Tạp chí Khoa học Công nghệ Xây dựng, ĐHXDHN, 2026, 20 (6V): 28–41

ĐÁNH GIÁ ẢNH HƯỞNG CỦA NGUY CƠ NGẬP LỤT ĐẾN NHÀ Ở BẰNG MÔ HÌNH HỌC MÁY VÀ PHÂN TÍCH KHÔNG GIAN, THÍ ĐIỂM TẠI TỈNH THÁI NGUYÊN (CŨ)

Vũ Quốc Hưng[a,*], Tô Ngọc Huy[b], Hà Thị Hằng[c], Lưu Thị Diệu Chinh[a]

[a]*Khoa Công trình Thủy, Trường Đại học Xây dựng Hà Nội,*
55 đường Giải Phóng, phường Bạch Mai, Hà Nội, Việt Nam
[b]*Ban Đào tạo Kỹ sư chất lượng cao, Trường Đại học Xây dựng Hà Nội,*
55 đường Giải Phóng, phường Bạch Mai, Hà Nội, Việt Nam
[c]*Khoa Cầu đường, Trường Đại học Xây dựng Hà Nội,*
55 đường Giải Phóng, phường Bạch Mai, Hà Nội, Việt Nam

Nhận ngày 08/6/2026, Sửa xong 09/11/2026, Chấp nhận đăng 20/11/2026

Tóm tắt

Nghiên cứu này đề xuất tích hợp các mô hình học máy tiên tiến với kỹ thuật phân tích không gian địa lý nhằm định lượng mức độ ảnh hưởng của ngập lụt đến nhà ở tại tỉnh Thái Nguyên (cũ). Dữ liệu đầu vào bao gồm một bộ dữ liệu ngập lụt lịch sử giai đoạn 2016–2024 và 12 yếu tố đặc trưng cho địa hình, môi trường, khí hậu. Hai mô hình học máy tiên tiến, bao gồm CatBoost và Random Forest, đã được sử dụng để xây dựng bản đồ nguy cơ ngập lụt. Kết quả kiểm chứng cho thấy mô hình CatBoost có hiệu suất dự báo vượt trội (AUC = 0,973) nên được chọn để xây dựng bản đồ nguy cơ ngập lụt tốt nhất. Cuối cùng, dữ liệu nhà ở được chồng xếp với bản đồ này trong môi trường GIS để định lượng số nhà bị ảnh hưởng bởi nguy cơ ngập lụt trong từng huyện. Kết quả đạt được cho thấy những huyện có nguy cơ ngập lụt cao và rất cao thì cũng đồng thời là những huyện có số lượng nhà bị ảnh hưởng nhiều nhất, bao gồm huyện Phổ Yên (28.058 nhà), TP. Thái Nguyên (20.565 nhà), huyện Phú Bình (8.028 nhà) và huyện Đại Từ (7.983 nhà).

Từ khoá: nguy cơ ngập lụt; số nhà bị ảnh hưởng; Categorical Boosting; Random Forests; tỉnh Thái Nguyên (cũ).

ASSESSING THE IMPACTS OF FLOOD SUSCEPTIBILITY ON RESIDENTIAL BUILDINGS USING MACHINE LEARNING AND SPATIAL ANALYSIS, PILOT IN THAI NGUYEN PROVINCE (OLD)

Abstract

This study proposes integrating advanced machine learning models with geospatial analysis techniques to quantify the impact of flooding on housing in Thai Nguyen province (old). The input data consists of a historical flood dataset covering the 2016–2024 period and 12 variables characterizing terrain, environmental, and climatic conditions. Two machine learning models - CatBoost and Random Forest - were employed to develop flood susceptibility maps. Validation results indicate that the CatBoost model demonstrated superior predictive performance (AUC = 0.973) and was therefore selected to generate the optimal flood hazard map. Finally, housing data were overlaid with this best flood susceptibility map in GIS to quantify the number of houses exposed to flood hazard in each district. The results indicate that districts with high and very high flood susceptibility also have the highest number of affected houses, including Pho Yen district (28,058 houses), Thai Nguyen city (20,565 houses), Phu Binh district (8,028 houses), and Dai Tu district (7,983 houses).

Keywords: flood susceptibility; exposed housing; Categorical Boosting; Random Forests; Thai Nguyen province (old).

https://doi.org/10.31814/stce.huce2026-20(6V)-03 © 2026 Trường Đại học Xây dựng Hà Nội (ĐHXDHN)

1. Giới thiệu

Lũ lụt là một hiện tượng tự nhiên, vừa mang tính quy luật, vừa mang tính ngẫu nhiên. Các trận lũ lớn và đặc biệt lớn trên sông, nhất là khi lũ lớn tràn bờ và thậm chí vỡ đê, sẽ gây ngập lụt các vùng trũng ven sông, gây ra thiệt hại rất lớn về người, tài sản, sinh kế, cơ sở hạ tầng, và tác động xấu đến môi trường tự nhiên [1]. Lũ lụt gây thiệt hại nghiêm trọng với tác động kinh tế - xã hội toàn cầu đáng

*Tác giả đại diện. *Địa chỉ e-mail:* hungvq@huce.edu.vn (Hưng, V. Q.)

kể, chiếm hơn 50% thiệt hại do thiên tai trong 50 năm qua [2]. Theo ước tính, lũ lụt ảnh hưởng tới khoảng 20 đến 300 triệu người mỗi năm. Mức độ nghiêm trọng và tần suất của các sự kiện lũ lụt được dự báo sẽ gia tăng lên đáng kể trong những thập kỷ tới do tình trạng biến đổi khí hậu và các sự kiện khí hậu cực đoan [3]. Từ trước tới nay, bản đồ nguy cơ ngập lụt vẫn được xem là một công cụ quan trọng nhằm hỗ trợ hiệu quả việc nhận diện các khu vực có nguy cơ ngập lụt trong tương lai theo không gian [4]. Trong khi đó, việc đánh giá và xác định số lượng nhà bị ảnh hưởng bởi nguy cơ ngập lụt có thể giúp xác định mức độ thiệt hại tiềm tàng, lập kế hoạch phân bổ tối ưu nguồn lực và ngân sách, cũng như cho phép lập kế hoạch chủ động để phòng ngừa và di dời nhằm đảm bảo an toàn cho cộng đồng và giảm thiểu thiệt hại kinh tế - xã hội trong tương lai [5].

Trong những năm gần đây, các thuật toán học máy (Machine Learning) tiên tiến đã thu hút được sự quan tâm rộng rãi của các nhà nghiên cứu do những ưu điểm của chúng trong việc mô hình hóa các sự kiện phức tạp như sạt lở đất và lũ lụt [6]. Nhờ khả năng khai thác dữ liệu lớn, tự động nhận diện quy luật tiềm ẩn và mô phỏng các mối quan hệ phi tuyến giữa các yếu tố đầu vào, các thuật toán học máy đang được ứng dụng ngày càng rộng rãi trong đánh giá rủi ro lũ lụt [3]. Đặc biệt, Categorical Boosting (CatBoost) và Random Forest (RF) là những thuật toán học máy mạnh mẽ, có khả năng xử lý hiệu quả dữ liệu phức tạp với cả biến liên tục và biến phân loại, rất phù hợp trong việc mô hình hóa các sự kiện phức tạp như lũ lụt. Chính vì vậy, hai thuật toán CatBoost và RF được sử dụng khá nhiều trong các nghiên cứu thành lập bản đồ nguy cơ ngập lụt trong những năm gần đây. Trong đó, một số nghiên cứu sử dụng đơn thuần mỗi thuật toán RF cùng với các thuật toán học máy khác trong đánh giá nguy cơ ngập lụt [7–9]. Một số nghiên cứu khác thì so sánh hiệu suất dự báo của mỗi thuật toán CatBoost với các thuật toán học máy khác [2, 10]. Cùng với đó, một số nghiên cứu khác thì sử dụng cả thuật toán CatBoost, RF và một số thuật toán khác trong đánh giá nguy cơ ngập lụt [6, 11].

Ở Việt Nam, hai thuật toán học máy RF và CatBoost cũng được sử dụng tách biệt trong một số nghiên cứu đánh giá nguy cơ ngập lụt tại một số địa phương. Điển hình, có thể kể đến một số nghiên cứu như Cường và cs. đã đề xuất sử dụng thuật toán RF cùng nhiều thuật toán học máy tiên tiến khác để đánh giá nguy cơ ngập lụt cho tỉnh Thừa Thiên Huế (cũ) [12]. Tiếp đến là nghiên cứu của Phong và cs. thì sử dụng mỗi thuật toán CatBoost để thành lập bản đồ nguy cơ ngập lụt cho huyện Quế Sơn, tỉnh Quảng Nam (cũ) [13].

Nhìn chung, cả trên thế giới và tại Việt Nam, hai thuật toán Random Forest và Categorical Boosting đã được sử dụng phổ biến trong các nghiên cứu xây dựng bản đồ nguy cơ ngập lụt. Việc đối sánh đồng thời hiệu quả dự báo giữa hai thuật toán này vẫn còn khá hạn chế. Bên cạnh đó, có rất ít nghiên cứu ứng dụng đồng thời các thuật toán học máy này và các kỹ thuật phân tích không gian trong GIS để định lượng số lượng nhà bị ảnh hưởng bởi nguy cơ ngập lụt. Chính vì vậy, nghiên cứu này đề xuất tích hợp hai thuật toán học máy phổ biến là CatBoost và RF và các kỹ thuật phân tích không gian trong GIS để vừa đánh giá nguy cơ ngập lụt, vừa ước tính số lượng nhà ở bị ảnh hưởng bởi nguy cơ ngập lụt trên địa bàn tỉnh Thái Nguyên (cũ). Kết quả của nghiên cứu này sẽ cung cấp cơ sở khoa học và thực tiễn quan trọng cho công tác quy hoạch sử dụng đất, quản lý phát triển đô thị, xây dựng phương án phòng chống thiên tai và nâng cao khả năng thích ứng với ngập lụt tại địa phương.

2. Khu vực và quy trình nghiên cứu

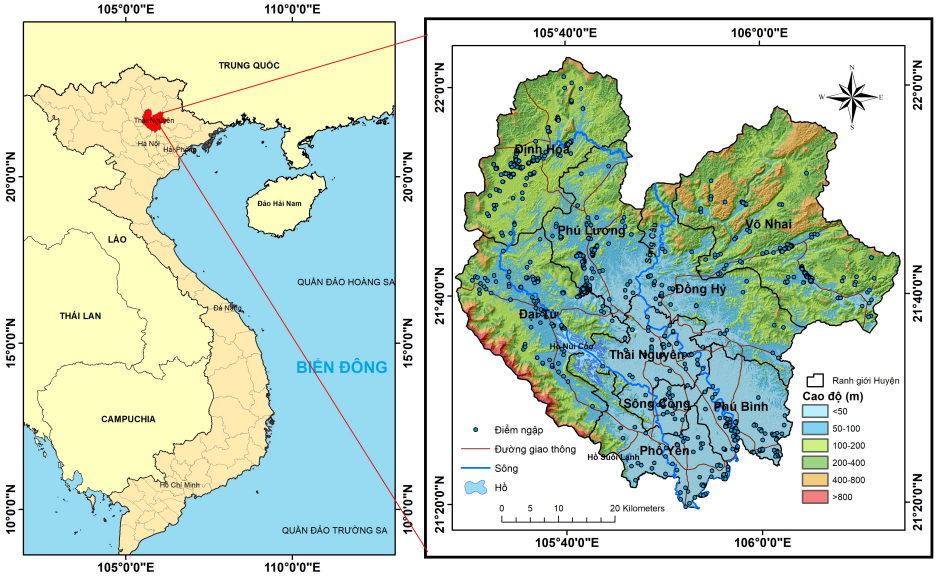
2.1. Khu vực nghiên cứu

Tỉnh Thái Nguyên (cũ) nằm ở vị trí địa lý từ 21°20′ đến 22°03′ vĩ độ Bắc và từ 105°52′ đến 106°14′ kinh độ Đông (Hình 1). Tỉnh này có 178 đơn vị hành chính cấp xã, bao gồm 32 phường, 9 thị trấn và 137 xã [14].

Địa hình tỉnh Thái Nguyên chủ yếu là đồi núi thấp, thấp dần từ Bắc xuống Nam và từ Tây sang Đông, trong đó, đồi núi cao trên 100 m chiếm 2/3 diện tích toàn tỉnh, còn lại là vùng có độ cao dưới

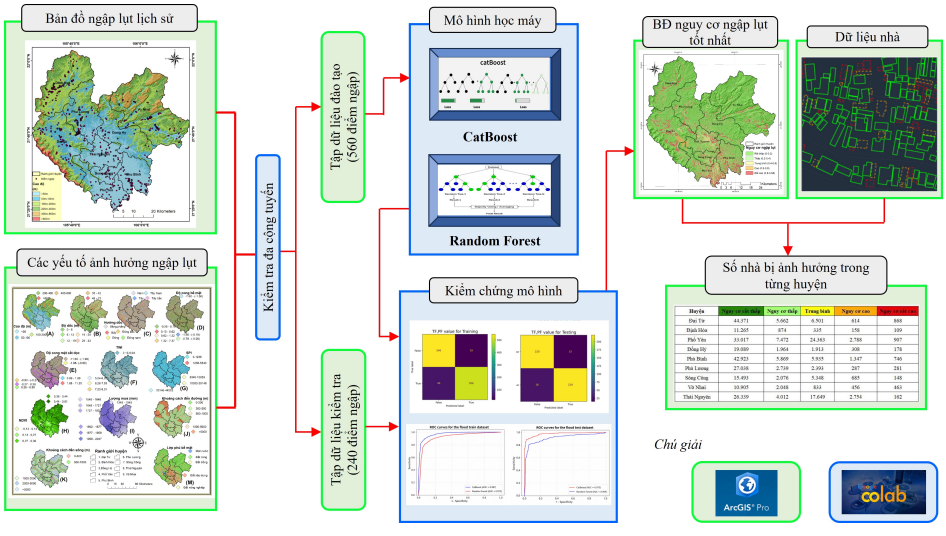
100 m. Khí hậu chia thành 2 mùa rõ rệt, mùa mưa kéo dài từ tháng 5 đến tháng 10, trong khi mùa khô diễn ra từ tháng 10 đến tháng 5 năm sau. Tổng lượng mưa trung bình năm dao động khoảng 2.000–2.500 mm. Lượng mưa tập trung nhiều nhất vào tháng 8 và thấp nhất vào tháng 1 [14].

Hệ thống sông suối trên địa bàn tỉnh Thái Nguyên tương đối phong phú, với các sông chính như sông Cầu và sông Công cùng nhiều suối nhỏ phân bố rộng khắp. Do chịu ảnh hưởng của địa hình đồi núi và khí hậu nhiệt đới gió mùa, sông ngòi trong tỉnh có sự phân hóa rõ rệt theo mùa, do đó, thường xảy ra lũ vào mùa mưa [15].



Hình 1. Bản đồ vị trí điểm ngập lụt giai đoạn 2016–2024 thuộc tỉnh Thái Nguyên (cũ)

2.2. *Quy trình nghiên cứu*



Hình 2. Quy trình nghiên cứu

Quy trình trong nghiên cứu này bao gồm 7 bước sau: (1) Chuẩn bị và thu thập dữ liệu, (2) Kiểm tra hiện tượng đa cộng tuyến của các yếu tố ảnh hưởng đến ngập lụt, (3) Dữ liệu ngập lụt lịch sử được chia thành hai nhóm: 70% tập dữ liệu đào tạo (tương ứng với 560 điểm ngập) được sử dụng để xây

dựng mô hình dự báo nguy cơ ngập lụt và 30% tập dữ liệu kiểm tra (tương ứng với 240 điểm ngập) được sử dụng để kiểm chứng mô hình dự báo, (4) Phát triển các mô hình học máy tiên tiến, bao gồm CatBoost và RF trên nền tảng Google Colab, (5) Xây dựng các mô hình dự báo nguy cơ ngập lụt, (6) So sánh hiệu suất dự báo của các mô hình nguy cơ ngập lụt để lựa chọn và thành lập bản đồ nguy cơ ngập lụt có độ chính xác và độ tin cậy cao nhất, và (7) Chồng xếp bản đồ nguy cơ ngập lụt tốt nhất với dữ liệu nhà ở để định lượng số nhà bị ảnh hưởng trong từng huyện (Hình 2).

3. Dữ liệu và phương pháp sử dụng

3.1. Dữ liệu sử dụng

a. Dữ liệu ngập lụt lịch sử

Dữ liệu ngập lụt lịch sử cung cấp thông tin về vị trí không gian của điểm ngập lụt trong quá khứ, điều này không chỉ giúp giải mã các nguy cơ trong tương lai mà còn đóng vai trò bảo chứng độ tin cậy và khả năng áp dụng trong thực tế của các bản đồ nguy cơ ngập lụt [6]. Trong nghiên cứu này, việc xác định vết lũ lịch sử được thực hiện dựa trên chuỗi dữ liệu ảnh vệ tinh radar Sentinel-1 trong giai đoạn từ năm 2016 đến năm 2024, do chúng chỉ được cung cấp trong những năm gần đây. Quy trình xử lý dữ liệu ảnh Sentinel-1 chiết tách thông tin vùng ngập lũ được tiến hành theo nghiên cứu của [16]. Theo đó, quy trình chiết tách được điều chỉnh bằng cách sử dụng lớp nền mặt nước thường xuyên trước mùa lũ, tổ hợp theo giá trị trung vị nhằm tách các khu vực mặt nước ổn định như sông, kênh, rạch và ao nuôi thủy sản, từ đó loại bỏ các vùng không phải do ngập lụt trong phân tích ảnh vệ tinh đa thời gian [16].

Quá trình chiết xuất được thực hiện theo từng sự kiện lũ riêng biệt trong giai đoạn 2016–2024, tương ứng với mùa mưa hàng năm tại khu vực nghiên cứu; sau đó, các lớp bản đồ ngập của từng năm được chồng xếp không gian trong môi trường GIS. Các vị trí được ghi nhận là điểm ngập lụt lịch sử đáng tin cậy khi xuất hiện từ 5 lần ngập trở lên, nhằm loại bỏ nhiễu do sai số ngẫu nhiên của ảnh SAR (như bóng địa hình, mặt nước tạm thời) và tăng độ tin cậy của điểm chọn. Kết quả của quy trình thu được 800 điểm vết lũ lịch sử tại tỉnh Thái Nguyên giai đoạn 2016–2024 (Hình 1).

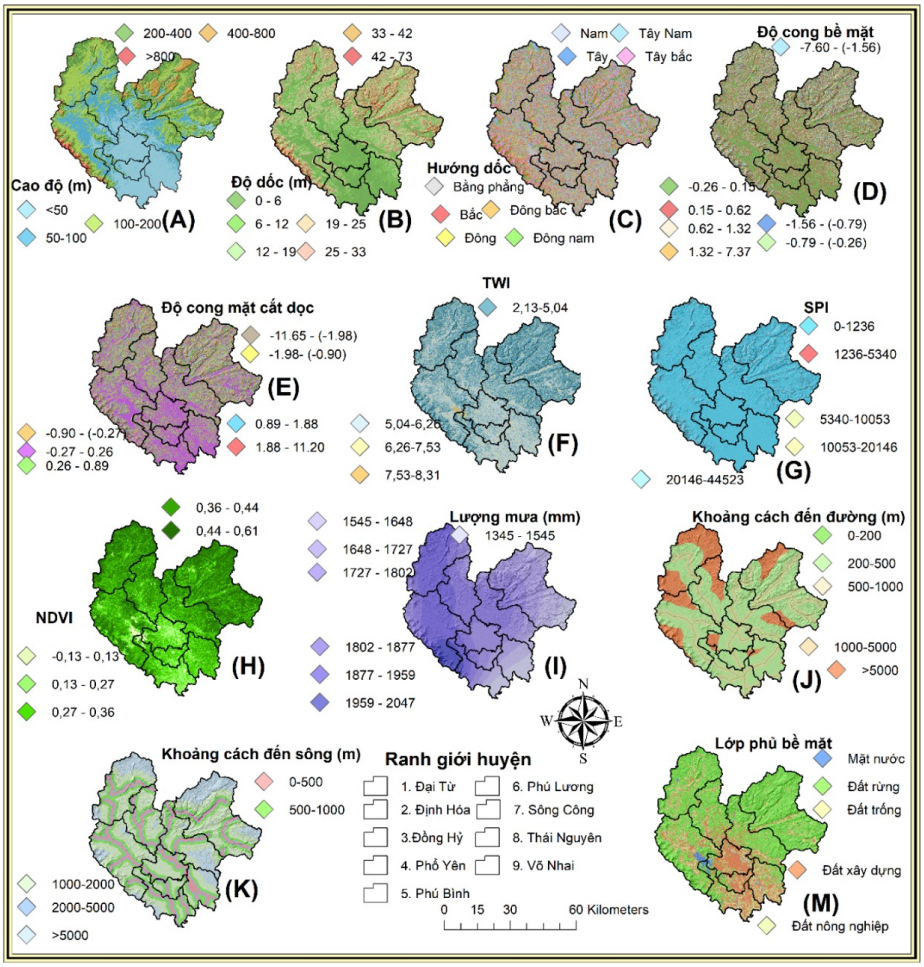
b. Các yếu tố ảnh hưởng tới ngập lụt

Hiện tượng ngập lụt thường xảy ra tại những khu vực địa hình trũng thấp, thoát nước kém, đặc biệt là tại lưu vực các con sông lớn [6]. Trong đó, lượng mưa lớn kéo dài được xem là nguyên nhân chính kích hoạt cơ chế hình thành ngập lụt [3]. Những khu vực có mật độ đô thị hóa cao kèm theo hệ thống thoát nước trong khu vực kém cũng thường xuyên phải đối mặt với ngập lụt [11]. Sự suy giảm của lớp phủ thực vật do tiến trình đô thị hóa cũng được xem là yếu tố thúc đẩy sự hình thành lũ [16]. Trong nghiên cứu này, 12 yếu tố ảnh hưởng tới ngập lụt được lựa chọn dựa trên mức độ liên quan, khả năng tiếp cận và tính sẵn có, bao gồm: độ cao, góc dốc, hướng sườn, độ cong bề mặt, mật độ chia cắt sâu, Chỉ số độ ẩm địa hình (TWI), Chỉ số năng lượng dòng chảy (SPI), Chỉ số khác biệt chuẩn hóa thực vật (NDVI), khoảng cách đến đường, khoảng cách đến sông, lượng mưa trung bình năm, và hiện trạng sử dụng đất.

Trong đó, dữ liệu độ cao được khai thác từ dữ liệu SRTM DEM toàn cầu với độ phân giải không gian 30 m, tải xuống vào tháng 12 năm 2024 từ trang web https://earthexplorer.usgs.gov/. Các yếu tố khác liên quan đến địa hình, như góc dốc, hướng sườn, độ cong bề mặt, mật độ chia cắt sâu, TWI và SPI, được tính toán từ dữ liệu SRTM DEM trong môi trường GIS, với độ phân giải không gian 30 m. Khoảng cách tới đường và khoảng cách tới sông được chiết tách từ bản đồ địa hình tỷ lệ 1:50.000 và được cung cấp bởi Cục Khảo sát, Cục Đo đạc, Bản đồ và Hệ thống Thông tin Địa lý (thuộc Bộ Tài nguyên và Môi trường Việt Nam) vào năm 2024. Lượng mưa trung bình từ năm 2011 đến năm 2024 được thu thập từ 17 trạm đo mưa của Trung tâm Dự báo Khí tượng Thủy văn Quốc gia Việt Nam. Dữ

liệu NDVI được thu thập và tính toán dựa trên dữ liệu ảnh Landsat-8 vào tháng 7 năm 2023. Bản đồ lớp phủ bề mặt khu vực nghiên cứu có tỷ lệ 1:50.000 và được cung cấp bởi Bộ Tài nguyên và Môi trường Việt Nam vào năm 2024.

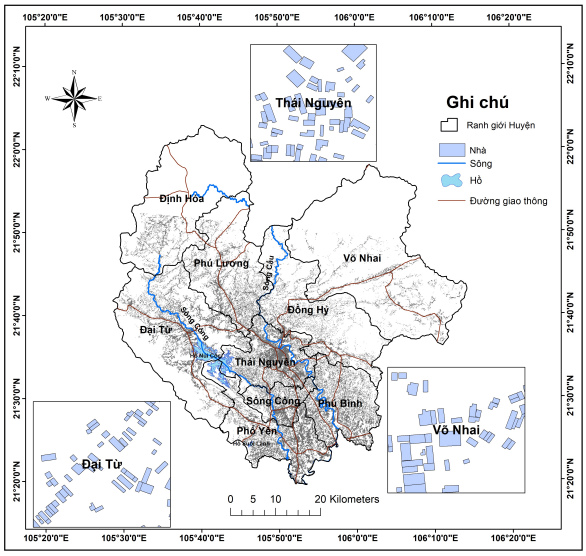
Tất cả các yếu tố ảnh hưởng tới ngập lụt này đều được chuyển đổi sang định dạng raster với độ phân giải không gian 30 m trong phần mềm ArcGIS Pro (Hình 3).



Hình 3. Bản đồ chuyên đề 12 lớp yếu tố ảnh hưởng tới ngập lụt

c. Dữ liệu nhà ở

Trong nghiên cứu này, dữ liệu về nhà ở được trích xuất từ ảnh vệ tinh có độ phân giải không gian siêu cao (50 cm) do Google Research Open Buildings cung cấp (https://developers.google.com/earth-engine/datasets/catalog/GOOGLE_Research_open-buildings_v3_polygons) và công bố vào tháng 5 năm 2023. Nền tảng này cung cấp hình dạng chi tiết của các tòa nhà ở định dạng đa giác, cùng với điểm tin cậy đạt được nhằm xác nhận độ tin cậy của dữ liệu [5]. Để đảm bảo độ chính xác, nghiên cứu này chỉ giữ lại các tòa nhà có điểm tin cậy trên 80% và diện tích tối thiểu 30 m^2 để tiến hành phân tích và đánh giá ảnh hưởng của nguy cơ ngập lụt tới nhà ở trên địa bàn khu vực nghiên cứu (Hình 4).



Hình 4. Bản đồ dữ liệu nhà ở sử dụng trong nghiên cứu

3.2. Phương pháp sử dụng

a. Kiểm tra tính đa cộng tuyến của các yếu tố ảnh hưởng

Hiện tượng đa cộng tuyến (multicollinearity) xảy ra khi các yếu tố đầu vào trong mô hình có mối tương quan mạnh với nhau, tức là một biến có thể được giải thích bởi một hoặc nhiều biến khác. Điều này dẫn tới việc khó xác định tách bạch tác động của từng yếu tố ảnh hưởng cũng như làm giảm độ tin cậy của mô hình dự báo. Hệ số phóng đại phương sai (VIF) và giá trị dung sai là những chỉ số thống kê quan trọng và thường được dùng để phát hiện sự tồn tại của đa cộng tuyến giữa các biến đầu vào [17]. Giá trị VIF đo lường mức độ mà phương sai của hệ số hồi quy ước lượng tăng lên do đa cộng tuyến. Dung sai là nghịch đảo của giá trị VIF và phản ánh tỷ lệ phương sai của một biến dự báo mà không được dự đoán bởi tổ hợp của các biến dự báo khác [18]. Đa cộng tuyến đáng kể giữa các biến đầu vào được phát hiện nếu giá trị VIF vượt quá 10 hoặc giá trị dung sai giảm xuống dưới 0,1.

b. Các thuật toán học máy sử dụng

- Thuật toán Categorical Boosting (CatBoost)

Thuật toán Categorical Boosting (CatBoost) là một phương pháp tăng cường gradient sử dụng các cây quyết định được xây dựng theo trình tự liên tiếp nhằm giảm sai số và nâng cao độ chính xác của mô hình dự đoán [19]. Cơ chế huấn luyện của CatBoost tương tự như các thuật toán boosting khác, đó là xây dựng tuần tự các cây quyết định. Ở mỗi bước, cây mới được điều chỉnh để khử dần sai số (gradient âm của hàm mất mát) của mô hình trước đó. Nội dung của các thuật toán Boosting có thể tóm lược như sau [20]:

Nguyên lý chung của các thuật toán Boosting là xây dựng tuần tự các mô hình học yếu nhằm giảm sai số dự đoán trên tập huấn luyện. Đối với bài toán phân loại nhị phân, biến mục tiêu được ký hiệu là $y \in \{-1, 1\}$, và hàm dự đoán của mô hình đối với đầu vào x_i được biểu diễn là $\hat{f}(x_i)$. Sai số phân loại trên tập huấn luyện được xác định như sau:

$$r = \frac{1}{N} \sum_{i=1}^{N} I\left(y_i \neq \hat{f}(x_i)\right) \tag{1}$$

trong đó: $I(.)$ là hàm chỉ thị nhận giá trị 1 khi điều kiện bên trong đúng và nhận giá trị 0 trong trường hợp ngược lại.

- Thuật toán Random Forest (RF)

Thuật toán Random Forest (RF) được giới thiệu và phát triển bởi Breiman (2001). Thuật toán này hoạt động dựa trên việc kết hợp nhiều cây quyết định nhằm nâng cao độ chính xác dự đoán và giảm hiện tượng quá khớp [8]. Nó được sử dụng cho cả nhiệm vụ phân loại và hồi quy. Nội dung của thuật toán này có thể được mô tả như sau: Từ tập dữ liệu huấn luyện ban đầu $D = (X, Y)$, trong đó X là tập biến đầu vào và Y là biến mục tiêu, thuật toán tạo ra nhiều tập mẫu bootstrap bằng phương pháp lấy mẫu có hoàn lại từ dữ liệu gốc. Mỗi cây quyết định được xây dựng trên một tập bootstrap riêng. Tại mỗi nút phân chia, thuật toán chọn ngẫu nhiên một tập con gồm m đặc trưng từ tổng số M đặc trưng để tìm phép chia tối ưu, qua đó làm tăng tính đa dạng giữa các cây. Theo đó, các cây quyết định được phát triển theo cơ chế phân chia đệ quy dựa trên các đặc trưng được chọn và tiêu chí phân tách phù hợp [7]. Kết quả cuối cùng của RF được xác định bằng cách tổng hợp dự đoán từ tất cả các cây quyết định. Đối với bài toán phân loại, lớp có số phiếu bầu nhiều nhất được chọn làm kết quả dự đoán cuối cùng, còn đối với bài toán hồi quy, giá trị dự đoán được tính bằng trung bình của các cây [7, 8].

- Quá trình lựa chọn siêu tham số

Trong nghiên cứu này, các siêu tham số của hai mô hình CatBoost và Random Forest được xác định thông qua quá trình thử nghiệm lặp lại (trial-and-error) nhằm cân bằng giữa độ chính xác của mô hình và khả năng khái quát hóa, tránh hiện tượng quá khớp (overfitting) trên tập dữ liệu huấn luyện trong môi trường Google Colab. Đối với mô hình CatBoost, quá trình khởi tạo được thực hiện với các tham số mặc định trước, sau đó, tinh chỉnh 3 siêu tham số chính bao gồm tốc độ học (*learning_rate*), số lượng cây (*n_estimators*), giá trị khởi tạo ngẫu nhiên (*random_seed*) nhằm đảm bảo tính ổn định và khả năng tái lặp của kết quả. Đối với mô hình Random Forest, các siêu tham số được lựa chọn bao gồm số lượng cây quyết định (*n_estimators*) và độ sâu tối đa của mỗi cây (*max_depth*), đồng thời giá trị khởi tạo ngẫu nhiên (*random_seed*) được giữ cố định để đảm bảo kết quả có thể tái lập giữa các lần chạy. Nhìn chung, quá trình thử nghiệm lặp lại nhằm tối ưu hóa tham số của hai mô hình được thực hiện với số tổ hợp tham số thử nghiệm *n_iter = 100* và 95, kết hợp kiểm định chéo 10-fold (*cv = 10*), trong đó bộ dữ liệu được chia thành 10 phần bằng nhau; mỗi lần lặp sử dụng 9 phần để huấn luyện và 1 phần để kiểm định mô hình, cùng với giá trị *random_seed = 42* và 40.

c. Kiểm chứng độ chính xác dự báo của các thuật toán học máy

Việc kiểm chứng độ chính xác dự báo của các mô hình học máy nhằm đảm bảo độ tin cậy và khả năng ứng dụng trong thực tiễn của chúng [21]. Trong nghiên cứu này, các chỉ số thống kê phổ biến đã được sử dụng để đánh giá hiệu suất dự báo của mô hình học máy, bao gồm: Độ bao phủ (Recall) hay độ nhạy cảm (Sensitivity), Độ chính xác dự đoán (Precision), Độ chính xác tổng thể (Accuracy), Giá trị F1 (F1-score), chỉ số Jaccard, Sai số trung phương - Root Mean Squared Error (RMSE), đường cong ghi nhận tín hiệu ROC – (Receiver Operating Characteristic curve - ROC), và diện tích dưới đường cong ROC (Area Under the ROC Curve - AUC). Những chỉ số này được xác định bằng cách áp dụng các công thức từ (2) đến (7) như sau:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$F\text{-}measure = \frac{2TP}{2TP + FP + FN} \tag{4}$$

$$Jaccard = \frac{TP}{TP + FP + FN} \tag{5}$$

$$Accuracy = \frac{TN + TP}{TN + FN + TP + FP} \tag{6}$$

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(X_{Real} - X_{For.})^2}{N}} \tag{7}$$

trong đó TP (True Positive) và TN (True Negative) là số pixel trong sự kiện ngập lụt được phân loại chính xác, FP (False Positive) và FN (False Negative) là số pixel không phải ngập lụt được phân loại chính xác, N là tổng số trường hợp trong tập dữ liệu đào tạo/kiểm tra, X_{Real} là giá trị thực của trường hợp thứ i và $X_{For.}$ là giá trị dự đoán của trường hợp thứ i.

Đường cong đặc tính hoạt động của bộ thu ROC (Receiver Operating Characteristic) được xây dựng bằng cách biểu diễn mối quan hệ giữa tỷ lệ dương tính thật (độ nhạy) và tỷ lệ dương tính giả (1 − độ đặc hiệu) tại nhiều ngưỡng phân loại khác nhau [22]. Độ nhạy định lượng khả năng của mô hình trong việc xác định chính xác các khu vực dễ bị ảnh hưởng, trong khi độ đặc hiệu đo lường khả năng xác định chính xác các khu vực không dễ bị ảnh hưởng [21]. Giá trị AUC là thước đo trong việc định lượng chất lượng của mô hình dự đoán. Giá trị AUC nằm trong khoảng từ 0 đến 1, trong đó giá trị AUC từ 0,5–0,6 phản ánh hiệu suất dự đoán thấp, từ 0,7–0,8 được coi là hiệu suất dự đoán trung bình, từ 0,8–0,9 cho thấy hiệu suất dự đoán tốt, và từ 0,9–1,0 biểu thị hiệu suất dự đoán xuất sắc [21].

d. Phân tích tầm quan trọng của các yếu tố đầu vào

Việc đánh giá mức độ quan trọng của các biến đầu vào đóng vai trò quan trọng trong việc nhận diện những yếu tố thực sự chi phối kết quả dự báo, đồng thời góp phần cải thiện độ tin cậy và tính minh bạch của mô hình được xây dựng [23]. Phương pháp SHAP (SHapley Additive exPlanations) thuộc lĩnh vực trí tuệ nhân tạo có thể giải thích và định lượng được mức độ đóng góp của từng biến đầu vào tới kết quả dự đoán của các mô hình học máy [24]. Dựa trên lý thuyết trò chơi hợp tác, kỹ thuật Shapley phân phối tổng lợi ích thu được thông qua sự hợp tác giữa những người tham gia trò chơi dựa trên đóng góp biên của mỗi người tham gia. Từ đó, giá trị SHAP biểu diễn dự đoán của mô hình dưới dạng tổng của các hiệu ứng từ từng biến đầu vào [23, 24].

4. Kết quả và thảo luận

4.1. Xây dựng bản đồ nguy cơ ngập lụt

a. Kiểm tra tính đa cộng tuyến của các yếu tố ảnh hưởng đến ngập lụt

Trong nghiên cứu này, hệ số phóng đại phương sai (VIF) và dung sai (TOL) được sử dụng để kiểm tra tính đa cộng tuyến của 12 yếu tố ảnh hưởng tới ngập lụt. Kết quả nhận được cho thấy 12 yếu tố ảnh hưởng này đều có giá trị VIF dao động trong khoảng từ 1,01 đến 1,42; các giá trị VIF tính được đều nhỏ hơn 10. Trong khi đó, dung sai TOL nhỏ nhất tính được là 0,70, tất cả giá trị TOL tính được đều lớn hơn 0,10. Do vậy, tất cả 12 yếu tố này đều được lựa chọn để xây dựng mô hình dự báo nguy cơ ngập lụt cho khu vực tỉnh Thái Nguyên (Bảng 1).

Bảng 1. Kết quả kiểm tra đa cộng tuyến của 12 yếu tố ảnh hưởng tới ngập lụt

STT	Các yếu tố ảnh hưởng	VIF	TOL
1	Lớp phủ bề mặt	1,42	0,70
2	Chỉ số độ ẩm địa hình (TWI)	1,36	0,74
3	Chỉ số khác biệt chuẩn hóa thực vật (NDVI)	1,29	0,78
4	Chỉ số năng lượng dòng chảy (SPI)	1,20	0,83

STT	Các yếu tố ảnh hưởng	VIF	TOL
5	Khoảng cách đến đường	1,14	0,88
6	Khoảng cách đến sông	1,12	0,89
7	Góc dốc	1,12	0,89
8	Mật độ chia cắt sâu	1,07	0,94
9	Độ cong bề mặt	1,05	0,95
10	Độ cao	1,03	0,97
11	Hướng sườn	1,02	0,98
12	Lượng mưa	1,01	0,99

b. So sánh hiệu suất của các mô hình dự báo ngập lụt

Trong nghiên cứu này, việc so sánh và đánh giá hiệu suất dự báo của hai mô hình CatBoost và RF được thực hiện thông qua các chỉ số thống kê, ma trận nhầm lẫn của từng mô hình và giá trị AUC của đường cong nhận diện tín hiệu ROC (Bảng 2 và Hình 5).

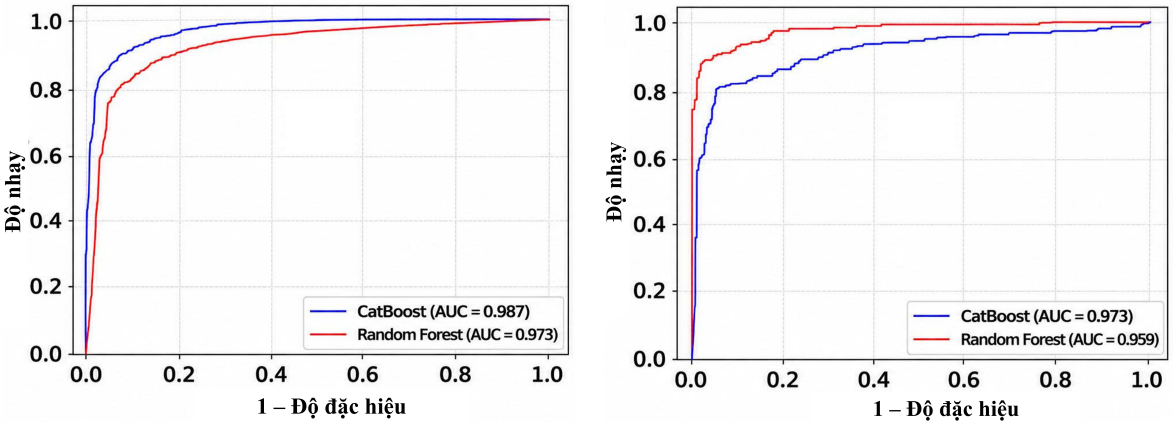
Kết quả đánh giá từ ma trận nhầm lẫn và các chỉ số thống kê cho thấy mô hình CatBoost có hiệu suất dự báo tốt hơn so với mô hình RF trên cả tập dữ liệu đào tạo và kiểm tra. Trên tập dữ liệu đào tạo, mô hình CatBoost đạt độ chính xác (Accuracy) 93,76%, cao hơn so với 91,71% của mô hình RF. Đồng thời, các chỉ số Precision, Recall và F1-Score của CatBoost cũng đạt giá trị cao hơn, lần lượt là 96,38%, 91,17% và 93,70%, trong khi mô hình RF chỉ đạt 94,26%, 88,83% và 91,47%. Bên cạnh đó, giá trị sai số RMSE của CatBoost đạt 0,250, thấp hơn so với giá trị 0,288 của Random Forest, cho thấy sai số dự báo của mô hình CatBoost nhỏ hơn. Đối với tập dữ liệu kiểm tra, mô hình CatBoost tiếp tục duy trì hiệu suất ổn định với Accuracy đạt 91,46%, Precision đạt 93,59%, Recall đạt 89,39% và F1-Score đạt 91,44%, đồng thời RMSE đạt 0,292. Trong khi đó, mô hình Random Forest đạt Accuracy 89,38%, Precision 91,81%, Recall 86,94%, F1-Score 89,31% và RMSE 0,326 (Bảng 2).

Bảng 2. Các chỉ số thống kê dùng trong đánh giá hiệu suất mô hình dự báo

Các chỉ số thống kê	Mô hình CatBoost		Mô hình RF	
	Dữ liệu đào tạo	Dữ liệu kiểm tra	Dữ liệu đào tạo	Dữ liệu kiểm tra
TN	516	220	525	216
FP	19	15	30	19
FN	49	26	62	32
TP	506	219	493	213
Accuracy (%)	93,76	91,46	91,71	89,38
Precision (%)	96,38	93,59	94,26	91,81
Recall (%)	91,17	89,39	88,83	86,94
F1-Score (%)	93,70	91,44	91,47	89,31
RMSE	0,250	0,292	0,288	0,326

Kết quả đánh giá đường cong ROC trên Hình 5 cho thấy, cả hai mô hình học máy CatBoost và RF đều có khả năng phân loại rất tốt trên cả tập dữ liệu đào tạo và kiểm tra. Đối với tập dữ liệu đào tạo, mô hình CatBoost đạt giá trị AUC là 0,987, cao hơn so với mô hình Random Forest (AUC = 0,973) (Hình 5(a)). Đối với tập dữ liệu kiểm tra, mô hình CatBoost tiếp tục duy trì hiệu suất dự báo cao với giá trị AUC đạt 0,973, trong khi mô hình Random Forest đạt giá trị AUC là 0,959 (Hình 5(b)). Mặc dù cả hai mô hình học máy đều có hiệu suất dự báo rất tốt với giá trị AUC đều lớn hơn 0,95, song mô

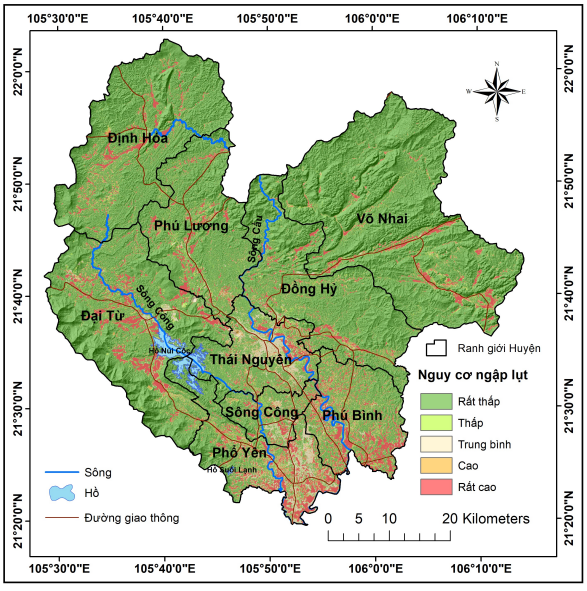
hình CatBoost thể hiện hiệu suất dự báo vượt trội hơn hẳn so với mô hình Random Forest trên cả hai tập dữ liệu.



(a) Tập dữ liệu đào tạo (b) Tập dữ liệu kiểm tra

Hình 5. Kết quả tính toán đường cong ROC của mô hình CatBoost và RF

c. Lựa chọn và xây dựng bản đồ nguy cơ ngập lụt tốt nhất



Hình 6. Bản đồ nguy cơ ngập lụt tỉnh Thái Nguyên (cũ) theo mô hình CatBoost

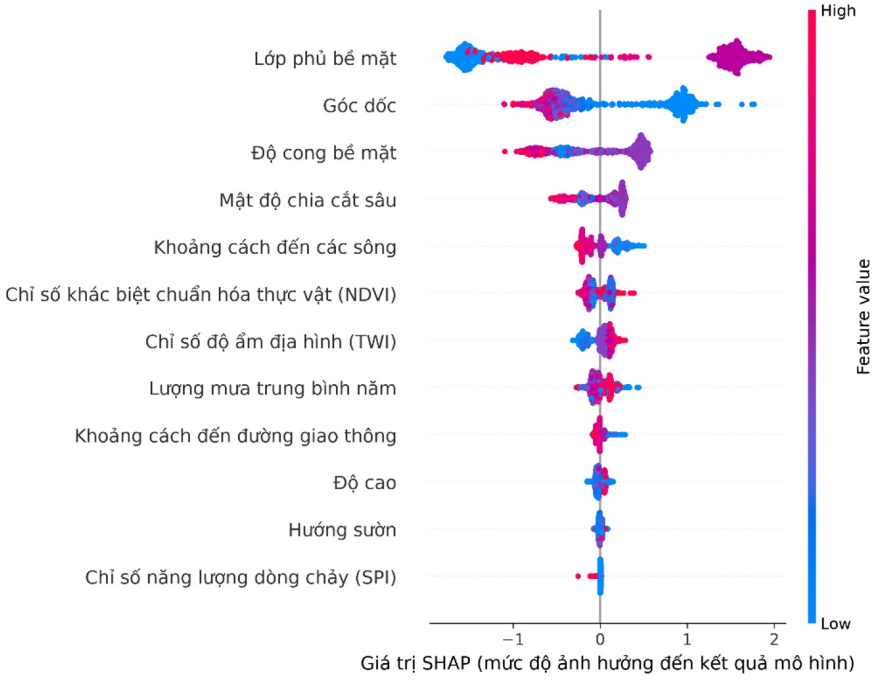
Dựa vào kết quả so sánh hiệu suất dự báo của hai mô hình học máy, nghiên cứu này lựa chọn mô hình CatBoost để xây dựng bản đồ nguy cơ ngập lụt cho tỉnh Thái Nguyên (cũ). Theo đó, bản đồ nguy cơ ngập lụt tỉnh Thái Nguyên theo mô hình CatBoost được phân loại thành năm cấp độ: rất cao, cao, trung bình, thấp và rất thấp (Hình 6). Dựa trên kết quả phân tích bản đồ nguy cơ ngập lụt theo mô hình CatBoost, nếu chỉ xét diện tích các vùng nằm trong nguy cơ từ trung bình đến rất cao, thứ tự giảm dần giữa các huyện trên địa bàn tỉnh Thái Nguyên (cũ) được xác định như sau: Huyện Phổ Yên là khu vực có nguy cơ ngập lụt lớn nhất với 113,29 km^2, tiếp theo là huyện Đại Từ với 103,43 km^2, ở vị trí thứ ba là huyện Phú Bình với 87,15 km^2, tiếp đến là TP. Thái Nguyên với 81,18 km^2, huyện

Võ Nhai với 56,07 km^2, huyện Định Hóa với 49,40 km^2, huyện Đồng Hỷ với 48,52 km^2, huyện Phú Lương với 45,19 km^2, và cuối cùng là thị xã Sông Công với 30,53 km^2.

Bản đồ phân vùng nguy cơ ngập lụt theo mô hình CatBoost cho thấy các khu vực có nguy cơ ngập cao và rất cao tập trung chủ yếu tại các khu vực trung tâm, nơi có địa hình trũng thấp, ven sông suối và nơi có mật độ đô thị hóa cao, như tại thành phố Thái Nguyên, thành phố Sông Công và thành phố Phổ Yên. Đây là những khu vực tập trung đông dân cư, mật độ xây dựng lớn, tốc độ đô thị hóa nhanh, nên làm gia tăng khả năng tích tụ dòng chảy mặt khi xảy ra mưa lớn. Ngoài ra, các khu vực ven sông và vùng trũng thấp, như huyện Phú Bình, cũng xuất hiện nhiều điểm nguy cơ cao, cho thấy ảnh hưởng đáng kể của địa hình thấp và mạng lưới thủy văn đến khả năng ngập lụt.

Tại các vùng núi cao như Võ Nhai, Định Hóa và phần lớn diện tích Phú Lương chủ yếu thuộc nhóm nguy cơ thấp đến trung bình. Nguyên nhân là do các khu vực này có địa hình cao, độ dốc lớn và mật độ xây dựng thấp, giúp quá trình thoát nước diễn ra nhanh hơn và hạn chế tình trạng tích nước kéo dài. Tuy nhiên, một số dải nguy cơ trung bình đến cao vẫn xuất hiện cục bộ dọc theo các thung lũng và khu vực ven suối, cho thấy nguy cơ ngập cục bộ vẫn có thể xảy ra khi xuất hiện mưa cực đoan. Nhận định này cũng phù hợp với những công bố gần đây của Vũ Đình Chiều và cs. [25] hay của Vũ Quốc Hưng và cs. [21].

d. Phân tích tầm quan trọng của các yếu tố đầu vào



Hình 7. Giá trị SHAP của 12 yếu tố ảnh hưởng đến ngập lụt trong mô hình CatBoost

Phương pháp SHAP (SHapley Additive exPlanations) được sử dụng trong việc đánh giá định lượng mức độ đóng góp của từng biến đầu vào trong mô hình dự báo nguy cơ ngập lụt tốt nhất theo thuật toán học máy CatBoost. Trên biểu đồ, mỗi điểm tương ứng với một mẫu dữ liệu, trục hoành thể hiện mức độ và chiều hướng tác động của biến đó đến kết quả dự báo của mô hình. Theo đó, giá trị dương cho thấy dự báo về phía nhạy cảm ngập lụt cao hơn và ngược lại. Thang màu biểu thị mức độ tác động của từng yếu tố đầu vào, trong đó màu đỏ thể hiện các yếu tố có tác động tích cực tới ngập lụt và màu xanh thể hiện các yếu tố có tác động tiêu cực tới ngập lụt. Xét theo mức độ ảnh hưởng

giảm dần tới nguy cơ ngập lụt, thứ tự 12 yếu tố đầu vào được sắp xếp như sau: Lớp phủ bề mặt > Góc dốc > Độ cong bề mặt > Mật độ chia cắt sâu > Khoảng cách đến các sông > Chỉ số khác biệt chuẩn hóa thực vật (NDVI) > Chỉ số độ ẩm địa hình (TWI) > Lượng mưa trung bình năm > Khoảng cách đến đường giao thông > Độ cao > Hướng sườn > Chỉ số năng lượng dòng chảy (SPI) (Hình 7).

Trong đó, các yếu tố (lớp phủ bề mặt, góc dốc, độ cong bề mặt, mật độ chia cắt sâu và khoảng cách đến các sông) là những biến có biên độ giá trị SHAP lớn nhất, cho thấy vai trò chi phối rõ rệt đối với đầu ra của mô hình. Ngược lại, các yếu tố (khoảng cách đến đường giao thông, độ cao, hướng sườn và chỉ số SPI) là những biến có tác động nhỏ hơn, thể hiện qua dao động SHAP hẹp và tập trung gần giá trị 0.

4.2. Đánh giá ảnh hưởng của nguy cơ ngập lụt tới khu dân cư

Để đánh giá định lượng số lượng nhà bị ảnh hưởng bởi nguy cơ ngập, nghiên cứu này đã chồng xếp bản đồ nguy cơ ngập lụt tốt nhất theo mô hình CatBoost với dữ liệu nhà ở trong môi trường GIS. Kết quả cho thấy, nếu chỉ tính riêng số lượng nhà bị ảnh hưởng bởi ngập lụt trong vùng nguy cơ từ trung bình tới rất cao thì thứ tự giảm dần giữa các huyện sẽ bao gồm: Huyện Phổ Yên (28.058 nhà) > TP. Thái Nguyên (20.565 nhà) > Huyện Phú Bình (8.028 nhà) > Huyện Đại Từ (7.983 nhà) > thị xã Sông Công (6.181 nhà) > Huyện Đồng Hỷ (2.961 nhà) > Huyện Võ Nhai (1.752 nhà) > Huyện Định Hóa (602 nhà), Bảng 3.

Bảng 3. Kết quả thống kê số lượng nhà bị ảnh hưởng bởi nguy cơ ngập lụt

Huyện	Số lượng nhà bị ảnh hưởng bởi nguy cơ ngập lụt theo các mức				
	Rất thấp	Thấp	Trung bình	Cao	Rất cao
Đại Từ	44.371	5.662	6.501	614	868
Định Hóa	11.265	874	335	158	109
Phổ Yên	33.017	7.472	24.363	2.788	907
Đồng Hỷ	19.089	1.964	1.913	308	178
Phú Bình	42.923	5.869	5.935	1.347	746
Phú Lương	27.038	2.739	2.393	287	281
Sông Công	15.493	2.076	5.348	685	148
Võ Nhai	10.905	2.048	833	456	463
TP. Thái Nguyên	26.339	4.012	17.649	2.754	162

Nhìn chung, số lượng nhà ở bị ảnh hưởng bởi nguy cơ ngập lụt phân bố không đồng đều giữa các huyện trên địa bàn tỉnh Thái Nguyên (cũ). Trong đó, huyện Phổ Yên, TP. Thái Nguyên, huyện Phú Bình và huyện Đại Từ là những khu vực có số lượng nhà ở bị ảnh hưởng nặng nhất. Đây là những khu vực có địa hình trũng thấp, nằm ven sông nên đối diện với nguy cơ ngập lụt thường xuyên và hứng chịu hậu quả nặng nề hơn. Ngược lại, các huyện miền núi như Định Hóa, Đồng Hỷ và Phú Lương có số lượng nhà nằm trong vùng nguy cơ cao và rất cao tương đối thấp, phù hợp với đặc điểm địa hình cao và phân bố dân cư phân tán.

5. Kết luận

Nghiên cứu này đã tích hợp hiệu quả các mô hình học máy tiên tiến với kỹ thuật phân tích không gian địa lý nhằm định lượng mức độ ảnh hưởng của ngập lụt đến nhà ở trên địa bàn tỉnh Thái Nguyên (cũ). Hai mô hình học máy tiên tiến, bao gồm CatBoost và Random Forest, đã được sử dụng để xây dựng bản đồ nguy cơ ngập lụt dựa trên bộ dữ liệu ngập lụt lịch sử giai đoạn 2016–2024 và 12 yếu tố ảnh hưởng. Kết quả phân tích đường cong ROC trên tập dữ liệu kiểm tra cho thấy, mô hình CatBoost

có hiệu suất dự báo vượt trội so với mô hình Random Forest, với giá trị AUC là 0,973, do đó, mô hình này được lựa chọn để xây dựng bản đồ nguy cơ ngập lụt tốt nhất cho khu vực tỉnh Thái Nguyên (cũ). Sau đó, số lượng nhà ở bị ảnh hưởng bởi nguy cơ ngập lụt được định lượng bằng cách chồng xếp bản đồ nguy cơ ngập lụt tốt nhất với dữ liệu nhà ở trong môi trường GIS. Những kết quả đạt được trong nghiên cứu này không chỉ cung cấp cơ sở khoa học cho việc nhận diện và đánh giá nguy cơ ngập lụt, mà còn hỗ trợ hiệu quả cho công tác quy hoạch sử dụng đất, quản lý phát triển đô thị và xây dựng chiến lược giảm nhẹ thiên tai tại địa phương. Tuy nhiên, dữ liệu nhà được công bố vào tháng 05/2023 nên dẫn tới việc có thể bỏ sót những công trình nhà ở được xây dựng mới sau thời điểm này, vì vậy, nghiên cứu tiếp theo cần sử dụng bộ dữ liệu nhà có tính cập nhật hơn nhằm cải thiện độ chính xác và tính đầy đủ của đánh giá này.

Lời cảm ơn

Nghiên cứu này được tài trợ bởi Bộ Giáo dục và Đào tạo trong đề tài mã số B2024-XDA-06.

Tuyên bố về việc sử dụng trí tuệ nhân tạo tạo sinh và các công nghệ hỗ trợ bởi AI

Các tác giả xác nhận không sử dụng bất kỳ công cụ trí tuệ nhân tạo tạo sinh hoặc công nghệ hỗ trợ bởi AI nào trong quá trình chuẩn bị và hoàn thiện bản thảo.

Tài liệu tham khảo

[1] Chính, L. T. D., Nam, D. Đ., Lai, T. Đ., Quỳnh, B. D. (2021). Nghiên cứu xây dựng bản đồ hiểm họa lũ lụt sử dụng vết lũ lịch sử và mô hình độ cao số. *Tạp chí Khoa học Công nghệ Xây dựng (TCKHCNXD) - ĐHXDHN*, 15(3V):152–164.

[2] Razavi-Termeh, S. V., Pourzangbar, A., Sadeghi-Niaraki, A., Franca, M. J., Choi, S.-M. (2025). Metaheuristic-driven enhancement of categorical boosting algorithm for flood-prone areas mapping. *International Journal of Applied Earth Observation and Geoinformation*, 136:104357.

[3] Chinh, L. T. D., Hang, H. T. (2023). An analysis of the relative variable importance to flood fatality using a machine learning approach. *Journal of Science and Technology in Civil Engineering (JSTCE) - HUCE*, 17(1):125–136.

[4] J., S. R., J., M. (2024). Machine learning algorithms for natural disaster prediction and management. Satishkumar, D., Sivaraja, M., editors, *Utilizing AI and Machine Learning for Natural Disaster Management*, Hershey, PA, USA, IGI Global, 24–38.

[5] Ha, H., Bui, Q. D., Nguyen, V.-P., Tran, X. T., Nguyen, D. Q., Luu, C. (2025). Assessing housing exposure to landslide hazards using hybrid machine learning and spatial modeling. *Modeling Earth Systems and Environment*, 11(4):251.

[6] Seydi, S. T., Kanani-Sadat, Y., Hasanlou, M., Sahraei, R., Chanussot, J., Amani, M. (2023). Comparison of machine learning algorithms for flood susceptibility mapping. *Remote Sensing*, 15(1):192.

[7] Lee, S., Kim, J.-C., Jung, H.-S., Lee, M. J., Lee, S. (2017). Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. *Geomatics, Natural Hazards and Risk*, 8(2):1185–1203.

[8] Wu, Y., Zhang, Z., Qi, X., Hu, W., Si, S. (2024). Prediction of flood sensitivity based on Logistic Regression, eXtreme Gradient Boosting, and Random Forest modeling methods. *Water Science & Technology*, 89(10):2605–2624.

[9] Al-Aizari, A. R., Alzahrani, H., AlThuwaynee, O. F., Al-Masnay, Y. A., Ullah, K., Park, H.-J., Al-Areeq, N. M., Rahman, M., Hazaea, B. Y., Liu, X. (2024). Uncertainty reduction in flood susceptibility mapping using random forest and eXtreme gradient boosting algorithms in two tropical desert cities, Shibam and Marib, Yemen. *Remote Sensing*, 16(2):336.

[10] Kurugama, K. M., Kazama, S., Hiraga, Y., Samarasuriya, C. (2024). A comparative spatial analysis of flood susceptibility mapping using boosting machine learning algorithms in Rathnapura, Sri Lanka. *Journal of Flood Risk Management*, 17(2):e12980.

[11] Aydin, H. E., Iban, M. C. (2023). Predicting and analyzing flood susceptibility using boosting-based ensemble machine learning algorithms with SHapley Additive exPlanations. *Natural Hazards*, 116(3): 2957–2991.

[12] Ha, M. C., Vu, P. L., Nguyen, H. D., Hoang, T. P., Dang, D. D., Dinh, T. B. H., Şerban, G., Rus, I., Breţcan, P. (2022). Machine learning and remote sensing application for extreme climate evaluation: Example of flood susceptibility in the Hue Province, Central Vietnam region. *Water*, 14(10):1617.

[13] Phong, T. V., Nguyen, D. D., Pham, B. T. (2023). Modeling and mapping of flood susceptibility at Que Son District, Quang Nam Province, Vietnam using CatBoost. *IOP Conference Series: Materials Science and Engineering*, 1289(1):012019.

[14] Open Development Mekong (2025). *Thai Nguyen Portal – Disasters*. Truy cập ngày 22/10/2025.

[15] Mai Sy, H., Luu, C., Bui, Q. D., Ha, H., Nguyen, D. Q. (2023). Urban flood risk assessment using Sentinel-1 on the Google Earth Engine: A case study in Thai Nguyen city, Vietnam. *Remote Sensing Applications: Society and Environment*, 31:100987.

[16] Chính, L. T. D., Hằng, H. T., Linh, T. H., Quỳnh, B. D. (2022). Đánh giá ảnh hưởng của lũ lụt khu vực tỉnh Quảng Bình sử dụng Google Earth Engine và các phân tích không gian. *Tạp chí Khoa học Công nghệ Xây dựng (TCKHCNXD) - ĐHXDHN*, 16(5V):101–113.

[17] Arabameri, A., Pradhan, B., Rezaei, K., Yamani, M., Pourghasemi, H. R., Lombardo, L. (2018). Spatial modelling of gully erosion using evidential belief function, logistic regression, and a new ensemble of evidential belief function–logistic regression algorithm. *Land Degradation & Development*, 29(11):4035–4049.

[18] Bui, Q. D., Ha, H., Khuc, D. T., Nguyen, D. Q., von Meding, J., Nguyen, L. P., Luu, C. (2023). Landslide susceptibility prediction mapping with advanced ensemble models: Son La province, Vietnam. *Natural Hazards*, 116(2):2283–2309.

[19] Bentéjac, C., Csörgő, A., Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3):1937–1967.

[20] Anh, T. V., Khiên, H. T., Đông, K. T., Nghị, L. T., Hạnh, T. H., Phong, D. H. (2024). Nghiên cứu khả năng của mô hình học máy GB và SVR trong thành lập bản đồ nguy cơ lún đất khu vực bán đảo Cà Mau, Việt Nam. *Vietnam Journal of Hydrometeorology*, (757):60–73.

[21] Vu, Q.-H., Bui, Q. D., Vu, D. C., Luong, N.-D., Duong, C.-H., Luu, C., Ha, H. (2026). Metaheuristic-enhanced XGBoost models for improving flood and landslide susceptibility assessment. *Earth Systems and Environment*, 10(7):8371–8397.

[22] Carter, J. V., Pan, J., Rai, S. N., Galandiuk, S. (2016). ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery*, 159(6):1638–1645.

[23] Lee, Y.-G., Oh, J.-Y., Kim, D., Kim, G. (2023). SHAP value-based feature importance analysis for short-term load forecasting. *Journal of Electrical Engineering & Technology*, 18(1):579–588.

[24] Li, M., Sun, H., Huang, Y., Chen, H. (2024). Shapley value: From cooperative game to explainable artificial intelligence. *Autonomous Intelligent Systems*, 4:2.

[25] Vu, C. D., Ha, H. T., Luong, D. N., Bui, Q. D., Luu, C. D. T., Tran, H. H., Truong, T. X. (2025). Advanced machine learning models for flood susceptibility mapping: A case study in Thai Nguyen Province (Old), Vietnam. *Inżynieria Mineralna*, 1(2).