Ứng dụng học máy nhận dạng tướng đất đá khu vực bồn trũng An Châu

Doãn Ngọc San[1*], Nguyễn Văn Thắng[2], Hoàng Hữu Hiệp[2], Nguyễn Anh Hào[2], Trần Thị Oanh[1], Nguyễn Thị Minh Hồng[3]

[1]Trường Đại học Dầu khí Việt Nam, 762 Cách Mạng Tháng Tám, phường Long Toàn, TP Bà Rịa, tỉnh Bà Rịa - Vũng Tàu, Việt Nam
[2]Công ty TNHH MTV Dầu khí Sông Hồng, 117 Trần Duy Hưng, phường Trung Hòa, quận Cầu Giấy, Hà Nội, Việt Nam
[3]Trường Đại học Mỏ - Địa chất, 18 phố Viên, phường Đức Thắng, quận Bắc Từ Liêm, Hà Nội, Việt Nam

Ngày nhận bài 22/2/2023; ngày chuyển phản biện 24/2/2023; ngày nhận phản biện 7/3/2023; ngày chấp nhận đăng 11/3/2023

Tóm tắt:

Bồn trầm tích An Châu (bồn An Châu) là một cấu trúc địa chất kéo dài theo phương Tây Bắc - Đông Nam, phân bố ở vùng Đông Bắc Việt Nam. Bồn An Châu được cho là bồn có quy mô lớn và có tiềm năng dầu khí. Thực tế này cho thấy, nhiều khả năng sẽ phát hiện được các tích tụ dầu khí trong bồn An Châu thuộc địa phận của Việt Nam. Mặc dù tiềm năng dầu khí của bồn này được nhận định từ rất sớm, nhưng vì nhiều lý do mà cho đến nay công tác điều tra, khảo sát địa chất và thăm dò ở khu vực bồn An Châu còn rất sơ sài, chưa đáp ứng được các yêu cầu của công tác tìm kiếm - thăm dò dầu khí. Việc ứng dụng mô hình học máy (machine learning - ML) vào nhận dạng tướng đá là một phương pháp mới giúp giảm thiểu thời gian xử lý, tổng hợp cơ sở dữ liệu về cả số lượng và định dạng, phát hiện các mối quan hệ ẩn sâu giữa các lớp thông tin nhận dạng. Mục tiêu chính của nghiên cứu này là nhận dạng tướng đất đá khu vực bồn An Châu từ phần số liệu đầy đủ đã được huấn luyện bởi cấu trúc mạng cây quyết định (DT) kết hợp thuật toán gradient boosting (XGB) để đánh giá cấu trúc và xác định tiềm năng dầu khí khu vực này. Điều kiện tiên quyết để nâng cao độ chính xác của học máy là phải làm giàu cơ sở dữ liệu thông qua tích hợp số liệu địa chất - địa chấn và tính toán thêm các thuộc tính để xây dựng "mô hình học" - huấn luyện ML và sử dụng kết quả huấn luyện đó để nhận dạng tướng đất đá trong khu vực bồn An Châu.

Từ khóa: **cây quyết định, cơ sở dữ liệu địa chất - địa vật lý, học máy, trí tuệ nhân tạo.**

Chỉ số phân loại: **1.2, 1.8, 2.1**

Application of machine learning for facies recognition in An Chau basin

Ngoc San Doan[1*], Van Thang Nguyen[2], Huu Hiep Hoang[2], Anh Hao Nguyen[2], Thi Oanh Tran[1], Thi Minh Hong Nguyen[3]

[1]Petroleum Vietnam University (PVU), 762 Cach Mang Thang Tam Street, Long Toan Ward, Ba Ria City, Ba Ria - Vung Tau Province, Vietnam
[2]PVEP Song Hong, 117 Tran Duy Hung Street, Trung Hoa Ward, Cau Giay District, Hanoi, Vietnam
[3]Hanoi University of Mining and Geology, 18 Vien Street, Duc Thang Ward, Bac Tu Liem District, Hanoi, Vietnam

Received 22 February 2023; revised 7 March 2023; accepted 11 March 2023

Abstract:

The An Chau basin (An Chau basin) is a geological structure extending in a northwest-southeast direction, located in the northeastern region of Vietnam. The An Chau basin is considered a large-scale basin with significant petroleum potential. This fact suggests a high likelihood of discovering hydrocarbon accumulations within the An Chau basin within Vietnam's territory. Although the petroleum potential of this basin was identified early, for various reasons, geological surveys and exploration activities in the An Chau basin remain rudimentary and have not yet met the requirements for effective petroleum exploration and production. The application of machine learning (ML) models to lithofacies identification offers a novel approach to reducing processing time, consolidating large and diverse datasets, and uncovering hidden relationships among layers of identification information. The primary objective of this study is to identify lithofacies in the An Chau basin using comprehensive datasets trained with a decision tree (DT) structure combined with a gradient boosting algorithm (XGB) to evaluate the structure and assess the petroleum potential of this area. A prerequisite for improving the accuracy of machine learning is enriching the database through the integration of geological and seismic data, along with the calculation of additional attributes to build the "learning model" training the ML system and applying the trained model to identify lithofacies in the An Chau basin.

Keywords: **artificial intelligence, decision tree, geophysical database, machine learning.**

Classification numbers: **1.2, 1.8, 2.1**

[*]Tác giả liên hệ: Email: doanngocsan@gmail.com

1. Đặt vấn đề

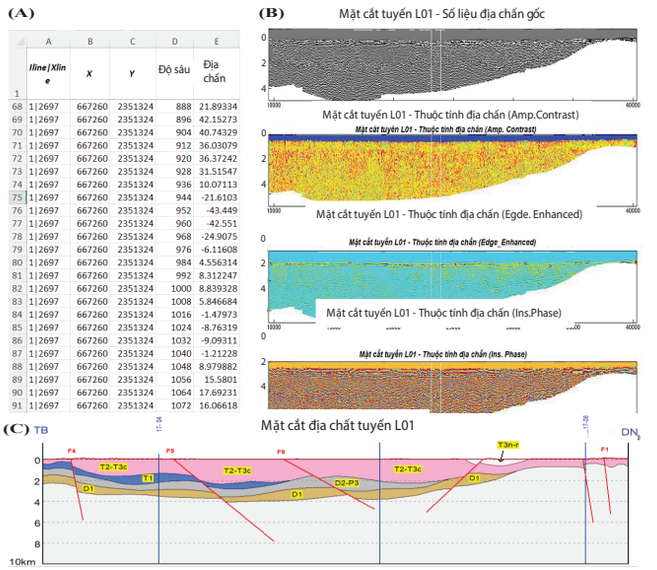
Đòi hỏi đổi mới công nghệ tìm kiếm thăm dò dầu khí (khảo sát và xử lý minh giải) để tìm ra các mỏ dầu khí mới là rất cấp bách trong điều kiện trữ lượng dầu khí ngày càng cạn kiệt, môi trường địa chất rất phức tạp của các vùng tiềm năng còn lại thì việc đổi mới công nghệ tìm kiếm thăm dò dầu khí (khảo sát và xử lý minh giải) để tìm ra các mỏ mới là rất cấp bách. Công tác xử lý số liệu yêu cầu đẩy mạnh việc phát triển về thuật toán, đòi hỏi việc sử dụng và tối ưu hóa các hệ thống phần cứng và đáp ứng việc xử lý Big data [1]. Với việc các bài toán cơ bản như xử lý hình ảnh và xử lý ngôn ngữ tự nhiên ngày càng tiếp cận với trình độ cơ bản của con người, trí tuệ nhân tạo (artificial intelligence - AI) sẽ là một giải pháp mang tính đột phá để khai phá lượng dữ liệu địa chất - địa vật lý để tìm ra các mối quan hệ ẩn giữa các lớp thông tin địa chất - địa vật lý và dầu khí, qua đó tìm ra các bộ thuộc tính đặc trưng để đánh giá triển vọng dầu khí [2]. Khu vực trũng An Châu hiện đang được đánh giá là một trong những vị trí có tiềm năng dầu khí cao trên đất liền ở Việt Nam [3].

2. Nội dung nghiên cứu

2.1. Tích hợp cơ sở dữ liệu phục vụ đầu vào cho hệ thống học máy

Sự phát triển mạnh mẽ của công nghệ thông tin, sự nâng cấp về thiết bị và các chương trình ứng dụng cho phép chiết xuất được khối lượng thông tin ngày càng nhiều từ tài liệu địa chấn, làm tăng số lượng tham số có thể sử dụng được trong quá trình minh giải. Thuộc tính địa chấn (Seismic Attributes) được hiểu là những đặc trưng động lực học như: pha, tần số, biên độ hay các thông số phụ trợ khác như thuộc tính đa mạch (coherency...). Các thuộc tính địa chấn có thể được tính theo mặt cắt, theo bề mặt hoặc theo khối [4]. Các thuộc tính địa chấn bao gồm cả các đặc điểm hình học động (thời gian, tốc độ...) và đặc điểm động lực (pha, biên độ, tần số, độ suy giảm năng lượng). Thuộc tính địa chấn còn có thể được xác định theo đơn mạch (được tính cho từng mạch địa chấn), liên kết giữa các mạch hoặc thông qua tần số, biên độ, pha, tần số tức thời, pha tức thời, cường độ phản xạ... Các thuộc tính đa mạch được tính trên cơ sở hàm tương quan liên kết theo một nhóm mạch địa chấn và trên nhiều xung địa chấn (coherency).

Trong các thuộc tính địa chấn, thuộc tính được quan tâm nhiều hiện nay là biên độ. Từ các dị thường biên độ như các "điểm sáng", "điểm tối"... có thể phát hiện các dấu hiệu liên quan đến dầu khí. Dị thường biên độ thường được sử dụng để thành lập các bản đồ phản ảnh sự biến đổi tướng đất đá và tham số vỉa của tầng chứa. Sự biến đổi biên độ theo chiều ngang có thể là cơ sở phân biệt sự khác nhau của các loại tướng đất đá [5].



Hình 1. Tích hợp số liệu địa chất - địa chấn. (A) Số liệu địa chấn, **(B)** Các thuộc tính địa chấn, **(C)** Mặt cắt địa chất.

Hình 1 là các thông tin địa chấn và kết quả minh giải theo phương pháp truyền thống tại tuyến L01. Theo đó, đã minh giải được 5 tầng trầm tích (biểu hiện theo các màu) theo các mặt phản xạ địa chấn, có các tuổi khác nhau từ dưới lên như sau (hình 1C): Địa tầng tuổi Devon sớm (D_1); Địa tầng tuổi Devon giữa - Pecmi muộn (D_2-P_3); Địa tầng tuổi Trias sớm (T_1); Trias giữa - muộn (T_2-T_3c) và Trias muộn (T_3n-r).

Hiện nay, có trên 50 thuộc tính địa chấn khác nhau có thể chiết xuất được từ tài liệu địa chấn và được sử dụng để minh giải các cấu trúc địa chất, địa tầng, tính chất chất lưu, cơ lý đá [6]. Có thể nói, sự phát triển của phân tích thuộc tính địa chấn gắn liền với phát triển trong lĩnh vực máy tính - điện toán. Phân tích thuộc tính địa chấn trong tìm kiếm dầu khí là quá trình sử dụng các kỹ thuật phân tích dữ liệu địa chất và địa chấn để đánh giá tiềm năng của các khu vực địa chất có khả năng chứa dầu khí. Trong quá trình phân tích thuộc tính địa chấn, các chuyên gia sử dụng các phương pháp phân tích dữ liệu như phân tích tương quan, phân tích phân cụm, phân tích thành phần chính và các phương pháp khác để phân tích và tìm kiếm các mô hình địa chấn. Các mô hình này cung cấp thông tin về các tính chất địa chất và địa chấn của khu vực đó, bao gồm các đặc điểm về cấu trúc, độ sâu, hướng và hình dạng của các lớp đất và đá.

Phân tích thuộc tính địa chấn rất quan trọng trong tìm kiếm dầu khí. Nó giúp xác định các khu vực có tiềm năng khai thác dầu khí và giảm thiểu các rủi ro và chi phí không cần thiết trong quá trình tìm kiếm. Nó cũng giúp các chuyên gia địa chất và địa chấn tìm hiểu sâu hơn về đặc điểm địa chất và địa chấn của các khu vực khai thác dầu khí, để có thể phát triển các phương pháp khai thác hiệu quả và bền vững hơn. Khả năng quản trị dữ liệu lớn cũng là một ưu điểm nổi trội của hệ thống AI. Độ chính xác "nhận dạng" đối tượng của ML càng tăng lên khi chúng ta sử dụng càng nhiều "features" đặc trưng cho

đối tượng cần nghiên cứu [7]. Chính vì vậy, nhằm mục đích nâng cao độ chính xác của thuật toán nhận dạng bằng ML, bằng phần mềm PETREL đã tiến hành phân tích các thuộc tính địa chấn như: ‹amp_contrast›, ‹cur_contour›, ‹Cur_max›, ‹Curvature_Az_min›, 'Dominant_Freq', 'Edge_Enhanced', 'GLCM_X','ins_fre', 'ins_phase', 'Polarity', 'RAI' ,'trace_gradient'… (hình 1).

Tuy nhiên, việc sử dụng số lượng feature một cách hợp lý, loại bỏ những features quá "liên kết chặt" với nhau sẽ làm tăng tính khách quan của tập hợp số liệu đầu vào. Để loại bỏ các thuộc tính có hệ số tương quan cao, tác giả đã sử dụng thuật toán 'Pearson' (linear correlation). Trong hình 2, các thuộc tính có R>|0,7| đã được loại bỏ khỏi đầu vào của hệ thống ML.

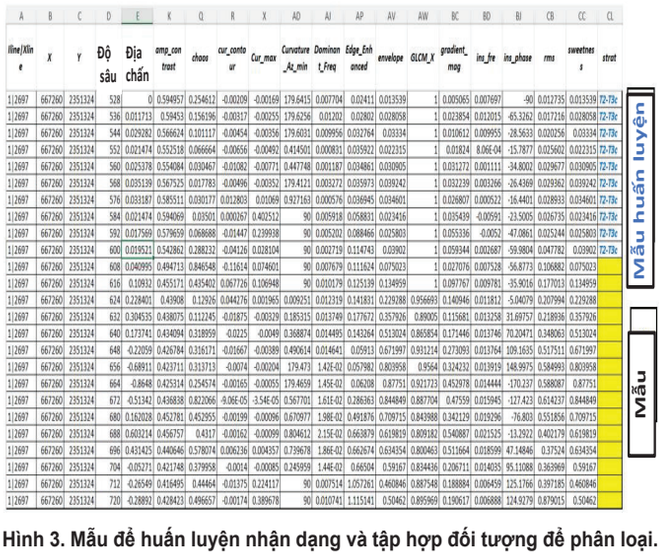
(A)

	org	amp_contrast	chaos	cur_contour	Cur_max	Cur_max_avg	Cur_max_std	Cur_max_dif	Cur_max_1derv	Cur_max_curvature	Curvature_Az_min	Dominant_Freq	Edge_Enhanced	envelope	GLCM_X	gradient_mag	ins_fre	ins_phase	Polarity	RAI	rms	sweetness	trace_gradient
org	1.00	-0.01	-0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.29	0.04	-0.13	0.01	0.00	-0.01	0.06	0.20	0.03	0.04	0.00
amp_contrast	-0.01	1.00	-0.66	-0.71	-0.09	-0.11	0.02	0.00	-0.02	0.01	-0.23	0.43	0.01	0.00	-0.37	0.02	0.39	0.05	-0.05	0.00	0.00	0.00	0.00
chaos	-0.02	-0.66	1.00	0.80	0.05	0.06	0.09	0.00	-0.01	0.11	0.18	-0.45	-0.10	-0.11	0.41	-0.12	-0.43	-0.01	0.05	-0.01	-0.11	-0.11	-0.01
cur_contour	0.00	-0.71	0.80	1.00	0.01	0.00	-0.10	0.08	0.10	-0.08	0.10	-0.51	-0.08	-0.08	0.38	-0.08	-0.48	0.00	0.00	0.00	-0.09	-0.08	0.00
Cur_max	0.00	-0.09	0.05	0.01	1.00	0.94	0.44	0.23	-0.19	0.45	0.29	-0.14	-0.06	-0.06	0.06	-0.07	-0.13	-0.04	-0.02	0.00	-0.06	-0.06	0.00
Cur_max_avg	0.00	-0.11	0.06	0.00	0.94	1.00	0.47	0.01	-0.19	0.43	0.33	-0.15	-0.06	-0.07	0.07	-0.08	-0.14	-0.04	-0.03	0.00	-0.07	-0.07	0.00
Cur_max_std	0.00	0.02	0.09	-0.10	0.44	0.47	1.00	-0.01	-0.18	0.86	0.11	-0.12	-0.05	-0.06	0.07	-0.06	-0.11	-0.06	0.09	0.00	-0.06	-0.06	0.00
Cur_max_dif	0.00	0.00	0.00	0.08	0.23	0.01	-0.01	1.00	0.29	0.07	-0.03	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00
Cur_max_1derv	0.00	-0.02	-0.01	0.10	-0.19	-0.19	-0.18	0.29	1.00	-0.06	-0.03	0.01	-0.01	-0.01	-0.01	-0.01	0.00	0.02	0.01	-0.01	-0.01	-0.01	0.00
Cur_max_curvature	0.00	0.01	0.11	-0.08	0.45	0.43	0.86	0.07	-0.06	1.00	0.10	-0.16	-0.07	-0.07	0.08	-0.07	-0.15	-0.05	0.11	0.00	-0.07	-0.07	0.00
Curvature_Az_min	-0.01	-0.23	0.18	0.10	0.29	0.33	0.11	-0.03	-0.03	0.10	1.00	-0.05	0.00	0.01	0.09	0.00	-0.02	0.00	0.06	-0.01	0.01	0.01	0.00
Dominant_Freq	0.00	0.43	-0.45	-0.51	-0.14	-0.15	-0.12	0.00	0.01	-0.16	-0.05	1.00	0.18	0.20	-0.17	0.22	0.95	0.03	0.04	-0.02	0.21	0.20	0.01
Edge_Enhanced	0.29	0.01	-0.10	-0.08	-0.06	-0.06	-0.05	0.00	-0.01	-0.07	0.00	0.18	1.00	0.91	-0.09	0.81	0.20	0.03	0.05	0.14	0.90	0.91	-0.15
envelope	0.04	0.00	-0.11	-0.08	-0.06	-0.07	-0.06	0.00	-0.01	-0.07	0.01	0.20	0.91	1.00	-0.01	0.93	0.22	0.00	0.03	-0.06	1.00	1.00	0.03
GLCM_X	-0.13	-0.37	0.41	0.38	0.06	0.07	0.07	0.00	-0.01	0.08	0.09	-0.17	-0.09	-0.01	1.00	-0.01	-0.15	-0.51	-0.03	-0.22	-0.02	-0.01	0.17
gradient_mag	0.01	0.02	-0.12	-0.08	-0.07	-0.08	-0.06	0.00	-0.01	-0.07	0.00	0.22	0.81	0.93	-0.01	1.00	0.23	0.01	0.02	-0.10	0.95	0.93	0.11
ins_fre	0.00	0.39	-0.43	-0.48	-0.13	-0.14	-0.11	0.00	0.00	-0.15	-0.02	0.95	0.20	0.22	-0.15	0.23	1.00	0.04	0.04	-0.02	0.23	0.22	0.01
ins_phase	-0.01	0.05	-0.01	0.00	-0.04	-0.04	-0.06	0.00	0.02	-0.05	0.00	0.03	0.03	0.00	-0.51	0.01	0.04	1.00	-0.01	0.21	0.01	0.00	-0.18
Polarity	0.06	-0.05	0.05	0.00	-0.02	-0.03	0.09	-0.01	0.01	0.11	0.06	0.04	0.05	0.03	-0.03	0.02	0.04	-0.01	1.00	-0.02	0.03	0.03	0.03
RAI	0.20	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	-0.01	-0.02	0.14	-0.06	-0.22	-0.10	-0.02	0.21	-0.02	1.00	-0.03	-0.06	-0.73
rms	0.03	0.00	-0.11	-0.09	-0.06	-0.07	-0.06	0.00	-0.01	-0.07	0.01	0.21	0.90	1.00	-0.02	0.95	0.23	0.01	0.03	-0.03	1.00	1.00	0.02
sweetness	0.04	0.00	-0.11	-0.08	-0.06	-0.07	-0.06	0.00	-0.01	-0.07	0.01	0.20	0.91	1.00	-0.01	0.93	0.22	0.00	0.03	-0.06	1.00	1.00	0.03
trace_gradient	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	-0.15	0.03	0.17	0.11	0.01	-0.18	0.03	-0.73	0.02	0.03	1.00

(B)

	org	amp_contrast	chaos	Cur_max	Curvature_Az_min	Dominant_Freq	GLCM_X	ins_phase	Polarity	sweetness	trace_gradient
org	1										
amp_contrast		1									
chaos			1								
Cur_max				1							
Curvature_Az_min					1						
Dominant_Freq						1					
GLCM_X							1				
ins_phase								1			
Polarity									1		
sweetness										1	
trace_gradient											1

Hình 2. Ma trận PC tương quan các thuộc tính địa chấn. (A) gốc và **(B)** đã loại bỏ các thuộc tính có R cao.

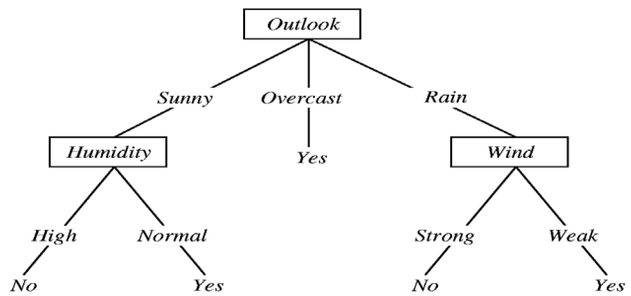


Hình 3. Mẫu để huấn luyện nhận dạng và tập hợp đối tượng để phân loại.

Dữ liệu đầu vào cho cấu trúc mạng DT gồm 2 phần: 1) mẫu để huấn luyện nhận dạng và 2) tập hợp đối tượng để phân loại. Cả hai loại số liệu phải đảm bảo tính thống nhất về cấu trúc dữ liệu như nhau. Nghĩa là các "mẫu" được đặc trưng bởi các thuộc tính (features) nào thì các đối tượng cũng phải có các thuộc tính như vậy (hình 3). Cả hai loại dữ liệu này được tích hợp thành một hệ thống GIS-DB đảm bảo tính thống nhất trong không gian [2].

2.2. Mô hình cây quyết định và thuật toán XGBoost

DT là một thuật toán máy học phổ biến được sử dụng để giải quyết các bài toán phân loại và dự đoán. Nó được gọi là DT bởi vì nó bắt đầu từ một nút gốc và phân nhánh ra thành các nút lá dựa trên các quyết định được đưa ra tại các nút nội bộ trên đường đi từ gốc đến lá. Một DT thường được xây dựng bằng cách chia tập dữ liệu thành các tập con nhỏ hơn dựa trên các đặc trưng của dữ liệu (hình 4). Các đặc trưng này được chọn dựa trên mức độ quan trọng và khả năng phân loại của chúng. Việc chia tập dữ liệu tiếp tục được thực hiện cho đến khi không thể phân loại các mẫu dữ liệu tốt hơn hoặc cây đạt đến độ sâu được xác định trước DT thường được sử dụng trong phân tích quyết định, để giúp xác định một chiến lược có nhiều khả năng đạt được mục tiêu nhất, nhưng cũng là một công cụ phổ biến trong học máy.



Hình 4. Minh họa mô hình cây quyết định.

Mục đích là tạo ra một mô hình dự đoán giá trị của một biến mục tiêu dựa trên một số biến đầu vào. DT là một biểu diễn đơn giản để phân loại số liệu. Mỗi nút trên cây đại diện cho một quyết định và nó bao gồm các thuộc tính như: đặc trưng được sử dụng để phân loại mẫu dữ liệu; ngưỡng giá trị đặc trưng được sử dụng để phân loại; số lượng mẫu dữ liệu đã được phân loại dựa trên quyết định đó; số lượng mẫu dữ liệu không được phân loại dựa trên quyết định đó.

Cây quyết định thường được sử dụng để phân loại dữ liệu dạng rời rạc (discrete) và liên tục (continuous), có thể được sử dụng cho cả bài toán phân loại đa lớp (multiclass) và nhị phân (binary).

Việc xây dựng DT có thể dẫn đến hiện tượng quá khớp (overfitting), khi mô hình quá phức tạp và không tổng quát hóa được cho các dữ liệu mới. Để giải quyết vấn đề này, các kỹ thuật như cắt tỉa cây (pruning) và thay thế các nút bằng phân bố xác suất (probabilistic splitting) có thể được sử dụng.

Cây quyết định được xây dựng bằng cách chia tập dữ liệu gốc, bắt đầu từ nút gốc của cây, sau đó tạo ra các tập con kế tiếp. Việc chia tách này dựa trên các quy tắc phân chia được thiết lập dựa trên các đặc điểm phân loại. Quá trình này được lặp đi lặp lại một cách đệ quy, được gọi là phân vùng đệ quy. Quá trình đệ quy kết thúc khi tập hợp con tại một nút có tất cả các giá trị giống nhau của biến mục tiêu hoặc khi việc phân tách không mang lại cải thiện cho dự đoán. Phương pháp suy diễn từ trên xuống của DT là một ví dụ về thuật toán tham lam và cho đến nay vẫn là chiến lược phổ biến nhất để xây dựng cây quyết định từ dữ liệu.

Ưu điểm của DT là:

- Dễ hiểu và diễn giải: DT có cấu trúc rõ ràng, dễ hiểu và diễn giải, do đó dễ dàng giải thích cho những người không có kinh nghiệm về ML.

- Khả năng xử lý các dữ liệu lớn: DT có khả năng xử lý các tập dữ liệu lớn, có thể xử lý các tập dữ liệu chứa hàng triệu bản ghi.

- Tính linh hoạt: DT có tính linh hoạt cao trong việc chọn các tính năng và tham số để tối ưu hóa kết quả.

- Hiệu quả tính toán: DT có thể được xây dựng và tối ưu hóa bằng các thuật toán hiệu quả, do đó, thời gian tính toán và tài nguyên cần thiết để xây dựng một DT là tương đối thấp.

- Có thể xử lý dữ liệu không đầy đủ: DT có khả năng xử lý các tập dữ liệu không đầy đủ hoặc bị thiếu thông tin.

Nhược điểm của DT:

- Dễ bị quá khớp: DT có thể trở nên quá phức tạp và dễ bị quá khớp với dữ liệu huấn luyện nếu không được tối ưu hóa một cách chính xác.

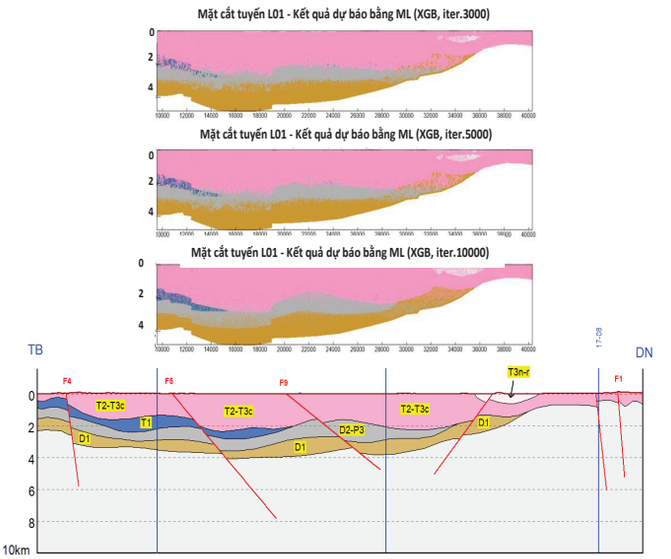
- Nhạy cảm với nhiễu và các giá trị ngoại lai: DT có thể dễ bị ảnh hưởng bởi các giá trị ngoại lai hoặc nhiễu trong dữ liệu.

- Không hiệu quả với các dữ liệu dạng liên tục: DT không hiệu quả trong việc xử lý các tập dữ liệu dạng liên tục.

- Khó khăn trong việc xử lý các mối quan hệ phức tạp giữa các tính năng: DT có thể không hiệu quả trong việc xử lý các mối quan hệ phức tạp giữa các trường thông tin.

Thuật toán XGBoost [8] là viết tắt của eXtreme Gradient Boosting. "eXtreme" nói đến các cải tiến về tốc độ như tính toán song song và nhận biết bộ nhớ cache giúp XGBoost nhanh hơn khoảng 10 lần so với tăng cường Gradient truyền thống. Ngoài ra, XGBoost còn bao gồm một thuật toán tìm kiếm phân tách độc đáo để tối ưu hóa cây, cùng với tính năng chính quy tích hợp giúp giảm việc trang bị quá mức. Nói chung, XGBoost là phiên bản Gradient Boost nhanh hơn, chính xác hơn. Vì XGBoost là phiên bản nâng cao của Gradient Boosting và kết quả của nó là không song song nên nó có thể được cho là ML tốt nhất chúng ta có hiện nay.

3. Kết quả và bàn luận



Hình 5. Kết quả nhận dạng tướng đất đá theo DT-XGBoost.

Nghiên cứu đã sử dụng các lát cắt địa chất truyền thống đã xây dựng để làm "mẫu luyện mạng" (hình 5). Số lượng mẫu học chỉ chiếm 60% và hoàn toàn ngẫu nhiên về vị trí XY. Cấu trúc mạng DT-XGBoost đã được sử dụng thử nghiệm với các số iteration lần lượt là 3.000, 5.000 và 10.000. Mục đích sử dụng DT-XGBoost là xác định mối liên hệ giữa mặt cắt địa chất đã biết (mẫu học) và dữ liệu địa chấn (số liệu gốc và các thuộc tính địa chấn). Kết quả nhận dạng được trình bày trong hình 6.

Hình 6. Sai số nhận dạng tướng đất đá với số lần lặp. (A) >3.000, **(B)** >5.000, **(C)** >10.000.

Các tham số thống kê đánh giá sai số được trình bày trong hình 6 cho thấy, sai số nhận dạng phân loại với 3000 lần lặp còn thấp: nhận dạng sai 16%, đặc biệt đất đá T_3n-r và T1 có sai số cao (75-80%). Khi phân loại với số lần lặp trên 10.000 lần thì độ chính xác đã được cải thiện đáng kể (hình 6C).

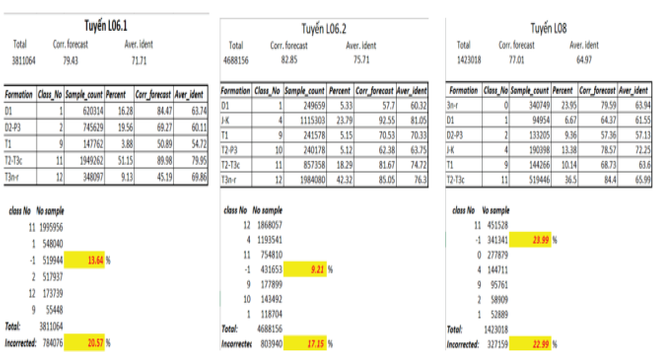
Các tham số thống kê của số liệu địa chấn và 11 thuộc tính địa chấn cho thấy nếu chỉ sử dụng các tham số thống kê với các thuật toán thống kê truyền thống để phân loại đất đá thì rất khó khăn, vì các nhóm tham số có giá trị rất gần nhau và cấu trúc số liệu không hoàn toàn phản ánh cấu trúc địa chất của các tuyến (hình 1). Các thuật toán ML đã chứng tỏ được khả năng vượt trội của mình so với minh giải theo các phương pháp truyền thống trong công tác nhận dạng/phân loại đối tượng theo các thuộc tính. Đó là khả năng quản trị dữ liệu lớn: Bảng số liệu tích hợp các tuyến địa chấn có dung lượng khoảng xấp xỉ 10TB (hàng triệu dòng và gần 80 cột); khả năng khai thác các mối quan hệ ẩn sâu giữa các cột thuộc tính từ đó nhận dạng các tướng đá với độ chính xác cao.

Các tham số thống kê mức độ chính xác nhận dạng phân loại đất đá bằng ML và truyền thống cho thấy (hình 7 và 8) [9]:

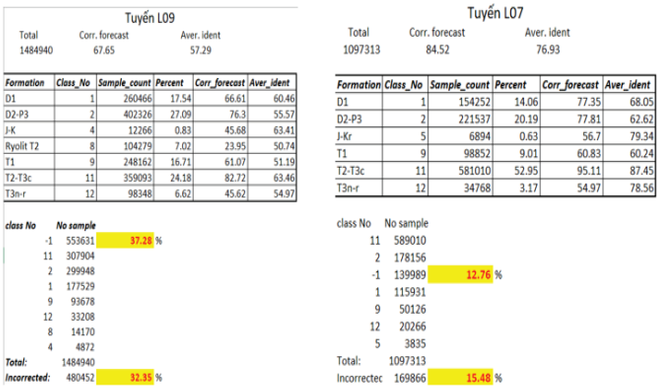
Hệ tầng Nà Khuất, Mẫu Sơn (T_2-T_3c) gồm sét vôi, bột kết, cát kết, đá phiến sét, đá vôi; sạn kết, cát kết, sét vôi, cát kết màu đỏ; bột kết xen cát kết màu đỏ; Phân hệ tầng dưới: cát kết, cát kết dạng quarzit, thấu kính cuội kết, bột kết màu đỏ được nhận dạng chính xác (~90% đúng).

Hệ tầng Tam Hoa, Bắc Sơn, Đồng Đăng (D_2-P_3) với thành phần đất đá gồm cuội kết, sạn kết, cát kết arko, đá phiến sét, đá vôi, đá vôi dolomit, bauxit, bột kết, đá vôi, sét silic được nhận dạng với độ chính xác trung bình cao (70-90% đúng).

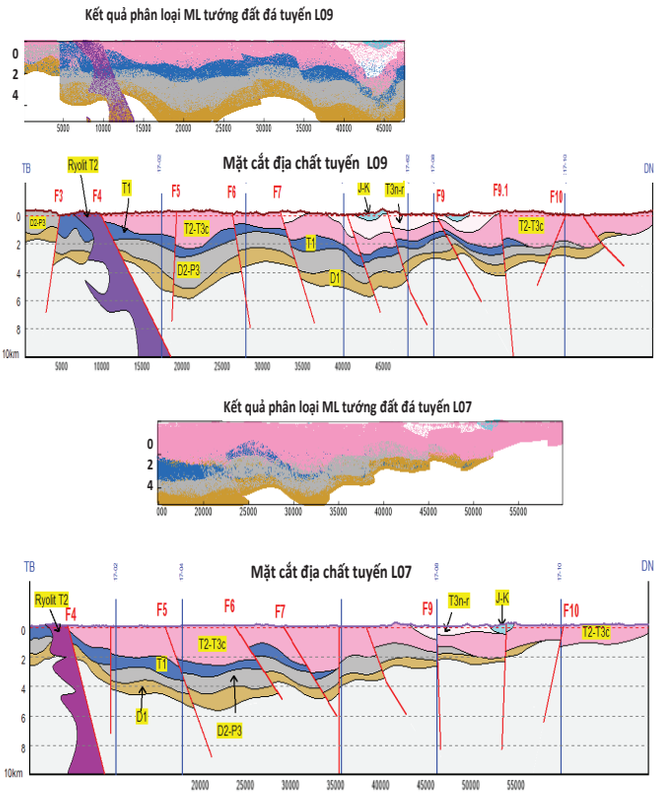
Ngược lại, hệ tầng Văn Lãng (T_3n-r) gồm cát kết, bột kết, sét - bột kết màu đen, xám đen, đá vôi - silic, sét than lại có độ chính xác nhận dạng trung bình thấp nhất (~45-50%).



Hình 7. Sai số nhận dạng tướng đất đá trên các mặt cắt L06.1, L06.2 và L08.



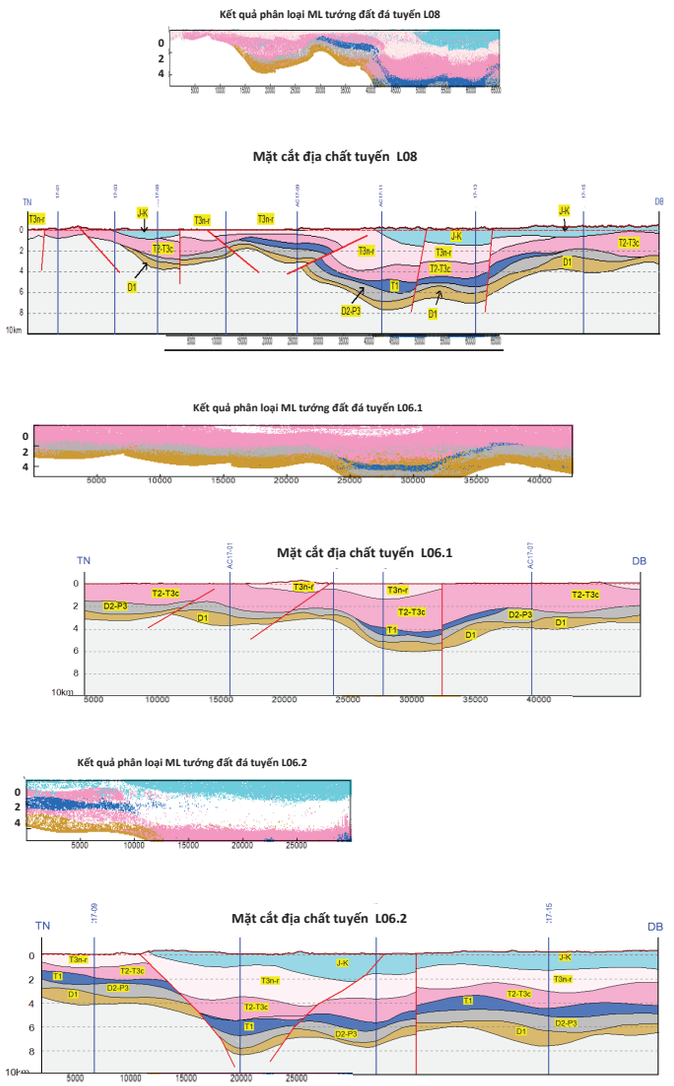
Hình 8. Sai số nhận dạng tướng đất đá trên các mặt cắt L09 và L07.



Hình 9. Mặt cắt địa chất và kết quả nhận dạng tướng đất đá trên các mặt cắt L09 và L07.

Theo kết quả xây dựng mặt cắt địa chấn - địa chấn trên cơ sở nhận dạng theo dữ liệu địa chấn (hình 9 và 10) có thể thấy: ML đã khẳng định có mối liên quan chặt chẽ giữa các hệ tầng trên mặt cắt địa chất do các chuyên gia truyền thống xây dựng và dữ liệu địa chấn, đặc biệt là dữ liệu thuộc tính địa chấn. Đa số các đoạn tuyến đều có tính phù hợp cao giữa mặt cắt địa chất đã biết và kết quả dự báo bằng ML.

Tuy nhiên, DT (XGB) cũng phát hiện một vài chỗ không phù hợp giữa kết quả của truyền thống và liên kết địa chấn [10]:



Hình 10. Mặt cắt địa chất và kết quả nhận dạng tướng đất đá trên các mặt cắt L08, L06.1 và L06.2.

Trên tuyến L09, đoạn 40.000-47.000 m, độ sâu 2.000-3.000 m. Khả năng hệ tầng Lạng Sơn ($T_{l,2}$, cát kết, bột kết, đá phiến sét, đá vôi sét, sét vôi) có thể nằm ngang, không tạo nên đứt đẫy.

Trên tuyến L08, đoạn 40.000-45.000 m, hệ tầng Lạng Sơn không có biểu hiện rõ và vị trí đứt gẫy được xác định ở độ sâu nông hơn.

Trên dữ liệu địa chấn, các khối ryorit thể hiện khá mờ nhạt và không rõ ràng với độ nhận dạng không cao (29%).

4. Kết luận

Kết quả cho thấy, các phần mềm được phát triển trong phạm vi nghiên cứu cũng như hệ thống máy tính công suất cao (High Performance Computer - HPC) đáp ứng được khả năng xử lý cơ sở dữ liệu khổng lồ bằng các thuật toán tiên tiến. Tuy nhiên, để đáp ứng được công tác xử lý một lượng lớn số liệu địa chất - địa vật lý rất cần các thuật toán mới, phương pháp mới, hướng đi mới để khắc phục được các yếu điểm của hệ thống truyền thống hiện nay. Đó là khả năng tổng hợp cơ sở dữ liệu nhiều cả về số lượng và định dạng; xác định các mối liên kết giữa dầu khí và thông tin địa chất - địa vật lý.

Với thử nghiệm mô hình cấu trúc ML dạng DT và thuật toán Gradient Boosting (GB) để phân loại có kiểm định (có mẫu học) trên cơ sở tài liệu địa chất - địa chấn bước đầu đã thu được một số kết quả nhất định:

Thử nghiệm thành công tích hợp dữ liệu địa chất - địa chấn thành cơ sở dữ liệu lớn theo nguyên tắc GIS.

Phát hiện các mối quan hệ ẩn sâu giữa các lớp thông tin qua đó nhận dạng với độ chính xác chấp nhận được các đối tượng địa chất theo mẫu đã biết.

Với DT - GB số lần lặp phải trên 10.000 lần thì kết quả nhận dạng mới chấp nhận được. Thời gian tính toán còn quá lâu (~60-90 giờ chạy máy) để đạt được kết quả cao.

LỜI CẢM ƠN

Nghiên cứu này là kết quả đề tài "Nghiên cứu xây dựng hệ thống trí tuệ nhân tạo tích hợp dữ liệu địa chất dầu khí đánh giá triển vọng dầu khí" (mã số KC-4.0-01/19-25) thuộc Chương trình trọng điểm quốc gia "Hỗ trợ nghiên cứu, phát triển và ứng dụng công nghệ của công nghiệp 4.0". Các tác giả xin trân trọng cảm ơn.

TÀI LIỆU THAM KHẢO

[1] D.N. San (2003), *Conducting and Completing Technology for Synthesizing and Analyzing Data using Geological Geographic Information Systems and Geological Expert Systems, Applied for in the Cho Don Area*, Ministerial level research project (in Vietnamese).

[2] D.N. San, N.D. Nuong (2019), *Research and Development of an Artificial Intelligence System Integrated with Petroleum Geological Data for Petroleum Prospect Evaluation*, Government level research project (in Vietnamese).

[3] N.T. San (1975), "How should the petroleum prospects in the An Chau Basin be assessed", *Geological Journal*, **122(11-12)**, 6pp (in Vietnamese).

[4] Schlumberger (2015), *3D Seismic Facies Generated From Independent Seismic Attributes Using The Neural Network Technique*, 56pp.

[5] M.T. Tan (2011), *Seismic Exploration,* Transport Publishing House, 45pp (in Vietnamese).

[6] S. Chopra, K.J. Marfurt (2005), *Seismic Attribute Mapping of Structure and Stratigraphy,* EAGE, 30pp.

[7] L. Zhang, C. Zhan (2017), "Machine learning in rock facies classification: An application of XGBoost", *International Geophysical Conference*, Qingdao, China, Society of Exploration Geophysicists and Chinese Petroleum Society, DOI: 10.1190/IGC2017-351.

[8] Boosting Algorithm (2020), *XGBoost*, Towards Data Science, 2017-05-14.

[9] PVEP Song Hong (2019), *Phase I Report of The An Chau Basin Research Project* (in Vietnamese).

[10] PVEP Song Hong (2013), *Petrographic Analysis Report of The An Chau Basin Region* (in Vietnamese).