XÂY DỰNG MÔ HÌNH THẠCH HỌC CHO MỎ ĐÁ VÔI TÀ THIẾT BẰNG SỬ DỤNG KỸ THUẬT ĐỊA THỐNG KÊ VÀ HỌC MÁY

Vũ Đình Trọng[1,*], Nguyễn Văn Đức[1]
[1]*Trường Đại học Công nghiệp Quảng Ninh*
[*]*Email: trongvu@qui.edu.vn*

TÓM TẮT

Việc xây dựng mô hình thạch học 3D tại mỏ đá vôi Tà Thiết gặp nhiều khó khăn do mạng lưới lỗ khoan thưa thớt và dữ liệu mất cân bằng giữa lớp đá vôi (đa số) và các lớp kẹp như sét, đá ong (thiểu số). Nghiên cứu này đề xuất mô hình lai (Hybrid RF), kết hợp thuật toán Random Forest (RF) và kỹ thuật Indicator Kriging (IK). Bằng cách sử dụng các phân phối xác suất không gian từ IK làm đặc trưng đầu vào, mô hình Hybrid RF giúp thuật toán nắm bắt được bối cảnh địa chất cục bộ. Kết quả cho thấy mô hình lai đạt độ chính xác tổng thể 93,0% (vượt trội so với RF tiêu chuẩn 80,0% và IK 61,0%), đồng thời cải thiện 145% chỉ số F1-score cho các lớp đá thiểu số. Mô hình lai cho phép tái hiện chính xác các ranh giới thạch học sắc nét. Đây là công cụ tin cậy hỗ trợ khai thác chọn lọc và tối ưu hóa phối trộn nguyên liệu xi măng.

Từ khóa: *Mô hình thạch học 3D, Địa thống kê, Học máy, Indicator Kriging, Random Forest, Dữ liệu mất cân bằng.*

1. ĐẶT VẤN ĐỀ

Trong nhiều năm qua, các mô hình địa chất 3D đã được sử dụng như một công cụ quan trọng trong quản trị tài nguyên [1, 2]. Chúng giúp các nhà địa chất và kỹ sư mỏ hiểu rõ sự phức tạp của các cấu trúc địa chất dưới bề mặt, hỗ trợ quá trình đánh giá và lập kế hoạch để quản lý, sử dụng hiệu quả nguồn tài nguyên khoáng sản [3, 4]. Độ chính xác của các mô hình này có tầm quan trọng sống còn, vì nó ảnh hưởng trực tiếp đến các quá trình sản xuất phía sau từ khía cạnh kinh tế, kỹ thuật đến môi trường [5]. Đặc biệt, trong công nghiệp sản xuất xi măng, quy trình quản trị chất lượng phụ thuộc chặt chẽ vào một mô hình thạch học 3D chính xác [3, 6]. Mô hình này cho phép lập kế hoạch khai thác hiệu quả thông qua việc khai thác chọn lọc các đới đá vôi chất lượng cao và phối trộn các loại vật liệu khác nhau để đảm bảo nguồn cung cấp nguyên liệu ổn định cho nhà máy [3, 7].

Việc xây dựng mô hình thạch học chính xác từ dữ liệu lỗ khoan thưa thớt và mất cân bằng đối mặt với những rào cản kỹ thuật đáng kể khi áp dụng các phương pháp đơn lẻ. Mặc dù là công cụ phổ biến trong ước tính tài nguyên, các phương pháp địa thống kê truyền thống thường tạo ra hiệu ứng "làm trơn" (smoothing effects) các thuộc tính địa chất [8-10]. Điều này dẫn đến việc mô hình khó tái tạo được các ranh giới thạch học sắc nét hoặc các chi tiết địa chất có độ phân giải cao [11, 12]. Trong điều kiện mạng lưới lỗ khoan thăm dò thưa thớt (khoảng cách từ 50m đến 100m), khả năng xác định tính liên tục không gian của IK bị hạn chế đáng kể [12].

Các thuật toán học máy (ML) như Random Forest (RF), khi chỉ được huấn luyện dựa trên các tọa độ không gian thô (x, y, z), thường gặp hiện tượng nhầm lẫn về tính liên tục không gian [14]. Do không có các ràng buộc về quy luật địa chất toàn cục, mô hình ML tiêu chuẩn dễ tạo ra các dự đoán rời rạc, thiếu sự kết nối tự nhiên giữa các thực thể thạch học. Một thách thức lớn khác là vấn đề dữ liệu mất cân bằng. Các thuật toán ML tiêu chuẩn có xu hướng ưu tiên học các lớp chiếm đa số (như Đá vôi) và thường bỏ qua hoặc dự báo sai các lớp chiếm tỷ trọng nhỏ (như Sét, Laterite và Đất) [14]. Mặc dù các thuật toán ML tiêu chuẩn có độ chính xác cao hơn IK , nhưng nó vẫn gặp khó khăn trong việc học các

bối cảnh địa chất phức tạp nếu không được bổ sung các đặc trưng bổ trợ từ địa thống kê.

Nhằm vượt qua những rào cản kỹ thuật của các phương pháp đơn lẻ, nghiên cứu này đề xuất một chiến lược mô hình hóa lai mới, kết hợp chặt chẽ giữa kỹ thuật Địa thống kê và thuật toán Học máy. Ý tưởng cốt lõi của phương pháp là sử dụng Indicator Kriging (IK) không phải như một công cụ dự báo cuối cùng, mà như một bộ lọc đặc trưng không gian. Cụ thể, các phân phối xác suất không gian được trích xuất từ IK cho từng loại thạch học sẽ được sử dụng làm biến đầu vào bổ trợ cho thuật toán Random Forest (RF), bên cạnh các tọa độ không gian truyền thống.

Sự kết hợp này cho phép mô hình tận dụng khả năng mô hình hóa cấu trúc không gian và tính liên tục địa chất của IK, đồng thời phát huy sức mạnh của RF trong việc xử lý các mối quan hệ phi tuyến phức tạp và dữ liệu mất cân bằng. Phương pháp này được kỳ vọng sẽ giải quyết triệt để bài toán dự báo thạch học tại mỏ đá vôi Tà Thiết, nơi có mạng lưới lỗ khoan thưa thớt và sự hiện diện của các thấu kính sét, laterite chiếm tỷ trọng nhỏ nhưng có tác động lớn đến chất lượng nguyên liệu.

Nghiên cứu tập trung vào việc chứng minh rằng mô hình lai không chỉ cải thiện đáng kể độ chính xác tổng thể so với các mô hình tiêu chuẩn, mà còn có khả năng tái hiện các cấu trúc địa chất thực tế hơn, đặc biệt là nâng cao hiệu suất dự báo cho các nhóm thạch học thiểu số. Đây một yếu tố then chốt trong việc giảm thiểu rủi ro địa chất và tối ưu hóa quy trình phối trộn nguyên liệu tại nhà máy xi măng.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Khu vực nghiên cứu và dữ liệu

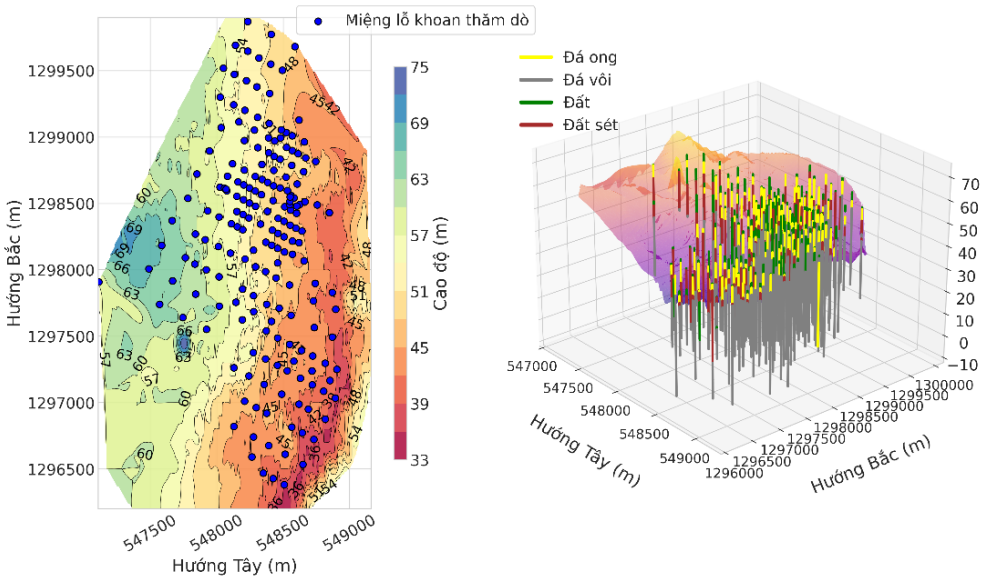
Đối tượng nghiên cứu của bài báo này là mỏ đá vôi Tà Thiết, một nguồn tài nguyên quan trọng tọa lạc tại tỉnh Bình Phước, Việt Nam. Hiện nay, mỏ đang được khai thác bằng phương pháp lộ thiên bởi Công ty Xi măng Hà Tiên nhằm cung cấp nguyên liệu thô trực tiếp cho nhà máy xi măng Bình Phước, nằm cách khu vực khai thác khoảng 7km. Do yêu cầu khắt khe về chất lượng đầu vào của dây chuyền sản xuất xi măng, việc quản lý chất lượng tại mỏ dựa trên một mô hình thạch học 3D chính xác là yếu tố quyết định để lập kế hoạch khai thác chọn lọc và phối trộn nguyên liệu tối ưu.

Tập dữ liệu sử dụng trong nghiên cứu này được thu thập từ mạng lưới thăm dò gồm 194 lỗ khoan. Trên Hình 1 (trái) có thể thấy các lỗ khoan này được bố trí theo dạng lưới không đồng nhất với khoảng cách trung bình dao động từ 50m × 50m đến 100m × 100m, tạo nên một hệ thống dữ liệu thăm dò thưa thớt. Cấu trúc dữ liệu đầu vào bao gồm: Mã số định danh duy nhất cho từng vị trí thăm dò (hole_id), Vị trí tuyệt đối theo các trục x, y, và z, chiều sâu tối đa của lỗ khoan và các khoảng phân lớp thạch học chi tiết dọc theo thân lỗ khoan.

Table 1. Phân bố đất đá theo chiều dài mẫu lõi khoan thăm dò.

Loại đất đá	Tổng chiều dài mẫu lõi khoan (m)	Tỷ lệ (%)
Đá vôi	4.897	81,31
Đá ong	545	9,06
Sét	451	7,49
Đất	130	2,15

Hình 1 (phải) cho thấy cấu trúc địa chất tại mỏ Tà Thiết bao gồm 4 loại đất đá chính: đất, sét , đá ong và đá vôi. Các loại đá thể hiện một xu hướng phân lớp theo chiều đứng khá rõ rệt; trong đó đá vôi tập trung chủ yếu ở phần dưới sâu, trong khi các lớp đất, đá ong và sét thường xuất hiện ở phần trên của cấu trúc địa tầng. Trong Bảng 1, tập dữ liệu tồn tại sự chênh lệch lớn về số lượng mẫu giữa các lớp thạch học. Đá vôi đóng vai trò là loại đá chiếm đa số, trong khi đất, sét và đá ong được coi là các nhóm thiểu số. Sự mất cân bằng nghiêm trọng này, kết hợp với mạng lưới thăm dò thưa thớt, là nguyên nhân chính khiến các phương pháp dự báo truyền thống thường bỏ sót hoặc dự báo sai các ranh giới thạch học của các lớp kẹp (sét, đá ong), gây khó khăn cho việc kiểm soát chất lượng nguyên liệu.



Hình 1. Bố trí lỗ khoan thăm dò (trái) và phân bố thạch học trong các lỗ khoan (phải) tại mỏ đá vôi Tà Thiết, Bình Phước.

2.2. Chuẩn bị dữ liệu

Để áp dụng hiệu quả kỹ thuật Địa thống kê (cụ thể là Indicator Kriging - IK) và tạo ra các đặc trưng không gian cho mô hình Học máy, dữ liệu thạch học ban đầu cần được chuyển đổi sang định dạng số học. Trong nghiên cứu này, quá trình chuyển đổi được thực hiện qua hai bước chính nhằm chuẩn hóa dữ liệu lỗ khoan thưa thớt:

Bước 1: Trước khi mã hóa, dữ liệu các khoảng phân lớp thạch học dọc theo thân lỗ khoan được chuẩn hóa thành các đoạn mẫu có chiều dài đều nhau là 1 mét. Việc này giúp đảm bảo mỗi điểm dữ liệu có trọng số không gian tương đương nhau trong quá trình nội suy và tính toán, tránh hiện tượng thiên lệch do các lớp đá có độ dày mỏng khác nhau.

Bước 2: Dữ liệu phân loại thạch học từ các đoạn mẫu 1 mét sau đó được chuyển đổi thành bốn biến chỉ báo nhị phân riêng biệt tương ứng với 4 loại đất đá: Đất, Sét, Đá ong và Đá vôi. Về mặt toán học, hàm chỉ báo Ik(x) cho một loại thạch học k $\in$ { Đất, Sét, Đá ong, Đá vôi} tại vị trí không gian x được xác định bằng công thức:

$$I_k(x) = \begin{cases} 1, & khi\ Z(x)=k \\ 0, & khi\ Z(x) \neq k \end{cases} \quad (1)$$

Quá trình này sẽ tạo ra 4 trường dữ liệu độc lập, tại mỗi vị trí không gian, nếu xuất hiện loại đá k thì biến chỉ báo nhận giá trị 1, ngược lại nhận giá trị 0.

2.3. Định lượng tính liên tục không gian bằng Indicator Kriging và mô hình Variogram

Sau khi đã có các biến chỉ báo nhị phân, bước quan trọng tiếp theo là định lượng hóa tính liên tục không gian của các thành phần đất đá thông qua việc phân tích Variogram. Đây là cơ sở khoa học để mô hình có thể hiểu được quy luật phân bố của đất đá giữa các khoảng trống của mạng lưới lỗ khoan thưa thớt. Một variogram thường bao gồm 3 thông số quan trọng:

- Hiệu ứng hạt (Nugget): Nếu Nugget lớn hơn 0, nó thể hiện sai số đo đạc hoặc sự biến đổi cực nhỏ không thể quan sát được.

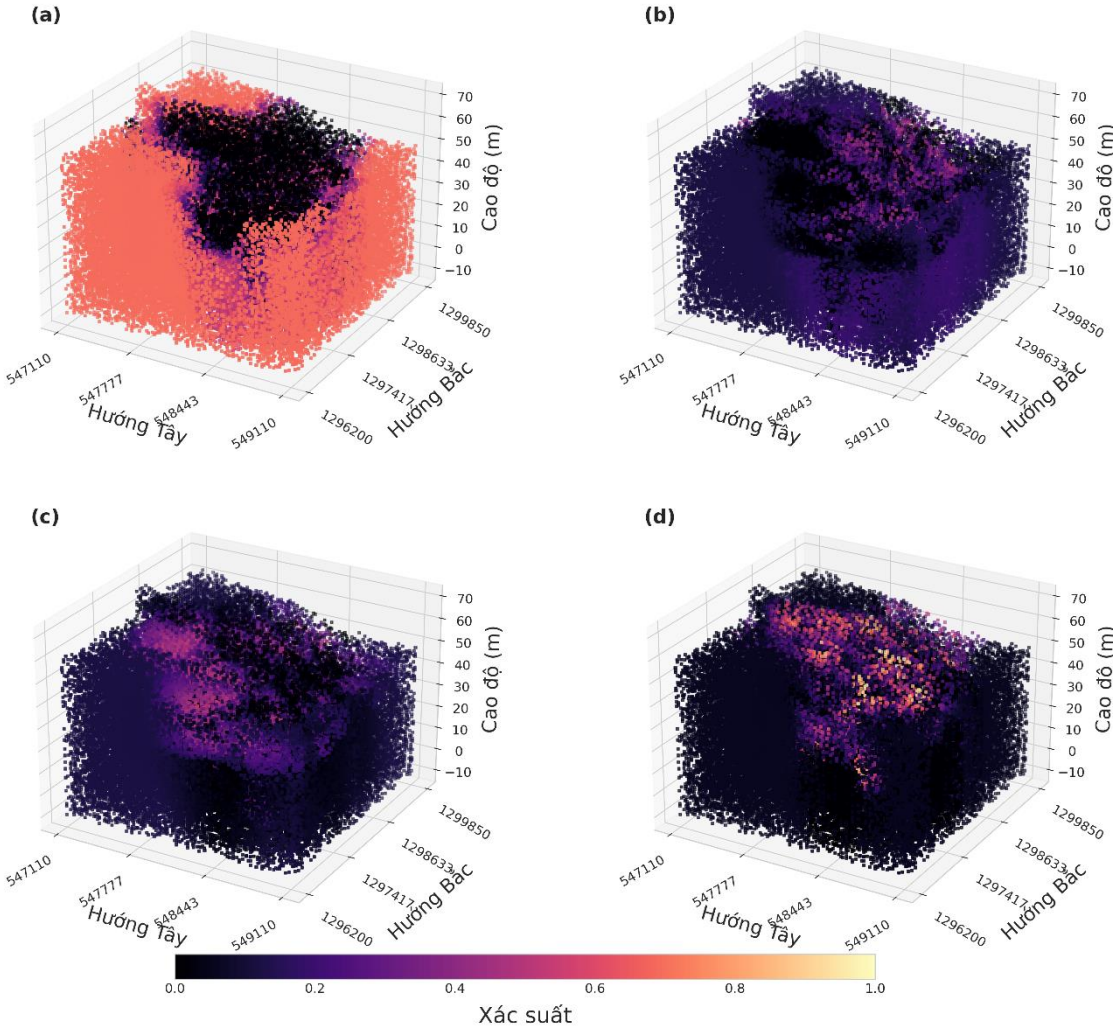
- Ngưỡng (Sill): là mức độ biến thiên tối đa của dữ liệu.

- Phạm vi ảnh hưởng (Range): Các điểm nằm trong khoảng cách này có tương quan không gian với nhau. Ngoài khoảng này, chúng được coi là độc lập.

Để tính toán nội suy, các dữ liệu thực nghiệm sẽ được khớp vào các mô hình toán học bền vững như: mô hình hình cầu (Spherical), mô hình hàm mũ (Exponential), mô hình Gaussian hoặc mô hình lũy thừa (Power).

Để phản ánh chính xác cấu trúc địa chất tại mỏ Tà Thiết, các mô hình variogram thực

nghiệm được thiết lập theo hai hướng chính: hướng ngang và hướng dọc. Việc phân tách hướng này là cần thiết do tính dị hướng (anisotropy) phổ biến trong địa chất, nơi mà tính liên tục của các lớp đá theo phương nằm ngang thường kéo dài hơn so với phương thẳng đứng. Hướng dọc phản ánh xu hướng địa chất rõ rệt theo chiều sâu, nơi đá vôi chiếm ưu thế ở tầng đáy và các lớp đất, sét, đá ong phân bố dần lên phía trên. Trong khi đó, hướng ngang: Giúp xác định phạm vi ảnh hưởng của dữ liệu giữa các lỗ khoan có khoảng cách từ 50m đến 100m. Bảng 3 tóm tắt các thông số quan trọng của mô hình variogram hình cầu được áp dụng cho từng loại đất đá. Phần mềm SGEMS được sử dụng để thu được các kết quả này.



Hình 2. Xác suất xuất hiện của các loại đất đá trong khu vực tính toán bằng phương pháp IK: (a) Đá vôi, (b) Đá ong, (c) Sét, và (d) Đất.

Dựa trên các mô hình variogram đã được hiệu chỉnh, thuật toán Indicator Kriging (IK) được thực hiện để ước tính xác suất xuất hiện của từng loại đá tại mọi vị trí trên lưới không gian 3D của toàn bộ khu vực mỏ. Thay vì chỉ đưa ra một loại đá duy nhất tại một vị trí, IK cung cấp một phân phối xác suất cho cả 4 loại thạch học. Hình 2 biểu diễn xác suất xuất hiện của từng loại đất đá trong khu vực nghiên cứu sử dụng phương pháp IK. Nhìn chung, sự phân bố của đất đá tương đồng với dữ liệu lỗ khoan với đá vôi thường nằm ở dưới cùng và chiếm tỷ trọng nhiều nhất. Trong khi đó, lớp phủ cũng là sự phân bố xen kẽ của các lớp đất, sét và đá ong. Dữ liệu này đóng vai trò là một đặc trưng quan trọng trong huấn luyện các mô hình học máy tiếp theo.

Bảng 3. Các thông số phân tích variogram cho từng loại đất đá

Loại đất đá	Hướng	Nugget	Sill	Range (m)	Cấu trúc
Đá vôi	Ngang 90^0	0,02	0,18	250	Hình cầu
	Ngang 180^0			100	
	Dọc			18	
Sét	Ngang 165^0	0,01	0,07	200	Hình cầu
	Ngang 75^0			150	
	Dọc			10	
Đá ong	Ngang 75^0	0,02	0,08	125	Hình cầu

2.3. Xây dựng Mô hình Lai (Hybrid Random Forest - Hybrid RF)

Trọng tâm của nghiên cứu này là việc thiết lập mô hình Hybrid RF, một giải pháp tích hợp sâu giữa khả năng mô hình hóa cấu trúc không gian của địa thống kê và sức mạnh phân loại phi tuyến của học máy.

Trong các mô hình học máy truyền thống, dữ liệu đầu vào thường chỉ giới hạn ở các tọa độ không gian thô X = {x, y, z}. Tuy nhiên, phương pháp Hybrid RF mở rộng không gian đặc trưng bằng cách tích hợp các giá trị xác suất không gian (Pk) thu được từ quá trình IK ở bước trước.

Véctơ đầu vào cho quá trình huấn luyện thuật toán RF lúc này được định nghĩa là:

$$X = \{x, y, z, P_{đất}, P_{sét}, P_{đá\ ong}, P_{đá\ vôi}\} \quad (2)$$

Trong đó, mỗi giá trị Pk đại diện cho xác suất địa thống kê của loại thạch học k tại vị trí tương ứng. Việc bổ sung các tham số này giúp chuyển đổi các thông tin ẩn về tính liên tục địa chất (đã được định lượng qua Variogram) thành các dữ liệu số trực tiếp mà thuật toán RF có thể xử lý được.

Trong nghiên cứu này, thuật toán RF được lựa chọn để phát triển cả hai mô hình Tiêu chuẩn (RF tiêu chuẩn) và mô hình lai (RF lai). RF được ưu tiên nhờ khả năng xử lý các cấu trúc địa chất phức tạp và làm việc hiệu quả với dữ liệu dạng bảng. Đây dữ liệu thường gặp trong mô hình hóa địa chất. Việc triển khai cả hai mô hình RF được thực hiện trong môi trường Python (phiên bản 3.10), sử dụng thư viện Scikit-learn với mô-đun RandomForestClassifier. Để tập trung vào việc đánh giá hiệu quả của các đặc trưng địa thống kê, cả hai mô hình đều sử dụng số lượng cây quyết định n_estimators = 100 và tiêu chí Gini (Gini impurity) cho việc phân tách các nút, trong khi tất cả các siêu tham số khác được giữ ở giá trị mặc định của Scikit-learn.

Để đảm bảo kết quả so sánh là khách quan và công bằng, nghiên cứu sẽ sử dụng cùng một chiến lược huấn luyện cho cả 2 mô hình RF. Đầu tiên, dữ liệu huấn luyện cả 2 mô hình sẽ được chia thành 5 phần, trong đó các mô hình sẽ được huấn luyện trên 4 phần và kiểm tra trên 1 phần còn lại. Quá trình này lặp lại 5 lần. Mỗi lỗ khoan được coi là một đơn vị riêng biệt trong quá trình chia tách. Điều này đảm bảo rằng tất cả các điểm mẫu từ cùng một lỗ khoan sẽ chỉ nằm hoàn toàn trong tập huấn luyện hoặc tập kiểm tra. Nhờ đó, mô hình không có thông tin về các lỗ khoan thử nghiệm và buộc phải học bối cảnh địa chất từ các lỗ khoan huấn luyện thay vì dựa vào các điểm dữ liệu trực tiếp phía trên hoặc phía dưới trong cùng một lỗ khoan.

Mặc dù độ chính xác tổng thể là thước đo phổ biến, nhưng nó thường gây hiểu lầm đối với tập dữ liệu mất cân bằng như tại mỏ Tà Thiết. Việc dự báo đúng các loại đá chiếm đa số có thể tạo ra độ chính xác cao, nhưng lại thất bại trong việc nhận diện các loại đá thiểu số. Do đó,

chúng tôi sử dụng ma trận nhầm lẫn (Confusion Matrix) bao gồm các chỉ số Precision, Recall, F1-Score và đường cong ROC (Receiver Operating Characteristic) để hiểu sâu hơn về hiệu suất mô hình:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

Trong đó: TP (True Positive): Số lần mô hình dự báo đúng một loại đất đá cụ thể; FP (False Positive): Số lần mô hình dự báo sai loại đất đá đó; và FN (False Negative): Số lần mô hình không nhận diện được loại đất đá thực tế đang hiện diện.

Mô hình so sánh thứ ba là mô hình IK truyền thống, thực hiện dự báo loại thạch học tại bất kỳ vị trí nào dựa trên giá trị xác suất cao nhất ước tính được. Để đảm bảo đánh giá không thiên vị, chiến lược phân chia dữ liệu huấn luyện cũng được áp dụng cho mô hình IK. Sau khi hoàn thành dự báo cho tất cả các điểm trong tập dữ liệu, các chỉ số hiệu suất được tính toán theo các công thức trên để so sánh trực tiếp với hai mô hình dựa trên RF.

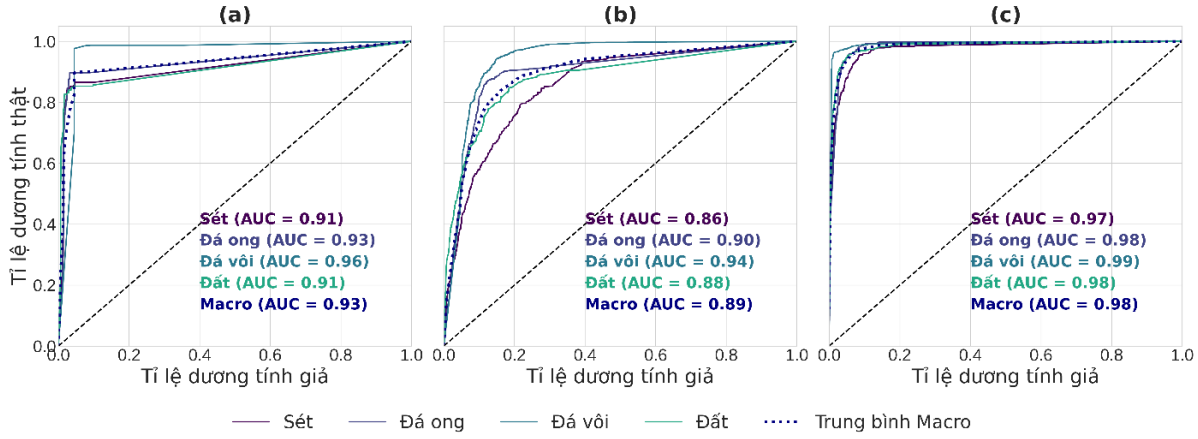
3. KẾT QUẢ VÀ THẢO LUẬN

Kết quả thực nghiệm trên tập dữ liệu kiểm tra cho thấy sự khác biệt rõ rệt về hiệu suất giữa ba phương pháp tiếp cận. Mô hình RF laiđạt độ chính xác tổng thể cao nhất là 93,0%, vượt trội hoàn toàn so với mô hình RF tiêu chuẩn (80,0%) và phương pháp IK truyền thống (61,0%). Sự chênh lệch đáng kể này (hơn 13% so với RF tiêu chuẩn và 32% so với IK) khẳng định rằng việc tích hợp các đặc trưng địa thống kê vào mô hình học máy đã giúp khắc phục hiệu quả những hạn chế của từng phương pháp đơn lẻ. Một trong những thành công lớn nhất của mô hình RF lai là khả năng dự báo chính xác các loại đá chiếm tỷ trọng thấp như sét, đất và đá ong. Trong khi mô hình IK thường gặp hiệu ứng làm trơn dẫn đến việc bỏ sót các thấu kính nhỏ, và RF tiêu chuẩn dễ bị thiên lệch về phía lớp chiếm đa số (đá vôi), thì RF lai đã cải thiện chỉ số F1-score cho các nhóm thiểu số này lên tới 145% so với các mô hình tiêu chuẩn.

Table 2. Kết quả huấn luyện 3 mô hình

Thông số	Loại đất dá	IK	RF tiêu chuẩn	RF lai
F1-Score	Sét	0,44	0,47	0,83
	Đá ong	0,43	0,55	0,84
	Đá vôi	0,83	0,94	0,98
	Đất	0,36	0,32	0,79
Precision	Sét	0,64	0,50	0,83
	Đá ong	0,57	0,54	0,83
	Đá vôi	0,93	0,92	0,98
	Đất	0,41	0,42	0,80
Recall	Sét	0,34	0,45	0,83
	Đá ong	0,35	0,56	0,83
	Đá vôi	0,75	0,96	0,98
	Đất	0,32	0,26	0,78
Độ chính xác		0,61	0,80	0,93

Để đánh giá sâu hơn khả năng phân loại của ba mô hình, các đường cong đặc tính hoạt động (ROC) và diện tích dưới đường cong (AUC) của mô hình đã được thiết lập cho từng loại đất đá (Hình 3). Chỉ số AUC cung cấp một cái nhìn khách quan về năng lực nhận diện giữa các lớp đất đá của mô hình, độc lập với các ngưỡng phân loại cụ thể. Diện tích này càng lớn hay tiệm cận về giá trị 1, chứng tỏ mô hình có hiệu suất càng cao. Đường chéo (nét đứt) đại diện cho một mô hình dự đoán ngẫu nhiên. Một mô hình có giá trị thực tế phải nằm phía trên đường chéo này. Sự dịch chuyển của đường cong ROC của mô hình Hybrid RF về sát góc trên bên trái của biểu đồ so với hai mô hình còn lại là bằng chứng thực nghiệm mạnh mẽ nhất cho thấy đây là chiến lược tối ưu để xây dựng mô hình thạch học 3D trong điều kiện dữ liệu mỏ Tà Thiết thưa thớt và mất cân bằng. Trong khi đó, đường trung bình macro (Macro) đóng vai trò là thước đo then chốt để so sánh hiệu suất tổng thể. Kết quả cho thấy mô hình RF lai đạt chỉ số macro AUC cao nhất là 0,98, minh chứng cho khả năng phân loại vượt trội trên tất cả các loại

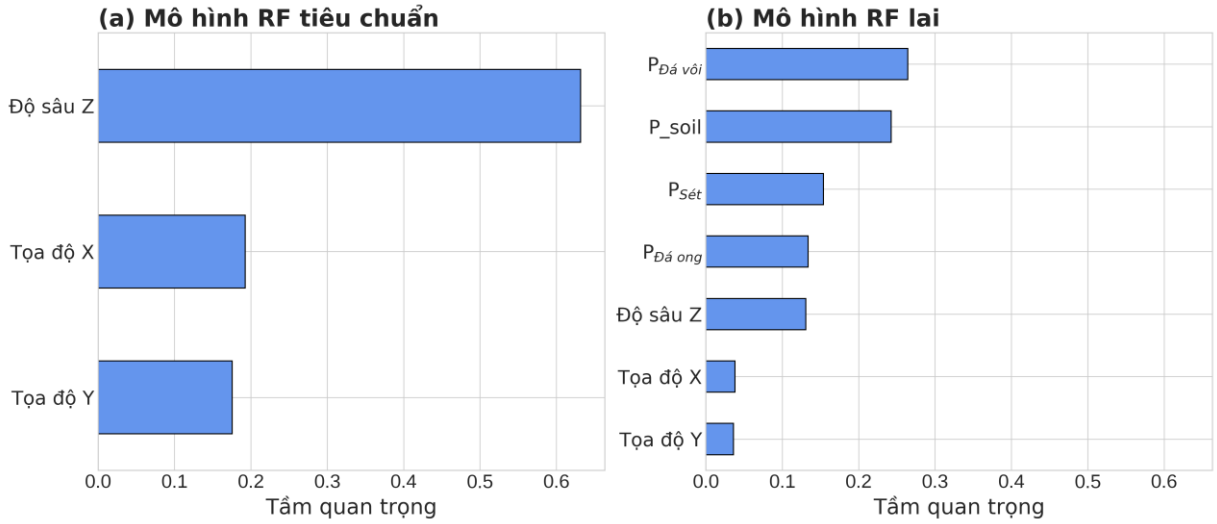


Hình 3. Đồ thị đánh giá khả năng phân loại đất đá của các mô hình dự báo (a) IK, (b) RF tiêu chuẩn, và (c) RF lai.

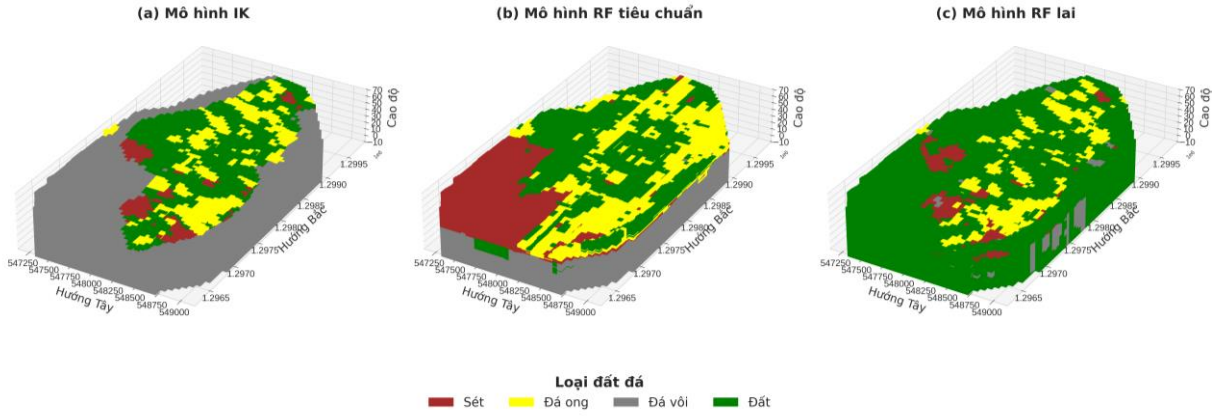
thạch học. Trong khi đó, mô hình RF tiêu chuẩn có chỉ số này thấp hơn (0,89) và mô hình IK đạt 0,93. Sự ưu việt của mô hình lai thể hiện rõ nét nhất ở các loại đá chiếm tỷ trọng nhỏ. Trong khi mô hình RF tiêu chuẩn gặp thách thức lớn với các lớp đá hiếm khi chỉ đạt AUC 0,86 cho sét; 0,90 cho đá ong và 0,88 cho đất. Điều này cho thấy nếu không có các đặc trưng địa thống kê bổ trợ, thuật toán RF tiêu chuẩn khó có thể phân biệt chính xác các ranh giới thạch học của các lớp kẹp thưa thớt. Mô hình RF lai đạt được các giá trị AUC gần như hoàn hảo cho cả nhóm thiểu số: sét (0,97), đá ong (0,98) và đất (0,98). Kết quả này khẳng định việc tích hợp xác suất không gian từ IK đã cung cấp tín hiệu quan trọng giúp RF nhận diện chính xác các nhóm đá dễ bị nhầm lẫn. Một điểm đáng chú ý là mô hình IK đạt chỉ số macro AUC khá cao (0,93), bao gồm cả các loại đá chiếm tỷ trọng nhỏ, dù độ chính xác tổng thể và F1-score của nó rất thấp (chỉ đạt 61,0%). Điều này cho thấy mặc dù IK có khả năng xếp hạng xác suất tương đối tốt giữa các lớp, nhưng khả năng đưa ra dự báo phân loại cuối cùng lại kém hiệu quả do hiệu ứng làm trơn dữ liệu. Mô hình RF lai đã tận dụng thành công lợi thế về xác suất này của IK để chuyển hóa thành độ chính xác dự báo thực tế. Đối với đá vôi, mô hình RF lai gần như đạt tới độ chính xác tuyệt đối với AUC là 0,99, cao hơn đáng kể so với RF tiêu chuẩn (0,94).

Để hiểu rõ cơ chế nội tại và cách thức các mô hình đưa ra quyết định phân loại, nghiên cứu đã tiến hành trích xuất và so sánh mức độ quan trọng của các biến đầu vào giữa các mô hình (Hình 4). Trong mô hình RF tiêu chuẩn, khi tập đặc trưng chỉ bao gồm các tọa độ không gian (x, y, z), kết quả phân tích cho thấy độ sâu (biến z) đóng vai trò áp đảo hoàn toàn so với tọa độ mặt bằng (x và y). Điều này phản ánh đúng thực tế địa chất cơ bản tại mỏ Tà Thiết, nơi các lớp đất đá có xu hướng phân bố theo chiều thẳng đứng (đất và đá ong ở trên, đá vôi ở dưới sâu). Tuy nhiên, việc phụ thuộc quá mức vào biến z chính là nguyên nhân khiến mô hình này dễ dự báo sai các thấu kính sét hay đá ong nằm xen kẹp bất thường, bởi nó thiếu các tín hiệu về tính liên tục theo phương ngang. Khi chuyển sang mô hình RF lai, bức tranh phân bổ tầm quan trọng của các đặc trưng có sự thay đổi mang tính bước ngoặt. Phân tích cho thấy thuật toán RF lai đã chủ động ưu tiên sử dụng các giá trị xác suất địa thống kê (Pđá vôi, Pđá ong, Psét, và Pđất) làm các yếu tố dự đoán cốt lõi, mạnh mẽ hơn nhiều so với việc chỉ dùng tọa độ không gian thô. Các biến xác suất từ IK vươn lên chiếm giữ những vị trí có trọng số cao nhất trong quá trình phân tách các nút của cây quyết định. Mặc dù biến độ sâu (z) vẫn giữ một vai trò nhất định để duy trì xu hướng địa tầng vĩ mô, nhưng nó không còn là yếu tố duy nhất chi phối toàn bộ kết quả. Điều này cho thấy, thuật toán RF đã tự động nhận diện được rằng các xác suất do Indicator Kriging cung cấp chứa đựng lượng thông tin giàu có hơn về bối cảnh địa chất cục bộ và tính liên tục không gian so với các tọa độ đơn thuần. Thêm vào đó, Bằng cách gán trọng số cao cho Psét và Pđất, thuật toán có thể phát hiện và khoanh vùng chính xác các loại đá thiểu số ngay cả khi chúng nằm ngoài quy luật phân bố độ sâu thông thường. Sự kết hợp giữa

Địa thống kê và Học máy trong nghiên cứu này không phải là việc gộp dữ liệu một cách cơ học. Kết quả phân tích tầm quan trọng là bằng chứng định lượng rõ ràng nhất cho thấy IK đã đóng vai trò như một bộ trích xuất đặc trưng, chuyển hóa các quy luật biến thiên không gian phức tạp (Variogram) thành ngôn ngữ cấu trúc mà thuật toán Học máy có thể khai thác tối đa.

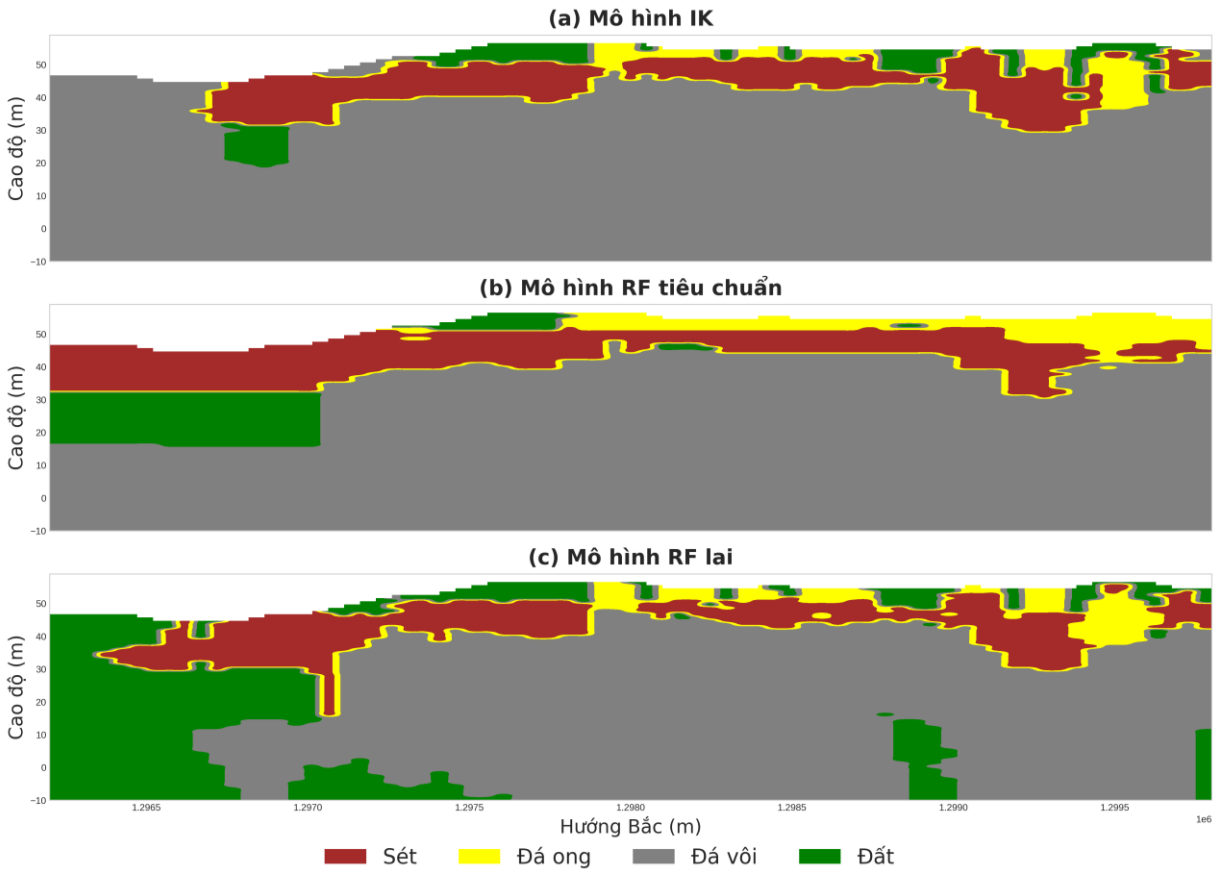


Hình 4. Tầm quan trọng của các đặc trưng sử dụng để xây dựng mô hình RF tiêu chuẩn và RF lai (*P* là đặc trưng xác suất được tạo bởi kỹ thuật *IK* cho từng loại đất đá).

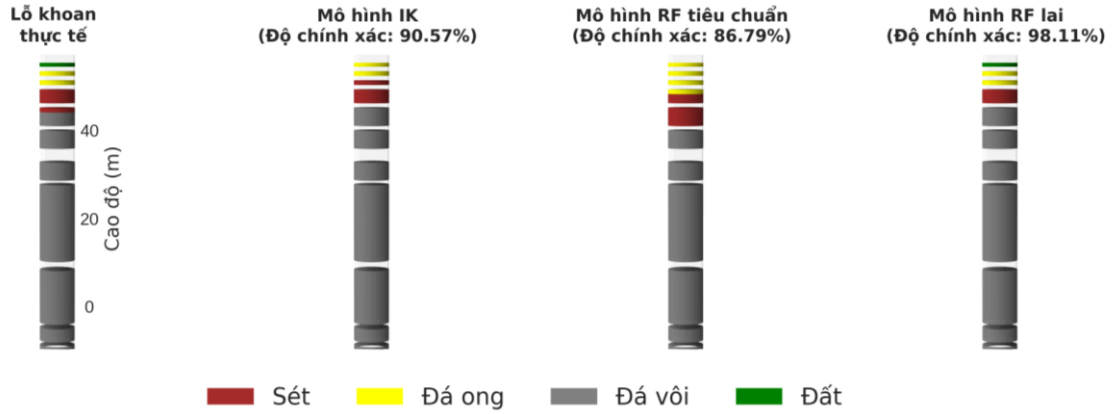


Hình 5. Kết quả mô hình hóa 3D đất đá tại khu vực nghiên cứu với các thuật toán khác nhau.

Hình 5 và 6 trình bày kết quả mô hình hóa 3D và 2D các loại đất đá của toàn bộ khu vực nghiên cứu được dự báo bởi ba phương pháp tiếp cận khác nhau. Mô hình IK truyền thống thể hiện rõ đặc trưng của các phương pháp nội suy khoảng cách: hiệu ứng làm trơn (smoothing effect). Các khối thạch học được tạo ra có ranh giới quá mềm mại và lan tỏa dạng hình cầu. Mặc dù mô hình duy trì được xu hướng vĩ mô, nhưng nó thất bại trong việc tái tạo các ranh giới địa chất sắc nét tự nhiên. Đối với các lớp đá chiếm đá số như đá vôi, IK có xu hướng phóng đại thể tích của chúng, dẫn đến sự sai lệch nghiêm trọng nếu sử dụng cho mục đích tính toán trữ lượng. Trái ngược với sự làm trơn của IK, mô hình RF tiêu chuẩn phân chia các lớp đá chủ yếu theo trục z (cao độ) nhưng lại thiếu khả năng liên kết theo phương ngang. Mô hình lai RF thiết lập được các khối thạch học vững chắc với ranh giới sắc nét (kế thừa từ đặc tính phân loại của RF), đồng thời duy trì được hình khối mạch lạc, các lớp đất, sét nằm xen kẹp giữa các lớp đá vôi, phù hợp với với quy luật trầm tích và phong hóa tại mỏ Tà Thiết.



Hình 6. Mặt cắt 2D (X = 548000.0 m) đất đá tại khu vực nghiên cứu với các thuật toán khác nhau.



Hình 7. So sánh kết quả dự đoán của các mô hình với lỗ khoan thực tế (mã hiệu BSLK19-2)

Để minh chứng rõ nét hơn cho độ tin cậy và khả năng ứng dụng thực tiễn của phương pháp đề xuất, nghiên cứu đã tiến hành trích xuất dữ liệu dự báo dọc theo trục Z và đối chiếu trực tiếp với nhật ký lõi khoan thực tế của lỗ khoan mang mã hiệu BSLK19-2 (Hình 7). Mặc dù RF tiêu chuẩn dự báo tương đối tốt các đoạn đá vôi dày ở phần đáy lỗ khoan, nhưng nó lại thể hiện hiệu suất thấp nhất trong ba phương pháp tại vị trí này. Nguyên nhân cốt lõi là do sự thiếu hụt thông tin không gian. Thuật toán gặp khó khăn lớn tại các đới chuyển tiếp và các lớp kẹp xen kẽ ở phần trên của lỗ khoan, dẫn đến hiện tượng dự báo phân mảnh và nhầm lẫn cục bộ giữa các loại đá thiểu số. Phương pháp IK truyền thống cho kết quả khả quan hơn so với RF tiêu chuẩn tại vị trí lỗ khoan này. Nhờ có mô hình Variogram, IK đã nắm bắt được xu hướng

biến thiên tổng thể của địa tầng. Tuy nhiên, mức độ sai số ~9,4% chủ yếu xuất phát từ hiệu ứng làm trơn (smoothing effect). Tại các vị trí có sự xuất hiện đột ngột của các thấu kính sét mỏng nằm kẹp trong khối đá vôi, IK có xu hướng bỏ qua các lớp mỏng này hoặc làm mờ ranh giới chuyển tiếp, khiến độ chính xác cục bộ bị suy giảm. Mô hình Hybrid RF đạt độ trùng khớp gần như hoàn hảo (98,11%) so với cột địa tầng thực tế của lỗ khoan BSLK19-2. Sự kết hợp giữa tọa độ 3D và các đặc trưng xác suất từ IK đã cung cấp một bối cảnh không gian hoàn chỉnh.

4. KẾT LUẬN VÀ KIẾN NGHỊ

Nghiên cứu này đã giải quyết thành công những thách thức trong việc lập mô hình thạch học 3D tại mỏ đá vôi Tà Thiết, nơi có mạng lưới thăm dò thưa thớt và dữ liệu phân bố mất cân bằng. Việc đề xuất và áp dụng chiến lược mô hình hóa lai (RF lai), tích hợp các xác suất không gian từ kỹ thuật Indicator Kriging (IK) vào thuật toán Học máy Random Forest (RF), đã mang lại những bước tiến vượt trội so với việc sử dụng các phương pháp đơn lẻ.

Mô hình RF lai đạt độ chính xác tổng thể lên tới 93,0%, cao hơn đáng kể so với mô hình RF tiêu chuẩn (80,0%) và IK truyền thống (61,0%). Chỉ số Macro AUC của mô hình lai đạt mức 0,98, minh chứng cho khả năng phân loại tốt và đồng đều trên toàn bộ các loại thạch học. Sự kết hợp các đặc trưng địa thống kê đã giúp thuật toán Học máy khắc phục nhược điểm bỏ sót dữ liệu thiểu số. Chỉ số F1-score của các lớp kẹp như sét, đất và đá ong được cải thiện tới 145%, với giá trị AUC đều đạt từ 0,97 đến 0,98. Phân tích tầm quan trọng của đặc trưng chỉ ra rằng các xác suất phân bố từ IK đóng vai trò dẫn dắt mạnh mẽ hơn hẳn các tọa độ không gian thô (x, y, z) trong quá trình ra quyết định của thuật toán Học máy. IK đã đóng vai trò như một bộ trích xuất đặc trưng xuất sắc để chuyển hóa quy luật không gian thành ngôn ngữ số liệu. Mô hình RF lai tạo ra các khối 3D với ranh giới sắc nét, thể hiện rõ tính liên tục theo phương ngang của các thấu kính sét và đá ong xen kẹp, hoàn toàn phù hợp với quy luật trầm tích tự nhiên tại mỏ. Khi đối chiếu thực tế tại lỗ khoan BSLK19-2, mô hình lai đạt độ trùng khớp lên tới 98,11%, khắc phục hoàn toàn hiệu ứng làm trơn của IK và sự phân mảnh rời rạc của RF tiêu chuẩn.

Tóm lại, mô hình thạch học 3D được xây dựng từ phương pháp RF lai là một giải pháp cung cấp một nền tảng cơ sở dữ liệu không gian có độ tin cậy cao. Đây là công cụ đắc lực hỗ trợ các kỹ sư trong việc lập kế hoạch khai thác chọn lọc, quản trị rủi ro địa chất và tối ưu hóa quy trình phối trộn nguyên liệu, từ đó đảm bảo nguồn cung cấp thô ổn định và chất lượng cao cho dây chuyền sản xuất của nhà máy xi măng Bình Phước.

TÀI LIỆU THAM KHẢO

1. P. Goovaerts (1999), *Geostatistics for Natural Resources Evaluation by Pierre Goovaerts*, Oxford University Press, City.
2. Z. Hang, T. Xue, J. Chen, Y. Shi, Z. Yin, Z. Cui and G. Zhou (2025), *A 3D Geological Modeling Method Using the Transformer Model: A Solution for Sparse Borehole Data*, Minerals, 15 3, 301-301.
3. T. B. Afeni, V. O. Akeju and A. E. Aladejare (2021), *A comparative study of geometric and geostatistical methods for qualitative reserve estimation of limestone deposit*, Geoscience Frontiers, 12 1, 243-253.
4. G. Ji, Q. Wang, X. Zhou, Z. Cai, J. Zhu and Y. Lu (2023), *An automated method to build 3D multi-scale geological models for engineering sedimentary layers with stratum lenses*, Engineering Geology, 317 107077-107077.
5. S. A. Hosseini, O. Asghari, X. Emery and M. J. B. D. G. T. E. A. Maleki (2017), *Forecasting the grade-tonnage curves and their uncertainty at the Mehdiabad deposit-Yazd, central Iran*, Bollettino

Di Geofisica Teorica Ed Applicata.

6. T. Y. Yünsel (2018), *Simulation of cement raw material deposits using plurigaussian technique*, 10 1, 889-901.
7. T. Vu, C. Drebenstedt and T. Bao (2020), *Assessing geological uncertainty of a cement raw material deposit, southern Vietnam, based on hierarchical simulation*, International Journal of Mining Science and Technology,
8. M. M. Tahernejad, R. KhaloKakaie and M. Ataei (2018), *Analyzing the effect of ore grade uncertainty in open pit mine planning; A case study of Rezvan iron mine, Iran*, International Journal of Mining and Geo-Engineering, 52 1, 111-120.
9. R. Goodfellow, F. A. Consuegra, R. G. Dimitrakopoulos and T. Lloyd (2012), *Quantifying multi-element and volumetric uncertainty, Coleman McCreedy deposit, Ontario, Canada*, Computer Geoscience, 42 71-78.
10. Y. Dagasan, O. Erten, P. Renard, J. Straubhaar and E. Topal (2019), *Multiple-point statistical simulation of the ore boundaries for a lateritic bauxite deposit*, Stochastic Environmental Research and Risk Assessment, 33 3, 865-878.
11. P. Pereira, M. N. Rabelo, C. C. Ribeiro and H. S. Diniz-Pinto (2017), *Geological modeling by an indicator kriging approach applied to a limestone deposit in Indiara city - Goiás*, REM - International Engineering Journal, 70 3, 331-37.
12. N. K. Dumakor-Dupey and S. Arya (2021), *Machine learning—a review of applications in mineral resource estimation, Energies*, 14 14, 4079-4079.

Thông tin của tác giả:

Vũ Đình Trọng
BM Khai thác khoáng sản, Khoa Mỏ & Công Trình, ĐH Công Nghiệp Quảng Ninh
Điện thoại: 0869437970 - Email: trongvu@qui.edu.vn

Nguyễn Văn Đức
BM Khai thác khoáng sản, Khoa Mỏ & Công Trình, ĐH Công Nghiệp Quảng Ninh
Điện thoại: 0359990865 - Email: ducnguyen@qui.edu.vn

DEVELOPMENT OF A LITHOLOGICAL MODEL FOR TA THIET LIMESTONE DEPOSIT USING GEOSTATISTICS TECHNIQUE AND MACHINE LEARNING

Information about authors:

Vu Dinh Trong, Department of Mining, Faculty of Mining & Construction, Quang Ninh University of Industry
email: trongvu@qui.edu.vn

Nguyen Van Duc, Department of Mining, Faculty of Mining & Construction, Quang Ninh University of Industry

ABSTRACT:

The establishment of a 3D lithological model is a challenging task in Ta Thiet quarry limestone due to the sparseness of the exploration drilling pattern along with the imbalance of drill log data. This study proposes a hybrid Random Forest (RF) which is the combination of standard RF and geostatistical Indicator Kriging (IK). Using the spatial probability generated by IK, the RF model

captures the geological context. The result shows that the hybrid RF achieved an overall accuracy of 93.0%, which outperformed the standard RF (80,0%) and IK (61,0%) and improved its ability in detecting minor rock types. The hybrid RF model also generates a high solution of lithological boundaries. This is potentially a reliable tool for selective mining and blending of cement raw mmaterials.

Keywords: *3D rocktype model, geostatistics, machine learning, Indicator Kriging, Random Forest, imbalanced data.*

Ngày nhận bài: 07/04/2026;
Ngày nhận bài sửa: 16/04/2026;
Ngày chấp nhận đăng: 17/04/2026.