Journal of Mining and Earth Sciences

Website: https://jmes.humg.edu.vn

Forest fire risk prediction using geospatial data and machine learning techniques, a case study in the western region of Nghe An province

Phuong Nam Thi Doan [1,*], Hung Le Trinh [2], Trung Van Nguyen [1], Ha Thu Thi Le [1], Phu Van Le [2]

[1] *Hanoi University of Mining and Geology, Hanoi, Vietnam*
[2] *Le Quy Don Technical University, Hanoi, Vietnam*

ARTICLE INFO

Article history:
Received 27th Mar. 2024
Revised 28th July 2024
Accepted 26th Aug. 2024

Keywords:
Forest fire risk prediction model,
Geospatial data,
Machine learning,
Nghe An province.

ABSTRACT

Nghe An is the province with the largest area of forests and forestry land in the country with more than 1 million hectares of forest, coverage rate reaching 58,33%. Due to the influence of climate change and human activities, forest cover in Nghe An has profound fluctuations, of which forest fires are one of the main causes. This article presents the results of developing a forest fire risk prediction model in the western region of Nghe An province from geospatial data and machine learning algorithms. From the analysis of natural and social conditions in the study area, 9 input data layers include: (1) elevation, (2) slope, (3) aspect, (4) vegetation cover density, (5) population density, (6) land surface temperature, (7) evapotranspiration, (8) wind speed and (9) average monthly rainfall is used to build a forest fire risk prediction model. In the study, we tested with 02 machine learning algorithms, including Random Forest (RF) and Gradient Tree Boosting (GTB), then selected the appropriate algorithm by evaluating accuracy using the fire point data set as well as model performance. The obtained results showed that the AUC (Area Under the Curve) value of the GTB(350) algorithm reached 0,948, higher than the RF(100) (0,947). From this result, the study used the GTB algorithm with 350 trees to create a forest fire risk prediction map in the western region of Nghe An province.

[*] *Corresponding author*
E - mail: doanthinamphuong@humg.edu.vn
DOI: 10.46326/JMES.2024.65(5).06

Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất

Trang điện tử: https://tapchi.humg.edu.vn

Ứng dụng dữ liệu địa không gian và kỹ thuật học máy trong dự báo nguy cơ cháy rừng, thử nghiệm tại khu vực phía tây tỉnh Nghệ An

Đoàn Thị Nam Phương [1,*], Trịnh Lê Hùng [2], Nguyễn Văn Trung [1], Lê Thị Thu Hà [1], Lê Văn Phú [2]

[1] *Trường Đại học Mỏ Địa chất, Hà Nội, Việt Nam*
[2] *Trường Đại học Kỹ thuật Lê Quý Đôn, Hà Nội, Việt Nam*

THÔNG TIN BÀI BÁO

Quá trình:
Nhận bài 27/3/2024
Sửa xong 28/7/2024
Chấp nhận đăng 26/8/2024

Từ khóa:
Dữ liệu địa không gian,
Học máy,
Mô hình dự báo nguy cơ cháy rừng,
Tỉnh Nghệ An.

TÓM TẮT

Nghệ An là tỉnh có diện tích rừng và đất lâm nghiệp lớn nhất cả nước với hơn 1 triệu ha rừng, tỉ lệ che phủ đạt 58,33%. Do ảnh hưởng của biến đổi khí hậu và hoạt động của con người, lớp phủ rừng ở Nghệ An có sự biến động sâu sắc, trong đó cháy rừng là một trong những nguyên nhân chính. Bài báo này trình bày kết quả ứng dụng dữ liệu địa không gian và các kỹ thuật học máy nhằm dự báo nguy cơ cháy rừng khu vực phía Tây tỉnh Nghệ An. Từ phân tích điều kiện tự nhiên-xã hội khu vực nghiên cứu, chín lớp dữ liệu bao gồm: (1) độ cao, (2) độ dốc, (3) hướng sườn, (4) mật độ che phủ, (5) mật độ dân cư, (6) nhiệt độ bề mặt, (7) độ bốc thoát hơi nước, (8) tốc độ gió và (9) lượng mưa trung bình tháng được sử dụng để mô hình hóa nguy cơ cháy rừng. Trong nghiên cứu đã thử nghiệm với 02 thuật toán học máy khác nhau, bao gồm Random Forest (RF) và Gradient Tree Boosting (GTB), từ đó lựa chọn thuật toán phù hợp thông qua đánh giá độ chính xác bằng bộ dữ liệu điểm cháy cũng như hiệu suất mô hình. Kết quả nhận được cho thấy, giá trị AUC (Area Under the Curve) của thuật toán GTB(350) đạt 0,948, cao hơn so với thuật toán RF(100) (0,947). Từ kết quả này, trong nghiên cứu đã sử dụng thuật toán GTB với số lượng cây 350 để thành lập bản đồ dự báo nguy cơ cháy rừng khu vực phía tây tỉnh Nghệ An.

[*] *Tác giả liên hệ*
E - mail: doanthinamphuong@humg.edu.vn
DOI: 10.46326/JMES.2024.65(5).06

1. Mở đầu

Theo công bố hiện trạng rừng năm 2023, Việt Nam có diện tích rừng trên 14 triệu ha, tỉ lệ che phủ đạt 42,02% (MARD, 2024). Mặc dù tổng diện tích rừng và tỉ lệ che phủ ở Việt Nam có xu hướng gia tăng trong giai đoạn gần đây, tuy nhiên diện tích rừng gia tăng chủ yếu là rừng trồng, trong khi rừng tự nhiên tiếp tục suy giảm. Có nhiều nguyên nhân khác nhau gây nên sự suy giảm diện tích và chất lượng rừng ở Việt Nam, trong đó cháy rừng là một trong những nguyên nhân quan trọng (Trần, 2017; Nguyễn và nnk., 2017). Cháy rừng là một hiện tượng phức tạp bị ảnh hưởng bởi sự kết hợp của các yếu tố. Thời tiết và các hoạt động của con người là các yếu tố chính góp phần vào việc xảy ra và lan rộng của cháy rừng. Bên cạnh đó, biến đổi khí hậu và sự thay đổi về địa hình có thể ảnh hưởng đến hành vi và đẩy nhanh quá trình cháy (Hoang và nnk., 2020).

Lớp phủ rừng thường phân bố ở các khu vực có địa hình phức tạp, khó tiếp cận, do vậy các phương pháp truyền thống có nhiều hạn chế trong phát hiện và dự báo nguy cơ cháy rừng. Những hạn chế này có thể được khắc phục khi sử dụng công nghệ địa không gian, trong đó chủ đạo là dữ liệu viễn thám và GIS. Một số hệ thống cảnh báo cháy rừng từ dữ liệu viễn thám và GIS được áp dụng như Hệ thống thông tin cháy rừng châu Âu (EFFIS) sử dụng dữ liệu ảnh MODIS, hệ thống INPAS (Croatia) sử dụng kết hợp đa nguồn dữ liệu (video, dữ liệu khí tượng, dữ liệu GIS). Với sự phát triển mạnh mẽ của trí tuệ nhân tạo, thời gian gần đây nhiều nghiên cứu đã sử dụng các mô hình học máy, học sâu như mạng neural nhân tạo, random forest, support vector machine,... để nâng cao độ chính xác kết quả dự báo nguy cơ cháy rừng (Vasilakos và nnk., 2009; Oliveira và nnk., 2012; Dieu và nnk., 2016; Enod và nnk., 2021; Iban và Sekertekin, 2022; Nguyen và nnk, 2018). Các kỹ thuật hồi quy như hồi quy đa biến (multiple regression) (Oliveira và nnk., 2012), hồi quy logistic (Pourghasemi, 2015), hồi quy trọng số địa lý (geographically weighted regression - GWR) (Fernandez và nnk., 2012), kỹ thuật khai phá dữ liệu (data mining) (Arpaci và nnk., 2014), các mô hình tổng quát (GLMs, GAMs) (Ruano và nnk., 2022) cũng được sử dụng để đánh giá và dự báo nguy cơ cháy rừng từ bộ dữ liệu đầu vào đại diện cho các yếu tố tự nhiên và kinh tế - xã hội.

Các nghiên cứu dự báo nguy cơ cháy rừng ở Việt Nam được thực hiện từ những năm cuối thế kỷ XX trên cơ sở sử dụng các chỉ số tổng hợp, chủ yếu là chỉ số Nesterov - chỉ số P (Phạm, 1988; Võ, 1995). Phương pháp truyền thống này tiếp tục được sử dụng trong các nghiên cứu thời gian sau, trong đó mô hình dự báo được bổ sung thêm các lớp thông tin đầu vào cũng như điều chỉnh giá trị chỉ tiêu P trong phân cấp nguy cơ cháy rừng (Lê và Vương, 2014; Nguyễn, 2019). Thời gian gần đây, một số nghiên cứu đã kết hợp dữ liệu viễn thám, GIS và các mô hình học máy để nâng cao độ chính xác kết quả dự báo nguy cơ cháy rừng (Đặng và nnk,, 2017 Hoang và nnk., 2020). Các nghiên cứu này đã sử dụng một số mô hình học máy như RF, SVM và Classification and Regression Tree (CART) để dự báo nguy cơ cháy rừng trên cơ sở xác suất xảy ra cháy của từng điểm ảnh. Kết quả nhận được cho thấy, thuật toán RF có độ chính xác cao nhất trong dự báo nguy cơ cháy rừng, trong khi đó thuật toán CART có độ chính xác thấp nhất (Đoàn, 2023).

Bài báo này trình bày kết quả xây dựng bản đồ dự báo nguy cơ cháy rừng khu vực phía tây tỉnh Nghệ An sử dụng dữ liệu địa không gian và kỹ thuật học máy. Ba thuật toán học máy thông dụng, đã được chứng minh hiệu quả trong các nghiên cứu khác nhau bao gồm RF, SVM và GTB được thử nghiệm để lựa chọn thuật toán phù hợp với điều kiện cụ thể khu vực nghiên cứu. Bộ dữ liệu đầu vào bao gồm các lớp đại diện cho yếu tố địa hình, lớp phủ, khí hậu, điều kiện kinh tế-xã hội được lựa chọn và xây dựng từ dữ liệu viễn thám, GIS và các cơ sở dữ liệu quốc tế. Độ chính xác kết quả dự báo nguy cơ cháy rừng được đánh giá thông qua bộ dữ liệu điểm cháy trong quá khứ cũng như đánh giá hiệu năng của mô hình.

2. Dữ liệu và phương pháp nghiên cứu

2.1. Dữ liệu và khu vực nghiên cứu

Ảnh Sentinel 2 MSI chụp trong khoảng thời gian từ 15/11/2021 đến 16/01/2022 được sử dụng để tạo ảnh không mây và xây dựng lớp dữ liệu về mật độ che phủ. Trong khi đó, ảnh Landsat 8 trong cùng thời gian trên được sử dụng để tính nhiệt độ bề mặt.

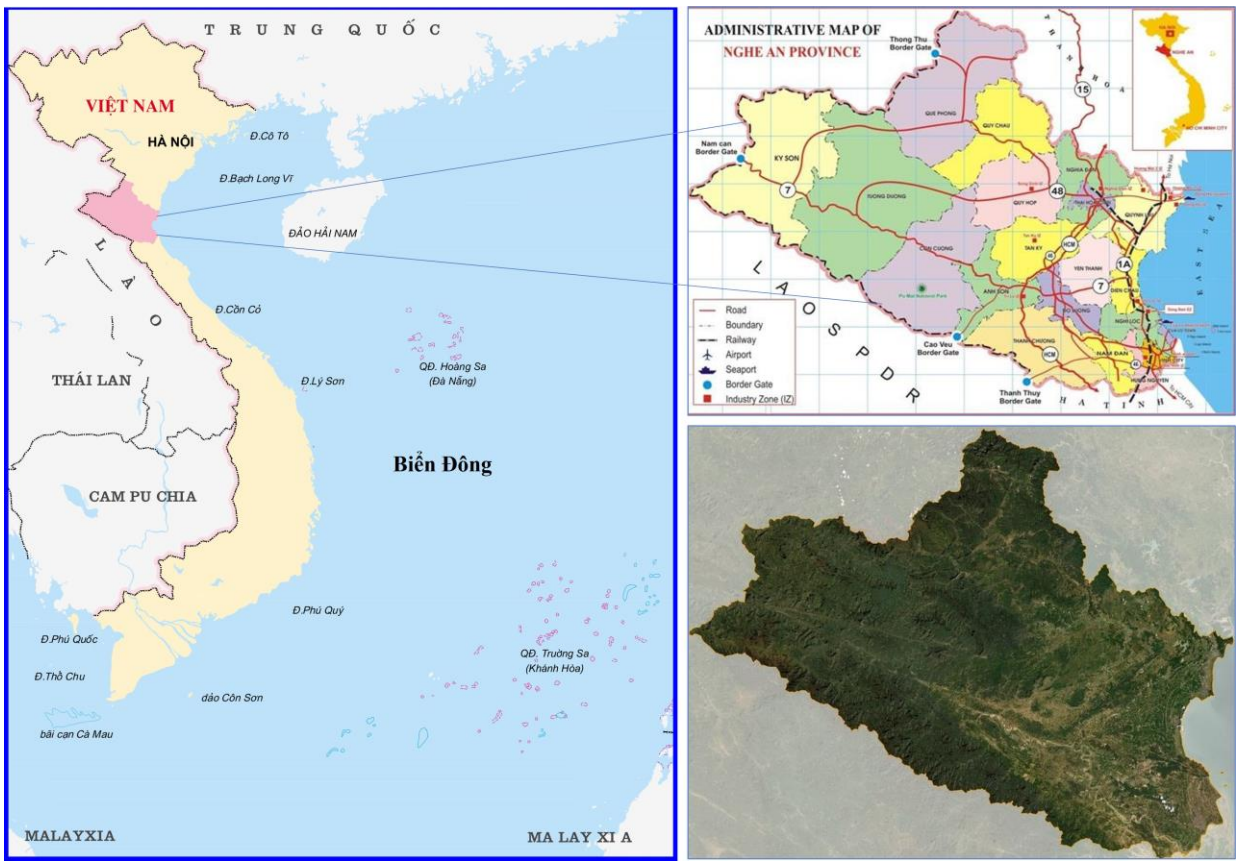
Dữ liệu DEM toàn cầu SRTM với độ phân giải không gian 30 m cung cấp bởi Cơ quan Hàng không vũ trụ Mỹ (NASA) được sử dụng để chiết

xuất thông tin các yếu tố địa hình như độ dốc, độ cao, hướng sườn.

Bên cạnh đó, dữ liệu về khí hậu và thời tiết, bao gồm: tốc độ gió, lượng mưa trung bình tháng, độ bốc thoát hơi nước cũng được sử dụng trong nghiên cứu. Các lớp dữ liệu này được thu thập từ cơ sở dữ liệu quốc tế về khí hậu WorldClim (https://www.worldclim.org/). Lớp dữ liệu Mật độ dân cư được thu thập và xây dựng từ cơ sở dữ liệu nhân khẩu học WorldPop (https://data.worldpop.org/). Các lớp dữ liệu được thu thập tương ứng trong giai đoạn thu thập dữ liệu điểm cháy rừng.

Bộ dữ liệu điểm cháy trong quá khứ được thu thập trong giai đoạn 2018÷2023 dựa trên hồ sơ cháy của Cục Kiểm lâm (Bộ Nông nghiệp và Phát triển Nông thôn) và dữ liệu hỏa hoạn từ hệ thống FIRMS (NASA). Bên cạnh đó, trong nghiên cứu cũng xây dựng bộ dữ liệu các điểm «không cháy» được lấy theo các vị trí có mật độ cây thấp, mặt nước, ít hoạt động của con người để huấn luyện và kiểm định mô hình dự báo.

Khu vực nghiên cứu được lựa chọn là địa bàn phía tây tỉnh Nghệ An bao gồm các huyện Con Cuông, Kỳ Sơn, Quế Phong, Quỳ Châu, Quỳ Hợp và Tương Dương (Hình 1). Theo báo cáo hiện trạng rừng Việt Nam năm 2023 của Bộ Nông nghiệp và Phát triển nông thôn, Nghệ An có diện tích rừng và đất lâm nghiệp lớn nhất cả nước, chiếm gần 70% diện tích đất tự nhiên toàn tỉnh, mật độ che phủ rừng đạt 58,36%. Nghệ An cũng là một trong các địa phương đứng đầu trên toàn quốc về nguy cơ cháy rừng ở cấp V, cấp cực kỳ nguy hiểm. Phía tây tỉnh Nghệ An nằm trong vùng núi và có đặc điểm địa hình đa dạng, bao gồm khu vực núi cao, thung lũng và sông suối. Do đặc điểm địa hình phức tạp và thảm phủ rừng dày đặc, phía tây tỉnh Nghệ An có nguy cơ cháy rừng cao trong mùa khô. Một số vụ cháy rừng điển hình như cháy rừng tại xã Châu Bính, huyện Quỳ Châu (14/04/2022) đã gây thiệt hại hơn 500 ha rừng, cháy rừng tại xã Mường Típ, huyện Kỳ Sơn (25/04/2022) đã gây thiệt hại hơn 200 ha rừng, cháy rừng tại xã Lục Dạ, huyện Con Cuông (05/05/2022) gây thiệt hại hơn 100 ha rừng.



Hình 1. Mô tả vị trí khu vực nghiên cứu.

2.2. Phương pháp nghiên cứu

a) Xây dựng bộ dữ liệu đầu vào và chuẩn hoá dữ liệu

Để lựa chọn và xây dựng bộ dữ liệu đầu vào của mô hình, trong bài báo tiến hành phân tích, đánh giá ảnh hưởng của các điều kiện tự nhiên, kinh tế - xã hội đến nguy cơ xảy ra cháy rừng. Phân tích đặc điểm cháy rừng ở khu vực nghiên cứu cho thấy, nguyên nhân chính của tình trạng này là do ảnh hưởng của điều kiện tự nhiên (đặc điểm lớp phủ, các yếu tố thời tiết, khí hậu, địa hình) cũng như hoạt động của con người (du lịch rừng, đốt nương làm rẫy, sự mở rộng đô thị và nông thôn). Từ kết quả này, trong bài báo đã lựa chọn chín lớp dữ liệu để lựa chọn xây dựng mô hình dự báo nguy cơ cháy rừng, bao gồm: (1) độ cao, (2) độ dốc, (3) hướng sườn, (4) mật độ che phủ thực vật, (5) mật độ dân cư, (6) nhiệt độ bề mặt, (7) độ bốc thoái hơi nước bề mặt, (8) tốc độ gió, (9) lượng mưa trung bình tháng.

Các lớp dữ liệu về địa hình (độ cao, độ dốc, hướng sườn) được xây dựng từ DEM SRTM (30m). Các lớp dữ liệu về thời tiết, khí hậu, dân cư được thu thập từ cơ sở dữ liệu WorldClim và WorldPop. Nhiệt độ bề mặt được xác định từ ảnh vệ tinh Landsat 8 theo phương pháp do NASA cung cấp (Landsat 8 data users handbook). Trong khi đó, mật độ che phủ được xác định thông qua chỉ số thực vật NDVI theo công thức (1) (Trinh and Zablotskii, 2017):

$$Pv = \frac{NDVI - NDVI_{\min}}{NDVI_{\max} - NDVI_{\min}} \quad (1)$$

Do các lớp dữ liệu đầu vào được xây dựng từ các nguồn khác nhau với thang đo không thống nhất, để có thể đưa vào mô hình dự báo, tất cả các lớp dữ liệu này được chuyển đổi sang định dạng raster với độ phân giải 10 m (phù hợp với độ phân giải không gian của ảnh Sentinel 2 MSI). Tiếp theo, các lớp dữ liệu được chuẩn hóa về phạm vi [0÷1] theo công thức (2) (Dieu và nnk., 2012):

$$N_v = \frac{Fa_i - Min(Fa)}{Max(Fa) - Min(Fa)} \times [0,99 - 0,01] + 0,01 \quad (2)$$

Trong đó: Fa_i - giá trị của hệ số được xem xét, Min(Fa) và Max(Fa) là giá trị tối thiểu và giá trị tối đa của hệ số được xem xét; N_v - giá trị tính toán mới cho hệ số được xem xét.

Mức độ quan trọng của từng yếu tố đầu được xác định trên cơ sở hệ số tương quan Pearson (hệ số *r*). Hệ số (*r*) có giá trị trong khoảng từ -1÷1, trong đó giá trị r dương thể hiện tương quan thuận, r âm thể hiện tương quan nghịch. Trong nghiên cứu này, hệ số *r* được xác định bằng phần mềm QGIS 2.18.

b) Lựa chọn thuật toán học máy

Để dự báo nguy cơ cháy rừng khu vực phía tây tỉnh Nghệ An từ bộ dữ liệu đầu vào bao gồm chín yếu tố, trong bài báo thử nghiệm với một số thuật toán học máy thông dụng như Random Forest (RF), Support Vector Machine (S SVM VM) và Gradient Tree Boosting (GTB).

Random Forest

RF là một thuật toán học máy có giám sát được sử dụng phổ biến trong hồi quy và phân loại, đồng thời tạo ra kết quả phân loại tốt ngay cả khi không điều chỉnh bộ siêu tham số. RF hoạt động trên cơ sở xây dựng nhiều cây quyết định (decision tree) trên các mẫu huấn luyện, mỗi cây quyết định sẽ khác nhau (có yếu tố random). Ở bước tiếp theo, đối với mỗi cây quyết định sẽ đi từ trên xuống theo các nút điều kiện để được các dự đoán, sau đó kết quả cuối cùng được tổng hợp từ kết quả của các cây quyết định. Như vậy, RF lấy ngẫu nhiên dữ liệu và thuộc tính để xây dựng cây quyết định (Breiman, 2001).

Gradient Tree Boosting

GTB là một thuật toán học máy kết hợp sức mạnh của cây quyết định với kỹ thuật tối ưu hóa giảm dần độ dốc. Đây là một thuật toán linh hoạt và mạnh mẽ được sử dụng rộng rãi cho cả bài toán phân loại và hồi quy (Friedman, 2001).

GTB hoạt động bằng cách xây dựng một tập hợp các cây quyết định, trong đó mỗi cây được huấn luyện để cải thiện dự đoán của các cây trước đó. Thuật toán sử dụng kỹ thuật tối ưu hóa giảm độ dốc để giảm thiểu hàm mất mát, đây là thước đo sai số giữa giá trị dự đoán và giá trị thực.

Về mặt toán học, giả sử *X* và *Y* lần lượt là đầu vào và mục tiêu của *N* mẫu. Nghiên cứu cần xây dựng một hàm *f(x)* ánh xạ các đặc trưng đầu vào *X* tới các biến mục tiêu *y*. Hàm mất mát được xác định là sự khác biệt giữa các biến thực tế và dự đoán theo công thức (3) (Friedman, 2001; Sharma and Ghosh, 2023):

$$L(f) = \sum_{i=1}^{N} L\left(y_i, f\left(x_i\right)\right) \quad (3)$$

Mục tiêu của bài toán là muốn cực tiểu hoá hàm mất mát *L(f)* đối với *f* như công thức (4):

$$f_0(x) = \arg\min_f \left(L\left(f\right)\right) = \arg\min_f \left(\sum_{i=1}^{N} L\left(y_i, f\left(x_i\right)\right)\right) \quad (4)$$

Trong quá trình thực thi cây tăng cường độ dốc, tại bước thứ *m*, để cải thiện giá trị của f_m, thuật toán sẽ tích hợp các giá trị ước tính h_m (công thức (5))

$$y_i = f_{m+1}\left(x_i\right) = f_m\left(x_i\right) + h_m\left(x_i\right) \quad (5)$$

Đối với việc tăng cường độ dốc ở bước thứ *m*, độ dốc cao nhất tìm thấy là $h_m = -\rho_m g_m$. Trong đó ρ_m là giá trị không đổi được gọi là độ dài bước, g_m là độ dốc của hàm mất mát *L(f)*. g_m được xác định theo công thức (6):

$$g_{im} = -\left[\frac{\partial L\left(y_i, f\left(x_i\right)\right)}{\partial f\left(x_i\right)}\right]_{f(x_i)=f_{m-1}(x_i)} \quad (6)$$

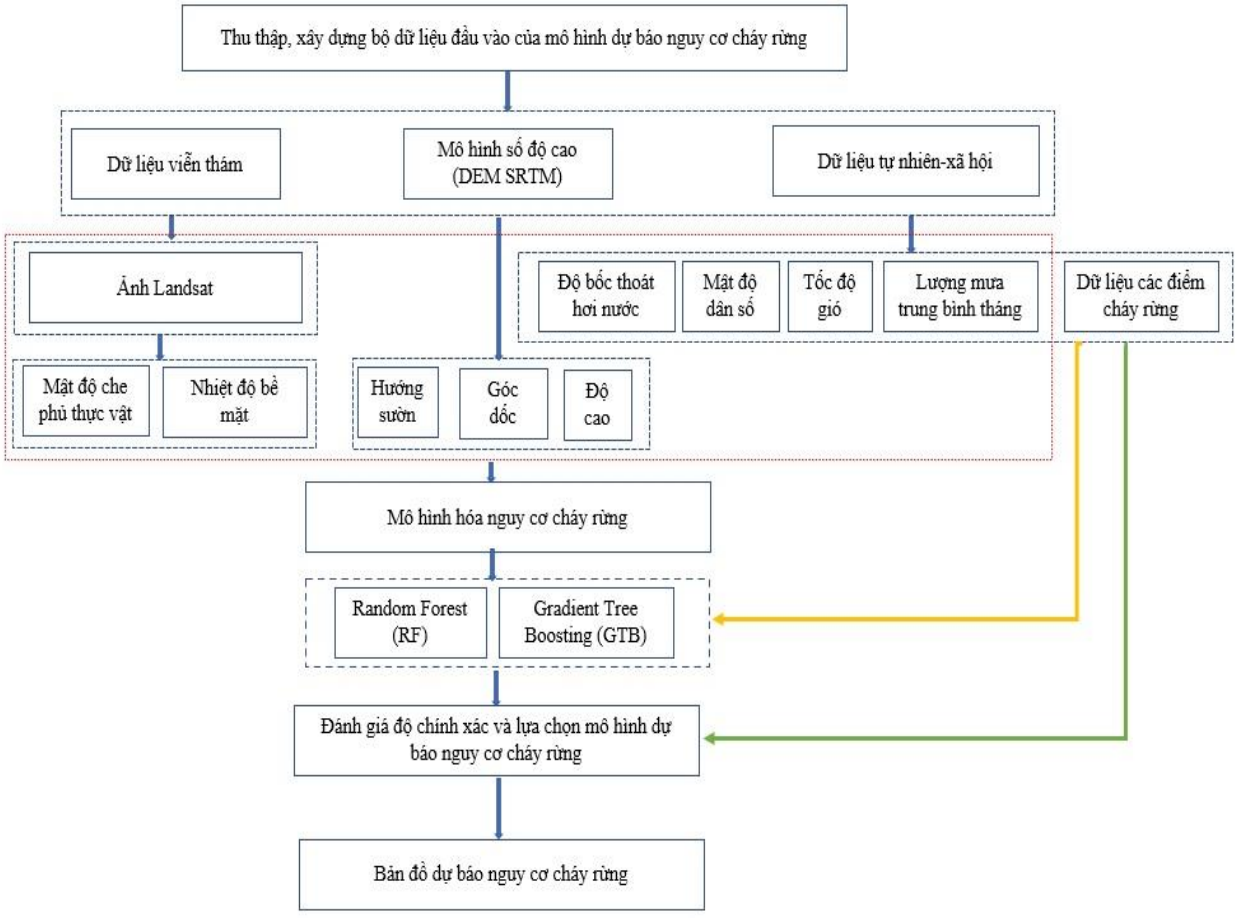
Thuật toán sẽ chạy cho tới khi giá trị h_m trở nên quá nhỏ. Khi đó, hàm *f(x)* được xây dựng và phù hợp với mô hình.

Sơ đồ quy trình dự báo nguy cơ cháy rừng từ dữ liệu địa không gian và mô hình học máy được mô tả trên Hình 2, bao gồm:

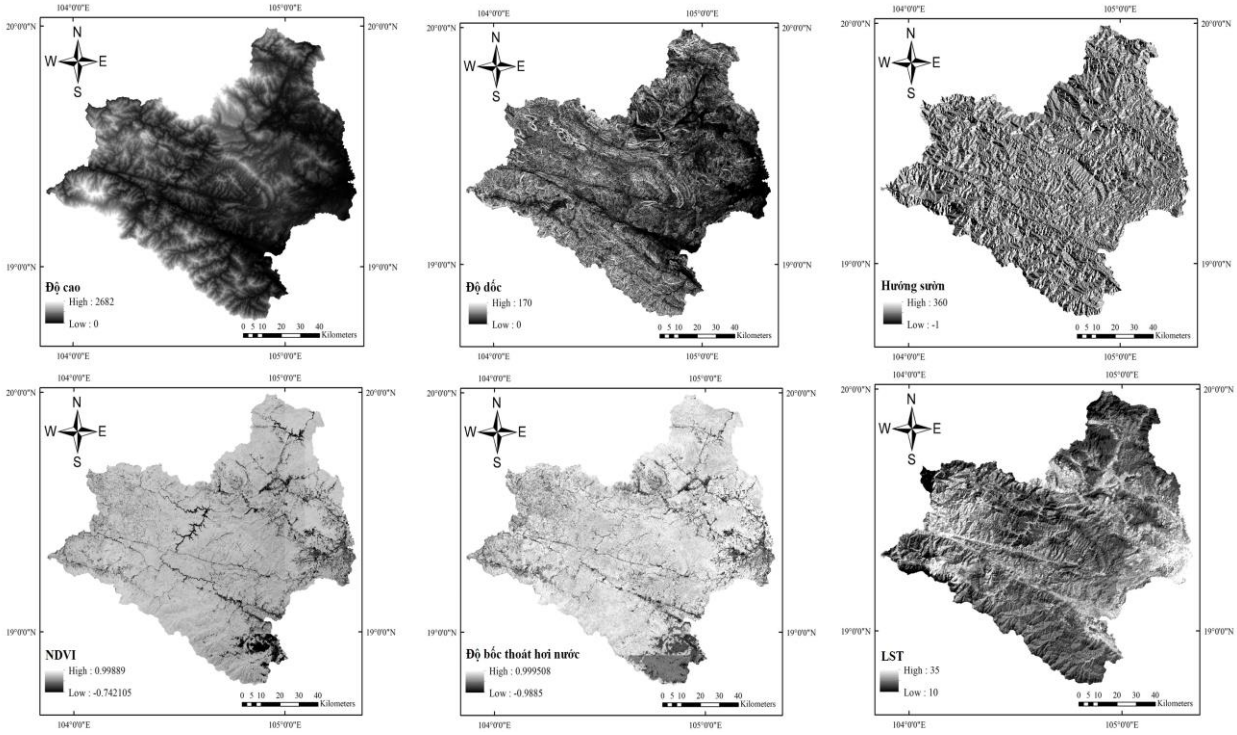
Bước 1: Nghiên cứu xây dựng bộ dữ liệu đầu vào và chuẩn hoá dữ liệu. Đồng thời, dữ liệu điểm cháy và không cháy được thu thập. Các dữ liệu mẫu được chia thành bộ dữ liệu huấn luyện và bộ dữ liệu kiểm tra.

Bước 2: Xây dựng mô hình nguy cơ cháy rừng với các thuật toán RF và GTB. Thiết lập các tham số ban đầu cho thuật toán.

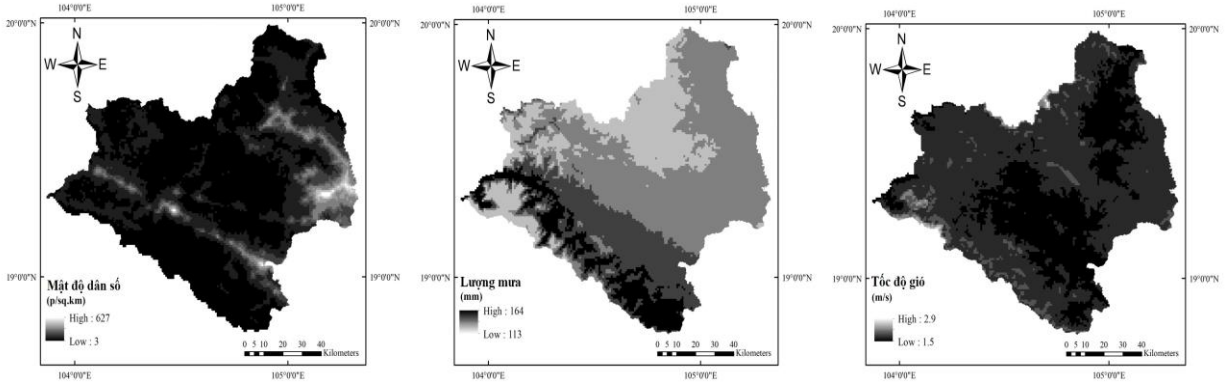
Bước 3: Thay đổi các tham số và đánh giá độ chính xác của tường mô hình. Lựa chọn mô hình có độ chính xác cao nhất làm mô hình tối ưu.



Hình 2. Sơ đồ mô hình dự báo nguy cơ cháy rừng.



Hình 3. Các lớp dữ liệu độ cao, độ dốc, hướng sườn, mật độ che phủ, độ bốc thoát hơi nước và nhiệt độ bề mặt khu vực nghiên cứu.



Hình 4: Các lớp dữ liệu mật độ dân cư, lượng mưa trung bình tháng và tốc độ gió khu vực nghiên cứu.

Bước 4: Sử dụng mô hình tối ưu, xây dựng lớp dữ liệu nguy cơ cháy rừng và biên tập bản đồ.

3. Kết quả và thảo luận

Trên Hình 3 và 4 trình bày kết quả xây dựng bộ dữ liệu đầu vào cho mô hình dự báo nguy cơ cháy rừng khu vực phía tây tỉnh Nghệ An, bao gồm chín lớp dữ liệu: độ cao, độ dốc, hướng sườn, mật độ che phủ, độ bốc thoát hơi nước bề mặt, nhiệt độ bề mặt (Hình 3) và mật độ dân cư, lượng mưa trung bình tháng, tốc độ gió (Hình 4). Các lớp dữ liệu này được nội suy về cùng độ phân giải không gian 10 m và chuẩn hóa về thang giá trị [0÷1] theo công thức (2).

Sau khi xây dựng các lớp thông tin đầu vào cho mô hình, trong nghiên cứu tiến hành xác định hệ số tương quan Pearson giữa các lớp dữ liệu đầu vào và bộ dữ liệu điểm cháy trong quá khứ để đánh giá mức độ quan trọng của từng yếu tố đối với mô hình dự báo. Bảng 1 trình bày giá trị tương quan so với dữ liệu dự báo cháy rừng. Kết quả nhận được cho thấy, lớp dữ liệu nhiệt độ bề mặt

có ảnh hưởng lớn nhất đến nguy cơ cháy rừng, thể hiện qua hệ số tương quan r đạt 0,582. Mật độ che phủ của thực vật, độ bốc thoát hơi nước bề mặt, độ dốc và lượng mưa trung bình tháng có tương quan ở mức trung bình với bộ dữ liệu điểm cháy (hệ số r từ 0,2 đến trên 0,4). Trong khi đó, lớp dữ liệu về tốc độ gió, hướng sườn và mật độ dân cư có mức độ quan trọng thấp hơn, thể hiện qua giá trị hệ số tương quan r xấp xỉ 0,1. Mặc dù có tương quan không cao với bộ dữ liệu điểm cháy trong quá khứ ở khu vực nghiên cứu, tuy nhiên với đặc thù khu vực thử nghiệm có địa hình dốc, chịu ảnh hưởng của gió phơn Tây Nam (gió Lào) cũng như nhiều vụ cháy rừng ở khu vực phía tây tỉnh Nghệ An có nguyên nhân từ hoạt động của con người, các lớp dữ liệu trên vẫn được lựa chọn để đưa vào các mô hình dự báo nguy cơ cháy rừng.

Bảng 1. Giá trị tương quan so với dữ liệu dự báo cháy rừng.

Lớp dữ liệu bổ sung	Giá trị
Hướng sườn	0,052
Mật độ dân số	0,089
Độ cao	-0,175
Tốc độ gió	0,098
Lượng mưa	-0,441
Độ dốc	-0,288
Độ bốc thoát hơi nước	0,257
Mật độ thực vật	0,273
Nhiệt độ	0,582

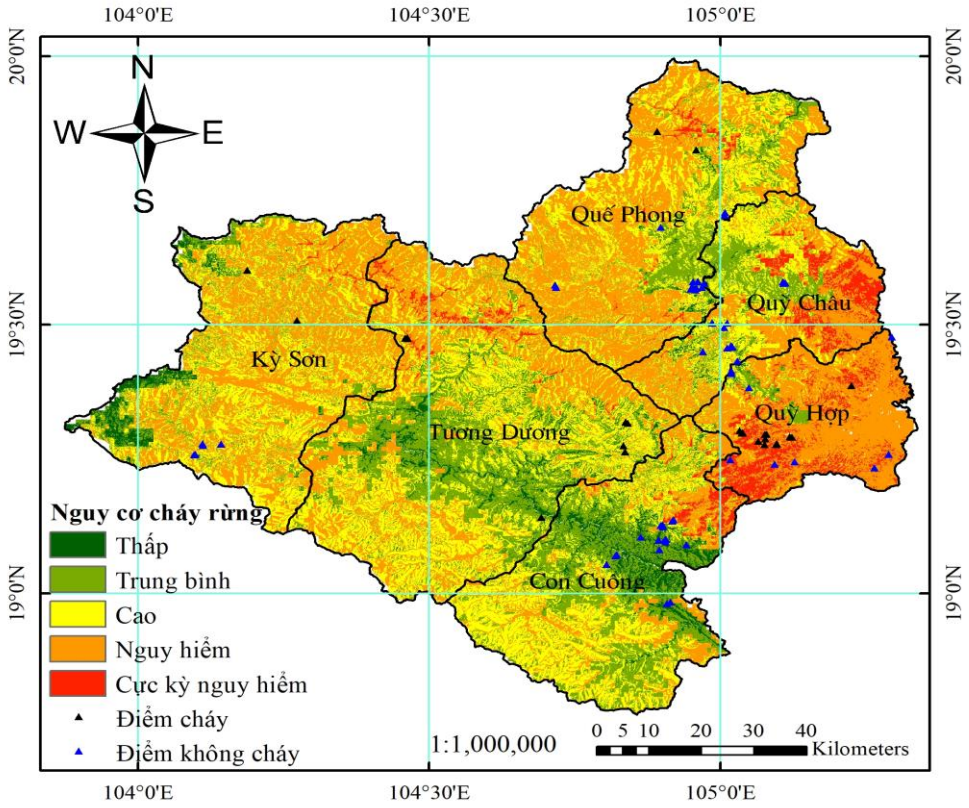
Bộ dữ liệu điểm cháy bao gồm 324 điểm được thu thập từ cơ sở dữ liệu của Cục Kiểm lâm, Bộ Nông nghiệp và Phát triển Nông thôn và các cơ sở dữ liệu quốc tế khác. Bộ dữ liệu này được chia theo tỉ lệ 7:3, trong đó 70% số điểm cháy (227 điểm) được sử dụng để huấn luyện mô hình, 30% số điểm (97 điểm, bao gồm 36 vị trí xảy ra cháy và 61 vị trí có báo cháy nhưng không xảy ra cháy) được sử dụng để đánh giá độ chính xác của mô hình. Bộ tham số đối với thuật toán học máy RF và GTB được thử nghiệm với nhiều giá trị khác nhau để lựa chọn tham số có độ chính xác cao nhất. Với thuật toán GTB, các tham số tỉ lệ giảm kích thước (shrinkage), tỉ lệ lấy mẫu (samplingRate), nhân (seed) được đặt cố định với giá trị lần lượt là 0,005, 0,7 và 0. Kết quả nhận được cho thấy, thuật toán GTB với số lượng cây 350 (GTB(350)) có độ chính xác cao nhất. Trong số 36 vị trí xảy ra cháy rừng trong quá khứ, có 27 điểm được phân bố vào khu vực có cấp độ dự báo « nguy hiểm » và « cực kỳ nguy hiểm », tương đương 75%. Không có điểm cháy nào phân bố ở khu vực có cấp độ dự báo « thấp », trong khi 9/36 điểm phân bố ở khu vực có cấp độ dự báo cháy rừng « trung bình » và « cao » (tương đương 25%).

Với thuật toán RF, tham số số lượng cây (numberOfTrees) là 100 (RF(100)) có độ chính xác cao nhất, thể hiện qua vị trí phân bố dữ liệu các điểm cháy, không cháy và kết quả đánh giá bằng đường cong ROC. Để lựa chọn mô hình phù hợp với điều kiện cụ thể khu vực thử nghiệm, trong nghiên cứu đã tiến hành so sánh hiệu suất của mô hình GTB(350) và các mô hình RF(100) và SVM(25). Bộ dữ liệu huấn luyện và bộ dữ liệu kiểm tra được so sánh, đánh giá thông qua giá trị AUC (Area under Curve). AUC là một kỹ thuật đánh giá hiệu quả của mô hình hồi quy trong phân loại dữ liệu. Giá trị AUC thay đổi từ 0 đến 1, trong đó giá trị AUC càng lớn thì mô hình phân loại dữ liệu càng tốt. Kết quả nhận được cho thấy, giá trị AUC của thuật toán GTB(350) đạt 0,948, tương đương với thuật toán RF(100) (0,947) và cao hơn đáng kể so với thuật toán SVM(25) (0,756).

Từ kết quả này, trong nghiên cứu đã lựa chọn thuật toán GTB(350) để dự báo nguy cơ cháy rừng khu vực thực nghiệm. Bản đồ dự báo nguy cơ cháy rừng bằng thuật toán GTB(350) được thể hiện trên Hình 5, trong đó nguy cơ cháy rừng được chia thành 5 mức độ theo quy định của Điều 46 Nghị định 156/2018/NĐ-CP, bao gồm thấp, trung bình, cao, nguy hiểm và cực kỳ nguy hiểm.

Phân tích các kết quả thu được cho thấy, gần một nửa diện tích khu vực nghiên cứu có mức độ nguy hiểm cháy rừng «nguy hiểm» và «cực kỳ nguy hiểm», tương ứng với 40,38% và 4,06% tổng diện tích khu vực nghiên cứu (thể hiện bởi màu cam và màu đỏ trên Hình 5). Những khu vực có nguy cơ cháy rừng «cực kỳ nguy hiểm» phân bố tập trung ở những vùng gần khu vực dân cư, lớp phủ rừng chủ yếu là rừng thứ sinh và rừng trồng.

Diện tích có nguy cơ cháy rừng «thấp» chỉ chiếm 3,94% tổng diện tích, phân bố chủ yếu ở phía Đông Nam tỉnh Nghệ An. 19,91% tổng diện tích khu vực nghiên cứu có nguy cơ cháy rừng ở mức độ «trung bình». Các khu vực có nguy cơ cháy rừng «cao» chiếm 31,71% tổng diện tích và phân bố đều trên toàn bộ khu vực nghiên cứu.



Hình 5. Kết quả xây dựng bản đồ dự báo nguy cơ cháy rừng khu vực phía tây tỉnh Nghệ An bằng mô hình học máy GTB(350).

4. Kết luận

Trong nghiên cứu này, dữ liệu địa không gian và ba thuật toán học máy (RF, GTB) được sử dụng để dự báo nguy cơ cháy rừng cho khu vực phía tây tỉnh Nghệ An. Chín yếu tố ảnh hưởng tới nguy cơ cháy rừng, bao gồm: (1) độ cao, (2) độ dốc, (3) hướng sườn, (4) mật độ che phủ thực vật, (5) mật độ dân cư, (6) nhiệt độ bề mặt, (7) độ bốc thoái hơi nước bề mặt, (8) tốc độ gió, (9) lượng mưa trung bình tháng được lựa chọn để xây dựng bộ dữ liệu đầu vào cho mô hình. Kết quả nhận được cho thấy, mô hình GTB với số lượng cây 350 (GTB350) có độ chính xác cao nhất khi so sánh với bộ dữ liệu điểm cháy trong quá khứ. Ngoài ra, GTB(350) cũng có hiệu suất cao khi so sánh với các mô hình học máy khác như RF và SVM thông qua giá trị AUC. Từ kết quả này, trong nghiên cứu đã xây dựng được bản đồ dự báo nguy cơ cháy rừng khu vực phía tây tỉnh Nghệ An với năm cấp độ khác nhau.

Kết quả nhận được trong nghiên cứu có thể sử dụng nhằm cung cấp thông tin giúp các nhà quản lý trong theo dõi, ứng phó và giảm thiểu thiệt hại do cháy rừng gây ra.

Lời cảm ơn

Bài báo có sử dụng một phần số liệu và kết quả của đề tài KH&CN cấp cơ sở "*Nghiên cứu mô hình dự báo nguy cơ cháy rừng bằng công nghệ Địa không gian, thử nghiệm cho khu vực phía Tây tỉnh Nghệ An*", mã số T23-38. Các tác giả xin chân thành cảm ơn Trường Đại học Mỏ Địa chất và Ban chủ nhiệm đề tài đã hỗ trợ nhóm nghiên cứu hoàn thành bài báo này.

Đóng góp của các tác giả

Đoàn Thị Nam Phương, Trịnh Lê Hùng - lên ý tưởng, viết bản thảo bài báo; Nguyễn Văn Trung, Lê Thị Thu Hà - đánh giá và chỉnh sửa; Lê Văn Phú - thu thập và xử lý dữ liệu.

Tài liệu tham khảo

Arpaci A., Malowerschnig, B., Sass, O., Vacik, H. (2014). Using multivariate data mining

techniques for estimating fire susceptibility of Tyrolean forests, *Applied Geography*, 53, 258 - 270.

Breiman, L. (2001). Random Forests, *Machine Learning* 45, 5-32, https://doi.org/10.1023/A:1010933404324.

Dieu, T. B., Pradhan, B., Lofman, O., Revhaug, I., Dick, O. B. (2012). Spatial prediction of landslide hazards in Hoa Binh province (Vietnam): A comparative assessment of the efficacy of evidential belief functions and fuzzy logic models, *CATENA* 96, 28-40.

Dieu, T. B., Thoa, L. T. K., Van, N. C., Duc, L. H., Revhaug, I. (2016). Tropical forest fire susceptibility mapping at the Cat Ba national park area, Hai Phong city, Vietnam, using GIS-based kernel logistic regression, *Remote Sensing*, 8, 347, doi:10.3390/rs8040347.

Đoàn, T. N. P. (2023). Lựa chọn mô hình dự báo nguy cơ cháy rừng từ dữ liệu viễn thám và hệ thông tin địa lý, Luận án tiến sĩ kỹ thuật, Hà Nội.

Đặng, N.B.T., Nguyễn, N.T., Phạm, X.C. (2017). Ứng dụng viễn thám và GIS thành lập bản đồ nguy cơ cháy rừng phục vụ phòng chống, giảm thiểu thiệt hại do cháy rừng tại tỉnh Sơn La, Việt Nam, *Hội thảo GIS toàn quốc*, tr.252-261.

Enoh, M., Okeke, U., Narinua, N. (2021). Identification and modelling of forest fire severity and risk zones in the Cross - Niger transition forest with remotely sensed satellite data, *The Egyptian Journal of Remote Sensing and Space Science, 24*(3), 879 - 887.

Fernandez, J., Chuvieco, E., Koutsias, N. (2012). Modelling long-term fire occurrence factors in Spain by accounting for local variations with geographically weighted regression, *Natural Hazards Earth System Sciences*, 12, 1-17.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine, *Annals of Statistics, 29*(5), 1189-1232.

Hoang, V. T., Chou, T., Fang, Y., Nguyen, N. T., Nguyen, Q. H., Pham, X. C., Dang, N. B. T., Nguyen, X. L., Meadows, M. (2020). Mapping forest fire risk and development of early warning system for NW Vietnam using AHP and MCA/GIS methods, *Applied Sciences, 10*(12), 4348.

Iban, M., Sekertekin, A. (2022). Machine learning based wildfire susceptibility mapping using remotely sensed fire data and GIS: A case study of Adana and Mersin provinces, Turkey, *Ecological Informatics*, 69, 101647.

Lê, S. D., Vương, V. Q. (2014). Phương pháp dự báo nguy cơ cháy rừng theo điều kiện khí hậu ở Việt Nam, *Tạp chí Khoa học Công nghệ lâm nghiệp,* số 1, trang 3-10.

Nguyễn, V. L., Trần, M. Đ., Nguyễn, P. V. (2017). Thực trạng và giải pháp quản lý cháy rừng ứng phó với biến đổi khí hậu tại tỉnh Quảng Bình, *Tạp chí Khoa học Lâm nghiệp*, số 4, trang 139 - 150.

Nguyễn, P.V (2019). Nghiên cứu thực trạng và đề xuất giải pháp quản lý cháy rừng thích ứng với biến đổi khí hậu tại tỉnh Quảng Bình, *Luận án tiến sĩ Lâm nghiệp.*

Nguyen, N.T., Dang, B.T.N, Pham, X.C., Nguyen, H.T., Bui, H.T., Hoang, N.D., Bui, D.T. (2018). Spatial pattern assessment of tropical forest fire danger at Thuan Chau area (Vietnam) using GIS-based advanced machine learning algorithms: A comparative study, *Ecological Informatics, vol.46*, pp.74-85.

Oliveira, S., Oehler, F., Ayanz, J., Camia, A., Pereira, J. (2012). Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest, *Forest Ecology and Management*, 275, 117 - 129

Phạm, N. H. (1988). Xây dựng phương pháp dự báo cháy rừng Thông nhựa (Pinus merkusii J.) ở Quảng Ninh, *Luận án PTS khoa học Nông nghiệp*, Hà Nội.

Pourghasemi, H. (2015). GIS-based forest fire susceptibility mapping in Iran: A comparison between evidential belief function and binary logistic regression models, *Scandinavian Journal of Forest Research*, 40 pp., DOI: 10.1080/02827581.2015.1052750.

Ruano, A., Jolly, W., Freeborn, P., Nieva, D., Vega, N., Herrera, C., Rodrigues, M. (2022). Spatial predictions of human and natural-caused

wildfire likelihood across Montana (USA), *Remote Sensing, 13*(8), 1200.

Sharma, V., Ghosh, S. K. (2023). Evaluating the potential of 8 band Planet scope dataset for crop classification using Random Forest and Gradient Tree Booting by Google Earth Engine, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-M-1-2023, 325-330.

Trần, Q. B. (2017). Nghiên cứu sử dụng công nghệ không gian địa lý (RS, GIS, GPS) trong phát hiện cháy rừng và giám sát tài nguyên rừng, Đề tài nghiên cứu khoa học cấp Bộ.

Trinh, L. H., Zablotskii, V. R. (2017). The application of Landsat multi-temporal thermal infrared data to identify coal fire in the Khanh Hoa coal mine, Thai Nguyen province, Vietnam, *Izvestiya, Atmospheric and Oceanic Physics, 53*(9), 11850 - 6088.

Vasilakos, C., Kalabokidis, K., Hatzopoulos, J., Matsinos, T. (2009). Identifying wildland fire ignition factors through sensitivity analysis of a neural network, *Natural Hazards*, 50, 125 - 143.

Võ, Đ. T. (1995). Phương pháp dự báo, lập bản đồ, khoanh vùng trọng điểm cháy rừng ở Bình Thuận, *Tạp chí Lâm nghiệp*, số 10, trang 11 - 14.