VNU Journal of Science: Earth and Environmental Sciences
Journal homepage: https://js.vnu.edu.vn/EES

Original Article

Optimizing Machine Learning Models for Event-conditioned Flood Susceptibility Mapping in Binh Dinh Region

Hoang Tich Phuc[1], Nguyen Thi Thanh Thao[2], Ha Minh Cuong[1], Vu Xuan Quan[1], Vu Phuong Lan[3,*], Vu Van Tich[4]

[1]*VNU University of Engineering and Technology, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam*
[2]*True Technology Co., Ltd, 117 Tran Duy Hung, Yen Hoa, Hanoi, Vietnam*
[3]*VNU University of Science, 334 Nguyen Trai, Thanh Xuan, Hanoi, Vietnam*
[4]*Vietnam Insitute of Science and Technology Strategy, Ministry of Science and Technology, 115 Tran Duy Hung, Yen Hoa, Hanoi, Vietnam*

Received 13th Januray 2026
Revised 27th March 2026; Accepted 8th April 2026

Abstract: Flooding is a major natural hazard in Vietnam, causing significant human and economic losses, particularly in Bình Định (now the eastern part of Gia Lai Province), where complex terrain and dense river networks increase vulnerability. This study integrates the Puma Optimization (PO) meta-heuristic algorithm with Random Forest (RF), Extreme Gradient Boosting (XGB), and Support Vector Machine (SVM) models to improve hyperparameter tuning, addressing limitations of conventional methods such as Random Search (RS) and Particle Swarm Optimization (PSO). The framework was implemented on Google Earth Engine using 16 conditioning factors and 1,978 flood/non-flood samples from the 06–10/2022 flood events. Results show that the PO–XGB model achieved the highest performance (R^2 = 0.84), outperforming PSO–XGB (0.81) and RS–XGB (0.55), with low errors (MAE = 0.07; RMSE = 0.11). The results demonstrate that the PO algorithm achieves stable convergence and high accuracy in capturing the complex nonlinear relationships inherent in flood processes. The integration of next-generation optimization algorithms shows strong potential to significantly enhance the reliability of flood susceptibility maps under the 2022 wet-season context, while effectively overcoming the limitations of standalone approaches. This study not only contributes methodologically to the application of artificial intelligence in hydrology but also provides a scientific basis and decision-support tools for sustainable spatial planning and disaster risk reduction at the local scale.

Keywords: Event-conditioned Flood susceptibility; Puma Optimization; Machine Learning; Remote Sensing; Binh Dinh, Gia Lai province.

* Corresponding author.
E-mail address: vuphuonglan@hus.edu.vn

https://doi.org/10.25073/2588-1094/vnuees.5480

Tối ưu hóa mô hình học máy trong thành lập bản đồ nguy cơ lũ lụt theo sự kiện tại khu vực Bình Định

Hoàng Tích Phúc[1], Nguyễn Thị Thanh Thảo[2], Hà Minh Cường[1], Vũ Xuân Quân[1], Vũ Phương Lan[3,*], Vũ Văn Tích[4]

[1]*Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội, 144 Xuân Thủy, Cầu Giấy, Hà Nội, Việt Nam*
[2]*Công ty Trách nhiệm hữu hạn Công nghệ Niềm tin, 117 Trần Duy Hưng, Yên Hòa, Hà Nội, Việt Nam*
[3]*Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội, 334 Nguyễn Trãi, Thanh Xuân, Hà Nội, Việt Nam*
[4]*Học viện Chiến lược Khoa học và Công nghệ, Bộ Khoa học và Công nghệ, 115 Trần Duy Hưng, Yên Hòa, Hà Nội, Việt Nam*

Nhận ngày 13 tháng 01 năm 2026
Chỉnh sửa ngày 27 tháng 3 năm 2026; Chấp nhận đăng ngày 8 tháng 4 năm 2026

Tóm tắt: Lũ lụt là một hiểm họa thiên nhiên lớn ở Việt Nam, gây thiệt hại đáng kể về người và kinh tế, đặc biệt là ở Bình Định (nay là phần phía đông của tỉnh Gia Lai), nơi địa hình phức tạp và mạng lưới sông ngòi dày đặc làm tăng tính dễ bị tổn thương. Nghiên cứu này đề xuất tích hợp thuật toán tối ưu hóa Puma (PO), một thuật toán meta-heuristic, với các mô hình RF, XGB và SVM để giải quyết bài toán tối ưu hóa siêu tham số, khắc phục những hạn chế của các phương pháp thông thường như Tìm kiếm ngẫu nhiên (RS) hay bầy đàn hạt (PSO). Khung nghiên cứu được triển khai trên Google Earth Engine sử dụng 16 yếu tố điều kiện và 1.978 mẫu lũ/không lũ từ các sự kiện lũ lụt ngày 06–10/2022. Kết quả cho thấy mô hình PO–XGB đạt hiệu suất cao nhất (R^2 = 0,84), vượt trội hơn PSO–XGB (0,81) và RS–XGB (0,55), với sai số thấp (MAE = 0,07; RMSE = 0,11). Kết quả nghiên cứu chứng minh thuật toán PO có khả năng hội tụ ổn định và đạt độ chính xác cao trong việc nắm bắt các mối quan hệ phi tuyến phức tạp của quá trình lũ lụt. Việc tích hợp các thuật toán tối ưu hóa thế hệ mới cho thấy tiềm năng nâng cao đáng kể độ tin cậy của bản đồ nhạy cảm lũ lụt theo bối cảnh mùa mưa năm 2022, đồng thời khắc phục hiệu quả những hạn chế của các phương pháp đơn lẻ. Nghiên cứu không chỉ đóng góp về phương pháp luận trong ứng dụng trí tuệ nhân tạo cho lĩnh vực thủy văn, mà còn cung cấp cơ sở khoa học và công cụ hỗ trợ ra quyết định, góp phần định hướng quy hoạch không gian bền vững và giảm thiểu rủi ro thiên tai ở cấp địa phương.

Từ khóa: Mức độ nhạy cảm lũ lụt theo sự kiện; Tối ưu hóa Puma; Học máy; Viễn Thám; Bình Định, tỉnh Gia Lai.

1. Mở đầu

Lũ lụt là một trong những loại hình thiên tai cực đoan phổ biến và có sức tàn phá nghiêm trọng nhất trên quy mô toàn cầu, gây tổn thất lớn về người, tài sản và làm gián đoạn các hoạt động kinh tế – xã hội [1-4]. Dưới tác động của biến đổi khí hậu, quá trình đô thị hóa nhanh và sự suy giảm khả năng điều tiết tự nhiên của lưu vực, tần suất cũng như cường độ các trận lũ lớn có xu

* Tác giả liên hệ.
Địa chỉ email: vuphuonglan@hus.edu.vn

https://doi.org/10.25073/2588-1094/vnuees.5480

hướng gia tăng rõ rệt trong những thập kỷ gần đây [5, 6]. Tại Việt Nam, quốc gia nằm trong khu vực nhiệt đới gió mùa với hệ thống sông ngòi dày đặc, lũ lụt chiếm tỷ trọng lớn trong tổng số các loại hình thiên tai xảy ra trong hơn hai thập kỷ qua và được xem là mối đe dọa thường trực đối với phát triển bền vững [3-6]. Trong đó, tỉnh Bình Định (phía đông tỉnh Gia Lai mới) thường xuyên phải đối mặt với các đợt mưa lũ dồn dập do có vị trí địa lý đặc thù chuyển tiếp giữa cao nguyên bazan và duyên hải miền Trung. Đặc biệt, sự tương phản địa hình mạnh mẽ cùng mạng lưới thủy văn phức tạp của các hệ thống sông Ba, sông Kôn làm gia tăng áp lực lên hạ du, gây thiệt hại nặng nề về người và tài sản, điển hình như đợt thiên tai tháng 11/2025 đã khiến nhiều người thương vong và hàng trăm nghìn ngôi nhà bị ngập nước [7]. Trước xu thế gia tăng của các hiện tượng thời tiết cực đoan, việc xây dựng bản đồ mức độ nhạy cảm lũ lụt có độ chính xác cao được xem là yêu cầu cấp thiết nhằm hỗ trợ hiệu quả cho công tác quản lý rủi ro thiên tai, quy hoạch không gian và bố trí dân cư bền vững.

Sự phát triển của công nghệ viễn thám và hệ thống thông tin địa lý (GIS) đã mở ra các hướng tiếp cận mới, cho phép giám sát và phân tích lũ lụt trên quy mô lớn với chi phí và thời gian hợp lý. Trong những năm gần đây, nhiều nghiên cứu đã chuyển dịch từ các mô hình thủy văn – thủy lực truyền thống (như MIKE 11, HEC-RAS) [2, 8-10]; vốn đòi hỏi dữ liệu đầu vào chi tiết và quá trình hiệu chỉnh phức tạp, sang việc ứng dụng các thuật toán học máy (ML) [1, 11-13]. Các mô hình phổ biến như Random Forest (RF), Extreme Gradient Boosting (XGB) và Support Vector Machine (SVM) đã chứng minh hiệu quả trong việc khai thác các tập dữ liệu địa không gian đa nguồn và mô hình hóa các quan hệ phi tuyến giữa các yếu tố địa hình, lượng mưa và lớp phủ bề mặt [14,[15]. Tuy nhiên, một thách thức lớn phát sinh là hiệu suất của các mô hình này phụ thuộc rất lớn vào việc tối ưu hóa siêu tham số (HPO). Các siêu tham số này đóng vai trò kiểm soát hành vi của thuật toán và ảnh hưởng đến khả năng tổng quát hóa của mô hình trên dữ liệu mới.

Trên thực tế, việc tinh chỉnh siêu tham số thủ công hoặc sử dụng các phương pháp tìm kiếm truyền thống như Tìm kiếm ngẫu nhiên (RS) thường tiêu tốn tài nguyên tính toán và không đảm bảo tìm được nghiệm tối ưu toàn cục. Ngay cả các thuật toán tối ưu hóa xấp xỉ phổ biến như Tối ưu bầy đàn hạt (PSO) cũng có xu hướng mắc kẹt tại các cực trị cục bộ khi xử lý các không gian tham số phức tạp. Theo Wolpert và cộng sự (1997), không có thuật toán tối ưu nào hoàn hảo cho mọi bài toán, nên việc tìm kiếm và ứng dụng các thuật toán thế hệ mới là vô cùng cần thiết để đạt được hiệu suất tối đa trong các bài toán thực tế phức tạp [16].

Xuất phát từ những vấn đề trên, nghiên cứu này đề xuất tích hợp thuật toán tối ưu hóa Puma (PO) nhằm giải quyết bài toán tối ưu siêu tham số cho các mô hình học máy trong thành lập bản đồ mức độ nhạy cảm lũ lụt (Flood Susceptibility Mapping – FSM). PO là một thuật toán meta-heuristic mới, sở hữu cơ chế chuyển pha thích nghi giúp cân bằng hiệu quả giữa quá trình khám phá và khai thác không gian tham số, từ đó cải thiện khả năng hội tụ và độ ổn định của mô hình. Nghiên cứu tập trung vào các mục tiêu chính: i) Xây dựng bộ dữ liệu địa không gian đa nguồn đồng bộ với 16 lớp thông tin; ii) Ứng dụng và so sánh hiệu quả của PO với RS và PSO trên nền tảng các mô hình RF, XGB, SVM; và iii) Thành lập bản đồ mức độ nhạy cảm lũ lụt chi tiết cho khu vực tỉnh Bình Định. Trong đó, vì biến mưa được lấy theo bối cảnh năm 2022 và nhãn ngập được đối chiếu theo cửa sổ thời gian sự kiện, kết quả được diễn giải là bản đồ nhạy cảm lũ lụt có điều kiện theo bối cảnh năm 2022, không phải lớp nhạy cảm nền dài hạn.

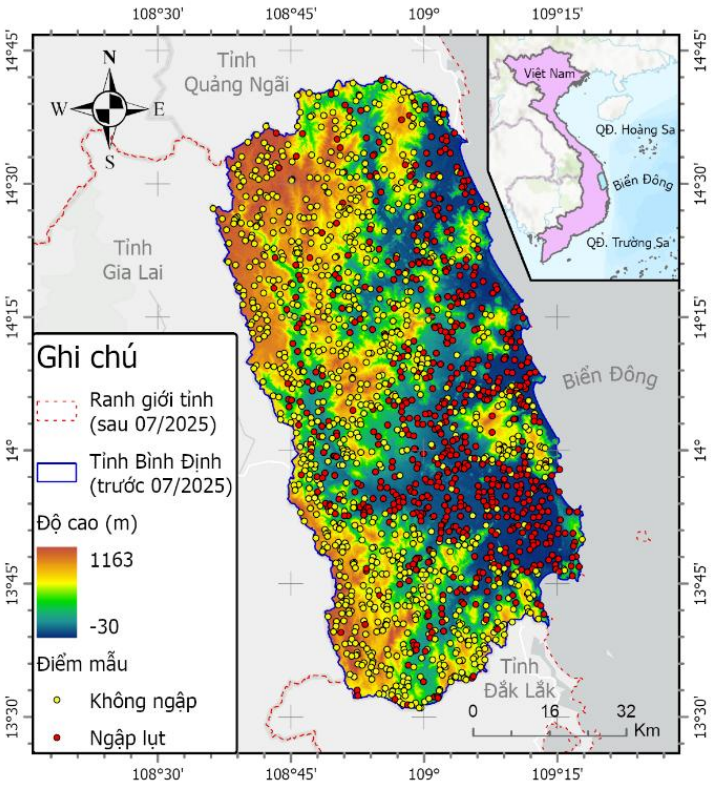
2. Khu vực nghiên cứu và bộ dữ liệu

2.1. Khu vực nghiên cứu

Khu vực nghiên cứu được lựa chọn là tỉnh Bình Định cũ, nay thuộc tỉnh Gia Lai, với tổng diện tích tự nhiên khoảng 6.000 km^2 (Hình 1). Trong nghiên cứu này cách gọi 'Bình Định' được sử dụng theo đơn vị hành chính tại thời điểm thu thập dữ liệu (năm 2022) để đảm bảo nhất quán với các lớp dữ liệu nguồn. Đây là một khu vực có vị trí địa lý đặc thù khi nằm trên vùng

chuyển tiếp giữa cao nguyên bazan Tây Nguyên và dải duyên hải Nam Trung Bộ. Điểm nổi bật ở đây là sự tương phản mạnh mẽ về địa hình: phía Tây chủ yếu là các cao nguyên rộng lớn với độ cao phổ biến từ 600–800 m, được bao bọc bởi các dãy núi cao trên 1.500 m; trong khi phía Đông là dải đồng bằng hẹp bị chia cắt mạnh bởi các sườn núi dốc. Sự chuyển tiếp đột ngột từ địa hình cao nguyên dốc thoải sang các thung lũng dốc đứng là yếu tố then chốt làm gia tăng tốc độ dòng chảy, tạo điều kiện hình thành các trận lũ quét và ngập lụt nghiêm trọng tại vùng hạ lưu.

Hệ thống thủy văn của khu vực nghiên cứu được chi phối bởi lưu vực sông Ba với vai trò là trục thoát nước chủ đạo, dẫn nước từ vùng thượng nguồn cao nguyên đổ dồn nhanh xuống vùng đồng bằng hạ lưu. Do mạng lưới sông ngòi có độ dốc lớn và nhiều nhánh nhỏ tập trung, lưu lượng nước thường tăng đột ngột trong mùa mưa, gây áp lực trực tiếp lên các khu vực dân cư đông đúc ven sông.



Hình 1. Khu vực nghiên cứu Bình Định (thuộc phía đông tỉnh Gia Lai mới).

Khu vực nghiên cứu chịu ảnh hưởng rõ rệt của khí hậu nhiệt đới gió mùa, với mùa mưa tập trung chủ yếu trong khoảng từ tháng 9 đến tháng 12. Do địa hình đón gió Đông Bắc, lượng mưa tại Bình Định rất lớn và phân bố không đều: vùng cao nguyên đạt từ 2.200–2.500 mm, trong khi vùng đồng bằng phía Đông có thể vượt mức 3.000 mm/năm [17]. Đáng chú ý, có đến 70–80% tổng lượng mưa năm tập trung vào mùa lũ, kết hợp với tình trạng biến đổi lớp phủ do hoạt động kinh tế (khai thác gỗ, xây dựng thủy điện) đã làm suy giảm khả năng giữ nước tự nhiên, khiến rủi ro lũ lụt diễn biến càng phức tạp và khó lường.

2.2. Bộ dữ liệu

Bộ dữ liệu nghiên cứu được xây dựng từ đa nguồn với tổng cộng 16 lớp thông tin địa không gian đầu vào, đảm bảo tính đồng bộ về không thời gian nhằm phục vụ bài toán đánh giá mức

độ nhạy cảm lũ lụt (Bảng 1). Dữ liệu nhãn của mô hình bao gồm 1978 điểm ghi nhận sự kiện ngập và không ngập, được trích xuất từ ảnh vệ tinh Landsat-9 thông qua phân tích chỉ số nước khác biệt chuẩn hóa hiệu chỉnh (MNDWI). Các biến độc lập được phân thành bốn nhóm đặc trưng chính: i) Nhóm địa hình trích xuất từ Copernicus DEM (bao gồm độ cao, độ dốc, hướng dốc, hướng dòng chảy, chỉ số ẩm ướt địa hình TWI và độ cong); ii) Nhóm chỉ số bề mặt từ Sentinel-2 và ESA (NDVI, NDBI, NDWI và lớp phủ LULC); iii) Nhóm khí tượng ghi nhận dữ liệu lượng mưa đợt lũ lịch sử năm 2022 từ Open-Meteo; và iv) Nhóm hạ tầng - thủy văn từ OpenStreetMap (khoảng cách và mật độ sông suối, đường giao thông). Biến lượng mưa được tổng hợp theo cả năm 2022, trong khi điểm mẫu ngập được thu nhận/đối chiếu theo giai đoạn mưa lũ từ tháng 6 đến tháng 10 nhằm kiểm chứng theo bối cảnh chuỗi sự kiện.

Trong đó, dữ liệu mẫu ngập/không ngập được xây dựng từ nhiều cảnh ảnh vệ tinh trong giai đoạn 06–10/2022 nhằm phản ánh đợt mưa–lũ kéo dài đến tháng 10/2022. Vùng ngập được xác định dựa trên chỉ số MNDWI (và/hoặc chênh lệch MNDWI trước–sau sự kiện) với ngưỡng phát hiện 0,05 để tạo mặt nạ vùng ngập phục vụ lấy mẫu. Tập huấn luyện có 1978 điểm mẫu gồm: 778 điểm ngập và 1200 điểm không ngập. Trong đó điểm không ngập được lấy trong vùng an toàn ngoài vùng ngập lụt, đồng thời ràng buộc cách ranh giới ngập tối thiểu 30 m để giảm nhiễu nhãn tại vùng biên. Ngoài ra, để tăng độ 'sạch' của lớp không ngập, chúng tôi loại trừ các khu vực có độ cao >250 m và độ dốc >30° khỏi không gian lấy mẫu do ít có khả năng tích nước/ngập trong bối cảnh sự kiện.

Bảng 1. Bộ dữ liệu và đặc trưng địa hình–khí tượng được sử dụng trong nghiên cứu

TT	Dữ liệu	Nguồn	ĐPG* thời gian	ĐPG* không gian
1	Điểm lũ	Landsat	06–10/2022	30 m
2	Độ cao	Copernicus	-	30 m
3	Độ dốc	Dẫn xuất từ Copernicus DEM		
4	TWI			
5	Hướng dòng chảy			
6	Độ cong địa hình			
7	Hướng dốc			
8	NDVI		2022	10 m
9	NDBI			
10	NDWI			
11	Lượng mưa	OpenMeteo	2022	9 km
12	LULC	ESA	2022	10 m
13	Khoảng cách tới sông	Open StreetMap	-	30 m
14	Mật độ sông			
15	Khoảng cách tới đường			
16	Mật độ đường			

* ĐPG: độ phân giải

Trước khi trích xuất giá trị tại các điểm mẫu, toàn bộ các lớp raster được chuẩn hóa về cùng hệ quy chiếu và cùng lưới không gian. Trong nghiên cứu, lưới mục tiêu được chọn là 30 m (phù hợp với DEM và nhãn ngập từ Landsat), các lớp 10 m (NDVI/NDBI/NDWI, LULC 10 m) được resample về 30 m; với biến liên tục sử dụng bilinear, trong khi biến phân loại (LULC) sử dụng nearest-neighbor để tránh tạo lớp giả. Biến lượng mưa có độ phân giải thô (~9 km) được đưa về cùng lưới để phục vụ trích xuất tại điểm; giá trị mưa vì vậy được hiểu là đại diện theo ô khí tượng trong năm 2022.

2.3. Dữ liệu thực địa

Nhằm kiểm chứng khả năng tổng quát hóa và dự đoán của mô hình, nghiên cứu đã thiết lập một

bộ dữ liệu kiểm định thực địa độc lập tại khu vực trọng điểm đầm Thị Nại. Khác với bộ dữ liệu huấn luyện thông thường, đây là biến phụ thuộc độc lập phản ánh độ sâu ngập thực tế trong trận lũ lịch sử năm 2020, hoàn toàn không tham gia vào quá trình học của mô hình nhằm đảm bảo tính khách quan trong đánh giá. Quy trình xây dựng bộ dữ liệu này được thực hiện nghiêm ngặt thông qua việc kết hợp khảo sát thực địa bằng thiết bị GPS/GNSS để xác định tọa độ và độ sâu thực đo, đồng thời tiến hành điều tra phỏng vấn nhân chứng tại địa phương và tổng hợp số liệu quan trắc từ các trạm thủy văn. Bộ dữ liệu thực địa được dùng để kiểm chứng/đối chiếu tính hợp lý không gian của bản đồ nhạy cảm; các chỉ số R^2, RMSE và MAE trong nghiên cứu được tính trên tập điểm ngập lụt/không ngập lụt (0/1) theo đầu ra điểm số liên tục của mô hình. Việc sử dụng bộ dữ liệu thực địa tách biệt này không chỉ giúp xác định độ chính xác của bản đồ nhạy cảm lũ lụt mà còn khẳng định độ tin cậy của mô hình khi triển khai tại các khu vực địa lý mới.

3. Phương pháp nghiên cứu

Quy trình tổng thể được cấu trúc thành hai giai đoạn xử lý (Hình 2), tương ứng với hai khối chức năng chính trong thiết kế hệ thống: i) Giai đoạn thu thập và xử lý dữ liệu; và ii) Giai đoạn huấn luyện và tối ưu mô hình học máy.

3.1. Các mô hình học máy

Các mô hình học máy được triển khai dựa trên nguyên lý sử dụng dữ liệu quan sát để tự điều chỉnh các tham số của một hệ thống thích nghi, từ đó học được các quan hệ tiềm ẩn để đưa ra dự báo cho dữ liệu mới. Nghiên cứu tập trung vào ba thuật toán học máy cơ sở có khả năng xử lý quan hệ phi tuyến phức tạp trong bài toán thành lập bản đồ mức độ nhạy cảm lũ lụt.

Rừng ngẫu nhiên (RF) là mô hình học tổ hợp kết hợp nhiều cây quyết định [18]. Về phương pháp, mỗi cây được xây dựng trên một tập con dữ liệu và tập con đặc trưng ngẫu nhiên nhằm cải thiện độ chính xác và giảm thiểu hiện tượng quá khớp. Kết quả cuối cùng là sự tổng hợp thông qua cơ chế bỏ phiếu đa số hoặc lấy giá trị trung bình. RF nổi bật với khả năng chống nhiễu và đánh giá được tầm quan trọng của các yếu tố đầu vào.

Tăng cường độ dốc cực đại (XGB) là mô hình xây dựng các cây quyết định theo cách tuần tự để sửa lỗi cho các cây trước đó [19]. XGB sử dụng kỹ thuật giảm dần độ dốc để tối ưu hóa hàm mất mát, kết hợp với các tham số điều chuẩn để ngăn chặn quá khớp. XGB có ưu thế về tốc độ tính toán và khả năng xử lý dữ liệu lớn, nhưng rất nhạy cảm với việc tinh chỉnh siêu tham số.

Máy Vector hỗ trợ (SVM) là mô hình tập trung vào việc xác định một siêu mặt phẳng tối ưu để phân tách các lớp dữ liệu với biên lớn nhất [20]. Đối với dữ liệu lũ lụt không tuyến tính, SVM sử dụng các hàm nhân như RBF để ánh xạ dữ liệu lên không gian cao chiều hơn nhằm tìm ra ranh giới phân tách. Hiệu suất của SVM phụ thuộc cực kỳ mạnh vào các siêu tham số như C, gamma và loại kernel.

3.2. Thuật toán tối ưu hóa Puma

Puma Optimizer (PO) là một thuật toán meta-heuristic thế hệ mới (2024), mô phỏng chiến lược săn mồi và hành vi di cư của loài báo Puma trong tự nhiên [21]. Trong nghiên cứu này, PO được ứng dụng để giải quyết bài toán tối ưu hóa siêu tham số bằng cách tìm kiếm vectơ siêu tham số tối ưu nhằm tối đa hóa hàm mục tiêu là hệ số xác định. PO vận hành dựa trên việc cân bằng hai giai đoạn: Pha khám phá và Pha khai thác.

Tại Pha khám phá, mô hình tập trung di chuyển ngẫu nhiên để tìm kiếm các vùng tiềm năng mới trong không gian tham số, giúp mô hình tránh bẫy cực trị cục bộ. Công thức sinh nghiệm mới trong pha này được xác định bởi phương trình (1) và (2).

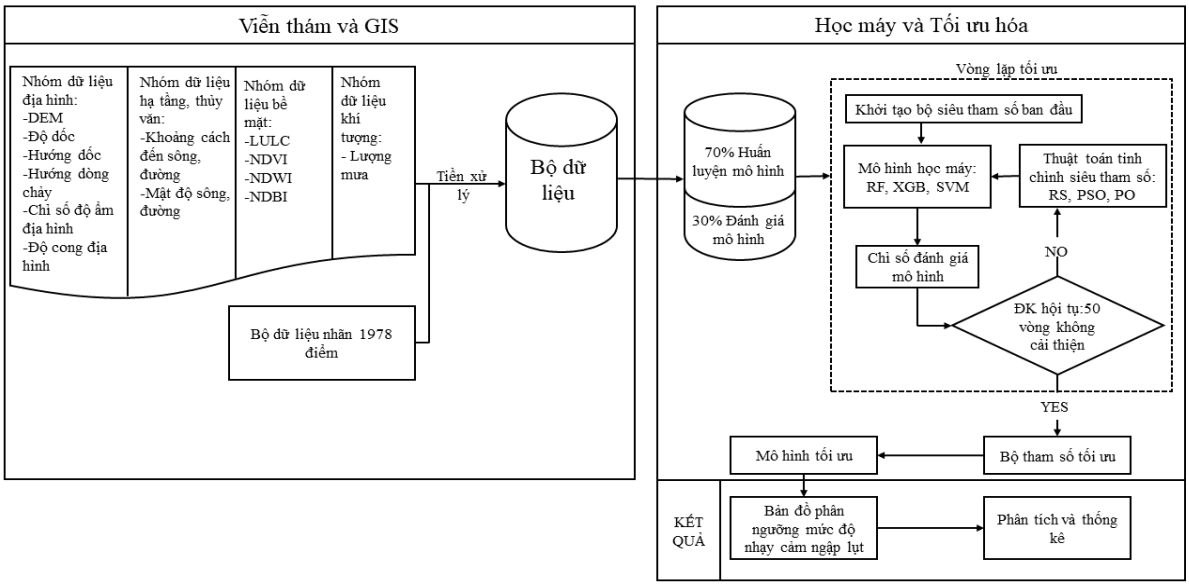
$$\begin{aligned} &\text{Nếu } rd_1 > 0{,}5: \\ &\lambda_{new} = R_{Dim} * (Ub - Lb) + Lb \\ &\text{Nếu } rd_1 \le 0{,}5: \\ &\lambda_{new} = X_a + G * (X_a - X_b) + G \\ &\qquad * \Big(\big((X_a - X_b) - (X_c - X_d) \big) \\ &\qquad + \big((X_c - X_d) - (X_e - X_f) \big) \Big) \end{aligned} \tag{1}$$

$$G = 2.rd_2 - 1 \tag{2}$$

Tiếp đó, pha khai thác sử dụng nghiệm tốt nhất hiện tại làm điểm dẫn đường để tinh chỉnh các siêu tham số xung quanh vùng tối ưu. Một trong những chiến lược khai thác chính được biểu diễn qua công thức (3), với N là kích thước quần thể, là các số ngẫu nhiên trong khoảng.

$$\lambda_{new} = \lambda^* + 2(rd_7).\exp(rdn_1).[round(1 + (N-1).rd_{10})] - \lambda_i \quad (3)$$

Điểm đặc biệt của PO là cơ chế chuyển pha thông minh. Trong 3 vòng lặp đầu (giai đoạn chưa có kinh nghiệm), thuật toán thực hiện đồng thời cả hai pha. Từ vòng lặp thứ 4, PO đánh giá hiệu suất lịch sử thông qua các hàm chi phí để tự động lựa chọn pha tối ưu, giúp tối ưu hóa thời gian tính toán và độ chính xác.



Hình 2. Khái quát quy trình nghiên cứu.

3.3. Quy trình tích hợp và đánh giá

Để đánh giá thuật toán Puma, nghiên cứu tiến hành so sánh thực nghiệm với hai phương pháp tối ưu hóa phổ biến trong các nghiên cứu thủy văn trước đây bao gồm Tối ưu hóa bầy đàn hạt và Tìm kiếm ngẫu nhiên. Tối ưu hóa bầy đàn hạt (PSO) là thuật toán trí tuệ bầy đàn mô phỏng hành vi đàn chim, trong đó các "hạt" di chuyển dựa trên kinh nghiệm cá nhân và kinh nghiệm của toàn bầy [22]. Mặc dù PSO có khả năng hội tụ tốt, nhưng nghiên cứu cho thấy nó dễ mắc kẹt tại cực trị cục bộ khi không gian tham số phức tạp.

Tìm kiếm ngẫu nhiên (RS) là phương pháp thử nghiệm các bộ tham số được chọn ngẫu nhiên trong không gian cấu hình. RS có ưu điểm về sự đơn giản nhưng hiệu suất thường thấp và không ổn định do thiếu cơ chế học tập từ các vòng lặp trước.

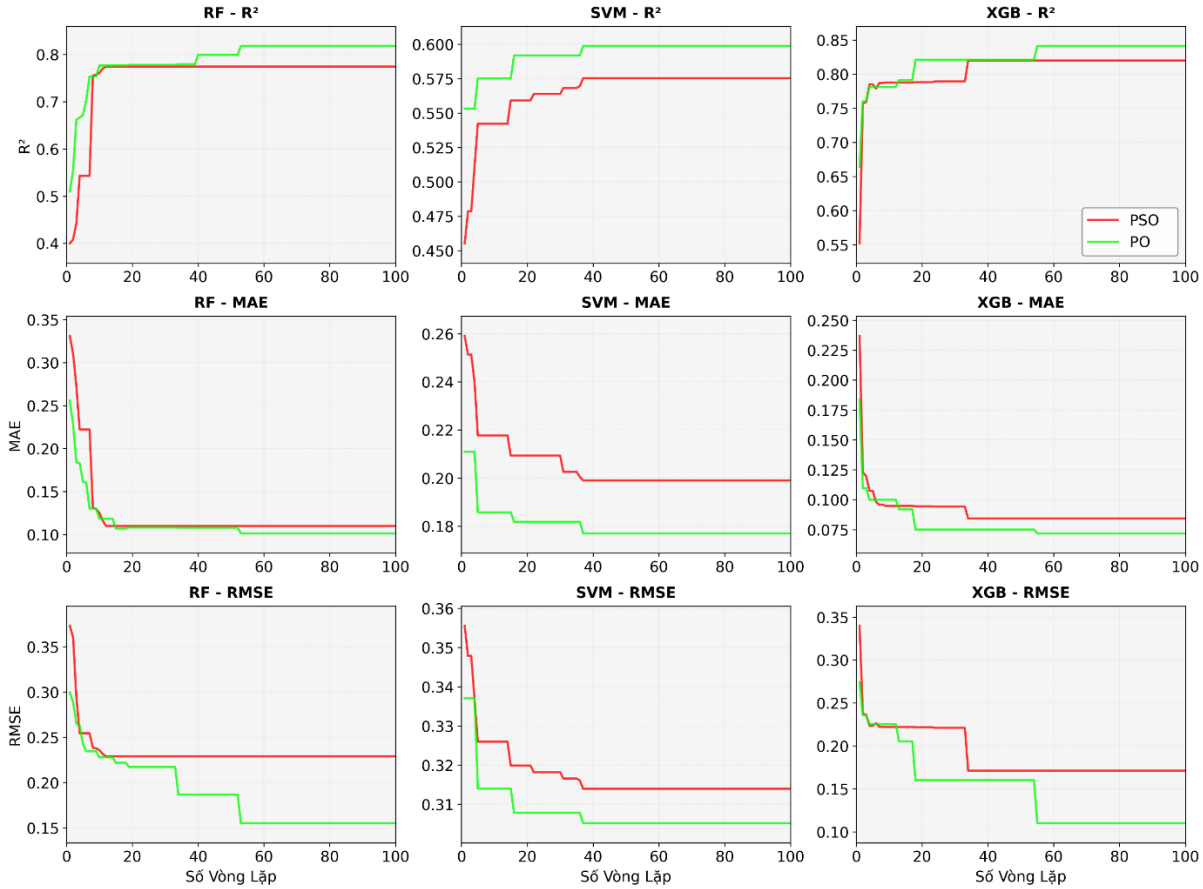
Trong nghiên cứu này, mô hình dự đoán một điểm số nhạy cảm liên tục trong khoảng 0–1 (soft classification/probability-like score) từ nhãn ngập lụt/không ngập lụt (0/1). Vì vậy các chỉ số MAE, RMSE và R^2 được dùng để đo mức độ sai lệch/độ khớp giữa điểm số dự báo và nhãn 0/1 trên tập kiểm tra. Trong quá trình huấn luyện, kỹ thuật xác thực chéo k lần (k=5) cũng được áp dụng nhằm tăng độ ổn định và giảm phụ thuộc vào một lần chia dữ liệu, các chỉ số được lấy trung bình trong các lần lặp.

4. Kết quả thực nghiệm

4.1. So sánh hiệu suất mô hình

Để đánh giá hành vi hội tụ của các thuật toán tối ưu siêu tham số, chúng tôi theo dõi sự thay

đổi của các chỉ số R^2, MAE và RMSE theo 100 vòng lặp đối với hai thuật toán PSO và PO. Kết quả đường hội tụ (Hình 3) cho thấy PO cho hiệu quả tối ưu tốt hơn PSO một cách nhất quán trên cả ba mô hình (RF, SVM và XGB): PO đạt R^2 cao hơn và MAE/RMSE thấp hơn, đồng thời cải thiện mạnh ở giai đoạn đầu và sớm bão hòa khi tăng số vòng lặp. Trong khi đó, do RS lựa chọn tham số ngẫu nhiên ở mỗi lần thử và không có cơ chế hội tụ theo vòng lặp, nghiên cứu sử dụng RS như baseline đối chứng và không trình bày đường hội tụ theo số vòng lặp. Trên cơ sở kết quả tối ưu, các giá trị chỉ số cuối cùng của từng mô hình được tổng hợp trong Bảng 2.



Hình 3. Biểu đồ so sánh hội tụ của thuật toán PO và PSO.

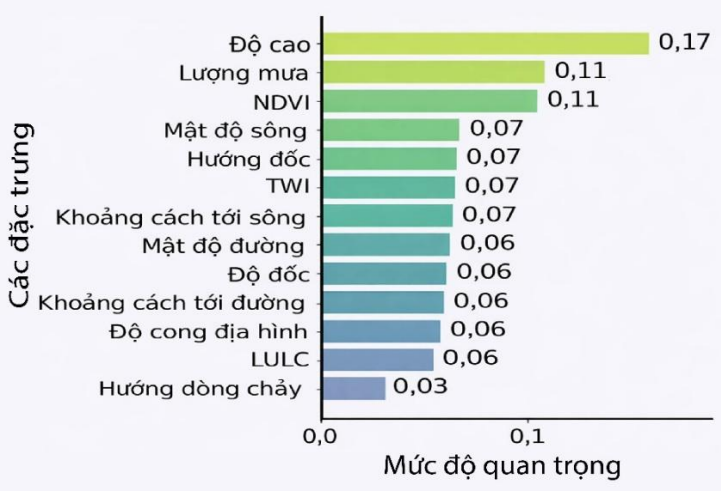
Bảng 2. Hệ số xác định của các mô hình tương ứng với từng thuật toán

Mô hình	RS	PSO	PO
RF	0,73	0,77	0,81
XGB	0,55	0,81	0,84
SVM	0,45	0,59	0,6

Dựa trên hệ số R^2, kết quả tổng hợp từ 9 tổ hợp mô hình cho thấy thuật toán PO luôn đạt hiệu suất cao nhất trên cả ba nền tảng học máy. Cụ thể, mô hình PO-XGB đạt R^2=0,84, vượt trội đáng kể so với phiên bản sử dụng PSO (R^2=0,81) và vượt xa phương pháp RS truyền thống (R^2=0,55). Mức cải thiện này tương đồng với các nghiên cứu trước đây cho thấy việc tích hợp các thuật toán tối ưu hóa siêu tham số tiên tiến có thể nâng cao đáng kể hiệu quả của các mô hình học máy trong các bài toán thủy văn và môi trường phức tạp [23, 24].

Xu hướng cải thiện hiệu suất tương tự cũng được ghi nhận đối với mô hình RF, trong đó PO

giúp nâng hệ số R^2 lên 0,81, so với 0,77 của PSO và 0,73 của RS. Kết quả này phù hợp với nhận định của các nghiên cứu gần đây rằng hiệu quả của RF phụ thuộc mạnh vào quá trình hiệu chỉnh siêu tham số, và các thuật toán tối ưu hóa thế hệ mới thường cho kết quả ổn định hơn so với các phương pháp tìm kiếm ngẫu nhiên hoặc heuristic truyền thống [24]. Đáng chú ý, ngay cả với mô hình SVM – vốn nhạy cảm với nhiễu và cấu trúc không gian phức tạp của dữ liệu đầu vào – thuật toán PO vẫn đạt giá trị R^2 =0,60. Mức cải thiện rõ rệt của PO trong các mô hình nhạy cảm với cấu trúc dữ liệu phức tạp như SVM cho thấy khả năng khai thác siêu tham số hiệu quả hơn, góp phần giảm hiện tượng tối ưu cục bộ và tăng khả năng khái quát hóa của mô hình [25].



Hình 4. Mức độ quan trọng của các biến đầu vào trong mô hình PO-XGB.

Bên cạnh các chỉ số xác định, các chỉ số sai số cũng cho thấy tính ổn định vượt trội của mô hình PO-XGB, với MAE đạt 0,07 và RMSE đạt 0,11. Giá trị RMSE thấp phản ánh khả năng kiểm soát các sai số lớn, cho thấy mô hình không chỉ tái hiện tốt xu thế tổng thể mà còn hạn chế hiệu quả các dự báo cực đoan sai lệch. Điều này đặc biệt quan trọng trong các ứng dụng cảnh báo thiên tai, nơi sai số lớn có thể dẫn đến báo động giả hoặc đánh giá sai mức độ rủi ro.

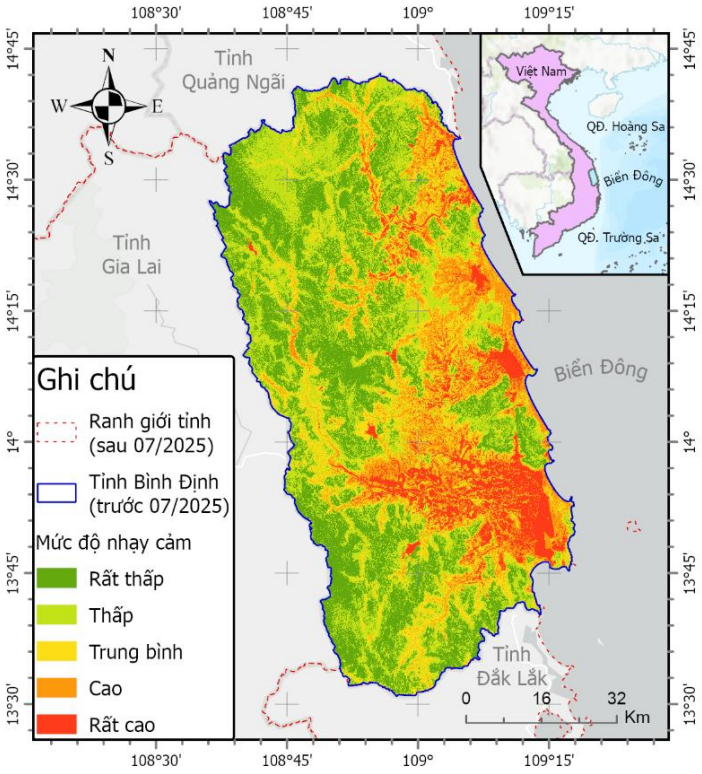
Ưu thế của PO so với PSO và RS có thể giải thích thông qua khả năng tìm kiếm nghiệm tối ưu toàn cục. Trong khi RS thiếu cơ chế học tập và PSO dễ mắc kẹt tại các cực trị cục bộ do hiệu ứng hội tụ đám đông, cơ chế chuyển pha thông minh của PO giúp thuật toán này cân bằng hiệu quả giữa việc khám phá không gian tham số mới và khai thác các vùng nghiệm tiềm năng đã biết. Kết quả này khẳng định rằng việc đầu tư vào các kỹ thuật tối ưu hóa tiên tiến như PO mang lại những cải tiến hiệu suất hữu hình và thiết thực cho các bài toán thủy văn thực tế.

Bên cạnh đó, để tăng tính giải thích của mô hình PO-XGB, chúng tôi đã phân tích mức độ quan trọng của các biến đầu vào. Hình 4 cho thấy nhóm biến địa hình–khí tượng (độ cao, lượng mưa) và thảm phủ (NDVI) chi phối rõ rệt, trong khi các biến liên quan mạng lưới sông và điều kiện tích nước (TWI, khoảng cách/mật độ sông, hướng dốc…) đóng vai trò bổ trợ. Nhìn chung, kết quả này cho thấy việc kết hợp PO giúp tăng độ ổn định tối ưu siêu tham số và cải thiện chất lượng dự đoán của XGB trong bối cảnh sự kiện mưa–lũ năm 2022.

4.2. Thành lập bản đồ nhạy cảm lũ lụt

Sau khi xác định được bộ siêu tham số tối ưu, nghiên cứu đã tiến hành triển khai mô hình trên quy mô của khu vực Bình Định dựa trên nền tảng Google Earth Engine. Bản đồ kết quả được thành lập dưới dạng xác suất thực liên tục từ 0 đến 1 và được phân cấp thành 5 mức độ nguy cơ (Rất thấp, Thấp, Trung bình, Cao, Rất cao) bằng phương pháp Natural Breaks (Jenks) (Hình 5).

Kết quả cho thấy, bản đồ nhạy cảm lũ lụt thành lập từ mô hình PO-XGB thể hiện sự phân hóa rõ rệt theo đặc điểm địa hình và mạng lưới thủy văn của khu vực. Vùng phía Tây với địa hình cao nguyên bazan chủ yếu thuộc mức nhạy cảm rất thấp và thấp, đóng vai trò như vùng đệm tự nhiên nhờ khả năng thấm nước tốt và địa hình dốc thoải. Ngược lại, các vùng nguy cơ cao và rất cao tập trung dày đặc tại dải đồng bằng phía Đông và Đông Nam, đặc biệt là hạ lưu sông Kôn và sông Hà Thanh. Sự chuyển tiếp đột ngột về độ cao từ 600m xuống vùng trũng thấp ven biển làm gia tăng tốc độ dòng chảy, kết hợp với lượng mưa cực đoan đã tạo ra các điểm nóng rủi ro dọc theo các hành lang sông chính.



Hình 5. Bản đồ mức độ nhạy cảm lũ lụt được thành lập dựa trên mô hình PO-XGB.

Đáng chú ý, mô hình PO-XGB không chỉ phản ánh các yếu tố tự nhiên mà còn nhận diện chính xác tác động của hoạt động nhân sinh thông qua chỉ số NDBI cao tại các khu vực đô thị. Các khu vực rủi ro rất cao tập trung chủ yếu tại các khu vực ven sông và trục giao thông chính, nơi bề mặt bê tông hóa làm giảm khả năng thoát nước tự nhiên và tăng dòng chảy bề mặt.

4.3. Kiểm chứng với dữ liệu thực địa

Để đánh giá khách quan khả năng và tính thực tiễn của mô hình, nghiên cứu thực hiện thống kê phân bố 41 điểm kiểm chứng độc lập theo các lớp nhạy cảm lũ lụt của bản đồ PO-XGB (Bảng 3). Kết quả cho thấy 36.6% điểm nằm trong lớp Rất cao (15/41), trong khi hai lớp Cao và Trung bình lần lượt chiếm 31,7% (13/41) và 31,7% (13/41). Đáng chú ý, không có điểm kiểm chứng rơi vào các lớp Thấp/Rất thấp, cho thấy các ghi nhận ngập thực địa chủ yếu trùng khớp với những khu vực mà mô hình dự báo có mức nhạy cảm từ Trung bình đến Rất cao.

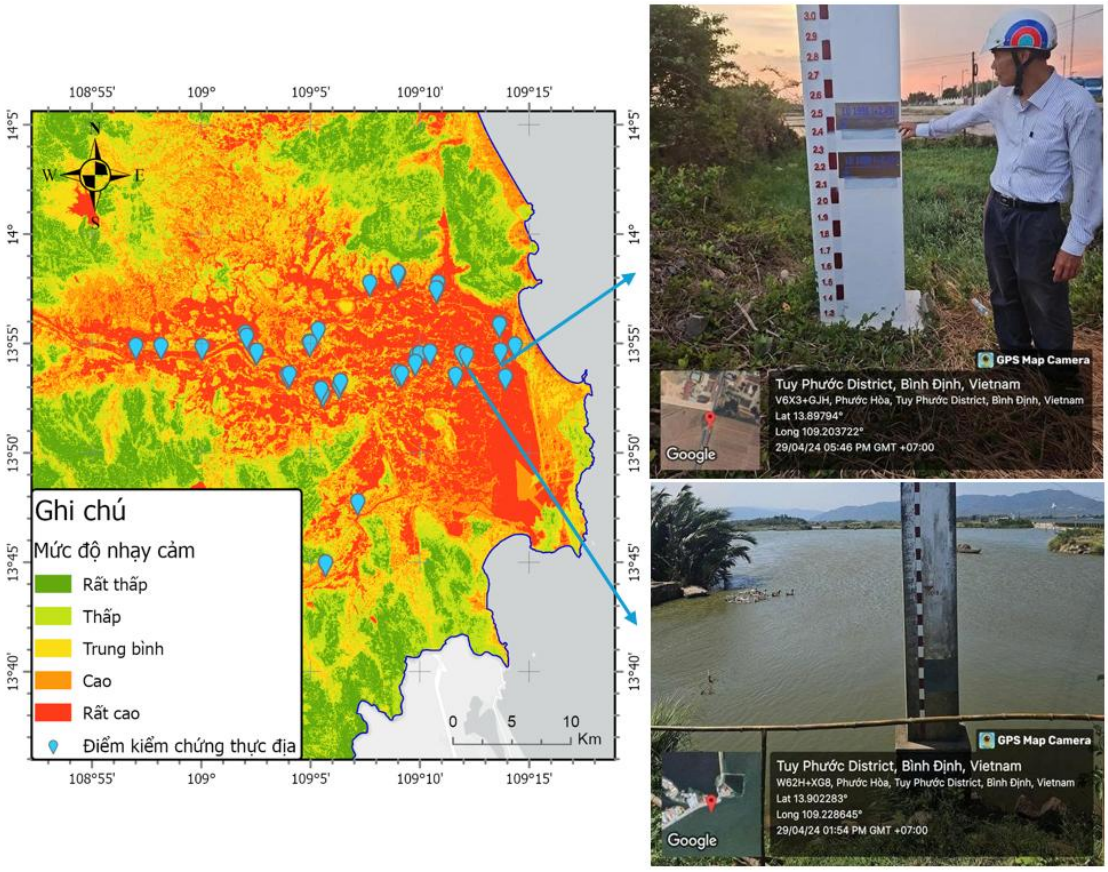
Về phân bố không gian (Hình 6), các điểm kiểm chứng tập trung dọc hành lang sông và vùng trũng thấp ven khu vực đầm và hạ lưu sông.

Phần lớn các điểm lũ quan sát đều tập trung trong các vùng nguy cơ này, thể hiện sự phù hợp rõ rệt giữa kết quả mô hình và hiện trạng thực địa. Đặc biệt, tại các khu vực hạ lưu sông Kôn, sông Hà Thanh và vùng ven đầm Thị Nại, nơi có địa hình trũng thấp và mật độ hội tụ dòng chảy lớn, các điểm thực địa xuất hiện với mật độ dày đặc và trùng khớp rõ ràng với các điểm nóng rủi ro được xác định trên bản đồ. Ngược lại, hầu như không ghi nhận điểm ngập thực tế nào thuộc các khu vực được phân loại ở mức rất thấp hoặc thấp. Điều này cung cấp bằng chứng thuyết phục cho thấy mô hình có độ đặc hiệu cao, qua đó hạn chế hiệu quả hiện tượng cảnh báo sai và nâng cao độ tin cậy của kết quả phân vùng nguy cơ lũ lụt.

Sự tương quan chặt chẽ này không chỉ củng cố độ tin cậy của hệ số xác định R^2=0,84 đã đạt được trong quá trình huấn luyện mà còn khẳng định khả năng tổng quát hóa vượt trội của thuật toán tối ưu hóa Puma (PO) khi áp dụng vào dữ liệu thực tế tại tỉnh Bình Định. Kết quả kiểm chứng thực địa cho thấy bản đồ nhạy cảm lũ lụt được thành lập không chỉ dừng lại ở các con số thống kê lý thuyết mà thực sự là công cụ hỗ trợ ra quyết định đáng tin cậy, giúp các cơ quan quản lý thiên tai khoanh vùng chính xác các khu vực ưu tiên để triển khai các biện pháp ứng phó và sơ tán dân cư trước các kịch bản mưa lũ cực đoan.

Bảng 3. Thống kê điểm kiểm chứng thực địa theo mức nhạy cảm lũ lụt

Mức độ nhạy cảm lũ lụt	Số điểm kiểm chứng	Tỷ lệ
Trung bình	13	31,7%
Cao	13	31,7%
Rất cao	15	36,6%



Hình 6. So sánh các điểm ngập lụt thực địa.

5. Kết luận và đề xuất

Nghiên cứu đã tích hợp thuật toán tối ưu hóa Puma (PO) với ba mô hình học máy phổ biến (RF, XGB, SVM) kết hợp dữ liệu viễn thám và GIS đa nguồn, nhằm xây dựng bản đồ mức độ nhạy cảm lũ lụt cho khu vực Bình Định theo sự kiện mùa mưa năm 2022. Kết quả thực nghiệm khẳng định ưu thế vượt trội của mô hình PO-XGB với hệ số xác định R^2=0,84, cao hơn đáng kể so với các phương pháp PSO (R^2=0,81) và RS (R^2=0,55), đồng thời duy trì các chỉ số sai số ở mức thấp nhất (MAE = 0,07; RMSE = 0,11). Về mặt không gian, bản đồ dự báo phản ánh chính xác quy luật địa lý với các vùng nguy cơ cao tập trung tại hạ lưu phía Đông (lưu vực sông Kôn, sông Hà Thanh) và hoàn toàn tương thích với các dữ liệu điểm lũ thực tế đã ghi nhận. Những kết quả này khẳng định vai trò của thuật toán PO trong việc nâng cao độ chính xác và tính ổn định của các mô hình học máy khi xử lý các quan hệ phi tuyến phức tạp trong bài toán lũ lụt.

Bên cạnh những kết quả đạt được, nghiên cứu vẫn tồn tại một số hạn chế. Việc đánh giá mô hình hiện tại dựa trên phép chia ngẫu nhiên theo điểm có thể chưa phản ánh đầy đủ ảnh hưởng của tự tương quan không gian trong dữ liệu, dẫn đến khả năng đánh giá hiệu suất có xu hướng lạc quan. Do đó, các nghiên cứu tiếp theo cần xem xét áp dụng các chiến lược xác thực chéo theo không gian nhằm đánh giá chặt chẽ hơn khả năng tổng quát hóa của mô hình.

Ngoài ra, phương pháp đề xuất hiện mới tập trung vào việc đánh giá mức độ nhạy cảm lũ lụt và chưa cung cấp thông tin định lượng về độ sâu ngập, vận tốc dòng chảy, cũng như còn phụ thuộc vào dữ liệu viễn thám quang học vốn chịu ảnh hưởng bởi mây mù trong điều kiện thời tiết cực đoan. Để nâng cao giá trị ứng dụng, các hướng nghiên cứu trong tương lai cần tập trung vào việc tích hợp dữ liệu radar (như Sentinel-1) nhằm khắc phục hạn chế về thời tiết, phát triển các chiến lược tối ưu hóa lai ghép, và tiến tới xây dựng hệ thống cảnh báo sớm thời gian thực dựa trên sự kết hợp giữa dữ liệu khí tượng động và các kịch bản biến đổi khí hậu.

Lời cảm ơn

Công trình hoàn thành với sự hỗ trợ đề tài mã số ĐTĐL.CN-93/21 của Bộ Khoa học và Công nghệ về tài chính khảo sát thực địa và thu thập dữ liệu cho nghiên cứu này.

Tài liệu tham khảo

[1] H. Farhadi, H. Ebadi, A. Kiani, A. Asgary, Near Real-Time Flood Monitoring Using Multi-Sensor Optical Imagery and Machine Learning by GEE: An Automatic Feature-Based Multi-Class Classification Approach, Remote Sensing, Vol. 16, No. 23, 2024, p. 4454, https://doi.org/10.3390/rs16234454.

[2] H. Belay, A. M. Melesse, G. Tegegne, S. M. Kassaye, Flood Inundation Mapping Using the Google Earth Engine and HEC-RAS Under Land Use/Land Cover and Climate Changes in the Gumara Watershed, Upper Blue Nile Basin, Ethiopia, Remote Sensing, Vol. 17, No. 7, 2025, pp. 1283, https://doi.org/10.3390/rs17071283.

[3] Y. Hirabayashi et al., Global Flood Risk Under Climate Change, Nature Clim Change, Vol. 3, No. 9, 2013, pp. 816-821, https://doi.org/10.1038/nclimate1911.

[4] S. Blenkinsop, L. M. Alves, A. J. P. Smith, Climate Change Increases Extreme Rainfall and the Chance of Floods, Zenodo, 2021, https://doi.org/10.5281/ZENODO.4779119.

[5] A. Akhyar et al., Deep Artificial Intelligence Applications for Natural Disaster Management Systems: A Methodological Review, Ecological Indicators, Vol. 163, 2024, pp. 112067, https://doi.org/10.1016/j.ecolind.2024.112067.

[6] E. Trégarot et al., Effects of Climate Change on Marine Coastal Ecosystems – A Review to Guide Research and Management, Biological Conservation, Vol. 289, 2024, pp. 110394, https://doi.org/10.1016/j.biocon.2023.110394.

[7] Gia Lai Government Portal, Gia Lai Province Has Declared a State of Emergency Due to Storms and Floods in 77 Communes and Wards, https://thonghat.gialai.gov.vn/tin-tuc/tin-tuc-van-hoa-xa-hoi/gia-lai-cong-bo-tinh-huong-khan-cap-ve-thien-tai-do-bao-va-lu-tai-77-xa-phuong.html, (accessed on: January 10th, 2026) (in Vietnamese).

[8] H. Q. Nguyen, J. Degener, M. Kappas, Flash Flood Prediction by Coupling KINEROS2 and HEC-RAS Models for Tropical Regions of Northern

Vietnam, Hydrology, Vol. 2, No. 4, 2015, pp. 242-265, https://doi.org/10.3390/hydrology2040242.

[9] H. D. Luong, Q. H. Le, V. T. Phan, T. H. Van, T. T. Tran, V. D. Nguyen, Application of HEC-RAS Model to Simulate Inundation for the Lower Ba River, Journal of Climate Change Science, No. 30, 2024, pp. 65-76, https://doi.org/10.55659/2525-2496/30.99719 (in Vietnamese).

[10] Vietnam Academy of Science and Technology, Portal of Vietnam Academy of Science and Technology, https://vast.gov.vn/tin-chi-tiet/-/chi-tiet/nghien-cuu-ung-dung-bo-mo-hinh-mike-flood-phuc-vu-canh-bao-lu-va-ngap-lut-luu-vuc-song-thu-bon-tinh-quang-nam-2300-463.html, 2026 (accessed on: January 10th, 2026) (in Vietnamese).

[11] S. Patro, C. Chatterjee, S. Mohanty, R. Singh, N. S. Raghuwanshi, Flood Inundation Modeling Using Mike Flood and Remote Sensing Data, J Indian Soc Remote Sens, Vol. 37, No. 1, 2009, pp. 107-118, https://doi.org/10.1007/s12524-009-0002-1.

[12] G. W. Brunner, HEC-RAS River Analysis System. Hydraulic User's Manual. Version 1.0., https://apps.dtic.mil/sti/html/tr/ADA311953/, 1995 (accessed on: January 10th, 2026).

[13] H. D. Nguyen et al., Flood Susceptibility Mapping Using Advanced Hybrid Machine Learning and CyGNSS: A Case Study of Nghe An Province, Vietnam, Acta Geophys., Vol. 70, No. 6, 2022, pp. 2785-2803, https://doi.org/10.1007/s11600-022-00940-2.

[14] M. Bashirgonbad, B. Farokhzadeh, V. Gholami, Enhancing Flood Mapping Through Ensemble Machine Learning in the Gamasyab Watershed, Western Iran, Environ Sci Pollut Res, Vol. 31, No. 38, 2024, pp. 50427-50442, https://doi.org/10.1007/s11356-024-34501-5.

[15] M. C. Ha et al., Machine Learning and Remote Sensing Application for Extreme Climate Evaluation: Example of Flood Susceptibility in the Hue Province, Central Vietnam Region, Water, Vol. 14, No. 10, 2022, p. 1617.

[16] D. H. Wolpert, W. G. Macready, No Free Lunch Theorems for Optimization, IEEE Transactions on Evolutionary Computation, Vol. 1, No. 1, 1997, pp. 67-82, https://doi.org/10.1109/4235.585893.

[17] D. T. Bui et al., A New Intelligence Approach Based on GIS-based Multivariate Adaptive Regression Splines and Metaheuristic Optimization for Predicting Flash Flood Susceptible Areas at High-frequency Tropical Typhoon Area, Journal of Hydrology, Vol. 575, 2019, pp. 314-326, https://doi.org/10.1016/j.jhydrol.2019.05.046.

[18] T. K. Ho, Random Decision Forests, Proceedings of 3rd International Conference on Document Analysis and Recognition, Vol. 1, 1995, pp. 278-282, https://doi.org/10.1109/ICDAR.1995.598994.

[19] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785-794, https://doi.org/10.1145/2939672.2939785.

[20] C. Cortes, V. Vapnik, Support-vector Networks, Mach Learn, Vol. 20, No. 3, 1995, pp. 273-297, https://doi.org/10.1007/BF00994018.

[21] B. Abdollahzadeh et al., Puma Optimizer (PO): A Novel Metaheuristic Optimization Algorithm and Its Application in Machine Learning, Cluster Comput, Vol. 27, No. 4, 2024, pp. 5235-5283, https://doi.org/10.1007/s10586-023-04221-5.

[22] J. Kennedy, R. Eberhart, Particle Swarm Optimization, Proceedings of ICNN'95 - International Conference on Neural Networks, Vol. 4, 1995, pp. 1942-1948, https://doi.org/10.1109/ICNN.1995.488968.

[23] L. Nabilah, L. Hakim, Hybrid PSO-XGBoost Model for Accurate Flood Risk Assessment, Journal of Applied Informatics and Computing, Vol. 9, No. 6, 2025, pp. 3681-3688, https://doi.org/10.30871/jaic.v9i6.11094.

[24] N. T. T. Linh et al., Flood Susceptibility Modeling Based on New Hybrid Intelligence Model: Optimization of XGboost Model Using GA Metaheuristic Algorithm, Advances in Space Research, Vol. 69, No. 9, 2022, pp. 3301-3318, https://doi.org/10.1016/j.asr.2022.02.027.

[25] S. Mirjalili, S. M. Mirjalili, A. Lewis, Grey Wolf Optimizer, Advances in Engineering Software, Vol. 69, 2014, pp. 46-61, https://doi.org/10.1016/j.advengsoft.2013.12.007.