DỰ BÁO NGUY CƠ TRƯỢT LỞ ĐẤT CHO HUYỆN A LƯỚI, TỈNH THỪA THIÊN HUẾ SỬ DỤNG MÔ HÌNH LOGISTIC REGRESSION

PREDICT LANDSLIDE SUSCEPTIBILITY USING LOGISTIC REGRESSION MODEL IN A LUOI DISTRICT, THUA THIEN HUE PROVINCE

Lê Trần Minh Đạt[1], Trương Thị Hồng Ngọc[2], Đoàn Viết Long[1], Nguyễn Chí Công[1]*

[1]Trường Đại học Bách khoa - Đại học Đà Nẵng
[2]Công ty Cổ phần Tư vấn Đầu tư và Xây dựng Thừa Thiên Huế

*Tác giả liên hệ: nccong@dut.udn.vn

(Nhận bài: 07/6/2022; Chấp nhận đăng: 07/9/2022)

Tóm tắt - Nghiên cứu này đề xuất một mô hình hồi quy Logistic (LR) hiệu quả trong việc dự báo nguy cơ trượt lở đất (TLĐ) cho huyện miền núi A Lưới. Cơ sở dữ liệu gồm 429 điểm sạt lở và 574 điểm không sạt lở được thu thập trong các năm 2006, 2009, 2020 với 11 yếu tố biến đầu vào ảnh hưởng đến xác xuất xảy ra được xem xét, bao gồm: Độ dốc, hướng phơi sườn, cao độ, chỉ số độ ẩm địa hình, loại đất, sử dụng đất, khoảng cách đến đường, khoảng cách đến sông, chỉ số thực vật và lượng mưa lớn nhất 3 ngày. Một mô hình LR tối ưu cũng được đề xuất để dự báo nguy cơ TLĐ. Đường cong ROC và diện tích dưới đường cong AUC được sử dụng để đánh giá hiệu suất của mô hình dự báo. Kết quả cho thấy, AUC ở tập huấn luyện đạt 0,8 và 0,81 ở tập kiểm tra. Cuối cùng, một bản đồ nguy cơ TLĐ cho huyện A Lưới với độ phân giải 30mx30m được xây dựng dựa trên kết quả dự báo của mô hình hồi quy LR.

Từ khóa - Học máy; logistic regression; trượt lở đất; ROC; AUC

Abstract - This study proposes an effective Logistic Regression (LR) model for predicting landslide susceptibility (LS) at A Luoi district. The dataset includes 429 landslide points and 574 non-landslide points collected in the years 2006, 2009 and 2020 with eleven input variables, affecting on landslide probability. They are considered, including slope, slope direction, elevation, topographic moisture index, soil type, land use, distance to road, distance to river, vegetation index (NVDI) and 3-day antecedent rainfall. An optimal LR model is also proposed to predict landslide susceptibility. The ROC curve and the area under the ROC curve (AUC) are used to evaluate the performance of the predictive model. The results show that, the AUC in the training set and testing set is 0.8 and 0.81, respectively. Finally, a LS predictive model with a resolution of 30mx30mfor A Luoi district is established basing on the prediction results of the LR model.

Key words - Machine learning; logistic regression; landslides; ROC; AUC.

1. Đặt vấn đề

Trượt lở đất (TLĐ) là loại hình thiên tai nguy hiểm, xảy ra phổ biến ở trên thế giới, gây ra nhiều hậu quả nghiêm trọng. Để góp phần giảm thiểu tác hại của loại hình thiên tai này, công tác nghiên cứu xây dựng bản đồ dự báo nguy cơ TLĐ là rất cần thiết. Bản đồ dự báo nguy cơ TLĐ cung cấp thông tin về mức độ nguy cơ xảy ra trượt lở đất ở mỗi khu vực trong tương lai. Đây là tài liệu hết sức quan trọng hỗ trợ công tác quy hoạch và phòng chống loại hình thiên tai đặc biệt nguy hiểm này [1]. Nghiên cứu xây dựng bản đồ dự báo nguy cơ TLĐ được các nhà khoa học trên thế giới chú trọng từ lâu. Vào những năm 1970, đã xuất hiện những nghiên cứu về đánh giá nguy cơ trượt lở đất [1]. Cho đến nay, có 2 phương pháp cơ bản để xây dựng bản đồ nguy cơ TLĐ là phương pháp định tính và phương pháp định lượng hoặc có thể phân làm 3 nhóm: Phương pháp phát hiện (heuristic); Phương pháp thống kê (statistical); và Phương pháp quyết định (deterministic) [5]. Phương pháp phát hiện dựa trên sự hiểu biết của các chuyên gia để đánh giá trọng số của các yếu tố ảnh hưởng, từ đó xây dựng chỉ số nguy cơ của từng vị trí trên bản đồ. Phương pháp này có nhược điểm lớn là phụ thuộc vào ý kiến chủ quan của con người [2], [8], [9]. Phương pháp quyết định là một phương pháp định lượng, dựa trên việc tính toán và phân tích điều kiện ổn định hoặc không ổn định của mái dốc. Đây là một phương pháp có độ chính xác cao, tuy nhiên cũng yêu cầu mức độ rất chi tiết của dữ liệu nên chỉ áp dụng trong phạm vi nhỏ [5]. Phương pháp thống kê dựa vào dữ liệu các vụ TLĐ trong quá khứ và tập hợp các yếu tố ảnh hưởng để xây dựng mô hình dự báo và thành lập bản đồ nguy cơ TLĐ, phương pháp này tỏ ra ưu việt đối với khu vực có diện tích rộng lớn [5]. Với sự phát triển của khoa học thống kê hiện đại, kỹ thuật học máy, học sâu đã được áp dụng trong những năm gần đây, kết hợp với công cụ GIS để xây dựng mô hình dự báo nguy cơ TLĐ dựa trên phương pháp thống kê với độ chính xác cao [3], [4], [10]. Nghiên cứu thống kê các bài báo uy tín viết về lĩnh vực này của Reichenbach [1] trong giai đoạn từ năm 1983 đến 2016 đã cho thấy có đến hơn 160 mô hình thống kê đã được áp dụng, trong đó mô hình hồi quy Logistic là loại được sử dụng phổ biến nhất. Nghiên cứu của Pourghasemi [5] cũng cho kết quả tương tự và lý giải rằng mô hình hồi quy Logistic được sử dụng nhiều nhất do có ưu điểm ít mắc lỗi, dễ sử dụng và phù hợp với đa số khu vực nghiên cứu.

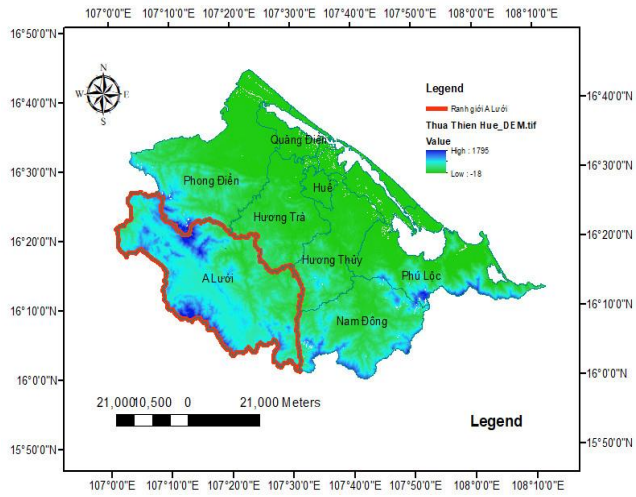
Ở nước ta, trượt lở đất chủ yếu xuất hiện vào các tháng mùa mưa, xảy ra chủ yếu ở các tỉnh miền núi phía Bắc và khu vực miền Trung - Tây Nguyên [6]. Theo báo cáo của Tổng cục phòng chống thiên tai - Bộ Nông nghiệp và Phát triển Nông thôn, thiên tai lũ quét và sạt lở đất ở Việt Nam giai đoạn 2000 đến 2009 xảy ra 108 trận làm 544 người chết

[1] The University of Danang - University of Science and Technology (Le Tran Minh Dat, Doan Viet Long, Nguyen Chi Cong)
[2] Thua Thien Hue Construction and Investment Consulting Joint Stock Company (Truong Thi Hong Ngoc)

và mất tích. Trong giai đoạn từ 2010 đến 2020, có đến 224 trận lũ quét và sạt lở đất xảy ra làm chết và mất tích 572 người, riêng trong tháng 10 năm 2020 đã có 18 trận trượt lở đất tại 4 tỉnh thành miền Trung làm 111 người chết và mất tích. Có thể nói, thiên tai TLĐ ngày càng xảy ra với mức độ nghiêm trọng, đòi hỏi cần nghiên cứu xác định ra các khu vực có nguy cơ, từ đó đưa ra các giải pháp phòng chống.

A Lưới là một huyện miền núi biên giới phía Tây của tỉnh Thừa Thiên Huế, Địa giới huyện A Lưới được giới hạn trong tọa độ địa lý từ $16^0$00'57" đến $16^0$27'30" vĩ độ Bắc và từ 107^0 0'3' đến $107^0$30'30" kinh độ Đông. Hàng năm huyện A Lưới gánh chịu rất nhiều rủi ro do thiên tai gây ra như: Bão, lũ lụt, hạn hán và TLĐ. Trong đó, TLĐ là một dạng thiên tai thường xuyên xảy ra vào mùa mưa. Trong thời gian qua, một số nghiên cứu về khảo sát, đánh giá nguy cơ TLĐ đã được áp dụng cho khu vực này [2], [7], [8], [9]. Nghiên cứu của [2], [8], [9] đã sử dụng các kỹ thuật thống kê cổ điển để đánh giá trọng số của các yếu tố nguy cơ, kết hợp với công cụ GIS để xây dựng bản đồ nguy cơ TLĐ, các nghiên cứu này cũng chưa đánh giá được độ chính xác của mô hình. Nguyen Thanh Long [7] đã áp dụng mô hình chỉ số thống kê (Statistical Index - SI), mô hình hồi quy Logistic và mô hình Certainty Factor (CF) để đánh giá nguy cơ TLĐ. Kết quả chỉ ra mô hình CF cho kết quả tốt nhất. Tuy nhiên, do chỉ dựa trên số điểm TLĐ hạn chế (181 điểm) nên vẫn áp dụng phương pháp thống kê truyền thống và chưa đưa ra được các cấp dự báo nguy cơ TLĐ.

Dựa trên những phân tích về tình hình nghiên cứu ở trên thế giới và khu vực, bài báo này áp dụng mô hình học máy sử dụng phương pháp hồi quy Logistic để xây dựng và đánh giá mô hình dự báo nguy cơ TLĐ cho địa bàn huyện A Lưới, tỉnh Thừa Thiên Huế. Mô hình này sau đó kết hợp với kỹ thuật GIS để xây dựng bản đồ dự báo nguy cơ TLĐ cho khu vực này.

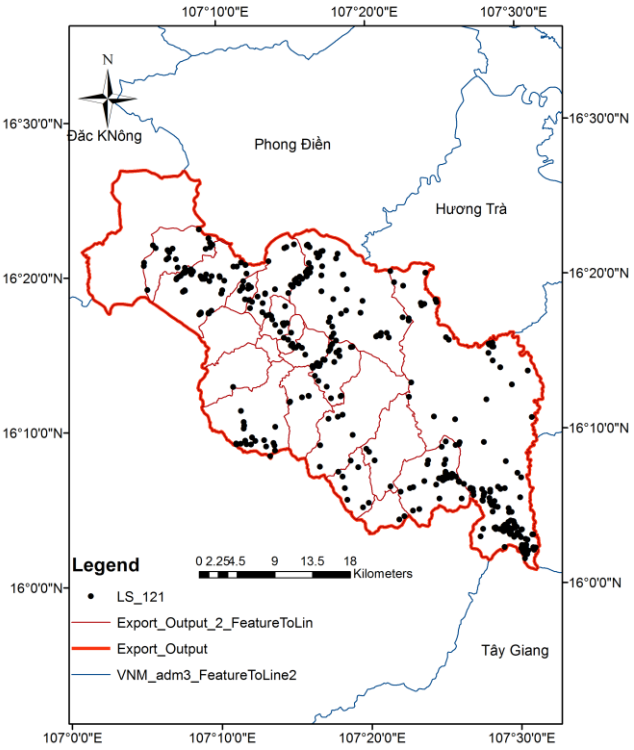


Hình 1. *Vị trí khu vực nghiên cứu (đường bao nét đậm)*

2. Dữ liệu và phương pháp nghiên cứu

2.1. Dữ liệu nghiên cứu

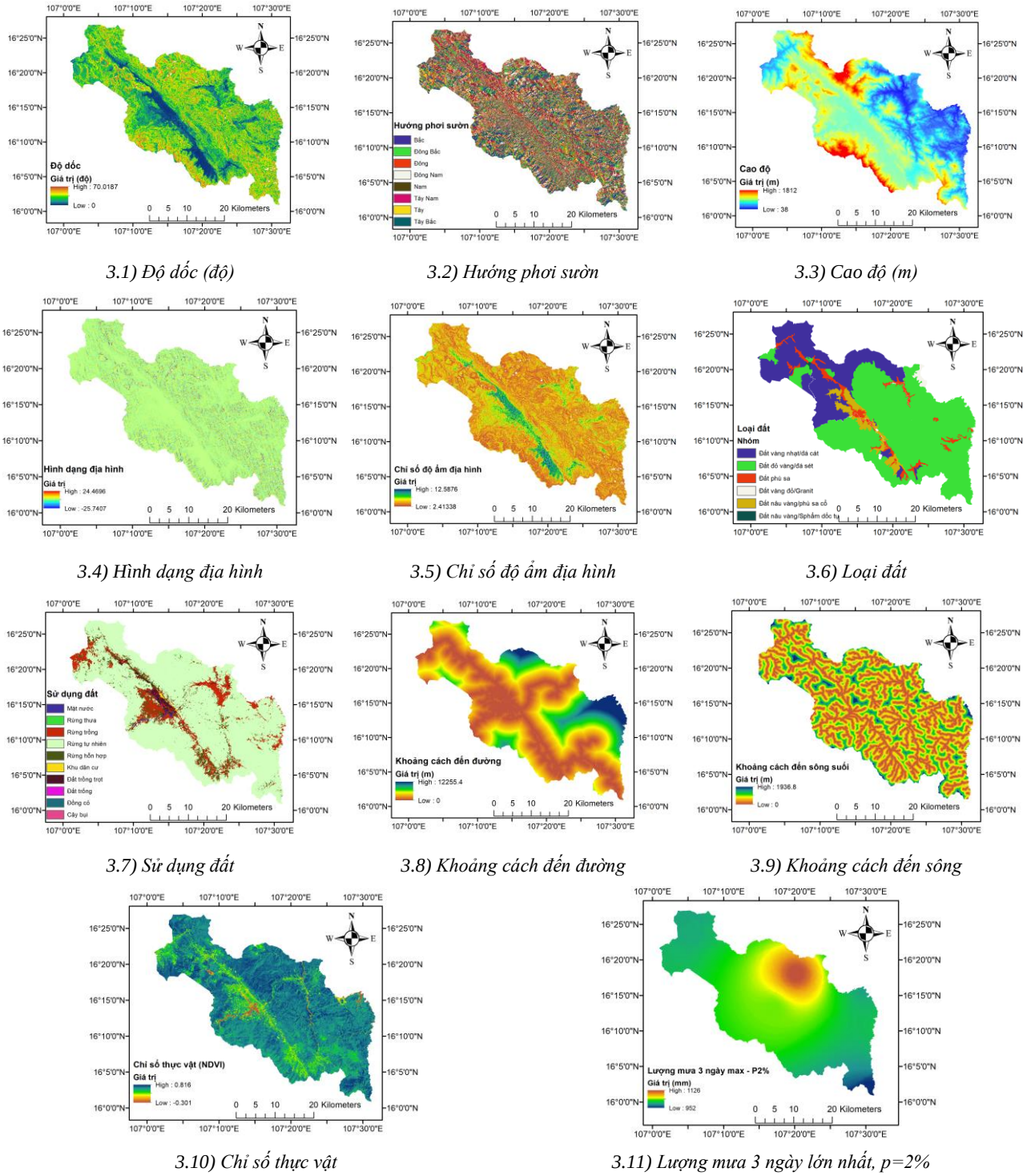
Các vị trí sạt ở tại vùng nghiên cứu được xác định dựa trên việc điều tra, khảo sát kết hợp phục hồi các điểm sạt lở sử dụng kỹ thuật viễn thám. Các vị trí sạt lở đất đã được số hóa bằng cách diễn giải trực quan bằng công cụ ArcGIS 10.2 (Hình 2). Dữ liệu điểm đại diện cho các vị trí sạt lở được chuyển đổi sang định dạng pixel, với độ phân giải 30x30 m. Ngoài ra, các pixel đại diện các điểm không trượt lở được chọn ngẫu nhiên từ các pixel không trượt lở trong khu vực nghiên cứu. Bộ dữ liệu được sử dụng để huấn luyện và kiểm tra mô hình dự báo LR bao gồm 429 điểm TLĐ và 574 điểm không sạt lở. Mười một yếu tố biến đầu vào ảnh hưởng đến xác suất xảy ra TLĐ được xem xét, bao gồm: Độ dốc (x_1), hướng phơi sườn (x_2), cao độ (x_3), hình dạng bề mặt địa hình (x_4), chỉ số độ ẩm địa hình (x_5), loại đất (x_6), sử dụng đất (x_7), khoảng cách đến đường (x_8), khoảng cách đến sông (x_9), chỉ số thực vật (NVDI) (x_{10}) và lượng mưa 3 ngày lớn nhất [1] ứng với tần suất 2% (x_{11}). Trạng thái trượt lở đất được chọn làm biến đầu ra (y) cho mô hình dự báo, chỉ nhận giá trị 0 nếu không trượt và 1 nếu trượt.



Hình 2. *Vị trí các điểm TLĐ thu thập (chấm đen)*

Bảng 1. *Mô tả 11 biến đầu vào của mô hình*

Biến đầu vào	Yếu tố ảnh hưởng đến TLĐ	Nguồn, tỷ lệ, độ phân giải
x_1	Độ dốc	NasaDEM, 30mx30m
x_2	Hướng phơi sườn	NasaDEM, 30mx30m
x_3	Cao độ	NasaDEM, 30mx30m
x_4	Hình dạng địa hình	NasaDEM, 30mx30m
x_5	Chỉ số độ ẩm địa hình	NasaDEM, 30mx30m
x_6	Loại đất	1/50.000
x_7	Sử dụng đất	landcovermapping.org, 30mx30m
x_8	Khoảng cách đến đường	1/50.000
x_9	Khoảng cách đến sông	1/50.000
x_{10}	Chỉ số thực vật	sentinel.esa.int, 30mx30m
x_{11}	Lượng mưa	[2], 30mx30m



3.1) Độ dốc (độ) *3.2) Hướng phơi sườn* *3.3) Cao độ (m)*

3.4) Hình dạng địa hình *3.5) Chỉ số độ ẩm địa hình* *3.6) Loại đất*

3.7) Sử dụng đất *3.8) Khoảng cách đến đường* *3.9) Khoảng cách đến sông*

3.10) Chỉ số thực vật *3.11) Lượng mưa 3 ngày lớn nhất, p=2%*

Hình 3. *Dữ liệu 11 biến đầu vào mô hình*

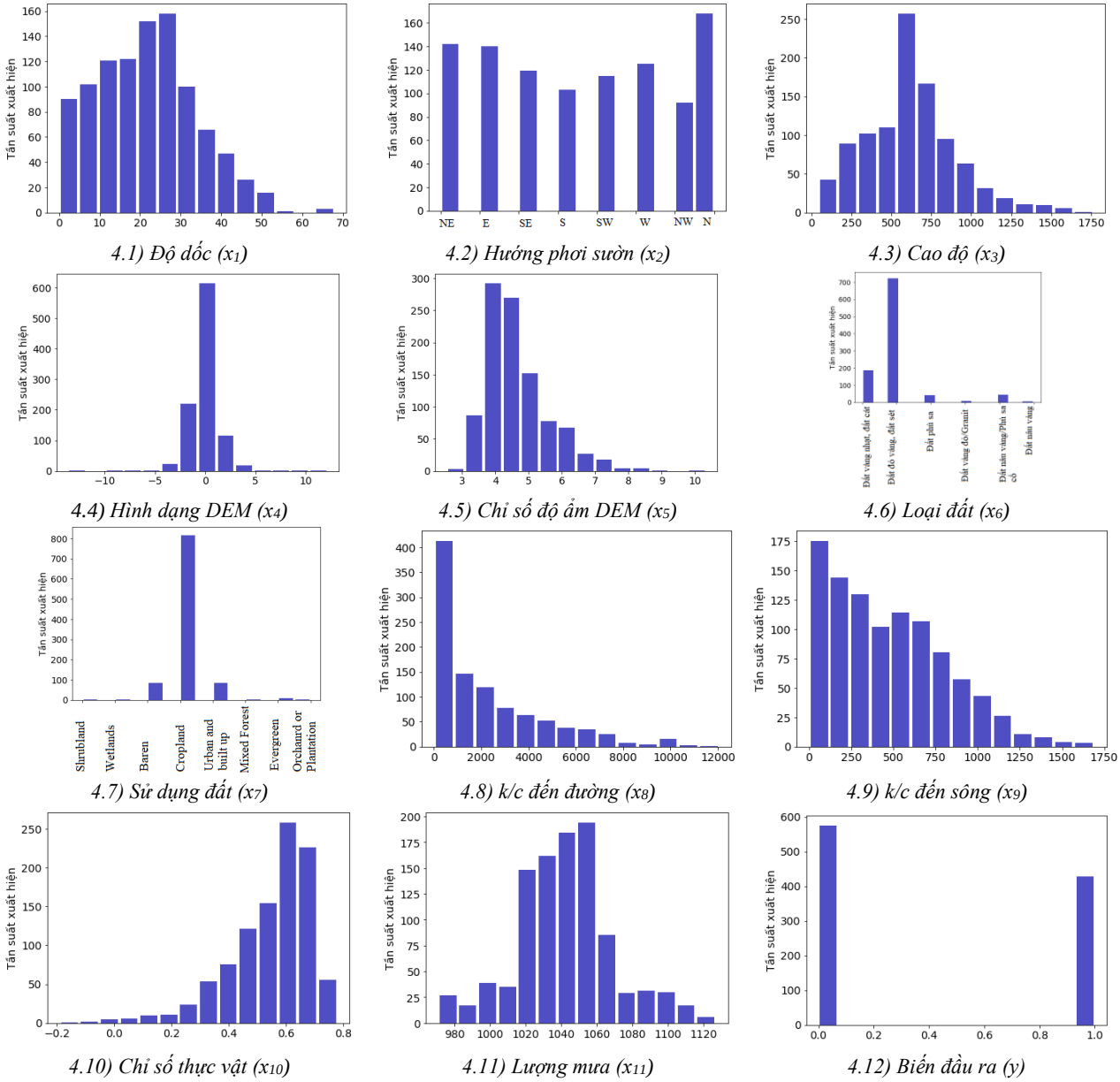
Để xác nhận hiệu quả của mô hình LR, phần dữ liệu trong tập kiểm tra chiếm tỉ lệ 30% (301 điểm) trong tổng số 1004 mẫu. Tập dữ liệu huấn luyện được sử dụng để xác định các trọng số (hoặc tham số) của mô hình LR chứa 70% bộ dữ liệu (702 điểm). Tần suất xuất hiện của các biến đầu vào và đầu ra trong bộ dữ liệu được thể hiện trong Hình 4.

Để giảm biên độ biến động trong mô hình LR, cũng như nâng cao tốc độ học tập của mô hình, hiệu suất, độ chính xác và tính ổn định của quá trình huấn luyện, biến đầu vào và đầu ra của tập dữ liệu đều được quy đổi lại tỷ lệ trong khoảng [0, 1]. Phương trình quy đổi tỷ lệ của các biến được biểu diễn bên dưới:

$$\hat{x}_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Trong đó: x_i là giá trị thực tế, $\hat{x}_i$ là giá trị quy đổi, x_{min}, x_{max} lần lượt là giá trị nhỏ nhất và lớn nhất của các biến đầu vào.

Dữ liệu thống kê của các biến đầu vào và đầu ra được tóm tắt trong Bảng 2.



4.1) Độ dốc (x_1) *4.2) Hướng phơi sườn (x_2)* *4.3) Cao độ (x_3)*

4.4) Hình dạng DEM (x_4) *4.5) Chỉ số độ ẩm DEM (x_5)* *4.6) Loại đất (x_6)*

4.7) Sử dụng đất (x_7) *4.8) k/c đến đường (x_8)* *4.9) k/c đến sông (x_9)*

4.10) Chỉ số thực vật (x_{10}) *4.11) Lượng mưa (x_{11})* *4.12) Biến đầu ra (y)*

Hình 4. *Tần suất xuất hiện của 11 biến đầu vào và đầu ra của mô hình*

Bảng 2. *Thống kê mô tả các biến đầu vào và biến đầu ra của bộ dữ liệu*

Biến	min	mean	max	sd	skewness
x_1	0	21,74	67,84	12,16	0,38
x_2	1,00	4,49	8,00	2,42	0,04
x_3	43,00	624,20	1763,00	282,49	0,62
x_4	-13,66	-0,029	12,18	1,52	-0,24
x_5	2,53	4,64	10,33	0,96	1,31
x_6	1,00	2,00	6,00	0,86	2,23
x_7	15,00	2193,21	12051,20	2412,49	1,38
x_8	0	473,90	1690,00	341,60	0,66
x_9	1,00	4,00	8,00	0,55	-1,37
x_{10}	-0,19	0,55	0,80	0,14	1,72
x_{11}	970,82	1042,21	1126,74	27,13	0,20
y	0	0,427	1,00	0,49	0,29

2.2. Phương pháp

Trong hồi quy logistic (LR), mối quan hệ định lượng giữa sự xuất hiện của trượt lở đất và sự phụ thuộc của nó vào một tập hợp các yếu tố ảnh hưởng được biểu thị dưới dạng một hàm logistic:

$$p = \frac{1}{1 + e^{-Z}} \quad (2)$$

Trong đó, p là xác suất của sự kiện trượt đất, nếu trượt thì p = 1 và không trượt thì p = 0. Z là hàm tuyến tính đa biến như sau:

$$Z = a_0 + \sum_{i=1}^{n} a_i x_i \quad (3)$$

Trong đó, a_0, a_i là các tham số của mô hình và x_i là các biến đầu vào.

Một yếu tố quan trọng của mô hình LR là việc xác định các các các tham số (a_0, a_i) của phương trình hồi quy phù

hợp để tối ưu hóa hàm mất mát. Mô hình LR được đề xuất thông qua việc tìm kiếm các tham số tối ưu sử dụng bộ công cụ GridSearchCV của scikit-learn. Thuật toán Broyden – Fletcher – Goldfarb – Shanno (lbfgs) bộ nhớ giới hạn kết hợp phương pháp điều chuẩn với hệ số C=1 được áp dụng để tối ưu hóa hàm mất mát. Kỹ thuật phân bố dữ liệu Stratified K-Fold được áp dụng để đảm bảo tỉ lệ phân chia tương đồng nhau giữa các biến trong bộ dữ liệu. Các thuật toán và code được triển khai trên Google Colab kết hợp với công cụ GIS.

Để giảm thiểu độ nhiễu trong mô hình và để đảm bảo sự kết hợp tuyến tính hoàn hảo giữa các biến, một phân tích đa cộng tuyến đã được tiến hành. Hệ số phóng đại phương sai và dung sai được áp dụng để kiểm tra tính đa cộng tuyến giữa 11 biến đầu vào.

Để đánh giá độ chính xác và hiệu suất của mô hình dự báo, biểu đồ đường cong ROC dựa trên mối liên hệ giữa độ nhạy và độ đặc hiệu và chỉ số AUC được sử dụng. Độ nhạy, độ đặc hiệu và độ chính xác được xác định bằng các phương trình sau:

$$\text{Độ } nhạy = \frac{TP}{TP + FN} \quad (4)$$

$$\text{Độ đặc } hiệu = \frac{TN}{FP + TN} \quad (5)$$

$$\text{Độ } chính\ xác = \frac{TP + TN}{TP + FN + FP + TN} \quad (6)$$

Trong đó:

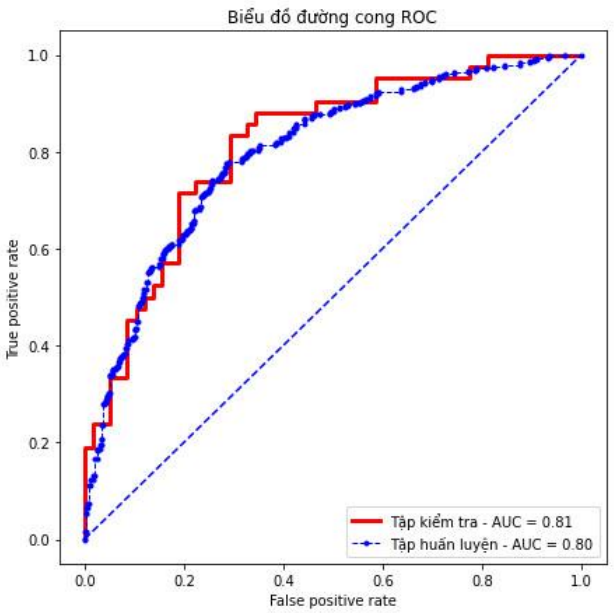
TP: là số điểm TLĐ mà mô hình dự báo đúng;

FP: là số điểm TLĐ mà mô hình dự báo sai;

FN: là số điểm không TLĐ mà mô hình dự báo sai;

TN: là số điểm không TLĐ mà mô hình dự báo đúng.

3. Kết quả và bàn luận



Hình 5. *Đường cong ROC của tập dữ liệu huấn luyện và kiểm tra*

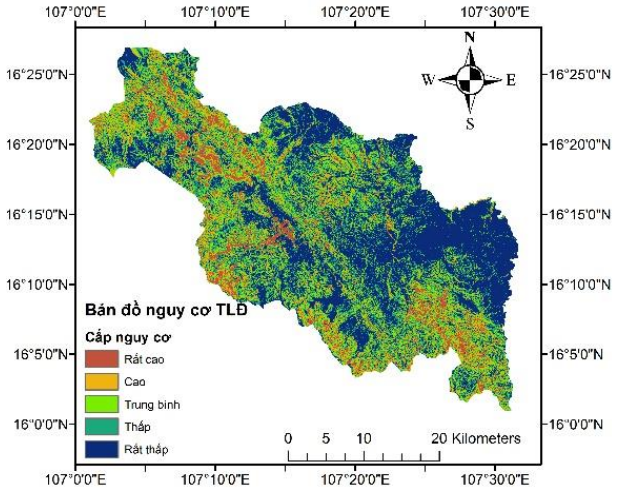
Mô hình hồi quy Logistic cho kết quả dự báo tốt. Điều này được thể hiện qua kết quả ở Hình 5 (đồ thị đường ROC) và giá trị AUC trong tập dữ liệu huấn luyện và kiểm tra quan sát đạt được giá trị tương ứng 0,80 và 0,81. Quan sát các giá trị trong các fold của quá trình phân bố dữ liệu cho thấy kết quả trên tập dữ liệu kiểm tra và huấn luyện là tương đồng nhau, điều này đảm bảo sự hoạt động ổn định của mô hình dự báo. Do đó với mô hình LR tối ưu được đề xuất trong nghiên cứu này có thể xem là một công cụ hữu hiệu trong việc dự báo nguy cơ LTĐ của vùng nghiên cứu.

Giá trị các hệ số trong phương trình hồi quy (3) của mô hình LR đề xuất được thể hiện trong Bảng 3.

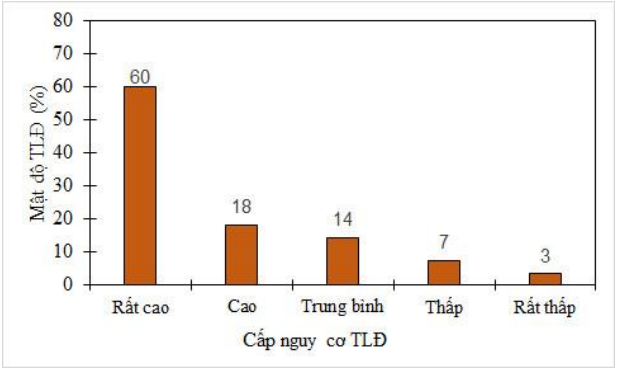
Bảng 3. *Hệ số của phương trình hồi quy trong mô hình LR.*

a_0	a_1	a_2	a_3	a_4	a_5
3,53	2,56	0,65	0,19	-0,53	-1,58
a_6	a_7	a_8	a_9	a_{10}	a_{11}
-0,39	-0,01	-1,66	0,25	4,28	-1,73

Hình 6 là bản đồ dự báo nguy cơ TLĐ tại huyện A Lưới được chia theo 5 mức cấp độ: nguy cơ rất cao, nguy cơ cao, nguy cơ trung bình, nguy cơ thấp và nguy cơ rất thấp. Hình 7 biểu diễn tỷ lệ % mật độ TLĐ với 5 cấp nguy cơ dự báo nêu trên. Đáng chú ý nhất là vùng dự báo mức nguy cơ TLĐ rất cao có tỷ lệ % mật độ TLĐ đạt 60%.



Hình 6. *Bản đồ dự báo nguy cơ TLĐ huyện A Lưới.*



Hình 7. *Mật độ TLĐ của huyện A Lưới*

4. Kết luận

Nghiên cứu này đã thu thập và cập nhật các điểm TLĐ cho huyện A Lưới với tổng số 429 điểm trong các năm 2006, 2009 và 2020. Dựa trên phân tích 11 biến đầu vào,

một mô hình hồi quy Logistic tối ưu được đề xuất để dự báo xác suất xảy ra TLĐ cho vùng nghiên cứu với độ tin cậy khá cao, giá trị AUC=0,80 cho tập huấn luyện và AUC=0,81 cho tập dữ liệu kiểm tra. Dựa trên mô hình dự báo đề xuất kết hợp với công cụ GIS, một bản đồ nguy cơ TLĐ chi tiết cho huyện A Lưới với độ phân giải 30mx30m đã được xây dựng.

Lời cảm ơn: Đoàn Viết Long được tài trợ bởi Tập đoàn Vingroup – Công ty CP và hỗ trợ bởi chương trình học bổng đào tạo thạc sĩ, tiến sĩ trong nước của Quỹ Đổi mới sáng tạo Vingroup (VINIF), Viện Nghiên cứu Dữ liệu lớn (VinBigdata), mã số VINIF.2021.TS.122.

TÀI LIỆU THAM KHẢO

[1] P. Reichenbach, M. Rossi, B. D. Malamud, M. Mihir, and F. Guzzetti, "A review of statistically-based landslide susceptibility models", *Earth-Science Rev.*, vol. 180, 2018, pp. 60–91.

[2] Vo Nguyen Duc Phuoc, Nguyen Quang Binh, Phan Dinh Hung, Doan Viet Long, Nguyen Chi Cong, "Study on the causes of landslides for mountainous regions in central region of VietNam" *Journal of science and technology*. ISSN 1859-1531, Vol. 17, No. 12.1, 2019, pp. 29-33.

[3] B. Thai Pham, D. Tien Bui, and I. Prakash, "Landslide susceptibility modelling using different advanced decision trees methods" *Civ. Eng. Environ. Syst.*, vol. 35, no. 1–4, 2018, pp. 139–157.

[4] B. T. Pham et al, "Ensemble modeling of landslide susceptibility using random subspace learner and different decision tree classifiers" *Geocarto Int*, 2020, pp. 1–23.

[5] H. R. Pourghasemi, Z. T. Yansari, P. Panagos, and B. Pradhan, "Analysis and evaluation of landslide susceptibility: a review on articles published during 2005–2016 (periods of 2005–2012 and 2013–2016)". *Arab. J. Geosci.*, vol. 11, no. 9, 2018, p. 193.

[6] Doan Viet long, Nguyen Chi Cong, Nguyen Quang Binh, Nguyen Tien Cuong, "Đánh giá thực trạng và giải pháp nghiên cứu về sạt lở đất ở Việt Nam giai đoạn 2010-2020", *Tạp chí Khoa học và Công nghệ Thủy lợi*. Số 61, 2020, pp. 119-128.

[7] Nguyen Thanh Long et al, "Analysis and mapping of rainfall-induced lamdslide susceptibility in A Luoi district, Thua Thien Hue province, Vietnam" *Water* 2019,11,5; doi: 10.3390.

[8] N. H. K. Linh, J. Degener, N. B. Ngoc, and T. T. M. Chau, "Mapping risk of landslide at A Luoi district, Thua Thien Hue province, Vietnam by GIS-based multi-criteria evaluation" *Asian J. Agric. Dev.*, vol. 15, no. 1362-2018–3543, 2018, pp. 87–105.

[9] M. T. Tan and N. Van Tao, "Studying landslides in Thua Thien-Hue province: *VIETNAM J. EARTH Sci.*, vol. 36, no. 2, 2014, pp. 121–130.

[10] D. T. Bui, P. Tsangaratos, V.-T. Nguyen, N. Van Liem, and P. T. Trinh, "Comparing the prediction performance of a Deep Learning Neural Network model with conventional machine learning models in landslide susceptibility assessment", *Catena*, vol. 188, 2020, pp. 104-426.