



Tạp chí điện tử
Khoa học và Công nghệ Giao thông
Trang website: <https://jstt.vn/index.php/vn>



Article info

Type of article:

Original research paper

DOI:

<https://doi.org/10.58845/jstt.utt.2025.vn.5.4.95-109>

***Corresponding author:**

Email address:

lehuyentrang0500@gmail.com

Received: 07/10/2025

Received in Revised Form:

10/12/2025

Accepted: 12/12/2025

Using Gradient Boosting Regression (GBR) model in predicting the optimum moisture content of soil used for roadbed compaction

Dam Duc Nguyen¹, Le Huyen Trang¹, Luu Thuy Duong², Vu Quang Dung¹,
Pham Quang Dung¹, Bui Thi Thanh Nga²

¹University of Transport Technology, 54 Trieu Khuc, Thanh Liet, Ha Noi, Viet Nam. Email: damnd@utt.edu.vn, lehuyentrang0500@gmail.com, dungvq@utt.edu.vn, dungpq@utt.edu.vn,

² Hanoi University of Natural Resources and Environment, Ha Noi, Viet Nam. Email: ltduong@hunre.edu.vn, bttnga.ph@hunre.edu.vn

Abstract: Optimum Water Content of soil (OWC) is a critical engineering parameter that directly affects compaction efficiency and the long-term stability of subgrade soil in road construction. Determination of OWC using conventional laboratory testing methods is time-consuming and costly, which poses limitations in practical construction. This study applies the Gradient Boosting Regressor (GBR), an advanced machine learning algorithm, to predict OWC based on the physical and mechanical properties of soil samples. Quantitative performance metrics, including the correlation coefficient (R), mean absolute error (MAE), and root mean square error (RMSE), are employed to evaluate model performance. The results show that the model achieves R of 0.916 for the training dataset and 0.776 for the testing dataset, along with low MAE and RMSE values, indicating high predictive accuracy and strong generalization capability. Model interpretability analysis using SHAP techniques and Partial Dependence Plots (PDPs) highlights the significant influence of variables such as plastic limit and gravel content on OWC. This research provides an effective supporting tool for subgrade design and construction, while also expanding the application of machine learning approaches in the field of geotechnical engineering.

Keywords: Optimum water content, soil compaction, Gradient Boosting Regressor, machine learning, SHAP analysis, Partial Dependence Plot (PDP), geotechnical engineering.



Nghiên cứu áp dụng mô hình GBR trong dự báo giá trị độ ẩm tốt nhất của đất sử dụng trong đầm nén nền đường

Thông tin bài viết

Dạng bài viết:

Bài báo nghiên cứu

DOI:

<https://doi.org/10.58845/jstt.utt.2025.vn.5.4.95-109>

*Tác giả liên hệ:

Địa chỉ Email:

lehuyentrang0500@gmail.com

Ngày nộp bài: 07/10/2025

Ngày nộp bài sửa: 10/12/2025

Ngày chấp nhận: 12/12/2025

Nguyễn Đức Đảm¹, Lê Huyền Trang¹, Lưu Thùy Dương², Vũ Quang Dũng¹, Phạm Quang Dũng¹, Bùi Thị Thanh Nga²

¹Trường Đại học Công nghệ Giao thông vận tải, 54 Triều Khúc, Thanh Liệt, Hà Nội, Việt Nam. Email: damnd@utt.edu.vn, lehuyentrang0500@gmail.com, dungvq@utt.edu.vn, dungpq@utt.edu.vn,

²Trường Đại học Tài nguyên và Môi trường Hà Nội, Hà Nội, Việt Nam. Email: ltduong@hunre.edu.vn, bttnga.ph@hunre.edu.vn

Tóm tắt: Độ ẩm tốt nhất của đất (Optimal Moisture Content - OWC) là một thông số kỹ thuật quan trọng ảnh hưởng trực tiếp đến hiệu quả đầm nén và độ bền vững của nền đất trong xây dựng đường bộ. Việc xác định OWC bằng phương pháp thí nghiệm truyền thống đòi hỏi nhiều thời gian và chi phí, gây hạn chế trong thực tế thi công. Nghiên cứu này ứng dụng mô hình Gradient Boosting Regressor (GBR) – một thuật toán học máy tiên tiến – để dự báo giá trị OWC dựa trên các đặc tính cơ lý của mẫu đất. Các chỉ số đánh giá định lượng được dùng để đánh giá hiệu suất của mô hình bao gồm: R, MAE và RMSE. Kết quả mô hình đạt hệ số xác định lần lượt là 0.916 trên tập huấn luyện và 0.776 trên tập kiểm tra, cùng với các chỉ số lỗi MAE và RMSE thấp, cho thấy độ chính xác và khả năng tổng quát hóa tốt. Phân tích giải thích mô hình bằng kỹ thuật SHAP và biểu đồ Partial Dependence Plot (PDP) làm rõ vai trò quan trọng của các biến như giới hạn dẻo và hàm lượng sỏi trong ảnh hưởng đến OWC. Nghiên cứu cung cấp công cụ hỗ trợ hiệu quả cho thiết kế và thi công nền đất, đồng thời mở rộng ứng dụng các phương pháp học máy trong lĩnh vực địa kỹ thuật.

Từ khóa: Độ ẩm tốt nhất, Đầm nén đất nền, Gradient Boosting Regressor, Học máy, Giải thích mô hình SHAP, Biểu đồ Partial Dependence Plot, Kỹ thuật địa chất.

1. Giới thiệu

Độ ẩm tốt nhất của đất (Optimal Water Content - OWC) là một trong những thông số kỹ thuật quan trọng trong lĩnh vực thiết kế và thi công nền đường, ảnh hưởng trực tiếp đến đặc tính cơ học cũng như độ bền vững của đất nền sau khi được đầm nén [1]. Việc xác định chính xác giá trị OWC không chỉ giúp tối ưu hóa quá trình thi công mà còn góp phần giảm thiểu hiện tượng lún sụt

không đều, từ đó nâng cao tuổi thọ và chất lượng công trình giao thông. Truyền thống, giá trị OWC thường được xác định thông qua các thí nghiệm đầm nén chuẩn như Proctor tiêu chuẩn hoặc Proctor cải tiến [2], đòi hỏi thời gian thực hiện khá dài, tốn kém chi phí, đồng thời dễ bị ảnh hưởng bởi các yếu tố thao tác và điều kiện thí nghiệm, đặc biệt là với những mẫu đất có tính chất phức tạp và đa dạng về thành phần địa chất ở các khu vực khác

nhau.

Trong những năm gần đây, sự phát triển vượt bậc của công nghệ học máy (Machine Learning) đã mở ra nhiều hướng đi mới trong việc dự báo các đặc tính cơ lý của đất nền như độ ẩm, độ chặt, sức chịu tải, góp phần nâng cao độ chính xác và hiệu quả của các phương pháp truyền thống [3]. Các phương pháp học máy, bao gồm mạng nơ-ron nhân tạo, máy véc-tơ hỗ trợ, cây quyết định, cùng các thuật toán tổ hợp như rừng ngẫu nhiên và tăng cường dần, đã và đang được nghiên cứu, ứng dụng một cách rộng rãi trong lĩnh vực địa kỹ thuật nhằm nâng cao độ chính xác trong phân tích và dự báo các đặc trưng của đất nền và công trình [3]. Gradient Boosting Regressor (GBR) nổi bật như một phương pháp học máy mạnh mẽ, có khả năng xử lý dữ liệu phi tuyến, đa chiều phức tạp và cung cấp kết quả dự báo chính xác trong nhiều bài toán kỹ thuật đa dạng [4].

Mặc dù đã có một số nghiên cứu áp dụng các mô hình học máy để dự báo giá trị OWC, phần lớn các công trình chủ yếu tập trung vào các mô hình truyền thống hoặc một số mô hình phổ biến, mà chưa khai thác hết tiềm năng của các thuật toán tiên tiến như GBR. Thêm vào đó, các nghiên cứu trước đây còn thiếu sự so sánh hệ thống, đánh giá toàn diện hiệu quả và tính ổn định của các mô hình trên các tập dữ liệu đa dạng, đặc biệt là trong điều kiện thực tiễn của các vùng địa chất Việt Nam. Do đó, nghiên cứu này hướng đến phát triển và đánh giá mô hình dự báo OWC dựa trên thuật toán GBR. Điểm mới của nghiên cứu này là việc ứng dụng GBR – một thuật toán học máy tiên tiến với khả năng xử lý hiệu quả các tập dữ liệu lớn, đa chiều và phi tuyến, cùng với tốc độ huấn luyện nhanh và khả năng tổng quát hóa tốt – vào dự báo OWC cho đất nền sử dụng trong đầm nén nền đường. Mô hình được xây dựng dựa trên việc phân tích và khai thác đồng thời các đặc tính vật lý, hóa học đa dạng của mẫu đất thu thập từ nhiều khu vực khác nhau, nhằm nâng cao độ chính xác và độ tin cậy của kết quả dự báo. Kết quả nghiên cứu không chỉ góp phần rút ngắn thời gian và chi phí thí nghiệm, mà còn cung cấp công cụ hỗ trợ đắc lực cho các

kỹ sư trong việc thiết kế và thi công nền đường một cách hiệu quả, tiết kiệm và bền vững, đồng thời mở rộng hướng nghiên cứu áp dụng học máy tiên tiến trong lĩnh vực địa kỹ thuật tại Việt Nam.

2. Dữ liệu và phương pháp nghiên cứu

2.1. Dữ liệu nghiên cứu

Trong nghiên cứu này, chúng tôi sử dụng dữ liệu thu thập được từ dự án xây dựng cao tốc Vân Đồn - Móng Cái. Dữ liệu nghiên cứu được thu thập ngẫu nhiên, bao gồm 214 bộ kết quả thí nghiệm, rải đều tại các vị trí khác nhau trên toàn tuyến cao tốc. Các mẫu đất được lấy trực tiếp từ công trường và mô vật liệu, sau đó đem về phòng thí nghiệm để xác định các chỉ tiêu cơ lý như thành phần hạt, độ ẩm tốt nhất (OWC), giới hạn chảy (LL), giới hạn dẻo (PL), chỉ số dẻo (PI), tỷ số CBR. Các mẫu thí nghiệm được bảo quản, vận chuyển theo đúng quy trình.

Quy trình thí nghiệm mẫu được thực hiện theo tiêu chuẩn hiện hành của Việt Nam, có tham khảo thêm tiêu chuẩn ASTM, AASHTO của Mỹ. Áp dụng tiêu chuẩn TCVN4198, AASHTO-T-27-14, and AASHTO-M-145-91 [5-7] để xác định thành phần hạt và phân loại mẫu đất theo các cỡ hạt khác nhau, nhóm đất khác nhau (đá dăm sỏi (G), cát hạt thô (CS), cát hạt mịn (FS), sét bụi (SC), các nhóm nhất như A-2-6, A-4, A-6, A-7-6). Áp dụng tiêu chuẩn AASHTO-T-267-86 [8] để xác định hàm lượng hữu cơ (O). Áp dụng tiêu chuẩn TCVN4197 [9] để xác định giới hạn chảy (LL), giới hạn dẻo (PL), chỉ số dẻo (PI). Áp dụng tiêu chuẩn TCVN12790 [10] để xác định độ ẩm tốt nhất (OWC), khối lượng thể tích khô lớn nhất (MDD). Áp dụng tiêu chuẩn TCVN12792 [11] để xác định tỷ số CBR.

Độ ẩm tốt nhất (OWC) được xác định qua các thí nghiệm đầm nén như thí nghiệm Proctor. Các nghiên cứu của các tác giả trước đây [12-18] đã chỉ ra các tham số đầu vào chính ảnh hưởng đến giá trị OWC bao gồm thành phần hạt, hàm lượng hữu cơ, giới hạn chảy (LL), giới hạn dẻo (PL), chỉ số dẻo (PI), trọng lượng riêng, phương pháp đầm nén. Thứ nhất: Thành phần hạt hay sự phân bố kích thước hạt hay, kích thước và sự phân

bổ của các hạt đất (cát, silt, sét) ảnh hưởng đến cách chúng sắp xếp trong quá trình đầm nén. OMC cũng thay đổi tùy thuộc vào loại đất, với đất sét thường yêu cầu lượng nước cao hơn để đạt độ chặt tối ưu. Thứ hai: Đặc tính dẻo, với đất dính (như đất sét), các giới hạn Atterberg (giới hạn chảy và giới hạn dẻo) quyết định lượng nước cần thiết để đạt OWC. Đất có độ dẻo cao thường có OWC lớn hơn, do cần nhiều nước để "bôi trơn" các hạt trong quá trình đầm nén. Thứ ba: Trọng lượng riêng của các hạt đất, trọng lượng riêng ảnh hưởng đến OWC, vì nó liên quan đến khối lượng chất rắn

trong một đơn vị thể tích. Thứ tư: Phương pháp đầm nén, năng lượng đầm nén (ví dụ: đầm nén theo phương pháp Proctor tiêu chuẩn hay Proctor cải tiến) ảnh hưởng trực tiếp đến OWC.

Việc lựa chọn các tham số đầu vào như đá dăm sỏi (G), cát hạt thô (CS), cát hạt mịn (FS), hàm lượng hữu cơ (O), sét bụi (SC), giới hạn chảy (LL), giới hạn dẻo (PL), chỉ số dẻo (PI) để dự đoán OWC của đất dựa trên việc xem xét tài liệu về các công trình đã công bố có liên quan [3, 19-23]. Bảng 1 cho thấy phân tích ban đầu về các biến được thu thập và sử dụng để lập mô hình.

Bảng 1. Phân tích ban đầu của dữ liệu được sử dụng

Tham số	Giá trị trung bình	Độ lệch chuẩn	Giá trị nhỏ nhất	25%	50%	75%	Giá trị lớn nhất
G	22.05706	13.29548	0	9.075	24.75	31.7	51.4
CS	24.10103	7.017101	3	20.7	23.7	27.775	46.3
FS	9.034766	6.467583	2.5	4.6	7.25	11	41.5
SC	44.8071	10.44727	17.87	37.75	44.55	49.2	88.7
O	1.508879	0.37256	0.12	1.2525	1.51	1.77	2.94
LL	39.51453	6.173263	2.08	36.6375	39.99	43.5075	48.45
PL	20.31841	3.067936	1.17	19.2925	20.835	21.8875	28.49
PI	19.19799	4.077597	0.91	16.83	18.435	22.32	27.48
OWC	14.00953	2.618564	9.3	12.19	14.275	15.4	21.5

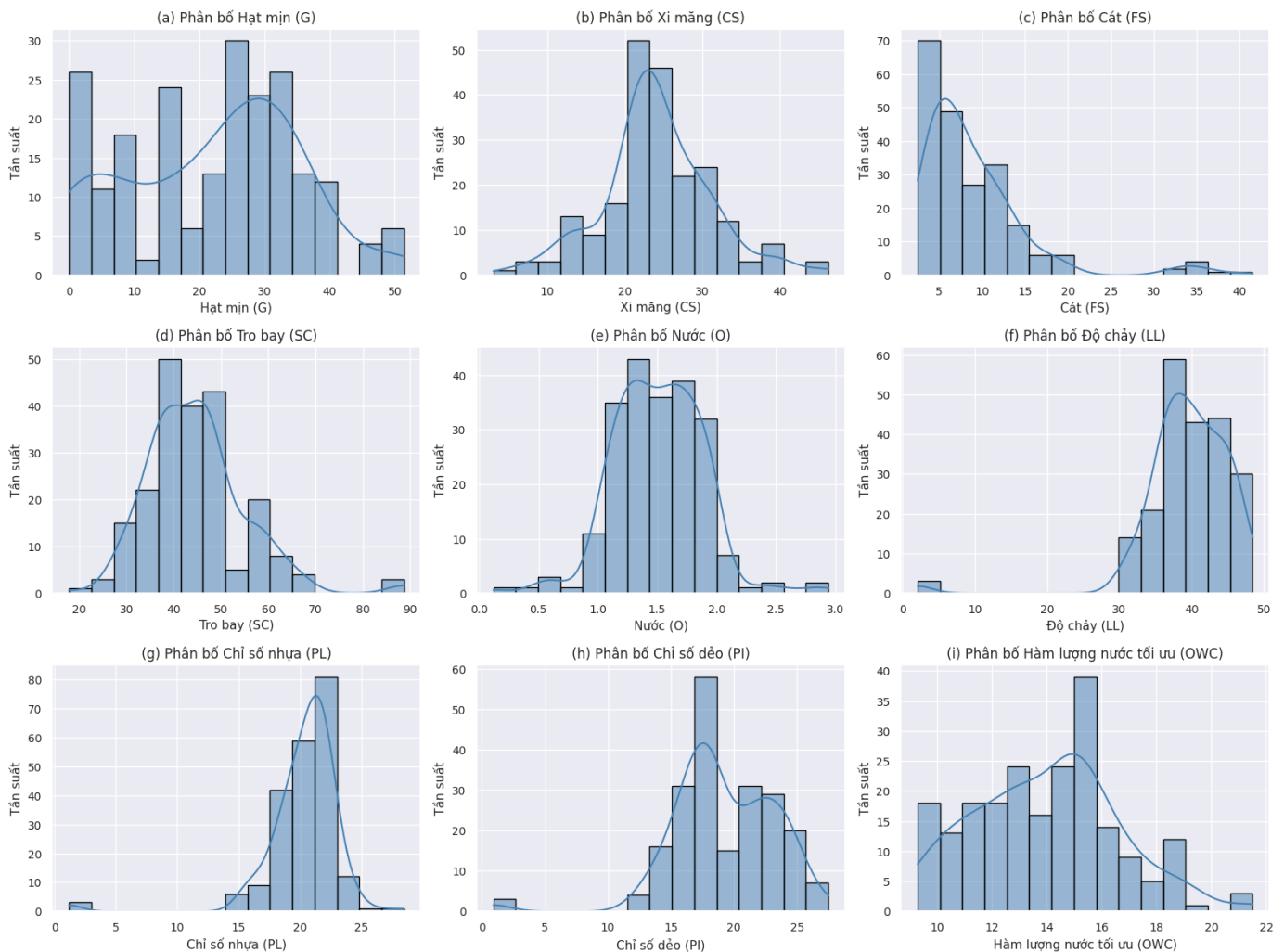
Bảng 1 trình bày phân tích thống kê mô tả ban đầu của bộ dữ liệu được sử dụng trong nghiên cứu, bao gồm các thông số trung bình (mean), độ lệch chuẩn (std), giá trị nhỏ nhất (min), các phân vị thứ 25%, 50%, 75% và giá trị lớn nhất (max) của từng biến. Hàm lượng sỏi (G) có giá trị trung bình là 22.06%, dao động từ 0% đến 51.4%, cho thấy sự đa dạng về thành phần hạt lớn trong các mẫu đất khảo sát. Hàm lượng cát thô (CS) trung bình đạt 24.10%, với độ lệch chuẩn là 7.02%, phản ánh sự phân bố khá đồng đều của thành phần này trong bộ dữ liệu. Hàm lượng cát mịn (FS) có giá trị trung bình thấp hơn, khoảng 9.03%, nhưng cũng có sự biến động đáng kể với giá trị nhỏ nhất là 2.5% và giá trị lớn nhất là 41.5%. Hàm lượng hữu cơ (O) chiếm tỷ lệ thấp trong các mẫu, với giá trị trung bình chỉ 1.51%, dao động từ 0.12% đến 2.94%, phù hợp với tiêu chuẩn loại bỏ đất hữu cơ

trong các công trình xây dựng nền đường. Hàm lượng bụi sét (SC) có giá trị trung bình cao nhất, đạt 44.81% với độ lệch chuẩn là 10.44%, cho thấy nhiều mẫu đất có hàm lượng hạt mịn vượt trội. Các chỉ tiêu giới hạn Atterberg như giới hạn chảy (LL) có giá trị trung bình 39.51% và giới hạn dẻo (PL) trung bình là 20.32%, với độ lệch chuẩn lần lượt là 6.17% và 3.07%, phản ánh sự đa dạng về tính dẻo của các mẫu đất nghiên cứu. Chỉ số dẻo (PI) có giá trị trung bình là 19.20%, dao động từ 0.91% đến 27.48%, cho thấy phạm vi trạng thái dẻo khá rộng trong bộ dữ liệu. Biến mục tiêu là OWC có giá trị trung bình là 14.01%, với độ lệch chuẩn là 2.62%, dao động từ 9.3% đến 21.5%. Các giá trị phân vị cho thấy phần lớn các mẫu đất có độ ẩm tốt nhất tập trung quanh giá trị trung bình, chứng tỏ bộ dữ liệu có sự phân bố hợp lý và phù hợp cho việc xây dựng mô hình dự báo. Phân tích thống kê này

không chỉ giúp nhận diện đặc điểm phân bố và sự biến thiên của từng biến mà còn hỗ trợ đánh giá chất lượng và tính đại diện của bộ dữ liệu trước khi

tiến hành các bước xử lý dữ liệu và xây dựng mô hình học máy nhằm dự báo độ ẩm tốt nhất cho đất nền.

Biểu đồ phân bố các biến đầu vào và đầu ra



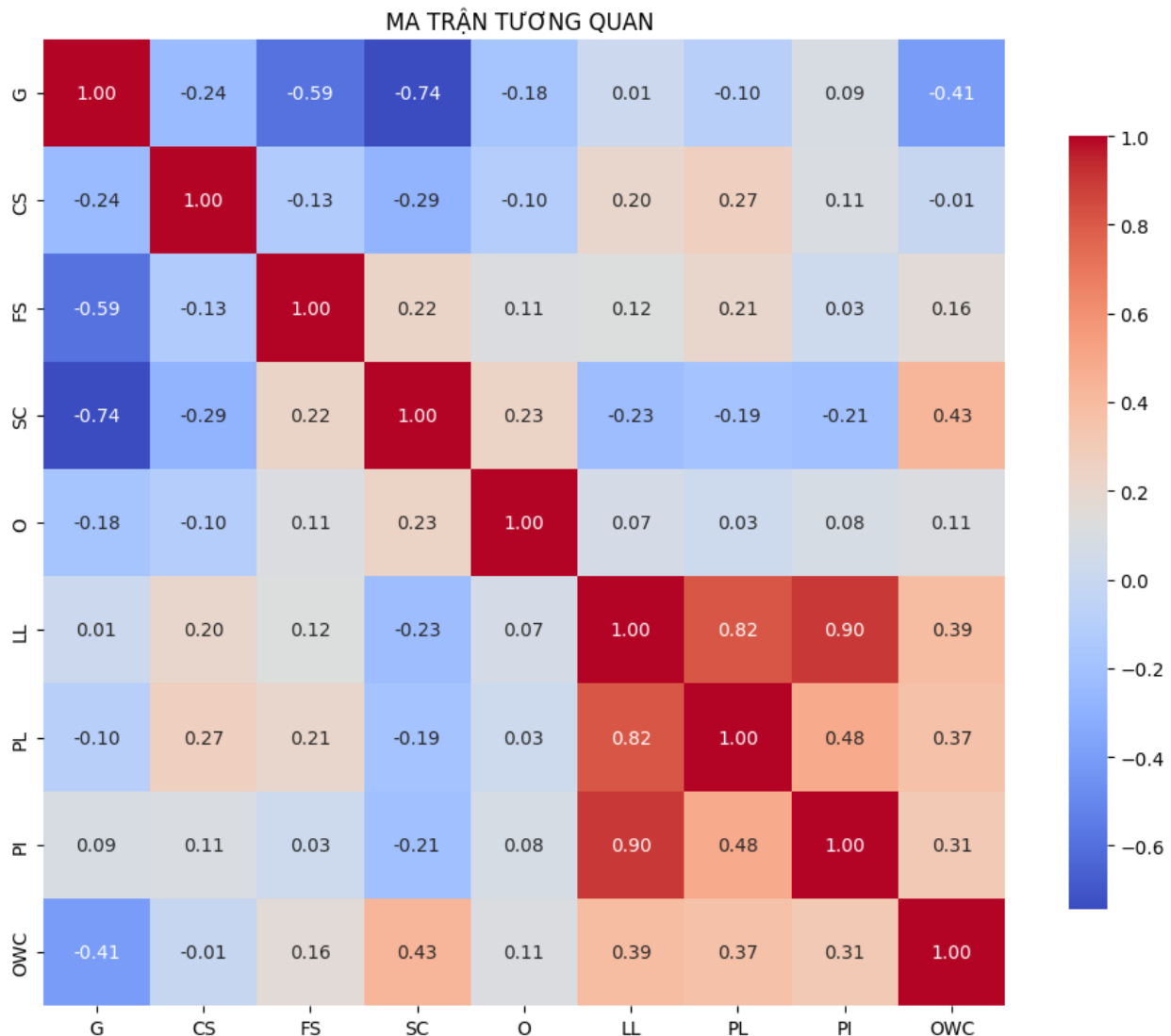
Hình 1. Phân phối dữ liệu của các biến được sử dụng trong mô hình

Hình 1 thể hiện phân phối dữ liệu của các biến đầu vào được sử dụng trong mô hình dự báo. Qua các biểu đồ, có thể thấy sự đa dạng và biến thiên rõ nét trong từng đặc tính vật lý và hóa học của mẫu đất nền. Biến hàm lượng sỏi (Gravel content - G) có phân bố khá rộng, trải dài từ 0% đến hơn 50%, với phần lớn các mẫu tập trung ở mức trung bình khoảng 10–40%, thể hiện sự khác biệt lớn về thành phần hạt thô trong các mẫu khảo sát. Điều này phản ánh tính chất không đồng nhất của đất nền và yêu cầu mô hình phải xử lý tốt các biến động lớn trong dữ liệu. Hàm lượng cát thô (Coarse sand content - CS) thể hiện phân phối lệch phải, với đa số mẫu nằm trong khoảng 15–30%, số

lượng mẫu có hàm lượng cát thô vượt quá 40% khá hạn chế. Đây là đặc trưng phổ biến của các loại đất nền đắp được kiểm soát nghiêm ngặt về thành phần hạt thô nhằm đảm bảo độ ổn định. Hàm lượng cát mịn (Fine sand content - FS) phần lớn tập trung ở các giá trị thấp dưới 15%, điều này phù hợp với mục tiêu hạn chế tỷ lệ hạt nhỏ nhằm tăng khả năng chịu lực và giảm hiện tượng co ngót đất. Phân bố hàm lượng bụi sét (Silt clay content - SC) khá rộng, với phần lớn mẫu nằm trong khoảng 35–50% và một số mẫu có giá trị rất cao lên đến gần 90%. Hàm lượng hữu cơ (Organic content - O) duy trì ở mức thấp, chủ yếu dưới 2.5%, đảm bảo tính ổn định và chất lượng của nền đất trong thi công

công trình. Đặc tính dẻo của đất được thể hiện qua các biến giới hạn lỏng (Liquid limit - LL), giới hạn dẻo (Plastic limit - PL) và chỉ số độ dẻo (Plasticity Index - PI), với giá trị phân bố đồng đều và đa dạng, phản ánh sự khác biệt về tính chất vật lý giữa các mẫu đất và ảnh hưởng trực tiếp đến OWC. Cuối cùng, biến mục tiêu OWC có phân bố tập trung quanh giá trị trung bình từ khoảng 12% đến 16%,

với một số mẫu có giá trị cao hơn 20%, phản ánh phạm vi độ ẩm tối ưu đa dạng cần thiết để đạt được độ đầm chặt tối ưu cho các loại đất nền khác nhau. Tổng thể, sự phân bố đa dạng và đặc trưng của các biến này cho thấy bộ dữ liệu nghiên cứu có tính đại diện cao và phù hợp để xây dựng các mô hình dự báo có độ chính xác và khả năng tổng quát hóa tốt.



Hình 2. Ma trận tương quan của các biến được sử dụng trong mô hình

Hình 2 trình bày ma trận tương quan giữa các biến đầu vào và biến mục tiêu được sử dụng trong mô hình dự báo. Các giá trị tương quan được biểu diễn bằng bản đồ nhiệt với thang màu từ đỏ đậm đến xanh dương, phản ánh mức độ tương quan dương và âm giữa các biến. Qua ma trận này, có thể quan sát thấy mối quan hệ rõ ràng giữa một số biến đầu vào với biến mục tiêu OWC. Cụ

thể, các biến đặc trưng về tính dẻo của đất như giới hạn lỏng (LL), giới hạn dẻo (PL) và chỉ số độ dẻo (PI) có tương quan dương mạnh mẽ với OWC, với hệ số tương quan lần lượt đạt khoảng 0.73, 0.65 và 0.68. Điều này cho thấy những mẫu đất có tính dẻo cao thường có độ ẩm tối ưu lớn hơn, phù hợp với đặc tính giữ nước và khả năng đầm nén tốt hơn của đất nền. Ngược lại, hàm lượng sỏi (G)

thể hiện tương quan âm nhẹ với OWC, với hệ số khoảng -0.25, cho thấy đất chứa nhiều sỏi thường có độ ẩm tốt nhất thấp hơn do đặc tính giữ nước kém của các hạt thô. Các biến khác như hàm lượng bụi sét (SC) có tương quan dương vừa phải (khoảng 0.45), trong khi hàm lượng hữu cơ (O) có tương quan thấp hơn, gần bằng 0.1, cho thấy ảnh hưởng không quá nổi bật của thành phần hữu cơ đến OWC trong bộ dữ liệu này. Ma trận tương quan cũng cho thấy mối liên hệ chặt chẽ giữa một số biến đầu vào với nhau, đặc biệt là nhóm các biến liên quan đến tính dẻo (LL, PL, PI) có tương quan rất cao, điều này cảnh báo về khả năng đa cộng tuyến và cần được xem xét trong quá trình lựa chọn biến và xây dựng mô hình nhằm tránh ảnh hưởng đến hiệu quả dự báo. Tổng thể, ma trận tương quan cung cấp cái nhìn tổng quan về mối quan hệ giữa các biến vật lý và hóa học của đất nền với độ ẩm tốt nhất, hỗ trợ quan trọng cho việc lựa chọn biến đầu vào phù hợp và hiểu sâu hơn về các yếu tố ảnh hưởng trong mô hình dự báo.

Cơ sở dữ liệu sau khi thu thập và phân tích được chia thành hai phần: 70% dữ liệu được dùng để đào tạo mô hình và 30% dữ liệu còn lại được dùng để kiểm chứng mô hình.

2.2. Phương pháp nghiên cứu

2.2.1. Mô hình Gradient Boosting Regressor (GBR)

GBR là một phương pháp học máy thuộc nhóm boosting, được phát triển dựa trên ý tưởng của thuật toán Gradient Boosting do Friedman giới thiệu vào năm 1999. Phương pháp này tập trung xây dựng một mô hình dự báo thông qua quá trình kết hợp tuần tự các cây quyết định nhỏ (weak learners), nhằm giảm thiểu lỗi dự báo bằng cách tối ưu hóa trực tiếp hàm mất mát theo hướng gradient của lỗi. GBR đã trở thành một trong những công cụ mạnh mẽ và phổ biến trong các bài toán hồi quy phức tạp do khả năng xử lý dữ liệu phi tuyến và tính năng tăng cường hiệu suất mô hình qua từng bước huấn luyện [24].

Nguyên lý chính của GradientBoostingRegressor là xây dựng mô hình

hồi quy $F_m(x)$ theo dạng tổ hợp tuyến tính của các mô hình con $h_m(x)$, mỗi mô hình con được huấn luyện trên phần dư lỗi còn lại từ mô hình trước đó. Cụ thể, quá trình huấn luyện diễn ra qua các bước lặp, tại bước thứ m , mô hình con $h_m(x)$ được huấn luyện để dự đoán gradient âm của hàm mất mát L tại mô hình hiện thời $F_{m-1}(x)$. Hàm dự báo được cập nhật theo công thức:

$$F_m(x) = F_{m-1}(x) + v \cdot h_m(x) \quad (1)$$

trong đó, v là hệ số học (learning rate), điều chỉnh mức độ đóng góp của từng mô hình con vào mô hình tổng thể nhằm tránh quá khớp [25].

Trong bài toán dự báo giá trị độ ẩm tối ưu của đất nền sử dụng trong đầm nén, mô hình GBR có khả năng khai thác mối quan hệ phi tuyến phức tạp giữa các đặc tính vật lý của đất (như độ chặt, hàm lượng hạt mịn, độ dẻo...) và giá trị độ ẩm cần dự báo. Bằng cách tối ưu hóa hàm mất mát dựa trên sai số bình phương trung bình (MSE):

$$L = \frac{1}{N} \sum_{i=1}^N (y_i - F_m(x_i))^2 \quad (2)$$

với y_i là giá trị độ ẩm thực tế và $F_m(x_i)$ là giá trị dự báo tại điểm dữ liệu i , mô hình GBR được huấn luyện để giảm thiểu sai số này qua các vòng lặp [26].

Ưu điểm nổi bật của GBR bao gồm khả năng xử lý hiệu quả các dữ liệu phi tuyến và đa chiều, tính linh hoạt trong lựa chọn hàm mất mát, đồng thời tránh được vấn đề quá khớp nhờ việc điều chỉnh hệ số học và số lượng mô hình con. Tuy nhiên, nhược điểm của phương pháp là chi phí tính toán cao, thời gian huấn luyện dài và sự phụ thuộc vào việc điều chỉnh các siêu tham số như độ sâu cây, learning rate và số lượng vòng lặp, đòi hỏi quá trình tinh chỉnh kỹ lưỡng để đạt hiệu quả tốt nhất.

2.2.2 Đánh giá năng lực dự báo: Hệ số tương quan R, RMSE và MAE

Trong nghiên cứu này, để đánh giá khả năng dự báo của mô hình GBR, các chỉ số phổ biến như hệ số tương quan tuyến tính (R), sai số trung bình bình phương căn (RMSE) và sai số trung bình

tuyệt đối (MAE) đã được áp dụng. Các chỉ số này không chỉ phản ánh mức độ khớp giữa kết quả dự báo và dữ liệu thực nghiệm mà còn cho thấy khả năng mô hình vận hành hiệu quả trên các bộ dữ liệu mới, qua đó đảm bảo tính tổng quát và độ tin cậy của mô hình trong việc ước lượng giá trị độ ẩm tối ưu của đất.

Hệ số R biểu thị mức độ phù hợp giữa giá trị dự báo và giá trị thực tế, với giá trị nằm trong khoảng từ -1 đến 1. Giá trị càng gần 1 hoặc -1 cho thấy sự tương quan tuyến tính càng mạnh, trong khi giá trị gần 0 cho thấy mô hình không dự báo tốt. Công thức tính hệ số tương quan tuyến tính được biểu diễn như sau:

$$R = \frac{\sum_{i=1}^N (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2 \sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2}} \quad (3)$$

trong đó y_i là giá trị quan sát thực tế, $\hat{y}_i$ là giá trị dự báo, $\bar{y}$ và $\bar{\hat{y}}$ lần lượt là giá trị trung bình của y_i và $\hat{y}_i$.

Chỉ số RMSE đo lường mức độ sai lệch trung bình giữa giá trị dự báo và thực tế, cho biết sai số dự báo trung bình theo đơn vị của biến được quan tâm, được tính bằng công thức:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (4)$$

Giá trị RMSE càng nhỏ chứng tỏ mô hình có năng lực dự báo càng chính xác và sai số dự báo càng thấp. Việc kết hợp sử dụng cả R và RMSE giúp đánh giá toàn diện hơn về năng lực dự báo của mô hình, bao gồm cả mức độ phù hợp về mặt tương quan và mức độ sai số thực tế trong các dự báo [27].

Sai số tuyệt đối trung bình (MAE) là giá trị trung bình của các sai số tuyệt đối giữa giá trị thực tế và giá trị dự báo:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5)$$

MAE cũng càng nhỏ càng tốt, và thường dễ hiểu hơn RMSE do không bị ảnh hưởng bởi các giá trị ngoại lai lớn.

Ưu điểm của các số liệu này là trực quan, dễ tính toán và cho phép so sánh hiệu quả giữa các mô hình khác nhau trên cùng một bộ dữ liệu. Tuy nhiên, nhược điểm là các chỉ số như Corr và R^2 chỉ phản ánh mối quan hệ tuyến tính, trong khi RMSE và MAE có thể bị ảnh hưởng bởi phân phối dữ liệu hoặc các giá trị ngoại lai. Do đó, việc sử dụng đồng thời nhiều chỉ số xác thực giúp đánh giá toàn diện và khách quan hơn về hiệu quả của mô hình dự báo.

2.2.3. SHAP (SHapley Additive exPlanations)

SHAP là một phương pháp giải thích mô hình học máy dựa trên lý thuyết trò chơi Shapley, được phát triển bởi Lundberg và Lee vào năm 2017. Phương pháp này nhằm cung cấp các giá trị giải thích đóng góp từng biến đầu vào vào kết quả dự báo của mô hình một cách minh bạch và có cơ sở lý thuyết vững chắc. SHAP tận dụng khái niệm giá trị Shapley từ lý thuyết trò chơi để phân bổ công bằng giá trị đóng góp của từng đặc tính trong mô hình phức tạp, giúp người dùng hiểu rõ ảnh hưởng từng biến đến quyết định cuối cùng của mô hình [28].

Nguyên lý chính của SHAP là xây dựng mô hình giải thích dưới dạng hàm cộng tuyến của các giá trị Shapley cho từng biến đặc trưng. Cho một mô hình dự báo f và đầu vào x , giá trị dự báo $f(x)$ được biểu diễn dưới dạng tổng các giá trị SHAP của từng biến cộng với giá trị trung bình mô hình ϕ_0 :

$$f(x) = \phi_0 + \sum_{i=1}^M \phi_i \quad (6)$$

Trong đó:

- ϕ_0 là giá trị dự báo trung bình của mô hình,
- M là số lượng đặc trưng,
- ϕ_i là giá trị SHAP của đặc trưng i .

Giá trị Shapley ϕ_i được tính dựa trên trung bình trọng số của đóng góp thay đổi dự báo khi biến i được thêm vào các tập con biến đầu vào khác nhau, theo công thức:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (7)$$

Trong đó:

F là tập hợp tất cả các đặc trưng,

S là một tập con của F không chứa i ,

$f_s(x_s)$ là giá trị dự báo của mô hình khi chỉ sử dụng các đặc trưng trong S ,

$|S|$ là số lượng phần tử trong tập S .

Trong bài toán dự báo giá trị OWC, SHAP được sử dụng để phân tích và làm rõ vai trò của từng đặc tính vật lý của đất như hàm lượng hạt mịn, độ chặt, chỉ số dẻo, trong việc ảnh hưởng đến giá trị dự báo của mô hình GBR. Qua đó, SHAP giúp cải thiện tính minh bạch và giải thích mô hình, tạo điều kiện thuận lợi cho việc kiểm chứng và áp dụng mô hình trong thực tiễn [29].

Ưu điểm của SHAP là khả năng cung cấp giải thích mô hình với tính công bằng, nhất quán và có cơ sở lý thuyết vững chắc, đồng thời hỗ trợ đa dạng các loại mô hình học máy. Tuy nhiên, nhược điểm của phương pháp là tính toán giá trị Shapley có thể rất phức tạp và tốn thời gian đối với mô hình có nhiều biến đầu vào, mặc dù các thuật toán tối ưu đã được phát triển nhằm giảm thiểu vấn đề này. Ngoài ra, việc diễn giải giá trị SHAP cũng đòi hỏi người dùng có hiểu biết nhất định về mô hình và dữ liệu để tránh hiểu nhầm thông tin [30].

2.2.4. Partial Dependence Plots (PDP)

PDP là một kỹ thuật giải thích mô hình học máy được sử dụng để trực quan hóa và phân tích ảnh hưởng trung bình của từng biến đầu vào đến kết quả dự báo của mô hình. Phương pháp này được phát triển trong bối cảnh tăng cường khả năng giải thích các mô hình phức tạp như Gradient Boosting hay Random Forest, giúp người dùng hiểu rõ hơn về mối quan hệ giữa biến đặc trưng và biến mục tiêu một cách trực quan và dễ tiếp cận [24].

Nguyên lý của PDP dựa trên việc tính trung bình kết quả dự báo của mô hình khi giữ cố định một hoặc một số biến đầu vào ở các giá trị nhất định, trong khi các biến còn lại được phép thay đổi tự do theo phân phối trong tập dữ liệu huấn luyện. Cụ thể, với mô hình dự báo $f(x)$ và tập biến đặc trưng được phân tích là x_s , giá trị PDP cho biến tại

giá trị z được tính theo công thức:

$$\hat{f}_s(z) = \frac{1}{N} \sum_{i=1}^N f(z, x_{c_i}) \quad (8)$$

trong đó, N là số lượng mẫu dữ liệu, x_{c_i} đại diện cho các giá trị biến còn lại không thuộc tập S của mẫu thứ i . Giá trị $\hat{f}_s(z)$ thể hiện ảnh hưởng trung bình của biến đến kết quả dự báo tại điểm z .

Trong bài toán dự báo giá trị độ ẩm tối ưu của đất nền, PDP được sử dụng để minh họa và đánh giá ảnh hưởng của từng đặc tính vật lý như độ chặt, hàm lượng hạt mịn, đối với giá trị dự báo của mô hình GBR. Kết quả phân tích PDP giúp tăng tính minh bạch cho mô hình và hỗ trợ các kỹ sư trong việc đưa ra các quyết định thiết kế phù hợp hơn với đặc tính đất thực tế [29].

Ưu điểm của PDP là dễ hiểu, dễ triển khai và trực quan hóa mối quan hệ giữa biến đầu vào và đầu ra, đặc biệt hữu ích với các mô hình phức tạp không tuyến tính. Tuy nhiên, nhược điểm của phương pháp là không thể thể hiện được các tương tác phức tạp giữa nhiều biến cùng lúc, và kết quả có thể bị ảnh hưởng khi các biến đầu vào có sự phụ thuộc mạnh với nhau, dẫn đến giải thích không chính xác trong một số trường hợp [31].

3. Kết quả và thảo luận

Bảng 2 trình bày các chỉ số đánh giá hiệu suất của mô hình GradientBoostingRegressor trên cả hai tập dữ liệu huấn luyện và kiểm tra. Trên tập huấn luyện, mô hình đạt hệ số tương quan (R) rất cao là 0.960, cho thấy giá trị dự báo có sự tương quan chặt chẽ và gần như tuyến tính với giá trị thực tế. Hệ số xác định đạt 0.916, cho thấy mô hình giải thích được tới 91.6% sự biến thiên của dữ liệu huấn luyện, phản ánh khả năng mô hình nắm bắt tốt mối quan hệ giữa các biến đầu vào và biến mục tiêu. Các chỉ số lỗi dự báo cũng rất thấp, với lỗi tuyệt đối trung bình (MAE) là 0.491, lỗi bình phương trung bình căn bậc hai (RMSE) là 0.747, cùng độ lệch chuẩn (STD) của sai số khoảng 0.7459, biểu thị sự phân tán sai số nhỏ và độ ổn định cao của mô hình trên dữ liệu đã học.

Trên tập kiểm tra, các chỉ số hiệu suất có sự

giảm nhẹ nhưng vẫn duy trì mức độ cao với hệ số tương quan là 0.895 và bằng 0.776. Điều này chứng tỏ mô hình vẫn duy trì khả năng tổng quát hóa tốt và có thể dự báo chính xác trên dữ liệu mới chưa từng xuất hiện trong quá trình huấn luyện. Mức sai số trên tập kiểm tra tăng lên so với tập

huấn luyện, với MAE là 0.957 và RMSE là 1.282, cho thấy một số độ lệch trong dự báo, nhưng vẫn nằm trong giới hạn chấp nhận được cho các ứng dụng thực tiễn. Độ lệch chuẩn sai số trên tập kiểm tra cũng tăng lên, đạt khoảng 1.2578, phản ánh sự đa dạng và khó đoán hơn của dữ liệu ngoài mẫu.

Bảng 2. Kết quả của mô hình GBR

Chỉ số đánh giá	R	MAE	RMSE	STD
Dữ liệu đào tạo	0.960	0.491	0.747	0.74588
Dữ liệu kiểm chứng	0.895	0.957	1.282	1.25782

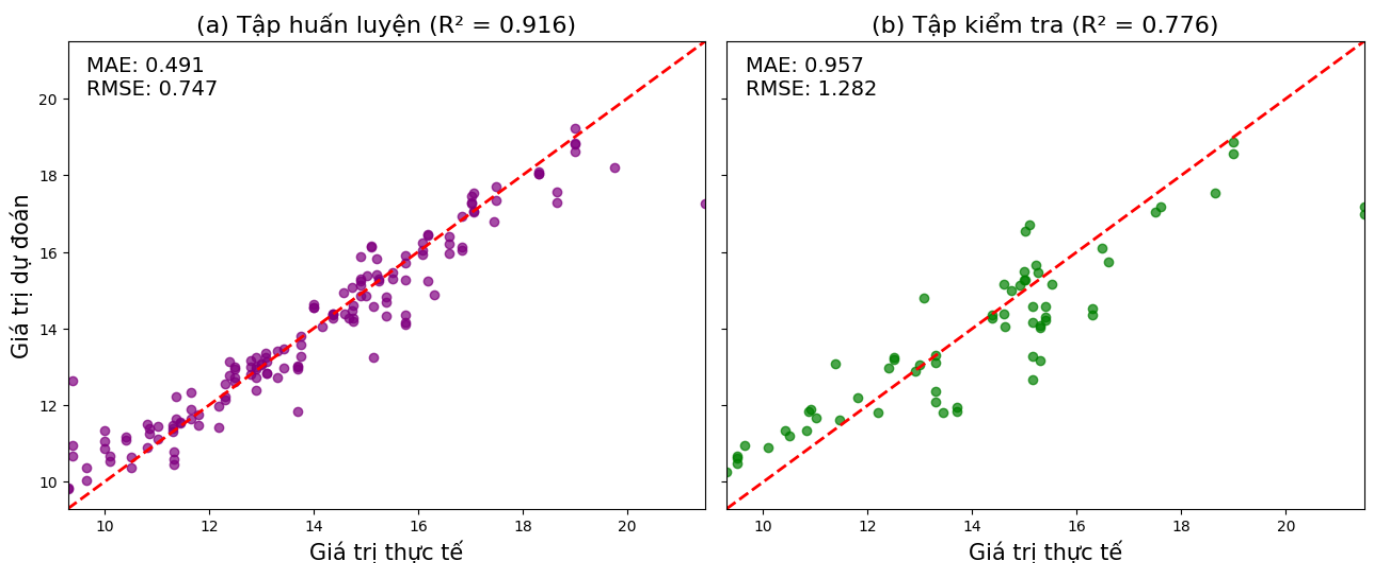
Nhìn chung, các kết quả trong bảng cho thấy mô hình GBR có hiệu suất dự báo vượt trội, thể hiện qua khả năng phù hợp tốt trên tập huấn luyện và khả năng tổng quát hóa cao trên tập kiểm tra. Điều này chứng minh mô hình phù hợp để áp dụng trong dự báo độ ẩm tốt nhất của đất nền trong các điều kiện khảo sát đa dạng và phức tạp.

Hình 3 trình bày biểu đồ phân tán so sánh giá trị dự đoán và giá trị thực tế của độ ẩm tốt nhất (OWC) được dự báo bởi mô hình GBR trên hai tập dữ liệu huấn luyện và kiểm tra.

Ở biểu đồ bên trái (tập huấn luyện), các điểm dữ liệu phân bố rất sát gần đường chéo $y = x$, biểu thị sự tương quan chặt chẽ giữa giá trị dự đoán và giá trị thực tế. Hệ số xác định cùng với lỗi tuyệt đối trung bình (MAE) là 0.491 và lỗi bình phương trung

bình căn bậc hai (RMSE) là 0.747 cho thấy mô hình có khả năng dự báo rất chính xác trên dữ liệu đã học. Độ tập trung cao của các điểm dữ liệu cho thấy mô hình học tốt và không có dấu hiệu overfitting trong giai đoạn huấn luyện.

Ở biểu đồ bên phải (tập kiểm tra), các điểm dữ liệu có sự phân tán rộng hơn, tuy nhiên phần lớn vẫn nằm gần đường $y = x$, thể hiện mô hình duy trì được khả năng tổng quát hóa tốt khi áp dụng trên dữ liệu chưa từng được học trước đó. Giá trị, MAE = 0.957 và RMSE = 1.282 phản ánh sai số dự báo tăng nhẹ so với tập huấn luyện nhưng vẫn ở mức chấp nhận được cho ứng dụng thực tế. Một số điểm lệch ra ngoài đường chéo có thể do đặc tính phức tạp của mẫu đất hoặc các yếu tố ngoại lai trong dữ liệu thực nghiệm.



Hình 3. Biểu đồ phân tán so sánh giá trị dự đoán và giá trị thực tế của độ ẩm tốt nhất

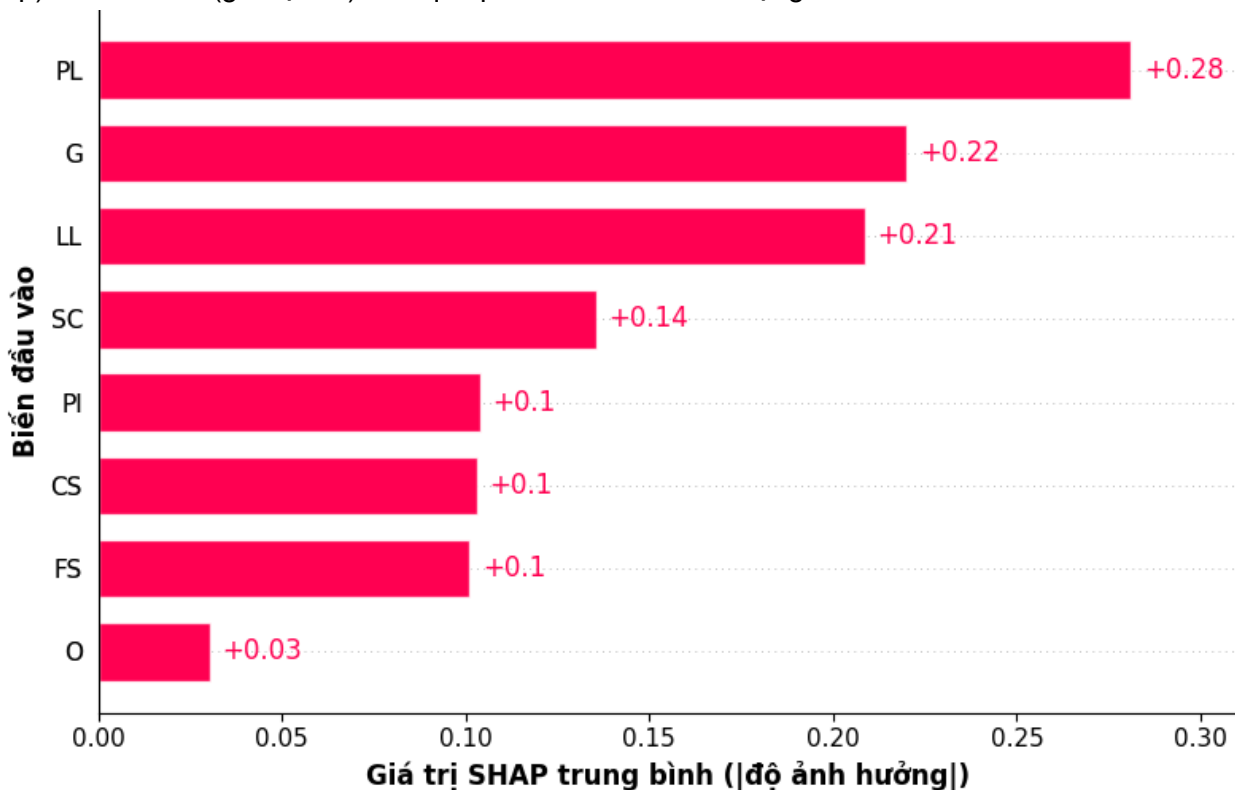
Kết quả của Hình 3 chứng minh rằng mô hình GBR có khả năng dự báo chính xác và ổn định giá

trị độ ẩm tốt nhất của đất nền, phù hợp với đa dạng đặc tính vật lý và hóa học trong bộ dữ liệu nghiên

cứu.

Hình 4 thể hiện biểu đồ SHAP dùng để phân tích mức độ ảnh hưởng của các biến đầu vào đến kết quả dự báo OWC trong mô hình GBR. Biểu đồ này cho thấy từng điểm dữ liệu được biểu diễn theo giá trị SHAP, phản ánh mức đóng góp của mỗi biến vào giá trị dự báo cuối cùng. Trục hoành thể hiện giá trị SHAP, trong khi màu sắc đại diện cho giá trị thực tế của biến đầu vào, từ màu xanh (giá trị thấp) đến màu đỏ (giá trị cao). Kết quả phân tích

cho thấy biến PL là biến có ảnh hưởng lớn nhất đến mô hình, với các giá trị cao của PL (màu đỏ) thường liên quan đến giá trị SHAP dương, tức là làm tăng dự báo OWC. Ngược lại, biến G có tác động ngược chiều, khi giá trị biến tăng cao (màu đỏ) thì giá trị SHAP giảm, làm giảm dự báo OWC. Các biến khác như LL, SC và CS cũng đóng vai trò quan trọng nhưng mức độ ảnh hưởng thấp hơn so với PL và G. Trong khi đó, các biến như FS và O có tác động nhỏ hơn.



Hình 4. Giá trị SHAP của các tham số ảnh hưởng

Biểu đồ SHAP giúp minh bạch hóa cấu trúc mô hình, cho phép hiểu rõ cách mỗi biến đầu vào tương tác và ảnh hưởng đến dự báo OWC. Việc xác định được biến quan trọng nhất cũng hỗ trợ tối ưu mô hình và giúp các kỹ sư địa kỹ thuật tập trung vào các đặc tính đất nền chủ chốt khi áp dụng mô hình dự báo trong thực tế.

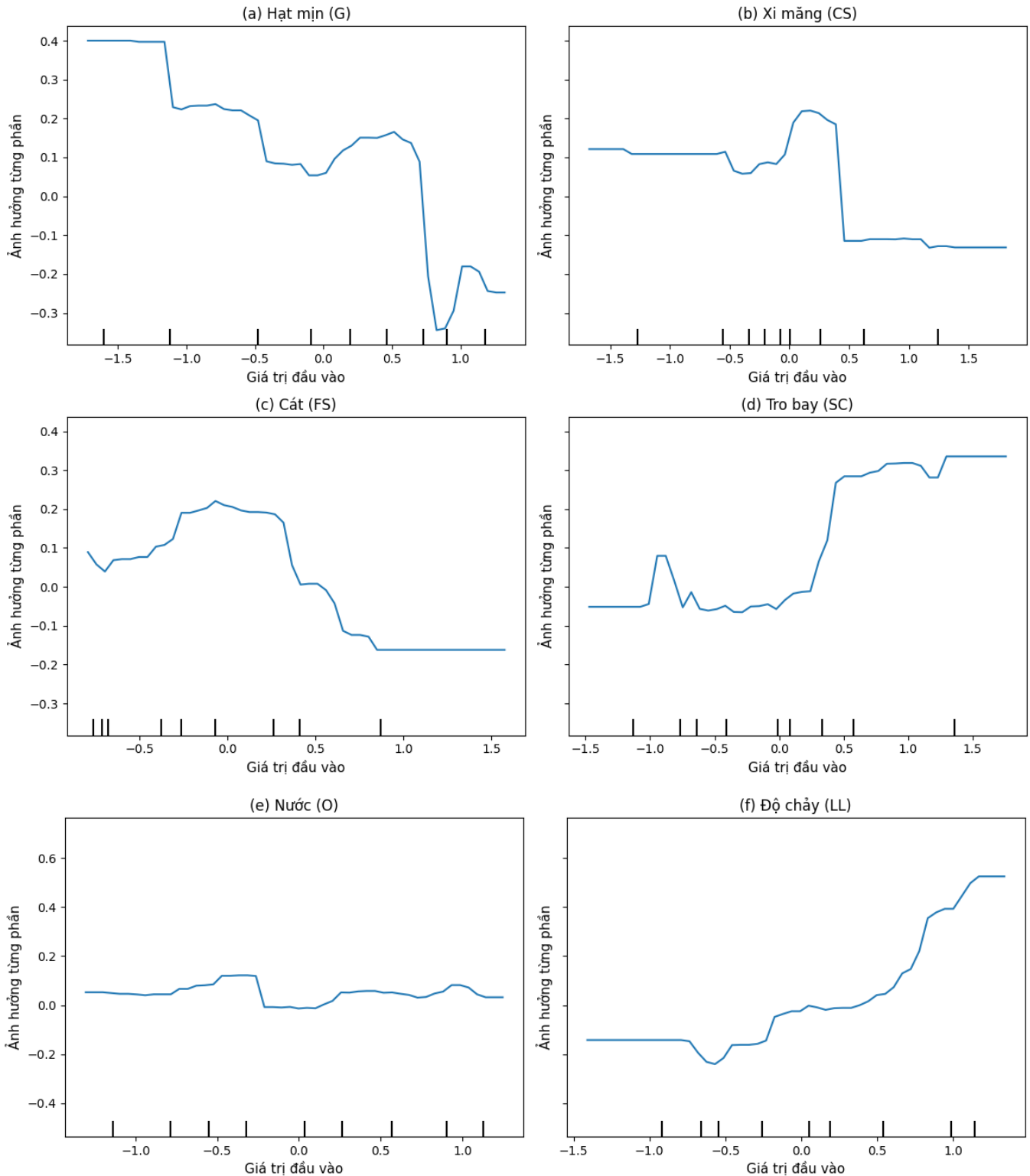
Hình 5 trình bày các biểu đồ PDP của các biến đầu vào nhằm mô tả mối quan hệ phi tuyến giữa từng biến và giá trị dự báo OWC của mô hình GBR. Biểu đồ PDP cho biến G cho thấy khi hàm lượng sỏi trong đất tăng từ mức thấp đến khoảng 30%, giá trị dự báo OWC tăng dần, đạt đỉnh tại khoảng giá trị 18%, sau đó có xu hướng giảm nhẹ

khi hàm lượng sỏi vượt quá mức này. Mối quan hệ này phản ánh đặc điểm kỹ thuật rằng hàm lượng sỏi vừa phải có thể giúp đất giữ nước tốt hơn, trong khi hàm lượng quá cao làm giảm khả năng giữ ẩm và ảnh hưởng tiêu cực đến giá trị OWC. Đối với biến CS, PDP biểu thị mối quan hệ phức tạp hơn, với giá trị dự báo OWC tăng nhẹ ở vùng giá trị thấp, sau đó giảm dần và có các dao động nhỏ khi hàm lượng cát thô thay đổi. Sự biến động này cho thấy tác động không đồng nhất của cát thô đến OWC, có thể phụ thuộc vào sự tương tác với các đặc tính đất khác. Biểu đồ PDP của biến FS cho thấy khi hàm lượng cát mịn tăng lên vượt quá mức 10–15%, giá trị dự báo OWC có xu hướng giảm nhẹ.

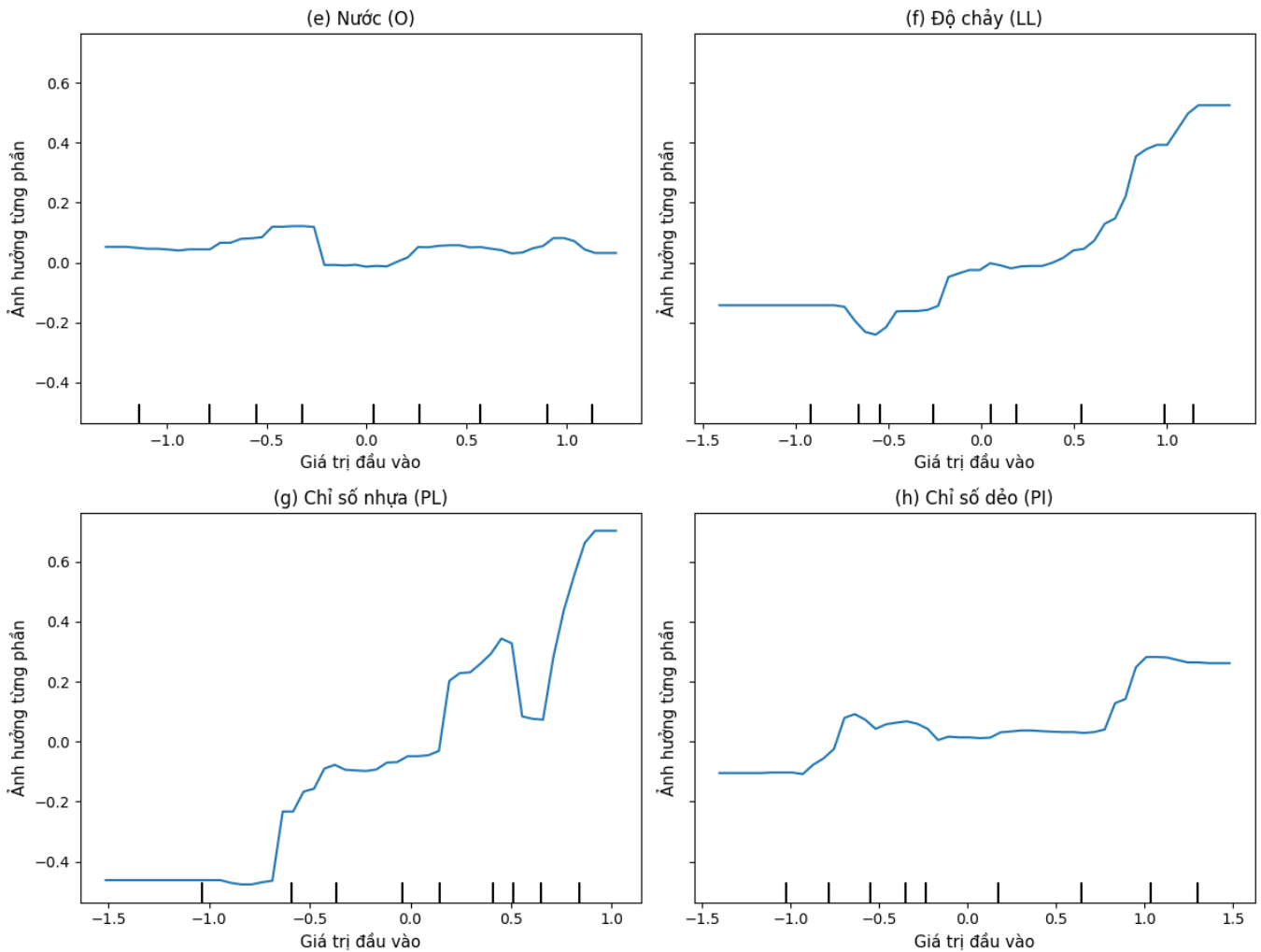
Điều này phù hợp với nhận định rằng đất chứa nhiều hạt mịn có thể giảm khả năng thoát nước và ảnh hưởng đến tính chất đầm nén, từ đó tác động lên độ ẩm tối ưu cần thiết.

Kết quả PDP cung cấp cái nhìn trực quan về ảnh hưởng phi tuyến và phức tạp của từng biến

đầu vào lên kết quả dự báo OWC. Những thông tin này không chỉ giúp làm rõ cơ chế hoạt động của mô hình GradientBoostingRegressor mà còn hỗ trợ các kỹ sư địa kỹ thuật trong việc hiểu rõ hơn về vai trò của từng đặc tính đất khi thiết kế và thi công nền đường.



Hình 5. Biểu đồ phân tích ảnh hưởng của các tham số đầu vào tới kết quả đầu ra của mô hình



Hình 5. (tiếp)

4. Kết luận

Nghiên cứu đã thành công trong việc ứng dụng mô hình GBR để dự báo giá trị OWC của đất nền sử dụng trong thi công đầm nén nền đường. Qua phân tích chi tiết bộ dữ liệu đa dạng về đặc tính vật lý và hóa học của mẫu đất, mô hình đã chứng minh khả năng xử lý tốt dữ liệu phức tạp, đa chiều và phi tuyến, từ đó nâng cao đáng kể độ chính xác dự báo so với các phương pháp truyền thống. Kết quả đánh giá trên tập huấn luyện và tập kiểm tra đều đạt các chỉ số hiệu suất cao, trong đó hệ số R lượt đạt 0.916 và 0.776, cùng với các chỉ số lỗi thấp như MAE và RMSE, chứng tỏ mô hình không chỉ học tốt dữ liệu đã biết mà còn giữ được tính ổn định và khả năng tổng quát hóa trên dữ liệu mới. Phân tích giải thích mô hình bằng phương pháp SHAP làm rõ vai trò quan trọng của các biến như giới hạn dẻo (PL) và hàm lượng sỏi (G), đồng

thời biểu đồ PDP cung cấp cái nhìn trực quan về mối quan hệ phi tuyến giữa các biến đầu vào và độ ẩm tối nhất. Những kết quả này không chỉ giúp rút ngắn thời gian và chi phí thí nghiệm trong xác định OWC mà còn cung cấp công cụ hỗ trợ tin cậy cho các kỹ sư trong việc thiết kế và thi công nền đường, góp phần nâng cao chất lượng và tuổi thọ công trình giao thông. Đồng thời, nghiên cứu mở ra hướng phát triển ứng dụng sâu rộng các thuật toán học máy tiên tiến trong lĩnh vực địa kỹ thuật và xây dựng, đặc biệt phù hợp với điều kiện đa dạng, phức tạp của các vùng đất nền tại Việt Nam. Trong các nghiên cứu tiếp theo, việc kết hợp thêm các nguồn dữ liệu mới và tối ưu hóa tham số mô hình hứa hẹn sẽ tiếp tục nâng cao hiệu quả dự báo, đồng thời mở rộng ứng dụng sang các đặc tính kỹ thuật khác của đất nền, nhằm đóng góp toàn diện cho phát triển bền vững ngành xây dựng cơ sở hạ tầng.

Tài liệu tham khảo

- [1] F. Karandish, J. Šimůnek. (2016). A comparison of numerical and machine-learning modeling of soil water content with limited input data. *Journal of Hydrology*, 543, 892-909. <https://doi.org/10.1016/j.jhydrol.2016.11.007>
- [2] A.R. Barzegar, M. Asoodar, M. Ansari. (2000). Effectiveness of sugarcane residue incorporation at different water contents and the Proctor compaction loads in reducing soil compactibility. *Soil and Tillage Research*, 57(3), 167-172. [https://doi.org/10.1016/S0167-1987\(00\)00158-6](https://doi.org/10.1016/S0167-1987(00)00158-6)
- [3] A. Baghbani, T. Choudhury, S. Costa, J. Reiner. (2022). Application of artificial intelligence in geotechnical engineering: A state-of-the-art review. *Earth-Science Reviews*, 228, 103991. <https://doi.org/10.1016/j.earscirev.2022.103991>
- [4] A. Keprate, R.M.C. Ratnayake. (2017). Using gradient boosting regressor to predict stress intensity factor of a crack propagating in small bore piping. *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, IEEE, pp. 1331-1336. doi: 10.1109/IEEM.2017.8290109
- [5] Bộ Khoa học và Công nghệ. (2014). TCVN4198, Đất xây dựng - Các phương pháp xác định thành phần hạt trong phòng thí nghiệm. Viện Khoa học Công nghệ Xây dựng - Bộ Xây dựng.
- [6] AASHTO. (2014). AASHTO-T-27-14, Standard Method of Test for Determining the Liquid Limit of Soils. American Association of State and Highway Transportation Officials.
- [7] AASHTO. (2004). AASHTO-M-145-91, Classification of Soils and Soil-Aggregate Mixtures For Highway Construction Purposes, American Association of State and Highway Transportation Officials.
- [8] AASHTO. (2018). AASHTO-T-267-86, Standard Method of Test for Determination of Organic Content in Soils by Loss of Ignition, American Association of State and Highway Transportation Officials.
- [9] Bộ Khoa học và Công nghệ. (2012). TCVN 4197, Đất xây dựng - Phương pháp xác định giới hạn dẻo và giới hạn chảy trong phòng thí nghiệm. Viện Khoa học Công nghệ Xây dựng - Bộ Xây dựng.
- [10] Bộ Khoa học và Công nghệ. (2020). TCVN12790, Đất, đá dăm dùng trong công trình giao thông - Đầm nén Proctor. Viện Khoa học và Công nghệ Giao thông vận tải.
- [11] Bộ Khoa học và Công nghệ. (2020). TCVN12792, Vật liệu nền, móng mặt đường - Phương pháp xác định tỷ số CBR trong phòng thí nghiệm. Viện Khoa học và Công nghệ Giao thông vận tải.
- [12] T.W. Lambe, R.V. Whitman. (2008). Soil mechanics SI version. *John Wiley & Sons*.
- [13] B. Das. (2011). Principles of geotechnical engineering. *Cengage Learning*.
- [14] D.P. Coduto, M.R. Yeung, W.A. Kitch. (2011). Geotechnical engineering: principles and practices. *Pearson*.
- [15] M. Chen, G. Wu, B. Gan, W. Jiang, J. Zhou. (2018). Physical and compaction properties of granular materials with artificial grading behind the particle size distributions. *Advances in Materials Science and Engineering*, 2018(1), 8093571. <https://doi.org/10.1155/2018/8093571>
- [16] R.F. Howard, M.J. Singer, G.A. Frantz. (1981). Effects of soil properties, water content, and compactive effort on the compaction of selected California forest and range soils. *Soil Science Society of America Journal*, 45(2), 231-236. <https://doi.org/10.2136/sssaj1981.03615995004500020001x>
- [17] C.J. Miller, N. Yesiller, K. Yaldo, S. Merayyan. (2002). Impact of soil type and compaction conditions on soil water characteristic. *Journal of Geotechnical and Geoenvironmental Engineering*, 128(9), 733. [https://doi.org/10.1061/\(ASCE\)1090-0241\(2002\)128:9\(733\)](https://doi.org/10.1061/(ASCE)1090-0241(2002)128:9(733))
- [18] Y. Wei, X. Wu, C. Cai. (2015). Splash erosion of clay-sand mixtures and its relationship with soil physical properties: The effects of particle size distribution on soil structure. *Catena*, 135, 254-262. <https://doi.org/10.1016/j.catena.2015.08.003>
- [19] S.K. Sinha, M.C. Wang. (2008). Artificial neural

- network prediction models for soil compaction and permeability. *Geotechnical and Geological Engineering*, 26, 47-64. <https://doi.org/10.1007/s10706-007-9146-3>
- [20] M.N. Navidi, J. Seyedmohammadi, S.A.S. Jalali. (2022). Predicting soil water content using support vector machines improved by meta-heuristic algorithms and remotely sensed data. *Geomechanics and Geoengineering*, 17(3), 712-726. <https://doi.org/10.1080/17486025.2020.1864032>
- [21] S. Maroufpoor, E. Maroufpoor, O. Bozorg-Haddad, J. Shiri, Z.M. Yaseen. (2019). Soil moisture simulation using hybrid artificial intelligent model: Hybridization of adaptive neuro fuzzy inference system with grey wolf optimizer algorithm. *Journal of Hydrology*, 575, 544-556. <https://doi.org/10.1016/j.jhydrol.2019.05.045>
- [22] M.N. Duc, A.H. Sy, T.N. Ngoc, T.L.H. Thi. (2022). An artificial intelligence approach based on multi-layer perceptron neural network and random forest for predicting maximum dry density and optimum moisture content of soil material in Quang Ninh Province, Vietnam. *CIGOS 2021, Emerging Technologies and Applications f1232or Green Infrastructure: Proceedings of the 6th international conference on Geotechnics, Civil Engineering and Structures, Springer, 2022, pp. 1745-1754.*
- [23] W.Z. Taffese, K.A. Abegaz. (2021). Artificial intelligence for prediction of physical and mechanical properties of stabilized soil for affordable housing. *Applied Sciences*, 11(16), 7503. <https://doi.org/10.3390/app11167503>
- [24] J.H. Friedman. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232. DOI: 10.1214/aos/1013203451
- [25] T. Hastie, R. Tibshirani, J. Friedman. (2009). The elements of statistical learning: Data Mining, Inference, and Prediction. *Springer*.
- [26] T. Chen, C. Guestrin. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794. <https://doi.org/10.1145/2939672.2939785>
- [27] T. Chai, R.R. Draxler. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific model development*, 7(3), 1247-1250. <https://doi.org/10.5194/gmd-7-1247-2014>
- [28] S.M. Lundberg, S.-I. Lee. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768 - 4777.
- [29] C. Molnar. (2020). Interpretable machine learning. *Lulu. com*.
- [30] E. Štrumbelj, I. Kononenko. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41, 647-665. <https://doi.org/10.1007/s10115-013-0679-x>
- [31] A. Goldstein, A. Kapelner, J. Bleich, E. Pitkin. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), 44-65. <https://doi.org/10.1080/10618600.2014.907095>