# Event-based Flood Susceptibility Mapping for Ha Tinh Province, Vietnam: Methodology and Preliminary Results

(Draft for academic discussion)

## 1 Tổng quan bài toán và mục tiêu nghiên cứu

Lũ lụt là một trong những rủi ro thiên tai nghiêm trọng nhất tại khu vực Bắc Trung Bộ Việt Nam, đặc biệt là tỉnh Hà Tĩnh, nơi có địa hình phân hóa mạnh giữa vùng núi phía Tây và đồng bằng ven sông, ven biển. Thay vì tập trung vào việc dự báo thời điểm xảy ra lũ (flood forecasting), nghiên cứu này hướng tới bài toán **đánh giá nguy cơ lũ lụt không gian (flood susceptibility mapping)**, tức là xác định *những khu vực nào có khả năng bị ngập cao hơn các khu vực khác*, dựa trên đặc điểm địa hình, thủy văn, mặt phủ và khí hậu.

Mục tiêu của nghiên cứu là:

- Xây dựng bản đồ nguy cơ lũ (Flood Susceptibility Index) dưới dạng xác suất liên tục từ 0 đến 1.
- Đảm bảo kết quả vừa có độ chính xác cao về mặt thống kê, vừa phù hợp với quy luật vật lý – địa mạo.
- Cung cấp khả năng giải thích (explainability) rõ ràng cho mô hình học máy.

## 2 Xây dựng dữ liệu nhãn: Flood Baseline

### 2.1 Nguyên tắc xây dựng baseline

Thay vì sử dụng dữ liệu thiệt hại hành chính hoặc bản đồ lũ đơn lẻ, nghiên cứu này xây dựng **baseline lũ dựa trên dữ liệu Sentinel-1 SAR**, khai thác đặc tính phản xạ thấp của mặt nước trên ảnh radar.

Baseline lũ được xác định theo từng sự kiện lũ trong giai đoạn 2016–2025, với các nguyên tắc bảo thủ:

- Chỉ sử dụng ảnh Sentinel-1 VH, quỹ đạo DESCENDING để giảm nhiễu địa hình.
- Áp dụng kỹ thuật *minimum composite* trong cửa sổ thời gian sự kiện để đảm bảo bắt được trạng thái ngập lớn nhất.
- Ngưỡng phân tách nước được cố định ở -19 dB, kết hợp với lọc địa hình (slope $< 10^{\circ}$).
- Loại bỏ các cụm nhiễu nhỏ bằng lọc hình thái học (connected pixel count).

Kết quả của bước này là tập các bản đồ nhị phân (0/1) thể hiện sự xuất hiện ngập lụt cho từng sự kiện. Các bản đồ này được tổng hợp để tạo ra bản đồ tần suất ngập, đóng vai trò là dữ liệu nhãn cho mô hình học máy.

*(Đính kèm: Sơ đồ pipeline baseline + ví dụ bản đồ tần suất lũ)*

# 3 Xây dựng tập biến đầu vào (Feature Engineering)

## 3.1 Nhóm địa hình

Các biến địa hình được trích xuất từ SRTM 30m, bao gồm:

- Độ dốc (Slope)
- Hướng sườn (Aspect)
- Độ cong địa hình (Curvature – Laplacian proxy)
- Chỉ số địa hình tương đối (Relief_2km), đại diện cho độ cao tương đối so với đáy thung lũng trong bán kính 2 km

Trong đó, Relief_2km đóng vai trò như một biến thay thế cho HAND, giúp phân biệt rõ vùng trũng ngập và vùng đồi cao.

## 3.2 Nhóm thủy văn

Hai biến thủy văn chính được sử dụng:

- TWI (Topographic Wetness Index), tính từ Flow Accumulation và độ dốc.
- Khoảng cách tới mặt nước (Dist_Water), tính từ lớp mặt nước của ESA WorldCover.

Biến Dist_Water được biến đổi bằng hàm logarit tự nhiên (log1p) nhằm giảm ảnh hưởng của các giá trị ngoại lai rất lớn ở vùng xa sông.

## 3.3 Nhóm khí hậu

Lượng mưa trung bình mùa lũ (tháng 9–11) trong giai đoạn 2016–2025 được trích xuất từ bộ dữ liệu CHIRPS. Biến này đại diện cho điều kiện mưa nền (climatological forcing) thay vì mưa sự kiện.

## 3.4 Nhóm mặt phủ

Lớp sử dụng đất (LULC) được lấy từ ESA WorldCover và xử lý bằng kỹ thuật One-Hot Encoding để đảm bảo mô hình không áp đặt thứ tự giả tạo lên các lớp phân loại.

# 4 Phân tích tương quan và xử lý đa cộng tuyến

Trước khi huấn luyện mô hình, mối tương quan giữa các biến đầu vào được phân tích bằng hệ số Pearson. Kết quả cho thấy Elevation và Relief_2km có tương quan rất cao (r 0.90), dẫn đến nguy cơ đa cộng tuyến.

Do đó, Elevation được loại bỏ khỏi tập biến đầu vào, trong khi Relief_2km được giữ lại vì mang ý nghĩa địa mạo trực tiếp hơn đối với ngập lụt.

*(Đính kèm: Ma trận tương quan và scatter plot Elevation vs Relief_ 2km)*

# 5 Xây dựng tập dữ liệu huấn luyện

Tập dữ liệu huấn luyện được tạo bằng cách:

- Lấy mẫu ngẫu nhiên từ raster đặc trưng và raster nhãn.
- Cân bằng dữ liệu theo tỷ lệ 50% ngập – 50% không ngập.
- Tổng số mẫu sử dụng: 40,000 pixel.

Chiến lược này giúp mô hình học tốt ranh giới phân loại mà không bị thiên lệch do mất cân bằng lớp.

# 6 Mô hình học máy và chiến lược huấn luyện

## 6.1 Lựa chọn mô hình

XGBoost Classifier được lựa chọn do:

- Khả năng mô hình hóa quan hệ phi tuyến.
- Hoạt động tốt với dữ liệu hỗn hợp (liên tục + phân loại).
- Tương thích tốt với các phương pháp giải thích như SHAP.

Bài toán được thiết lập là phân loại nhị phân, nhưng đầu ra của mô hình là xác suất, được diễn giải như **Flood Susceptibility Index**.

## 6.2 Hiệu chỉnh xác suất

Để đảm bảo xác suất đầu ra có ý nghĩa thực, mô hình được hiệu chỉnh bằng phương pháp calibration, giúp cải thiện độ tin cậy của giá trị xác suất.

# 7 Đánh giá mô hình

## 7.1 Đánh giá thống kê

Mô hình đạt được các chỉ số sau trên tập test:

- Accuracy 0.92
- ROC-AUC 0.97

- PR-AUC  0.96
- Brier Score  0.06

Các chỉ số này cho thấy mô hình có khả năng phân biệt rất tốt và xác suất đầu ra được hiệu chỉnh hợp lý.

### 7.2 Giải thích mô hình bằng SHAP

Phân tích SHAP cho thấy các biến quan trọng nhất bao gồm:

- Slope
- Relief_2km
- Dist_Water
- Precipitation

Chiều tác động của các biến hoàn toàn phù hợp với quy luật vật lý: vùng thấp, dốc nhỏ và gần sông có nguy cơ ngập cao hơn.
*(Đính kèm: SHAP Bar Plot và SHAP Beeswarm)*

## 8 Dự báo không gian và kiểm định kết quả

Mô hình được áp dụng cho toàn bộ raster Hà Tĩnh để tạo bản đồ nguy cơ lũ liên tục. Phân phối giá trị đầu ra cho thấy:

- Khoảng 72% diện tích có nguy cơ thấp ($<0.2$).
- Khoảng 17% diện tích nằm trong nhóm nguy cơ rất cao ($>0.8$).

Kết quả kiểm định theo đơn vị hành chính cấp huyện cho thấy các huyện đồng bằng ven sông có tỷ lệ diện tích nguy cơ cao lớn hơn đáng kể so với các huyện miền núi, phù hợp với thực tế lịch sử lũ lụt.
*(Đính kèm: Bản đồ nguy cơ lũ và biểu đồ so sánh theo huyện)*

## 9 Nhận xét tổng hợp

Toàn bộ pipeline từ xây dựng baseline, thiết kế biến đầu vào, huấn luyện mô hình, giải thích và kiểm định đều cho kết quả nhất quán, không xuất hiện dấu hiệu học máy phi vật lý hoặc overfitting. Các kết quả đạt được đủ độ tin cậy để sử dụng làm nền tảng cho các phân tích nâng cao trong các nghiên cứu tiếp theo.