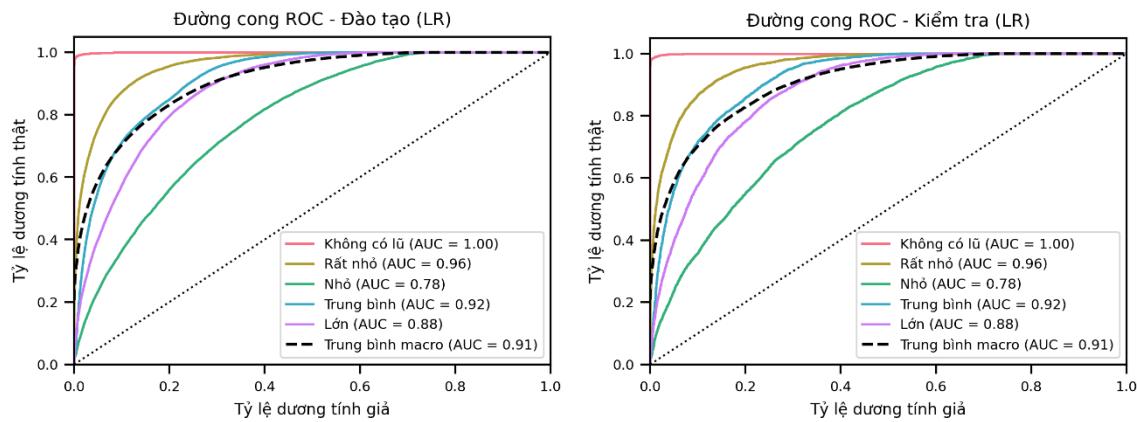


Hình 3-58. Đường cong ROC mô hình SVM

Support Vector Machine (SVM) chỉ đạt 69.47% accuracy, thấp nhất 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:

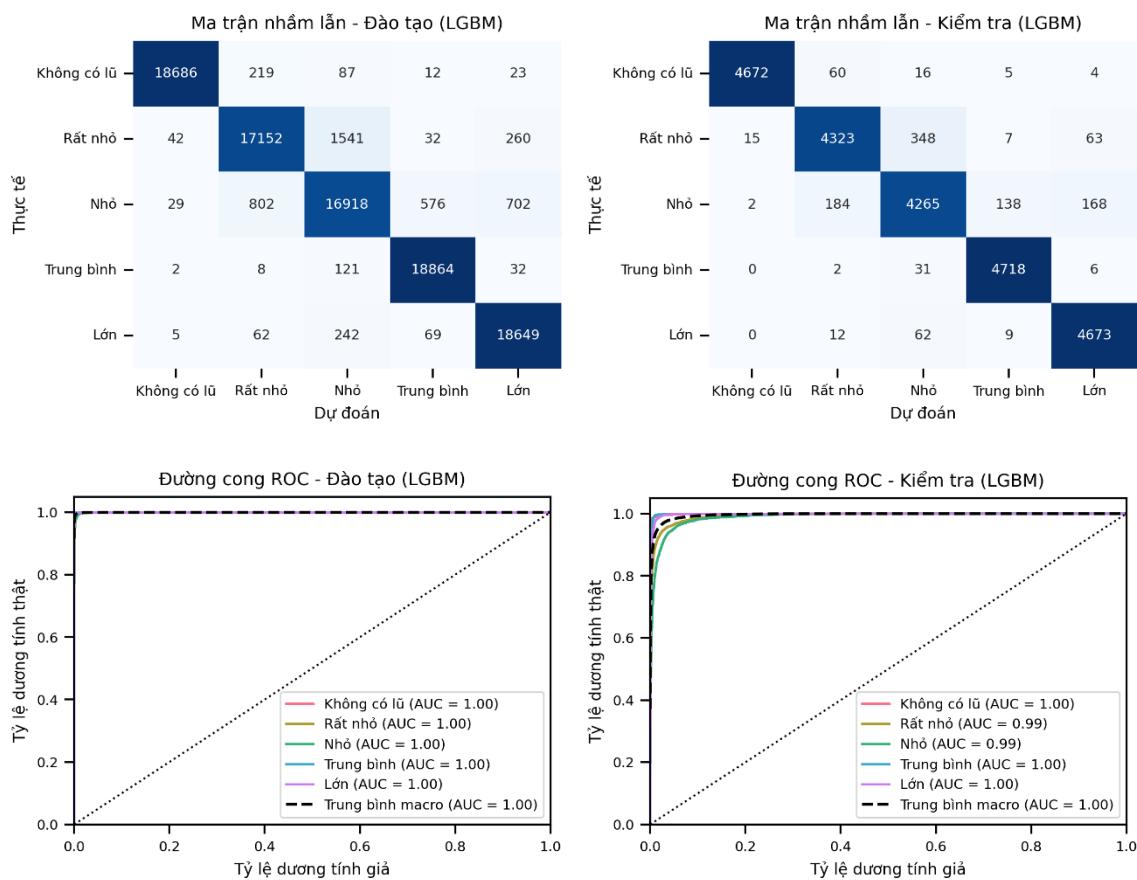
Ma trận nhầm lẩn - Đào tạo (LR)						Ma trận nhầm lẩn - Kiểm tra (LR)					
Thực tế	Không có lũ	223	79	25	35	Không có lũ	4664	67	15	5	6
	Rất nhỏ	52	14996	3007	247	Rất nhỏ	15	3768	724	46	203
	Nhỏ	56	3888	7530	3787	Nhỏ	13	997	1828	937	982
	Trung bình	17	64	2286	13406	Trung bình	3	23	563	3368	800
	Lớn	5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



Hình 3-59. Ma trận nhầm lẫn và đường cong ROC mô hình LR

Logistic Regression (LR) có hiệu suất gần bằng SVM (68.97% accuracy), nhưng ma trận nhầm lẫn 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:

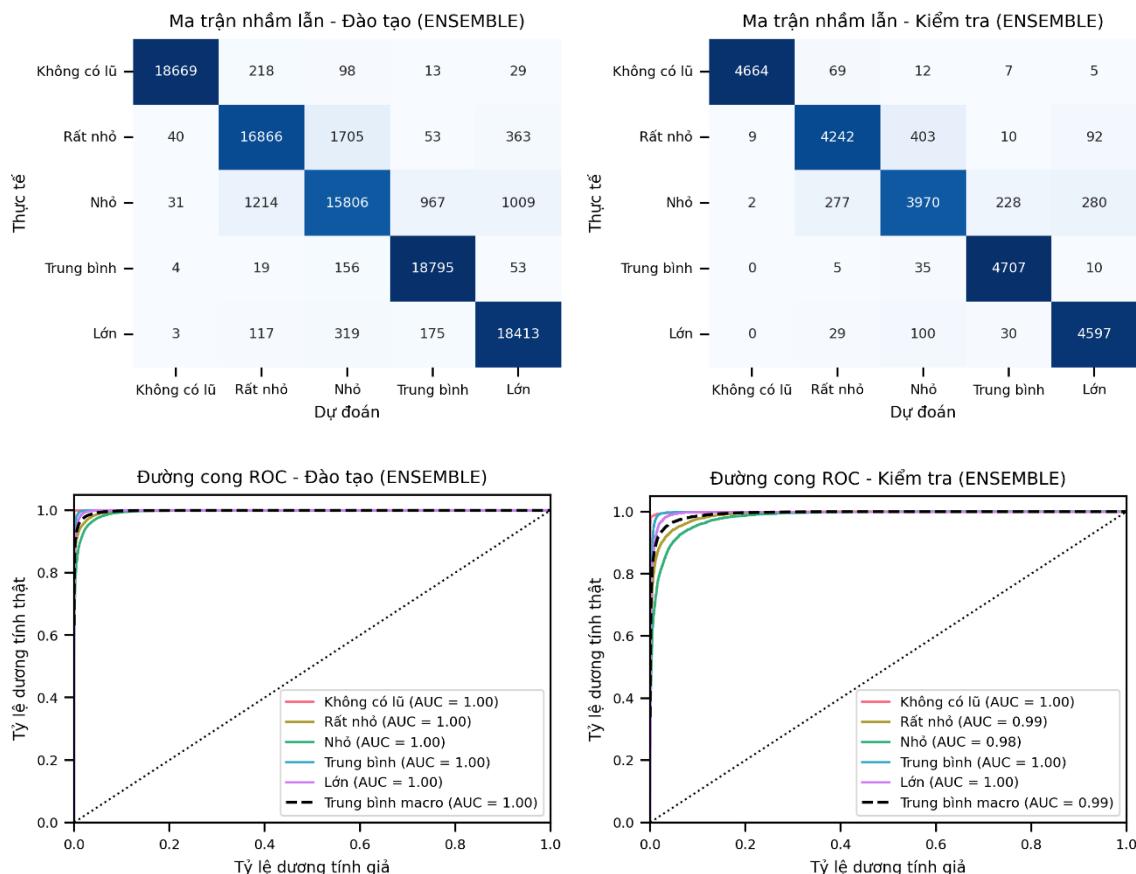


Hình 3-60. Ma trận nhầm lẫn và đường cong ROC mô hình LGBM

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 lẫn 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 3-61. Ma trận nhầm lẫn 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

hắn. 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.

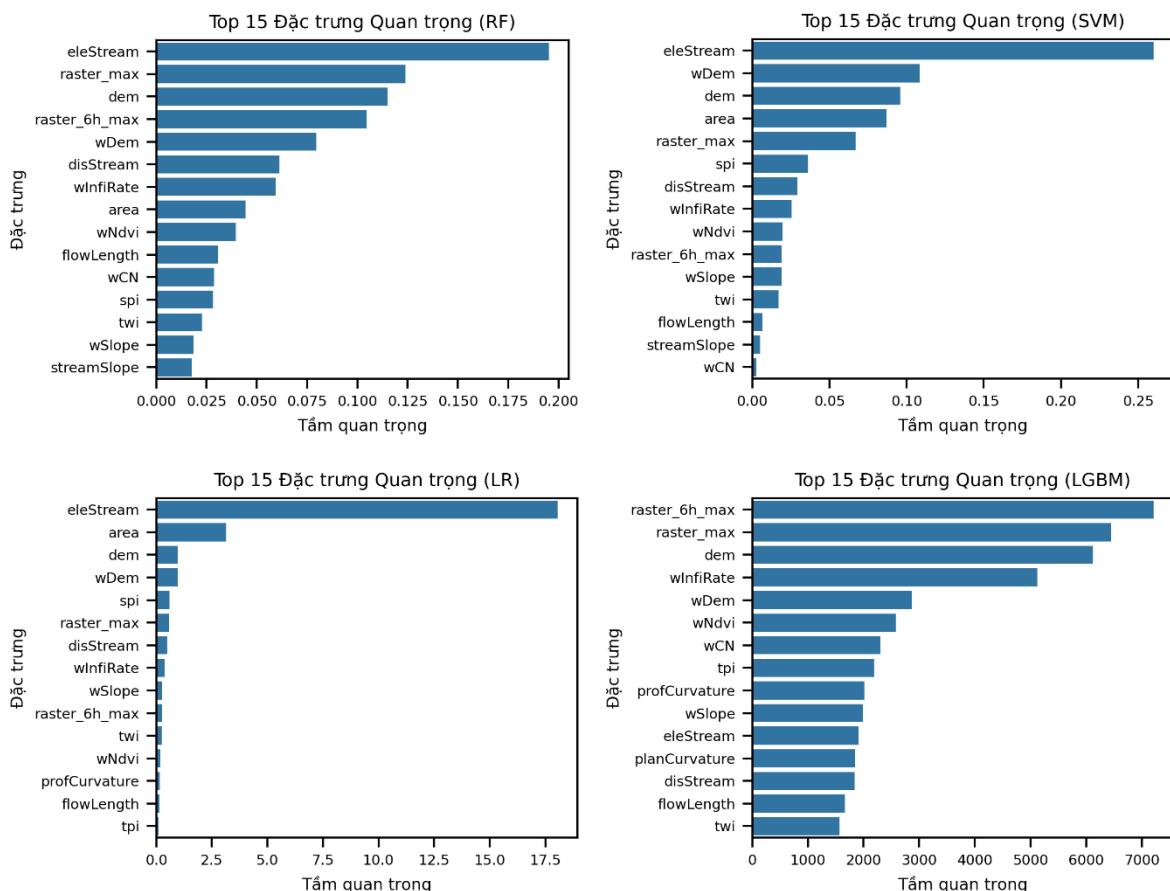
Hai mô hình SVM và Logistic Regression cho kết quả kém hơn hẳn với độ chính xác chỉ khoảng 69%. Điều này phản ánh hạn chế của các phương pháp tuyến tính khi xử lý bài toán có ranh giới phức tạp như phân loại mức độ lũ. Các mô hình này gần như không phân biệt được giữa các mức độ trung gian, đặc biệt 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

trung 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 wInfiRate (chỉ số tốc độ thẩm) và disStream (khoảng cách đến sông suối) cũng xuất hiện thường xuyên, phản ánh vai trò của môi trường tự nhiên.



Hình 3-62. 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.

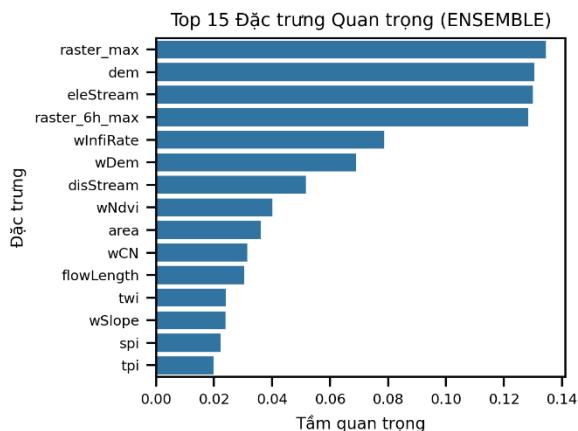
- LGBM: Ưu tiên cao độ địa hình (dem) và các đặc trưng mưa với thang đo lớn (lên đến 4000), phản ánh khả năng xử lý dữ liệu phức tạp tốt hơn. Tuy nhiên việc thể hiện cao độ của điểm (giá trị cao độ địa hình) có liên quan trực tiếp đến lũ quét là một trong những nhận định chưa tốt của mô hình này. Một 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

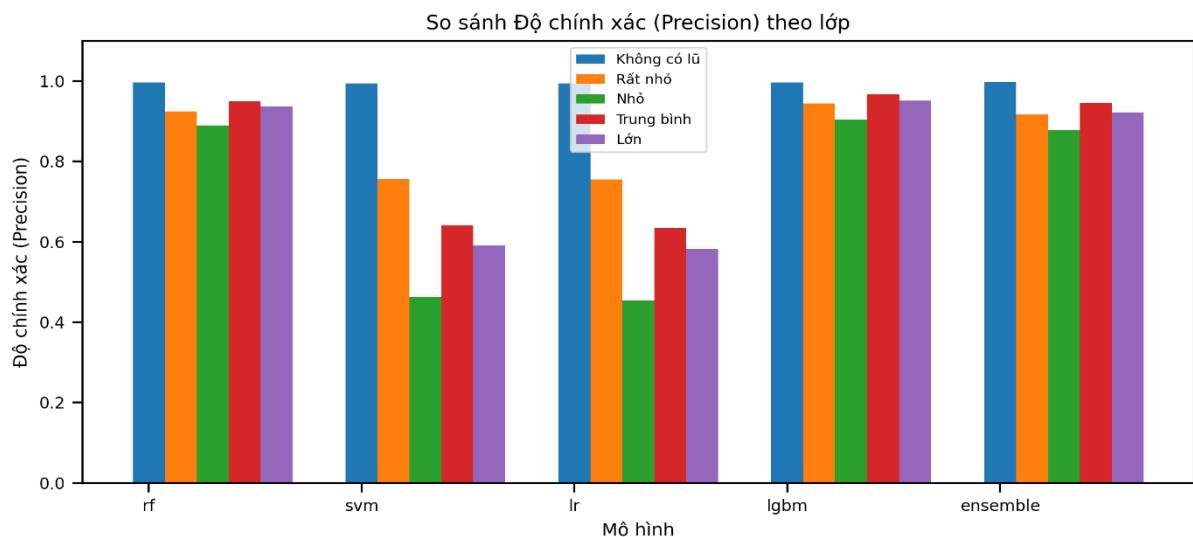
Độ chính xác (Precision), độ nhạy (Recall) và F1 Score là ba chỉ số quan trọng trong việc đánh giá hiệu suất của các mô hình phân loại, đặc biệt trong các bài toán mà dữ liệu có thể không cân bằng giữa các lớp. Độ chính xác đo lường tỷ lệ các dự đoán dương tính thực sự đúng, thể hiện khả năng mô hình 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



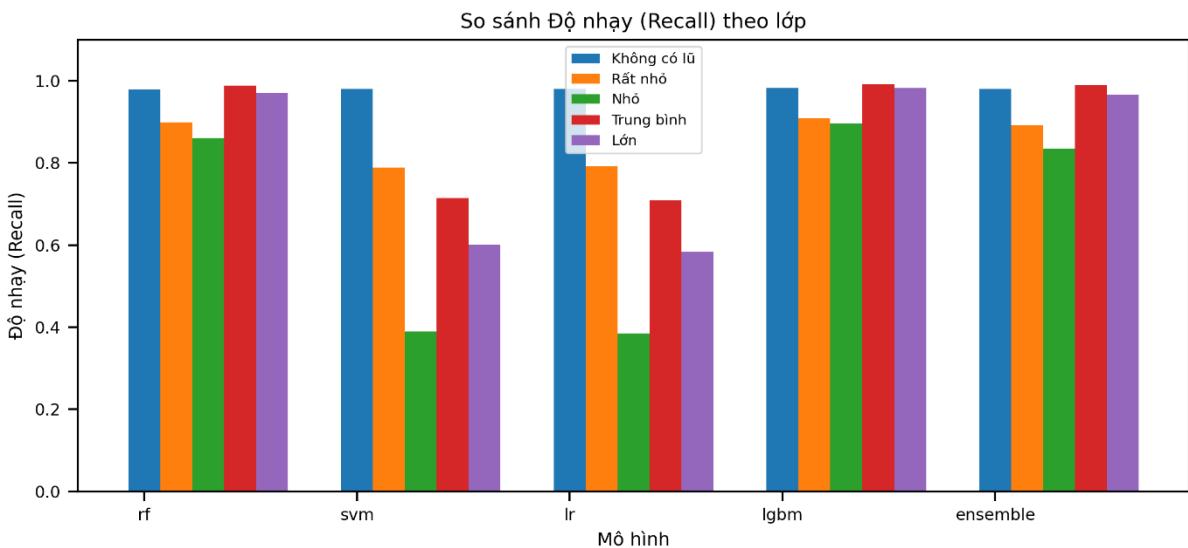
ranh giới giữa các mức độ lũ này chưa thực sự rõ ràng trong dữ liệu, khiến các mô hình tuyến tính như SVM và LR khó phân biệt chính xác.

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, cung cấp cho nhận định về sự chênh lệch đáng kể giữa các mức độ lũ liền kề trong không gian đặc trưng.



Hình 3-63. Độ chính xác của các mô hình học máy theo các lớp dữ liệu

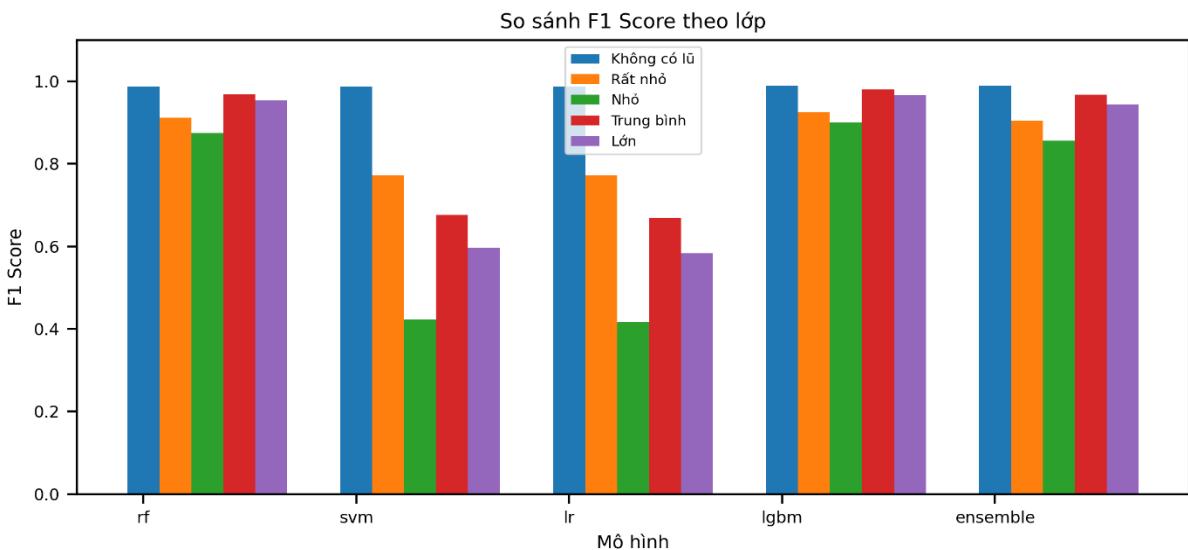
Các mô hình thể hiện sự khác biệt rõ rệt về khả năng phát hiện chính xác từng mức độ lũ. LGBM tiếp tục dẫn đầu với độ nhạy recall ấn tượng trên 0.95 cho hầu hết các lớp, chứng tỏ khả năng bắt gặp gần như toàn bộ các trường hợp lũ thực tế. Đặc biệt ở lớp “Lũ trung bình”, LGBM đạt recall gần như hoàn hảo (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 3-64. 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 3-65. 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 “Lũnhỏ”. 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 chòng lấn 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 cố hữu 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 3-32. 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.

### 3.3.3 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  $3 \times 3$  đến  $11 \times 11$ , kết quả lựa chọn được mô hình CNN được lấy vùng lân cận  $5 \times 5$  và mô hình DNN được lấy đặc trưng lân cận  $7 \times 7$  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  $5 \times 5$ , trích xuất từ các tệp raster địa lý. Mỗi tệp raster tương ứng với một tham số trong danh sách được nêu ở Bảng 3-28, 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  $5 \times 5$  tạo ra một mảng dữ liệu có kích thước  $5 \times 5 \times 20$  (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 từ trước.

Kích thước vùng lân cận  $5 \times 5$  được chọn để cân bằng giữa việc cung cấp đủ thông tin không gian và giảm chi phí tính toán. Một vùng lân cận  $5 \times 5$  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  $5 \times 5$  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  $9 \times 9$  (81 điểm dữ liệu) hoặc  $11 \times 11$  (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  $3 \times 3$ , mô hình có thể bỏ qua các mẫu không gian quan trọng do cửa sổ quá nhỏ. Kích thước  $5 \times 5$  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  $5 \times 5$  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  $7 \times 7$  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 x 5 đặc trưng thống kê).

Lý do sử dụng đặc trưng thống kê thay vì dữ liệu không gian (như vùng lân cận  $5 \times 5$  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  $7 \times 7$ , 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  $7 \times 7$  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 thống kê, 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 5x5 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 gamma=2.0 và 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.

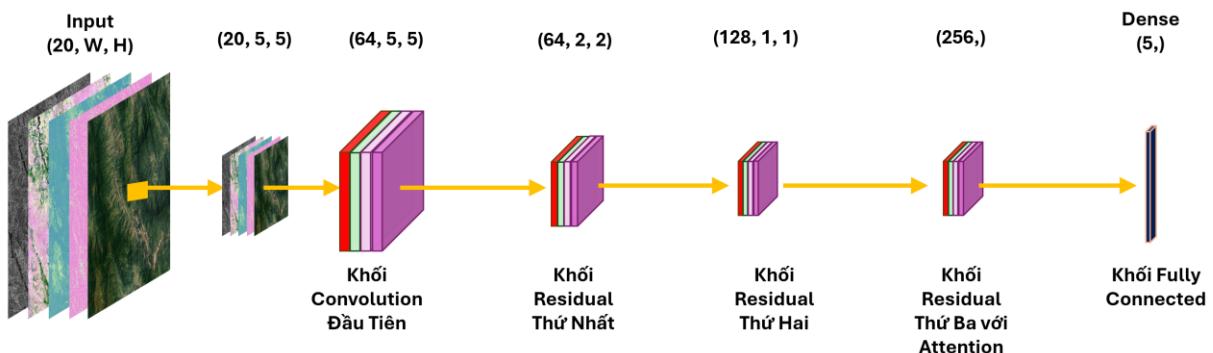
### Xây dựng mô hình CNN:

#### ***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  $5 \times 5$ . 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  $5 \times 5$  xuống  $2 \times 2$  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 3-66. 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  $2 \times 2$ , mô hình tập trung vào việc tích hợp thông tin từ các vùng lân cận để hiểu được bối cảnh 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 7x7, 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**

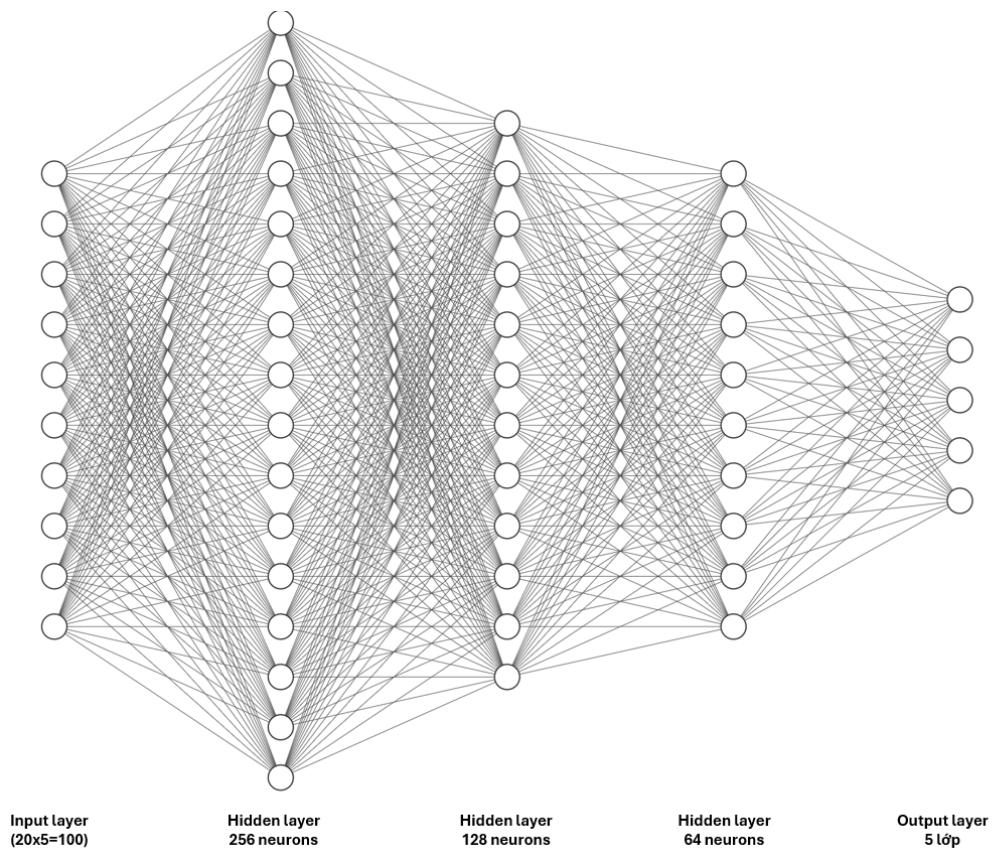
Mô hình DNN nhận đầu vào là vector một chiều với kích thước input\_dim, chứa các đặc trưng đã được tổng hợp và trích xuất từ vùng nghiên cứu. Khác với CNN xử lý dữ liệu không gian 2D, DNN làm việc với các chỉ số thống kê tổng hợp như giá trị trung bình, độ lệch chuẩn, giá trị min/max của 16 yếu tố địa 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 Trưng**

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 3-67. 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.

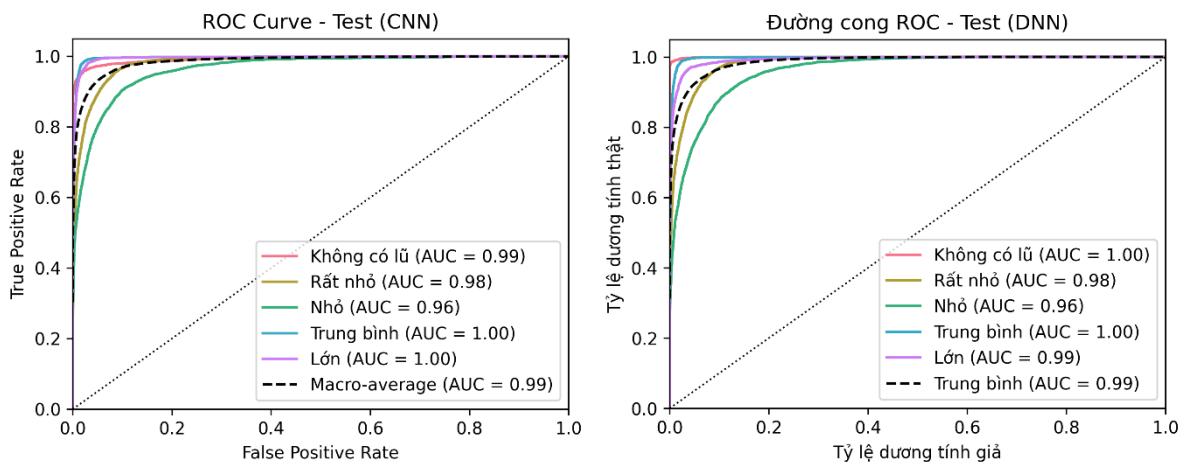
### **Triết Lý Thiết Kế Tổng Thể**

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 3-68. 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						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		
	Rất nhỏ	Nhỏ	Trung bình	Lớn		Rất nhỏ	Nhỏ	Trung bình	Lớn			

Hình 3-69. 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

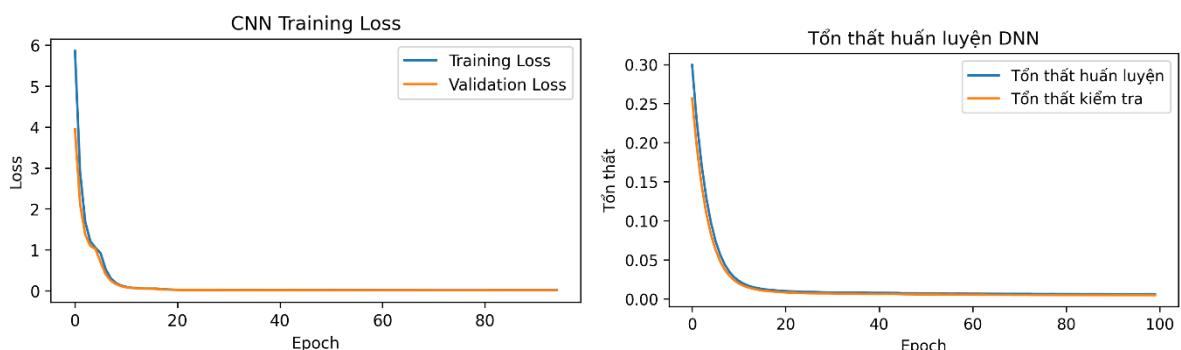
được độ chính xác cao với 4149 và 4685 trường hợp được phân loại đúng tương ứng. Điều này cho thấy mô hình có khả năng mạnh mẽ trong việc nhận diện các tình huống cực đoan - từ không có lũ đến lũ ở mức độ nghiêm trọng.

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. Cá 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 3-70. 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 3-33. 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.

### 3.3.4 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 phác tạp nhưng lại cho kết quả thua kém 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 3-34. 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 mất 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.

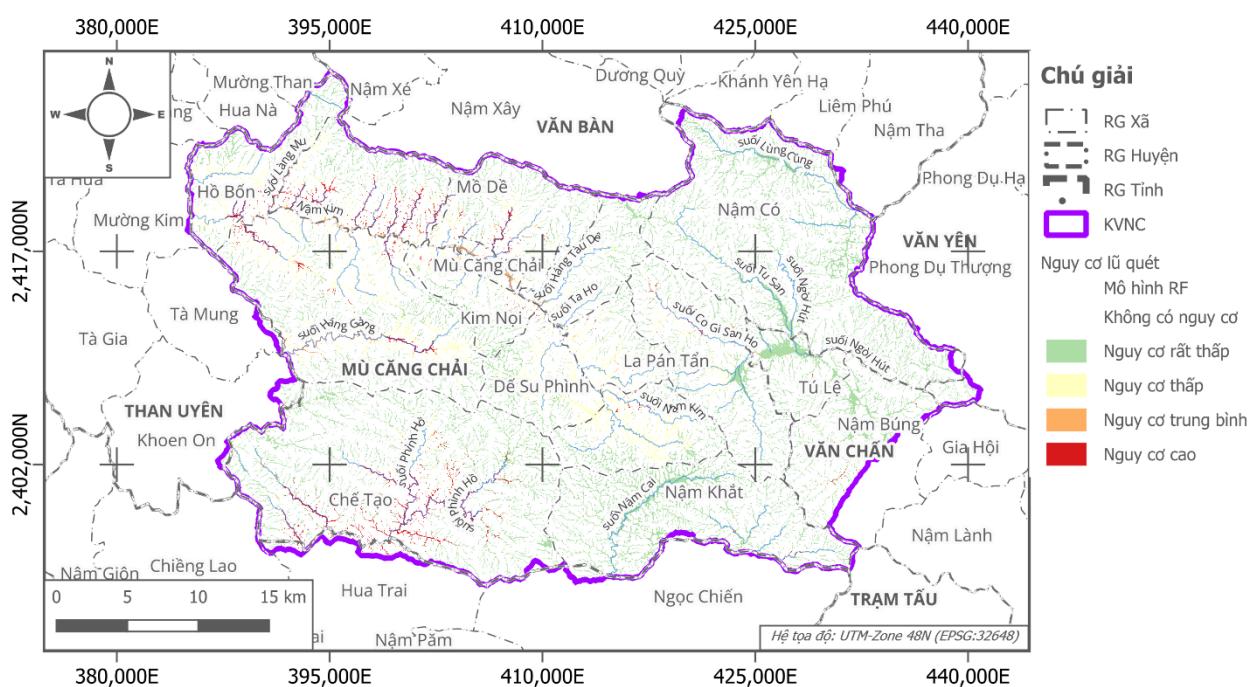
Kết quả này cho thấy tầm quan trọng của việc lựa chọn thuật toán phù hợp với bản chất dữ liệu và bài toán cụ thể, thay vì áp dụng một cách máy móc các mô hình "hiện đại" nhất. Trong lĩnh vực dự báo thiên tai, nơi tính chính xác, tốc độ và khả năng diễn giải đều quan trọng, các mô hình học máy truyền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

### 3.4. Kết quả phân vùng lũ quét cho khu vực nghiên cứu

### 3.4.1 Kết quả phân vùng lũ quét

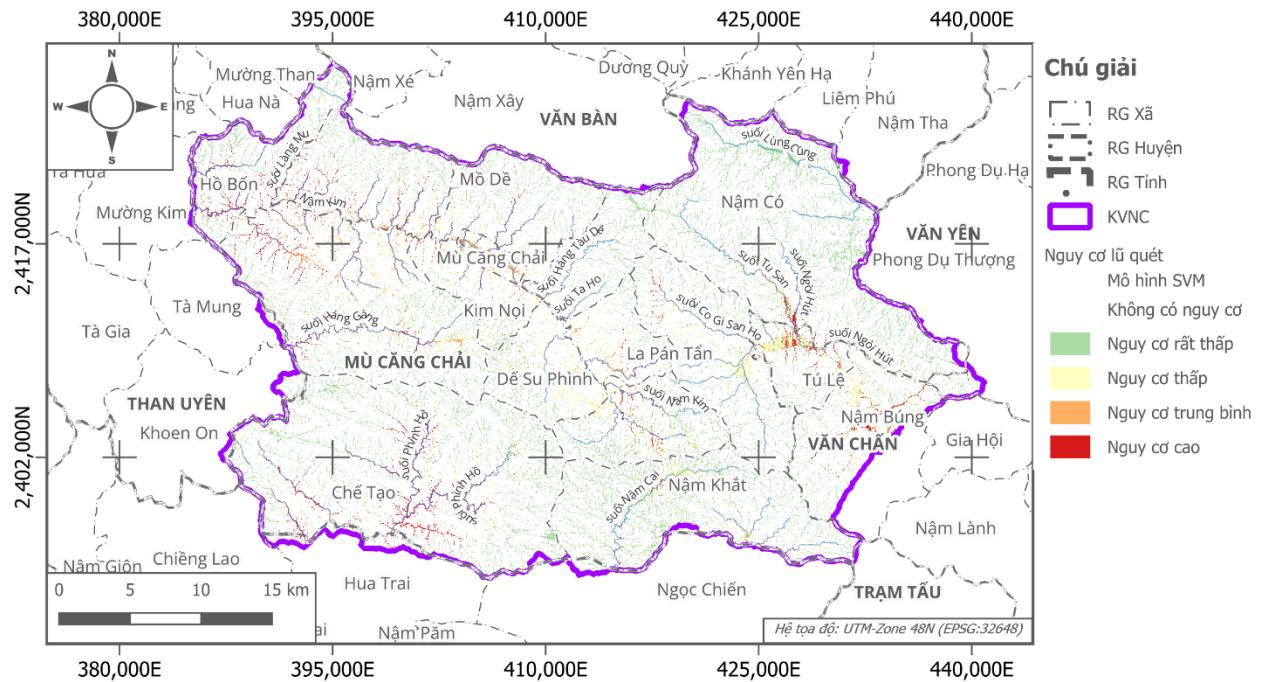
Bản đồ với độ phân giải cao được lưu trữ tại:  
<https://github.com/tamthat/MuCangChai>

## 1. Mô hình RF (rừng ngẫu nhiên)



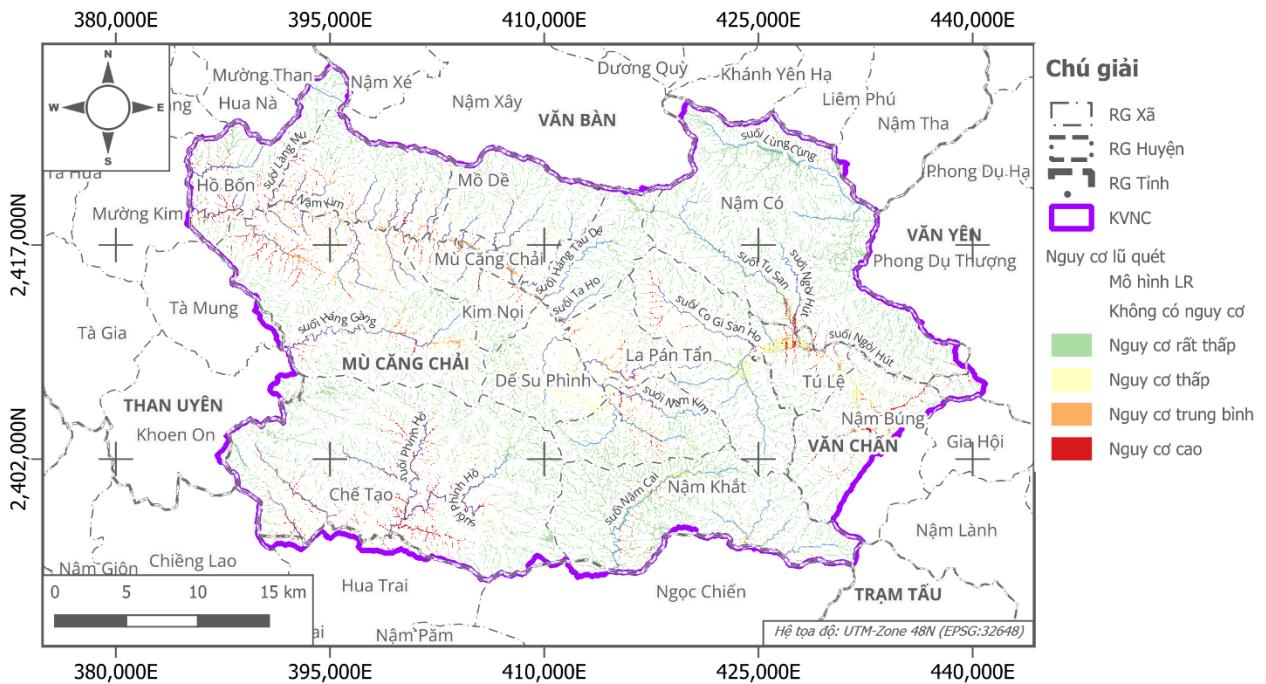
Hình 3-71. Kết quả xác định nguy cơ lũ quét bằng mô hình RF

## 2. Mô hình SVM



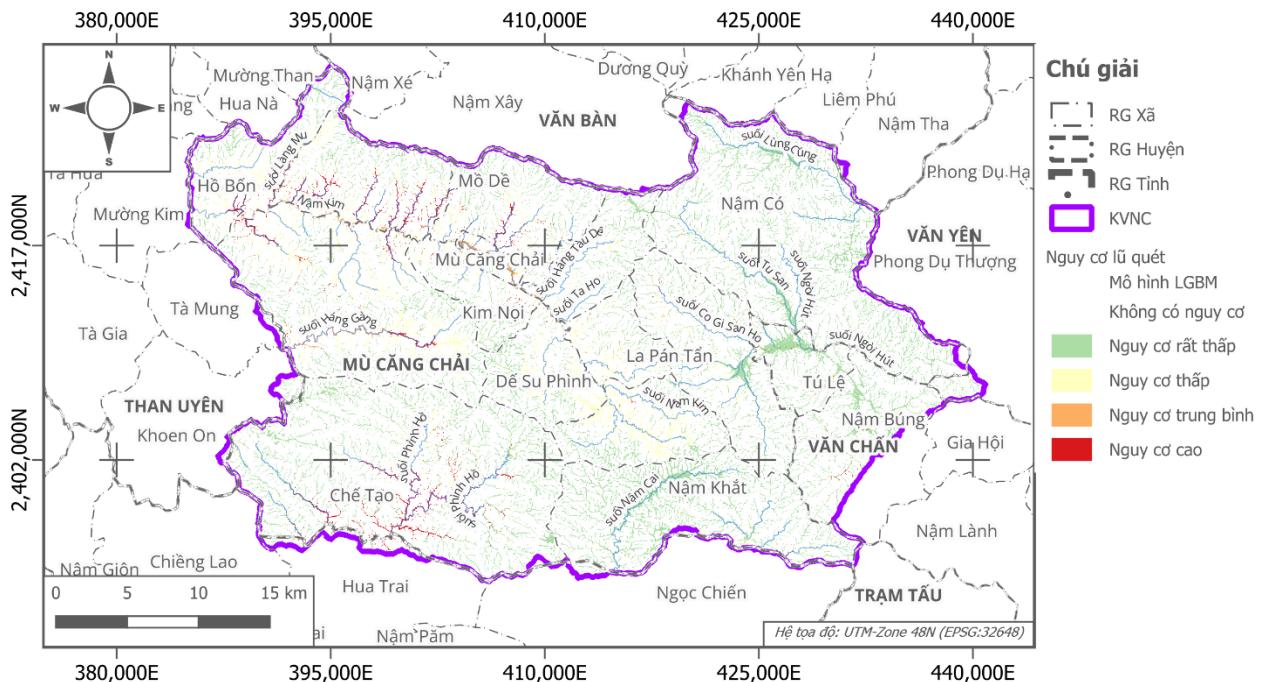
Hình 3-72. Kết quả xác định nguy cơ lũ quét bằng mô hình SVM

## 3. Mô hình hồi quy logistic (LR)



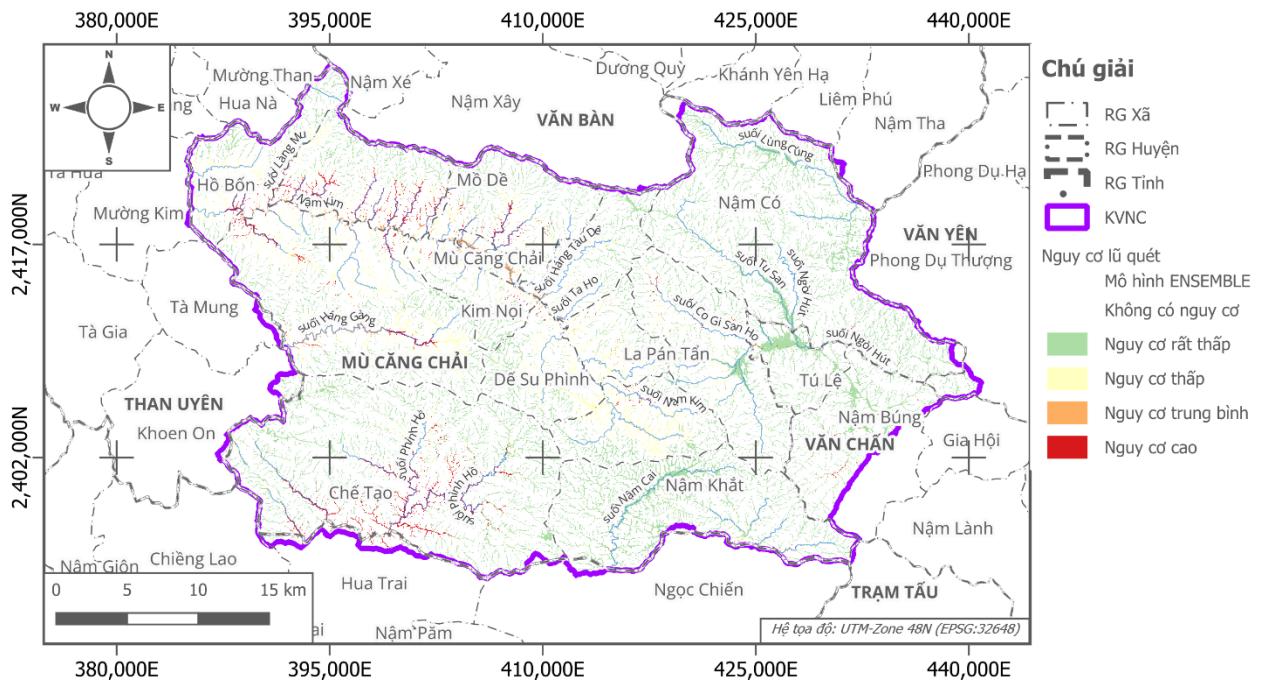
Hình 3-73. Kết quả xác định nguy cơ lũ quét bằng mô hình LR

#### 4. Mô hình LGBM



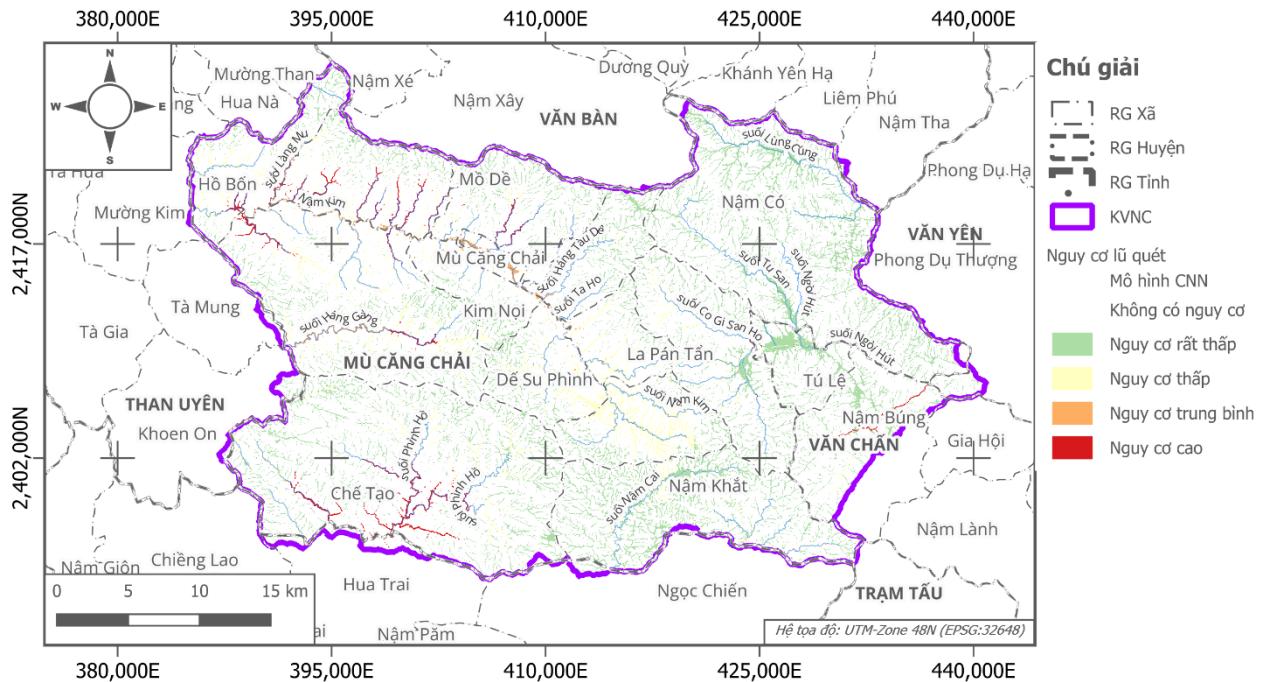
Hình 3-74. Kết quả xác định nguy cơ lũ quét bằng mô hình LGBM

#### 5. Mô hình ENSEMBLE



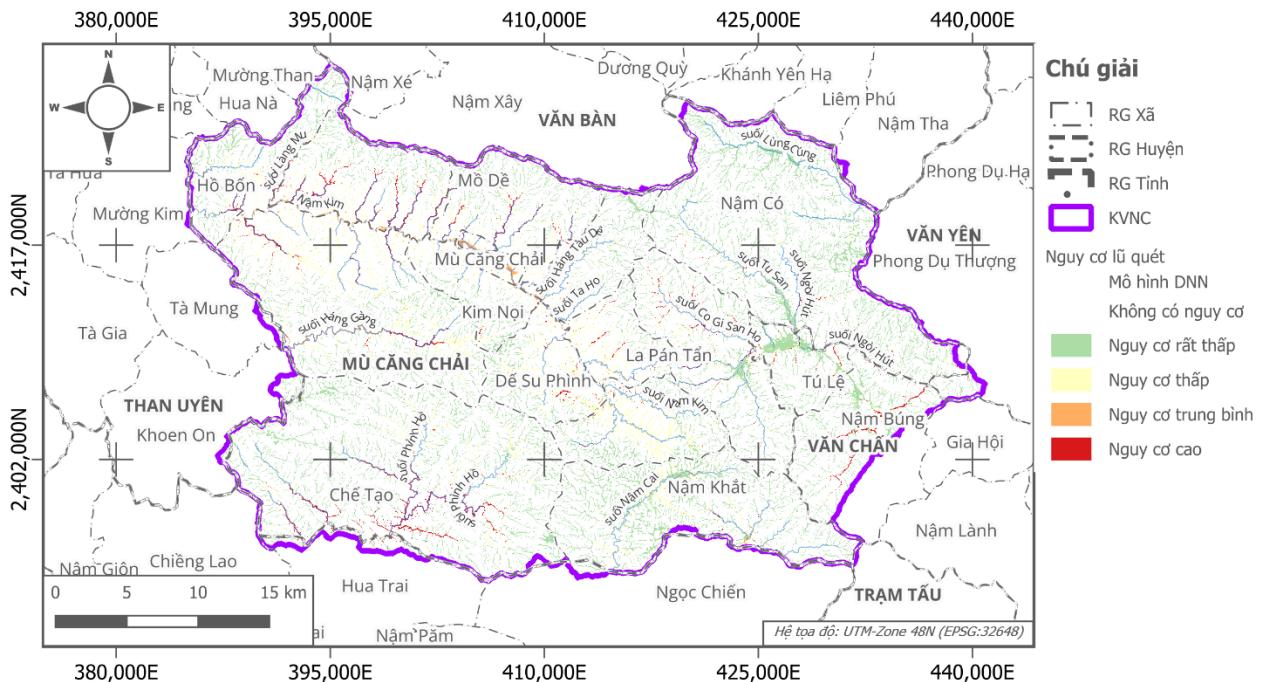
Hình 3-75. Kết quả xác định nguy cơ lũ quét bằng mô hình ENSEMBLE

## 6. Mô hình CNN



Hình 3-76. Kết quả xác định nguy cơ lũ quét bằng mô hình CNN

## 7. Mô hình DNN



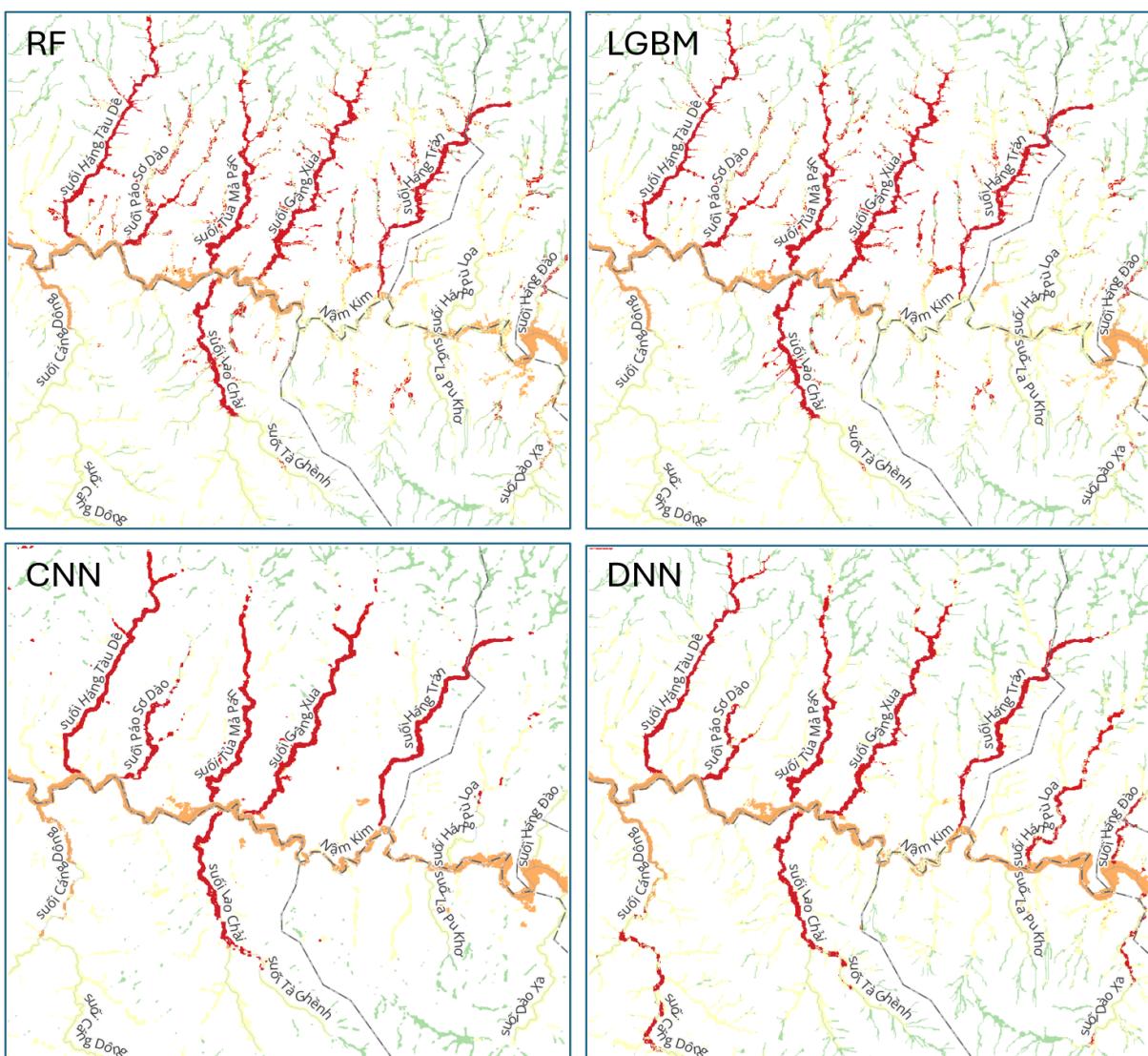
Hình 3-77. Kết quả xác định nguy cơ lũ quét bằng mô hình DNN

### 3.4.2 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 Bồn; (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”.



Hình 3-78. Sự khác biệt phân vùng lũ quét giữa 2 nhóm cây quyết định (RF, LGBM) và nhóm học sâu (CNN, DNN)

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.

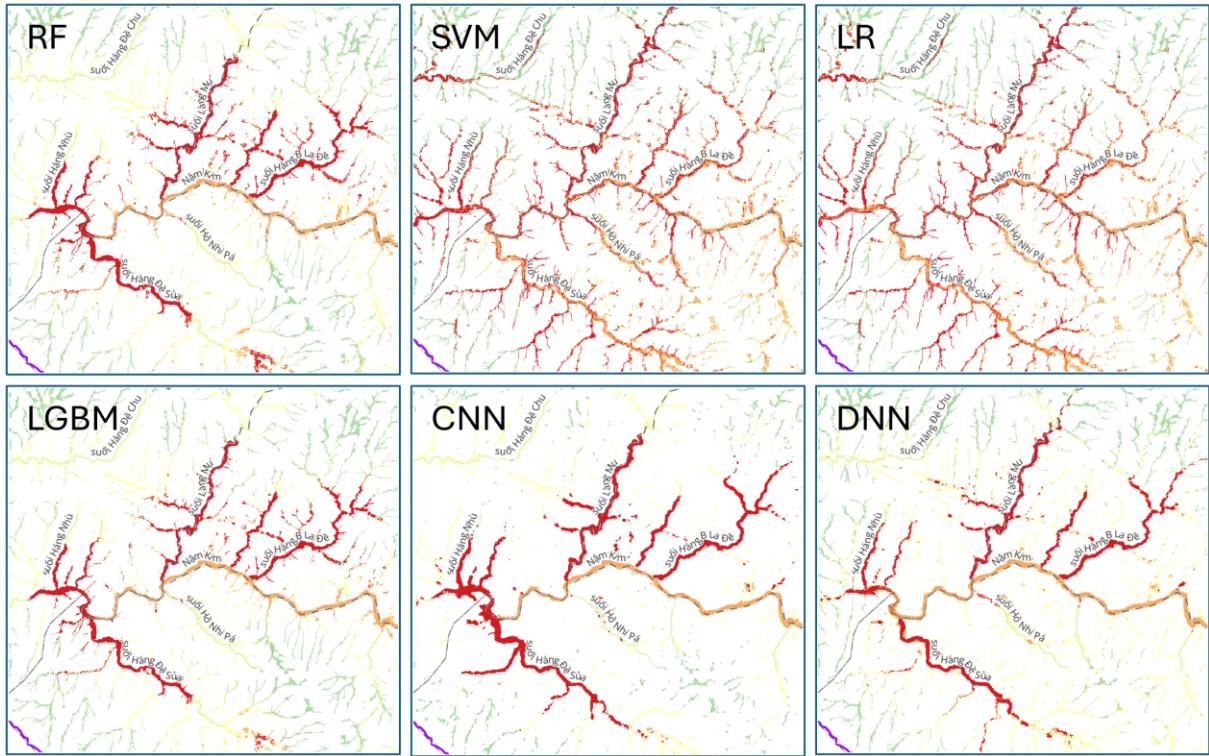
Đặc biệt, các nhánh suối đổ vào các suối nhỏ là các nhánh chỉ có diện tích lưu vực thượng nguồn vài trăm m<sup>2</sup>, khó hoặc không thể có khả năng sinh lũ quét, trong khi đó ở nhóm mô hình cây quyết định, các nhánh suối này vẫn hiển thị nguy cơ cao trước nhập lưu trong khi nhóm mô hình học sâu đánh 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 dãy 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 Xέo 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.

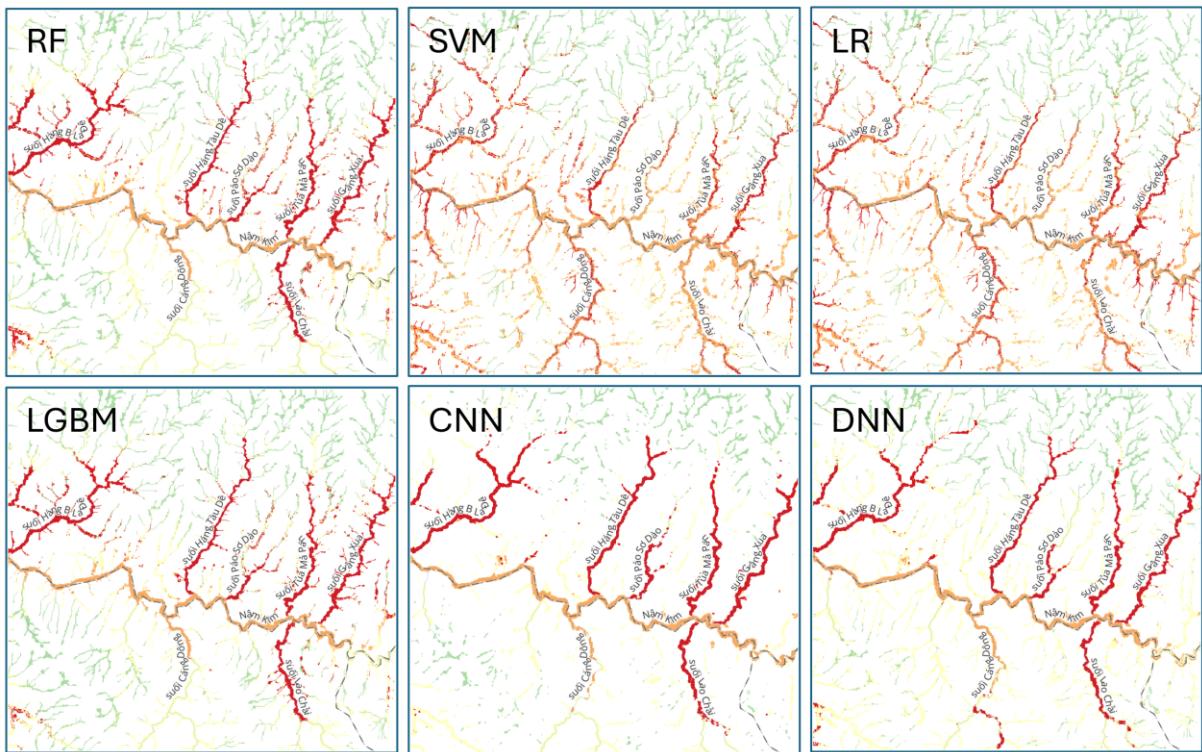


Hình 3-79. Phân vùng lũ quét khu vực xã Hồ Bốn

Như vậy, với khu vực xã Hồ Bốn, 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, DNN, LGBM, RF, CNN, 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 Đè 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 nỗi ở một số nhánh suối thương 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.

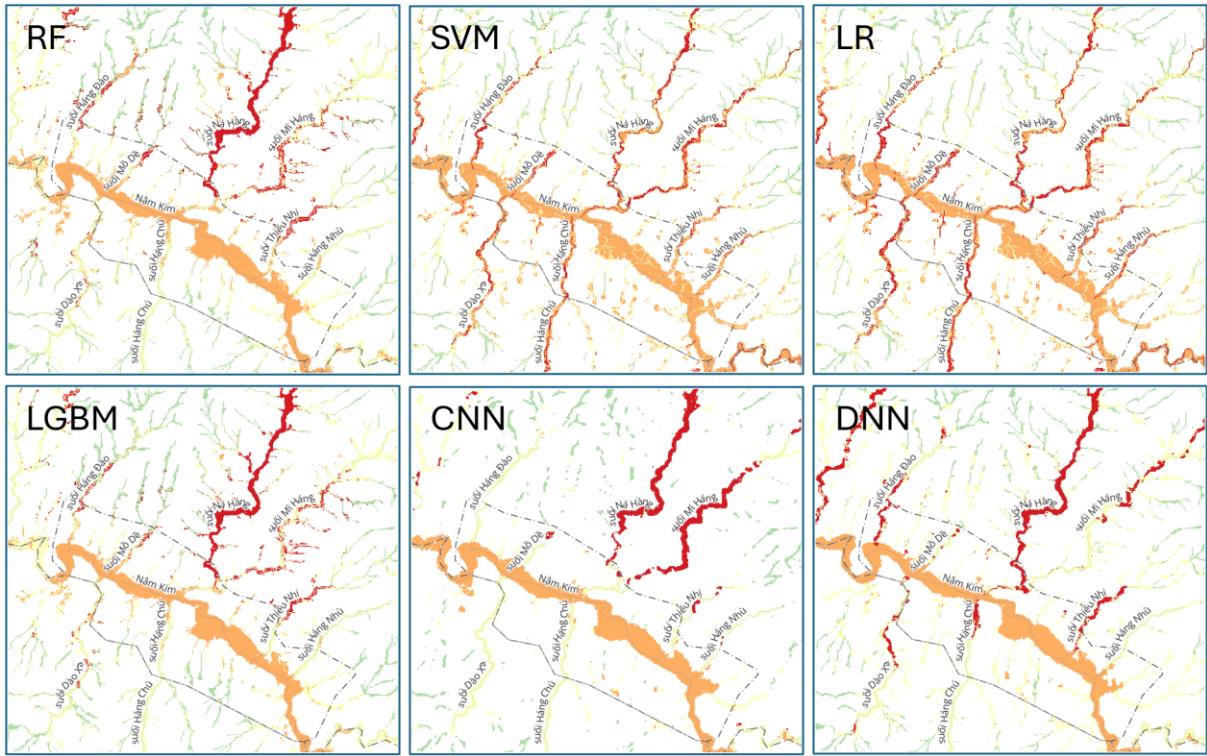


Hình 3-80. Phân vùng lũ quét khu vực xã Khao Mang

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 Cảng Đông 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

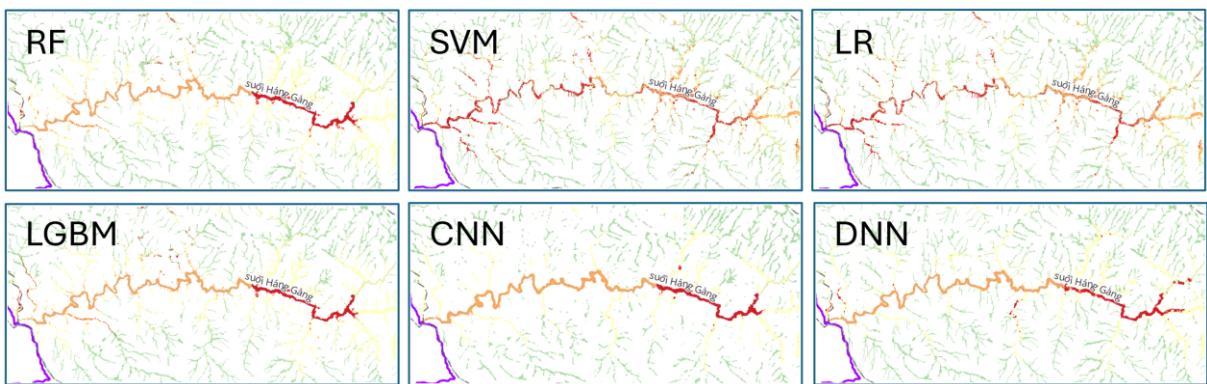
Tại khu vực xã Mò Dè và thị trấn Mù Cang Chải, nơi có xảy ra lũ lớn tại suối Nà Háng và Mỹ Háng. Bên cạnh đó, suối Háng Chủ, nơi xảy ra lũ quét năm 2017 chỉ ghi nhận lũ trung bình. Các mô hình phân loại nguy cơ có sự biến đổi rõ rệt, mỗi mô hình cho ra một kết quả khác nhau



Hình 3-81. 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 Chú 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

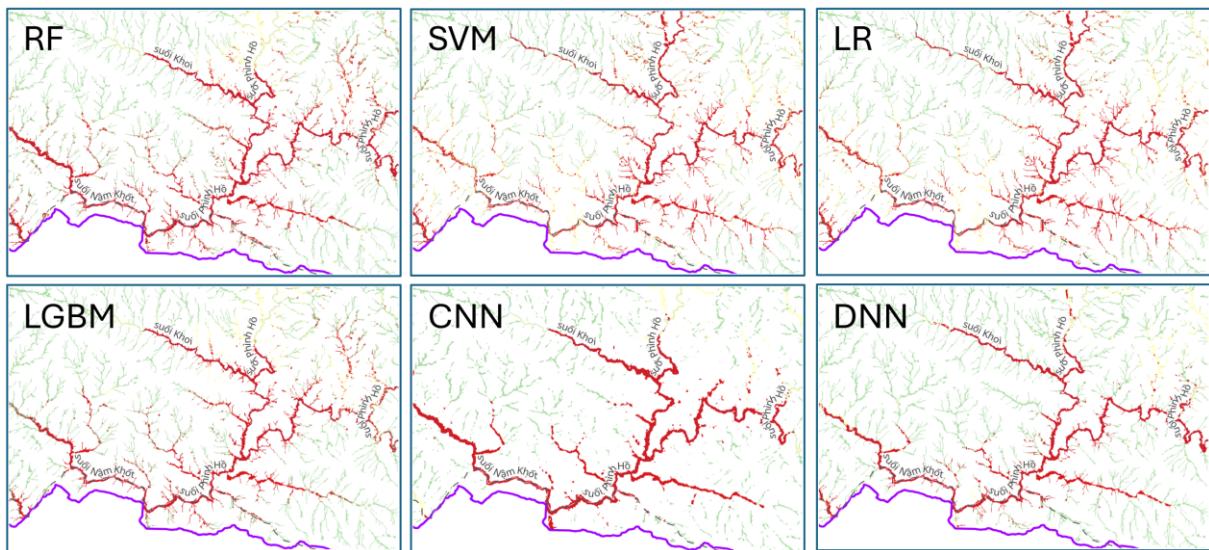


Hình 3-82. Phân vùng lũ quét khu vực xã Lao Chải (suối Háng Gàng)

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 Ché 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 3-83. Phân vùng lũ quét khu vực xã Ché 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

### 3.4.3 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 mơ 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

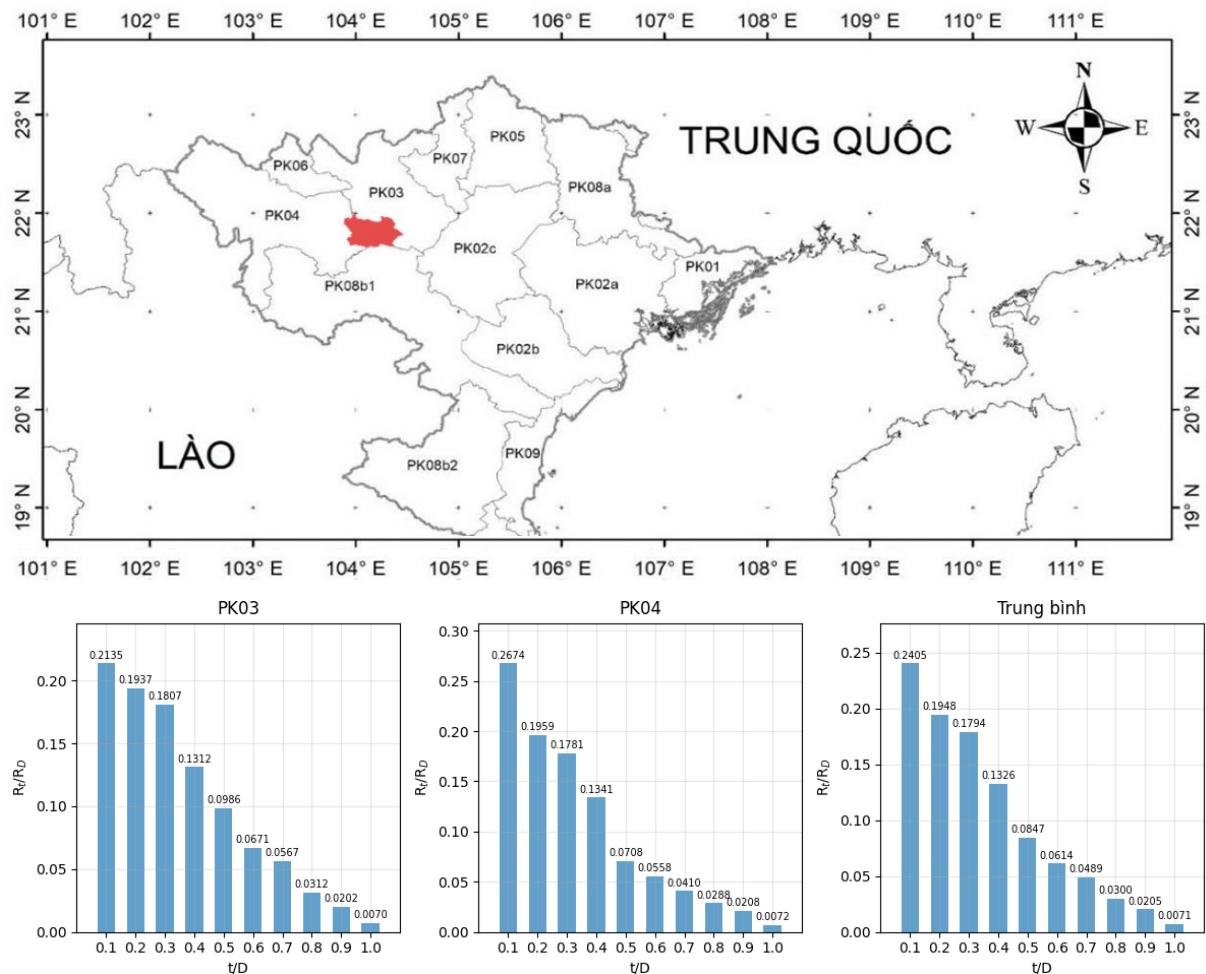
hiệu suất cao hơn, có thể cho ra các điểm tin cậy thiếu chính xác đối với các mẫu khó phân loại.

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 tốt. 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.

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

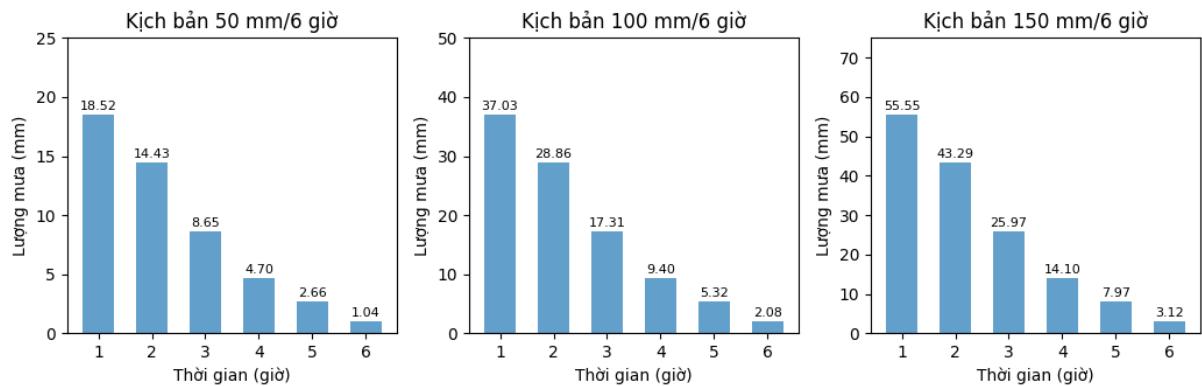
#### 3.5.1 Xây dựng kịch bản mưa

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 3-84. Phân vùng mưa rào theo TCVN 13615:2022 và khu vực nghiên cứu

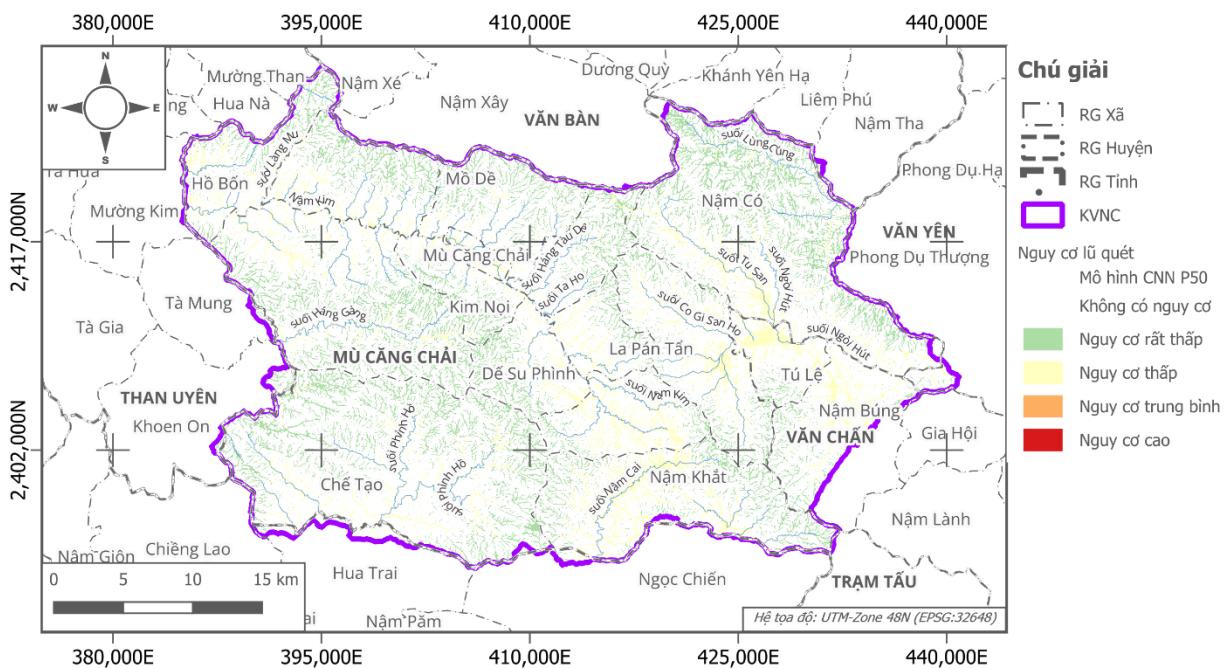
Khu vực nghiên cứu nằm trong vùng PK04 và PK03. Do đó, nghiên cứu này sẽ lấy bình quân phân vùng mưa rào bất lợi nhất của cả hai vùng làm cơ sở để xây dựng biểu đồ phân bố mưa cho các kịch bản mưa. Các kịch bản thể hiện như sau:



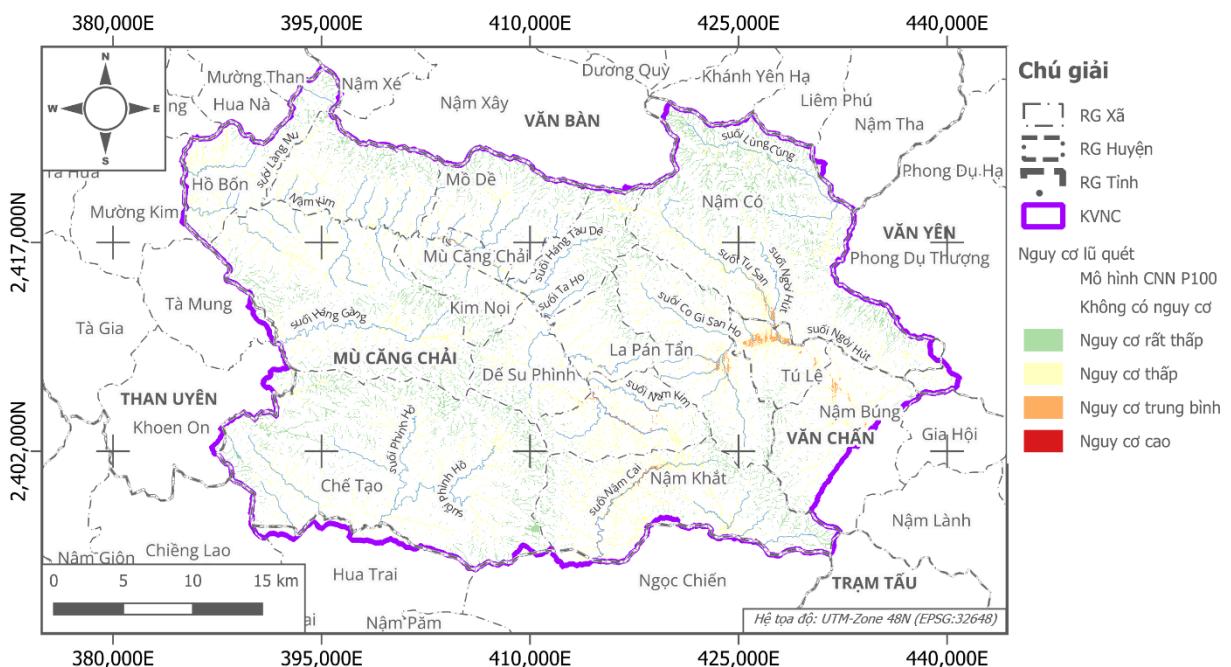
Hình 3-85. Phân bố mưa tương ứng với mô hình mưa bất lợi nhất cho KVNC

Tham số	Kịch bản降雨 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

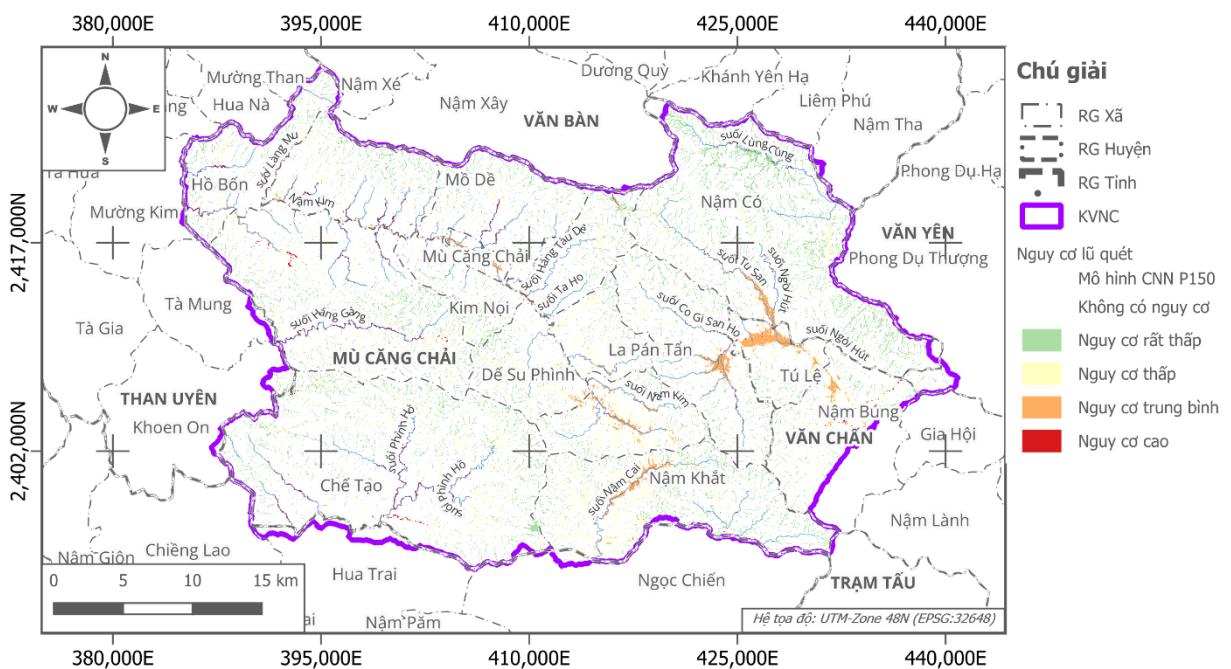
### 3.5.2 Xây dựng bản đồ phân vùng nguy cơ bằng mô hình CNN



Hình 3-86. Kết quả xây dựng bản đồ phân vùng nguy cơ kịch bản mưa 50mm/6 giờ



Hình 3-87. Kết quả xây dựng bản đồ phân vùng nguy cơ kịch bản mưa 100mm/6 giờ



Hình 3-88. Kết quả xây dựng bản đồ phân vùng nguy cơ kịch bản mưa 150mm/6 giờ

## KẾT LUẬN, KIẾN NGHỊ

### Kết luận

Kết quả của nghiên cứu đã đạt được những thành tựu 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 các chỉ số đánh giá tốt. 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 tiềm năng triển khai ứng dụng thực tế với chi phí tính toán hợp lý.

Sự thành công của các mô hình trí tuệ nhân tạo còn thể hiện qua việc phát triển hệ thống phân loại thảm phủ sử dụng mô hình CNN, đạt độ chính xác huấn luyện 98.68% và độ chính xác xác thực 98.32% sau 50 epoch trong việc phân loại 7 lớp thảm phủ khác nhau. Kết quả này không chỉ chứng minh khả năng ứng dụng mạnh mẽ của học sâu trong xử lý dữ liệu viễn thám mà còn tạo nên tầm quan trọng cho việc đánh giá rủi ro lũ quét, vì thảm phủ là một trong những yếu tố then chốt ảnh hưởng đến khả năng thảm nước và dòng chảy bờ mặt. Sự chênh lệch nhỏ giữa độ chính xác huấn luyện và xác thực (chỉ 0.36%) cho thấy mô hình có khả năng tổng quát hóa tốt, tránh được hiện tượng overfitting thường gặp trong các mô hình học sâu, đặc biệt các mô hình học sâu rất phù hợp sử dụng trong phân loại hình ảnh.

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 các nguyên tắc cơ bản 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.

Một điểm đáng chú ý khác là các mô hình phi tuyến sẽ có sự phù hợp tốt hơn các mô hình tuyến tính trong bài toán phân vùng lũ quét. 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. Tuy nhiên, các mô hình tuyến tính như hồi quy logistic hay máy hỗ trợ vectơ lại chưa thể hiện tốt với những tương quan phức tạp và phi tuyến.

## Những hạn chế còn tồn tại

Mặc dù đạt được những kết quả khả quan, nghiên cứu vẫn còn những hạn chế đáng chú ý. Sự chênh lệch lớn về hiệu suất giữa các mô hình cho thấy tính không ổn định trong việc lựa chọn thuật toán phù hợp. Trong khi LGBM đạt 95.24% thì Logistic Regression chỉ đạt 68.97%, điều này phản ánh sự phức tạp trong việc lựa chọn mô hình tối ưu cho từng điều kiện địa lý và khí hậu cụ thể.

Thời gian huấn luyện của các mô hình học sâu như CNN và DNN tương đối dài (4-5 giờ) nhưng hiệu suất lại không vượt trội so với các mô hình truyền thống, đặt ra câu hỏi về hiệu quả kinh tế trong việc triển khai thực tế. Sự khác biệt về hiệu suất này có thể được giải thích qua bản chất khác nhau giữa bài toán phân vùng lũ quét và phân loại thảm phủ. Trong khi phân loại thảm phủ dựa trên các đặc trưng không gian liên tục và có tính chất tương quan cao giữa các pixel lân cận - điều mà CNN rất giỏi trong việc trích xuất thông qua các bộ lọc tích chập, thì phân vùng lũ quét lại phụ thuộc vào sự tương tác phức tạp giữa nhiều yếu tố địa hình, thủy văn và khí tượng mà không nhất thiết có tính liên tục không gian mạnh.

Hơn nữa, dữ liệu đầu vào cho phân vùng lũ quét thường là các giá trị số đơn lẻ (như độ cao, độ dốc, lượng mưa) thay vì dữ liệu hình ảnh có cấu trúc không gian rõ ràng như trong phân loại thảm phủ. CNN và DNN đòi hỏi một lượng lớn dữ liệu để học được các mẫu hình phức tạp, trong khi các mô hình truyền thống như LGBM và Random Forest lại có khả năng xử lý tốt hơn các dữ liệu dạng bảng với số lượng mẫu hạn chế và có thể nắm bắt được mối quan hệ phi tuyến giữa các biến một cách hiệu quả hơn.

Tuy nhiên, về mức độ phù hợp trong phân loại, các mô hình CNN và DNN lại tỏ ra phù hợp hơn dù độ chính xác hay hiệu suất kém hơn các mô hình cây quyết định. Do đó, trong quá trình kiểm chứng, đánh giá cho các trận lũ tiếp theo, cần có sự so sánh và kiểm chứng độc lập nhằm khẳng định độ tin cậy của mô hình phân vùng lũ quét.

Một hạn chế quan trọng khác là việc đánh giá chỉ dựa trên các chỉ số thống kê mà chưa xem xét đến tính khả thi trong ứng dụng thực tế. Độ chính xác cao trên tập dữ liệu thử nghiệm không hoàn toàn đảm bảo hiệu quả khi triển khai trong môi trường thực tế với những biến động không lường trước về khí hậu và địa hình. Ngoài ra, việc đánh giá mức độ lũ chỉ là sự đánh giá chủ quan và thiếu phân định rõ ràng (như thế nào là lớn, thế nào là trung bình...), điều này gây khó khăn thực sự trong phân vùng lũ quét nói chung và phân loại lũ nói riêng. Tuy nhiên, sự nhầm lẫn giữa các cấp độ phân vùng liên tục (giữa các cấp độ và các cấp độ lân cận) là có thể chấp nhận được do thiếu ranh giới rõ ràng.

## Kiến nghị

Bất cập lớn nhất trong việc sử dụng mô hình trí tuệ nhân tạo trong phân vùng/dự báo/cảnh báo nguy cơ lũ quét là dữ liệu, đặc biệt là nhãn dữ liệu. Các trận lũ quét xảy ra

thường không phổ biến, do đó có rất ít mẫu được lấy cho việc huấn luyện mô hình trí tuệ nhân tạo. Ở Việt Nam, thông thường các trận lũ có thiệt hại về người hoặc cuốn trôi nhà cửa mới được xem là lũ quét, do đó các trận lũ quét không gây thiệt hại về người và tài sản, hoặc xảy ra ở các khu vực hẻo lánh không được ghi nhận. Điều này làm hạn chế rất lớn đối với việc xác định lũ quét cho mô hình. Đặc biệt, lũ quét tự nhiên sinh ra bởi mưa lớn, trong khi đó, lượng mưa phân bố ở khu vực miền núi là rất cục bộ, điều này dẫn đến tình trạng lượng mưa quan trắc không đủ độ tin cậy hoặc không phản ánh được chính xác nguyên nhân gây lũ quét, đặc biệt là khoảng trước năm 2021, khi số lượng các trạm quan trắc là không đáng kể (có những trạm cách xa khu vực xảy ra lũ quét hàng vài chục km). Do đó, cần tiến hành điều tra hàng năm về các trận lũ quét với các thông số đầy đủ bao gồm: (1) Phạm vi xảy ra và mức độ lũ quét: Đó không phải là một điểm, đó là một đoạn suối, hoặc một khu vực. Cần định lượng được phạm vi này dựa trên điều tra tại địa phương bằng tọa độ thể hiện trên GIS; (2) Lượng mưa sinh lũ quét: Đánh giá được lượng mưa giờ, lượng mưa tích lũy của đợt lũ quét nhằm phản ánh đúng nguyên nhân sinh lũ.

Ngoài ra, các nghiên cứu về lũ quét sử dụng dữ liệu địa không gian và mô hình trí tuệ nhân tạo trước đây thường sử dụng giá trị nội tại một điểm (độ dốc nội tại, chỉ số thực vật nội tại...), điều này chưa phản ánh đúng về sự hình thành lũ quét (mà phù hợp hơn đối với loại hình sạt lở đất hoặc dòng chảy sinh ra bởi sạt lở). Do đó, cần tiếp cận dựa trên nguyên tắc lưu vực đối với loại hình thiên tai lũ quét nhằm diễn tả quá trình vật lý về sự hình thành lũ. Đây cũng là những hạn chế đã tồn tại nhiều năm trong nghiên cứu lũ quét cần được nghiên cứu chuyên sâu hơn.

Qua nghiên cứu này, nhóm nghiên cứu đề xuất hướng nghiên cứu tiếp theo: **Nghiên cứu xây dựng hệ thống cảnh báo lũ quét ngắn hạn độ phân giải cao cho các lưu vực vừa và nhỏ: tích hợp radar-AI-thủy văn.**

Hướng nghiên cứu này hướng đến một hệ thống cảnh báo hoàn chỉnh cho một khu vực cụ thể, dựa vào dữ liệu radar hiện tại để dự báo lượng mưa trong khoảng 3÷6 giờ tiếp theo (sử dụng mô hình trí tuệ nhân tạo), làm đầu vào cho mô hình thủy văn (có khả năng dự báo đến từng vị trí – pixel) nhằm đánh giá lũ, mức độ lũ phục vụ cảnh báo nguy cơ lũ quét. Hệ thống này có khả năng chạy 24/7 theo thời gian thực nhằm đáp ứng nhu cầu phòng, chống thiên tai với thời gian cảnh báo sớm từ 3÷6 giờ dựa vào thời đoạn mưa dự báo và thời gian tập trung dòng chảy của mỗi lưu vực. Để thực hiện được điều này, mỗi pixel trong khu vực nghiên cứu được xem như cửa ra của một lưu vực con. Từ đó, lưu vực thượng nguồn của mỗi pixel sẽ được xác định và mô hình thủy văn dự báo lũ quét sẽ được xây dựng cho từng lưu vực này. Hệ thống này có tính khả thi cao nhờ kế thừa và phát huy các điểm mạnh của trí tuệ nhân tạo và kết hợp với mô hình thủy văn truyền thống (mô phỏng vật lý), từ đó đưa ra những cảnh báo phù hợp có độ chi tiết cao.

## TÀI LIỆU THAM KHẢO

A Saleh, N Sabtu & M R Bunmi, 2022. Flash Flood Susceptibility Mapping in Sungai Pinang catchment using Weight of Evidence. *IOP Conference Series: Earth and Environmental Science*, 1091(1), p. 012017.

Alarifi, Saad S., Abdelkareem, Mohamed, Abdalla, Fathy & Alotaibi, Mislat, 2022. Flash Flood Hazard Mapping Using Remote Sensing and GIS Techniques in Southwestern Saudi Arabia. *Sustainability*, 14(21), p. 14145.

Amoroch, J. & Brandstetter, A., 1971. Determination of Nonlinear Functional Response Functions in Rainfall-Runoff Processes. *Water Resources Research*, 7(5), pp. 1087-1101.

Anon., 2011. *Yên Bai: Hiệu quả từ mô hình phòng chống, giảm nhẹ thiên tai tại trường THPT Mù Cang Chải*. [Online] Available at: <https://baochinhphu.vn/yen-bai-hieu-qua-tu-mo-hinh-phong-chong-giam-nhe-thien-tai-tai-truong-thpt-mu-cang-chai-102104536.htm> [Accessed 17 4 2025].

Anon., n.d. *Ứng dụng khoa học công nghệ trong phòng, chống thiên tai trên địa bàn tỉnh Yên Bái còn nhiều hạn chế*. [Online] Available at: <https://phongchongthientai.mard.gov.vn/Pages/Ung-dung-khoa-hoc-cong-nghe-trong-phong-chong-thie-5821976939.aspx> [Accessed 17 4 2025].

Aqil Tariq, Jianguo Yan, Bushra Ghaffar & et al., 2022. Flash Flood Susceptibility Assessment and Zonation by Integrating Analytic Hierarchy Process and Frequency Ratio Model with Diverse Spatial Data. *Water*, 14(19), p. 3069.

Artigue, G., Johannet, A., Borrell, V. & Pistre, S., 2011. *Flash floods forecasting without rainfalls forecasts by recurrent neural networks. Case study on the Mialet basin (Southern France)*. s.l., s.n.

Bá Thao, Vũ, 2020. Phương pháp xác định khu vực rủi ro lũ bùn đá dựa vào bản đồ địa hình. *Vietnam Journal of Hydrometeorology*, 713(5), pp. 37-46.

Bakri Bashir, Mohammed, Latiff, Muhammad Shafie Bin Abd, Coulibaly, Yahaya & Yousif, Adil, 2016. A survey of grid-based searching techniques for large scale distributed data. *Journal of Network and Computer Applications*, Volume 60, p. 170÷179.

Band, Shahab S., et al., 2020. Flash Flood Susceptibility Modeling Using New Approaches of Hybrid and Ensemble Tree-Based Machine Learning Algorithms. *Remote Sensing*, 12(21), p. 3568.

Báo Nông Nghiệp Và Môi Trường & Thanh Ngà - Trần Nam, 2023. *Mưa lũ ở Mù Cang Chải, Yên Bái: Nhiều thôn của xã Hồ Bón vẫn chưa tiếp cận*. [Online] Available at: <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](http://www.cua-xa-ho-bon-van-chua-the-tiep-can-i717002.html)

[Accessed 17 4 2025].

Boukharouba, Khaled, Roussel, Pierre, Dreyfus, Gerard & Johannet, Anne, 2013. *Flash flood forecasting using Support Vector Regression: An event clustering based approach.* s.l., s.n.

Box G. E. P. & Cox D. R., 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2), pp. 211-252.

Cao Đăng Dư & Lương Tuấn Anh, 1995. Phân vùng khả năng xuất hiện lũ quét. *Tạp chí Khí tượng thủy văn*, p. 1÷8.

Carpenter, T.M., et al., 1999. National threshold runoff estimation utilizing GIS in support of operational flash flood warning systems. *Journal of Hydrology*, 224(1-2), p. 21–44.

Chang, Tiao J. & Sun, Hong Y., 1997. Study of Potential Flash Floods by Kriging Method. *Journal of Hydrologic Engineering*, 2(3), p. 104–108.

Chen Cao, Peihua Xu, Yihong Wang & et al., 2016. Flash Flood Hazard Susceptibility Mapping Using Frequency Ratio and Statistical Index Methods in Coalmine Subsidence Areas. *Sustainability*, 8(9), p. 948.

Christopher M. Bishop, 1995. *Neural Networks for Pattern Recognition*. Birmingham: Oxford.

Christopher Turner, n.d. *George R. Lawrence, Aeronaut Photographer*. [Online] Available at: <https://www.cabinetmagazine.org/issues/32/turner.php> [Accessed 02 07 2023].

Costache, Romulus, et al., 2021. Flash-Flood Potential Mapping Using Deep Learning, Alternating Decision Trees and Data Provided by Remote Sensing Sensors. *Sensors*, 21(1), p. 280.

Costache, Romulus, Pham, Quoc Bao, Sharifi, Ehsan & et al., 2019. Flash-Flood Susceptibility Assessment Using Multi-Criteria Decision Making and Machine Learning Supported by Remote Sensing and GIS Techniques. *Remote Sensing*, Volume 12, p. 106.

Davidson, Michael W., 2010. Pioneers in Optics: Louis Daguerre and George Eastman. *Microscopy Today*, 18(2), p. 48÷49.

Đinh Sơn, 2018. *Mưa lũ làm 3 người chết và 11 nạn nhân mất tích ở Yên Bai*. [Online]

Available at: <https://znews.vn/mua-lu-lam-3-nguo-chet-va-11-nan-nhan-mat-tich-o-yen-bai-post861935.html> [Accessed 17 4 2025].

Đương Thị Lợi & Đặng Phương Lan, 2021. 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*, Volume 721, p. 31÷45.

F. Silvestro, N. Rebora, G. Cummings & L. Ferraris, 2015. 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*, 10(4), pp. 446-462.

Fausto Guzzetti, Silvia Peruccacci, Mauro Rossi & Colin P. Stark, 2007. The rainfall intensity-duration control of shallow landslides and debris flows: an update. *Landslides*, 5(1), pp. 3-17.

Forest Service, 1931. CAUSES OF FLASHY FLOODS AND MUD FLOWS IN UTAH. *Monthly Weather Review*, Volume 59, p. 122–122.

Fussell, J., Rundquist & D., & Harrington, J. A., 1986. On defining remote sensing. *Photogrammetric Engineering and Remote Sensing*, 92(9), p. 1507÷1511.

Georgakakos, Konstantine P., 1986. A generalized stochastic hydrometeorological model for flood and flash-flood forecasting: 1. Formulation. *Water Resources Research*, 22(13), p. 2083–2095.

Geraldo Moura Ramos Filho, Victor Hugo Rabelo Coelho, Emerson da Silva Freitas & et al., 2020. An improved rainfall-threshold approach for robust prediction and warning of flood and flash flood hazards. *Natural Hazards*, 105(3), pp. 2409-2429.

Ha, Hang, et al., 2022. A machine learning approach in spatial predicting of landslides and flash flood susceptible zones for a road network. *Modeling Earth Systems and Environment*, 8(4), p. 4341÷4357.

Hales, John E., 1978. The Kansas City Flash Flood of 12 September 1977. *Bulletin of the American Meteorological Society*, 59(6), p. 706–710.

Han J., Kamber M. & Pei J., 2012. *Data Mining: Concepts and Techniques*. s.l.:Elsevier.

Hastie, T., Tibshirani, R. & Friedman, J., 2008. Model Assessment and Selection. In: *The Elements of Statistical Learning*. s.l.:Springer New York, pp. 219-259.

He, Fei, Liu, Suxia, Mo, Xingguo & Wang, Zhonggen, 2025. Interpretable flash flood susceptibility mapping in Yarlung Tsangpo River Basin using H2O Auto-ML. *Scientific Reports*, 15(1), p. .

Heppener, Marc, 2008. Spaceward ho!: The future of humans in space. *EMBO reports*, 9(S1).

Hoang, Duc-Vinh & Liou, Yuei-An, 2024. Assessing the influence of human activities on flash flood susceptibility in mountainous regions of Vietnam. *Ecological Indicators*, 158(), p. 111417.

Hsu, Kuo-lin, Gupta, Hoshin Vijai & Sorooshian, Soroosh, 1995. Artificial Neural Network Modeling of the Rainfall-Runoff Process. *Water Resources Research*, 31(10), pp. 2517-2530.

I. K. Westerberg, J.-L. Guerrero, P. M. Younger & et al., 2011. Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences*, 15(7), pp. 2205-2227.

Ilia, Ioanna, et al., 2022. Flash flood susceptibility mapping using stacking ensemble machine learning models. *Geocarto International*, 37(27), p. 15010÷15036.

In-Kwon Yeo & Richard A. Johnson, 2000. A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika*, 87(4), pp. 954-959.

Izyan 'Izzati Abdul Rahman & Nik Mohd Asrol Alias, 2011. *Rainfall forecasting using an artificial neural network model to prevent flash floods*. s.l., s.n.

Janál, Petr & Starý, Miloš, 2009. Fuzzy model use for prediction of the state of emergency of river basin in the case of flash flood. *Journal of Hydrology and Hydromechanics*, 57(3).

JICA, 2021. *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*, Hà Nội: Cơ quan hợp tác quốc tế Nhật Bản (JICA).

Jonathan D Phillips, 2002. Geomorphic impacts of flash flooding in a forested headwater basin. *Journal of Hydrology*, 269(3-4), pp. 236-250.

Kairong Lin, Haiyan Chen, Chong-Yu Xu & et al., 2020. Assessment of flash flood risk based on improved analytic hierarchy process method and integrated maximum likelihood clustering algorithm. *Journal of Hydrology*, Volume 584, p. 124696.

Kock, Winston E., 1978. Radar. In: *The Creative Engineer*. s.l.:Springer US, p. 219÷238.

Kong A Siou, L., Johannet, A., Pistre, S. & Borrell, V., 2010. Flash Floods Forecasting in a Karstic Basin Using Neural Networks: the Case of the Lez Basin (South of France). In: *Environmental Earth Sciences*. s.l.:Springer Berlin Heidelberg, pp. 215-221.

Kuz'min, K. K., 1974. The catastrophic flash flood of 1973 and the medeo dam. *Hydrotechnical Construction*, 8(3), p. 203–206.

L. Alfieri, D. Velasco & J. Thielen, 2011. Flash flood detection through a multi-stage probabilistic warning system for heavy precipitation events. *Advances in Geosciences*, Volume 29, pp. 69-75.

Lã Thanh Hà, 2009. 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*, 578(2), pp. 11-15.

Lal, Aleena B, et al., 2024. Flash Flood Detection and Alert System Using Machine Learning. *International Journal for Research in Applied Science and Engineering Technology*, 12(6), p. 45÷51.

Lamovec, Peter, Veljanovski, Tatjana, Mikoš, Matjaž & Oštir, Krištof, 2013. Detecting flooded areas with machine learning techniques: case study of the Selška Sora river flash flood in September 2007. *Journal of Applied Remote Sensing*, 7(1), p. 073564.

Landsat Science, n.d. *Landsat* 1. [Online] Available at: <https://landsat.gsfc.nasa.gov/satellites/landsat-1/> [Accessed 23 07 2023].

Lorenzo Alfieri, Marc Berenguer, Valentin Knechtl & et al., 2015. Handbook of Hydrometeorological Ensemble Forecasting. In: *Flash Flood Forecasting Based on Rainfall Thresholds*. Berlin: Springer Berlin Heidelberg, pp. 1-38.

Lugt, Dorien, van Hoek, Mattijn, Meirink, Jan Fokke & van der Kooij, Eva, 2021. Nowcasting for urban flash floods in Africa: a machine-learning and satellite-observation based model. , (), p. .

Mạnh Cường, 2019. *Thiệt hại do mưa, lũ gây ra trên địa bàn huyện Mù Cang Chải* ước khoảng 920 triệu đồng. [Online] Available at: <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> [Accessed 17 4 2025].

Marco Borga, et al., 2014. Hydrogeomorphic response to extreme rainfall in headwater systems: Flash floods and debris flows. *Journal of Hydrology*, Volume 518, pp. 194-205.

Maxwell, James Clerk, 2010. *A Treatise on Electricity and Magnetism*. s.l.:Cambridge University Press.

MDE, n.d. *Method for Designing Infiltration Structures*. [Online] Available at: [https://mde.maryland.gov/programs/water/StormwaterManagementProgram/Document%20s/www.mde.state.md.us/assets/document/sedimentstormwater/Appnd\\_D13.pdf](https://mde.maryland.gov/programs/water/StormwaterManagementProgram/Document%20s/www.mde.state.md.us/assets/document/sedimentstormwater/Appnd_D13.pdf) [Accessed 15 6 2023].

Minh Đức, Đào, et al., 2022. 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*, EME4(1), pp. 341-354.

Mohamed Saber & Koray Yilmaz, 2018. Evaluation and Bias Correction of Satellite-Based Rainfall Estimates for Modelling Flash Floods over the Mediterranean region: Application to Karpuz River Basin, Turkey. *Water*, 10(5), p. 657.

Mohammed Sadek & Xuxiang Li, 2019. 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*, Volume 2019, pp. 1-15.

Moore, I. D., Grayson, R. B. & Ladson, A. R., 1991. Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. *Hydrological Processes*, jan, 5(1), pp. 3-30.

Nejc Bezak, Mojca vSraj & Matjavz Mikovs, 2016. Copula-based IDF curves and empirical rainfall thresholds for flash floods and rainfall-induced landslides. *Journal of Hydrology*, Volume 541, pp. 272--284.

Nel Caine, 1980. The Rainfall Intensity: Duration Control of Shallow Landslides and Debris Flows. *Geografiska Annaler. Series A, Physical Geography*, 62(1/2), p. 23.

Ngô Thị Phương Thảo, 2024. *Luận án: Nghiên cứu phát triển mô hình trí tuệ nhân tạo trong phân vùng nguy cơ lũ quét ở Việt Nam*, s.l.: Trường đại học Mỏ - Địa chất.

Nguyễn Viết Nghĩa & Nguyễn Cao Cường, 2020. 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*, 44(6), p. 56÷64.

NOAA, 1979. *National Weather Service forecasting handbook (No. 1 - 1979)*. s.l.:University of Michigan Library (January 1, 1979).

NRCS, 2020. Part 630 - Hydrology. [Online] Available at: <https://directives.nrcs.usda.gov/sites/default/files2/1712930634/Part%20630%20-%20Hydrology.pdf> [Accessed 11 12 2023].

NRCS, n.d. Soil Infiltration. [Online] Available at: [https://web.archive.org/web/20240301064123/https://cropwatch.unl.edu/documents/USDA\\_NRCS\\_infiltration\\_guide6-4-14.pdf](https://web.archive.org/web/20240301064123/https://cropwatch.unl.edu/documents/USDA_NRCS_infiltration_guide6-4-14.pdf) [Accessed 11 6 2023].

Pedregosa, Fabian, 2012. Scikit-learn: Machine Learning in Python.

Petr Sercl, et al., 2023. *Flash Flood Indicator*, Prague: Czech Hydrometeorological Institute.

Phạm Thị Hương Lan & Vũ Minh Cát, 2008. 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*, 556(2), pp. 11-16.

Philip McCouat, 2016. The adventures of Nadar: photography, ballooning, invention and the impressionists. *Journal of Art in Society*.

Piotrowski, A., Napiórkowski, J. J. & Rowiński, P.M., 2006. Flash-flood forecasting by means of neural networks and nearest neighbour approach – a comparative study. *Nonlinear Processes in Geophysics*, 13(4), pp. 443-448.

*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).*

Rahmati, Mehdi, et al., 2018. Development and analysis of the Soil Water Infiltration Global database. *Earth System Science Data*, 10(3), pp. 1237-1263.

Razavi-Termeh, Seyed Vahid, Seo, MyoungBae, Sadeghi-Niaraki, Abolghasem & Choi, Soo-Mi, 2023. 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*, 41(), p. 100595.

Romdani, R. P. et al., 2018. Development of Flash Flood Hazard Map in Bima City (NTB) using Analytical Hierarchy Process. *IOP Conference Series: Earth and Environmental Science*, Volume 166, p. 012035.

Romulus Costache & Liliana Zaharia, 2017. 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*, 126(4), pp. 1-19.

ROSS, C. et al., 2018. *Global Hydrologic Soil Groups (HYSOGs250m) for Curve Number-Based Runoff Modeling*. s.l.:ORNL Distributed Active Archive Center.

S. Talha, M. Maanan, H. Atika & H. Rhinane, 2019. 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*, Volume XLII-4/W19, pp. 407-414.

Sahoo, G.B., Ray, C. & De Carlo, E.H., 2006. Use of neural network to predict flash flood and attendant water qualities of a mountainous stream on Oahu, Hawaii. *Journal of Hydrology*, 327(3-4), pp. 525-538.

Seann Reed, John Schaake & Ziya Zhang, 2007. A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. *Journal of Hydrology*, 337(3-4), pp. 402-420.

Sellami, El Mehdi & Rhinane, Hassan, 2024. A modern method for building damage evaluation using deep learning approach - Case study: Flash flooding in Derna, Libya. *E3S Web of Conferences*, 502(), p. 03010.

SELLAMI, EL Mehdi & Rhinane, Hassan, 2024. Google Earth Engine and Machine Learning for Flash Flood Exposure Mappingâ€”Case Study: Tetouan, Morocco. *Geosciences*, 14(6), p. 152.

Shahin Khosh Bin Ghomash, Daniel Bachmann, Daniel Caviedes-Voulli`eme & Christoph Hinz, 2022. Impact of Rainfall Movement on Flash Flood Response: A Synthetic Study of a Semi-Arid Mountainous Catchment. *Water*, 14(12), p. 1844.

Sharada, D., Devi, D. Kaveri, Prasad, S. & Kumar, Seelan Santhosh, 1997. Modelling flash flood hazard to a railway line: A GIS approach. *Geocarto International*, 12(3), p. 77–82.

Sliney, David H., Bitran, Maurice & Murray, William, 2012. *Infrared, Visible, and Ultraviolet Radiation*. s.l.:Wiley.

Stock, Kristin & Guesgen, Hans, 2016. Geospatial Reasoning With Open Data. In: *Automating Open Source Intelligence*. s.l.:Elsevier, p. 171÷204.

Sweeney, Timothy L., 1992. *Modernized Areal Flash Flood Guidance*, s.l.: NOAA Technical Memorandum NWS HYDRO 44.

T. Turkington, J. Ettema, C. J. van Westen & K. Breinl, 2014. Empirical atmospheric thresholds for debris flows and flash floods in the southern French Alps. *Natural Hazards and Earth System Sciences*, 14(6), pp. 1517-1530.

Thanh Thủy, n.d. *Lũ quét ở Mù Cang Chải: Uớc thiệt hại khoảng 150 tỷ đồng*. [Online]

Available at: [https://baoyenbai.com.vn/12/151798/Luquet\\_o\\_Mu\\_Cang\\_Chai\\_Uoc\\_thiet\\_hai\\_khoan\\_g\\_150\\_ty\\_dong.htm](https://baoyenbai.com.vn/12/151798/Luquet_o_Mu_Cang_Chai_Uoc_thiet_hai_khoan_g_150_ty_dong.htm)

[Accessed 17 4 2025].

Theo Vietnamplus, n.d. *Yên Bái: Bảy người mất tích do lũ ống, lũ quét ở Mù Cang Chải*. [Online]

Available at: <https://daidoanket.vn/yen-bai-bay-nguo-mat-tich-do-lu-ong-lu-quet-o-mu-cang-chai-10078324.html>

[Accessed 17 4 2025].

Thị Huyền, Nguyễn, et al., 2023. Kết quả khoanh định các khu vực nhạy cảm về trượt lở, lũ quét khu vực Thành phố Đà Nẵng. *Vietnam Journal of Hydrometeorology*, 1(745), pp. 21-33.

Thị Phương Thảo, Ngô, Hùng Long, Ngô, Anh Tuấn, Trần & Minh Hằng, Lê, 2024. Sử dụng ảnh Sentinel-1A đa thời gian để phát hiện lũ quét, thử nghiệm tại tỉnh Lào Cai. *Journal of Hydro-meteorology*, 8(764), pp. 29-37.

Thomas L Saaty, 1977. A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, 15(3), pp. 234-281.

Thủy Thanh, 2020. *Yên Bái: Chủ động ứng phó với mưa lớn diện rộng từ chiều tối ngày 14 đến 16/10*. [Online]  
Available at: <https://baoyenbai.com.vn/PrintPreview/198871/>  
[Accessed 15 6 2023].

Toukourou, Mohamed Samir, Johannet, Anne & Dreyfus, Gérard, 2009. Flash Flood Forecasting by Statistical Learning in the Absence of Rainfall Forecast: A Case Study. In: *Communications in Computer and Information Science*. s.l.:Springer Berlin Heidelberg, pp. 98-107.

Vũ Bá Thao & Bùi Xuân Việt, 2023. Phân tích ngưỡng mưa phát sinh một số trận lũ quét, lũ bùn đá thuộc các tỉnh Lai Châu, Điện Biên, Yên Bái, Sơn La. *Tạp chí Khí tượng thủy văn*, Volume 749, pp. 96-110.

Wenlin Yuan, Xinyu Tu, Chengguo Su & et al., 2021. Research on the Critical Rainfall of Flash Floods in Small Watersheds Based on the Design of Characteristic Rainfall Patterns. *Water Resources Management*, 35(10), pp. 3297-3319.

White, Jack, 2012. Herschel and the Puzzle of Infrared. *American Scientist*, 100(3), p. 218.

Wikipedia, n.d. Nadar. [Online] Available at: <https://en.wikipedia.org/wiki/Nadar> [Accessed 15 06 2023].

William W. Emmett, 1975. *The Channels and Waters of the Upper Salmon River Area, Idaho*, Washington: United States Government Printing Office.

Xiaoyan Zhai, Liang Guo, Ronghua Liu & Yongyong Zhang, 2018. Rainfall threshold determination for flash flood warning in mountainous catchments with consideration of antecedent soil moisture and rainfall pattern. *Natural Hazards*, 94(2), pp. 605-625.

Zhao, Gang, et al., 2022. Large-scale flash flood warning in China using deep learning. *Journal of Hydrology*, Volume 604, p. 127222.