



Original Article

Research on the Application of CYGNSS Data and Machine Learning for Monitoring and Predicting Soil Salinity in the Ben Tre Area, Vinh Long Province

Pham Minh Quang¹, Hoang Tich Phuc¹, Vu Phuong Lan²,
Nguyen Huu Duy², Ha Minh Cuong^{1,*}

¹*VNU University of Engineering and Technology, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam*

²*VNU University of Science, 334 Nguyen Trai, Thanh Xuan, Hanoi, Vietnam*

Received 8th December 2025

Revised 10th April 2026; Accepted 19th June 2026

Abstract: Soil salinity is a serious environmental problem, directly affecting crop productivity and sustainable agricultural development, especially in the context of climate change. This study develops a low-cost method for mapping soil electrical conductivity (EC), based on machine learning combined with multi-source remote sensing data for the Ben Tre area, Vinh Long province, a coastal plain region of the Mekong Delta, Vietnam. Three machine learning models (CatBoost (CB), Random Forest (RF), XGBoost (XGB)) were used with 17 input variables extracted from MODIS imagery, ancillary data, and GNSS-R reflectivity data from the CYGNSS satellite. Feature importance analysis revealed that Surface Reflectivity (SR) derived from CYGNSS was one of the three most important predictors. The performance comparison showed a distinction between the models' learning capabilities and their generalization performance. On the training set, CatBoost demonstrated superior learning capability ($R = 0.89$), followed by XGB ($R = 0.84$) and RF ($R = 0.83$). However, on the validation set, RF model showed the best generalization performance, achieving $R = 0.71$. The resulting salinity distribution map indicates a clear spatial trend: salinity increases from the inland Western areas to the coastal Eastern areas. The research results confirm the feasibility of integrating machine learning and GNSS-R data for soil salinity monitoring.

Keywords: Soil salinity, CYGNSS, Machine learning, Remote sensing, Ben Tre, Vinh Long.

* Corresponding author.

E-mail address: cuonghm@vnu.edu.vn

<https://doi.org/10.25073/2588-1094/vnuces.5462>

Nghiên cứu ứng dụng dữ liệu CYGNSS và mô hình học máy trong giám sát và dự báo nhiễm mặn đất tại khu vực Bến Tre, tỉnh Vĩnh Long

Phạm Minh Quang¹, Hoàng Tích Phúc¹, Vũ Phương Lan²,
Nguyễn Hữu Duy², Hà Minh Cường^{1,*}

¹*Trường Đại học Công Nghệ, Đại học Quốc gia Hà Nội, 144 Xuân Thủy, Cầu Giấy, Hà Nội, Việt Nam*

²*Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội,
344 Nguyễn Trãi, Thanh Xuân, Hà Nội, Việt Nam*

Nhận ngày 8 tháng 12 năm 2025

Chỉnh sửa ngày 10 tháng 4 năm 2026; Chấp nhận đăng ngày 19 tháng 6 năm 2026

Tóm tắt: Nhiễm mặn đất là một trong những vấn đề môi trường nghiêm trọng, ảnh hưởng trực tiếp đến năng suất cây trồng và phát triển nông nghiệp bền vững, đặc biệt trong bối cảnh biến đổi khí hậu. Nghiên cứu này xây dựng phương pháp lập bản đồ nhiễm mặn đất (EC) chi phí thấp, dựa trên mô hình học máy kết hợp với dữ liệu viễn thám đa nguồn cho khu vực Bến Tre, tỉnh Vĩnh Long, vùng đồng bằng ven biển thuộc Đồng bằng sông Cửu Long, Việt Nam. Ba mô hình học máy (CatBoost (CB), Random Forest (RF), XGBoost (XGB)) được sử dụng với 17 biến đầu vào trích xuất từ ảnh MODIS, dữ liệu bổ trợ và dữ liệu phản xạ GNSS-R từ vệ tinh CYGNSS. Phân tích tầm quan trọng của biến cho thấy hệ số phản xạ bề mặt (SR) từ CYGNSS là một trong ba yếu tố dự báo quan trọng nhất. Kết quả so sánh hiệu suất cho thấy sự khác biệt giữa khả năng học và khả năng tổng quát hóa của các mô hình. Trên tập huấn luyện, CB thể hiện khả năng học vượt trội ($R = 0,89$), tiếp theo là XGB ($R = 0,84$) và RF ($R = 0,83$). Tuy nhiên, trên tập kiểm định, mô hình RF cho thấy khả năng tổng quát hóa tốt nhất, đạt $R = 0,71$. Bản đồ kết quả phân bố nhiễm mặn chỉ ra xu hướng không gian rõ rệt: độ mặn phân bố tăng dần từ nội đồng phía Tây (nhiễm mặn thấp) sang khu vực ven biển phía Đông (nhiễm mặn cao). Kết quả nghiên cứu khẳng định tính khả thi của việc tích hợp học máy và dữ liệu GNSS-R trong giám sát nhiễm mặn đất.

Từ khóa: Nhiễm mặn đất, CYGNSS, Học máy, Viễn thám, Bến Tre, Vĩnh Long.

1. Mở đầu

Hệ sinh thái đất là một trong những hệ thống phức tạp và đa dạng nhất trên thế giới, không chỉ cung cấp tới 98,8% nguồn lương thực cho con người mà còn đóng vai trò quan trọng trong việc lưu trữ carbon, điều hòa khí hậu và giảm thiểu biến đổi khí hậu [1]. Tuy nhiên, quá trình thâm canh nông nghiệp, áp lực của các hoạt động nhân

sinh và biến đổi khí hậu đã làm gia tăng nhanh chóng mức độ suy thoái đất trên toàn cầu [2]. Trong số các dạng suy thoái như xói mòn, ô nhiễm, nhiễm mặn đất được xem là một trong những vấn đề nghiêm trọng nhất đe dọa an ninh lương thực toàn cầu. Ước tính có khoảng 10–30% diện tích đất tưới tiêu trên thế giới bị ảnh hưởng bởi mặn hoặc kiềm, tương đương khoảng 76 triệu ha, trong đó 69% nằm ở khu vực châu Á

* Tác giả liên hệ.

Địa chỉ email: cuonghm@vnu.edu.vn

<https://doi.org/10.25073/2588-1094/vnuees.5462>

[3]. Mỗi năm, quá trình mặn hoá làm suy giảm năng suất trên khoảng 1,5 triệu ha đất canh tác, gây thiệt hại gần 10% sản lượng lương thực toàn cầu [4].

Tại Việt Nam, xâm nhập mặn đã và đang là vấn đề nghiêm trọng tại vùng Đồng bằng sông Cửu Long (ĐBSCL), khu vực nông nghiệp trọng điểm của cả nước. Theo Bộ Nông nghiệp và Phát triển nông thôn, xâm nhập mặn mùa khô năm 2019–2020 ở khu vực ĐBSCL được ghi nhận ở mức nghiêm trọng nhất trong lịch sử, ảnh hưởng đến 10/13 tỉnh trong vùng. Ranh giới độ mặn 4 g/l đã làm 42,5% diện tích tự nhiên của toàn vùng (tương đương 1.688.600 ha) bị ảnh hưởng, cao hơn năm 2016 là 50.376 ha. Trên cây ăn trái, hạn và xâm nhập mặn đã làm 6.650 ha tại sáu tỉnh (Long An, Tiền Giang, Bến Tre, Vĩnh Long, Trà Vinh và Sóc Trăng cũ nay là các tỉnh: Tây Ninh, Đồng Tháp, Vĩnh Long và thành phố Cần Thơ) thiếu nước tưới, trong đó 355 ha bị thiệt hại mất trắng [5]. Hệ quả của xâm nhập mặn kéo dài khiến cho đất bị nhiễm mặn tại khu vực Đồng bằng sông Cửu Long.

Riêng tại khu vực Bến Tre, chỉ trong lĩnh vực nông nghiệp và thủy sản, đã có 27.985 ha cây ăn quả, 1,2 triệu hoa kiểng các loại, 600 ha cây giống tại huyện Chợ Lách, 168 ha hoa màu và 3.097 ha nuôi thủy sản bị thiệt hại và ảnh hưởng. Ước tổng giá trị thiệt hại lên tới khoảng 1.660 tỷ đồng. Theo Bộ Nông nghiệp và Phát triển nông thôn thiệt hại do xâm nhập mặn năm 2019–2020 là rất lớn, tác động đến hầu hết các lĩnh vực kinh tế – xã hội, ảnh hưởng trực tiếp đến đời sống và sinh kế của hàng chục nghìn hộ dân ven sông, ven biển [6].

Các phương pháp truyền thống thường dựa vào đo đạc tại chỗ (có độ chính xác cao nhưng hạn chế về phạm vi không gian và chi phí tốn kém [7, 8]). Trong khi đó, viễn thám đã chứng minh là công cụ hữu hiệu cho quan trắc diện rộng, đặc biệt khi kết hợp với các mô hình học máy [1, 9]. Dữ liệu quang học như MODIS và Landsat cho phép trích xuất các chỉ số thực vật (NDVI, EVI,...) và chỉ số mặn (SI, NDSI,...) phản ánh gián tiếp tình trạng nhiễm mặn đất thông qua biến đổi sinh lý của thảm thực vật [10–12]. Tuy nhiên, các ảnh quang học phụ thuộc vào

điều kiện thời tiết và ánh sáng mặt trời, dẫn đến nhiều khoảng trống dữ liệu khi có mây hoặc mưa kéo dài.

Trong bối cảnh đó, dữ liệu GNSS-R từ vệ tinh CYGNSS (Cyclone Global Navigation Satellite System) mở ra một hướng tiếp cận mới. CYGNSS thu nhận tín hiệu phản xạ GPS ở băng tần L, cho phép ghi nhận độ phản xạ bề mặt (Surface Reflectivity - SR), thông tin nhạy cảm với độ nhám bề mặt [13], độ ẩm đất và độ dẫn điện của đất [14, 15]. So với các cảm biến quang học, CYGNSS có tần suất quan sát cao (trung bình 7,2 giờ/lần) và khả năng hoạt động trong điều kiện mây mù, rất phù hợp cho khu vực ven biển nhiệt đới như khu vực nghiên cứu [16]. Mặc dù CYGNSS đã được ứng dụng để giám sát độ ẩm đất [17], lũ lụt và đặc tính bề mặt, nghiên cứu sử dụng dữ liệu CYGNSS để lập bản đồ nhiễm mặn đất ở Việt Nam vẫn còn rất hạn chế [13, 14, 18–20].

Vì vậy, nghiên cứu này được thực hiện với mục tiêu khai thác dữ liệu phản xạ bề mặt từ CYGNSS kết hợp với các chỉ số độ mặn, chỉ số thực vật của ảnh quang học [21] để thành lập bản đồ nhiễm mặn đất cho khu vực Bến Tre. Dữ liệu được tích hợp và phân tích bằng các thuật toán học máy hiện đại gồm Random Forest (RF) [22], XGBoost (XGB) [23] và CatBoost (CB) [24], cho phép mô hình hóa các mối quan hệ phi tuyến phức tạp giữa các biến viễn thám và độ mặn thực tế. Cách tiếp cận này giúp cải thiện độ chính xác, giảm nhiễu và tăng khả năng khái quát hóa cho mô hình dự báo, góp phần hỗ trợ công tác quản lý tài nguyên nước và quy hoạch nông nghiệp thích ứng với biến đổi khí hậu.

2. Bộ dữ liệu và khu vực nghiên cứu

2.1. Khu vực nghiên cứu

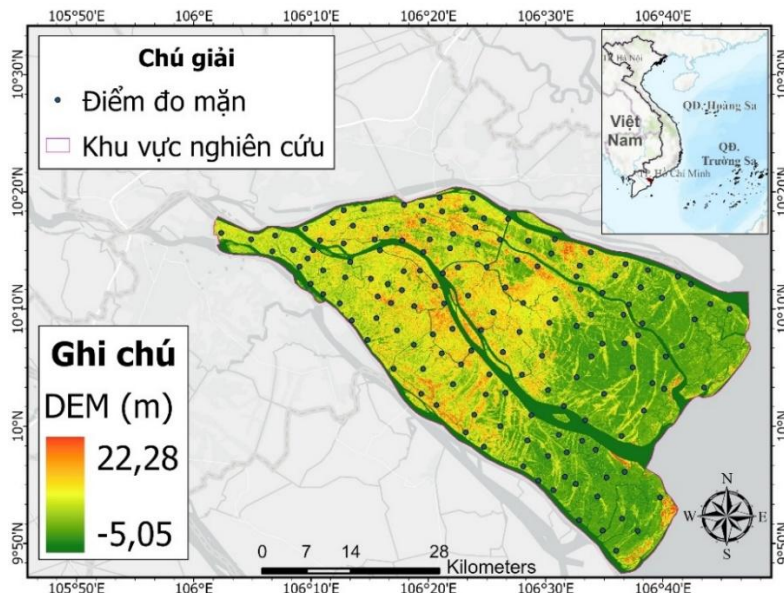
Khu vực Bến Tre (tỉnh Bến Tre cũ), nay là xã phường Bến Tre gồm 48 đơn vị hành chính cấp xã, tỉnh Vĩnh Long, nằm ở hạ lưu sông Mê Kông. Diện tích tự nhiên của khu vực khoảng 2.360 km², dân số tính đến năm 2024 đạt hơn 1,3 triệu người. Địa hình chủ yếu là đồng bằng thấp, cao trung bình từ 1 đến 3 mét so với mực nước

biển, bị chia cắt bởi mạng lưới sông ngòi dày đặc. Đây là khu vực chịu ảnh hưởng mạnh của chế độ bán nhật triều, với hai lần nước lên và xuống trong một ngày, gây biến động lớn về mực nước và độ mặn.

Khí hậu Bến Tre mang đặc trưng nhiệt đới gió mùa, có hai mùa rõ rệt: mùa mưa kéo dài từ tháng 5 đến tháng 11 và mùa khô từ tháng 12 đến tháng 4 năm sau. Lượng mưa trung bình hàng năm vào khoảng 1.400–1.600 mm, trong đó phần lớn tập trung trong mùa mưa. Đất đai tại khu vực này khá đa dạng, bao gồm đất phù sa, đất phèn và đất mặn, tạo điều kiện thuận lợi cho nông nghiệp và nuôi trồng thủy sản. Tuy nhiên, chính đặc điểm địa hình thấp và phụ thuộc vào nguồn

nước sông Mê Kông khiến khu vực này dễ bị ảnh hưởng nặng nề bởi nhiễm mặn đất trong mùa khô.

Trong nghiên cứu này, phạm vi không gian được xác định là toàn bộ ranh giới 48 xã phường Bến Tre (theo ranh giới tỉnh Bến Tre cũ), tỉnh Vĩnh Long (Hình 1). Phạm vi thời gian nghiên cứu tập trung từ tháng 2 đến tháng 4 năm 2020 – giai đoạn cao điểm của mùa khô, khi hiện tượng nhiễm mặn đất diễn ra mạnh nhất. Việc đánh giá và mô hình hóa nhiễm mặn đất trong khoảng thời gian này giúp làm rõ mối quan hệ giữa các yếu tố khí hậu, thủy văn và đặc trưng bề mặt, đồng thời hỗ trợ công tác quy hoạch, quản lý và ứng phó với biến đổi khí hậu trong tương lai.



Hình 1. Khu vực nghiên cứu xã phường Bến Tre (48 đơn vị cấp xã), tỉnh Vĩnh Long.

2.2. Bộ dữ liệu

Nghiên cứu này sử dụng một bộ dữ liệu địa không gian đa nguồn được thu thập trong giai đoạn mùa khô năm 2020, bao phủ toàn bộ khu vực Bến Tre, nơi chịu ảnh hưởng mạnh của nhiễm mặn đất. Cách tiếp cận này tích hợp nhiều nguồn dữ liệu nhằm phân tích các yếu tố môi trường và đặc trưng bề mặt liên quan đến quá trình nhiễm mặn đất (Bảng 1, Hình 2).

Một bước tiền xử lý quan trọng là đồng bộ hóa không gian, do các tập dữ liệu đầu vào có độ

phân giải gốc rất khác nhau (Bảng 1). Để đảm bảo tính nhất quán cho mô hình học máy, tất cả các lớp dữ liệu biến độc lập (Hình 2) đã được thống nhất độ phân giải về một lưới không gian chung là 1 km.

2.2.1. CYGNSS L1 v3.1

Dữ liệu CYGNSS Level 1 phiên bản 3.1 được sử dụng trong nghiên cứu này được thu thập trong giai đoạn từ tháng 2 đến tháng 4 năm 2020 và được tải về từ công dữ liệu NASA PODAAC. Sau khi thu thập, dữ liệu được tiến

hành tiền xử lý nhằm đảm bảo chất lượng và độ tin cậy cho các bước phân tích tiếp theo. Cụ thể, chỉ những quan sát có góc tới $\leq 65^\circ$ được giữ lại nhằm loại bỏ các phép đo có sai số cao do góc quan sát quá lớn.

Tiếp theo, quá trình lọc dữ liệu dựa trên hai biến cờ chất lượng (quality_flags và quality_flags_2) được áp dụng nhằm loại bỏ các phép đo có khả năng bị lỗi hoặc có độ tin cậy thấp. Đối với biến quality_flags, các phép đo bị loại bỏ bao gồm: khởi động bộ phát S-band, sai số vị trí vệ tinh, tín hiệu DDM thân đen, DDM là mẫu kiểm thử, tín hiệu trực tiếp và độ tin cậy thấp trong giá trị EIRP. Tương tự, với biến quality_flags_2, các phép đo bị loại bỏ khi tồn tại các lỗi như ăng-ten DDMI, lỗi tính toán ngưỡng, hoặc hệ số khuếch đại thấp ở góc thiên đỉnh thấp.

Việc áp dụng các bước tiền xử lý này giúp loại bỏ nhiễu hệ thống, giảm sai số và nâng cao độ tin cậy của tín hiệu phản xạ bề mặt thu được từ vệ tinh CYGNSS, phục vụ hiệu quả cho quá trình phân tích và mô hình hóa nhiễm mặn đất trong nghiên cứu [16, 25].

2.2.2. Chỉ số viễn thám

Các chỉ số viễn thám được tính toán từ ảnh MODIS (MOD09GA) [26], phản ánh các đặc trưng quang học của bề mặt đất và thảm thực vật trong điều kiện chịu ảnh hưởng của nhiễm mặn đất. Trước khi tính toán, dữ liệu được hiệu chỉnh khí quyển và lọc mây nhằm đảm bảo chất lượng phản xạ bề mặt đầu vào. Quá trình lọc mây được thực hiện thông qua mặt nạ chất lượng pixel (QC_500m), trong đó chỉ giữ lại các pixel có hai bit đầu bằng 00, tương ứng với các giá trị có chất lượng. Sau khi xử lý, các chỉ số được trích xuất bao gồm: chỉ số thảm thực vật (NDVI) thể hiện mức độ phát triển của thảm phủ; chỉ số mặn (NDSI) phản ánh đặc trưng khô hạn và độ sáng bề mặt; các chỉ số mặn (SI1–SI5) được thiết lập nhằm tăng cường khả năng nhận dạng khu vực bị ảnh hưởng bởi mặn thông qua sự thay đổi trong phản xạ phổ của các băng hồng ngoại và cận hồng ngoại. Ngoài ra, hai băng hồng ngoại gần (SWIR và SWIR2) được sử dụng để hỗ trợ tính toán và hiệu chỉnh các chỉ số, nhờ độ nhạy cao với hàm lượng nước và muối trên bề mặt. Các chỉ số này đóng vai trò quan trọng trong việc

mô hình hóa và phân tích không gian nhằm đánh giá mức độ và phạm vi ảnh hưởng của nhiễm mặn trên bề mặt đất.

Bảng 1. Bộ dữ liệu nghiên cứu

TT	Dữ liệu	Nguồn	Độ phân giải
1	CYGNSS L1 v3.1	PODAAC	1 km
2	Chỉ số thực vật (NDVI)	MOD09GA	500 m
3	Chỉ số mặn (NDSI)	MOD09GA	500 m
4	Chỉ số mặn (SI1)	MOD09GA	500 m
5	Chỉ số mặn (SI2)	MOD09GA	500 m
6	Chỉ số mặn (SI3)	MOD09GA	500 m
7	Chỉ số mặn (SI4)	MOD09GA	500 m
8	Chỉ số mặn (SI5)	MOD09GA	500 m
9	Băng phổ SWIR	MOD09GA	500 m
10	Băng phổ SWIR 2	MOD09GA	500 m
11	Độ cao (Elevation)	SRTM	30 m
12	Độ ẩm đất (Soil Moisture)	SMAP L3	9 km
13	Nhiệt độ bề mặt (LST)	MYD11A1	1 km
14	Lớp phủ bề mặt (LULC)	MCD12Q1	500 m
15	Hàm lượng cát (Sand)	SoilGrids	250 m
16	Hàm lượng sét (Clay)	SoilGrids	250 m
17	Khối lượng riêng đất (Bulk Density)	SoilGrids	250 m
18	Điểm đo mặn thực tế	Tổng hợp	Độ sâu 0-30 cm

2.2.3. Dữ liệu khí tượng

Dữ liệu khí tượng bao gồm độ ẩm đất từ sản phẩm SMAP L3 (độ phân giải 9 km) và nhiệt độ bề mặt (LST) từ MODIS (MYD11A1, độ phân giải 1 km), phản ánh trạng thái độ ẩm của đất và điều kiện nhiệt của bề mặt trong mùa khô – các

yếu tố có ảnh hưởng trực tiếp đến quá trình khuếch tán và lan truyền của mặn. Độ ẩm đất đóng vai trò quan trọng trong việc xác định khả năng giữ nước và tính dẫn truyền ion mặn trong tầng đất mặt, trong khi nhiệt độ bề mặt ảnh hưởng đến tốc độ bốc hơi và sự thay đổi cân bằng ẩm của môi trường đất.

2.2.4. Dữ liệu địa hình

Dữ liệu địa hình được sử dụng trong nghiên cứu là Mô hình số độ cao (DEM) của SRTM với độ phân giải không gian 30 m, thể hiện sự biến thiên độ cao bề mặt trên toàn khu vực nghiên cứu. Địa hình là một trong những yếu tố chi phối quá trình lan truyền và phân bố của nhiễm mặn đất, đặc biệt tại các khu vực đồng bằng ven biển có địa hình thấp như khu vực Bến Tre. Việc sử dụng dữ liệu DEM có độ chi tiết cao cho phép mô tả chính xác cấu trúc địa hình, từ đó hỗ trợ đánh giá xu thế di chuyển của dòng chảy mặn từ vùng cửa sông vào sâu nội đồng.

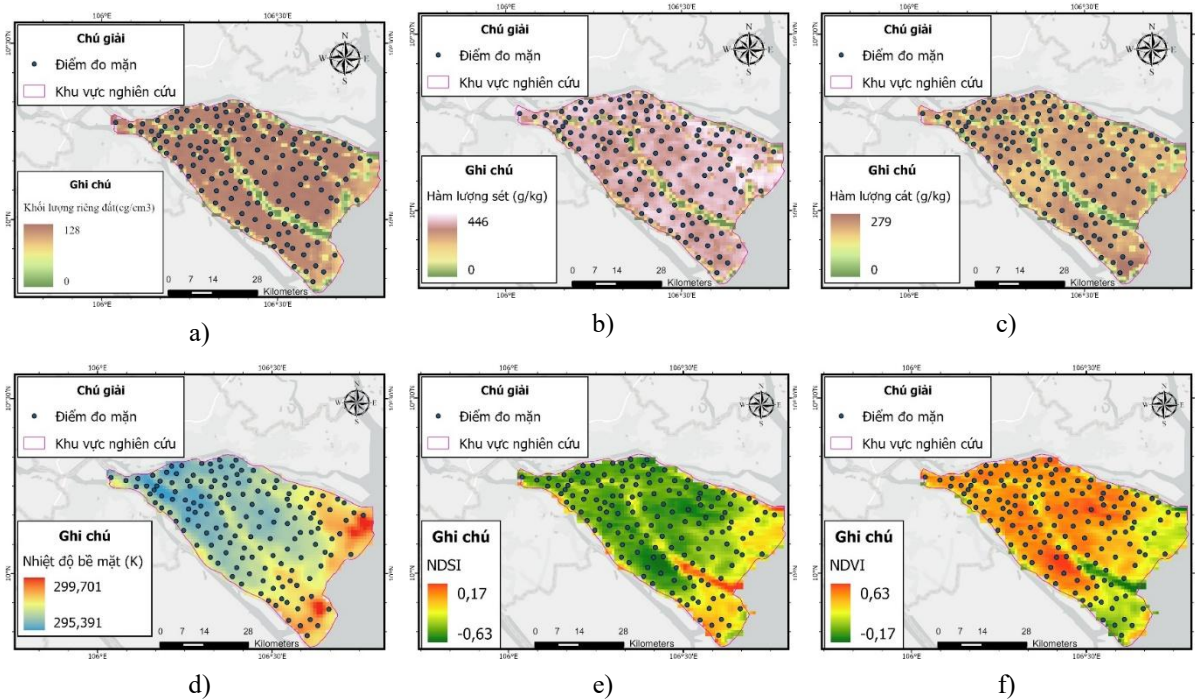
2.2.5. Dữ liệu thổ nhưỡng và lớp phủ bề mặt

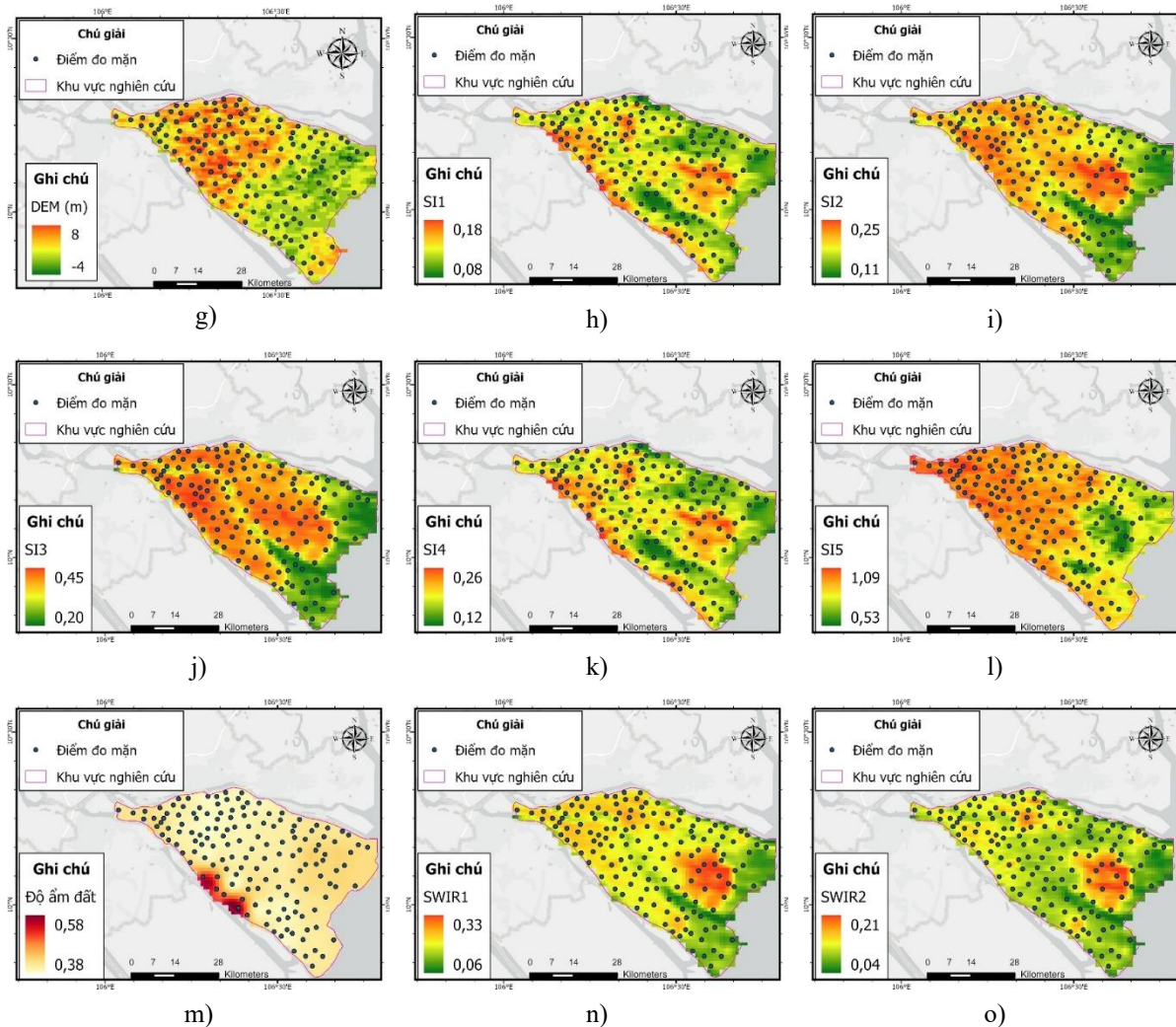
Dữ liệu lớp phủ bề mặt được lấy từ sản phẩm MODIS Land Cover (MCD12Q1) nhằm phân loại bề mặt ở khu vực Bến Tre. Dữ liệu thổ

nhưỡng từ SoilGrids (hàm lượng cát, hàm lượng sét và khối lượng riêng đất) được dùng để đánh giá tính thấm và khả năng giữ nước của đất, ảnh hưởng tới quá trình lan truyền nhiễm mặn đất.

2.2.6. Dữ liệu điểm mặn thực tế

Dữ liệu điểm mặn thực tế được thu thập thông qua khảo sát thực địa của nhóm tác giả tại khu vực Bến Tre. Tổng cộng 150 điểm mẫu đã được thu thập trong giai đoạn cao điểm của mùa khô, từ tháng 2 đến tháng 4 năm 2020, trùng với thời gian thu nhận của dữ liệu vệ tinh. Các điểm mẫu được bố trí phân bố đều trên toàn khu vực nghiên cứu nhằm đảm bảo tính đại diện không gian cho các khu vực chịu ảnh hưởng khác nhau của nhiễm mặn đất (Hình 2). Tại mỗi điểm, mẫu đất được lấy ở độ sâu từ 0-30 cm để xác định giá trị độ dẫn điện (EC) – tiêu chí phản ánh mức độ mặn trong đất. Quá trình lấy mẫu được tiến hành trong điều kiện thời tiết khô ráo, không có mây che phủ, nhằm hạn chế ảnh hưởng của các yếu tố khí tượng và đảm bảo tính đồng bộ với ảnh viễn thám [27]. Tập dữ liệu này được sử dụng làm biên phụ thuộc để huấn luyện, hiệu chỉnh và kiểm định các mô hình mô phỏng, góp phần nâng cao độ tin cậy của kết quả phân tích không gian.





Hình 2. Bộ dữ liệu nghiên cứu: a) Khối lượng riêng đất ở độ sâu 0-30 cm, b) Hàm lượng đất sét ở độ sâu 0-30cm, c) Hàm lượng cát ở độ sâu 0-30 cm, d) Nhiệt độ bề mặt, e) Chỉ số mặn NDSI, f) Chỉ số thảm thực vật, g) Độ cao địa hình, h)-l) Chỉ số mặn (SI1-SI5), m) Độ ẩm đất, n)-o): Hai băng phổ SWIR.

3. Phương pháp

3.1. Quy trình giám sát và dự báo nhiễm mặn đất

Việc dự báo nhiễm mặn đất tại khu vực Bến Tre được thực hiện thông qua bốn bước chính (Hình 3): i) Thu thập và xử lý dữ liệu; ii) Xây dựng mô hình dự báo độ mặn; iii) Kiểm định mô hình; và iv) phân tích – lập bản đồ phân bố độ mặn.

i) Thu thập và xử lý dữ liệu: Trong nghiên cứu này, dữ liệu được chia thành hai nhóm chính: (1) 150 mẫu EC thu thập qua khảo sát thực địa

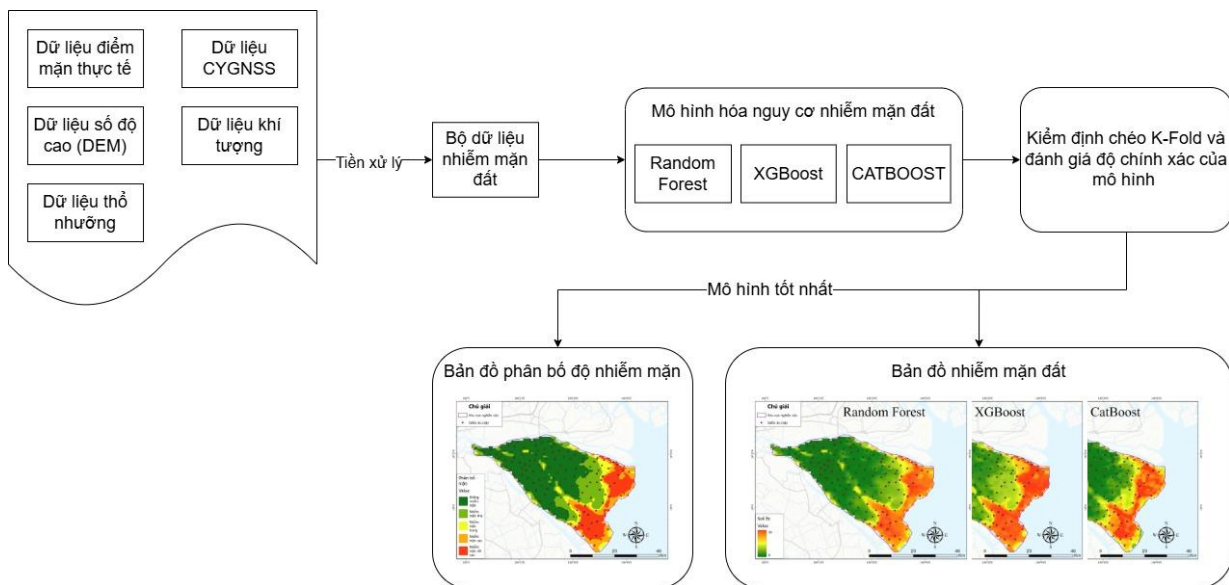
trong giai đoạn từ tháng 2 đến tháng 4 năm 2020, được sử dụng làm biến phụ thuộc (biến mục tiêu); và (2) 17 biến độc lập được tổng hợp từ nhiều nguồn dữ liệu khác nhau (như đã nêu ở mục 2.2), bao gồm hệ số SR trích xuất từ vệ tinh CYGNSS, 9 chỉ số viễn thám tính toán từ ảnh MODIS, cùng các biến hỗ trợ khác (địa hình, thổ nhưỡng, khí tượng). Việc thu thập dữ liệu thực địa và dữ liệu vệ tinh trong cùng một khoảng thời gian mùa khô nhằm đảm bảo tính đồng nhất về mặt thời gian cho quá trình phân tích.

ii) Xây dựng mô hình: Việc lựa chọn các biến độc lập phù hợp đóng vai trò quan trọng trong việc quyết định độ chính xác của mô hình, bởi sự dư thừa dữ liệu có thể làm tăng thời gian tính toán và giảm hiệu quả dự báo của mô hình hồi quy. Trong nghiên cứu này, 17 biến độc lập đã được lựa chọn từ các nhóm dữ liệu bao gồm dữ liệu thổ nhưỡng, địa hình, khí tượng, chỉ số viễn thám và đặc biệt là hệ số SR từ vệ tinh CYGNSS. Tập dữ liệu sau khi xử lý được sử dụng để xây dựng ba mô hình học máy gồm XGBoost (XGB), Random Forest (RF) và CatBoost (CB). Quá trình huấn luyện và kiểm định mô hình được thực hiện bằng phương pháp kiểm định chéo K-Fold ($K = 5$), giúp đảm bảo độ tin cậy và tính ổn định của kết quả dự báo, đồng thời giảm thiểu sai lệch do phân chia ngẫu nhiên dữ liệu.

iii) Đánh giá mô hình: Độ chính xác của các mô hình dự báo nhiễm mặn đất được đánh giá

thông qua ba chỉ tiêu thống kê chính gồm căn sai số bình phương trung bình (RMSE), sai số tuyệt đối trung bình (MAE) và hệ số tương quan (R). Các chỉ số này được sử dụng để định lượng mức độ phù hợp giữa giá trị dự báo và giá trị quan sát thực tế, qua đó phản ánh hiệu quả và khả năng tổng quát hóa của từng mô hình học máy trong việc ước lượng nhiễm mặn đất.

iv) Thành lập bản đồ nhiễm mặn đất: Sau khi các mô hình được hiệu chỉnh và đánh giá, mô hình có hiệu suất dự báo cao nhất được lựa chọn để thành lập bản đồ phân bố nhiễm mặn đất cho khu vực Bến Tre. Bản đồ này được xây dựng bằng cách tính toán giá trị EC cho từng pixel trên toàn bộ khu vực nghiên cứu, từ đó thể hiện không gian biến thiên của mức độ nhiễm mặn trong giai đoạn khảo sát.



Hình 3. Sơ đồ thành lập bản đồ nhiễm mặn đất.

3.2. Phương pháp ước tính phản xạ bề mặt từ dữ liệu CYGNSS

Khu vực Bến Tre nằm trong vùng đồng bằng ven biển có địa hình bằng phẳng, biên độ dao động địa hình nhỏ. Do đó, yếu tố địa hình hầu như không ảnh hưởng đáng kể đến độ chính xác của kết quả, và tín hiệu GNSS-R thu được bởi vệ

tinh CYGNSS chủ yếu mang đặc trưng phản xạ gương.

Hệ thống quan sát của CYGNSS hoạt động theo cấu trúc thu-phát tách biệt (bistatic), trong đó tín hiệu GNSS được phát ra từ vệ tinh phát và sau khi phản xạ trên bề mặt Trái Đất sẽ được vệ tinh thu nhận lại. Trong quá trình này, tín hiệu thay đổi phân cực từ phân cực tròn tay phải

(RHCP) ở pha phát sang phân cực tròn tay trái (LHCP) ở pha thu. Năng lượng phản xạ (P_r^c) thu được tuân theo phương trình radar bistatic [28]:

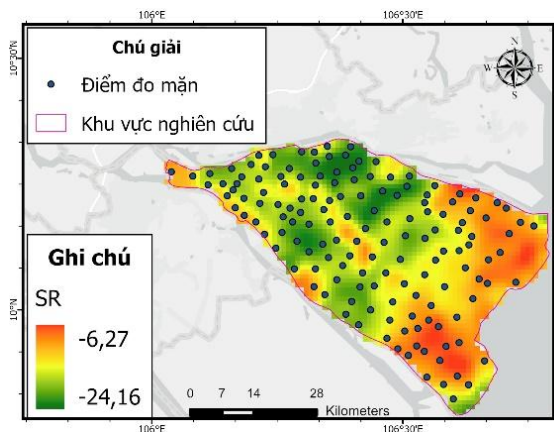
$$P_r^c = \frac{P_t \lambda^2 G_t G_r}{(4\pi)^2 (r_{st} + r_{sr})^2} \Gamma_{rl}(\theta_i) \quad (1)$$

Trong đó, λ là bước sóng của tín hiệu vệ tinh GPS (19 cm); $\Gamma_{rl}(\theta_i)$ là hệ số phản xạ bề mặt (SR) tại góc tới θ_i ; P_t là công suất phát (gps_tx_power_db_w); G_t là độ lợi của anten phát (gps_ant_gain_db_i); G_r là độ lợi của anten thu theo hướng điểm phản xạ SP (sp_rx_gain); còn r_{st} và r_{sr} lần lượt là khoảng cách từ vệ tinh phát đến điểm phản xạ và từ điểm phản xạ đến vệ tinh thu.

Trong thực tế, năng lượng phản xạ bề mặt P_r^c được tính toán thông qua giá trị đỉnh của ma trận công suất tán xạ mô phỏng DDM, ký hiệu là P_{DDM} , sau khi đã trừ đi lớp nhiễu nền N . Do đó, bằng cách đảo ngược phương trình (1), hệ số phản xạ bề mặt Γ_{rl} được tính toán theo công thức:

$$\Gamma_{rl}(\theta_i) = \frac{(P_{DDM} - N)(4\pi)^2 (r_{st} + r_{sr})^2}{P_t G_t G_r \lambda^2} \quad (2)$$

Trong đó, N là mức nhiễu nền trong ma trận DDM. Hai cột đầu tiên của ma trận DDM được xem là nhiễu, và giá trị trung bình của chúng được sử dụng để xác định N . Nếu giá trị công suất đỉnh P_{DDM} nằm trong vùng nhiễu (hai cột đầu tiên), điểm đó được xem là không hợp lệ và bị loại bỏ. Việc loại bỏ các điểm nhiễu này giúp xác định chính xác các điểm phản xạ hiệu lực, từ đó tính được hệ số phản xạ thực của bề mặt [28].



Hình 4. Bản đồ ước tính hệ số phản xạ Γ_{rl} (SR) khu vực Bến Tre.

Toàn bộ các giá trị trong công thức (2) đều được quy đổi về đơn vị tuyến tính trước khi tính toán. Hệ số phản xạ SR phản ánh tỷ lệ năng lượng tín hiệu GNSS bị phản xạ so với năng lượng tín hiệu tới. Hình 4 minh họa giá trị trung bình dạng lưới của SR trên lưới 1 km để thể hiện sự biến thiên theo khu vực.

Tín hiệu GNSS ở băng tần L chịu ảnh hưởng đáng kể của các điều kiện bề mặt, đặc biệt tại các khu vực có mặt nước hoặc vùng đồng bằng ven biển. Do đó, hệ số phản xạ SR đóng vai trò quan trọng, phản ánh trực tiếp mức suy hao của tín hiệu GNSS do quá trình tán xạ và khuếch tán sóng vô tuyến xảy ra tại bề mặt phản xạ.

3.3. Tính toán chỉ số viễn thám từ ảnh vệ tinh MODIS

Từ dữ liệu phản xạ bề mặt MODIS, các chỉ số viễn thám đặc trưng phản ánh tình trạng thảm thực vật và mức độ nhiễm mặn của bề mặt được tính toán nhằm phục vụ cho quá trình phân tích không gian. Trong đó, NDVI được sử dụng để biểu thị mức độ xanh và khả năng quang hợp của thảm thực vật; NDSI phản ánh sự biến đổi phổ do ảnh hưởng của muối hòa tan; trong khi các chỉ số SI1-SI5 được xây dựng nhằm tăng cường khả năng phát hiện và phân biệt các khu vực chịu tác động của nhiễm mặn đất thông qua sự kết hợp giữa các kênh phổ xanh, đỏ và cận hồng ngoại [29].

Các chỉ số này được tính toán dựa trên giá trị phản xạ bề mặt đã được hiệu chỉnh khí quyển của từng băng phổ trong ảnh MODIS, theo các công thức phổ biến được trình bày trong Bảng 2.

Bảng 2. Các chỉ số viễn thám

Các chỉ số viễn thám	Biểu thức tính toán
NDVI	$\frac{(NIR - RED)}{(NIR + RED)}$
NDSI	$\frac{(RED - NIR)}{(RED + NIR)}$
SI1	$\sqrt{GREEN \times RED}$
SI2	$\sqrt{NIR \times RED}$
SI3	$\sqrt{(GREEN)^2 \times (RED)^2 \times (NIR)^2}$
SI4	$\sqrt{(GREEN)^2 \times (RED)^2}$
SI5	$\frac{BLUE}{RED}$

3.4. Các thuật toán học máy

3.4.1 Thuật toán XGBoost

Thuật toán XGBoost (XGB) là một biến thể cải tiến của thuật toán Gradient Boosting, được đề xuất bởi Friedman [30, 31]. Đây là thuật toán học máy mạnh mẽ, thường được ứng dụng để giải quyết các bài toán hồi quy và phân loại phức tạp. Nguyên lý của XGB dựa trên học tăng cường tuần tự, trong đó nhiều mô hình học yếu, thường là các cây quyết định nông, được kết hợp lại để tạo thành một mô hình học mạnh. Quá trình huấn luyện được thực hiện theo cơ chế cộng dồn sai số: mô hình đầu tiên được huấn luyện trên dữ liệu gốc, sau đó mô hình tiếp theo tập trung dự đoán phần dư (sai số) của mô hình trước, và kết quả cuối cùng được tổng hợp từ tất cả các mô hình con [32].

XGB sử dụng thuật toán hạ dốc (gradient descent) để giảm thiểu hàm mất mát khi thêm các mô hình mới, giúp cải thiện hiệu suất dự đoán một cách liên tục. Một ưu điểm quan trọng của XGB là khả năng xử lý dữ liệu đa chiều và quy mô lớn, nên rất phù hợp cho các ứng dụng Big Data như dự báo hoặc phân vùng độ mặn đất [33]. Mô hình hạn chế hiện tượng quá khớp [34], cụ thể, XGB bổ sung thêm hai tham số điều chỉnh là L1 và L2 trong hàm mục tiêu, giúp kiểm soát độ phức tạp của mô hình và duy trì khả năng khái quát hóa.

Trong nghiên cứu này, việc tối ưu siêu tham số được thực hiện bằng phương pháp dò lưới (grid-search), một quy trình tự động thử nghiệm các tổ hợp tham số khác nhau để xác định cấu hình mang lại hiệu suất tối ưu [35]. Trong bước tối ưu này, nhiều dải giá trị siêu tham số đã được khảo sát nhằm đánh giá tác động của từng thành phần lên hiệu suất dự báo. Cụ thể, các tham số như số lượng mô hình tăng cường ($n_estimators$), độ sâu của cây (max_depth), tốc độ học ($learning_rate$) cùng với các tham số điều chuẩn và tỷ lệ lấy mẫu ($subsample$, $colsample_bytree$) được kết hợp theo nhiều cấu hình khác nhau trong quá trình dò lưới. Kết quả thử nghiệm cho thấy mô hình đạt chất lượng dự báo ổn định nhất khi sử dụng tổ hợp tham số gồm $n_estimators = 100$, $max_depth = 4$, $learning_rate = 0,03$, cùng với các giá trị điều chuẩn và lựa chọn đặc trưng

tương ứng. Cấu hình này giúp mô hình giảm sai số rõ rệt và duy trì khả năng khái quát hóa tốt hơn so với các thiết lập khác. Các nghiên cứu trước đây đã cho thấy rằng XGB không chỉ mang lại độ chính xác cao, mà còn rút ngắn đáng kể thời gian huấn luyện nhờ khả năng tối ưu hóa song song và xác định số vòng lặp tăng cường tối ưu trong một lần chạy [36, 37]. Nhờ những đặc tính này, mô hình XGB đã được áp dụng rộng rãi trong nhiều nghiên cứu đánh giá và dự báo mức độ nhạy cảm của các hiện tượng tự nhiên, bao gồm cả nhiễm mặn đất [38].

3.4.2. Thuật toán Random Forest

Thuật toán Rừng Ngẫu nhiên (Random Forest – RF) là một phương pháp học máy linh hoạt dựa trên cây quyết định. Nguyên lý cơ bản của RF là tạo ra nhiều cây quyết định độc lập từ các mẫu dữ liệu được chọn ngẫu nhiên, sau đó tổng hợp kết quả của các cây này để tạo ra một dự đoán duy nhất cho toàn bộ mô hình [22]. Một trong những ưu điểm nổi bật của mô hình RF là khả năng xử lý hiệu quả dữ liệu bị thiếu thông qua việc ước lượng giá trị bị khuyết dựa trên trung bình của các mẫu lân cận trong không gian đặc trưng. Nhờ đặc tính này, RF có thể duy trì độ ổn định và độ chính xác cao ngay cả khi tập dữ liệu không hoàn chỉnh.

Hiệu suất của mô hình phụ thuộc chủ yếu vào số lượng cây quyết định được tạo ra trong rừng và số lượng biến được chọn ngẫu nhiên tại mỗi lần phân tách - hai siêu tham số quyết định khả năng tổng quát hóa và mức độ phức tạp của mô hình. Trong nghiên cứu này, các siêu tham số của mô hình RF được tối ưu hóa bằng phương pháp dò lưới nhằm xác định cấu hình mang lại độ chính xác cao nhất. Quá trình tìm kiếm được thực hiện với nhiều giá trị khác nhau của số lượng cây quyết định $ntree$, từ 100 tới 1000, cùng với các tổ hợp giá trị của $mtry$. Kết quả cho thấy mô hình đạt hiệu suất dự báo tốt nhất khi $ntree = 500$, tại đó sai số giảm đáng kể và độ chính xác của RF cao hơn so với các cấu hình khác.

3.4.3. Thuật toán CatBoost

Thuật toán CatBoost (Categorical Boosting - CB) là một thuật toán học tăng cường (boosting) hiệu suất cao, được thiết kế để giải quyết các bài

toán hồi quy và phân loại [24]. Điểm khác biệt cốt lõi của CB nằm ở hai cơ chế tiên tiến: i) Phương pháp xử lý hiệu quả các đặc trưng dạng phân loại; và ii) Một biến thể của Gradient Boosting gọi là "Ordered Boosting" (Tăng cường theo thứ tự) để chống lại hiện tượng quá khớp.

Thay vì yêu cầu người dùng phải tiền xử lý, CB có thể xử lý trực tiếp các biến phân loại. Để làm điều này, nó tính toán các thống kê mục tiêu cho mỗi loại. Tuy nhiên, việc tính toán này nếu thực hiện một cách trực tiếp sẽ dẫn đến hiện tượng rò rỉ mục tiêu và gây ra quá khớp. CB giải quyết vấn đề này bằng cách sử dụng một cơ chế hoán vị ngẫu nhiên: đối với mỗi mẫu dữ liệu, các thống kê mục tiêu của nó được tính toán chỉ dựa trên các mẫu quan sát xuất hiện trước nó trong một tập dữ liệu đã được hoán vị ngẫu nhiên.

Bên cạnh đó, CB triển khai Ordered Boosting để giải quyết vấn đề dịch chuyển dự đoán vốn có trong các thuật toán boosting truyền thống. Thay vì tính toán gradient trên cùng một tập dữ liệu đã được sử dụng để xây dựng mô hình trước đó, CatBoost duy trì nhiều mô hình con và tính toán gradient cho mỗi mẫu trên một mô hình mà mẫu đó chưa được huấn luyện. Cơ chế này đảm bảo rằng các gradient được ước tính một cách không chệch, giúp mô hình duy trì khả năng tổng quát hóa cao và chống quá khớp hiệu quả [24].

Hiệu suất của CB phụ thuộc chủ yếu vào các siêu tham số như số vòng lặp (*iterations*), độ sâu cây (*depth*), tốc độ học (*learning rate*) và mức độ điều chuẩn (*l2 leaf reg*). Trong nghiên cứu này, các siêu tham số của mô hình được tối ưu hóa bằng phương pháp dò lưới (*grid-search*) nhằm xác định cấu hình mang lại độ chính xác dự báo tốt nhất. Quá trình tìm kiếm được thực hiện với nhiều giá trị khác nhau của *iterations* (từ 200 đến 600), *depth* (từ 2 đến 6), *learning rate* (từ 0,01 đến 0,1) và *l2 leaf reg*. Kết quả cho thấy mô hình đạt hiệu suất tối ưu tại cấu hình *iterations* = 450, *depth* = 2, *learning rate* = 0,02 và *l2 leaf reg* = 10,0, tại đó sai số dự báo giảm đáng kể và mô hình thể hiện khả năng khái quát hóa ổn định hơn so với các tổ hợp tham số khác.

3.5. Phương pháp đánh giá mô hình

Độ chính xác của các mô hình dự báo nhiệm vụ đất được đánh giá thông qua ba chỉ tiêu

thống kê chính, bao gồm: sai số bình phương trung bình (RMSE), sai số tuyệt đối trung bình (MAE) và hệ số tương quan (R). Các chỉ số này được sử dụng để định lượng mức độ phù hợp giữa giá trị dự báo và giá trị quan sát thực tế, qua đó phản ánh hiệu quả và khả năng tổng quát hóa của từng mô hình học máy trong việc ước lượng nhiệm vụ đất. Các chỉ số được tính toán theo công thức sau:

Sai số bình phương trung bình RMSE:

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (3)$$

Sai số tuyệt đối trung bình (MAE):

$$MAE = \frac{1}{n} \times \sum_{i=1}^n |Y_i - \hat{Y}_j| \quad (4)$$

Hệ số tương quan Pearson (R):

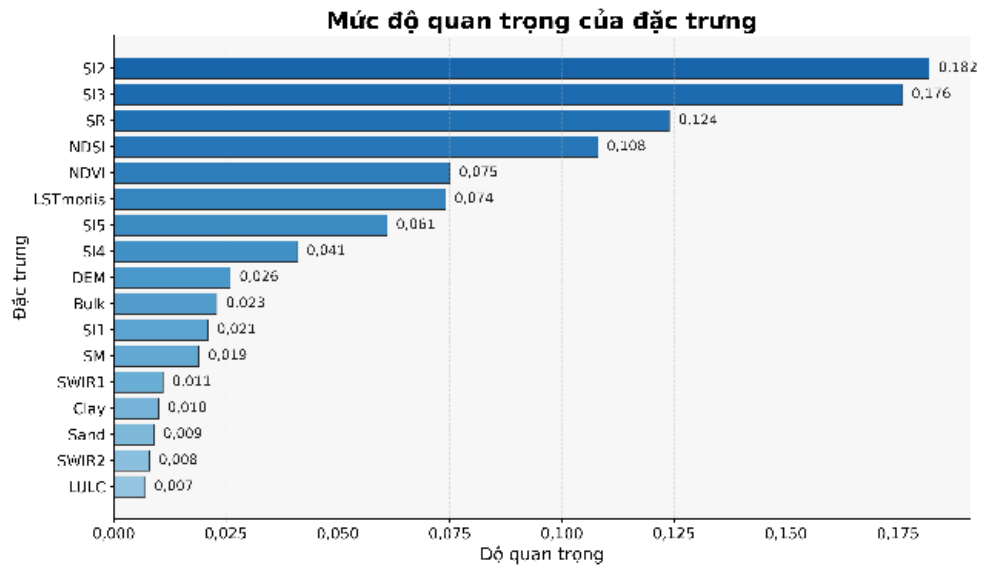
$$R = \sum_{i=1}^n \frac{(x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (5)$$

Trong đó, y_i là giá trị quan sát thực tế, $\hat{y}_i$ là giá trị dự báo của mô hình, và n là tổng số mẫu quan sát.

4. Kết quả

4.1. Tầm quan trọng của các biến đầu vào

Việc xác định và đánh giá mức độ đóng góp của các biến đầu vào là một bước quan trọng trong quá trình xây dựng mô hình dự báo nhiệm vụ đất. Để đảm bảo hiệu quả mô hình và tránh hiện tượng dư thừa dữ liệu, phương pháp RF được sử dụng để tính toán mức độ quan trọng của từng biến. Kết quả thể hiện trên Hình 5 cho thấy các chỉ số viễn thám và dữ liệu GNSS-R có ảnh hưởng lớn nhất đến khả năng dự báo nhiệm vụ đất. Cụ thể, các chỉ số SI2 (~0,18), SI3 (~0,175) và SR (~0,13) (hệ số phản xạ bề mặt từ CYGNSS) được xác định là ba biến có mức độ ảnh hưởng lớn nhất. Các biến NDSI (~0,11), NDVI (~0,09), LST (~0,09) (nhiệt độ bề mặt) và SI5 (~0,07) cũng thể hiện vai trò quan trọng trong mô hình. Nhóm biến gồm SI1, SM (độ ẩm đất), SWIR1, Clay, Sand, SWIR2 và LULC (lớp phủ bề mặt) có giá trị tầm quan trọng thấp hơn, phản ánh mối quan hệ gián tiếp với biến mục tiêu.



Hình 5. Tầm quan trọng các biến đầu vào

4.2. So sánh hiệu suất và lựa chọn mô hình

Kết quả so sánh hiệu suất của ba mô hình học máy RF, XGB, và CB trên tập huấn luyện và tập kiểm định được trình bày chi tiết trong Bảng 3.

Trên tập huấn luyện, mô hình CB thể hiện khả năng học vượt trội khi đạt hệ số tương quan cao nhất $R=0,89$ và các chỉ số lỗi thấp nhất $RMSE = 2,97$, $MAE = 1,93$.

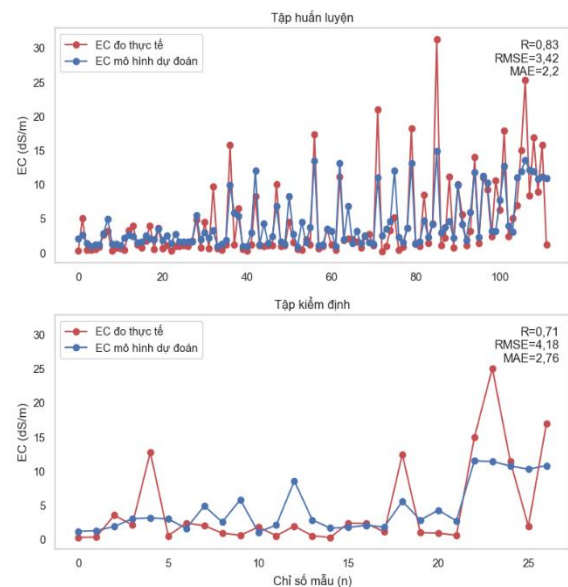
Bảng 3. Hiệu suất mô hình và so sánh

Mô hình huấn luyện	Tập huấn luyện		
	RMSE	MAE	R
RF	3,42	2,2	0,83
XGBoost	3,36	2,07	0,84
CatBoost	2,97	1,93	0,89
	Tập kiểm định		
	RMSE	MAE	R
RF	4,18	2,76	0,71
XGBoost	4,2	2,74	0,70
CatBoost	4,67	2,83	0,70

Tuy nhiên, hiệu suất trên tập kiểm định là yếu tố then chốt để đánh giá khả năng tổng quát hóa. Trên tập này, mô hình RF cho thấy hiệu suất tổng thể tốt nhất, đạt hệ số tương quan cao nhất R là 0,71 và sai số $RMSE$ thấp nhất. Ngược lại, mô hình CB, dù học tốt nhất trên tập huấn luyện,

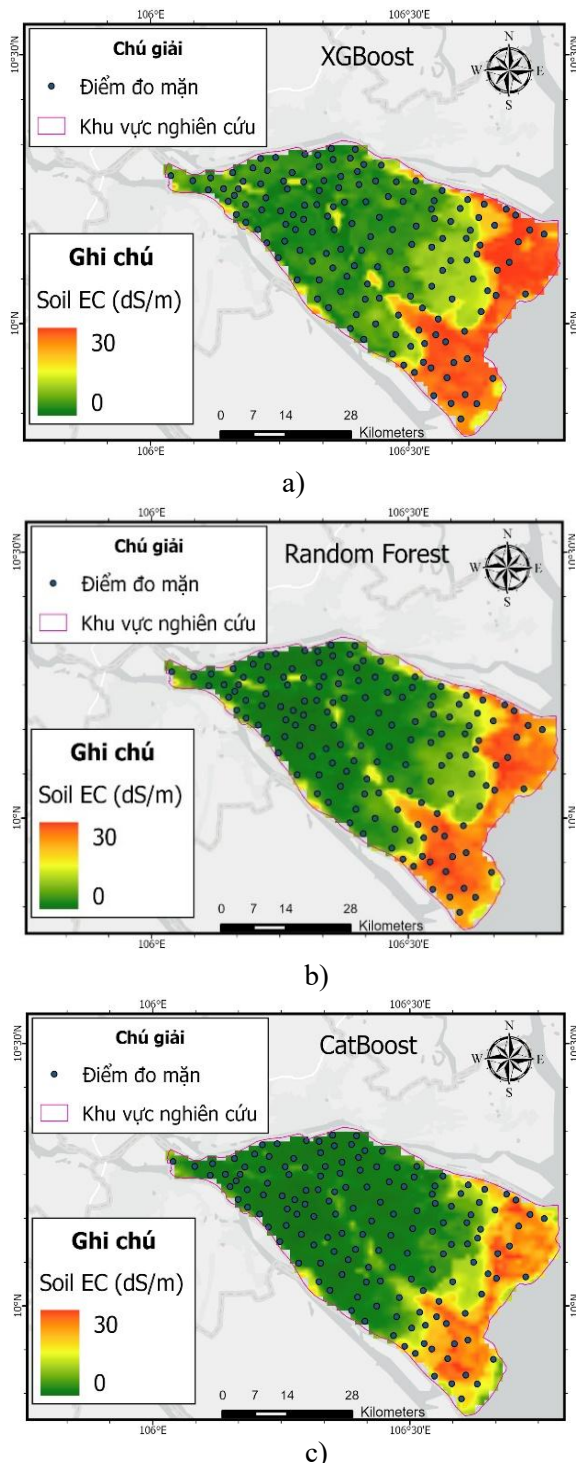
lại cho thấy hiệu suất trên tập kiểm định kém hơn ($R = 0,70$; $RMSE = 4,67$).

Do đó, RF được xác định là mô hình tối ưu và được lựa chọn để thành lập bản đồ nhiễm mặn đất trong nghiên cứu này. Hiệu suất chi tiết của mô hình RF được minh họa tại Hình 6, cho thấy sự tương quan giữa giá trị thực tế và giá trị dự đoán trên cả hai tập dữ liệu.



Hình 6. Hiệu suất mô hình RF.

4.3. Bản đồ phân bố độ mặn (EC)



Hình 7. Bản đồ nhiễm mặn đất ở khu vực Bến Tre sử dụng mô hình học máy: a) RF, b) XGB, c) CB.

Dựa trên mô hình RF đã được lựa chọn, bản đồ phân bố EC dưới dạng giá trị liên tục đã được thành lập cho toàn bộ khu vực nghiên cứu (Hình 7a).

Kết quả trực quan quan trọng nhất là cả ba bản đồ sử dụng các mô hình học máy khác nhau đều cho thấy một sự tương đồng rõ rệt về xu hướng phân bố không gian tổng thể của độ mặn (Hình 7a, 7b, 7c). Cả ba mô hình đều chỉ ra một quy luật phân bố rõ ràng, đặc trưng cho địa hình ven biển của khu vực Bến Tre: khu vực nội đồng ở phía Tây, nơi có địa hình tương đối cao hơn và xa biển, duy trì được giá trị EC ở mức thấp; ngược lại, khi di chuyển về phía Đông và Đông Nam, các khu vực ven biển tiếp giáp trực tiếp với Biển Đông, giá trị EC tăng lên rõ rệt. Cả ba bản đồ đều ghi nhận giá trị EC ở mức cao tại các vùng cửa sông và dọc theo toàn bộ đường bờ biển.

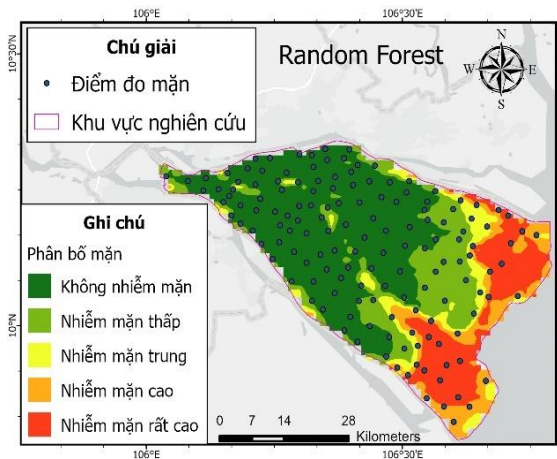
Tuy nhiên, khi so sánh chi tiết, có thể thấy sự khác biệt nhỏ trong việc biểu diễn giá trị và mức độ chi tiết của sự phân bố. Mô hình CB (Hình 7c) cho thấy khả năng phân bố chi tiết và nhạy cảm hơn, với sự phân hóa mạnh mẽ giữa các vùng có EC rất cao và các vùng lân cận, tạo ra các mảng màu có độ tương phản lớn. Ngược lại, mô hình XGB (Hình 7b) thể hiện kết quả có tính ổn định và khái quát hóa cao hơn, với các vùng chuyển tiếp mượt mà hơn và ít xuất hiện các đốm giá trị cực đoan. Mô hình RF (Hình 7a) cho kết quả trung gian, thể hiện sự cân bằng giữa hai mô hình còn lại. Bản đồ dự báo từ RF thể hiện rõ sự phân hóa tại khu vực ven biển đồng thời duy trì tính liên tục không gian, phù hợp với hiệu suất tổng quát hóa được ghi nhận trong quá trình kiểm định.

4.4. Bản đồ phân loại cấp độ nhiễm mặn

Để phục vụ trực tiếp cho công tác cảnh báo và quy hoạch nông nghiệp, bản đồ EC từ RF đã được phân loại thành các cấp độ nhiễm mặn cụ thể. Nghiên cứu sử dụng phương pháp phân loại Jenks để chia dữ liệu thành 5 cấp độ: Không nhiễm mặn, Nhiễm mặn nhẹ, Nhiễm mặn trung, Nhiễm mặn cao và Nhiễm mặn rất cao. Các ngưỡng giá trị này được xác định hoàn toàn dựa trên sự phân cụm thống kê của chính dữ liệu.

Hình 8 là bản đồ phân loại cấp độ nhiễm mặn cuối cùng thể hiện rõ nét sự phân vùng không gian. Khu vực nhiễm mặn cao và rất cao tập

trung dày đặc, bao phủ toàn bộ dải ven biển phía Đông và Đông Nam, là nơi chịu ảnh hưởng trực tiếp của thủy triều từ các cửa sông lớn. Khu vực nhiễm mặn trung bình xuất hiện rải rác, đóng vai trò như vùng đệm và phân bố dọc theo các nhánh sông chính, cho thấy đây là các kênh dẫn mặn chủ yếu vào sâu nội đồng.



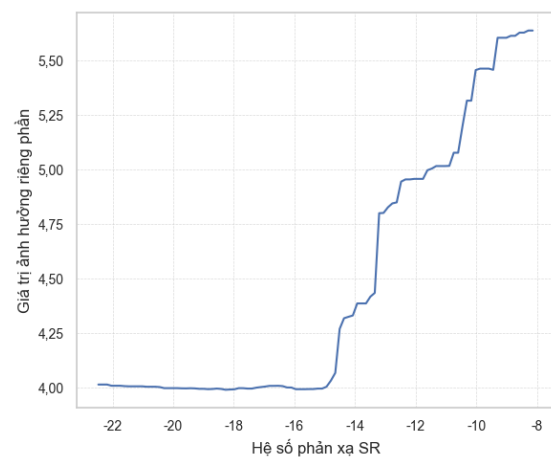
Hình 8. Bản đồ phân bố cấp độ nhiễm mặn khu vực Bến Tre.

5. Thảo luận

5.1. Ý nghĩa của các biến đầu vào và vai trò của CYGNSS trong mô hình

Phân tích tầm quan trọng của các biến (Hình 5) chỉ ra rằng hệ số SR trích xuất từ dữ liệu CYGNSS là một trong ba biến ảnh hưởng lớn nhất đến khả năng dự báo độ mặn, cùng với các chỉ số quang học SI2 và SI3. Kết quả này khẳng định giả thuyết ban đầu của nghiên cứu, cho thấy tín hiệu GNSS-R ở băng tần L, vốn nhạy cảm với đặc tính điện môi của đất, có mối liên hệ mạnh mẽ với sự thay đổi của độ mặn và độ ẩm bề mặt. Trong khi các chỉ số SI2 và SI3 từ MODIS rất hiệu quả nhưng lại bị hạn chế bởi điều kiện mây che phủ. Việc hệ số SR nổi lên như một yếu tố dự báo hàng đầu minh chứng tiềm năng của dữ liệu GNSS-R, với ưu điểm quan sát xuyên mây và tần suất cao, giúp khắc phục hạn chế của ảnh quang học trong giám sát các khu vực ven biển nhiệt đới.

Biểu đồ ảnh hưởng riêng phần cho biến SR ở trên cho thấy khi giá trị SR tăng, giá trị dự báo độ mặn tiếp cận kết quả dự báo của mô hình. Sự thay đổi của SR phản ánh rõ rệt sự biến động của độ mặn bề mặt, khẳng định vai trò quan trọng của tín hiệu GNSS-R trong việc giám sát và dự báo nhiễm mặn đất, nhất là ở những khu vực mà dữ liệu quang học bị hạn chế do mây che phủ. Kết quả này phù hợp với phân tích tầm quan trọng của các biến, cho thấy SR là một trong những yếu tố dự báo nổi bật trong mô hình Random Forest.



Hình 9. Biểu đồ ảnh hưởng riêng từng phần.

5.2. Sai số và tính không đồng nhất không gian

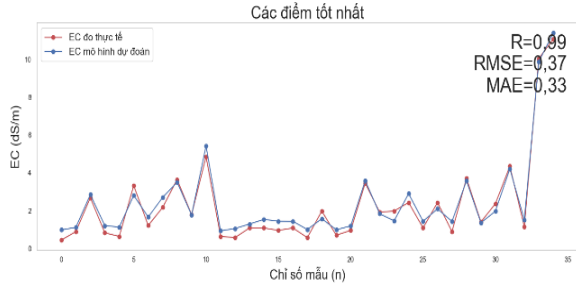
Mặc dù mô hình RF cho kết quả tổng quát hóa tốt ($R = 0,71$), hiệu suất này vẫn cho thấy một phần phương sai chưa được giải thích. Biểu đồ hiệu suất (Hình 6) chỉ ra rõ ràng rằng mô hình gặp khó khăn trong việc dự đoán các điểm có giá trị EC thực tế cao đột biến.

Nguyên nhân cốt lõi của sai số này không nằm ở thuật toán, mà ở bản chất của dữ liệu không gian. Giả thuyết của chúng tôi là sai số xuất phát từ hai vấn đề chính: i) Sự không tương thích về độ phân giải và; ii) Tính không đồng nhất của lớp phủ.

Mô hình cũng tăng rõ rệt, đặc biệt từ ngưỡng khoảng -14 trở lên. Kết quả này cho thấy SR có tác động mạnh và trực

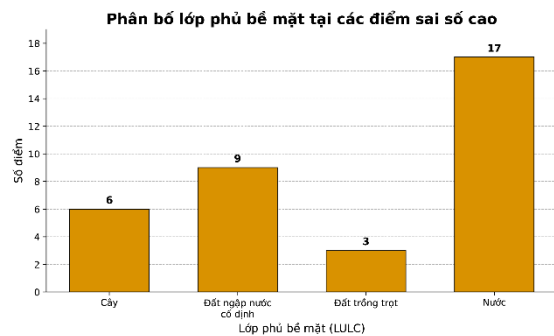
được huấn luyện để tìm mối liên hệ giữa dữ liệu đầu vào (đặc trưng trung bình của một ô lưới

1km x 1km) và dữ liệu đầu ra (một điểm đo EC thực địa). Một điểm đo có giá trị EC tăng vọt cục bộ có thể không đại diện cho đặc tính trung bình của cả ô lưới 1 km đó.



Hình 10. Biểu đồ hiệu suất mô hình tại các vị trí lớp phủ có tính đồng nhất.

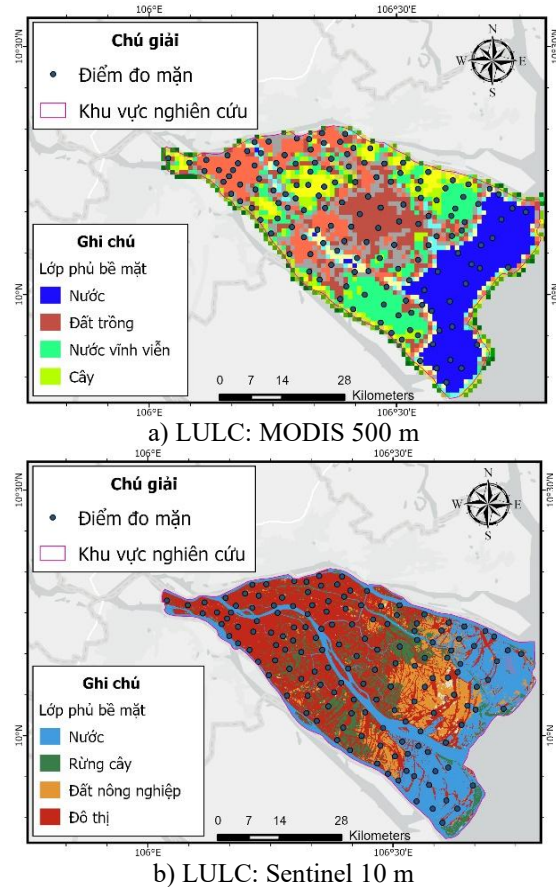
Khi lọc ra các điểm mẫu tại các khu vực được xác định là nằm ở các khu vực có tính đồng nhất không gian cao (khu vực trồng lúa và hoa màu), kết quả mô hình dự đoán đạt hiệu suất cải thiện vượt bậc đạt $R = 0,99$ (Hình 10). Kết quả này cho thấy khi dữ liệu đầu vào và đầu ra đồng nhất, mô hình RF có khả năng dự đoán độ chính xác cao và rất cao. Ngược lại, tại các điểm tập trung chủ yếu ở các lớp phủ không đồng nhất theo kết quả thống kê Hình 11, như vùng nước hay khu vực đô thị, hiệu suất mô hình tương ứng chỉ đạt $R=0,59$.



Hình 11. Thống kê các vị trí điểm tại khu vực lớp phủ bề mặt có tính không đồng nhất.

Bằng chứng rõ ràng nhất được thể hiện khi so sánh trực quan bản đồ lớp phủ (Hình 12 a,b). Một điểm đo mặn thực tế nằm trên *đất liền* (thể hiện trên bản đồ 10m) nhưng lại nằm hoàn toàn bên trong một pixel lớn được gán nhãn "Các

vùng nước" (trên bản đồ MODIS 500 m). Điều này tạo ra một xung đột tín hiệu: mô hình nhận dữ liệu đầu vào phản ánh đặc tính của nước, nhưng bị buộc phải học giá trị EC thực tế đo trên đất. Mô hình không thể dung hòa mâu thuẫn này, dẫn đến sai số dự đoán lớn.



Hình 12. Bản đồ phân bố điểm đo thực địa trên bản đồ phân loại đất LULC a) MODIS 500 m, và b) Sentinel 10 m.

Mặc dù ảnh MODIS với độ phân giải 500 m cho phép quan sát diện rộng và phù hợp cho đánh giá xu thế nhiễm mặn ở quy mô vùng, độ phân giải này vẫn còn tương đối thô đối với các khu vực có lớp phủ bề mặt phân mảnh. Do đó, kết quả từ MODIS thích hợp cho mục đích giám sát và phân tích ở quy mô tỉnh hoặc vùng, trong khi các ứng dụng chi tiết ở quy mô canh tác cần kết hợp thêm dữ liệu có độ phân giải cao hơn. Đối với khu vực nghiên cứu tại Bến Tre, nơi hiện

tượng nhiễm mặn chịu ảnh hưởng rõ rệt của yếu tố thủy văn và phân bố theo các vùng không gian tương đối rộng, dữ liệu MODIS vẫn cho thấy khả năng phản ánh khá tốt xu thế phân bố nhiễm mặn ở quy mô vùng, do đó có thể được xem là đủ tin cậy cho mục đích phân tích trong nghiên cứu này.

5.3. Sự phân bố không gian nhiễm mặn đất khu vực Bến Tre

Các bản đồ kết quả Hình 7 không chỉ là minh chứng cho hiệu quả dự báo của mô hình, mà còn phản ánh rõ nét đặc điểm phân bố hiện trạng nhiễm mặn đất tại khu vực Bến Tre trong thời điểm nghiên cứu mùa khô năm 2020. Mặc dù các bản đồ này chủ yếu mô tả trạng thái phân bố tĩnh tại một thời điểm nhất định, nhưng sự phân bố không gian với độ mặn tăng dần từ Tây sang Đông có thể được xem là hệ quả tích lũy từ sự tác động cộng hưởng của nhiều yếu tố.

Yếu tố tức thời và có ảnh hưởng mạnh nhất trong nghiên cứu này là điều kiện thời tiết trong thời kỳ khảo sát. Dữ liệu được thu thập vào giai đoạn mùa khô lịch sử 2019–2020, thời điểm khu vực Đồng bằng sông Cửu Long trải qua một trong những đợt hạn hán và nhiễm mặn đất nghiêm trọng nhất trong nhiều năm gần đây. Trong giai đoạn này, lượng mưa suy giảm mạnh so với trung bình nhiều năm, trong khi nhiệt độ cao kéo dài làm gia tăng quá trình bốc hơi nước trên diện rộng. Đồng thời, lưu lượng nước ngọt từ thượng nguồn sông Mê Kông đổ về hạ lưu bị giảm đáng kể do sự thiếu hụt nguồn nước từ các lưu vực phía thượng nguồn. Khi dòng chảy nước ngọt suy yếu, khả năng rửa trôi và pha loãng muối trong đất giảm rõ rệt, tạo điều kiện để áp lực nước biển từ phía Đông gia tăng. Dưới tác động của thủy triều, nước mặn theo các cửa sông lớn xâm nhập vào hệ thống sông rạch và kênh mương, làm tăng hàm lượng muối trong nguồn nước tưới và dần tích tụ trong đất canh tác. Vì vậy, ranh giới giữa các vùng đất mặn và đất ngọt có xu hướng dịch chuyển sâu hơn vào nội đồng, làm mở rộng phạm vi các khu vực bị nhiễm mặn đất. Hiện tượng này được thể hiện khá rõ trên các bản đồ phân bố độ mặn đất, khi các dải giá trị mặn cao tập trung dọc theo các trục sông lớn và lan rộng dần về phía nội đồng.

Bên cạnh các yếu tố mang tính thời điểm, những tác động dài hạn như biến đổi khí hậu và mực nước biển dâng cũng đang từng bước làm thay đổi chế độ thủy văn của toàn vùng. Sự gia tăng mực nước biển làm tăng áp lực nước mặn từ phía biển, đồng thời khiến quá trình tích tụ muối trong đất có xu hướng diễn ra mạnh hơn theo thời gian. Ngoài ra, sự phân hóa không gian thể hiện trên bản đồ cũng phản ánh khá rõ đặc điểm cơ cấu kinh tế nông nghiệp của khu vực nghiên cứu. Các vùng nội đồng phía Tây, nơi tập trung nhiều khu vực trồng cây ăn trái đặc sản và các vùng sản xuất nông nghiệp truyền thống, vẫn duy trì được điều kiện đất và nguồn nước ngọt tương đối ổn định nên mức độ nhiễm mặn đất thấp hơn. Ngược lại, các vùng ven biển phía Đông và Đông Nam chịu ảnh hưởng trực tiếp của thủy triều biển, đồng thời là khu vực phát triển mạnh các mô hình nuôi trồng thủy sản nước lợ và nước mặn, vì vậy đất tại đây thường có độ mặn cao. Điều này cho thấy sự liên hệ chặt chẽ giữa điều kiện tự nhiên và định hướng sử dụng đất trong khu vực. Qua đó, bản đồ phân bố nhiễm mặn đất không chỉ phản ánh hiện trạng môi trường đất mà còn có giá trị thực tiễn quan trọng trong việc hỗ trợ quy hoạch sử dụng đất, định hướng chuyển đổi cơ cấu sản xuất nông nghiệp và phân vùng canh tác phù hợp với điều kiện sinh thái của từng khu vực.

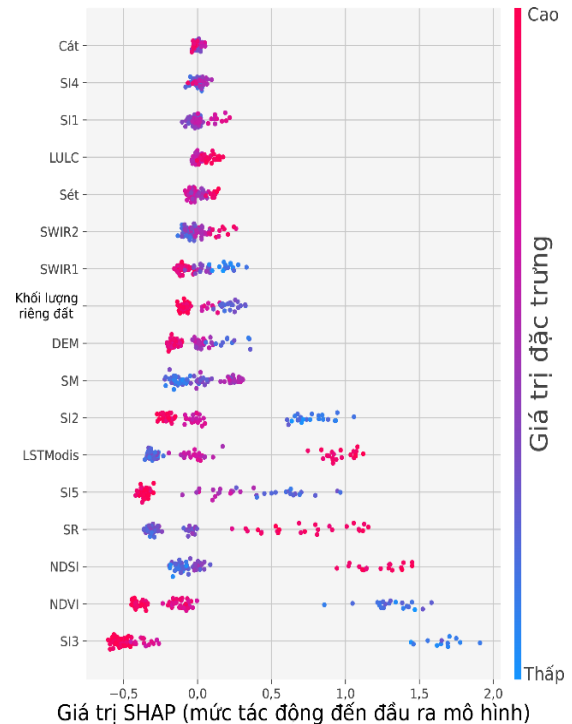
Mô hình được xây dựng trên cơ sở kết hợp dữ liệu viễn thám với tập hợp điểm đo thực địa có quy mô vẫn còn hạn chế, đồng thời một số điểm mẫu phân bố tại những vị trí không đồng nhất. Do đó, giá trị phản xạ phổ có thể chịu ảnh hưởng của nhiều thành phần bề mặt khác nhau, trong khi giá trị EC đo tại thực địa chỉ phản ánh điều kiện tại một vị trí cụ thể, từ đó có thể làm gia tăng sai số trong quá trình thiết lập mối quan hệ giữa dữ liệu viễn thám và độ mặn đất. Bên cạnh đó, sự phụ thuộc một phần vào dữ liệu thứ cấp và sự phân bố chưa đồng đều của các điểm mẫu cũng có thể ảnh hưởng đến mức độ đại diện của tập dữ liệu huấn luyện, qua đó tác động nhất định đến độ tin cậy của mô hình và khả năng tổng quát hóa kết quả cho các khu vực có điều kiện môi trường khác biệt. Vì vậy, mặc dù kết quả dự báo cho thấy mức độ phù hợp tương đối tốt, việc

tăng mật độ điểm đo thực địa và cải thiện thiết kế lấy mẫu trong các nghiên cứu tiếp theo sẽ góp phần nâng cao độ tin cậy và khả năng ứng dụng của các bản đồ nhiễm mặn đất.

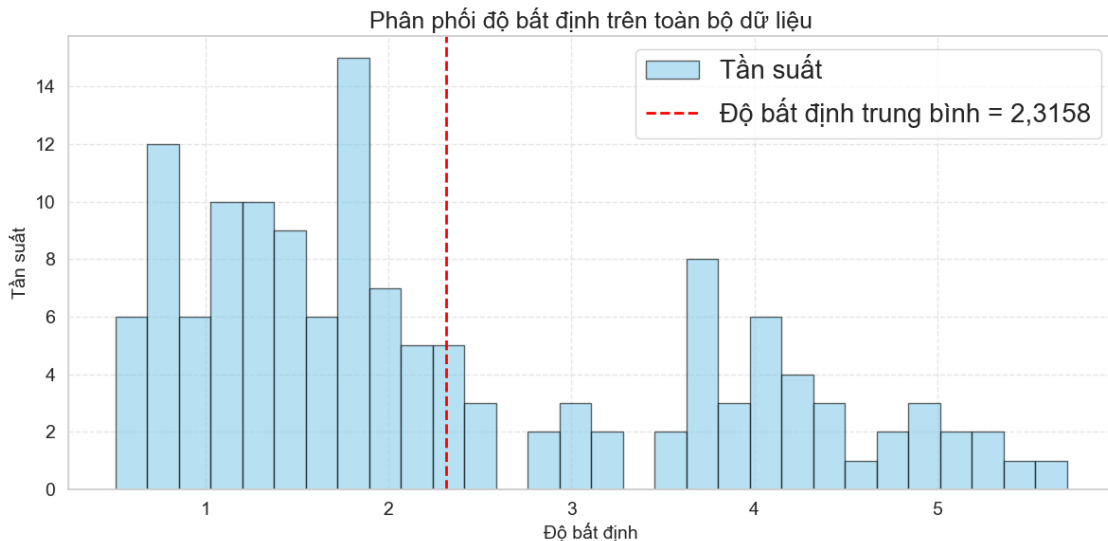
5.4. Độ nhạy và độ bất định của mô hình

Độ nhạy của mô hình trong nghiên cứu được đánh giá thông qua tầm quan trọng biến và phân tích SHAP, cho thấy Random Forest phản ứng mạnh với hệ số phản xạ bề mặt SR từ vệ tinh CYGNSS cùng các chỉ số viễn thám SI2, SI3, NDSI và NDVI (Hình 5). Biểu đồ SHAP ở Hình 13 cũng minh chứng rõ điều này khi các giá trị cao của SI3, NDVI, NDSI và SR tạo ra SHAP dương lớn, làm tăng mạnh giá trị dự đoán; trong khi các giá trị thấp của những biến này liên quan đến SHAP âm, kéo dự đoán giảm. Xu hướng này phản ánh sự nhạy cảm đặc trưng của tín hiệu băng tần L đối với sự thay đổi hằng số điện môi khi hàm lượng nước và muối hòa tan trong đất biến thiên. Ngược lại, các biến tĩnh như DEM và thuộc tính đất có mức ảnh hưởng thấp hơn, thể hiện qua phân bố SHAP chủ yếu tập trung quanh giá trị 0. Điều này hoàn toàn phù hợp với đặc điểm khu vực đồng bằng ven biển Bến Tre có địa hình bằng phẳng và thổ nhưỡng ít biến đổi, khiến

các biến này đóng vai trò nền hơn là yếu tố chi phối sự thay đổi độ mặn theo thời gian.



Hình 13. Biểu đồ SHAP.



Hình 14. Biểu đồ phân phối độ bất định.

Mặc dù mô hình đạt hiệu suất tổng quát hóa cao, độ bất định trong dự báo vẫn tồn tại do cả

dữ liệu đầu vào và thuật toán. Về dữ liệu, sự chênh lệch độ phân giải giữa ảnh vệ tinh thô như

CYGNSS (1 km) và MODIS (500 m) so với điểm đo thực địa dẫn đến hiện tượng điểm ảnh hỗn hợp, đặc biệt tại các khu vực có tính dị hướng cao như ranh giới giữa mặt nước và đất canh tác (Mục 5.2). Sự không tương thích không gian này làm tăng sai số dự báo và góp phần tạo nên độ bất định của mô hình. Về thuật toán, kiểm định chéo K-Fold cho thấy độ dao động phương sai giữa các vòng lặp, trong khi sự mất cân bằng phân phối mẫu khiến mô hình kém ổn định hơn khi dự báo các giá trị độ dẫn điện cực đoan. Độ bất định tổng hợp của mô hình được lượng hóa thông qua độ lệch chuẩn giữa các cây trong RF (Hình 14).

Phân bố độ lệch chuẩn cho thấy phần lớn các điểm dự báo tập trung trong khoảng từ khoảng 1,5 đến 3, phản ánh mức độ ổn định tương đối của mô hình trên phần lớn không gian nghiên cứu. Giá trị trung bình đạt khoảng 2,31 cho thấy mức bất định tổng thể ở mức trung bình. Tuy nhiên, vẫn tồn tại một nhóm điểm có độ lệch chuẩn cao hơn, đặc biệt vượt ngưỡng 4 và kéo dài đến trên 5, thể hiện mức độ thiếu chắc chắn đáng kể. Những khu vực này nhiều khả năng liên quan đến sự biến thiên mạnh của hệ số phản xạ bề mặt (SR) hoặc các điều kiện bề mặt phức tạp, khiến mô hình khó đưa ra dự báo nhất quán giữa các cây trong rừng ngẫu nhiên. Do đó, bản đồ nhiễm mặn cần được diễn giải như một phân bố giá trị kỳ vọng đi kèm với mức độ bất định theo không gian, thay vì một trường giá trị xác định tuyệt đối.

6. Kết luận

Nghiên cứu đã phát triển phương pháp giám sát nhiễm mặn đất tại tỉnh Bến Tre dựa trên tích hợp dữ liệu viễn thám đa nguồn, bao gồm dữ liệu quang học từ vệ tinh MODIS và dữ liệu phản xạ từ vệ tinh CYGNSS, kết hợp với các mô hình học máy cùng các thuật toán tăng cường độ dốc. Phương pháp được huấn luyện và kiểm định dựa trên 150 điểm đo thực địa thu thập đồng bộ với dữ liệu vệ tinh trong mùa khô năm 2020. Kết quả cho thấy mô hình RF đạt hiệu suất dự báo tốt nhất với hệ số tương quan đạt 0,71 và RMSE đạt 4,18 trên tập kiểm định. Bản đồ nhiễm mặn đất được

xây dựng từ mô hình này thể hiện rõ quy luật phân bố không gian với mức độ mặn tăng dần từ nội đồng ra khu vực ven biển, phù hợp với đặc điểm thủy văn và sử dụng đất của khu vực nghiên cứu.

Nghiên cứu đã chứng minh tiềm năng của dữ liệu phản xạ GNSS (CYGNSS) trong giám sát nhiễm mặn đất, khi biến phản xạ bề mặt từ CYGNSS được xác định là một trong những biến dự báo quan trọng nhất của mô hình. Kết quả này mở ra khả năng khai thác các nguồn dữ liệu GNSS-R trong nghiên cứu môi trường đất tại các khu vực nhiệt đới, nơi dữ liệu quang học thường bị hạn chế bởi mây. Các bản đồ nhiễm mặn được xây dựng có thể cung cấp thông tin hữu ích cho quản lý tài nguyên đất, hỗ trợ quy hoạch sử dụng đất và định hướng chuyển đổi cơ cấu sản xuất nông nghiệp tại các vùng chịu ảnh hưởng của xâm nhập mặn. Trong tương lai, việc tích hợp thêm dữ liệu radar có độ phân giải cao hơn và mở rộng tập dữ liệu thực địa sẽ giúp cải thiện độ chi tiết không gian của mô hình. Đồng thời, các phương pháp học sâu có thể được xem xét để mô phỏng tốt hơn biến động động lực của độ mặn theo mùa và theo chu kỳ thủy triều.

Tài liệu tham khảo

- [1] T. G. Nguyen, N. A. Tran, P. L. Vu, Q. H. Nguyen, H. D. Nguyen, Q. T. Bui, Salinity Intrusion Prediction Using Remote Sensing and Machine Learning in Data-Limited Regions: A Case Study in Vietnam's Mekong Delta, *Geoderma Regional*, Vol. 27, 2021, pp. e00424, <https://doi.org/10.1016/j.geodrs.2021.e00424>.
- [2] F. Khormali, M. Ajami, S. Ayoubi, C. Srinivasarao, S. P. Wani, Role of Deforestation and Hillslope Position on Soil Quality Attributes of Loess-Derived Soils in Golestan Province, Iran, *Agriculture, Ecosystems & Environment*, Vol. 134, No. 3, 2009, pp. 178-189, <https://doi.org/10.1016/j.agee.2009.06.017>.
- [3] B. Wicke et al., The Global Technical and Economic Potential of Bioenergy from Salt-Affected Soils, *Energy and Environmental Science*, Vol. 4, No. 8, 2011, pp. 2669-2681, <https://doi.org/10.1039/C1EE01029H>.
- [4] T. Gorji, E. Sertel, A. Tanik, Monitoring Soil Salinity via Remote Sensing Technology Under Data Scarce Conditions: A Case Study From

- Turkey, *Ecological Indicators*, Vol. 74, 2017, pp. 384-391, <https://doi.org/10.1016/j.ecolind.2016.11.043>.
- [5] Ministry of Agriculture and Rural Development, Summary Report on Drought and Salinity Intrusion in the Southern Region 2019-2020, <https://phongchongthientai.mard.gov.vn/Pages/ba-o-cao-tong-hop-tinh-hinh-han-han-xam-nhap-man-khu-vuc-mien-nam-2019--2020.aspx>, 2020 (accessed on: April 09, 2026) (in Vietnamese).
- [6] Ministry of Agriculture and Rural Development, Salinity Intrusion Causes Ben Tre to Lose Thousands of Billions of VND Every Year, <https://phongchongthientai.mard.gov.vn/Pages/xam-nhap-man-khien-ben-tre-thiet-hai-hang-ngan-ti-dong-moi-nam.aspx>, 2022 (accessed on: April 09, 2026) (in Vietnamese).
- [7] T. Gorji, A. Tanik, E. Sertel, Soil Salinity Prediction, Monitoring and Mapping Using Modern Technologies, *Procedia Earth and Planetary Science*, Vol. 15, 2015, pp. 507-512, <https://doi.org/10.1016/j.proeps.2015.08.062>.
- [8] V. L. Mulder, S. De Bruin, M. E. Schaepman, T. R. Mayr, The Use of Remote Sensing in Soil and Terrain Mapping - A Review, *Geoderma*, Vol. 162, No. 1, 2011, pp. 1-19, <https://doi.org/10.1016/j.geoderma.2010.12.018>.
- [9] H. D. Nguyen, D. K. Dang, Q. T. Bui, Estuary Salinity Prediction Using Machine Learning: Case Study in the Hau Estuary in Mekong River, Vietnam, *Water Supply*, Vol. 25, No. 2, 2025, pp. 327-342, <https://doi.org/10.2166/ws.2025.007>.
- [10] A. A. A. Aldabaa, D. C. Weindorf, S. Chakraborty, A. Sharma, B. Li, Combination of Proximal and Remote Sensing Methods for Rapid Soil Salinity Quantification, *Geoderma*, Vol. 239-240, 2015, pp. 34-46, <https://doi.org/10.1016/j.geoderma.2014.09.011>.
- [11] J. Ding, D. Yu, Monitoring and Evaluating Spatial Variability of Soil Salinity in Dry and Wet Seasons in the Werigan-Kuqa Oasis, China, Using Remote Sensing and Electromagnetic Induction Instruments, *Geoderma*, Vol. 235-236, 2014, pp. 316-322, <https://doi.org/10.1016/j.geoderma.2014.07.028>.
- [12] Y. Liu, et al., Estimating the Soil Salinity Over Partially Vegetated Surfaces From Multispectral Remote Sensing Image Using Non-Negative Matrix Factorization, *Geoderma*, Vol. 354, 2019, pp. 113887, <https://doi.org/10.1016/j.geoderma.2019.113887>.
- [13] S. H. Yueh, R. Shah, M. J. Chaubell, A. Hayashi, X. Xu, A. Colliander, A Semiempirical Modeling of Soil Moisture, Vegetation, and Surface Roughness Impact on CYGNSS Reflectometry Data, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, 2022, pp. 1-17, <https://doi.org/10.1109/TGRS.2020.3035989>.
- [14] H. Kim, V. Lakshmi, Use of Cyclone Global Navigation Satellite System (CyGNSS) Observations for Estimation of Soil Moisture, *Geophysical Research Letters*, Vol. 45, No. 16, 2018, pp. 8272-8282, <https://doi.org/10.1029/2018GL078923>.
- [15] J. Wang et al., A Novel Retrieval Model for Soil Salinity From CYGNSS: Algorithm and Test in the Yellow River Delta, *Geoderma*, Vol. 432, 2023, pp. 116417, <https://doi.org/10.1016/j.geoderma.2023.116417>.
- [16] C. Chew, E. Small, Description of the UCAR/CU Soil Moisture Product, *Remote Sensing*, Vol. 12, No. 10, 2020, <https://doi.org/10.3390/rs12101558>.
- [17] C. Chew, Soil Moisture Sensing Using Spaceborne GNSS Reflections: Comparison of CYGNSS Reflectivity to SMAP Soil Moisture, *Geophysical Research Letters*, 2018, <https://doi.org/10.1029/2018GL077905>.
- [18] P. Ghasemigoudarzi, W. Huang, O. De Silva, Q. Yan, D. T. Power, Flash Flood Detection From CYGNSS Data Using the RUSBoost Algorithm, *IEEE Access*, Vol. 8, 2020, pp. 171864-171881, <https://doi.org/10.1109/ACCESS.2020.3025302>.
- [19] W. Yang, F. Guo, X. Zhang, Y. Zhu, Z. Li, Z. Zhang, First Quasi-Global Soil Moisture Retrieval Using Fengyun-3 GNSS-R Constellation Observations, *Remote Sensing of Environment*, Vol. 321, 2025, pp. 114653, <https://doi.org/10.1016/j.rse.2025.114653>.
- [20] Y. Zhu, F. Guo, X. Zhang, Spaceborne GNSS-R Soil Moisture Retrieval From GPS/BDS-3/Galileo Satellites, *GPS Solutions*, Vol. 29, No. 1, 2024, pp. 10, <https://doi.org/10.1007/s10291-024-01767-8>.
- [21] T. T. Zhang, J. G. Qi, Y. Gao, Z. T. Ouyang, S. L. Zeng, B. Zhao, Detecting Soil Salinity With MODIS Time Series VI Data, *Ecological Indicators*, Vol. 52, 2015, pp. 480-489, <https://doi.org/10.1016/j.ecolind.2015.01.004>.
- [22] L. Breiman, Random Forests, *Machine Learning*, Vol. 45, No. 1, 2001, pp. 5-32, <https://doi.org/10.1023/A:1010933404324>.
- [23] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794, <https://doi.org/10.1145/2939672.2939785>.

- [24] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, CatBoost: Unbiased Boosting With Categorical Features, *Advances in Neural Information Processing Systems*, Vol. 31, 2018, <https://doi.org/10.48550/arXiv.1706.09516>.
- [25] O. Eroglu, M. Kurum, D. Boyd, A. C. Gurbuz, High Spatio-Temporal Resolution CYGNSS Soil Moisture Estimates Using Artificial Neural Networks, *Remote Sensing*, Vol. 11, No. 19, 2019, <https://doi.org/10.3390/rs11192272>.
- [26] E. Vermote, R. Wolfe, MOD09GQ MODIS/Terra Surface Reflectance Daily L2G Global 250m SIN Grid V006, NASA EOSDIS Land Processes Distributed Active Archive Center (DAAC) Data Set, 2015, pp. MOD09GQ.006, <https://doi.org/10.5067/MODIS/MOD09GQ.006>.
- [27] H. D. Nguyen et al., Soil Salinity Prediction Using Hybrid Machine Learning and Remote Sensing in Ben Tre Province on Vietnam's Mekong River Delta, *Environmental Science and Pollution Research*, Vol. 30, No. 29, 2023, pp. 74340-74357, <https://doi.org/10.1007/s11356-023-27516-x>.
- [28] A. Calabia, I. Molina, S. Jin, Soil Moisture Content From GNSS Reflectometry Using Dielectric Permittivity From Fresnel Reflection Coefficients, *Remote Sensing*, Vol. 12, No. 1, 2020, <https://doi.org/10.3390/rs12010122>.
- [29] Y. U. Haq, M. Shahbaz, S. Asif, K. Ouahada, H. Hamam, Identification of Soil Types and Salinity Using MODIS Terra Data and Machine Learning Techniques in Multiple Regions of Pakistan, *Sensors*, Vol. 23, No. 19, 2023, <https://doi.org/10.3390/s23198121>.
- [30] J. Friedman, T. Hastie, R. Tibshirani, Additive Logistic Regression: A Statistical View of Boosting (With Discussion and a Rejoinder by the Authors), *The Annals of Statistics*, Vol. 28, No. 2, 2000, pp. 337-407, <https://doi.org/10.1214/aos/1016218223>.
- [31] J. H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, *The Annals of Statistics*, Vol. 29, No. 5, 2001, pp. 1189-1232, <https://doi.org/10.1214/aos/1013203451>.
- [32] A. Zarei, M. Hasanlou, M. Mahdianpari, A Comparison of Machine Learning Models for Soil Salinity Estimation Using Multi-Spectral Earth Observation Data, *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. V-3, 2021, pp. 257-263, <https://doi.org/10.5194/isprs-annals-V-3-2021-257-2021>.
- [33] S. Aksoy, E. Sertel, R. Roscher, A. Tanik, N. Hamzehpour, Assessment of Soil Salinity Using Explainable Machine Learning Methods and Landsat 8 Images, *International Journal of Applied Earth Observation and Geoinformation*, Vol. 130, 2024, pp. 103879, <https://doi.org/10.1016/j.jag.2024.103879>.
- [34] A. Samat, E. Li, W. Wang, S. Liu, C. Lin, J. Abuduwaili, Meta-XGBoost for Hyperspectral Image Classification Using Extended MSER-Guided Morphological Profiles, *Remote Sensing*, Vol. 12, No. 12, 2020, <https://doi.org/10.3390/rs12121973>.
- [35] J. Bergstra, R. Bardenet, Y. Bengio, B. Kégl, Algorithms for Hyper-Parameter Optimization, *Advances in Neural Information Processing Systems*, Vol. 24, 2011.
- [36] Z. Shao, M. N. Ahmad, A. Javed, Comparison of Random Forest and XGBoost Classifiers Using Integrated Optical and SAR Features for Mapping Urban Impervious Surface, *Remote Sensing*, Vol. 16, No. 4, 2024, <https://doi.org/10.3390/rs16040665>.
- [37] Y. Liu, S. Zhang, J. Zhang, L. Tang, Y. Bai, Assessment and Comparison of Six Machine Learning Models in Estimating Evapotranspiration Over Croplands Using Remote Sensing and Meteorological Factors, *Remote Sensing*, Vol. 13, No. 19, 2021, <https://doi.org/10.3390/rs13193838>.
- [38] E. K. Sahin, Assessing the Predictive Capability of Ensemble Tree Methods for Landslide Susceptibility Mapping Using XGBoost, Gradient Boosting Machine, and Random Forest, *SN Applied Sciences*, Vol. 2, 2020, pp. 1308, <https://doi.org/10.1007/s42452-020-3060-1>.