

**ĐẠI HỌC QUỐC GIA HÀ NỘI  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**



**Phạm Minh Quang**

**NGHIÊN CỨU ỨNG DỤNG DỮ LIỆU VỆ TINH  
CYCLONE GNSS TRONG GIÁM SÁT XÂM NHẬP  
MẶN**

**ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY**

**Ngành: Công nghệ hàng không vũ trụ**

**HÀ NỘI - 2025**

ĐẠI HỌC QUỐC GIA HÀ NỘI  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Phạm Minh Quang

NGHIÊN CỨU ÚNG DỤNG DỮ LIỆU VỆ TINH  
CYCLONE GNSS TRONG GIÁM SÁT XÂM NHẬP  
MẶN

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ hàng không vũ trụ

Cán bộ hướng dẫn: TS. Hà Minh Cường

Cán bộ đồng hướng dẫn: ThS. Hoàng Tích Phúc

Hà Nội - 2025

## TÓM TẮT

Xâm nhập mặn đất đang trở thành một trong những thách thức môi trường nghiêm trọng nhất tại Đồng bằng sông Cửu Long – khu vực giữ vai trò trọng yếu trong sản xuất lúa gạo, cây ăn trái và thủy sản của Việt Nam. Dưới tác động của biến đổi khí hậu, nước biển dâng và sự suy giảm nguồn nước ngọt từ thượng nguồn, hiện tượng mặn xâm nhập ngày càng diễn biến phức tạp, lan sâu và kéo dài theo mùa, gây suy giảm chất lượng đất, giảm năng suất cây trồng và tạo ra áp lực lớn đối với chiến lược quản lý tài nguyên đất – nước của toàn vùng. Trong bối cảnh các phương pháp đo đặc truyền thống còn hạn chế về không gian và thời gian, yêu cầu về một hệ thống giám sát độ mặn đất có chi phí thấp, cập nhật thường xuyên và có khả năng áp dụng trên diện rộng trở nên đặc biệt cấp thiết. Nghiên cứu này phát triển một quy trình tích hợp dữ liệu viễn thám với mô hình học máy, sử dụng các chỉ số phản xạ từ ảnh MODIS, dữ liệu tần xạ vi sóng của CYGNSS cùng thông tin địa hình, thổ nhưỡng và khí tượng để ước lượng độ mặn đất ở quy mô toàn Đồng bằng sông Cửu Long. Ba mô hình học máy gồm CatBoost, Random Forest và XGBoost được triển khai để phân tích, trong đó mô hình có hiệu suất tốt nhất được sử dụng để tạo ra bộ bản đồ độ mặn đất theo tháng, phục vụ giám sát liên tục, nhận diện các biến động không gian – thời gian và hỗ trợ theo dõi quá trình mặn hóa trong các giai đoạn quan trọng của mùa khô. Việc ứng dụng đồng thời dữ liệu quang học và radar thụ động giúp mô hình phản ánh chính xác hơn các đặc trưng bề mặt như độ ẩm, mức độ phản xạ, tác động thủy triều và các biến động theo mùa vụ. Phương pháp này không chỉ bổ sung nguồn thông tin quan trọng cho công tác giám sát mặn đất, mà còn mở ra hướng tiếp cận khả thi để xây dựng các bản đồ, chuỗi thời gian và hệ thống cảnh báo phục vụ quản lý nông nghiệp trong bối cảnh biến đổi khí hậu ngày càng gia tăng tác động lên Đồng bằng sông Cửu Long.

**Từ khóa:** viễn thám, xâm nhập mặn, CYGNSS.

## **LỜI CAM ĐOAN**

Tôi xin cam đoan rằng đồ án tốt nghiệp “Nghiên cứu ứng dụng dữ liệu vệ tinh Cyclone GNSS trong giám sát xâm nhập mặn” được thực hiện dưới sự hướng dẫn của TS. Hà Minh Cường và ThS. Hoàng Tích Phúc là kết quả nghiên cứu độc lập của cá nhân tôi.

Các số liệu, kết quả tính toán và nội dung trình bày trong đồ án đều trung thực và không sử dụng, sao chép trái phép các tài liệu hay công trình nghiên cứu của người khác. Tôi hoàn toàn chịu trách nhiệm trước nhà trường về tính trung thực của nội dung trong đồ án. Nếu có bất kỳ sự vi phạm nào liên quan đến việc sao chép hoặc sử dụng trái phép kết quả nghiên cứu của người khác, tôi xin chịu mọi hình thức xử lý theo quy định.

Sinh viên

Phạm Minh Quang

## LỜI CẢM ƠN

Trong suốt quá trình thực hiện đồ án tốt nghiệp này, em đã nhận được sự hướng dẫn, giúp đỡ và động viên quý báu từ thầy cô, gia đình và bạn bè. Em xin được bày tỏ lòng biết ơn chân thành đến tất cả.

Trước hết, em xin gửi lời cảm ơn sâu sắc đến TS. Hà Minh Cường, giảng viên Viện Công nghệ Hàng không Vũ trụ, Trường Đại học Công nghệ, ĐHQGHN, người đã tận tình chỉ bảo và hướng dẫn em trong suốt thời gian thực hiện đồ án. Những ý kiến đóng góp quý báu của thầy đã giúp em hoàn thiện và phát triển các ý tưởng nghiên cứu của mình. Em cũng xin gửi lời cảm ơn sâu sắc đến ThS. Hoàng Tích Phúc, người đã đồng hành, hỗ trợ và hướng dẫn em trong từng bước nghiên cứu, góp phần quan trọng vào việc hoàn thiện đồ án này.

Em cũng xin chân thành cảm ơn các thầy cô giáo của Trường Đại học Công nghệ, ĐHQGHN nói chung, và các thầy cô trong Viện Công nghệ Hàng không Vũ trụ nói riêng. Chính những kiến thức cơ bản và chuyên sâu được truyền đạt trong quá trình học tập đã cung cấp cho em nền tảng vững chắc để thực hiện đồ án này.

Cuối cùng, em xin cảm ơn gia đình và bạn bè, những người luôn sát cánh bên em, động viên và ủng hộ em vượt qua mọi khó khăn. Sự quan tâm và chia sẻ của mọi người là nguồn động lực to lớn giúp em hoàn thành đồ án tốt nghiệp này.

Với điều kiện thời gian và kinh nghiệm còn hạn chế, đồ án này khó tránh khỏi những thiếu sót. Em mong nhận được những ý kiến đóng góp và sự chỉ dẫn từ các thầy cô để có thể tiếp tục hoàn thiện bản thân, phục vụ tốt hơn cho công việc thực tế sau này.

Em xin trân trọng cảm ơn!

## MỤC LỤC

TÓM TẮT .....	i
LỜI CAM ĐOAN .....	ii
LỜI CẢM ƠN .....	iii
MỤC LỤC .....	iv
DANH MỤC TỪ VIẾT TẮT .....	vii
DANH MỤC HÌNH ẢNH .....	ix
DANH MỤC BẢNG .....	x
Mở đầu .....	xii
1. Đặt vấn đề .....	xii
2. Mục tiêu và nội dung nghiên cứu. ....	xiii
3. Đối tượng và phạm vi nghiên cứu .....	xiii
4. Quan điểm và phương pháp nghiên cứu .....	xiii
5. Ý nghĩa khoa học và thực tiễn của đề tài.....	xiii
6. Cấu trúc của đồ án .....	xiv
CHƯƠNG 1: Tổng quan tài liệu và cơ sở lý luận về ứng dụng viễn thám, GIS và học máy trong nghiên cứu xâm nhập mặn.....	1
1.1. Khái quát về viễn thám .....	1
1.2. Khái quát về hệ thống thông tin địa lý.....	2
1.2.1. Khái niệm và công dụng của hệ thống thông tin địa lý .....	2
1.2.2. Vai trò của hệ thống thông tin địa lý trong chủ đề nghiên cứu .....	3
1.3. Khái quát về xâm nhập mặn .....	4
1.4. Khái quát về học máy và các phương pháp học máy .....	5
1.5. Công nghệ đo phản xạ tín hiệu của hệ thống vệ tinh định vị toàn cầu .....	7
1.5.1. Giới thiệu chung về công nghệ phản xạ tín hiệu GNSS-R .....	7
1.5.2. Nguyên lý hoạt động và cơ sở vật lý .....	7
1.5.3. Hệ thống vệ tinh Cyclone GNSS .....	10

1.5.4. Các kỹ thuật đo đạc và tham số quan trọng .....	10
1.6. Khái quát về bản đồ giám sát.....	13
1.7. Các phương pháp nghiên cứu đánh giá xâm nhập mặn.....	13
1.7.1. Trên thế giới.....	13
1.7.2. Tại Việt Nam.....	15
CHƯƠNG 2: Phương pháp nghiên cứu.....	18
2.1. Sơ đồ quy trình nghiên cứu.....	18
2.2. Các phương pháp học máy trong thành lập bản đồ giám sát xâm nhập mặn .....	19
2.2.1. Thuật toán XGBoost .....	19
2.2.2. Thuật toán Random Forest.....	24
2.2.3. Thuật toán CatBoost .....	26
2.3. Ước tính hệ số phản xạ bì mặt từ dữ liệu Cyclone GNSS .....	29
2.4. Các phương pháp đánh giá mô hình .....	30
2.4.1. Kiểm định chéo .....	30
2.4.2. Sai số tuyệt đối trung bình (MAE) .....	31
2.4.3. Sai số bình phương trung bình (RMSE) .....	31
2.4.4. Hệ số tương quan R .....	32
CHƯƠNG 3: Kết quả thực nghiệm ứng dụng viễn thám, GIS và học máy để thành lập bản đồ xâm nhập tại Đồng Bằng Sông Cửu Long .....	33
3.1. Đặc điểm tự nhiên và kinh tế - xã hội tại Đồng Bằng Sông Cửu Long....	33
3.1.1. Đặc điểm tự nhiên.....	33
3.1.2. Đặc điểm kinh tế - xã hội.....	35
3.2. Cơ sở dữ liệu.....	36
3.2.1. Dữ liệu vệ tinh CYGNSS .....	37
3.2.2. Chỉ số viễn thám .....	38
3.2.3. Dữ liệu địa hình .....	42
3.2.4. Dữ liệu khí tượng.....	43

3.2.5. Dữ liệu thô nhưỡng .....	44
3.2.6. Dữ liệu thực tế .....	45
3.3. Đánh giá và so sánh hiệu suất các mô hình .....	49
3.4. Đánh giá xâm nhập mặn tại Đồng Bằng Sông Cửu Long .....	54
CHƯƠNG 4: KẾT LUẬN VÀ ĐỀ XUẤT .....	67
TÀI LIỆU THAM KHẢO .....	69

## DANH MỤC TỪ VIẾT TẮT

Tiếng Anh	Viết tắt	Tiếng Việt
Cyclone Global Navigation Satellite System	ĐBSCL	Đồng bằng sông Cửu Long
Global Navigation Satellite System	CYGNSS	Hệ thống vệ tinh định vị toàn cầu Cyclone
Global Navigation Satellite System Reflectometry	GNSS	Hệ thống vệ tinh định vị toàn cầu
Synthetic Aperture Radar	GNSS-R	Kỹ thuật viễn thám phản xạ tín hiệu GNSS
Geographic Information System	SAR	Radar khẩu độ tổng hợp
Moderate Resolution Imaging Spectroradiometer	GIS	Hệ thống thông tin địa lý
Digital Elevation Model	MODIS	Máy đo quang phổ hình ảnh độ phân giải vừa phải
Normalized Difference Vegetation Index	DEM	Mô hình số độ cao
Normalized Difference Salinity Index	NDVI	Chỉ số thực vật khác biệt chuẩn hóa
Salinity Index	NDSI	Chỉ số mặn khác biệt chuẩn hóa
Electrical Conductivity	SI (SI1 - SI5)	Các chỉ số mặn chuyên biệt
Short-Wave Infrared	EC	Độ dẫn điện
Near Infrared	SWIR	Hồng ngoại sóng ngắn
Delay Doppler Map	NIR	Hồng ngoại gần
Random Forest	DDM	Bản đồ trễ Doppler
XGBoost	RF	Rừng ngẫu nhiên
CatBoost	XGB	
Root Mean Square Error	CB	
Mean Absolute Error	RMSE	Căn sai số bình phương trung bình
Soil Moisture Active Passive	MAE	Sai số tuyệt đối trung bình
Shuttle Radar Topography Mission	SMAP	Vệ tinh SMAP thu thập dữ liệu độ ẩm đất bề mặt
	SRTM	Nhiệm vụ SRTM cung cấp dữ liệu mô hình số độ cao

Physical Oceanography Distributed Active Archive Center	PODAAC	Trung tâm lưu trữ và phân phối dữ liệu hải dương
Support Vector Regression	SVR	Phương pháp hồi quy máy vector hỗ trợ
Artificial Neural Network	ANN	Mạng nơ-ron nhân tạo
Classification And Regression Trees	CART	Mô hình cây phân loại và hồi quy
Out-of-Bag	OOB	Kiểm định ngoài túi
Right Hand Circular Polarization	RHCP	Phân cực tròn tay phải
Left Hand Circular Polarization	LHCP	Phân cực tròn tay trái
Effective Isotropic Radiated Power	EIRP	Công suất bức xạ đẳng hướng hiệu dụng
Delay Doppler Map Instrument	DDMI	Thiết bị tạo bản đồ đồ trẽ

## DANH MỤC HÌNH ẢNH

Hình 1: Sơ đồ hình học của hệ thống phản xạ tín hiệu vệ tinh, minh họa đường truyền tín hiệu trực tiếp từ vệ tinh phát và tín hiệu phản xạ từ bề mặt Trái đất đến vệ tinh thu quỹ đạo tầm thấp.....	8
Hình 2: Các dạng sóng phản xạ tương ứng với các điều kiện bề mặt khác nhau, từ phản xạ gương lý tưởng trên bề mặt phẳng ở hình a đến tán xạ khuếch tán trên bề mặt gồ ghề ở hình b và c .....	9
Hình 3: Bản đồ trẽ Doppler và các biến dạng công suất tương ứng, với các đường đẳng tần số Doppler và các đường đẳng trẽ hình elip xoay quanh điểm phản xạ gương .....	11
Hình 4: Biểu đồ minh họa công suất tỷ số tín hiệu trên nhiễu chuẩn hóa theo thời gian, cho thấy phần dao động do tín hiệu phản xạ gây ra nằm chồng lên tín hiệu trực tiếp ..	12
Hình 5: Sơ đồ quy trình nghiên cứu .....	18
Hình 6: Khu vực nghiên cứu và phân bố của điểm 330 điểm đo mặn tại khu vực Đồng Bằng Sông Cửu Long .....	37
Hình 7: Hệ số phản xạ bề mặt từ dữ liệu CYGNSS trung bình tháng 1 đến tháng 5 năm 2025 .....	38
Hình 8: Bộ dữ liệu chỉ số viễn thám dưới dạng bản đồ:a) Chỉ số thực vật NDVI; b) Chỉ số mặn NDSI; c)-g) Chỉ số mặn SI1-SI5; h) Băng phổ SWIR1; i) Băng phổ SWIR2..	42
Hình 9: Bộ dữ liệu địa hình; a) Độ cao; b) Khoảng cách đến biển .....	42
Hình 10: Bộ dữ liệu khí tượng được trình bày dưới dạng bản đồ; a) Độ ẩm đất; b) Nhiệt độ bề mặt .....	43
Hình 11: Bộ dữ liệu thổ nhưỡng dưới dạng bản đồ: a) Lớp phủ bề mặt; b) Hàm lượng cát; c) Hàm lượng sét; d) Khối lượng riêng đất.....	45
Hình 12: Giá trị độ mặn tại các trạm đo .....	47
Hình 13: Biểu đồ đánh giá chỉ số tầm quan trọng của các biến trong 18 yếu tố.....	53
Hình 14: Bản đồ xâm nhập mặn được xây dựng từ mô hình RF; a)Tháng 1; b)Tháng 2; c)Tháng3; d)Tháng 4; e)Tháng 5 năm 2025.....	56
Hình 15: Bản đồ xâm nhập mặn được xây dựng từ mô hình XGB; a)Tháng 1; b)Tháng 2; c)Tháng3; d)Tháng 4; e)Tháng 5 năm 2025.....	58
Hình 16: Bản đồ xâm nhập mặn được xây dựng từ mô hình CB; a)Tháng 1; b)Tháng 2; c)Tháng3; d)Tháng 4; e)Tháng 5 năm 2025.....	60
Hình 17: Bản đồ mức độ xâm nhập mặn được xây dựng từ mô hình XGB; a)Tháng 1; b)Tháng 2; c)Tháng3; d)Tháng 4; e)Tháng 5 năm 2025 .....	64

## **DANH MỤC BẢNG**

Bảng 1: Các tham số dữ liệu Cyclone GNSS .....	29
Bảng 2: Dữ liệu sử dụng trong nghiên cứu .....	36
Bảng 3: Giá trị đo mặn tại các trạm.....	46
Bảng 4: Tham số được sử dụng trong ba mô hình học máy.....	50
Bảng 5: Số liệu đánh giá hiệu suất các mô hình.....	51
Bảng 6: Bảng thống kê diện tích xâm nhập mặn.....	65

## Mở đầu

### 1. Đặt vấn đề

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 tác động của biến đổi khí hậu [1]. Tuy nhiên, quá trình thâm canh nông nghiệp, áp lực từ các hoạt động nhân sinh và sự biến đổi khí hậu đã làm gia tăng đáng kể mức độ suy thoái đất trên toàn cầu [2]. Trong số các dạng suy thoái như xói mòn, ô nhiễm hay sa mạc hóa, xâm nhập mặn được xem là một trong những vấn đề nghiêm trọng nhất, đe dọa trực tiếp đến 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 riêng khu vực châu Á chiếm tới 69% [3]. Mỗi năm, tình trạng mặn hóa đất làm suy giảm năng suất trên khoảng 1,5 triệu ha đất canh tác, gây tổn thất gần 10% sản lượng lương thực toàn cầu [4].

Tại Việt Nam, xâm nhập mặn đã và đang là vấn đề môi trường đặc biệt nghiêm trọng tại vùng Đồng bằng sông Cửu Long – khu vực được xem là trung tâm sản xuất lúa, trái cây và thủy sản lớn nhất cả nước. Theo Bộ Tài nguyên và Môi trường, đợt xâm nhập mặn mùa khô 2019–2020 tại ĐBSCL được đánh giá là nghiêm trọng nhất trong lịch sử quan trắc, ảnh hưởng đến 10/13 tỉnh trong vùng. Ranh giới mặn 4 g/l đã xâm nhập tới 42,5% diện tích tự nhiên toàn vùng (tương đương 1.688.600 ha), cao hơn năm 2016 khoảng 50.376 ha. Đối với cây ăn trái, hạn và mặn đã khiến 6.650 ha tại các tỉnh Long An, Tiền Giang, Vĩnh Long, Trà Vinh, Sóc Trăng... thiêu nước tưới, trong đó 355 ha bị thiệt hại hoàn toàn. Ngoài ra, diện tích hoa màu, cây giống và cả diện tích nuôi trồng thủy sản cũng chịu ảnh hưởng nặng nề, khiến thiệt hại kinh tế lên tới hàng nghìn tỷ đồng và tác động trực tiếp tới sinh kế của hàng chục nghìn hộ dân trong vùng ven sông, ven biển [5].

Các phương pháp truyền thống dựa vào đo đạc tại chỗ, điển hình là đo điện dẫn suất (EC) [6], tuy có độ chính xác cao nhưng lại hạn chế về phạm vi không gian, khó mở rộng và đòi hỏi chi phí lớn. Trong khi đó, công nghệ viễn thám đã chứng minh được hiệu quả vượt trội trong giám sát và lập bản đồ các hiện tượng môi trường trên diện rộng, đặc biệt khi kết hợp cùng các mô hình học máy. Dữ liệu quang học như MODIS và Landsat cho phép trích xuất các chỉ số thực vật và chỉ số mặn có khả năng phản ánh gián tiếp tình trạng xâm nhập mặn thông qua sự biến đổi sinh lý của thảm thực vật [7], [8]. Tuy nhiên, các nguồn dữ liệu quang học này chịu ảnh hưởng mạnh từ điều kiện thời

tiết và chiêu sáng, dẫn tới hiện tượng thiếu dữ liệu vào những thời điểm có mây dày hoặc mưa kéo dài – vốn rất phổ biến trong mùa khô và giai đoạn chuyển mùa tại DBSCL.

Trong bối cảnh đó, dữ liệu GNSS-R từ hệ thống vệ tinh CYGNSS mở ra một hướng tiếp cận mới nhờ khả năng quan sát tín hiệu GPS phản xạ ở băng tần L. CYGNSS thu nhận thông tin phản xạ từ bề mặt đất, vốn nhạy cảm với những thay đổi về độ nhám bề mặt, độ ẩm và khả năng dẫn điện của lớp đất trên cùng – các yếu tố có mối liên hệ trực tiếp với quá trình xâm nhập mặn. So với dữ liệu quang học, GNSS-R có ưu thế lớn nhờ khả năng quan sát xuyên mây, không phụ thuộc vào ánh sáng và duy trì được tần suất quan trắc liên tục ngay cả trong điều kiện thời tiết phức tạp. Điều này đặc biệt phù hợp với vùng ven biển nhiệt đới như Đồng bằng sông Cửu Long, nơi mây mù, hơi nước và khí hậu ẩm đặc trưng thường cản trở việc thu nhận ảnh quang học. Mặc dù CYGNSS đã được ứng dụng trong các nghiên cứu giám sát độ ẩm đất, theo dõi lũ lụt và phân tích đặc tính bề mặt, nhưng việc khai thác dữ liệu này để lập bản đồ và giám sát xâm nhập mặn tại Việt Nam vẫn còn rất hạn chế. Khoảng trống này đặt ra nhu cầu cần thiết cho việc tận dụng nguồn dữ liệu GNSS-R nhằm nâng cao khả năng quan trắc không gian – thời gian của quá trình xâm nhập mặn trong khu vực.

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 mặn và chỉ số thực vật từ dữ liệu quang học để thiết lập mô hình giám sát xâm nhập mặn theo tháng cho vùng Đồng bằng sông Cửu Long. Bộ dữ liệu tích hợp được phân tích bằng các thuật toán học máy hiện đại như Random Forest, XGBoost và CatBoost nhằm mô hình hóa các quan hệ phi tuyến giữa dữ liệu viễn thám và độ mặn thực đo. Cách tiếp cận này không chỉ giúp tăng độ chính xác dự báo mà còn mở rộng khả năng giám sát theo thời gian, đáp ứng nhu cầu theo dõi liên tục diễn biến mặn hóa đất nhằm phục vụ quản lý tài nguyên nước, quy hoạch mùa vụ và thích ứng hiệu quả với biến đổi khí hậu tại khu vực Đồng bằng sông Cửu Long.

## 2. Mục tiêu và nội dung nghiên cứu.

Mục tiêu của nghiên cứu là phát triển một phương pháp tích hợp mới dựa trên dữ liệu phản xạ GNSS-R, các chỉ số mặn và thực vật trích xuất từ ảnh viễn thám quang học kết hợp với công nghệ GIS, đồng thời ứng dụng các thuật toán học máy để xây dựng công cụ giám sát xâm nhập mặn theo tháng cho khu vực Đồng bằng sông Cửu Long trong mùa khô năm 2025. Để đạt được mục đích này, nghiên cứu tập trung thực hiện: (1) Xây dựng cơ sở dữ liệu đồng bộ đa nguồn bằng cách tích hợp dữ liệu phản xạ bề mặt từ vệ tinh CYGNSS, các chỉ số thực vật và độ mặn từ ảnh quang học cùng dữ liệu địa hình, khí tượng, thổ nhưỡng làm đầu vào cho các mô hình; (2) Đánh giá và lựa chọn mô

hình tối ưu thông qua việc huấn luyện, so sánh hiệu suất của ba thuật toán Random Forest, XGBoost và CatBoost; (3) Thành lập bản đồ xâm nhập mặn theo chuỗi thời gian từ tháng 1 đến tháng 5 năm 2025 và ứng dụng mô hình tốt nhất để xây dựng bản đồ phân bố mức độ xâm nhập mặn, nhằm phân tích diễn biến và phân vùng rủi ro cho khu vực khảo sát.

### 3. Đối tượng và phạm vi nghiên cứu

Trong học máy, các thuật toán cơ bản được sử dụng trong nghiên cứu bao gồm: Rừng ngẫu nhiên, XGBoost và CatBoost

Phản viễn thám và GIS tập trung vào việc xác định mức độ xâm nhập mặn từ dữ liệu phản xạ GNSS-R và ảnh viễn thám quang học thông qua nền tảng Google Earth Engine (GEE), đồng thời thu thập và tích hợp dữ liệu đo mặn thực địa gồm 330 điểm khảo sát cùng số liệu từ 7 trạm quan trắc mặn, kết hợp với các lớp dữ liệu khí tượng, địa hình và thổ nhưỡng theo từng tháng nhằm phục vụ quá trình mô hình hóa và đánh giá diễn biến xâm nhập mặn tại Đồng bằng sông Cửu Long.

### 4. Quan điểm và phương pháp nghiên cứu

Đồ án được thực hiện dựa trên cơ sở lý luận về công nghệ viễn thám phản xạ GNSS-R và thực tiễn biến đổi khí hậu, thông qua việc thu thập và tổng hợp đồng bộ các nguồn dữ liệu vệ tinh đa nguồn cùng số liệu quan trắc thực địa. Tiếp theo, nghiên cứu kế thừa và tham khảo các phương pháp, kết quả từ những công trình trước đó của các tác giả trong và ngoài nước về ứng dụng học máy trong giám sát môi trường. Cuối cùng, các mô hình học máy được hiệu chỉnh tham số và áp dụng thực nghiệm cho bài toán cụ thể nhằm xây dựng bộ bản đồ giám sát xâm nhập mặn theo chuỗi thời gian, phục vụ phân tích và đánh giá diễn biến mặn hóa tại khu vực Đồng bằng sông Cửu Long.

### 5. Ý nghĩa khoa học và thực tiễn của đề tài

Đề tài “Nghiên cứu ứng dụng dữ liệu vệ tinh Cyclone GNSS trong giám sát xâm nhập mặn” được thực hiện với mong muốn kết quả nghiên cứu sẽ trở thành cơ sở khoa học và tài liệu tham khảo quan trọng phục vụ công tác quan trắc, dự báo và theo dõi diễn biến xâm nhập mặn tại Đồng bằng sông Cửu Long. Bản đồ giám sát xâm nhập mặn theo tháng là nguồn thông tin có giá trị nhằm xác định các khu vực chịu ảnh hưởng mạnh nhất, dựa trên đặc điểm bề mặt và điều kiện tự nhiên thể hiện xu hướng mặn hóa của khu vực. Từ đó, kết quả nghiên cứu góp phần hỗ trợ việc xây dựng các giải pháp quản lý tài nguyên nước, điều chỉnh mùa vụ và giảm thiểu thiệt hại do xâm nhập mặn gây ra trong bối cảnh biến đổi khí hậu ngày càng gia tăng.

## 6. Cấu trúc của đồ án

Nội dung chính của đồ án gồm 4 chương như sau:

Chương 1: Tổng quan tài liệu và cơ sở lý luận về ứng dụng viễn thám, GIS và học máy trong nghiên cứu xâm nhập mặn

Thảo luận về tính cấp thiết của vấn đề xâm nhập mặn tại Đồng bằng sông Cửu Long dưới tác động của biến đổi khí hậu và sự thay đổi nguồn nước thượng nguồn. Các khái niệm cơ bản về viễn thám, Hệ thống thông tin địa lý và Học máy được giới thiệu như những công cụ đắc lực trong giám sát môi trường. Đặc biệt, chương này đi sâu phân tích công nghệ viễn thám phản xạ tín hiệu vệ tinh định vị toàn cầu và hệ thống vệ tinh Cyclone GNSS, làm rõ nguyên lý vật lý và tiềm năng ứng dụng của chúng trong việc quan sát bề mặt Trái Đất xuyên qua lớp mây che phủ.

Chương 2: Phương pháp nghiên cứu

Tập trung vào quy trình xây dựng mô hình giám sát, bắt đầu từ khâu thu thập, tiền xử lý dữ liệu đa nguồn đến việc thiết lập các mô hình toán học. Chương giới thiệu chi tiết lý thuyết của ba thuật toán học máy tiên tiến được sử dụng là XGBoost, Random Forest và CatBoost, cùng phương pháp ước tính hệ số phản xạ bề mặt từ dữ liệu CYGNSS. Các chỉ số đánh giá hiệu suất mô hình như RMSE, MAE và hệ số tương quan R cũng được trình bày để làm cơ sở cho việc kiểm định độ chính xác.

Chương 3: Kết quả thực nghiệm ứng dụng viễn thám, GIS và học máy để thành lập bản đồ xâm nhập tại Đồng Bằng Sông Cửu Long

Trình bày đặc điểm tự nhiên, kinh tế - xã hội của khu vực nghiên cứu và phân tích bộ cơ sở dữ liệu đã xây dựng (bao gồm dữ liệu vệ tinh, địa hình, khí tượng, thổ nhưỡng và dữ liệu đo mặn thực tế). Kết quả huấn luyện và kiểm định của ba mô hình sẽ được so sánh để lựa chọn thuật toán tối ưu nhất (XGBoost). Chương này cũng đi sâu phân tích mức độ quan trọng của các biến đầu vào, đồng thời trình bày các bản đồ phân bố độ mặn theo không gian và thời gian từ tháng 1 đến tháng 5 năm 2025, qua đó đánh giá diễn biến xâm nhập mặn tại các tỉnh ven biển.

Chương 4: Kết luận và đề xuất

Tổng kết lại các kết quả chính đã đạt được, khẳng định hiệu quả của việc tích hợp dữ liệu phản xạ GNSS-R với dữ liệu quang học trong mô hình học máy để giám sát độ mặn. Chương cũng chỉ ra những hạn chế còn tồn tại của nghiên cứu và đưa ra những đề

xuất, hướng phát triển tiếp theo nhằm nâng cao độ chính xác và khả năng ứng dụng thực tiễn trong công tác quản lý tài nguyên nước và ứng phó biến đổi khí hậu.

## **CHƯƠNG 1: Tổng quan tài liệu và cơ sở lý luận về ứng dụng viễn thám, GIS và học máy trong nghiên cứu xâm nhập mặn**

### 1.1. Khái quát về viễn thám

Viễn thám là lĩnh vực khoa học công nghệ cho phép nghiên cứu, thu thập các thông tin về các đối tượng địa lý mà không cần tiếp xúc trực tiếp với đối tượng đó[9]. Công nghệ này hoạt động dựa trên việc ghi nhận năng lượng bức xạ điện từ phản xạ hoặc phát ra từ vật thể, từ đó phân tích và trích xuất thông tin về đặc điểm của bề mặt Trái Đất. Trong bối cảnh biến đổi khí hậu, viễn thám đã trở thành công cụ đắc lực để giám sát các tai biến môi trường, điển hình là xâm nhập mặn.

Hoạt động viễn thám bao gồm toàn bộ các quy trình liên quan đến hệ thống dữ liệu ảnh viễn thám như thu nhận, lưu trữ, xử lý, cung cấp, khai thác và sử dụng dữ liệu, đồng thời xây dựng và hoàn thiện cơ sở dữ liệu ảnh phục vụ nghiên cứu và quản lý. Đây là nền tảng quan trọng bảo đảm cho việc ứng dụng dữ liệu viễn thám trong thực tiễn được hiệu quả, chính xác và kịp thời.

Trong thực tế, viễn thám được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau, trải dài từ địa lý, đo đạc và khảo sát đất đai đến hầu hết các ngành khoa học Trái Đất như thủy văn, khí tượng, hải dương học và địa chất. Ngoài các ứng dụng dân sự, công nghệ này còn giữ vai trò then chốt trong quân sự, tình báo, quy hoạch, kinh tế và các hoạt động nhân đạo. Nhờ khả năng quan sát từ xa và bao quát những khu vực rộng lớn, viễn thám trở thành công cụ không thể thiếu trong công tác theo dõi, đánh giá và dự báo các hiện tượng tự nhiên cũng như các quá trình biến đổi của môi trường và tài nguyên thiên nhiên.

Thuật ngữ “viễn thám” thường được dùng để chỉ việc sử dụng các cảm biến đặt trên vệ tinh, máy bay hoặc các phương tiện bay không người lái nhằm phát hiện, nhận dạng và phân loại các vật thể trên bề mặt Trái Đất. Các cảm biến này cho phép thu nhận thông tin về bề mặt đất liền, khí quyển và đại dương thông qua sự truyền và phản xạ của sóng điện từ. Về nguyên lý hoạt động, viễn thám được chia thành hai loại chính là viễn thám chủ động và viễn thám thụ động. Viễn thám chủ động là hệ thống trong đó cảm biến chủ động phát ra tín hiệu, chẳng hạn như sóng radar hoặc laser, rồi ghi lại tín hiệu phản hồi từ đối tượng, nhờ đó dữ liệu có thể được thu nhận trong mọi điều kiện thời tiết và bất kỳ thời điểm nào trong ngày. Ngược lại, viễn thám thụ động dựa vào nguồn năng lượng tự nhiên, chủ yếu là bức xạ mặt trời, để ghi nhận phần bức xạ được đối tượng phản xạ hoặc phát xạ.

Gắn liền với hoạt động viễn thám là dữ liệu ảnh viễn thám, vốn là những thông tin và hình ảnh về các đối tượng địa lý được thu nhận từ vệ tinh hoặc các thiết bị thu khác. Tại Việt Nam, dữ liệu ảnh viễn thám được khai thác cho nhiều mục đích quan trọng. Trước hết, đây là nguồn thông tin thiết yếu phục vụ công tác quan trắc và giám sát các dạng ô nhiễm môi trường như ô nhiễm đất và nước do chất thải công nghiệp và sinh hoạt, ô nhiễm không khí do khí thải, hoặc các hiện tượng ô nhiễm phát sinh từ thiên tai hay hoạt động khai khoáng. Dữ liệu viễn thám cũng góp phần hỗ trợ kiểm kê khí nhà kính và đánh giá các tác động của biến đổi khí hậu.

Bên cạnh đó, dữ liệu ảnh viễn thám được sử dụng hiệu quả trong thu thập thông tin, phân tích và đánh giá diễn biến tài nguyên và môi trường, cả định kỳ lẫn đột xuất, nhằm phục vụ phát triển kinh tế – xã hội, phòng chống thiên tai và ứng phó với biến đổi khí hậu. Viễn thám cũng hỗ trợ giám sát hạn hán, cảnh báo cháy rừng, theo dõi diễn biến lũ lụt, phục vụ công tác cứu hộ cứu nạn và đánh giá hiện trạng sản xuất nông nghiệp. Một ứng dụng quan trọng khác là xây dựng và cập nhật các bản đồ chuyên đề, cơ sở dữ liệu về hiện trạng tài nguyên thiên nhiên, hiện tượng biến đổi khí hậu và các yếu tố môi trường khác.

Ngoài ra, dữ liệu viễn thám còn được khai thác để cập nhật cơ sở dữ liệu nền địa lý quốc gia và hệ thống bản đồ địa hình, đồng thời phục vụ các nhiệm vụ quốc phòng, an ninh và bảo vệ chủ quyền lãnh thổ. Nhờ phạm vi ứng dụng rộng lớn và khả năng cung cấp thông tin liên tục, viễn thám ngày càng khẳng định vai trò không thể thay thế trong quản lý tài nguyên, giám sát môi trường và phát triển bền vững.

## 1.2. Khái quát về hệ thống thông tin địa lý

### 1.2.1. Khái niệm và công dụng của hệ thống thông tin địa lý

Hệ thống thông tin địa lý là một công cụ công nghệ tích hợp bao gồm phần cứng, phần mềm, dữ liệu và con người, được thiết kế chuyên biệt để thu thập, lưu trữ, quản lý, phân tích và hiển thị các dạng thông tin liên quan đến vị trí trên bề mặt Trái Đất[10]. Khác biệt căn bản của công nghệ này so với các bản đồ giấy truyền thống nằm ở khả năng liên kết chặt chẽ giữa dữ liệu không gian xác định vị trí và dữ liệu thuộc tính mô tả đặc điểm của đối tượng, tạo nên một cơ sở dữ liệu động cho phép người dùng thực hiện các truy vấn phức tạp để giải quyết các vấn đề thực tiễn. Một hệ thống hoàn chỉnh vận hành thông qua chuỗi năm chức năng cốt lõi, bắt đầu từ việc thu thập dữ liệu đa nguồn như ảnh vệ tinh, số liệu đo đạc thực địa hoặc bản đồ số hóa, sau đó chuyển sang khâu quản lý dữ liệu thông qua hệ quản trị cơ sở dữ liệu để đảm bảo tính toàn vẹn và

khả năng truy xuất hiệu quả. Tuy nhiên, sức mạnh thực sự của hệ thống nằm ở chức năng phân tích và truy vấn, cho phép thực hiện các thao tác kỹ thuật chuyên sâu như chồng lớp bản đồ để xác định vùng giao thoa giữa các yếu tố môi trường, tạo vùng đệm quanh mạng lưới sông ngòi để đánh giá phạm vi ảnh hưởng của xâm nhập mặn, hay nội suy không gian để dự đoán giá trị tại các vị trí chưa có số liệu quan trắc. Kết quả của quá trình phân tích này sau đó được chuyển hóa qua chức năng hiển thị dữ liệu dưới dạng các bản đồ chuyên đề, biểu đồ hoặc mô hình ba chiều trực quan, giúp biến các con số khô khan thành thông tin dễ hiểu hỗ trợ đặc lực cho công tác ra quyết định. Cuối cùng, chức năng xuất bản cho phép chia sẻ kết quả dưới nhiều định dạng khác nhau hoặc thông qua các dịch vụ bản đồ trực tuyến, qua đó hệ thống đóng vai trò như xương sống liên kết các nguồn dữ liệu đa dạng từ viễn thám và thực địa để cung cấp cái nhìn tổng thể trong công tác giám sát tài nguyên và môi trường.

### 1.2.2. Vai trò của hệ thống thông tin địa lý trong chủ đề nghiên cứu

Hệ thống thông tin địa lý đóng vai trò trung tâm và không thể thiếu trong chủ đề nghiên cứu giám sát xâm nhập mặn bằng dữ liệu vệ tinh Cyclone GNSS. Nó hoạt động như một nền tảng tích hợp toàn diện, cho phép kết hợp và quản lý các nguồn dữ liệu không gian vô cùng đa dạng. Cụ thể, đây là công cụ để hợp nhất dữ liệu viễn thám radar, dữ liệu quang học, và dữ liệu viễn thám mới với các dữ liệu thực địa vốn là các điểm đo độ dẫn điện được thu thập tại các tọa độ cụ thể. Chức năng quan trọng đầu tiên của nó là chuẩn bị dữ liệu đầu vào cho các mô hình học máy. Hệ thống này được sử dụng để tính toán các yếu tố dẫn xuất quan trọng, ví dụ như trích xuất các yếu tố địa hình như độ cao hay độ dốc từ mô hình số độ cao, hoặc thực hiện các phân tích không gian như tính toán "khoảng cách đến sông" hay "khoảng cách đến biển". Sau đó, chức năng mạnh mẽ nhất của nó là "lồng ghép không gian", nơi nó lấy giá trị từ tất cả các lớp dữ liệu viễn thám và dữ liệu dẫn xuất này tại chính xác vị trí của các mẫu thực địa, qua đó xây dựng nên bộ dữ liệu huấn luyện hoàn chỉnh cho các thuật toán học máy. Cuối cùng, sau khi mô hình học máy dự đoán xong giá trị độ mặn cho toàn bộ khu vực, hệ thống thông tin địa lý chính là công cụ để tiếp nhận kết quả này và thực hiện chức năng trực quan hóa, biến hàng triệu điểm dữ liệu dự đoán thành một bản đồ xâm nhập mặn liên tục. Bản đồ này sau đó thường được phân loại thành các cấp độ mặn khác nhau, tạo ra sản phẩm đầu ra trực quan, hỗ trợ trực tiếp cho nông dân và các nhà hoạch định chính sách trong việc đưa ra các quyết định canh tác và quản lý tài nguyên bền vững.

### 1.3. Khái quát về xâm nhập mặn

Xâm nhập mặn là hiện tượng gia tăng nồng độ muối trong đất, bao gồm hai dạng chính là đất mặn và đất kiềm mặn. Đất mặn được đặc trưng bởi nồng độ muối hòa tan cao, giá trị điện dẫn suất lớn hơn 4 dS/m, tỷ lệ natri hấp phụ nhỏ hơn 15 và pH thấp hơn 8,5, trong khi đất kiềm mặn có điện dẫn suất nhỏ hơn 4 dS/m, tỷ lệ natri hấp phụ lớn hơn 15 và pH cao hơn 8,5 [11], [12]. Độ pH cao của đất kiềm chủ yếu do hàm lượng cacbonat lớn. Quá trình xâm nhập mặn gây ra sự thay đổi hoặc phá vỡ các đặc tính tự nhiên, sinh hóa và các quá trình địa hóa của đất, làm suy giảm chất lượng đất và hạn chế khả năng cung cấp dinh dưỡng cho cây trồng[13]. Khi nồng độ muối trong đất tăng cao, nguồn tài nguyên đất có thể bị mất đi đáng kể, ảnh hưởng trực tiếp đến sản xuất nông nghiệp và tính bền vững môi trường[14]. Nếu không được kiểm soát, xâm nhập mặn có thể phát triển thành một vấn đề kinh tế – xã hội nghiêm trọng và tác động tiêu cực đến đời sống con người trong dài hạn[15].

Nguyên nhân gây xâm nhập mặn được chia thành hai nhóm chính là nguyên nhân do con người và nguyên nhân tự nhiên. Các hoạt động như sử dụng nước tưới kém chất lượng trong thời kỳ hạn hán kéo dài, lạm dụng phân bón hóa học và quản lý tưới tiêu không hợp lý là những yếu tố nhân sinh quan trọng làm gia tăng xâm nhập mặn[16]. Tình trạng này trở nên nghiêm trọng hơn tại các khu vực có hệ thống thoát nước kém, khiến muối tích tụ lâu dài trong đất[17]. Ngoài ra, phong hóa khoáng vật, sự rò rỉ muối từ vật liệu mè và xâm nhập nước biển vào đất liền là những nguyên nhân tự nhiên phổ biến dẫn đến mặn hóa đất. Các vùng đất canh tác, đặc biệt là khu vực phụ thuộc vào tưới tiêu, thường chịu ảnh hưởng nặng nề và dễ bị thoái hóa. Ước tính có hơn 14 nghìn km<sup>2</sup> đất nông nghiệp màu mỡ trên thế giới bị mất đi mỗi năm do xâm nhập mặn, gây suy giảm năng suất và đe dọa an ninh lương thực[18].

Tại Việt Nam, xâm nhập mặn diễn ra rõ rệt nhất ở khu vực Đồng bằng sông Cửu Long, vốn là vùng đất thấp, chịu tác động mạnh của chế độ bán nhật triều và có mạng lưới sông ngòi dày đặc. Trong mùa khô, khi lượng nước từ thượng nguồn sông Mê Công giảm mạnh, dòng chảy yếu làm giảm khả năng đẩy lùi nước mặn từ biển Đông và biển Tây. Điều này tạo điều kiện để các khối nước mặn xâm nhập ngày càng sâu vào lục địa, ảnh hưởng lớn đến sản xuất nông nghiệp, nuôi trồng thủy sản, nguồn nước sinh hoạt và cấu trúc hệ sinh thái.

Theo bộ Tài Nguyên và Môi Trường, xâm nhập mặn mùa khô năm 2019–2020 được ghi nhận là một trong những đợt hạn, mặn nghiêm trọng nhất trong lịch sử tại Đồng bằng sông Cửu Long [19]. Ranh giới mặn tiến sâu bất thường do dòng chảy từ

thượng nguồn sông Mê Công suy giảm mạnh, kết hợp với thời gian khô hạn kéo dài, đã làm gia tăng nhanh nồng độ muối trên các tuyến sông chính và lan rộng vào sâu trong nội đồng. Hiện tượng này ảnh hưởng trực tiếp đến hầu hết các tỉnh trong khu vực, gây tác động trên diện rộng đến sản xuất nông nghiệp, nuôi trồng thủy sản, sinh hoạt và hạ tầng thủy lợi – giao thông.

Về sản xuất nông nghiệp, xâm nhập mặn làm thiệt hại lớn đối với diện tích lúa trong cả vụ mùa và vụ đông xuân, đặc biệt ở các tỉnh ven biển như Cà Mau, Trà Vinh, Tiền Giang và Sóc Trăng. Nhiều diện tích gieo trồng bị mất trắng do nồng độ mặn vượt ngưỡng chịu mặn của cây trồng, trong khi những vùng còn lại bị giảm năng suất do thiếu nước tưới. Cây ăn trái và cây màu cũng chịu tác động nặng nề, với hàng nghìn hecta bị thiếu nước trong thời gian dài, làm giảm chất lượng và sản lượng, thậm chí gây thiệt hại hoàn toàn ở nhiều khu vực[19].

Tổng thể, xâm nhập mặn đã tác động sâu rộng đến hầu hết các lĩnh vực kinh tế – xã hội của vùng Đồng bằng sông Cửu Long. Tuy nhiên, mức độ thiệt hại đã được giảm thiểu đáng kể nhờ sự chủ động của các địa phương trong công tác chỉ đạo, điều tiết sản xuất, vận hành công trình thủy lợi và các giải pháp phòng, chống hạn mặn kịp thời.

#### 1.4. Khái quát về học máy và các phương pháp học máy

Học máy là một lĩnh vực của trí tuệ nhân tạo, tập trung vào việc phát triển các thuật toán và mô hình thống kê cho phép hệ thống máy tính tự động "học" từ dữ liệu. Thay vì được lập trình một cách rõ ràng để thực hiện một tác vụ cụ thể, các hệ thống học máy sử dụng dữ liệu để xác định các mẫu, quy luật và các mối quan hệ ẩn bên trong. Khả năng này làm cho học máy trở nên đặc biệt hữu ích trong việc giải quyết các vấn đề phức tạp, nơi các mối quan hệ giữa các biến đầu vào và đầu ra có thể là phi tuyến tính, điều mà các mô hình hồi quy tuyến tính truyền thống gặp khó khăn. Các mô hình học máy theo hướng dữ liệu này có khả năng xử lý dữ liệu nhiều chiều, dữ liệu có nhiễu, và cũng có thể được sử dụng để ước tính tầm quan trọng tương đối của các biến đầu vào trong mô hình dự đoán[20].

Quá trình xây dựng một mô hình học máy thường bắt đầu bằng việc thu thập và chuẩn bị dữ liệu. Dữ liệu này bao gồm các biến độc lập hay còn gọi là đặc trưng hoặc yếu tố dự báo, ví dụ như các kênh phô, chỉ số thực vật, chỉ số mặn, và các yếu tố địa hình và một biến phụ thuộc biến mục tiêu. Bộ dữ liệu tổng thể sau đó thường được phân chia thành hai phần: một tập dữ liệu huấn luyện và một tập dữ liệu kiểm định. Mô hình được xây dựng hoặc "học" trên tập huấn luyện, sau đó hiệu suất của nó được đánh giá

trên tập kiểm định, vốn là dữ liệu mà mô hình chưa từng thấy trước đó, để đảm bảo tính tổng quát và khả năng dự đoán của nó. Đôi khi, các phương pháp đánh giá chéo, chẳng hạn như kiểm định chéo năm lần, cũng được sử dụng để tăng độ tin cậy của mô hình[20].

Các phương pháp học máy rất đa dạng, nhưng trong các ứng dụng lập bản đồ và dự đoán môi trường, các mô hình học có giám sát là phổ biến nhất. Trong phương pháp này, thuật toán được cung cấp một tập dữ liệu huấn luyện đã được "dán nhãn", nghĩa là mỗi điểm dữ liệu đầu vào đều đi kèm với một kết quả đầu ra đã biết. Học có giám sát có hai tác vụ chính. Thứ nhất là tác vụ hồi quy, được áp dụng khi đầu ra mong muốn là một giá trị số liên tục, chẳng hạn như dự đoán giá trị độ dẫn điện chính xác của đất. Thứ hai là tác vụ phân loại, được sử dụng khi đầu ra là một danh mục hay một lớp rời rạc.

Trong số các thuật toán học có giám sát, các mô hình tổ hợp thường cho thấy hiệu suất cao. Rừng ngẫu nhiên là một ví dụ, đây là thuật toán tích hợp nhiều cây quyết định. Nó hoạt động bằng cách xây dựng một số lượng lớn các cây quyết định trong quá trình huấn luyện và đưa ra dự đoán dựa trên kết quả trung bình trong hồi quy hoặc bỏ phiếu đa số trong phân loại của tất cả các cây[21]. Một nhóm mô hình tổ hợp mạnh mẽ khác là các mô hình tăng cường, chẳng hạn như XGBoost. XGBoost là một cải tiến của cây quyết định tăng cường và hoạt động bằng cách xây dựng các mô hình một cách tuần tự, trong đó mỗi mô hình mới được huấn luyện để sửa chữa những lỗi sai mà mô hình trước đó đã tạo ra[22]. Các thuật toán khác như CatBoost[23] cũng là một dạng mô hình tăng cường. Bên cạnh các mô hình cây, Mạng nơ-ron nhân tạo cũng là một phương pháp được áp dụng rộng rãi, mô phỏng cách thức xử lý thông tin của não người để học các mối quan hệ phi tuyến tính phức tạp. Các máy học khác như Máy Vector Hỗ trợ và Quy trình Gaussian cũng đã được sử dụng.

Một trong những thách thức khi sử dụng học máy là các vấn đề như quá khớp mô hình hoạt động tốt trên dữ liệu huấn luyện nhưng kém trên dữ liệu mới hoặc chưa khớp mô hình quá đơn giản để nắm bắt quy luật. Để cải thiện hiệu suất dự đoán và giải quyết các vấn đề này, các mô hình học máy lai đã được phát triển. Các mô hình này kết hợp một thuật toán học máy cơ sở với một thuật toán tối ưu hóa. Các thuật toán tối ưu hóa này, thường dựa trên các thuật toán tiên hóa hoặc trí tuệ bầy đàn, được sử dụng để tinh chỉnh các tham số nội bộ của mô hình học máy. Ví dụ, Thuật toán Di truyền đã được sử dụng để điều chỉnh trọng số của Mạng nơ-ron nhân tạo. Các thuật toán tối ưu hóa bầy đàn khác bao gồm Tối ưu hóa Diều hâu Harris, Thuật toán Tìm kiếm Sơn ca, Thuật toán Bầy chim, Thuật toán Tìm kiếm Bướm đêm, Thuật toán Tối ưu hóa Cào cào, và Tối ưu

hóa Bầy đàn Hạt. Các mô hình lai này thường cho thấy độ chính xác được cải thiện so với các mô hình cơ sở.

Việc đánh giá hiệu suất của bất kỳ mô hình nào, dù là cơ sở hay lai, đều dựa trên các chỉ số thống kê. Các chỉ số được sử dụng rộng rãi nhất để đo lường độ chính xác bao gồm Hệ số Tương Quan (R), Sai số Bình phương Trung bình Góc, và Sai số Tuyệt đối Trung bình. Một mô hình được xem là hiệu quả khi có giá trị R cao đồng thời có giá trị sai số thấp. Điều quan trọng cần lưu ý là không có mô hình học máy đơn lẻ nào được coi là tốt nhất trong mọi trường hợp. Hiệu suất của một thuật toán phụ thuộc vào đặc điểm cụ thể của bộ dữ liệu và bài toán, do đó, việc so sánh nhiều mô hình khác nhau là một bước phổ biến trong các nghiên cứu ứng dụng.

## 1.5. Công nghệ đo phản xạ tín hiệu của hệ thống vệ tinh định vị toàn cầu

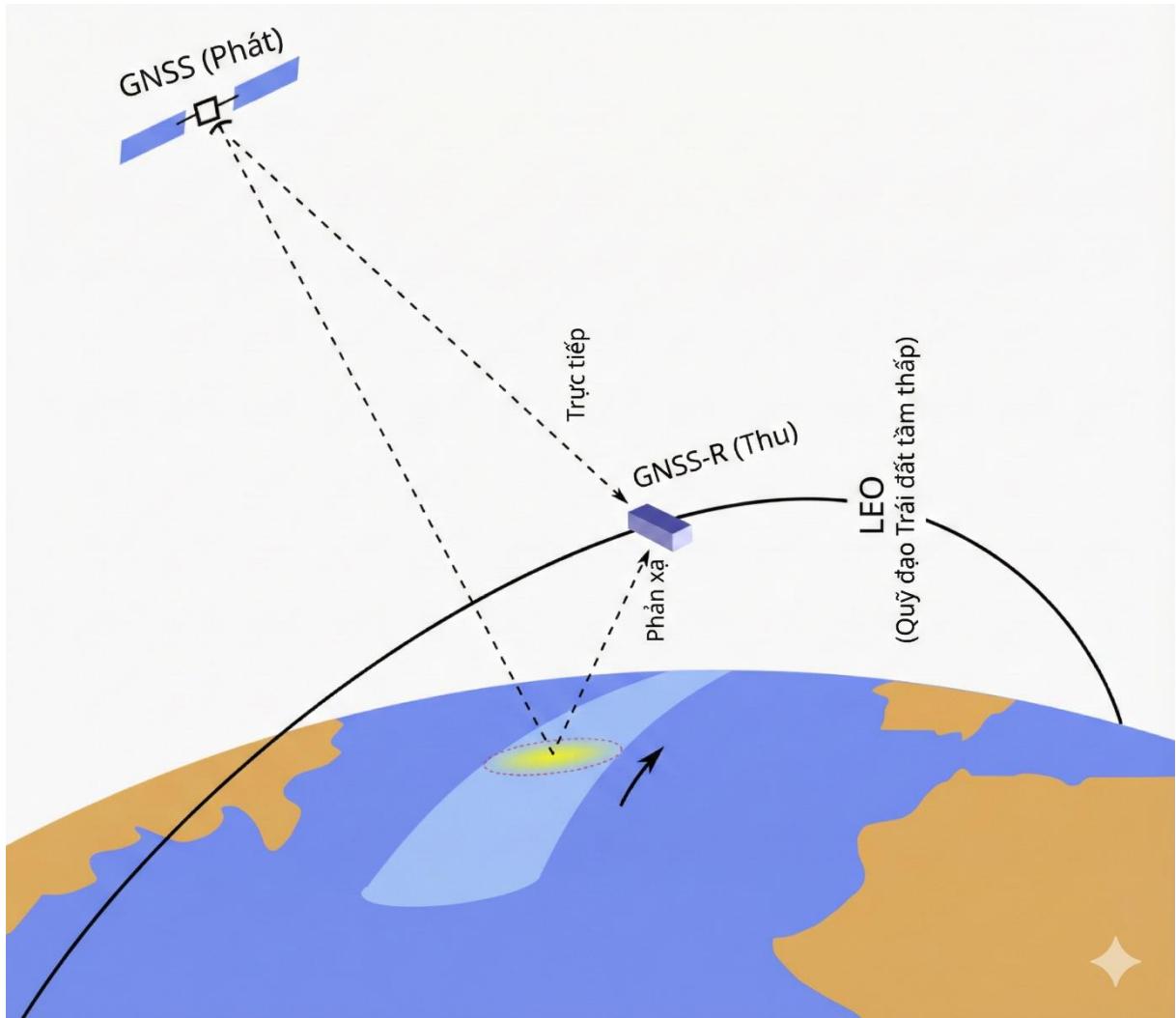
### 1.5.1. Giới thiệu chung về công nghệ phản xạ tín hiệu GNSS-R

Hệ thống vệ tinh định vị toàn cầu, vốn được biết đến rộng rãi với chức năng xác định vị trí và dẫn đường, hiện đang sở hữu chùm vệ tinh lớn nhất và được sử dụng phổ biến nhất trên thế giới. Bên cạnh các ứng dụng truyền thống trong trắc địa và khí tượng học, hệ thống này đã mở ra một hướng đi mới đầy tiềm năng cho ngành viễn thám thông qua việc khai thác các tín hiệu phản xạ. Phương pháp này, được gọi là viễn thám phản xạ tín hiệu vệ tinh định vị toàn cầu, tận dụng các tín hiệu vô tuyến được phát liên tục từ các vệ tinh định vị. Thay vì xem các tín hiệu phản xạ từ bề mặt Trái Đất là nhiều cần loại bỏ như trong các thiết bị định vị thông thường, công nghệ này thu nhận và phân tích chúng để trích xuất các thông tin vật lý về bề mặt phản xạ. Đây là một dạng radar thụ động, trong đó máy phát và máy thu nằm ở hai vị trí tách biệt, tạo thành cấu hình radar song tinh [24].

### 1.5.2. Nguyên lý hoạt động và cơ sở vật lý

Cơ sở vật lý của phương pháp này dựa trên sự tương tác giữa sóng điện từ và bề mặt Trái Đất. Các vệ tinh định vị phát tín hiệu ở băng tần L, với bước sóng trong khoảng từ 19 đến 25 cm. Một ưu điểm vượt trội của tín hiệu ở băng tần này là khả năng xuyên qua mây, mưa và hoạt động ổn định trong mọi điều kiện thời tiết, cả ngày lẫn đêm. Khi tín hiệu từ vệ tinh chạm tới bề mặt Trái Đất, nó sẽ bị phản xạ hoặc tán xạ trở lại không gian. Đặc tính của tín hiệu phản xạ này phụ thuộc chặt chẽ vào hai yếu tố chính của bề mặt là độ gồ ghề và tính chất điện môi. Độ gồ ghề của bề mặt quyết định cách thức tán xạ của sóng, trong khi tính chất điện môi, vốn chịu ảnh hưởng mạnh mẽ bởi độ ẩm và độ mặn, quyết định cường độ của tín hiệu phản xạ [24].

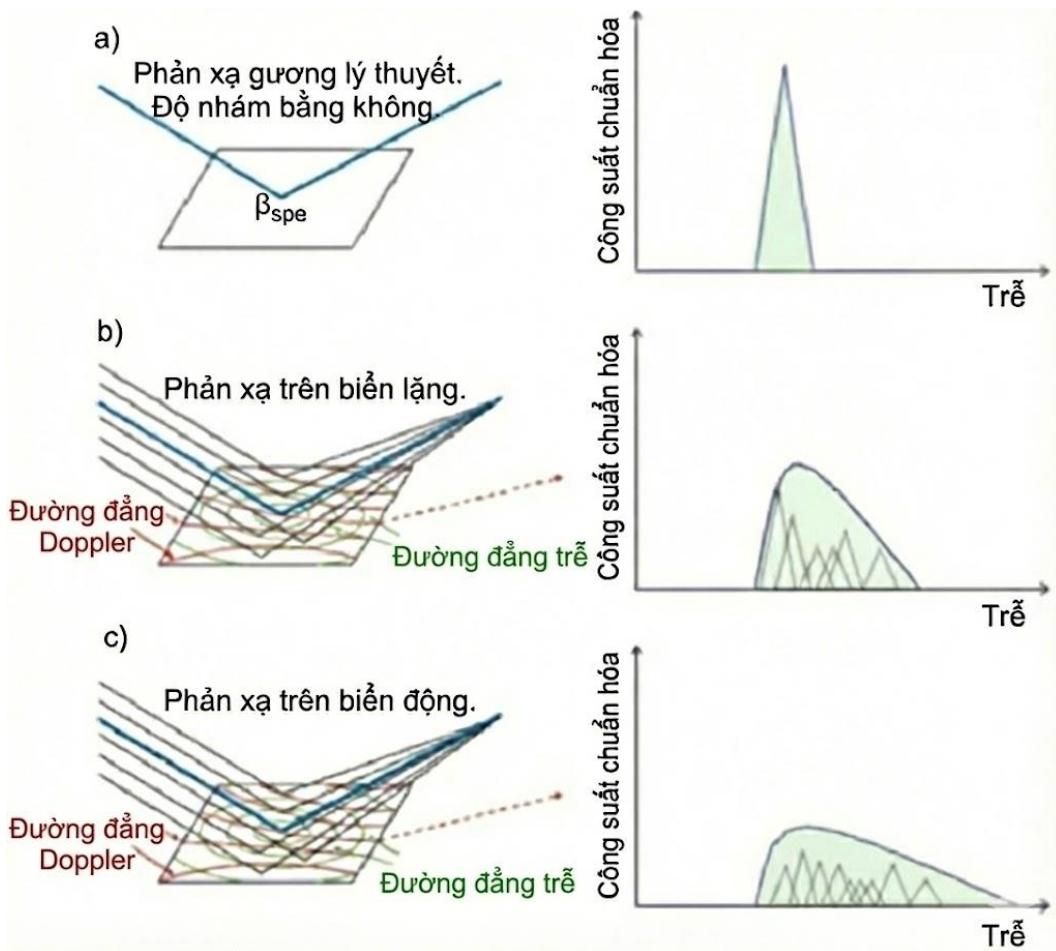
Trong cấu hình này, tín hiệu từ vệ tinh phát được chia thành hai đường truyền gồm đường truyền trực tiếp đến máy thu và đường truyền gián tiếp phản xạ từ bề mặt đất. Điểm phản xạ gương là điểm trên bề mặt mà tại đó góc tới bằng góc phản xạ, và đây là nơi tập trung năng lượng phản xạ mạnh nhất đối với các bề mặt phẳng.



Hình 1: Sơ đồ hình học của hệ thống phản xạ tín hiệu vệ tinh, minh họa đường truyền tín hiệu trực tiếp từ vệ tinh phát và tín hiệu phản xạ từ bề mặt Trái đất đến vệ tinh thu quỹ đạo tầm thấp

Tùy thuộc vào độ gồ ghề của bề mặt, cơ chế phản xạ sẽ thay đổi từ phản xạ gương sang tán xạ khuếch tán. Trên các bề mặt rất phẳng như mặt hồ yên tĩnh, sóng phản xạ giữ được tính chất kết hợp và năng lượng tập trung cao độ theo một hướng xác định. Ngược lại, trên các bề mặt gồ ghề như mặt biển động hay đất đai mấp mô, năng lượng bị tán xạ ra nhiều hướng khác nhau, tạo thành một vùng phản xạ rộng lớn xung quanh điểm phản xạ gương, được gọi là vùng phản xạ khuếch tán. Tín hiệu thu được tại máy

thu là tổng hợp của các sóng tán xạ từ các điểm khác nhau trong vùng này, mỗi điểm có độ trễ truyền dẫn và độ dịch chuyển tần số Doppler khác nhau.



Hình 2: Các dạng sóng phản xạ tương ứng với các điều kiện bề mặt khác nhau, từ phản xạ gương lý tưởng trên bề mặt phẳng ở hình a đến tán xạ khuếch tán trên bề mặt gồ ghề ở hình b và c

Hàng số điện môi là một đại lượng vật lý đo lường khả năng phân cực của vật liệu dưới tác động của điện trường. Nước lỏng có hàng số điện môi rất cao, trong khi đất khô có giá trị thấp hơn nhiều. Do đó, độ ẩm đất tăng lên sẽ làm tăng đáng kể hàng số điện môi hỗn hợp của đất, dẫn đến tín hiệu phản xạ mạnh hơn. Đặc biệt, sự hiện diện của muối hòa tan trong nước làm thay đổi tính chất dẫn điện và phản ảo của hàng số điện môi. Nước mặn có tính chất điện tử khác biệt so với nước ngọt, và sự thay đổi này ảnh hưởng đến hệ số phản xạ Fresnel. Hệ số này xác định tỷ lệ biên độ của sóng phản xạ so với sóng tới, phụ thuộc trực tiếp vào góc tới và hàng số điện môi phức của bề mặt. Vì vậy, tín hiệu thu được từ các vệ tinh phản xạ chứa đựng thông tin tổng hợp về cả độ ẩm và độ mặn của lớp đất mặt.

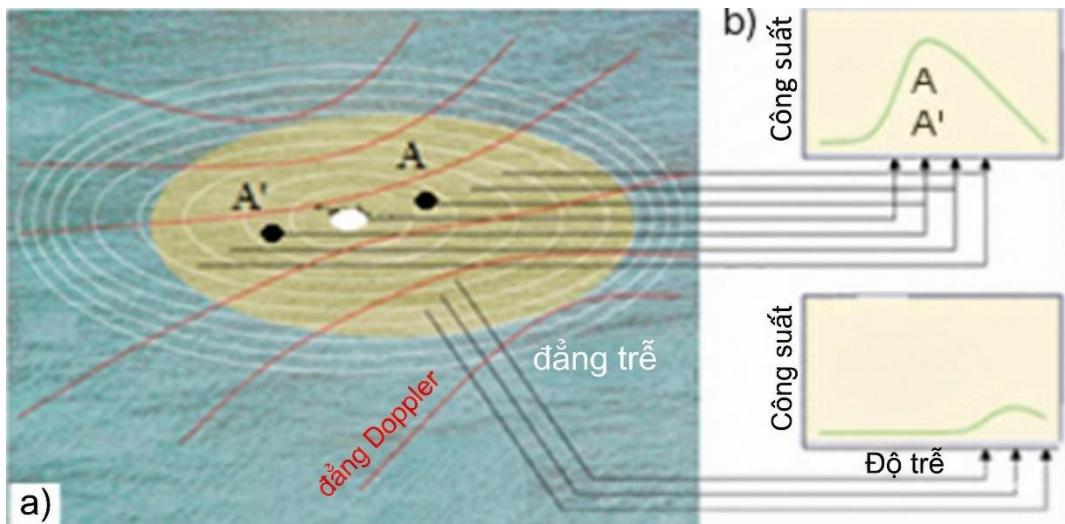
### 1.5.3. Hệ thống vệ tinh Cyclone GNSS

Hệ thống vệ tinh Cyclone GNSS là một chùm vệ tinh tiêu biểu ứng dụng công nghệ phản xạ tín hiệu định vị toàn cầu. Ban đầu được thiết kế cho mục đích khí tượng học để nghiên cứu bão, hệ thống này bao gồm một chùm 8 vệ tinh nhỏ bay ở quỹ đạo thấp quanh Trái Đất. Quỹ đạo của các vệ tinh này tập trung vào vùng nhiệt đới, cung cấp độ bao phủ dày đặc và tần suất cập nhật dữ liệu cao cho các khu vực như Đồng bằng sông Cửu Long. Với khả năng thu nhận tín hiệu phản xạ từ các vệ tinh GPS, Cyclone GNSS cung cấp một nguồn dữ liệu liên tục về đặc tính bề mặt Trái Đất[25].

Nhờ sử dụng tín hiệu băng tần L, Cyclone GNSS có khả năng xuyên qua các lớp thảm thực vật mỏng và trung bình để cảm nhận trực tiếp độ ẩm của lớp đất bên dưới, điều mà các cảm biến quang học và radar băng tần cao hơn thường gặp khó khăn. Dữ liệu từ hệ thống này được cung cấp dưới dạng các bản đồ trễ Doppler, chứa thông tin về công suất phản xạ tại các điểm phản xạ gương trên bề mặt. Đây là nguồn thông tin quý giá để trích xuất các tham số môi trường như độ ẩm đất và độ mặn, đặc biệt hữu ích cho các ứng dụng giám sát thiên tai và quản lý tài nguyên nước tại các khu vực thường xuyên bị mây che phủ[24].

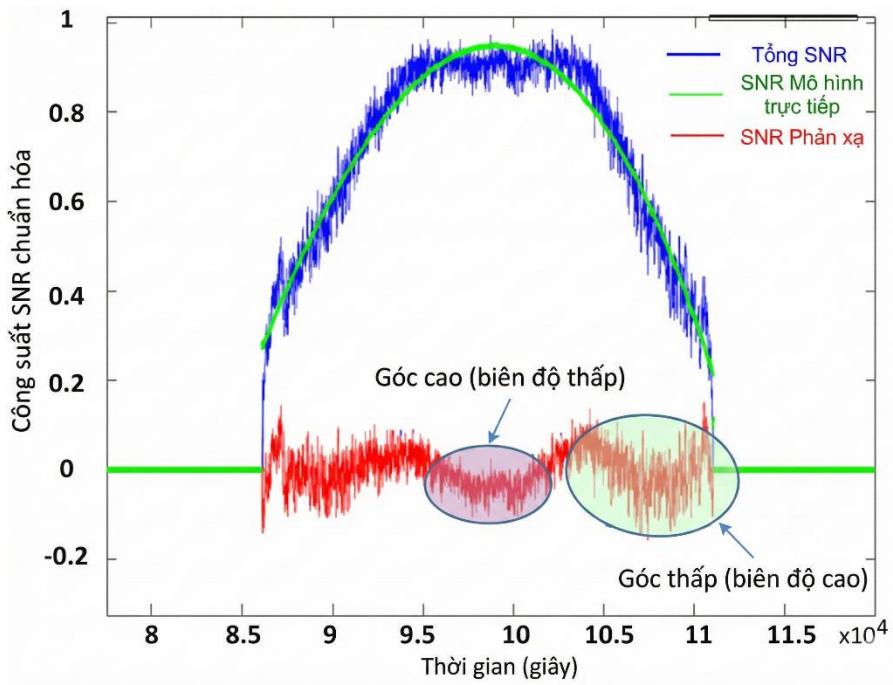
### 1.5.4. Các kỹ thuật đo đặc và tham số quan trọng

Để phân tích tín hiệu phản xạ phức tạp, hai kỹ thuật đo đặc chính thường được sử dụng là kỹ thuật phân tích dạng sóng và kỹ thuật giao thoa. Kỹ thuật phân tích dạng sóng, thường được áp dụng trên các vệ tinh như Cyclone GNSS, sử dụng các máy thu chuyên dụng để tạo ra bản đồ trễ Doppler. Đây là biểu diễn hai chiều của công suất tín hiệu phản xạ theo trực thời gian trễ và trực tần số Doppler. Máy thu thực hiện phép tương quan chéo giữa tín hiệu phản xạ thu được và bản sao của mã định danh vệ tinh. Kết quả là một bản đồ cho thấy sự phân bố năng lượng phản xạ, trong đó điểm có công suất cực đại thường tương ứng với điểm phản xạ gương[25].



Hình 3: Bản đồ trẽ Doppler và các biến dạng công suất tương ứng, với các đường đẳng tần số Doppler và các đường đẳng trẽ hình elip xoay quanh điểm phản xạ gương

Thông tin chứa trong bản đồ trẽ Doppler là chìa khóa để giải mã các tham số môi trường. Hình dạng của bản đồ phản ánh độ gồ ghề của bề mặt, trong khi cường độ hay công suất đỉnh của tín hiệu phản xạ tỷ lệ thuận với hệ số phản xạ của bề mặt, vốn được quyết định bởi hằng số điện môi của đất. Ngoài ra, kỹ thuật giao thoa thường được áp dụng cho các máy thu đặt trên mặt đất hoặc bay thấp, sử dụng một ăng-ten đơn để thu nhận tín hiệu tổng hợp của cả sóng trực tiếp và sóng phản xạ. Sự giao thoa giữa hai sóng này tạo ra các dao động trong tỷ số tín hiệu trên nhiều được ghi nhận bởi máy thu [24].



Hình 4: Biểu đồ minh họa công suất tý số tín hiệu trên nhiễu chuẩn hóa theo thời gian, cho thấy phản dao động do tín hiệu phản xạ gây ra nằm chồng lên tín hiệu trực tiếp  
 (Nguồn: [J.Darrozes](#))

Bằng cách phân tích tần số, biên độ và pha của các dao động tý số tín hiệu trên nhiễu này, người ta có thể ước tính được chiều cao của ăng-ten so với mặt phản xạ và các đặc tính của bề mặt như độ ẩm đất. Tuy nhiên, đối với các ứng dụng giám sát diện rộng từ không gian như sử dụng dữ liệu Cyclone GNSS, phương pháp phân tích bản đồ trễ Doppler vẫn là phương pháp chủ đạo, cho phép trích xuất thông tin về độ ẩm và độ mặn đất thông qua các mô hình đảo ngược hoặc các thuật toán học máy.

Tiềm năng của công nghệ này trong giám sát môi trường là rất lớn. Các hệ thống vệ tinh như Cyclone GNSS, với chùm vệ tinh nhỏ bay ở quỹ đạo thấp, cung cấp khả năng cập nhật dữ liệu với tần suất cao, cho phép theo dõi diễn biến của độ ẩm đất và xâm nhập mặn theo thời gian thực. Khả năng xuyên thấu của tín hiệu băng tần L cũng cho phép thu thập thông tin về độ ẩm đất dưới lớp thảm thực vật mỏng, điều mà các cảm biến quang học và radar băng tần cao hơn gặp khó khăn. Sự kết hợp giữa dữ liệu phản xạ vệ tinh định vị với các nguồn dữ liệu viễn thám truyền thống và các mô hình học máy hứa hẹn sẽ mang lại giải pháp giám sát toàn diện, giúp giải quyết bài toán phân tách tín hiệu độ mặn ra khỏi các yếu tố nhiễu như độ ẩm và độ gồ ghề, phục vụ hiệu quả cho công tác quản lý tài nguyên nước và nông nghiệp.

## 1.6. Khái quát về bản đồ giám sát

Bản đồ giám sát là một công cụ biểu diễn thông tin không gian được thiết kế để theo dõi sự thay đổi của một hiện tượng hoặc một đối tượng cụ thể theo thời gian. Khác với bản đồ tinh chỉ mô tả trạng thái tại một thời điểm duy nhất, bản đồ giám sát là một chuỗi các bản đồ được tạo ra theo các khoảng thời gian đều đặn chẳng hạn như hàng tháng, hàng quý, hoặc hàng năm để nắm bắt được động thái và diễn biến của hiện tượng. Mục đích cốt lõi của chúng là chuyển đổi dữ liệu thô thành một câu chuyện trực quan, cho phép người dùng quan sát, phân tích và hiểu rõ các quy luật thay đổi theo thời gian.

Trong nhiều lĩnh vực khoa học Trái Đất, các hiện tượng môi trường mang tính biến động cao. Ví dụ, trong giám sát xâm nhập mặn, độ mặn của đất không phải là một hằng số mà thay đổi rõ rệt giữa các mùa. Mùa khô, với lượng mưa ít và bốc hơi cao, có thể làm gia tăng nồng độ muối, trong khi mùa mưa có thể làm loãng và rửa trôi muối. Một bản đồ duy nhất được lập trong mùa khô sẽ không thể đại diện cho tình trạng của cả năm. Do đó, việc xây dựng một chuỗi bản đồ giám sát theo tháng hoặc theo mùa là cần thiết để nắm bắt toàn bộ chu kỳ biến động này, cung cấp thông tin kịp thời cho việc quản lý và ứng phó.

Giá trị của các bản đồ giám sát thể hiện ở khả năng hỗ trợ ra quyết định kịp thời. Đối với nông dân, các bản đồ này cung cấp thông tin trực quan về thời điểm và địa điểm mặn xâm nhập hoặc rút lui, giúp họ điều chỉnh lịch thời vụ, lựa chọn giống cây trồng chịu mặn thích hợp, hoặc tối ưu hóa lịch trình tưới tiêu nước ngọt. Đối với các nhà hoạch định chính sách và cơ quan quản lý tài nguyên nước, chuỗi bản đồ giám sát cho phép họ đánh giá hiệu quả của các công trình thủy lợi cần ưu tiên can thiệp, và phân biệt giữa các biến động mặn theo mùa tự nhiên với các xu hướng suy thoái lâu dài do biến đổi khí hậu hoặc các hoạt động của con người ở thượng nguồn.

## 1.7. Các phương pháp nghiên cứu đánh giá xâm nhập mặn

### 1.7.1. Trên thế giới

Các phương pháp truyền thống để đánh giá và lập bản đồ xâm nhập mặn từ lâu đã dựa trên nền tảng của việc khảo sát thực địa và phân tích mẫu đất trong phòng thí nghiệm của (J. D. Rhoades và cộng sự, 1993)[6]. Ưu điểm lớn nhất của các kỹ thuật này, điển hình như đo độ dẫn điện của dịch chiết bão hòa, là khả năng cung cấp các số liệu định lượng chuẩn xác về độ mặn tại các điểm đo cụ thể, do đó chúng thường được coi là dữ liệu tham chiếu chuẩn. Tuy nhiên, nhược điểm chí mạng của phương pháp này là tiêu tốn quá nhiều thời gian, kinh phí và nhân lực. Do tính biến động không gian

lớn của độ mặn, việc lấy mẫu điểm rời rạc khiến phương pháp này trở nên không thực tế để giám sát trên quy mô lớn hoặc theo dõi diễn biến xâm nhập mặn liên tục theo thời gian thực tại các vùng đồng bằng rộng lớn.

Để khắc phục những hạn chế về không gian của phương pháp truyền thống, công nghệ viễn thám đã được ứng dụng rộng rãi, các nghiên cứu áp dụng cảm biến quang học như Landsat(Aksoy và cộng sự, 2024)[26], Sentinel 2(Ma và cộng sự, 2021)[27]. Cụ thể, nghiên cứu của Aksoy và cộng sự sử dụng ảnh Landsat 8 đã chứng minh thuật toán XGBoost cho hiệu quả vượt trội hơn so với Random Forest, với hệ số xác định đạt tới 0,83 tại khu vực nghiên cứu Đông Nam hồ Urmia. Kết quả của họ chỉ ra rằng các dải sóng nhìn thấy đóng góp lớn trong việc nhận diện lớp vỏ muối, trong khi các chỉ số thực vật lại quan trọng hơn tại các khu vực đất mặn có thảm phủ. Tuy nhiên, hạn chế của nghiên cứu này là sự phụ thuộc vào số lượng mẫu thực địa hạn chế và khó khăn trong việc phân biệt đất mặn nhẹ hoặc trung bình do sự can nhiễu của các khoáng chất khác.

Tương tự, nghiên cứu của Ma và cộng sự đã ứng dụng dữ liệu Sentinel-2 để lập bản đồ mặn tại ốc đảo sông Ogan-Kuqa. Kết quả nghiên cứu khẳng định các chỉ số thực vật là biến số quan trọng nhất để dự báo độ mặn. Mặc dù vậy, nghiên cứu cũng chỉ ra rằng nếu chỉ dựa vào cảm biến quang học Sentinel-2, độ chính xác sẽ bị giới hạn bởi điều kiện thời tiết và độ che phủ thực vật.

Hiện nay, để giải quyết các mối quan hệ phi tuyến tính và phức tạp này, xu hướng chung của cộng đồng khoa học là tích hợp đa nguồn dữ liệu gồm quang học, radar và địa hình bằng cách sử dụng các mô hình học máy. Các thuật toán như Rừng ngẫu nhiên, Mạng nơ-ron nhân tạo và các mô hình tăng cường độ dốc đã cho thấy hiệu quả vượt trội trong việc xử lý dữ liệu đa chiều và dự báo độ mặn chính xác hơn so với các mô hình thông kê tuyến tính truyền thống. Trong bối cảnh đó, một hướng đi công nghệ mới và đầy hứa hẹn là sử dụng dữ liệu phản xạ tín hiệu hệ thống vệ tinh định vị toàn cầu từ các vệ tinh như Cyclone GNSS. Tương tự như radar, tín hiệu này hoạt động ổn định trong mọi điều kiện thời tiết và nhạy cảm với hằng số điện môi bề mặt. Điểm ưu việt của hệ thống này là khả năng thu thập dữ liệu với tần suất cao nhờ chùm vệ tinh lớn, mở ra tiềm năng to lớn trong việc theo dõi động thái biến đổi nhanh của độ ẩm và xâm nhập mặn. Việc ứng dụng dữ liệu Cyclone GNSS cho giám sát xâm nhập mặn vẫn còn là một lĩnh vực mới mẻ. Do đó, nghiên cứu tích hợp nguồn dữ liệu này vào các mô hình học máy tiên tiến, kết hợp cùng dữ liệu quang học và radar, được kỳ vọng sẽ khắc phục các nhược điểm còn tồn tại của các phương pháp trước đây, qua đó nâng cao độ chính xác và khả năng cảnh báo sớm xâm nhập mặn.

### 1.7.2. Tại Việt Nam

Ở Việt Nam, các nghiên cứu về xâm nhập mặn cũng được triển khai để giải quyết vấn đề. Các phương pháp điều tra thực địa và phân tích hóa lý mẫu đất trong phòng thí nghiệm thường được sử dụng làm dữ liệu tham chiếu chuẩn nhờ khả năng cung cấp số liệu định lượng chính xác về độ mặn. Các kỹ thuật địa vật lý như Chụp ảnh điện trở suất đã được áp dụng trong nghiên cứu của (Nguyen và cộng sự, 2023)[28] tại tỉnh Trà Vinh đã chứng minh khả năng của kỹ thuật là khả thi. Kết quả nghiên cứu chỉ ra mối liên hệ trực tiếp giữa độ sâu tầng nước mặn và khoảng cách đến biển đồng thời xác định giá trị điện trở suất nhỏ hơn  $3 \Omega\text{m}$  là chỉ thị rõ ràng cho sự hiện diện của nước mặn. Tuy nhiên, do tính biến động không gian lớn của độ mặn, việc phụ thuộc hoàn toàn vào lấy mẫu điểm rời rạc khiến các phương pháp này trở nên tốn kém về thời gian, kinh phí và khó áp dụng để giám sát liên tục trên quy mô lớn tại Đồng bằng sông Cửu Long.

Để khắc phục hạn chế trên, các nghiên cứu đã chuyển sang sử dụng dữ liệu viễn thám. Radar khẩu độ tổng hợp, điển hình là Sentinel-1 được sử dụng để ước tính độ mặn đất. Sóng radar có khả năng xuyên mây và rất nhạy cảm với hằng số điện môi của đất, vốn bị ảnh hưởng bởi cả độ ẩm và độ mặn. Nghiên cứu tại tỉnh Bến Tre của (Phạm Việt Hoa và cộng sự, 2019) đã ứng dụng các đặc trưng phân cực (VV, VH) và đặc trưng kết cấu từ Sentinel-1 để lập bản đồ mặn[29]. đạt hiệu suất dự báo tốt nhất với sai số toàn phương trung bình là  $2,885 \text{ dS/m}$  và hệ số tương quan đạt  $0,808$ , khẳng định tiềm năng của dữ liệu vệ tinh radar trong việc giám sát độ mặn tại khu vực nhiệt đới nhiều mây. Tuy nhiên, nghiên cứu này cũng chỉ ra một số hạn chế, cụ thể là số lượng mẫu đất thực địa dùng để huấn luyện mô hình còn ít dẫn đến hiện tượng quá khớp ở một số thuật toán, đồng thời các mô hình gặp khó khăn trong việc dự báo chính xác tại những khu vực có giá trị độ mặn quá cao. Mặc dù vậy, thách thức lớn nhất vẫn là tín hiệu tán xạ ngược của radar là một hàm phức tạp của nhiều yếu tố như độ ẩm, độ mặn, độ gồ ghề bề mặt, khiến việc bóc tách chính xác tín hiệu của độ mặn trở nên khó khăn.

Bên cạnh đó, cùng với xu hướng phát triển công nghệ hiện nay, các đề tài ứng dụng học máy và trí tuệ nhân tạo đã được thực hiện và chứng minh hiệu quả vượt trội trong việc giải quyết các mối quan hệ phi tuyến tính phức tạp này. Các thuật toán như Rừng ngẫu nhiên và Xgboost đang được sử dụng rộng rãi để tích hợp dữ liệu đa nguồn nhằm nâng cao độ chính xác dự đoán.

Tuy nhiên, các nghiên cứu nêu trên chủ yếu vẫn giới hạn trong dữ liệu quang học và SAR. Một công nghệ mới đang được quan tâm là dữ liệu phản xạ tín hiệu hệ thống vệ tinh định vị toàn cầu từ chùm vệ tinh Cyclone GNSS. Công nghệ này hoạt động trong

mọi điều kiện thời tiết và tín hiệu L-band mà nó sử dụng rất nhạy cảm với hằng số điện môi của bề mặt, cung cấp thông tin quý giá về cả độ ẩm [30] và độ mặn. Đối với kỹ thuật GNSS-R nói chung, nghiên cứu của (Hà Minh Cường và cộng sự, 2017) [31] đã sử dụng dữ liệu tỷ số tín hiệu trên nhiễu (SNR) từ một ăng-ten thương mại để thu hồi độ ẩm đất và chiều cao thảm thực vật lúa mì. Bằng phương pháp phân tích mẫu giao thoa (IPT), kết quả cho thấy độ tương quan cao với dữ liệu thực địa ( $R > 0,8$  đối với độ ẩm đất), tuy nhiên phương pháp này hạn chế về phạm vi không gian do phụ thuộc vào vị trí đặt trạm. Tương tự, nghiên cứu của (Hà Minh Cường, 2018) [32] cũng khẳng định khả năng theo dõi diễn biến độ ẩm đất thông qua phân tích tín hiệu đa đường dẫn. Nghiên cứu của Vũ Phương Lan và công sự vào năm 2022 cũng đã ứng dụng thành công kỹ thuật này để giám sát mực nước tại Phá Tam Giang, Thừa Thiên Huế với sai số RMSE chỉ từ 2,7 đến 3,6 cm [33], và sau đó vào năm 2023 tiếp tục ở rộng ứng dụng để phát hiện bão và lũ lụt với độ chính xác cao [34]. Những kết quả này chứng minh độ nhạy của tín hiệu GNSS phản xạ với sự thay đổi tính chất bề mặt (nước, độ ẩm), tạo tiền đề vững chắc cho việc mở rộng sang giám sát độ mặn.

Về ứng dụng vệ tinh CYGNSS các nghiên cứu tiên phong đã bắt đầu khai thác dữ liệu này cho các bài toán quy mô lớn. Nghiên cứu của (Huu Duy Nguyen và cộng sự, 2022) [35] tại tỉnh Nghệ An đã tích hợp dữ liệu độ ẩm đất từ CYGNSS với các thuật toán học máy lai ghép để lập bản đồ nguy cơ lũ lụt. Kết quả đạt được rất khả quan với diện tích dưới đường cong (AUC) lớn hơn 0,9, chứng minh rằng dữ liệu CYGNSS có thể cung cấp thông tin bề mặt quan trọng ở những khu vực bị hạn chế dữ liệu. Đối với bài toán độ mặn, nghiên cứu của Wang và cộng sự (2023) [36] tại Đồng bằng sông Hoàng Hà, kết quả thực nghiệm cho thấy tín hiệu CYGNSS có độ nhạy cao với EC của đất, đạt độ chính xác tốt với hệ số tương quan  $R=0,730$  và sai số RMSE=1,318 mS/cm. Kết quả cho thấy tính khả quan của vệ tinh CYGNSS trong việc giám sát xâm nhập mặn.

Xét thấy điểm chung của các nghiên cứu tại Việt Nam là đã ứng dụng viễn thám quang học, radar và các mô hình học máy cơ bản, nhưng việc tích hợp nguồn dữ liệu GNSS-R vào các mô hình học máy tiên tiến để giám sát xâm nhập mặn vẫn chưa được thực hiện. Đây là một khoảng trống nghiên cứu lớn, trong khi CYGNSS có khả năng cung cấp dữ liệu liên tục, bất chấp mây che, điều mà cả viễn thám quang học và SAR bị hạn chế.

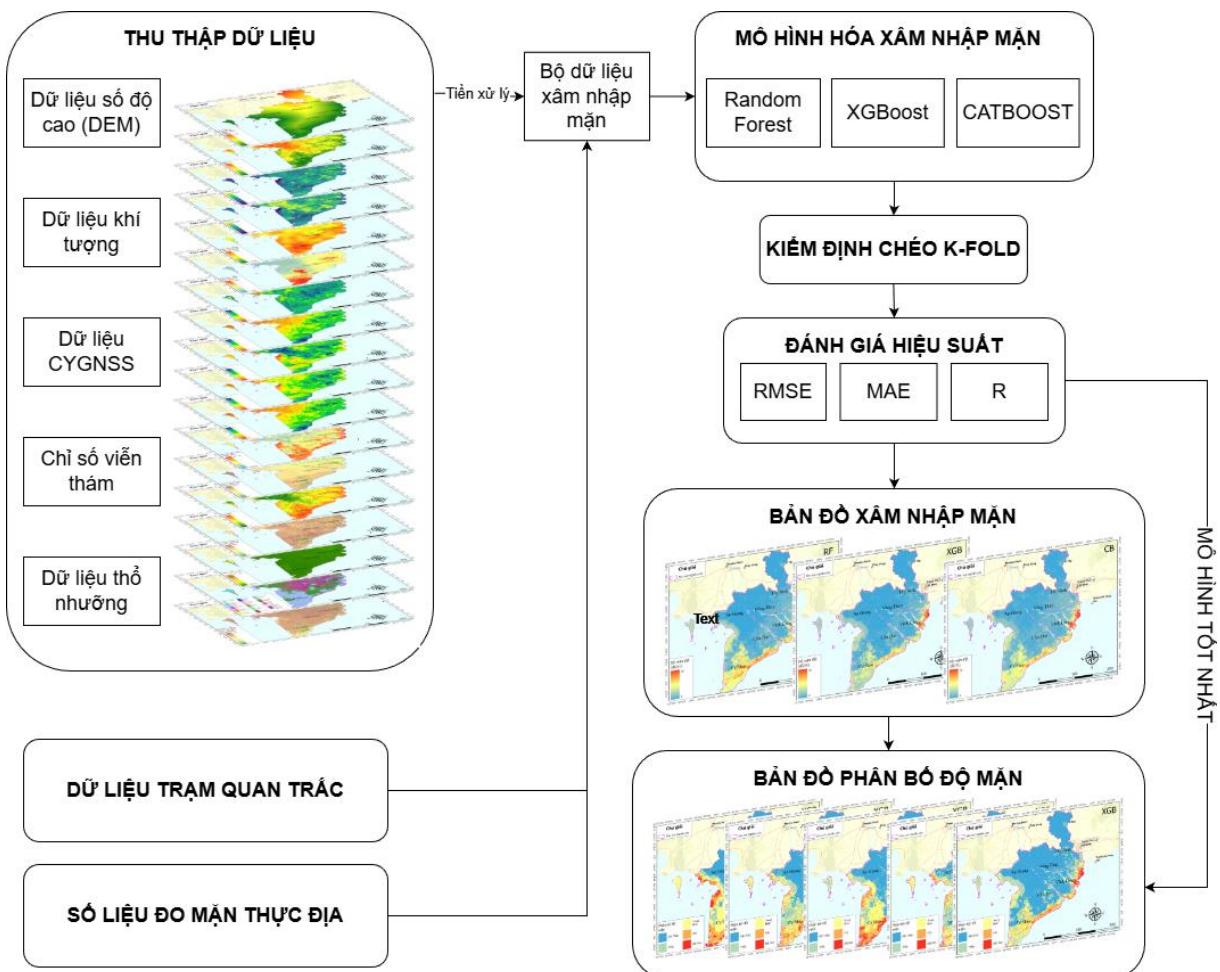
Vì vậy, mục tiêu của nghiên cứu này là phát triển các mô hình học máy, tích hợp đồng thời dữ liệu quang học và dữ liệu GNSS-R để xây dựng bản đồ giám sát xâm nhập mặn cho các khu vực ven biển. Giả thuyết được kiểm tra trong nghiên cứu này là việc

bổ sung dữ liệu CYGNSS có thể cải thiện đáng kể độ chính xác của các mô hình học máy trong việc dự đoán độ mặn, đặc biệt là ở những khu vực thường xuyên bị mây che hoặc có thảm thực vật phủ. Những phát hiện của nghiên cứu này có thể được coi là một công cụ hỗ trợ quan trọng cho các nhà hoạch định chính sách, giúp xác định các khu vực có nguy cơ xâm nhập mặn cao, từ đó có kế hoạch canh tác và sử dụng đất bền vững, giảm thiểu tác động của biến đổi khí hậu đến an ninh lương thực.

## CHƯƠNG 2: Phương pháp nghiên cứu

### 2.1. Sơ đồ quy trình nghiên cứu

Dưới đây là sơ đồ biểu diễn toàn bộ quy trình nghiên cứu thực nghiệm với dữ liệu trên Đồng Bằng Sông Cửu Long:



Hình 5: Sơ đồ quy trình nghiên cứu

Quy trình nghiên cứu được triển khai một cách hệ thống, bắt đầu với giai đoạn thu thập và xử lý dữ liệu. Trong nghiên cứu này, cơ sở dữ liệu được phân chia thành hai nhóm chính để phục vụ cho việc huấn luyện và kiểm chứng mô hình: nhóm thứ nhất bao gồm mẫu độ dẫn điện thu thập qua khảo sát thực địa trong giai đoạn từ năm 2019 đến năm 2025, được sử dụng làm biến phụ thuộc; nhóm thứ hai bao gồm 18 biến độc lập được tổng hợp từ đa dạng nguồn dữ liệu, bao gồm hệ số phản xạ từ vệ tinh CYGNSS, 9 chỉ số viễn thám trích xuất từ ảnh MODIS, cùng các biến bổ trợ quan trọng khác như địa hình, thổ nhưỡng và khí tượng. Sự đồng bộ về mặt thời gian giữa dữ liệu thực địa và dữ liệu vệ tinh trong cùng một giai đoạn mùa khô được đảm bảo nghiêm ngặt nhằm tối ưu hóa tính chính xác cho quá trình phân tích.

Tiếp nối giai đoạn chuẩn bị dữ liệu là bước xây dựng mô hình. Nhận thức được tầm quan trọng của việc lựa chọn đặc trưng đầu vào trong việc quyết định hiệu suất dự báo - bởi sự dư thừa dữ liệu có thể gây nhiễu, tăng thời gian tính toán và giảm hiệu quả của mô hình hồi quy - nghiên cứu đã sàng lọc và sử dụng 18 biến độc lập tiêu biểu từ các nhóm dữ liệu thổ nhưỡng, địa hình, khí tượng, chỉ số viễn thám quang học và đặc biệt là tín hiệu phản xạ từ CYGNSS. Tập dữ liệu sau xử lý được đưa vào huấn luyện trên ba thuật toán học máy tiên tiến là XGBoost, Random Forest và CatBoost. Để đảm bảo độ tin cậy và tính ổn định của kết quả, đồng thời giảm thiểu sai lệch do phân chia dữ liệu ngẫu nhiên, quá trình huấn luyện và kiểm định được thực hiện thông qua phương pháp kiểm định chéo K-Fold với K = 5.

Hiệu quả của các mô hình sau đó được đánh giá định lượng thông qua ba chỉ tiêu thống kê chính: 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 phản ánh mức độ phù hợp giữa giá trị dự báo của mô hình và giá trị quan sát thực tế, từ đó làm cơ sở để so sánh khả năng tổng quát hóa của từng thuật toán. Cuối cùng, dựa trên kết quả đánh giá, mô hình đạt hiệu suất cao nhất sẽ được lựa chọn để thành lập bản đồ phân bố độ mặn đất cho khu vực Đồng Bằng Sông Cửu Long. Bản đồ này được xây dựng thông qua việc tính toán giá trị EC dự báo trên toàn bộ vùng nghiên cứu, cung cấp cái nhìn trực quan về sự biến thiên không gian của mức độ nhiễm mặn trong giai đoạn khảo sát.

## 2.2. Các phương pháp học máy trong thành lập bản đồ giám sát xâm nhập mặn

### 2.2.1. Thuật toán XGBoost

XGBoost, viết tắt của Extreme Gradient Boosting, là một thuật toán học máy hiệu quả và có khả năng mở rộng cao, được xây dựng dựa trên kỹ thuật tăng cường độ dốc của cây. Được Tianqi Chen và Carlos Guestrin giới thiệu, XGBoost đã nhanh chóng trở thành một trong những phương pháp được ưa chuộng và sử dụng rộng rãi trong cộng đồng khoa học dữ liệu, đạt được nhiều kết quả hàng đầu trong các thách thức học máy. Về bản chất, đây là một mô hình tổ hợp sử dụng các cây quyết định cụ thể là cây hồi quy, hay CART làm mô hình cơ sở, và huấn luyện chúng theo một phương thức tăng cường tuần tự[22].

Để hiểu XGBoost, trước tiên cần nắm được nguyên lý của Tăng cường độ dốc, hay Gradient Boosting Machine. Không giống như Rừng ngẫu nhiên vốn xây dựng các cây một cách độc lập và song song, Tăng cường độ dốc xây dựng các cây một cách tuần tự.

Ý tưởng cốt lõi là mỗi cây mới được thêm vào mô hình sẽ có gắng sửa chữa những lỗi sai mà các cây trước đó đã tạo ra.

Quá trình này diễn ra một cách cộng tính. Giả sử tại vòng lặp thứ  $t$ , mô hình đã có dự đoán là  $\hat{y}_i^{(t-1)}$ . Mô hình sẽ xây dựng một cây mới,  $f_t(x_i)$ , sao cho cây này dự đoán được phần dư hoặc sai số của mô hình trước đó. Dự đoán mới sẽ được cập nhật:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (1)$$

Mục tiêu là tìm ra hàm  $f_t$  ở mỗi bước sao cho hàm mục tiêu tổng thể tổng sai số bình phương được giảm thiểu nhiều nhất. Tăng cường độ dốc tổng quát hóa ý tưởng này bằng cách sử dụng đạo hàm gradient của hàm mất mát. Thay vì chỉ khớp với phần dư, mỗi cây mới được huấn luyện để khớp với độ dốc âm của hàm mất mát, vốn chỉ ra hướng làm giảm sai số nhanh nhất.

XGBoost kế thừa ý tưởng này nhưng tối ưu hóa nó bằng một hàm mục tiêu được chính quy hóa một cách rõ ràng. Đối với một tập dữ liệu  $D$  có  $n$  mẫu và  $m$  đặc trưng, mô hình tổ hợp cây sử dụng  $K$  cây để dự đoán đầu ra:

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^K f_k(\mathbf{x}_i), f_k \in \mathcal{F} \quad (2)$$

trong đó  $\mathcal{F}$  là không gian của các cây hồi quy (CART). Mỗi cây  $f_k$  chứa một cấu trúc  $q$  ánh xạ một mẫu tới một nút lá và một véc-tơ trọng số  $w$  chứa điểm số liên tục tại mỗi nút lá.

XGBoost tối ưu hóa hàm mục tiêu sau, bao gồm hàm mất mát và một thành phần chính quy hóa:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (3)$$

trong đó  $l$  là hàm mất mát lồi khả vi - sai số bình phương đo lường sự khác biệt giữa dự đoán  $\hat{y}_i$  và giá trị thực  $y_i$ . Thành phần  $\Omega$  là yếu tố chính quy hóa, có vai trò phạt sự phức tạp của mô hình để tránh hiện tượng quá khớp. XGBoost định nghĩa cụ thể thành phần chính quy hóa này là:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (4)$$

Ở đây,  $T$  là số lượng lá trong cây, và  $w_j$  là trọng số của lá thứ  $j$ . Tham số  $\gamma$  kiểm soát mức phạt cho việc thêm một nút lá mới kiểm soát độ phức tạp của cấu trúc cây, trong khi  $\lambda$  là tham số chính quy hóa L2 cho các trọng số lá giúp làm mượt các dự đoán cuối cùng và tránh các trọng số quá lớn.

Vì mô hình được huấn luyện theo kiểu cộng tính, tại vòng lặp thứ  $t$ , chúng ta tìm cây  $f_t$  để tối ưu hóa hàm mục tiêu:

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (5)$$

Để tối ưu hóa hàm mục tiêu này một cách nhanh chóng, XGBoost sử dụng phép xấp xỉ Taylor bậc hai. Đặt  $g_i$  là đạo hàm bậc nhất gradient và  $h_i$  là đạo hàm bậc hai Hessian của hàm mất mát  $l$  tại điểm dự đoán

$$g_i = \partial_{\hat{y}^{(t-1)}} \ell(y_i, \hat{y}_i^{(t-1)}) \quad (6)$$

$$h_i = \partial_{\hat{y}^{(t-1)}}^2 \ell(y_i, \hat{y}_i^{(t-1)}) \quad (7)$$

Sau khi loại bỏ các thành phần hằng số, hàm mục tiêu tại bước  $t$  có thể được xấp xỉ là:

$$\tilde{\mathcal{L}}^{(t)} \simeq \sum_{i=1}^n \left[ g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (8)$$

Đây là một bước cải tiến quan trọng so với tăng cường độ dốc truyền thống chỉ sử dụng đạo hàm bậc nhất  $g_i$ . Việc sử dụng cả đạo hàm bậc hai  $h_i$  cung cấp thông tin về độ cong của hàm mất mát, giúp mô hình hội tụ nhanh hơn và chính xác hơn, tương tự như phương pháp Newton trong tối ưu hóa.

Bằng cách thay thế  $\Omega(f_t)$  và nhóm các mẫu theo từng nút lá  $j$  (với  $I_j = \{i \mid q(x_i) = j\}$  là tập hợp các mẫu rơi vào lá  $j$ ), hàm mục tiêu có thể được viết lại dưới dạng tổng theo từng lá:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{j=1}^T \left[ \left( \sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left( \sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \quad (9)$$

Đặt  $G_j = \sum_{i \in I_j} g_i$  (tổng gradient của các mẫu trong lá  $j$ ) và  $H_j = \sum_{i \in I_j} h_i$  (tổng Hessian của các mẫu trong lá  $j$ ). Đối với một cấu trúc cây  $q(x)$  đã cố định, trọng số tối ưu  $w_j^*$  cho mỗi lá  $j$  có thể được tính toán trực tiếp:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \quad (10)$$

Khi thay giá trị  $w_j^*$  này trở lại vào hàm mục tiêu, chúng ta thu được điểm chất lượng  $\tilde{\mathcal{L}}^{(t)}(q)$  dùng để đánh giá một cấu trúc cây q là tốt đến mức nào:

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (11)$$

Hàm tính điểm này đóng vai trò tương tự như chỉ số Gini hay Entropy trong cây quyết định truyền thống, nhưng nó được suy ra từ hàm mất mát tổng quát. Một giá trị điểm càng thấp do  $G_j^2$  lớn thì cấu trúc cây càng tốt.

Không thể duyệt qua tất cả các cấu trúc cây khả dĩ, XGBoost sử dụng một thuật toán tham lam để xây dựng cây. Bắt đầu từ một lá duy nhất, thuật toán lặp đi lặp lại việc thêm các nhánh vào cây. Tại mỗi nút, nó cố gắng tìm một phép chia một đặc trưng và một ngưỡng để tối đa hóa "lợi ích". Giả sử một nút I được chia thành hai nút con  $I_L$  trái và  $I_R$  phải, lợi ích của phép chia đó được tính bằng:

$$\mathcal{L}_{\text{split}} = \frac{1}{2} \left[ \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda} \right] - \gamma \quad (12)$$

trong đó  $G_L, H_L, G_R, H_R, G, H$  là tổng gradient và Hessian của các nút con trái, phải và nút cha tương ứng. Giá trị  $\lambda$  là chi phí cho việc thêm một nút lá mới vì phép chia thêm một lá. Phép chia chỉ được thực hiện nếu  $\mathcal{L}_{\text{split}} > 0$ .

Sự thành công của XGBoost không chỉ đến từ hàm mục tiêu chính quy hóa mà còn từ các tối ưu hóa hệ thống và thuật toán, giúp nó có khả năng mở rộng vượt trội:

Thuật toán Nhận biết Độ thưa: Trong các bài toán thực tế, dữ liệu đầu vào thường rất thưa, có thể do giá trị bị thiếu, giá trị 0, hoặc do kỹ thuật mã hóa one-hot. XGBoost xử lý tất cả các kiểu thưa thớt này một cách thống nhất. Thay vì duyệt qua các giá trị 0, thuật toán chỉ lặp qua các giá trị khác 0. Tại mỗi nút, nó học một hướng mặc định cho các giá trị bị thiếu. Bằng cách chỉ duyệt qua các mẫu không bị thiếu, độ phức tạp tính toán trở nên tuyếnh với số lượng mẫu khác 0, giúp tăng tốc độ lên đến 50 lần so với thuật toán thông thường trên dữ liệu thưa.

Thuật toán Xấp xỉ: Khi dữ liệu quá lớn không thể duyệt qua tất cả các phép chia khả dĩ Thuật toán Tham lam Chính xác, XGBoost sử dụng một thuật toán xấp xỉ. Thuật toán này để xuất một tập hợp hữu hạn các ngưỡng chia ứng cử dựa trên phân vị của phân bố đặc trưng. Sau đó, nó chỉ kiểm tra các phép chia tại các ngưỡng ứng cử này.

Phác thảo Phân vị có Trọng số: Một bước quan trọng trong thuật toán xấp xỉ là làm thế nào để đề xuất các ngưỡng chia một cách hiệu quả. XGBoost đưa ra một thuật toán mới gọi là phác thảo phân vị có trọng số, trong đó các đạo hàm bậc hai  $h_i$  Hessian được sử dụng làm trọng số cho các điểm dữ liệu. Điều này có cơ sở lý thuyết vững chắc vì hàm mục tiêu xấp xỉ  $\tilde{\mathcal{L}}^{(t)}$  chính là một hàm mất mát bình phương có trọng số với nhãn là  $\frac{-g_i}{h_i}$  và trọng số là  $h_i$ .

Các Kỹ thuật Chống Quá khóp Bổ sung: Ngoài hàm mục tiêu chính quy hóa  $\Omega$ , XGBoost còn sử dụng hai kỹ thuật phổ biến khác. Thứ nhất là co lại, tương tự như tốc độ học, trong đó trọng số của mỗi cây mới được thêm vào sẽ được nhân với một hệ số  $\eta < 1$ . Điều này làm giảm ảnh hưởng của mỗi cây riêng lẻ và chừa không gian cho các cây sau này cải thiện mô hình. Thứ hai là Lấy mẫu con theo cột, một kỹ thuật mượn từ Rừng ngẫu nhiên, giúp ngăn ngừa quá khóp thậm chí còn hiệu quả hơn so với lấy mẫu con theo hàng truyền thống.

Tối ưu hóa Hệ thống: XGBoost được thiết kế để có khả năng mở rộng cao. Nó sử dụng một cấu trúc dữ liệu gọi là khối để lưu trữ dữ liệu đã được sắp xếp trước theo từng đặc trưng. Dữ liệu này được nén và có thể được xử lý song song trên nhiều luồng. Hệ thống cũng được tối ưu hóa cho truy cập bộ nhớ đệm và tính toán ngoài lõi, cho phép nó xử lý các tập dữ liệu lên đến hàng tỷ mẫu trên một máy tính để bàn thông thường bằng cách sử dụng hiệu quả ổ đĩa.

### 2.2.2. Thuật toán Random Forest

Rừng ngẫu nhiên là một phương pháp học máy tổ hợp bao gồm một tập hợp các cây quyết định. Sự ra đời của Rừng ngẫu nhiên bắt nguồn từ những cải tiến đáng kể trong độ chính xác phân loại bằng cách phát triển một tập hợp các cây và đưa ra quyết định dựa trên sự bỏ phiếu của số đông[21].

Rừng ngẫu nhiên là một bộ phân loại bao gồm một tập hợp các cây có cấu trúc  $\{h(x, \Theta_k), k = 1, \dots, T\}$ , trong đó  $\Theta_k$  là các vectơ ngẫu nhiên độc lập có cùng phân phối, và mỗi cây sẽ đóng góp một phiếu bầu đơn vị cho lớp phổ biến nhất tại đầu vào  $x$ .

Các yếu tố ngẫu nhiên trong thuật toán này được kế thừa và phát triển từ các phương pháp trước đó như: Bagging - Trong đó mỗi cây được phát triển dựa trên một tập hợp ngẫu nhiên các mẫu từ dữ liệu huấn luyện, Lựa chọn phân chia ngẫu nhiên - Tại mỗi nút, việc phân chia được chọn ngẫu nhiên từ  $K$  phép phân chia tốt nhất, Không gian con ngẫu nhiên - Lựa chọn ngẫu nhiên một tập hợp con các đặc trưng để phát triển mỗi cây. Rừng ngẫu nhiên hội tụ nhờ vào Luật Số Lớn, điều này giải thích tại sao thuật toán này không bị hiện tượng quá khớp khi số lượng cây trong rừng tăng lên.

Mặc dù Rừng ngẫu nhiên thường được biết đến với bài toán phân loại, nhưng nó cũng cực kỳ hiệu quả trong các bài toán hồi quy. Rừng ngẫu nhiên cho hồi quy được hình thành bằng cách phát triển các cây phụ thuộc vào một vectơ ngẫu nhiên  $\Theta$ , sao cho bộ dự báo cây  $h(x, \Theta)$  nhận các giá trị số thay vì nhãn lớp.

Cơ chế dự báo và Sai số tổng quát hóa: Giả sử tập dữ liệu huấn luyện được rút ra độc lập từ phân phối của vectơ ngẫu nhiên  $Y, X$ . Sai số tổng quát hóa bình phương trung bình cho bất kỳ bộ dự báo số  $h(x)$  nào được định nghĩa là:

$$E_{X,Y}(Y - h(X))^2 \quad (13)$$

Khác với bài toán phân loại bộ dự báo của Rừng ngẫu nhiên trong hồi quy được hình thành bằng cách lấy giá trị trung bình của các cây  $h(X, \Theta_k)$  trên  $k$  cây. Khi số lượng cây trong rừng tiến tới vô cùng, sai số tổng quát hóa của rừng sẽ hội tụ theo công thức sau:

$$E_{X,Y} \left( Y - a_{v_k} h(X, \Theta_k) \right)^2 \rightarrow E_{X,Y} (Y - E_\Theta h(X, \Theta))^2 \quad (14)$$

Giá trị bên vế phải của phương trình trên được ký hiệu là  $PE^*(forest)$ , đại diện cho sai số tổng quát hóa của rừng.

Hiệu suất của Rừng ngẫu nhiên trong hồi quy phụ thuộc vào hai yếu tố chính: sai số của từng cây riêng lẻ và sự tương quan giữa các phần dư của chúng.

Gọi sai số tổng quát hóa trung bình của một cây đơn lẻ là  $PE^*(tree)$ , được định nghĩa như sau:

$$PE^*(tree) = E_{\theta}E_{X,Y}(Y - h(X, \theta))^2 \quad (15)$$

Giới hạn trên của sai số rừng được xác định bởi:

$$PE^*(forest) \leq \rho PE^*(tree) \quad (16)$$

Trong đó,  $\rho$  là sự tương quan có trọng số giữa các phần dư  $Y - h(X, \theta)$  và  $Y - h(X, \theta')$  (với  $\theta, \theta'$  là độc lập).

Công thức này chỉ ra yêu cầu cốt lõi để có một rừng hồi quy chính xác: Sự tương quan giữa các phần dư phải thấp và sai số của các cây đơn lẻ phải thấp. Rừng ngẫu nhiên làm giảm sai số trung bình của các cây đơn lẻ bằng một hệ số  $\rho$ . Do đó, quá trình ngẫu nhiên hóa được sử dụng trong thuật toán cần phải hướng tới việc giảm thiểu sự tương quan này.

Trong rừng hồi quy, kỹ thuật lựa chọn đặc trưng ngẫu nhiên được sử dụng kết hợp với bagging. Một sự khác biệt thú vị giữa hồi quy và phân loại trong Rừng ngẫu nhiên là cách phản ứng với số lượng đặc trưng được sử dụng để phân chia tại mỗi nút ký hiệu là  $F$ :

- Trong hồi quy, sự tương quan ( $\bar{p}$ ) tăng khá chậm khi số lượng đặc trưng  $F$  tăng lên.
- Tác động chính của việc tăng  $F$  là làm giảm đáng kể sai số của cây đơn lẻ  $PE^*(tree)$ .

Do đó, trong bài toán hồi quy, để đạt được sai số kiểm tra gần mức tối ưu, cần sử dụng một số lượng đặc trưng  $F$  tương đối lớn để giảm  $PE^*(tree)$ , mặc dù điều này có thể làm tăng nhẹ sự tương quan.

- Nếu số lượng đặc trưng quá nhỏ:  $PE^*(tree)$  trở nên quá lớn, dẫn đến sai số tổng thể tăng.

- Nếu số lượng đặc trưng quá lớn: Sự tương quan tăng lên, cũng làm sai số tổng thể tăng.

Một tính năng quan trọng của Rừng ngẫu nhiên là khả năng tự giám sát sai số, sức mạnh và sự tương quan thông qua ước lượng Out-of-Bag (OOB) mà không cần tập kiểm tra riêng biệt. Trong quá trình Bagging, mỗi cây được huấn luyện trên một tập dữ liệu bootstrap . Khoảng 1/3 số mẫu sẽ không xuất hiện trong tập huấn luyện của một cây cụ thể nào đó, những mẫu này gọi là "OOB". Trong hồi quy, ước lượng OOB cho sai số tổng quát hóa,  $PE^*(tree)$ , và sự tương quan ( $\bar{p}$ ) được tính toán tương tự như trong phân loại. Các ước lượng OOB cung cấp cái nhìn sâu sắc về khả năng dự báo của rừng và là công cụ hữu ích để tinh chỉnh tham số.

Rừng ngẫu nhiên là một công cụ hiệu quả trong dự báo. Nhờ vào Luật Số Lớn, chúng không bị quá khóp. Việc đưa vào đúng loại ngẫu nhiên làm cho chúng trở thành các bộ phân loại và hồi quy chính xác. Cơ sở lý thuyết của thuật toán dựa trên mối quan hệ giữa sức mạnh của từng cây đơn lẻ và sự tương quan giữa chúng: Rừng ngẫu nhiên hoạt động bằng cách giảm sai số thông qua việc giảm sự tương quan giữa các cây trong khi vẫn duy trì sức mạnh dự báo của từng cây. Điều này làm cho Rừng ngẫu nhiên trở thành một phương pháp mạnh mẽ, cạnh tranh tốt với các thuật toán như Boosting nhưng lại không làm thay đổi tập dữ liệu huấn luyện một cách lũy tiến[37].

### 2.2.3. Thuật toán CatBoost

CatBoost là một thuật toán học máy tiên tiến thuộc họ các phương pháp tăng cường độ dốc, được phát triển bởi các nhà nghiên cứu tại Yandex[23]. Thuật toán này được thiết kế đặc biệt để giải quyết hiệu quả các vấn đề liên quan đến dữ liệu có chứa các đặc trưng phân loại, đồng thời khắc phục các hạn chế về sai lệch dự đoán thường gặp trong các thuật toán tăng cường truyền thống. Về mặt nền tảng toán học, mục tiêu của quá trình học là tìm ra một hàm dự đoán  $F$  nhằm cực tiểu hóa hàm mất mát kỳ vọng

$$L(F) := \mathbb{E}L(y, F(x)), \quad (17)$$

trong đó  $L$  là một hàm mất mát trên cặp  $(x, y)$  là các cặp dữ liệu đầu vào và nhãn mục tiêu được lấy từ tập dữ liệu huấn luyện.

Quy trình tăng cường độ dốc xây dựng mô hình một cách tuần tự thông qua chuỗi các xấp xỉ, trong đó mô hình ở bước thứ  $t$  được cập nhật từ bước trước đó theo công thức cộng tính:

$$F^t = F^{t-1} + \alpha h^t. \quad (18)$$

Tại đây,  $\alpha$  là kích thước bước nhảy và  $h^t$  là một bộ dự báo có sở được chọn từ một họ các hàm nhầm xấp xỉ hướng của gradient âm của hàm mất mát.

Một trong những đóng góp kỹ thuật quan trọng nhất của CatBoost nằm ở phương pháp xử lý các biến phân loại, vốn là các biến có giá trị rời rạc không mang tính thứ tự. Các phương pháp truyền thống thường sử dụng kỹ thuật mã hóa one-hot, nhưng phương pháp này không hiệu quả khi số lượng danh mục lớn. Một phương pháp thay thế phổ biến là sử dụng thông kê mục tiêu, tức là thay thế giá trị danh mục bằng một giá trị số đại diện cho kỳ vọng của biến mục tiêu. Cách tiếp cận thông thường, được gọi là thông kê mục tiêu tham lam, ước tính giá trị này bằng cách lấy trung bình các giá trị mục tiêu của các mẫu có cùng danh mục. Công thức tính toán cho thông kê mục tiêu tham lam thường được biểu diễn như sau:

$$\hat{a}_k^i = \frac{\sum_{j=1}^n \mathbf{1}_{\{x_j^{cat}=k\}} \cdot y_j + aP}{\sum_{j=1}^n \mathbf{1}_{\{x_j^{cat}=k\}} + a} \quad (19)$$

Trong công thức trên,  $a$  là một tham số dương đóng vai trò làm tròn để tránh nhiễu khi số lượng mẫu ít,  $P$  là giá trị ưu tiên thường được lấy là trung bình của biến mục tiêu trên toàn bộ tập dữ liệu, và hàm chỉ báo  $I$  nhận giá trị 1 nếu danh mục của mẫu thứ  $j$  giống với mẫu thứ  $k$  đang xét, và bằng 0 nếu ngược lại.

Mặc dù phương pháp này giúp chuyển đổi hiệu quả các biến phân loại thành dạng số, nó gặp phải một vấn đề nghiêm trọng gọi là rò rỉ mục tiêu. Việc sử dụng chính nhãn mục tiêu của một mẫu để tính toán giá trị đặc trưng cho mẫu đó dẫn đến sự sai lệch có điều kiện giữa phân phối của dữ liệu huấn luyện và dữ liệu kiểm tra, gây ra hiện tượng quá khớp và giảm độ chính xác dự đoán trên dữ liệu mới. Để giải quyết triệt để vấn đề rò rỉ mục tiêu này, CatBoost giới thiệu một kỹ thuật mới gọi là thông kê mục tiêu có thứ tự. Kỹ thuật này dựa trên nguyên lý sắp xếp, trong đó thuật toán tạo ra một hoán vị ngẫu nhiên nhân tạo cho các mẫu trong tập dữ liệu huấn luyện. Khi tính toán giá trị thông kê mục tiêu cho một mẫu cụ thể, thuật toán chỉ sử dụng các mẫu xuất hiện trước nó trong thứ tự hoán vị này. Điều này mô phỏng quy trình học trực tuyến, nơi mô hình chỉ có thể học từ dữ liệu lịch sử đã quan sát được, đảm bảo rằng thông tin về biến mục tiêu của chính mẫu hiện tại không bị rò rỉ vào giá trị đặc trưng của nó. Để giảm phương sai do

việc chỉ sử dụng một thứ tự sắp xếp ngẫu nhiên, CatBoost thực hiện quy trình này trên nhiều hoán vị ngẫu nhiên khác nhau của tập dữ liệu trong các bước huấn luyện khác nhau.

Bên cạnh việc cải tiến xử lý biến phân loại, CatBoost cũng giải quyết một vấn đề mang tính hệ thống trong các thuật toán tăng cường độ dốc tiêu chuẩn, đó là hiện tượng dịch chuyển dự đoán. Trong quy trình chuẩn, tại mỗi bước lặp, bộ dự báo cơ sở  $h^t$  được huấn luyện để khớp với gradient âm của hàm mất mát. Thông thường, bộ dự báo này được chọn để cực tiểu hóa sai số bình phương trung bình:

$$h^{t*} = \arg \min_{h \in \mathcal{H}} \mathbb{E}[(-g^t(x, y) - h(x))^2] \quad (20)$$

Trong đó,  $g^t(x, y)$  là đạo hàm của hàm mất mát tại giá trị dự đoán hiện tại. Vấn đề này sinh do gradient được ước tính bằng cách sử dụng cùng một tập dữ liệu đã dùng để huấn luyện mô hình hiện tại. Điều này dẫn đến việc gradient bị chêch về phía hướng của dữ liệu huấn luyện, và do đó, mô hình học được sẽ bị sai lệch so với phân phối thực tế của dữ liệu kiểm tra. CatBoost khắc phục hạn chế này bằng kỹ thuật tăng cường có thứ tự. Thay vì sử dụng một mô hình duy nhất, thuật toán duy trì một tập hợp các mô hình hỗ trợ. Đối với một mẫu thứ k trong hoán vị ngẫu nhiên, mô hình được sử dụng để tính toán phần dư cho nó là mô hình chỉ được huấn luyện trên các mẫu xuất hiện trước k trong hoán vị. Điều này đảm bảo rằng phần dư được tính toán cho mẫu k là không chêch, vì mô hình chưa từng nhìn thấy mẫu này trong quá trình huấn luyện, từ đó giúp quá trình tăng cường hội tụ chính xác hơn tới nghiệm tối ưu thực sự.

Về mặt cấu trúc của các bộ dự báo cơ sở, CatBoost sử dụng một loại cây quyết định đặc biệt gọi là cây quyết định đối xứng hay cây không lăng quen. Khác với các cây quyết định thông thường có thể phát triển tự do và không cân bằng, trong cây đối xứng, cùng một tiêu chí phân chia bao gồm đặc trưng và ngưỡng được áp dụng cho toàn bộ các nút ở cùng một độ sâu của cây. Một cây quyết định  $h(x)$  chia không gian đặc trưng thành các vùng rời rạc  $R_j$  và có thể được biểu diễn dưới dạng tổng trọng số:

$$h(x) = \sum_{j=1}^J b_j \mathbf{1}_{\{x \in R_j\}} \quad (21)$$

Trong đó,  $b_j$  là giá trị dự đoán tại lá thứ  $j$  của cây và  $J$  là tổng số lá. Cấu trúc đối xứng này mang lại nhiều lợi ích quan trọng: nó giúp cây cân bằng hơn, làm giảm nguy cơ quá khớp, và đặc biệt là cho phép thực hiện dự đoán cực kỳ nhanh chóng trong quá trình kiểm tra nhờ cấu trúc lưu trữ và truy xuất hiệu quả. Ngoài ra, CatBoost còn tích

hợp khả năng xử lý các tổ hợp đặc trưng một cách tự động. Thuật toán tìm kiếm và kết hợp các đặc trưng phân loại lại với nhau để tạo ra các đặc trưng mới mạnh mẽ hơn trong quá trình xây dựng cây theo phương thức tham lam.

Tóm lại, CatBoost là một bước tiến đáng kể trong lĩnh vực máy học với các cây quyết định tăng cường độ dốc. Bằng cách giải quyết triệt để các vấn đề về rò rỉ mục tiêu và sai lệch dự đoán thông qua các kỹ thuật dựa trên sắp xếp cùng với việc sử dụng cây đối xứng và xử lý thông minh các biến phân loại, CatBoost cung cấp một giải pháp mạnh mẽ, chính xác và dễ sử dụng cho nhiều bài toán thực tế, đặc biệt là khi dữ liệu có chứa các đặc trưng phân loại phức tạp.

### 2.3. Ước tính hệ số phản xạ bề mặt từ dữ liệu Cyclone GNSS

Khu vực DBSCL có địa hình bằng phẳng, mạng lưới sông ngòi dày đặc và lớp phủ bề mặt chủ yếu là nước, tạo nên độ nhẵn cao đối với tín hiệu GNSS. Do đó, tín hiệu vệ tinh CYGNSS thu được tại đây mang đặc trưng điển hình của cơ chế phản xạ gương.

Để mô hình hóa và phân tích đặc tính bề mặt của khu vực, các tham số đầu vào quan trọng được trích xuất từ dữ liệu Cyclone GNSS được trình bày ở Bảng:

Tham số	Định nghĩa	Đơn vị
power_analog	Công suất định của DDM (công suất thực đo được tại máy thu)	Watt
sp_lat / sp_lon	Vĩ độ và Kinh độ của điểm phản xạ	Độ
rx_to_sp_range	Khoảng cách từ vệ tinh thu đến điểm phản xạ	Mét
tx_to_sp_range	Khoảng cách từ vệ tinh phát đến điểm phản xạ	Mét
sp_inc_angle	Góc tới tại điểm phản xạ	Độ
gps_tx_power_db_w	Công suất phát của vệ tinh GPS	dBW
gps_ant_gain_db_i	Độ lợi ăng-ten phát của vệ tinh GPS	dBi
sp_rx_gain	Độ lợi ăng-ten thu tại hướng của điểm phản xạ	dBi
gps_eirp	Công suất bức xạ đang hướng hiệu dụng	Watt
quality_flags	Cờ đánh dấu chất lượng dữ liệu DDM	

Bảng 1: Các tham số dữ liệu Cyclone GNSS

Dựa trên các tham số được liệt kê trong (Bảng 1), nghiên cứu tiến hành thiết lập mô hình toán học để tính toán các đặc trưng bề mặt. 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 ở pha phát sang phân cực tròn tay trái ở pha thu. Năng lượng phản xạ ( $P_r^c$ ) thu được tuân theo phương trình radar bistatic[38]:

$$P_r^c = \frac{P_t \lambda^2 G_t G_r}{(4\pi)^2 (r_{st} + r_{sr})^2} \Gamma_{rl}(\theta_i) \quad (22)$$

Trong đó,  $\lambda$  là bước sóng của tín hiệu vệ tinh GPS (19 cm);  $\Gamma_{rl}(\theta_i)$  là hệ số phản xạ bờ mặt tại góc tới  $\theta_i$ ;  $P_t$  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 bờ mặt  $P_r^c$  được tính toán thông qua giá trị định 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 bờ mặt  $\Gamma_{rl}$  được tính toán theo công thức:

$$\Gamma_{rl}(\theta_i) = \frac{(P_{DDM} - N)(4\pi)^2 (r_{st} + r_{sr})^2}{P_t G_t G_r \lambda^2} \quad (23)$$

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 định  $P_{DDM}$  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 bờ mặt[39].

Toàn bộ các giá trị trong công thức (23) đều được quy đổi về đơn vị tuyến tính trước khi tính toán. Tín hiệu GNSS ở băng tần L chịu ảnh hưởng đáng kể của các điều kiện bờ mặt, đặc biệt tại các khu vực có mặt nước hoặc vùng đồng bằng ven biển[40].

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 bờ mặt phản xạ.

## 2.4. Các phương pháp đánh giá mô hình

### 2.4.1. Kiểm định chéo

Kiểm định chéo K-Fold là một kỹ thuật được sử dụng rộng rãi và đóng vai trò quan trọng trong lĩnh vực học máy và thống kê để đánh giá hiệu suất của các mô hình dự đoán một cách khách quan và tin cậy. Phương pháp này giải quyết các hạn chế của việc chỉ chia dữ liệu thành hai tập cố định là tập huấn luyện và tập kiểm tra bằng cách chia tập dữ liệu ban đầu thành k phần nhỏ hơn có kích thước xấp xỉ nhau, được gọi là các fold, đảm bảo mỗi mẫu dữ liệu chỉ thuộc về một phần duy nhất. Quá trình đánh giá sau đó

được thực hiện lặp lại k lần, trong đó tại mỗi vòng lặp, một phần dữ liệu khác nhau được giữ lại để làm tập kiểm tra nhằm đánh giá chất lượng mô hình, trong khi các phần còn lại được gộp lại để làm dữ liệu huấn luyện. Kết quả cuối cùng về hiệu suất của mô hình được xác định bằng cách lấy trung bình cộng của các kết quả thu được từ k vòng lặp này, mang lại một ước lượng tổng quát và ít bị thiên lệch hơn về khả năng dự đoán của mô hình trên dữ liệu thực tế[41].

Ưu điểm nổi bật của kiểm định chéo K-Fold nằm ở khả năng tối ưu hóa việc sử dụng dữ liệu, đặc biệt trong các trường hợp lượng dữ liệu thu thập được còn hạn chế. Thay vì phải lãng phí một phần đáng kể dữ liệu chỉ để dành riêng cho việc kiểm tra và không bao giờ được mô hình học hỏi, phương pháp này cho phép toàn bộ các điểm dữ liệu đều có cơ hội được sử dụng cho cả quá trình huấn luyện lẫn đánh giá, giúp mô hình nắm bắt được các đặc điểm của dữ liệu tốt hơn. Bên cạnh đó, do kết quả đánh giá là trung bình của nhiều lần thử nghiệm trên các tập dữ liệu khác nhau, K-Fold giúp giảm thiểu sự phụ thuộc vào cách chia dữ liệu ngẫu nhiên ban đầu, qua đó cung cấp một cái nhìn chính xác hơn về khả năng tổng quát hóa của mô hình và giúp phát hiện sớm các vấn đề như quá khớp. Kỹ thuật này cũng đóng vai trò nền tảng trong việc lựa chọn mô hình, cho phép các nhà nghiên cứu so sánh hiệu quả giữa các thuật toán hoặc các bộ tham số khác nhau một cách công bằng để tìm ra cấu hình tối ưu nhất cho bài toán cụ thể[41].

#### 2.4.2. Sai số tuyệt đối trung bình (MAE)

Sai số tuyệt đối trung bình (Mean absolute error – MAE) là một chỉ số dùng để đo lường mức độ trung bình của các lỗi trong một tập hợp các dự đoán, mà không xem xét hướng của chúng[42].

MAE được tính bởi công thức:

$$MAE = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t| \quad (21)$$

Trong đó,  $\hat{Y}$  là vector của n giá trị dự báo và  $Y$  là vector của n giá trị quan sát được. Tức là MAE là trung bình của trị tuyệt đối các sai số.

#### 2.4.3. Sai số bình phương trung bình (RMSE)

Sai số bình phương trung bình (Root Mean Square Error – RMSE) là một trong những chỉ số thống kê được sử dụng rộng rãi nhất để đánh giá hiệu suất của các mô hình dự báo, đặc biệt là trong các bài toán hồi quy. Chỉ số này đo lường mức độ chênh lệch

giữa các giá trị dự đoán bởi mô hình và các giá trị thực tế quan sát được. Về mặt toán học, RMSE được tính bằng căn bậc hai của trung bình cộng các bình phương sai số giữa giá trị dự báo và giá trị thực tế. Công thức tính toán thường bao gồm việc lấy hiệu số giữa giá trị dự đoán và giá trị thực tại mỗi điểm dữ liệu, bình phương hiệu số này để loại bỏ dấu âm và tăng trọng số cho các sai số lớn, sau đó tính trung bình các bình phương này và cuối cùng là khai căn bậc hai của kết quả[42].

RMSE được tính bởi công thức:

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (22)$$

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. Giá trị của RMSE càng thấp thì mô hình càng có độ chính xác cao, và ngược lại, giá trị càng lớn cho thấy sự chênh lệch đáng kể giữa dự báo và thực tế. Do tính chất của phép bình phương trong công thức, RMSE rất nhạy cảm với các sai số lớn. Điều này có nghĩa là chỉ cần một vài điểm dữ liệu có sai số dự báo rất lớn cũng có thể làm tăng đáng kể giá trị RMSE của toàn bộ mô hình.

#### 2.4.4. Hệ số tương quan R

Hệ số tương quan Pearson (R) là một chỉ số thống kê dùng để đo lường độ mạnh của mối liên hệ tuyến tính giữa hai biến số. Để tính được hệ số này, dữ liệu đầu vào thường yêu cầu hai biến số phải là biến định lượng và ít nhất một biến phải có phân phối chuẩn[43]. Hệ số này dao động trong khoảng từ -1 đến +1, trong đó giá trị +1 biểu thị tương quan dương hoàn hảo giá trị -1 biểu thị tương quan âm hoàn hảo, và giá trị 0 cho thấy không có tương quan tuyến tính giữa hai biến.

Hệ số tương quan R được tính bởi công thức:

$$R = \sum_{i=1}^n \frac{(x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (23)$$

Trong đó,  $x_i$  là giá trị của biến độc lập tại quan sát thứ  $i$ ,  $y_i$  là giá trị của biến phụ thuộc tại quan sát thứ  $i$ ,  $\bar{x}$  và  $\bar{y}$  lần lượt là giá trị trung bình của các tập  $x$  và  $y$ , và  $n$  là tổng số mẫu quan sát.

## **CHƯƠNG 3: Kết quả thực nghiệm ứng dụng viễn thám, GIS và học máy để thành lập bản đồ xâm nhập tại Đồng Bằng Sông Cửu Long**

### **3.1. Đặc điểm tự nhiên và kinh tế - xã hội tại Đồng Bằng Sông Cửu Long**

Đồng bằng sông Cửu Long nằm ở cực Nam của Việt Nam, là vùng hạ lưu của sông Mekong một trong những hệ thống sông lớn nhất thế giới. Với tổng diện tích tự nhiên khoảng 40.577 km<sup>2</sup>, chiếm khoảng 12% diện tích cả nước, khu vực này bao gồm 13 tỉnh, thành phố trực thuộc Trung ương: Long An, Tiền Giang, Bến Tre, Trà Vinh, Vĩnh Long, Đồng Tháp, An Giang, Kiên Giang, Cần Thơ, Hậu Giang, Sóc Trăng, Bạc Liêu và Cà Mau. Đây là vùng đất có ý nghĩa chiến lược về an ninh lương thực quốc gia, đa dạng sinh học và phát triển kinh tế bền vững. Tuy nhiên, đến năm 2025, khu vực này đang đổi mới với những thách thức to lớn từ biến đổi khí hậu, nước biển dâng và sự thay đổi dòng chảy từ thượng nguồn[44].

#### **3.1.1. Đặc điểm tự nhiên**

Về địa hình và địa mạo, DBSCL có đặc điểm địa hình tương đối bằng phẳng và thấp, với độ cao trung bình phổ biến dao động từ 0,7m đến 1,2m so với mực nước biển. Cấu trúc địa hình có xu hướng thấp dần từ Tây Bắc xuống Đông Nam và từ các giòng đất ven sông vào nội đồng. Cụ thể, khu vực dọc biên giới Campuchia có địa hình cao hơn, trung bình từ 2,0m đến 4,0m; vùng giữa đồng bằng có độ cao trung bình từ 1,0m đến 1,5m; trong khi các khu vực duyên hải ven biển có độ cao thấp nhất, chỉ từ 0,3m đến 0,7m. Tuy nhiên, tại ven biển Đông lại tồn tại các giòng cát hình cánh cung chạy dọc bờ biển với cao độ khá cao, từ 2,0m đến 3,0m, tạo nên các đê chắn sóng tự nhiên nhưng cũng gây khó khăn cho việc thoát lũ. Đặc điểm địa hình thấp và bằng phẳng này là yếu tố chính khiến mức độ nhạy cảm với ngập lụt của vùng rất cao khi có triều cường hoặc nước biển dâng. Thổ nhưỡng của vùng rất phức tạp và đa dạng, ảnh hưởng trực tiếp đến khả năng thấm nước và bề mặt lớp phủ. Đất đai được chia thành ba nhóm chính gồm đất phù sa sông chiếm khoảng 30% diện tích tự nhiên, tập trung ở vùng trung tâm dọc sông Tiền và sông Hậu; đất phèn chiếm diện tích lớn nhất khoảng 40%, tập trung chủ yếu tại các vùng trũng như Đồng Tháp Mười, Tứ giác Long Xuyên và vùng trũng trung tâm Bán đảo Cà Mau; và đất nhiễm mặn chiếm khoảng 19%, phân bố dọc các dải ven biển Đông và vịnh Thái Lan. Nền đất của vùng chủ yếu là đất yếu, gây nhiều khó khăn và tốn kém cho việc xây dựng cơ sở hạ tầng giao thông và đê điều chống lũ [44].

Hệ thống sông ngòi là yếu tố quan trọng nhất chi phối chế độ ngập lụt của vùng. Mạng lưới sông kênh rạch chằng chịt với tổng chiều dài các sông lớn nhỏ khoảng 1.700

km và hệ thống kênh đào dày đặc, mật độ kênh rạch trung bình đạt 20–30 m/ha, chiếm tới 19% diện tích toàn vùng . Hai trực thoát nước chính là sông Tiền và sông Hậu phân lũ ra biển qua các cửa sông lớn. Chế độ thủy văn chịu tác động kép của lũ từ thượng nguồn và thủy triều. Mùa lũ thường kéo dài từ tháng 6 đến tháng 11, làm ngập từ 35-48% diện tích toàn vùng, với độ sâu ngập tại các vùng trũng như Đồng Tháp Mười có thể lên đến 2,5m - 4,0m . Chế độ bão nhiệt triều biển Đông có biên độ lớn từ 3,0m đến 3,5m và nhiệt triều biển Tây biên độ nhỏ hơn từ 0,8m đến 1,2m xâm nhập sâu vào nội đồng, làm giảm khả năng thoát lũ và gây ngập triều cho các đô thị ven sông[44].

Đáng chú ý, tình trạng xâm nhập mặn đang trở thành vấn đề cấp thiết hàng đầu, tác động sâu sắc đến cấu trúc tự nhiên và kinh tế - xã hội của vùng. Dưới tác động của biến đổi khí hậu, nước biển dâng và sự suy giảm dòng chảy từ thượng nguồn sông Mekong trong mùa khô, ranh mặn 4g/l có thể xâm nhập sâu vào đất liền từ 20-65 km, ảnh hưởng trực tiếp đến khoảng 45% diện tích toàn vùng, đặc biệt nghiêm trọng tại bán đảo Cà Mau và các tỉnh ven biển . Xâm nhập mặn làm ô nhiễm nguồn nước mặt, gây khó khăn lớn cho việc cấp nước sinh hoạt và tưới tiêu, đe dọa sức khỏe cộng đồng và buộc người dân phải gia tăng khai thác nước ngầm, dẫn đến nguy cơ sụt lún đất và nhiễm mặn lan sâu vào các tầng chúa nước phong phú . Về mặt kinh tế, mặn xâm nhập sâu làm giảm diện tích đất canh tác lúa và cây ăn trái nước ngọt, gây thiệt hại nghiêm trọng về năng suất như trong đợt hạn mặn lịch sử năm 2019-2020. Tuy nhiên, hiện tượng này cũng thúc đẩy sự chuyển dịch cơ cấu sản xuất, tạo điều kiện phát triển các mô hình kinh tế thích ứng như nuôi tôm nước lợ, nuôi trồng thủy sản kết hợp rừng ngập mặn, đòi hỏi công tác quy hoạch phải chuyển từ tư duy chống mặn sang thích ứng và kiểm soát mặn[44].

Thổ nhưỡng của vùng rất phức tạp và đa dạng, ảnh hưởng trực tiếp đến khả năng thấm nước và bề mặt lớp phủ. Đất đai được chia thành ba nhóm chính gồm đất phù sa sông chiếm khoảng 30% diện tích tự nhiên, tập trung ở vùng trung tâm dọc sông Tiền và sông Hậu; đất phèn chiếm diện tích lớn nhất khoảng 40%, tập trung chủ yếu tại các vùng trũng như Đồng Tháp Mười, Tứ giác Long Xuyên và vùng trũng trung tâm Bán đảo Cà Mau; và đất nhiễm mặn chiếm khoảng 19%, phân bố dọc theo các dải ven biển Đông và vịnh Thái Lan. Nền đất của vùng chủ yếu là đất yếu, gây nhiều khó khăn và tốn kém cho việc xây dựng cơ sở hạ tầng giao thông và đê điều chống lũ[44].

Xâm nhập mặn tại ĐBSCL diễn biến phức tạp do sự cộng hưởng chặt chẽ của ba yếu tố tự nhiên. Thứ nhất, địa hình thấp (phổ biến 0,7-1,2m) và bằng phẳng làm giảm lực cản thủy lực, khiến nước biển dễ dàng lấn sâu khi áp lực nước ngọt từ sông Mekong

suy giảm. Thứ hai, mạng lưới kênh rạch dày đặc (20–30 m/ha) đóng vai trò như các mao mạch dẫn truyền mặn nhanh chóng vào nội đồng dưới tác động của triều cường biển độ lớn. Thứ ba, cấu trúc đất phèn và nền đất yếu không chỉ dễ thâm thấu, lưu giữ muối khó rửa trôi mà còn chịu tác động kép của sụt lún, tạo thành vòng luẩn quẩn làm mất dần không gian trữ nước ngọt tự nhiên của vùng.

### 3.1.2. Đặc điểm kinh tế - xã hội

Về dân cư và xã hội, tính đến năm 2024, dân số toàn vùng khoảng 17,59 triệu người. Mặc dù là vùng đông dân, nhưng tỷ lệ đô thị hóa của Đồng bằng sông Cửu Long còn thấp, thấp hơn mức trung bình của cả nước là 33%. Mạng lưới đô thị phân bố phân tán, bám dọc theo các trục sông, kênh rạch và quốc lộ, hình thành nên cấu trúc đô thị nông nghiệp hay cư dân sông nước đặc thù. Đáng chú ý, vùng đang đổi mới với tình trạng di cư lao động lớn, với tỷ suất xuất cư bình quân khoảng 6,7%/năm, chủ yếu là lao động trẻ di chuyển về Thành phố Hồ Chí Minh và vùng Đông Nam Bộ để tìm kiếm việc làm, dẫn đến sự thiếu hụt lực lượng lao động tại chỗ[44].

Về kinh tế, Đồng bằng sông Cửu Long là vùng trọng điểm sản xuất lương thực, thực phẩm của cả nước. Nông nghiệp đóng vai trò chủ đạo, đóng góp 50% sản lượng lúa, 65% sản lượng nuôi trồng thủy sản và 70% trái cây của cả nước; đồng thời đóng góp 95% lượng gạo xuất khẩu và 60% sản lượng cá xuất khẩu. Công nghiệp của vùng chủ yếu là công nghiệp chế biến nông thủy sản và năng lượng, bao gồm các cụm khí - điện - đạm tại Cà Mau và nhiệt điện tại Trà Vinh, Sóc Trăng. Mặc dù có tiềm năng lớn về du lịch sinh thái sông nước và vận tải thủy, nhưng lĩnh vực thương mại và dịch vụ chưa phát triển tương xứng, hệ thống logistics còn yếu kém khi 80% hàng hóa xuất nhập khẩu vẫn phải trung chuyển qua các cảng tại Thành phố Hồ Chí Minh và Bà Rịa - Vũng Tàu.

Về cơ sở hạ tầng, mạng lưới giao thông đường bộ đang được đầu tư nhưng vẫn còn hạn chế do nền đất yếu và chi phí xây dựng cao. Trục xương sống là Quốc lộ 1A và các tuyến N1, N2, Quản Lộ - Phụng Hiệp, cùng với các tuyến cao tốc đang dần hình thành. Giao thông thủy đóng vai trò đặc biệt quan trọng, chiếm gần 70% tổng lượng hàng hóa vận chuyển của vùng, tuy nhiên hệ thống cảng biển nước sâu còn thiếu và các luồng lạch như luồng Định An, Quan Chánh Bố thường xuyên bị bồi lắng, hạn chế tàu trọng tải lớn ra vào. Hệ thống thủy lợi với hơn 20.000 km đê bao và công đập đã được xây dựng khá dày đặc để kiểm soát lũ và ngăn mặn, tuy nhiên việc phát triển đê bao khép kín triệt để để sản xuất lúa vụ 3 đã làm mất không gian trữ lũ tự nhiên, gây gia tăng ngập lụt cho các khu vực lân cận và làm suy thoái đất đai. Những đặc điểm về

hạ tầng và phân bố dân cư ven sông này làm tăng tính dễ bị tổn thương của khu vực trước các tai biến thiên nhiên như lũ lụt và sạt lở đất.

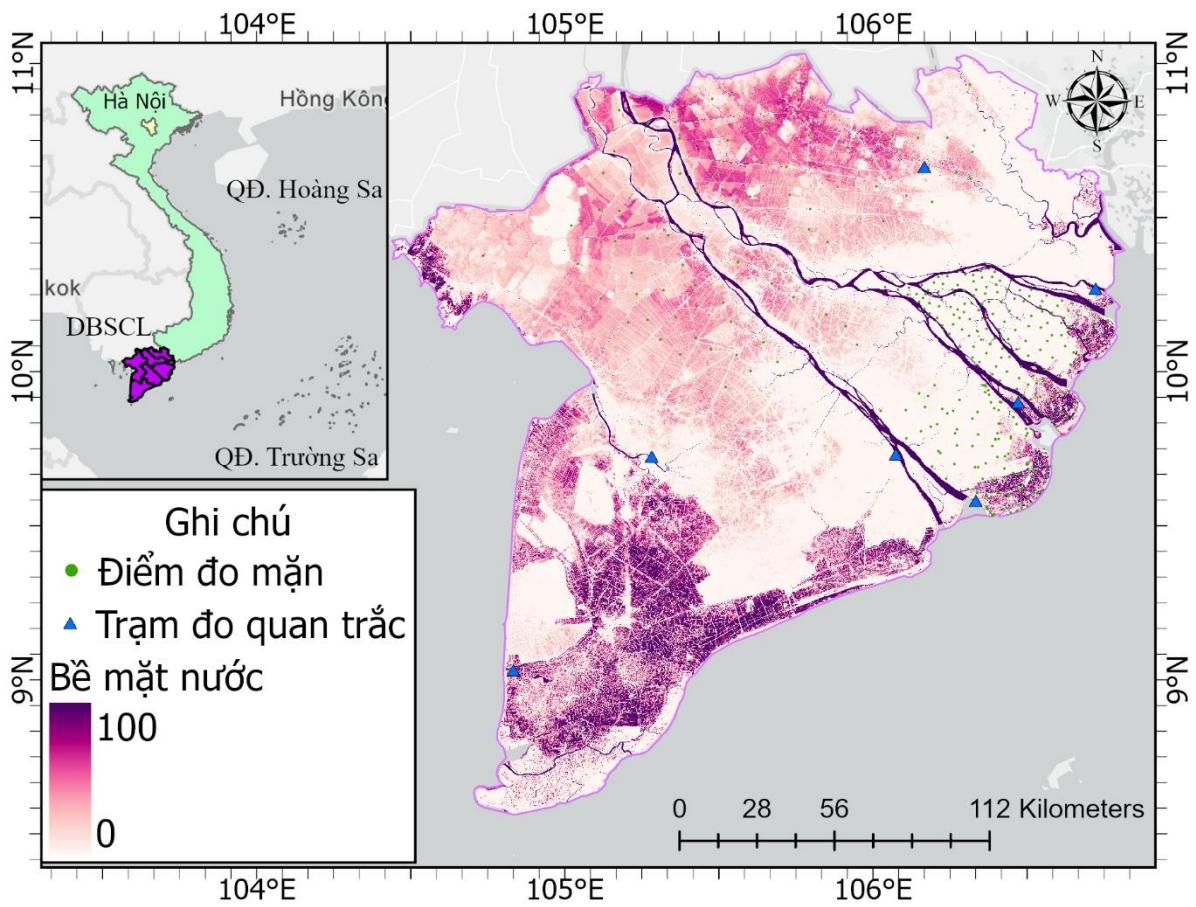
Các đặc điểm kinh tế - xã hội tại ĐBSCL đang là nhân tố chủ quan làm gia tăng tình trạng xâm nhập mặn. Việc thâm canh lúa vụ 3 và hệ thống đê bao khép kín đã triệt tiêu không gian trữ lũ, làm mất nguồn nước ngọt quan trọng để đầy mặn trong mùa khô. Song song đó, tình trạng khai thác nước ngầm quá mức cho sinh hoạt và thủy sản gây sụt lún nền đất, kết hợp với mạng lưới kênh rạch giao thông liên tục được khơi thông đã vô tình hạ thấp cao độ và tạo đường dẫn cho nước mặn lấn sâu vào nội đồng.

### 3.2. Cơ sở dữ liệu

Để đảm bảo tính chính xác và độ tin cậy cho quá trình xây dựng mô hình giám sát xâm nhập mặn bằng học máy, việc xây dựng một cơ sở dữ liệu đồng bộ và chất lượng là yêu cầu tiên quyết. Trong nghiên cứu này, cơ sở dữ liệu được thiết lập từ hai nguồn chính bao gồm dữ liệu khảo sát thực địa và dữ liệu thu thập từ công nghệ viễn thám. Dữ liệu thu thập từ công nghệ viễn thám được chia thành bốn nhóm chính dựa trên chức năng và bản chất của chúng: dữ liệu địa hình, dữ liệu khí tượng, dữ liệu thổ nhưỡng và các chỉ số viễn thám. Bảng dữ liệu được trình bày ở Bảng:

STT	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	SRTM	30 m
12	Khoảng cách đến biển	-	1 km
13	Độ ẩm đất	SMAP L3	9 km
14	Nhiệt độ bề mặt	MYD11A1	1 km
15	Lớp phủ bề mặt	MCD12Q1	500 m
16	Hàm lượng cát	SoilGrids	250 m
17	Hàm lượng sét	SoilGrids	250 m
18	Khối lượng riêng đất	SoilGrids	250 m
19	Điểm đo mặn thực tế	Tổng hợp	Độ sâu 0-30 cm
20	Trạm quan trắc	Tổng hợp	-

Bảng 2: Dữ liệu sử dụng trong nghiên cứu



Hình 6: Khu vực nghiên cứu và phân bố của điểm 330 điểm đo mặn tại khu vực Đồng Bằng Sông Cửu Long

### 3.2.1. Dữ liệu vệ tinh CYGNSS

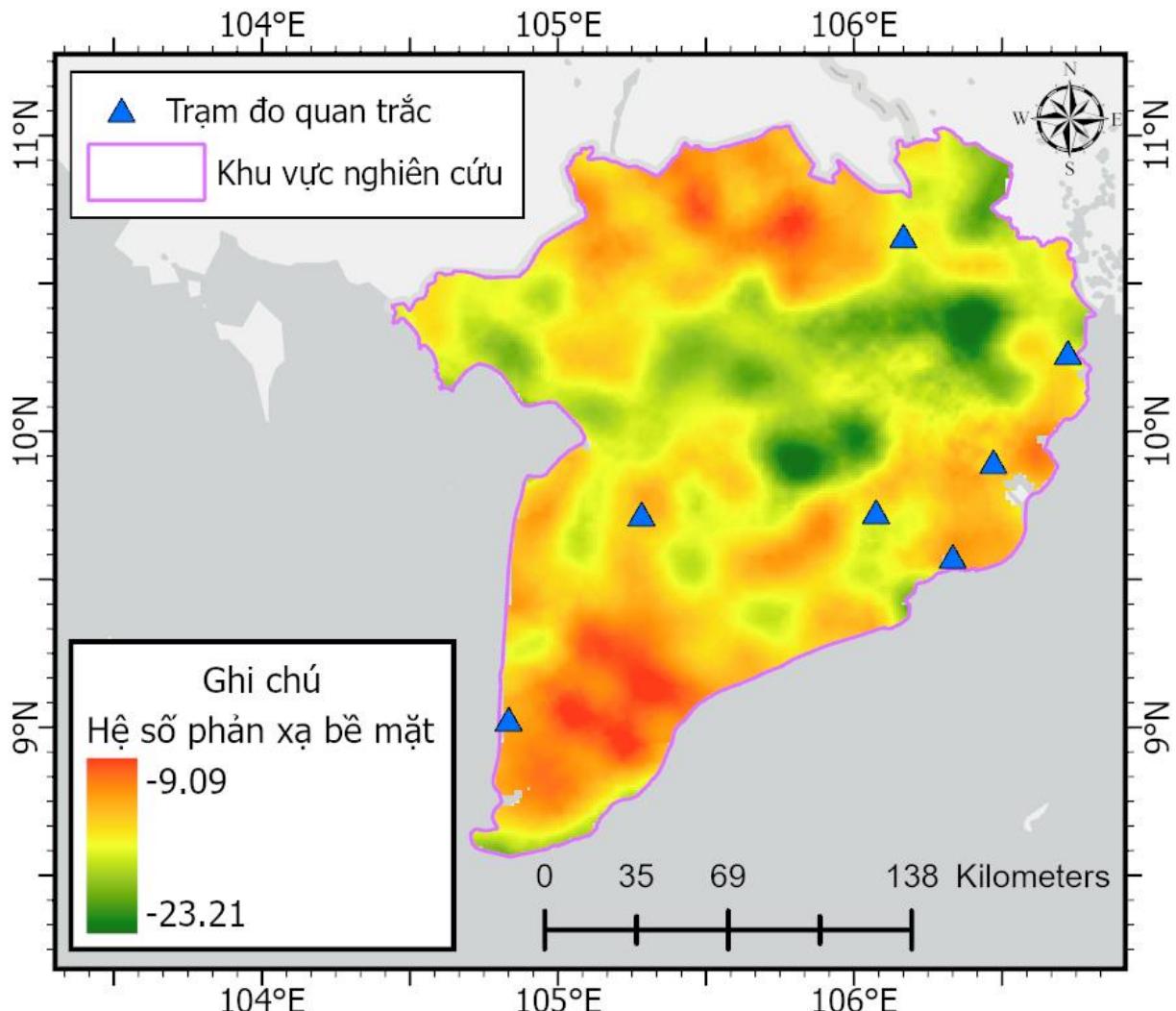
Dữ liệu CYGNSS Level 1 phiên bản 3.1 được sử dụng trong nghiên cứu này được thu thập trong giai đoạn từ tháng 1 đến tháng 5 năm 2025 và được tải về từ cổng dữ liệu NASA PODAAC. Sau khi thu thập, dữ liệu được tiến hành tiền xử lý nhằm đảm bảo chất lượng và độ tin cậy cho các bước phân tích tiếp theo. 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 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[45]. Tương tự, với biến quality\_flags\_2, các

phép đo bị loại khi tồn tại các lỗi như ăng-ten DDMI, lỗi tính toán nguồn, hoặc hệ số khuếch đại thấp ở góc thiên đỉnh thấp[45].

Việc áp dụng các bước tiền xử lý này giúp loại bỏ nhiều 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 xâm nhập mặn trong nghiên cứu.

Sau khi tiền xử lý, các biến của dữ liệu được lấy ra ở Bảng 1 (mục 2.3) để tính toán hệ số phản xạ bề mặt theo công thức (22)



Hình 7: Hệ số phản xạ bề mặt từ dữ liệu CYGNSS trung bình tháng 1 đến tháng 5 năm 2025

### 3.2.2. Chỉ số viễn thám

Đối với xâm nhập mặn nói chung và tại Đồng bằng Sông Cửu Long nói riêng, nguồn dữ liệu viễn thám đóng vai trò quan trọng trong việc mô tả hiện trạng bề mặt, giám sát biến động và xây dựng các mô hình dự báo. Toàn bộ dữ liệu phục vụ nghiên cứu được thu thập và xử lý từ nền tảng Google Earth Engine (GEE).

Nguồn dữ liệu chính được sử dụng trong nghiên cứu là sản phẩm MOD09GA, thuộc bộ ảnh phản xạ bề mặt của vệ tinh MODIS, với độ phân giải 500 m và chu kỳ lặp hàng ngày. Sản phẩm này cung cấp thông tin về phản xạ phổ đã được hiệu chỉnh khí quyển, phù hợp để tính toán các chỉ số liên quan đến thảm thực vật, tình trạng ẩm và dấu hiệu xâm nhập mặn. Từ sản phẩm MOD09GA, nhiều lớp thông tin chuyên đề đã được trích xuất và xây dựng, bao gồm chỉ số thực vật NDVI, chỉ số mặn NDSI, các chỉ số mặn tổng hợp SI1, SI2, SI3, SI4 và SI5, cùng hai băng phổ SWIR 1 và SWIR 2. Đây là những chỉ số phản ánh trực tiếp các đặc tính của bề mặt liên quan đến quá trình xâm nhập mặn, đồng thời có thể kết hợp linh hoạt trong mô hình dự báo không gian để nâng cao độ chính xác.

Chỉ số thực vật NDVI được tính toán nhằm mô tả mức độ phát triển của thảm thực vật và khả năng phản ứng của cây trồng trước sự thay đổi độ mặn trong đất và nước. Các vùng bị ảnh hưởng bởi xâm nhập mặn thường có sự suy giảm độ che phủ thực vật hoặc giảm sức sống rõ rệt, do đó NDVI là một chỉ số quan trọng giúp nhận dạng các khu vực nhạy cảm hoặc chịu tác động nặng nề.

$$NDVI = \frac{(NIR - RED)}{(NIR + RED)} \quad (24)$$

Tiếp theo, chỉ số mặn NDSI được trích xuất nhằm mô phỏng tín hiệu đặc trưng của các vùng bề mặt có khả năng bị ảnh hưởng bởi mặn. Cơ sở của chỉ số này nằm ở sự tương phản tín hiệu phản xạ giữa các băng phổ nhạy cảm với muối hòa tan.

$$NDSI = \frac{(RED - NIR)}{(RED + NIR)} \quad (25)$$

Để tăng cường độ chính xác và khả năng nhận diện xâm nhập mặn theo nhiều khía cạnh khác nhau, năm chỉ số mặn SI1 đến SI5 được xây dựng dựa trên các tổ hợp băng phổ khác nhau của MOD09GA. Các chỉ số này phản ánh nhiều góc độ thay đổi của độ ẩm, cấu trúc bề mặt, thảm thực vật và vật chất hòa tan. Việc sử dụng đồng thời nhiều chỉ số giúp giảm thiểu sai số do nhiều khí quyển, độ đục của nước hay yếu tố thảm phủ, từ đó mang lại bộ dữ liệu đa dạng và ổn định hơn trong mô hình phân tích.

$$SI1 = \sqrt{GREEN \times RED} \quad (26)$$

$$SI2 = \sqrt{NIR \times RED} \quad (27)$$

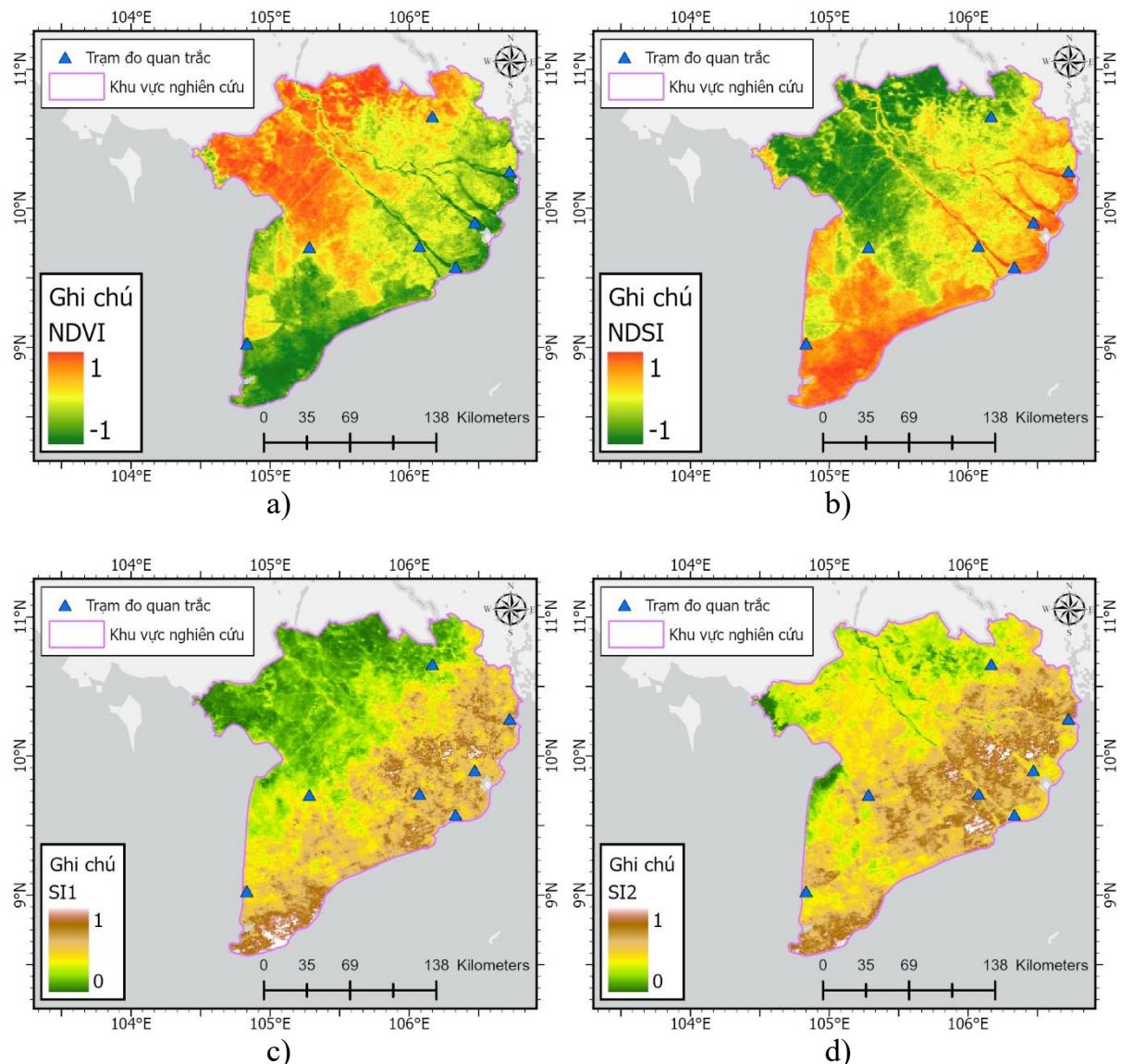
$$SI3 = \sqrt{(GREEN)^2 \times (RED)^2 \times (NIR)^2} \quad (28)$$

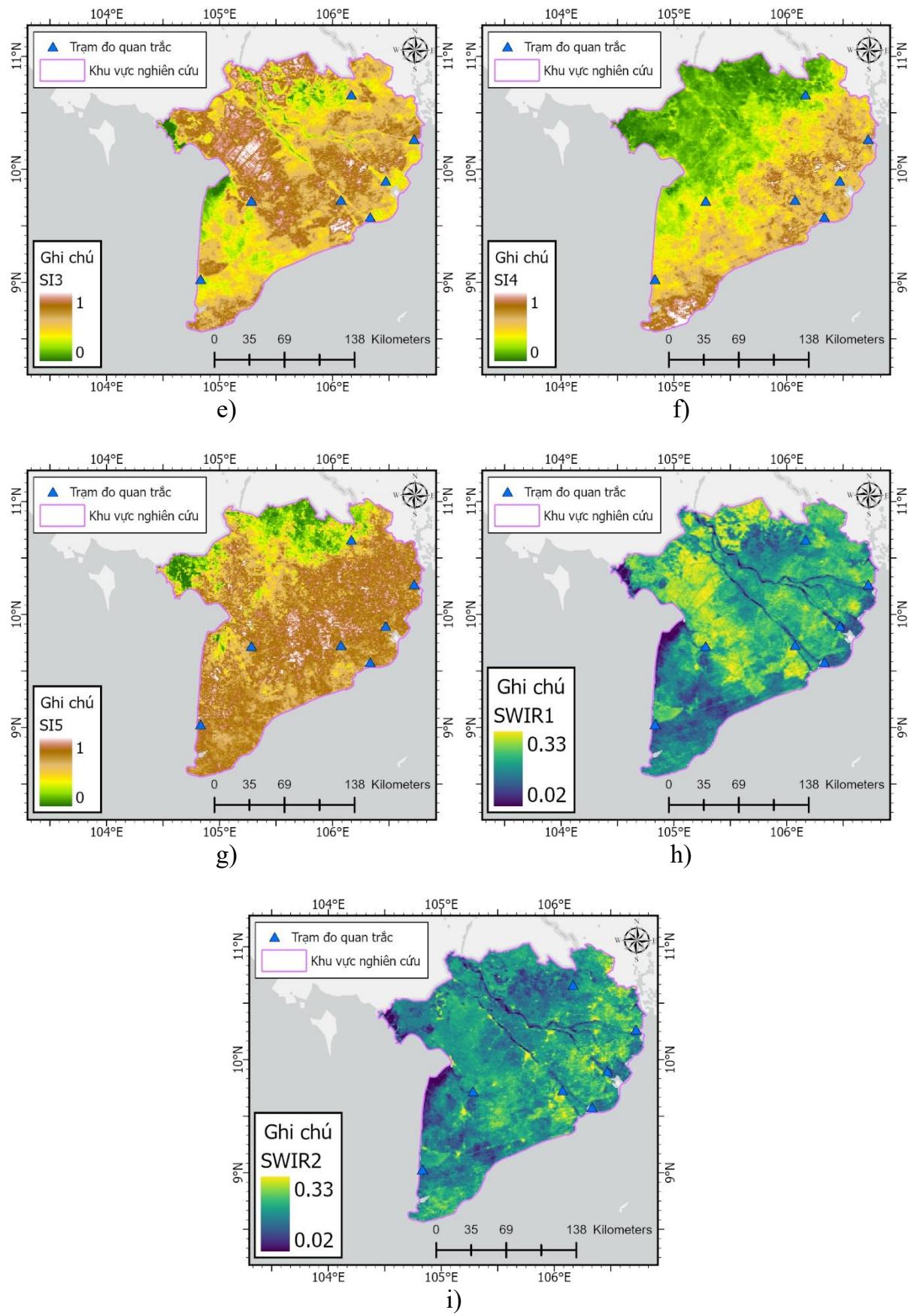
$$SI4 = \sqrt{(GREEN)^2 \times (RED)^2} \quad (29)$$

$$SI5 = \frac{BLUE}{RED} \quad (30)$$

Việc sử dụng các lớp dữ liệu viễn thám mang lại bộ dữ liệu ổn định, chi tiết và có độ phủ rộng trên toàn vùng Đồng bằng Sông Cửu Long. Các chỉ số NDVI, NDSI, SI1–SI5 và SWIR giúp mô hình không chỉ mô tả được trạng thái hiện tại của quá trình xâm nhập mặn mà còn hỗ trợ phân tích xu hướng theo thời gian.

Hình ảnh bộ dữ liệu chỉ số viễn thám được trình bày dưới dạng bản đồ:





Hình 8: Bộ dữ liệu chỉ số viễn thám dưới dạng bản đồ: a) Chỉ số thực vật NDVI; b) Chỉ số mặn NDSI; c)-g) Chỉ số mặn SII-SI5; h) Băng phổ SWIR1; i) Băng phổ SWIR2

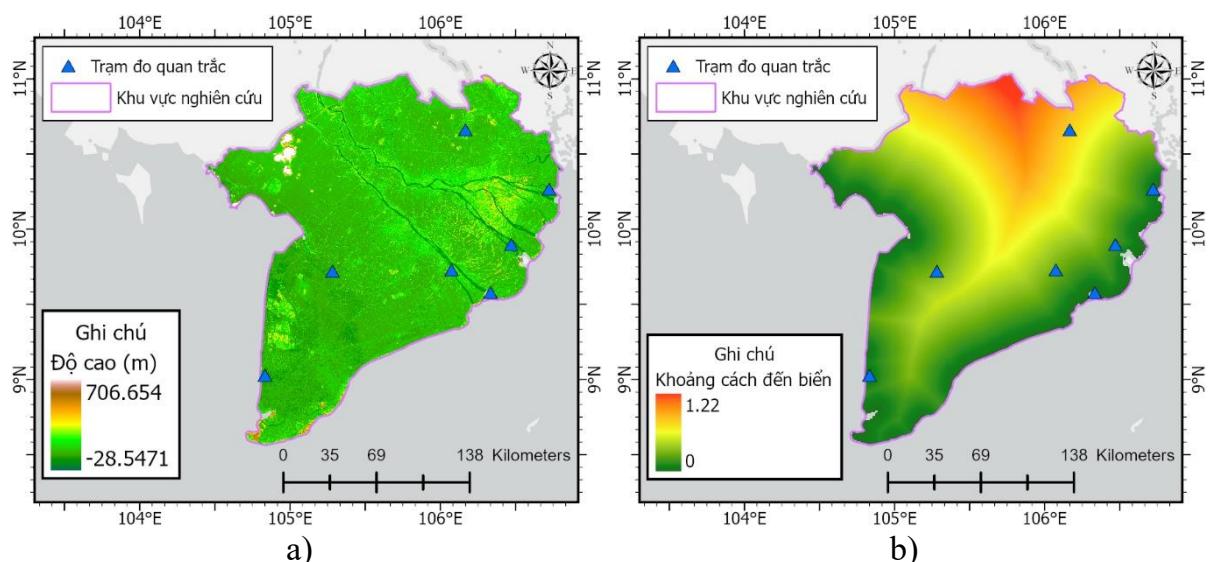
### 3.2.3. Dữ liệu địa hình

Trong nghiên cứu về giám sát xâm nhập mặn và các yếu tố môi trường liên quan tại khu vực Đồng bằng sông Cửu Long, dữ liệu địa hình đóng vai trò nền tảng quan trọng. Yếu tố này không chỉ cung cấp thông tin về hình thái bề mặt mà còn là cơ sở để tính toán các tham số dẫn xuất, ảnh hưởng trực tiếp đến sự phân bố dòng chảy và khả năng xâm nhập của nước mặn vào sâu trong đất liền.

Đối với khu vực nghiên cứu rộng lớn như Đồng bằng sông Cửu Long, dữ liệu DEM thường được khai thác từ các nguồn dữ liệu vệ tinh mở như SRTM với độ phân giải không gian phổ biến là 30 mét. Mặc dù địa hình của vùng đồng bằng này tương đối bằng phẳng, nhưng những sự chênh lệch nhỏ về độ cao lại có ý nghĩa quyết định đối với sự di chuyển của nước và muối. Các khu vực có độ cao thấp hơn thường dễ bị tổn thương hơn trước tác động của triều cường và xâm nhập mặn, trong khi các khu vực cao hơn có khả năng thoát nước tốt hơn.

Khoảng cách đến biển là yếu tố cực kỳ quan trọng đối với xâm nhập mặn tại Đồng bằng sông Cửu Long. Xâm nhập mặn ở đây chủ yếu do nước biển tiến sâu vào đất liền qua các cửa sông và hệ thống kênh rạch chằng chịt dưới tác động của thủy triều. Do đó, độ mặn thường có xu hướng giảm dần khi đi từ biển vào đất liền và từ sông chính vào nội đồng.

Hình ảnh bộ dữ liệu địa hình được trình bày dưới dạng bản đồ:



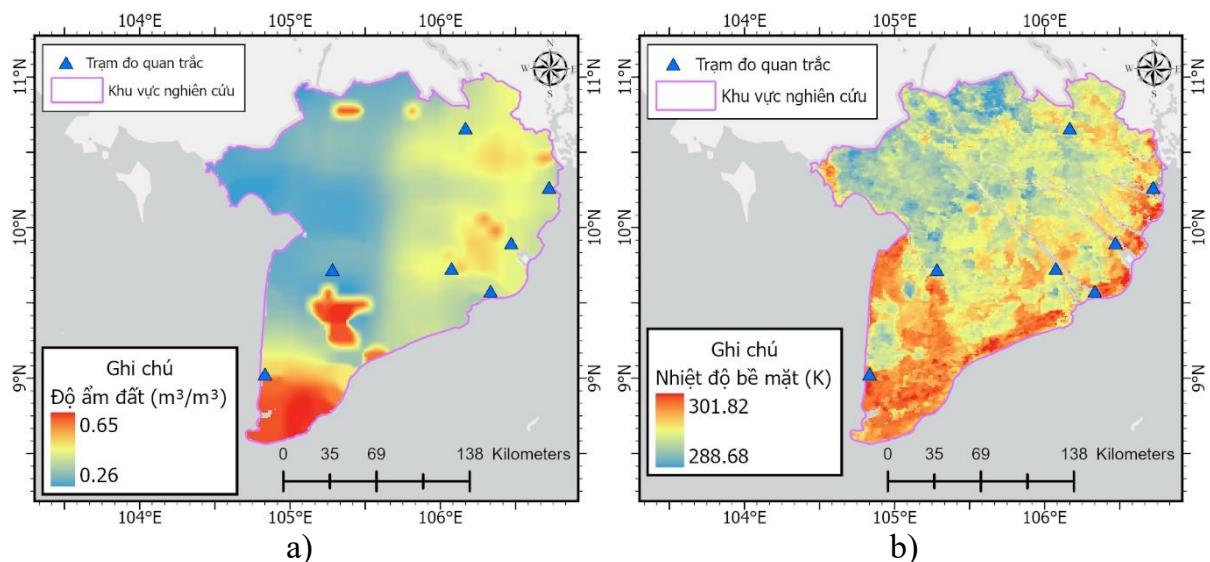
Hình 9: Bộ dữ liệu địa hình; a) Độ cao; b) Khoảng cách đến biển

### 3.2.4. Dữ liệu khí tượng

Bên cạnh dữ liệu địa hình và các chỉ số viễn thám, nghiên cứu còn bổ sung các nguồn dữ liệu khí tượng như độ ẩm đất và nhiệt độ bề mặt. Dữ liệu độ ẩm đất được thu thập từ sản phẩm SMAP cấp độ 3 với độ phân giải 9 km. Độ ẩm đất phản ánh lượng nước hiện có trong tầng đất mặt và là một yếu tố môi trường có vai trò quan trọng trong nghiên cứu xâm nhập mặn tại Đồng bằng Sông Cửu Long. Khi độ ẩm đất giảm, đặc biệt trong thời kỳ khô hạn kéo dài, dòng chảy sông ngòi suy giảm và lượng nước ngọt dự trữ giảm mạnh, làm tăng nguy cơ mặn tiến sâu vào nội đồng. Sự biến đổi không gian và thời gian của độ ẩm đất góp phần thể hiện mức độ cảng thẳng thủy văn của khu vực, đồng thời hỗ trợ xác định các vùng dễ bị tổn thương trước sự lan truyền của các nêm mặn.

Dữ liệu nhiệt độ bề mặt được lấy từ sản phẩm MYD11A1 với độ phân giải 1 km. Đây là dữ liệu phản ánh trạng thái năng lượng bề mặt và mức độ bốc hơi nước của khu vực. Tại Đồng bằng Sông Cửu Long, nhiệt độ bề mặt tăng cao trong mùa khô dẫn tới quá trình bốc hơi mạnh, làm giảm mực nước mặt và gây tích tụ muối trong đất và kênh rạch. Do sự liên hệ chặt chẽ giữa nhiệt độ cao, hạn hán và xâm nhập mặn, chỉ tiêu này giúp mô hình nhận diện các vùng có điều kiện vi khí hậu khắc nghiệt, nơi mặn hóa xảy ra sớm và kéo dài hơn trung bình toàn vùng. Nhiệt độ bề mặt vì vậy là thành phần không thể thiếu trong việc đánh giá mức độ và tốc độ lan truyền mặn, đặc biệt khi kết hợp cùng dữ liệu độ ẩm đất và các chỉ số mặn từ ảnh vệ tinh.

Hình ảnh bộ dữ liệu khí tượng được trình bày dưới dạng bản đồ:



Hình 10: Bộ dữ liệu khí tượng được trình bày dưới dạng bản đồ; a) Độ ẩm đất; b) Nhiệt độ bề mặt

### 3.2.5. Dữ liệu thổ nhưỡng

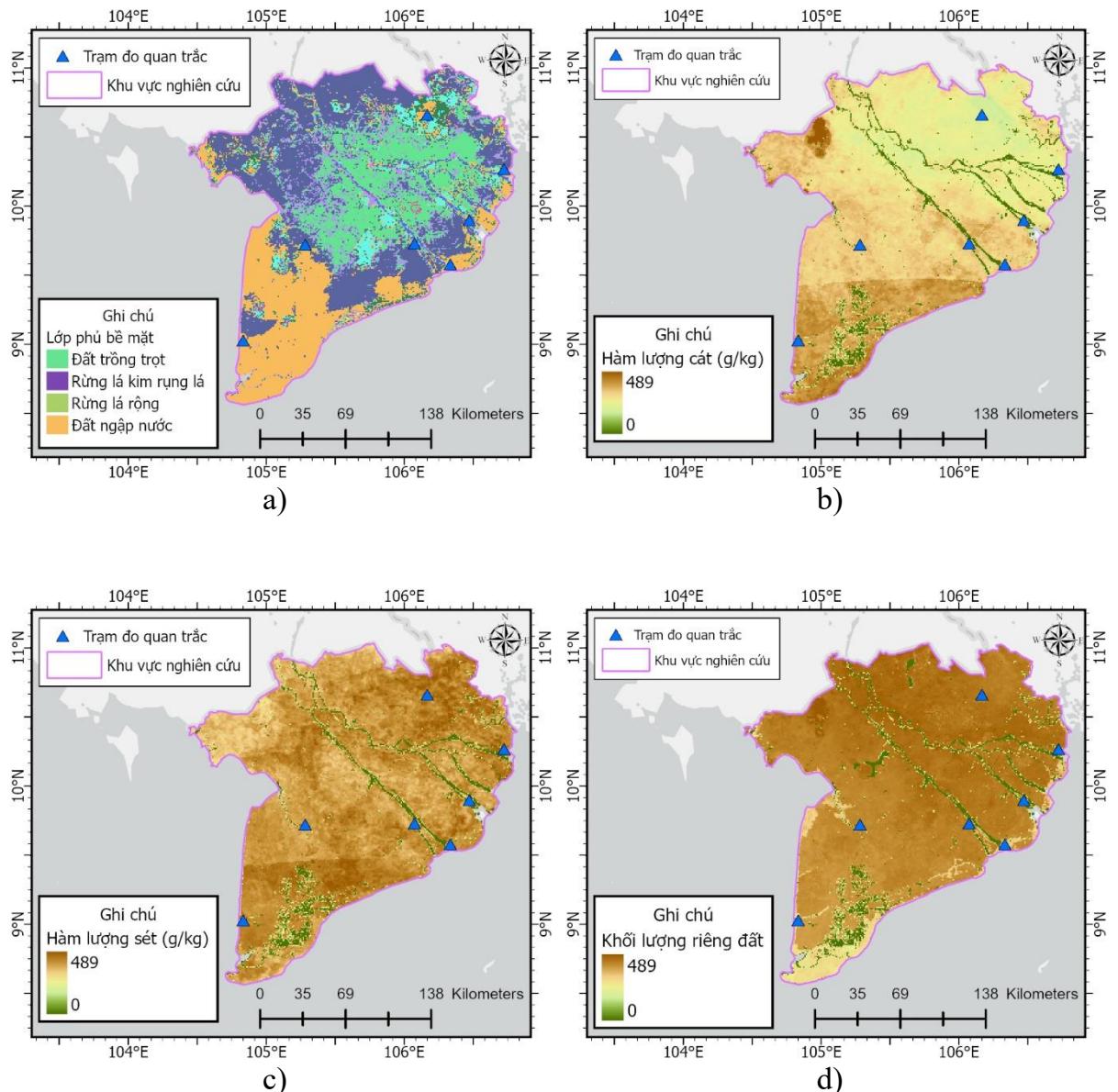
Dữ liệu lớp phủ bề mặt được thu từ sản phẩm MCD12Q1 với độ phân giải 500 m, cung cấp thông tin phân loại đất, thảm thực vật, đất nông nghiệp và mặt nước. Lớp phủ bề mặt ảnh hưởng trực tiếp đến khả năng giữ nước, giữ muối và khả năng phục hồi sau xâm nhập mặn. Các vùng đất trồng hoặc ít thảm phủ thường nhạy cảm hơn với mặn hóa, trong khi các khu vực có thảm thực vật dày hoặc canh tác ổn định có khả năng giảm thiểu tác động của muối thông qua giữ ẩm và che phủ bề mặt. Dữ liệu lớp phủ bề mặt giúp mô hình hóa được sự khác biệt sinh thái giữa các khu vực, qua đó đánh giá được mức độ dễ tổn thương của từng loại hình sử dụng đất đối với xâm nhập mặn.

Hàm lượng cát được trích xuất từ bộ SoilGrids với độ phân giải 250 m, phản ánh mức độ thấm nước và tính chất vật lý của đất. Đất giàu cát có khả năng thấm nhanh, giữ nước kém, rất dễ bị xâm nhập mặn khi nước biển hoặc nước mặn từ kênh rạch thâm nhập vào tầng đất mặt. Trong điều kiện mùa khô, các khu vực đất có hàm lượng cát cao thường có xu hướng mặn hóa mạnh và khó phục hồi. Việc sử dụng chỉ tiêu này trong mô hình giúp phân định rõ các vùng đất có nguy cơ mặn hóa cao do đặc tính thấm mạnh và giữ nước thấp.

Hàm lượng sét được trích xuất từ bộ SoilGrids với độ phân giải 250 m biểu thị mức độ kết dính và khả năng giữ nước của đất. Đất giàu sét có khả năng giữ nước tốt hơn đất cát, nhưng đồng thời lại dễ tích tụ muối khi đã bị xâm nhập mặn, làm cho quá trình rửa mặn trở nên khó khăn và kéo dài. Vì vậy, dữ liệu hàm lượng sét giúp mô hình đánh giá khả năng duy trì và lưu giữ muối trong đất theo thời gian. Các khu vực đất sét cao thường biểu hiện mức độ mặn hóa dai dẳng và nguy cơ suy thoái đất lớn hơn.

Khối lượng riêng đất được thu thập từ bộ dữ liệu SoilGrids với độ phân giải 250 m, phản ánh mức độ nén chặt và cấu trúc tổng thể của đất. Dữ liệu này thể hiện lượng khối lượng đất có trong một đơn vị thể tích, bao gồm cả phần hạt rắn và khoảng rỗng giữa các hạt đất. Đất có mật độ khối cao thường có cấu trúc chặt, ít lỗ rỗng, dẫn đến khả năng thấm và trao đổi nước kém. Trong điều kiện xâm nhập mặn, đặc điểm này khiến muối dễ bị giữ lại trong tầng đất mặt và khó được rửa trôi, làm cho tình trạng mặn hóa kéo dài hơn so với các khu vực đất tơi xốp. Ngược lại, những vùng đất có mật độ khối thấp thường có cấu trúc thoáng, khả năng thấm tốt hơn, từ đó hỗ trợ quá trình tiêu thoát muối trong mùa mưa hoặc khi có biện pháp rửa mặn. Việc tích hợp mật độ khối đất vào mô hình phân tích giúp nhận diện các khu vực dễ tích tụ muối lâu dài, đồng thời hỗ trợ đánh giá nguy cơ suy thoái đất trong bối cảnh xâm nhập mặn gia tăng tại Đồng bằng Sông Cửu Long.

Hình ảnh bộ dữ liệu thổ nhưỡng được trình bày dưới dạng bản đồ:



Hình 11: Bộ dữ liệu thổ nhưỡng dưới dạng bản đồ: a) Lớp phủ bì mặt; b) Hàm lượng cát; c) Hàm lượng sét; d) Khối lượng riêng đất

### 3.2.6. Dữ liệu thực tế

Dữ liệu điểm mặn thực tế được thu thập thông qua khảo sát thực địa của các nghiên cứu tại khu vực Đồng Bằng Sông Cửu Long của (Nguyễn Hữu Duy và cộng sự, 2023-2025) [46], [47], cùng với các báo cáo liên quan từ Cục Khí Tượng và Thuỷ Văn. Tổng cộng 330 điểm mẫu đã được thu thập trong giai đoạn mùa khô. 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 xâm nhập mặn (Hình 6). Tại mỗi điểm mẫu đất được lấy ở độ sâu từ 0–30 cm để xác định giá trị độ dẫn điện - tiêu chí phản ánh

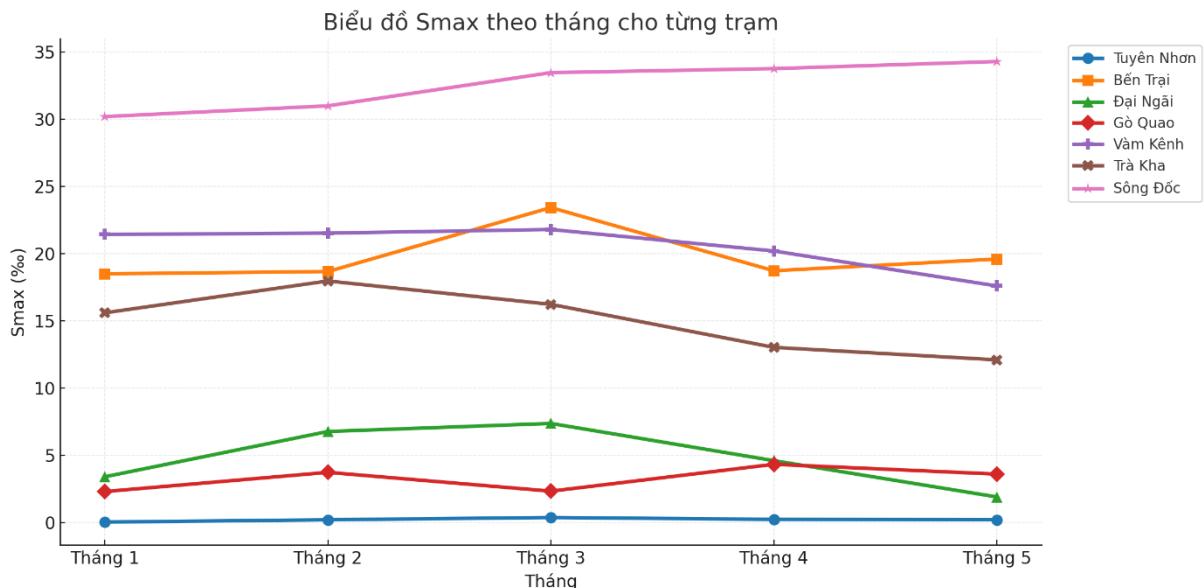
mức độ mặn trong đất. Quá trình lấy mẫu được tiến hành trong điều kiện thời tiết khô ráo, không có mây che phủ, nhằm hạn chế ảnh hưởng của các yếu tố khí tượng và đảm bảo tính đồng bộ với ảnh viễn thám[46]. Tập dữ liệu này được sử dụng làm biến phụ thuộc để huấn luyện, hiệu chỉnh và kiểm định các mô hình mô phỏng, góp phần nâng cao độ tin cậy của kết quả phân tích không gian.

Bên cạnh các điểm đo mặn thực tế, dữ liệu của 7 trạm quan trắc quanh Đồng Bằng Sông Cửu Long được tổng hợp từ Cục Khí Tượng Thuỷ Văn Nam Bộ [48] cũng được sử dụng để kiểm chứng xu hướng của mô hình. Các trạm được bố trí tại những vị trí đại diện cho các điều kiện thủy văn khác nhau trong khu vực, giúp phản ánh rõ sự thay đổi theo không gian của quá trình xâm nhập mặn. Trong nghiên cứu này, dữ liệu được thu thập từ 7 trạm quan trắc gồm: Trạm Tuyên Nhơn; Trạm Bến Tre; Trạm Đại Ngãi; Trạm Gò Quao; Trạm Vàm Kênh; Trạm Trà Kha; Trạm Sông Đốc. Các trạm này cung cấp chuỗi số liệu độ mặn đo đặc theo mùa khô từ tháng 1 đến tháng 5, tạo nên bộ dữ liệu liên tục mô tả diễn biến theo mùa và theo từng giai đoạn trong năm. Chuỗi số liệu độ mặn được tổng hợp và kiểm tra nhằm loại bỏ nhiều cung như đảm bảo tính thống nhất khi đối chiếu với các lớp dữ liệu viễn thám được xử lý theo cùng mốc thời gian. Bảng giá trị trung bình tháng của từng trạm được trình bày tại Bảng 3, thể hiện sự biến động độ mặn giữa các vị trí quan trắc và sự thay đổi đặc trưng theo từng tháng. Việc sử dụng hệ thống trạm quan trắc không chỉ giúp đánh giá khách quan hiệu quả mô hình mà còn cung cấp căn cứ thực nghiệm quan trọng, đảm bảo độ tin cậy của các kết quả phân tích và nâng cao giá trị ứng dụng của nghiên cứu trong thực tiễn.

Tên Trạm	Tỉnh/Phường	Tháng 1	Tháng 2	Tháng 3	Tháng 4	Tháng 5
Tuyên Nhơn	Long An	0.033	0.2	0.37	0.23	0.2
Bến Tre	Bến Tre	18.5	18.67	23.43	18.73	19.6
Đại Ngãi	Sóc Trăng	3.4	6.77	7.37	4.6	1.9
Gò Quao	Kiên Giang	2.3	3.73	2.33	4.33	3.6
Vàm Kênh	Tiền Giang	21.43	21.53	21.8	20.2	17.6
Trà Kha	Trà Vinh	15.6	17.97	16.23	13.03	12.1
Sông Đốc	Cà Mau	30.2	31	33.47	33.77	34.3

Bảng 3: Giá trị đo mặn tại các trạm

Các giá trị đo của các trạm được biểu diễn theo dạng biểu đồ được trình bày ở Hình 12:



Hình 12: Giá trị độ mặn tại các trạm đo

Dựa trên bảng số liệu thống kê (Bảng 3) về giá trị đo mặn và biểu đồ trực quan hóa độ mặn lớn nhất theo tháng (Hình 12), có thể thấy diễn biến xâm nhập mặn tại khu vực Đồng bằng sông Cửu Long trong năm tháng đầu năm có sự phân hóa rất rõ rệt cả về không gian giữa các trạm và thời gian giữa các tháng. Độ mặn đo được dao động trong biên độ rất lớn, trải dài từ mức thấp không đáng kể xấp xỉ 0 phần nghìn cho đến mức rất cao trên 34 phần nghìn, phản ánh tính chất đa dạng và phức tạp của chế độ thủy văn tại từng địa phương.

Về mặt tổng quan, trạm Sông Đốc thuộc tỉnh Cà Mau là nơi ghi nhận tình trạng xâm nhập mặn gay gắt nhất trong toàn bộ hệ thống quan trắc. Đường biểu diễn của trạm này nằm tách biệt hẳn lên phía trên so với các trạm còn lại và duy trì mức độ mặn rất cao ngay từ tháng 1 với chỉ số 30,2 phần nghìn. Xu hướng tại đây tăng dần đều theo thời gian và đạt đỉnh điểm vào tháng 5 với giá trị 34,3 phần nghìn. Điều này cho thấy khu vực Cà Mau chịu ảnh hưởng trực tiếp và mạnh mẽ của thủy triều biển mà ít chịu tác động đẩy mặn của nguồn nước ngọt từ thượng nguồn trong giai đoạn này, khiến độ mặn tích lũy và tăng cao liên tục.

Nằm ở nhóm có độ mặn trung bình cao là các trạm Bến Tre thuộc Bến Tre, Vầm Kênh thuộc Tiền Giang và Trà Kha thuộc Trà Vinh. Tại nhóm này, diễn biến độ mặn thể hiện rõ quy luật mùa khô đặc trưng của khu vực. Cụ thể, độ mặn tại Bến Tre tăng từ mức 18,5 phần nghìn ở tháng 1 và đạt đỉnh cực đại vào tháng 3 với giá trị 23,43 phần nghìn, sau đó giảm dần xuống còn 19,6 phần nghìn vào tháng 5. Tương tự, trạm Vầm

Kênh duy trì mức độ mặn ổn định quanh ngưỡng trên 21 phần nghìn trong ba tháng đầu năm và bắt đầu giảm nhẹ vào tháng 4 và tháng 5 xuống còn 17,6 phần nghìn. Trạm Trà Kha có xu hướng đạt đỉnh sớm hơn vào tháng 2 với 17,97 phần nghìn và giảm sâu dần về cuối chu kỳ quan trắc, xuống mức 12,1 phần nghìn vào tháng 5. Sự sụt giảm này là dấu hiệu của việc xuất hiện các cơn mưa chuyển mùa vào cuối tháng 4 và đầu tháng 5 giúp rửa trôi và đẩy lùi ranh mặn.

Đối với nhóm các trạm có độ mặn thấp hơn bao gồm Đại Ngãi thuộc Sóc Trăng và Gò Quao thuộc Kiên Giang, biến động độ mặn diễn ra phức tạp nhưng nằm trong phạm vi kiểm soát tốt hơn. Trạm Đại Ngãi cho thấy sự gia tăng độ mặn đáng kể từ 3,4 phần nghìn ở tháng 1 lên mức đỉnh 7,37 phần nghìn vào tháng 3, nhưng ngay sau đó giảm mạnh xuống chỉ còn 1,9 phần nghìn vào tháng 5. Trạm Gò Quao có diễn biến trồi sụt thất thường hơn khi đạt đỉnh vào tháng 4 với 4,33 phần nghìn. Mức độ biến động tại các trạm này cho thấy sự tranh chấp mạnh mẽ giữa triều biển và dòng chảy sông, tuy nhiên mức độ xâm nhập mặn tại đây không quá gay gắt như khu vực Cà Mau hay Bến Tre.

Đáng chú ý nhất ở nhóm độ mặn thấp là trạm Tuyên Nhơn thuộc tỉnh Long An. Số liệu cho thấy khu vực này hầu như không bị ảnh hưởng bởi xâm nhập mặn trong suốt năm tháng quan trắc. Giá trị đo được luôn ở mức rất thấp, dao động từ 0,033 đến cao nhất là 0,37 phần nghìn vào tháng 3. Trên biểu đồ, đường biểu diễn của trạm Tuyên Nhơn gần như nằm sát trực hoành và đi ngang, chứng tỏ nguồn nước tại đây vẫn giữ được tính chất ngọt quanh năm hoặc vị trí trạm nằm sâu trong nội đồng nên ít chịu tác động của thủy triều.

Xét về quy luật thời gian, tháng 3 là thời điểm ghi nhận sự xâm nhập mặn đạt đỉnh tại đa số các trạm quan trắc quan trọng như Bến Tre, Đại Ngãi và Vành Kênh. Đây là thời điểm giữa mùa khô khi lượng nước từ thượng nguồn sông Mekong đổ về thấp nhất kết hợp với gió chuvong hoạt động mạnh đẩy nước biển sâu vào đất liền. Tuy nhiên, sang đến tháng 4 và tháng 5, xu hướng giảm mặn đã xuất hiện ở hầu hết các trạm, ngoại trừ Sông Đốc vẫn tăng và Gò Quao biến động nhẹ, báo hiệu sự kết thúc của đợt hạn mặn gay gắt và bắt đầu chuyển sang mùa mưa.

Tổng hợp lại, các số liệu phản ánh một bức tranh rõ nét về hiện trạng xâm nhập mặn với mức độ nghiêm trọng tập trung chủ yếu ở các tỉnh ven biển như Cà Mau, Bến Tre và Tiền Giang, trong khi các khu vực sâu hơn trong nội địa như Long An vẫn đảm bảo được nguồn nước ngọt. Sự chênh lệch rất lớn về giá trị độ mặn cực đại giữa trạm cao nhất là Sông Đốc và thấp nhất là Tuyên Nhơn lên tới hơn 100 lần cho thấy tính cấp

thiết của việc xây dựng các kịch bản ứng phó riêng biệt cho từng tiêu vùng sinh thái khác nhau trong khu vực.

### 3.3. Đánh giá và so sánh hiệu suất các mô hình

Một bước quan trọng khác trong nghiên cứu là xây dựng mô hình với các tham số được điều chỉnh hợp lý nhằm đảm bảo khả năng áp dụng và thực nghiệm lại ở những khu vực khác. Trong quá trình triển khai, các mô hình hồi quy được lựa chọn làm đầu vào bao gồm XGBoost, CatBoost và Rừng Ngẫu Nhiên. Đây đều là những thuật toán học máy mạnh mẽ, có khả năng mô phỏng tốt các mối quan hệ phi tuyến giữa các yếu tố môi trường. Các mô hình được huấn luyện với việc tinh chỉnh các tham số quan trọng trong một khoảng giá trị xác định trước, nhằm tìm ra cấu hình tối ưu nhất cho độ chính xác dự báo. Việc tối ưu hóa tham số giúp nâng cao hiệu suất mô hình, đồng thời đảm bảo tính ổn định và khả năng mở rộng sang những vùng nghiên cứu khác.

Trong đó, đối với mô hình Hồi quy Rừng ngẫu nhiên, ta lựa chọn các tham số sau: số lượng cây trong rừng (`n_estimators = 600`); độ sâu tối đa của mỗi cây (`max_depth = 10`); số lượng mẫu tối thiểu cần thiết để tách một nút (`min_samples_split = 4`); số lượng mẫu tối thiểu cần có tại một nút lá (`min_samples_leaf = 2`); số lượng đặc trưng được xem xét tại mỗi lần phân chia (`max_features = "sqrt"`); cùng với tỷ lệ mẫu được sử dụng để huấn luyện mỗi cây (`max_samples = 0,8`) và tham số thiết lập tính ngẫu nhiên (`random_state`).

Đối với mô hình CatBoost, các tham số được chọn bao gồm: số vòng lặp (`iterations = 300`); độ sâu của cây (`depth = 5`); tốc độ học (`learning_rate = 0,02`); hệ số điều chuẩn L2 nhằm tránh hiện tượng quá khớp (`l2_leaf_reg = 8,0`); tỷ lệ mẫu được sử dụng trong mỗi lần xây dựng cây (`subsample = 0,85`); cùng với tham số ngẫu nhiên (`random_seed`) nhằm đảm bảo tính lặp lại trong huấn luyện.

Đối với mô hình XGBoost (XGB), ta lựa chọn các tham số sau: số lượng cây (`n_estimators = 300`); tốc độ học giúp mô hình cập nhật trọng số theo từng bước nhỏ để tránh vượt quá nghiệm tối ưu (`learning_rate = 0,03`); độ sâu tối đa của mỗi cây nhằm kiểm soát khả năng mô hình hóa quan hệ phi tuyến (`max_depth = 10`); trọng số tối thiểu trên mỗi nút để hạn chế việc tạo ra các nút ít ý nghĩa (`min_child_weight = 10`); tỷ lệ mẫu được sử dụng khi xây dựng mỗi cây nhằm giảm hiện tượng học quá chi tiết dữ liệu (`subsample = 0,6`); tỷ lệ đặc trưng được lấy ngẫu nhiên khi xây dựng cây (`colsample_bytree = 0,6`); cùng hai tham số điều chuẩn L1 (`reg_alpha = 5`) và L2

(reg\_lambda = 5) giúp tăng khả năng chống quá khóp của mô hình; cuối cùng là tham số ngẫu nhiên (random\_state) nhằm đảm bảo tính tái lập.

Các tham số được sử dụng trong ba mô hình học máy được trình bày dưới Bảng:

Mô hình	Tham số	Giá trị
Rừng ngẫu nhiên	Số lượng cây trong rừng	600
	Độ sâu tối đa của mỗi cây	10
	Số mẫu tối thiểu để tách một nút	4
	Số mẫu tối thiểu tại một nút lá	2
	Số lượng đặc trưng được xem xét khi phân chia	Sqrt
	Tỷ lệ mẫu dùng cho mỗi cây	0,8
CatBoost	Số vòng lặp học	300
	Độ sâu cây	5
	Tốc độ học	0,02
	Hệ số điều chuẩn L2	8,0
	Tỷ lệ mẫu sử dụng cho mỗi cây	0,85
XGBoost	Số lượng cây	300
	Tốc độ học	0,03
	Độ sâu tối đa của cây	10
	Trọng số tối thiểu tại mỗi nút	10
	Tỷ lệ mẫu sử dụng cho mỗi cây	0,6
	Tỷ lệ đặc trưng được chọn ngẫu nhiên	0,6
	Điều chuẩn L1	5
	Điều chuẩn L2	5

Bảng 4: Tham số được sử dụng trong ba mô hình học máy

Tiếp theo, sau khi xác định được bộ tham số phù hợp cho ba mô hình Random Forest, CatBoost và XGBoost, nghiên cứu tiến hành đánh giá hiệu suất mô hình thông qua phương pháp kiểm định chéo K-Fold. Toàn bộ dữ liệu được chia thành K phần bằng nhau; tại mỗi vòng lặp, một phần được sử dụng làm tập kiểm tra, trong khi K-1 phần còn lại được dùng làm tập huấn luyện. Cách tiếp cận này giúp đảm bảo rằng mỗi quan sát trong bộ dữ liệu đều được sử dụng để kiểm tra một lần, từ đó cung cấp đánh giá khách quan và ổn định hơn so với việc chia dữ liệu thành một lần huấn luyện – kiểm tra duy nhất.

Việc áp dụng kiểm định chéo đặc biệt quan trọng trong bối cảnh bộ dữ liệu có kích thước vừa phải, giúp giảm nguy cơ mô hình bị phụ thuộc vào cách chia dữ liệu ngẫu nhiên. Đồng thời, phương pháp này cho phép đánh giá độ ổn định của mô hình qua nhiều lần lặp lại, thể hiện qua sự biến thiên của các chỉ số sai số. Ba mô hình được huấn luyện

và kiểm định trên cùng một hệ đặc trưng bao gồm các biến viễn thám, thổ nhưỡng và khí tượng liên quan đến quá trình xâm nhập mặn. Điều này giúp đảm bảo sự công bằng trong so sánh và loại bỏ các khác biệt gây ra bởi chất lượng dữ liệu đầu vào.

Để đánh giá hiệu quả dự báo của các mô hình, ba chỉ số chính được sử dụng gồm RMSE, MAE và hệ số tương quan R, được tính trung bình trên toàn bộ K lần kiểm định. Trong đó, RMSE phản ánh mức độ sai lệch căn phương giữa giá trị quan trắc và giá trị mô hình dự báo, đồng thời nhấn mạnh mạnh hơn sai số lớn; MAE thể hiện sai lệch tuyệt đối trung bình giúp đánh giá mức độ sai lệch theo đúng đơn vị gốc; còn hệ số R phản ánh mức độ tương quan và sự phù hợp giữa mô hình với phân bố thực tế. Việc kết hợp đồng thời ba chỉ số này giúp đánh giá mô hình một cách toàn diện, bao gồm cả mức độ chính xác tổng thể, sai lệch cục bộ và chất lượng tương quan.

Kết quả kiểm định chéo cho ba mô hình RF, CB và XGB được tổng hợp trong bảng dưới đây. Việc so sánh trực tiếp giữa các thuật toán cho phép xác định mô hình có hiệu suất tổng thể tốt nhất, đồng thời giúp nhận diện các ưu điểm và hạn chế của từng phương pháp trong điều kiện bộ dữ liệu và khu vực nghiên cứu cụ thể.

Bảng dưới đây biểu diễn kết quả các chỉ số RMSE, MAE và R của từng mô hình:

Thuật toán	Tập huấn luyện			Tập kiểm tra		
	RMSE (dS/m)	MAE	R	RMSE (dS/m)	MAE	R
RF	1,59	0,77	0,94	2,73	1,37	0,78
XGB	1,37	0,69	0,95	2,55	1,31	0,81
CB	1,72	0,96	0,94	2,65	1,36	0,80

Bảng 5: Số liệu đánh giá hiệu suất các mô hình

Bảng kết quả cho thấy sự khác biệt rõ rệt giữa ba mô hình RF, XGB và CB trong khả năng dự báo độ mặn dựa trên các biến đầu vào. Khi phân tích đồng thời các chỉ số RMSE, MAE và hệ số tương quan R trên cả tập huấn luyện và tập kiểm tra, có thể nhận thấy mức độ phù hợp mô hình, xu hướng sai số và khả năng tổng quát hóa của từng thuật toán.

Đối với mô hình RF, kết quả huấn luyện cho thấy RMSE = 1,59; MAE = 0,77 và R = 0,94, phản ánh rằng mô hình học khá tốt các quan hệ phi tuyến trong bộ dữ liệu. Tuy nhiên, khi áp dụng cho tập kiểm tra, sai số tăng lên với RMSE = 2,73; MAE = 1,37 và hệ số tương quan giảm còn 0,78. Sự chênh lệch giữa hai tập cho thấy RF có xu hướng quá khớp, đặc biệt trong bối cảnh dữ liệu biến động mạnh theo không gian. Điều này phù hợp với đặc điểm của RF khi làm việc với các bộ dữ liệu có nhiều cao hoặc số lượng

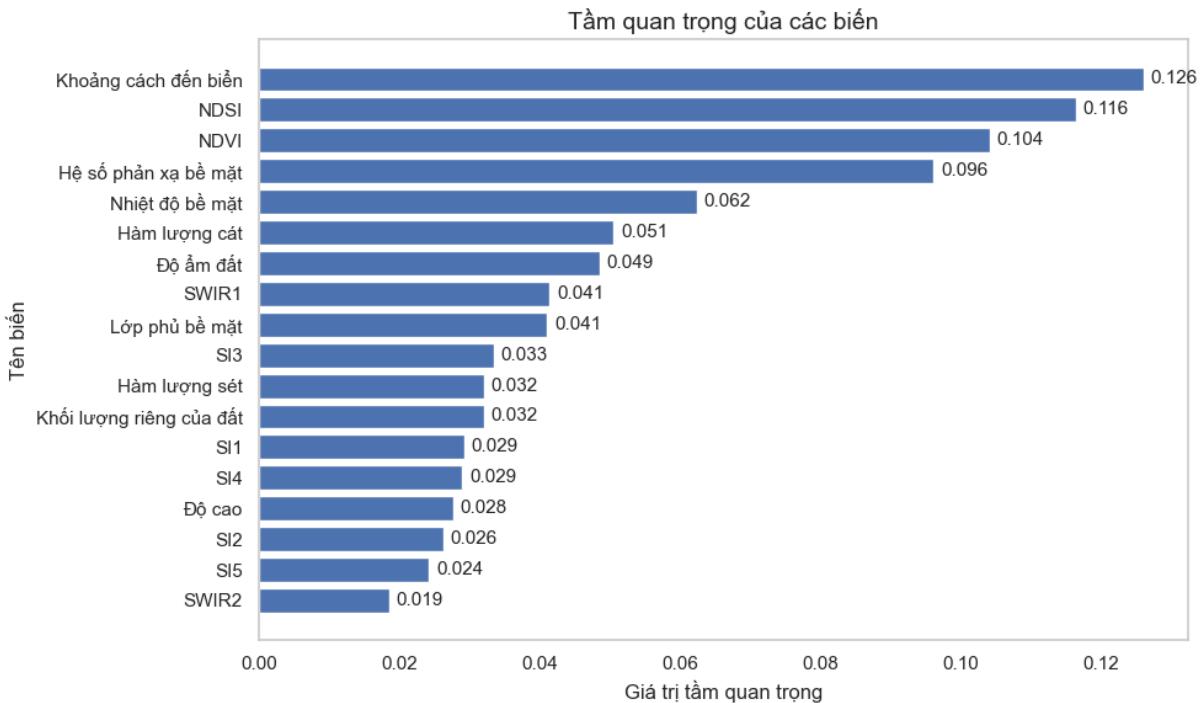
mẫu không lớn, khiến mô hình khó duy trì được tính ổn định khi dự báo trên những khu vực chưa từng gặp.

Mô hình XGB thể hiện hiệu suất vượt trội nhất trong ba mô hình. Trên tập huấn luyện, XGB đạt RMSE = 1,37; MAE = 0,69 và R = 0,95, cho thấy khả năng mô phỏng rất tốt. Đáng chú ý, trên tập kiểm tra, mô hình vẫn duy trì độ chính xác cao với RMSE = 2,55; MAE = 1,31 và hệ số tương quan đạt 0,81. Mức giảm hiệu suất giữa hai tập là không lớn, chứng tỏ khả năng tổng quát hóa tốt của XGB và khả năng xử lý hiệu quả các quan hệ phi tuyến phức tạp giữa các biến đầu vào. Sự ổn định của mô hình đến từ cơ chế tăng cường dần dần và khả năng tự điều chỉnh sai số, giúp XGB hạn chế được hiện tượng quá khớp và phù hợp với dữ liệu độ mặn có tính phân hóa mạnh theo không gian.

Đối với mô hình CB, kết quả huấn luyện đạt RMSE = 1,72; MAE = 0,96 và R = 0,94, phản ánh mức độ học khá tốt. Trên tập kiểm tra, sai số tăng lên với RMSE = 2,65; MAE = 1,36 và hệ số tương quan giảm còn 0,80. Mặc dù kém hơn XGB, CatBoost vẫn thể hiện khả năng tổng quát tốt hơn RF, nhờ cơ chế xử lý dữ liệu dạng bảng và khả năng khắc phục phân bố không đồng nhất. Tuy nhiên, mức hiệu quả thấp hơn XGB cho thấy CB chưa khai thác tối ưu sự biến động mạnh và tính phi tuyến đặc thù của dữ liệu mặn.

Khi so sánh ba mô hình, XGB là thuật toán cho hiệu suất tốt nhất, thể hiện qua RMSE và MAE thấp nhất cùng hệ số tương quan cao nhất trên tập kiểm tra. CB đứng thứ hai với khả năng dự báo tương đối ổn định và ít chênh lệch giữa hai tập. RF có hiệu suất thấp nhất, đặc biệt ở khả năng tổng quát hóa, cho thấy thuật toán này gặp khó khăn khi làm việc với dữ liệu mặn có sự thay đổi lớn giữa các vùng sinh thái và khoảng cách đến biển.

Tổng hợp các chỉ số trên tập kiểm tra – vốn là thước đo quan trọng nhất phản ánh khả năng ứng dụng thực tế – mô hình XGB chứng minh được độ tin cậy cao hơn cả. Điều này cho thấy mô hình XGB phù hợp để mô phỏng hiện tượng xâm nhập mặn, nơi các quan hệ giữa biến giải thích và độ mặn mang tính phi tuyến phức tạp, chịu ảnh hưởng đồng thời của nhiều yếu tố môi trường và đặc điểm đất – nước.



*Hình 13: Biểu đồ đánh giá chỉ số tầm quan trọng của các biến trong 18 yếu tố ảnh hưởng tới xâm nhập mặn*

Dựa trên mô hình Random Forest, chỉ số tầm quan trọng của các biến được tính toán nhằm đánh giá mức độ đóng góp của từng yếu tố vào khả năng dự báo độ mặn đất. Kết quả cho thấy tất cả 18 biến đều có vai trò nhất định, tuy nhiên mức độ ảnh hưởng có sự khác biệt rõ rệt, phản ánh đúng cơ chế lan truyền và phân bố mặn theo không gian tại Đồng bằng sông Cửu Long.

Biến có tầm quan trọng cao nhất là khoảng cách đến biển (0,126). Điều này hoàn toàn phù hợp với đặc trưng thủy văn – hình thái của khu vực: xâm nhập mặn lan truyền từ vùng cửa sông ven biển vào sâu trong nội đồng, do đó khoảng cách đến biển là yếu tố kiểm soát trực tiếp mức độ ảnh hưởng của mặn. Tiếp theo là chỉ số mặn NDSI (0,116) và chỉ số thực vật NDVI (0,104). NDSI mô tả đặc trưng phổ liên quan đến mặn nên giữ vai trò quan trọng trong việc phân biệt các vùng mặn – ngọt. NDVI phản ánh tình trạng sinh trưởng của thảm thực vật, vốn suy giảm mạnh khi đất bị nhiễm mặn, vì vậy mô hình có thể nhận diện khu vực bị mặn thông qua biến động của lớp phủ thực vật.

Đứng ở vị trí thứ tư là hệ số phản xạ bề mặt trích xuất từ dữ liệu CYGNSS (0,096). Đây là kết quả đáng chú ý, cho thấy tín hiệu phản xạ GNSS rất nhạy với sự thay đổi điện môi của bề mặt, từ đó phản ánh gián tiếp độ mặn đất. Việc biến này duy trì tầm quan trọng cao, dù thấp hơn một số chỉ số quang học, khẳng định tiềm năng của dữ liệu

CYGNSS trong giám sát khu vực ven biển nhiệt đới, nơi ảnh quang học thường bị hạn chế do mây che phủ.

Nhóm biển có ảnh hưởng trung bình bao gồm nhiệt độ bề mặt (0,062), hàm lượng cát (0,051), độ ẩm đất (0,049), băng phổ SWIR1 (0,041), lớp phủ bề mặt (0,041) và chỉ số SI3 (0,033). Các biến này góp phần mô tả cấu trúc đất – nước, ảnh hưởng đến quá trình giữ mặn, trao đổi ẩm và khả năng thâm thấu. Hàm lượng sét (0,032) và khối lượng riêng của đất (0,032) cũng có mức đóng góp tương đương, phản ánh vai trò của loại hình đất trong việc lưu giữ mặn trong tầng mặt đất.

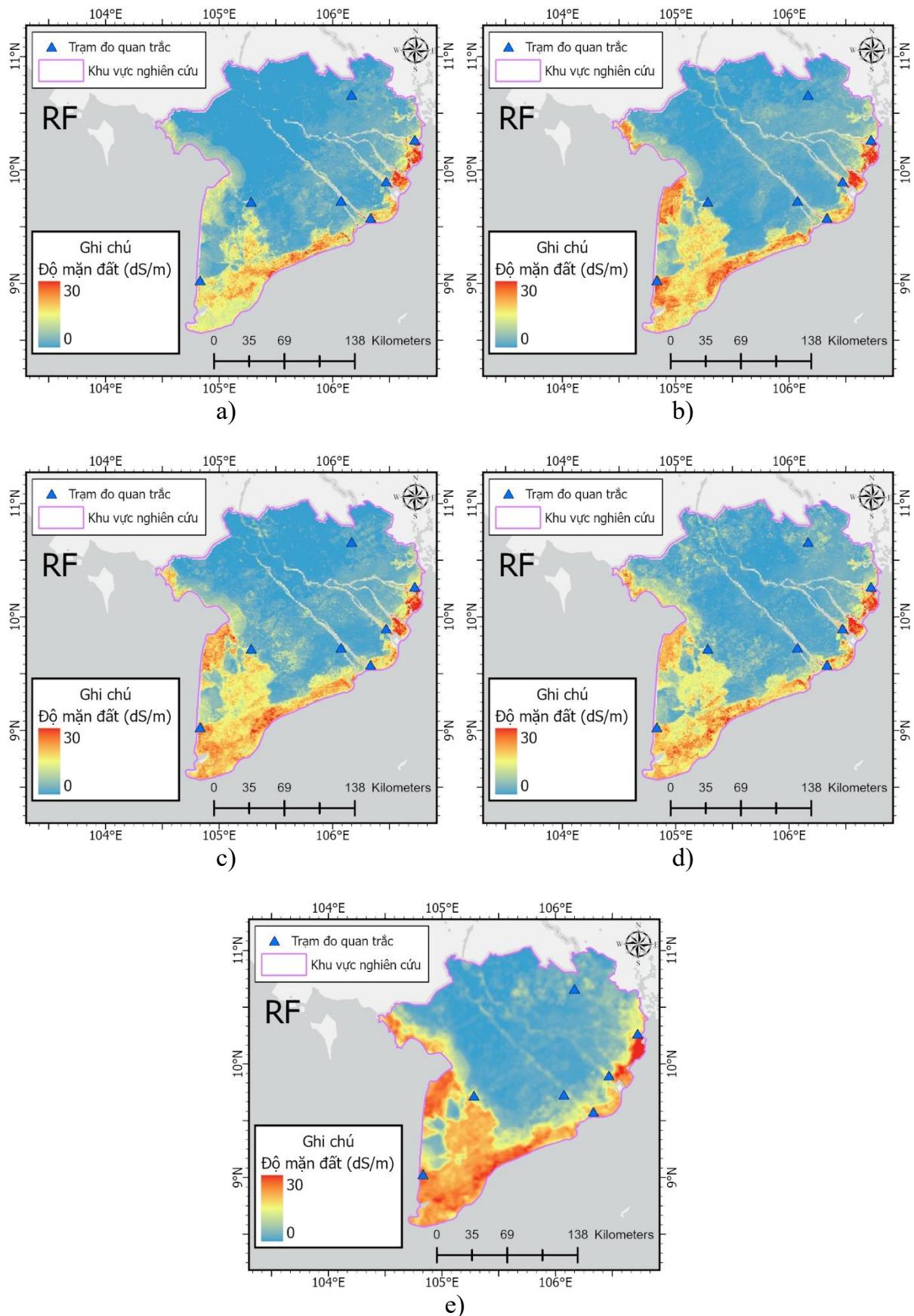
Nhóm biển có tầm quan trọng thấp hơn gồm SI1 (0,029), SI4 (0,029), độ cao địa hình (0,028), SI2 (0,026), SI5 (0,024) và SWIR2 (0,019). Mặc dù giá trị thấp, các biến này vẫn cung cấp thông tin bổ sung giúp mô hình phân biệt sự khác nhau giữa các tiểu vùng sinh thái. Độ cao giữ vai trò nhất định nhưng vì địa hình vùng nghiên cứu tương đối bằng phẳng nên ảnh hưởng không lớn. Băng phổ SWIR2 thể hiện khả năng phân biệt độ ẩm kém hơn so với SWIR1, do đó có mức tầm quan trọng thấp nhất trong mô hình.

Nhìn chung, kết quả tầm quan trọng của các biến phản ánh đúng cơ chế thực tế của hiện tượng xâm nhập mặn: lan truyền theo khoảng cách đến biển, tác động mạnh lên thảm thực vật, chịu ảnh hưởng bởi cấu trúc đất – nước, và cho thấy sự đóng góp rõ rệt của hệ số phản xạ bề mặt từ dữ liệu CYGNSS trong việc nâng cao khả năng dự báo.

### 3.4. Đánh giá xâm nhập mặn tại Đồng Bằng Sông Cửu Long

Dưới đây là bản đồ xâm nhập tại Đồng Bằng Sông Cửu Long từ tháng 1 đến tháng 5 của năm 2025 được xây dựng bởi ba mô hình: RF, XGB và CB. Mặc dù có sự khác biệt về hiệu suất giữa các mô hình, nhưng tất cả các mô hình đều cho thấy rằng, xâm nhập mặn tại Đồng Bằng Sông Cửu Long xuất hiện tại những tỉnh ven biển như: tỉnh Vĩnh Long, tỉnh Cà Mau.

Dưới đây là kết quả bản đồ xâm nhập mặn tại Đồng Bằng Sông Cửu Long được thành lập theo mô hình RF, XGB và CB:



*Hình 14: Bản đồ xâm nhập mặn được xây dựng từ mô hình RF; a) Tháng 1; b) Tháng 2; c) Tháng 3; d) Tháng 4; e) Tháng 5 năm 2025*

Kết quả mô phỏng bằng mô hình Random Forest cho giai đoạn từ tháng 1 đến tháng 5 phản ánh rõ nét quy luật biến động của xâm nhập mặn trong mùa khô tại khu vực nghiên cứu. Trong tháng 1, mặn chỉ xuất hiện rải rác dọc theo vùng ven biển phía Đông và Đông Nam, đặc biệt ở các dải duyên hải sát biển. Các khu vực nội đồng thượng nguồn vẫn duy trì mức mặn rất thấp, thể hiện giai đoạn đầu mùa khô khi dòng chảy thượng nguồn còn đủ mạnh để đẩy lùi nước mặn.

Sang tháng 2, phạm vi xâm nhập mặn mở rộng đáng kể. Các dải mặn đậm xuất hiện rõ dọc theo toàn bộ vùng ven biển Đông và ven biển Tây, đồng thời lan sâu vào đất liền thông qua các trực sông chính và hệ thống kênh rạch đổ ra biển. Đây là thời điểm dòng chảy từ thượng nguồn giảm nhanh, tạo điều kiện cho nước mặn xâm nhập sâu hơn vào các khu vực trũng thấp và ven sông.

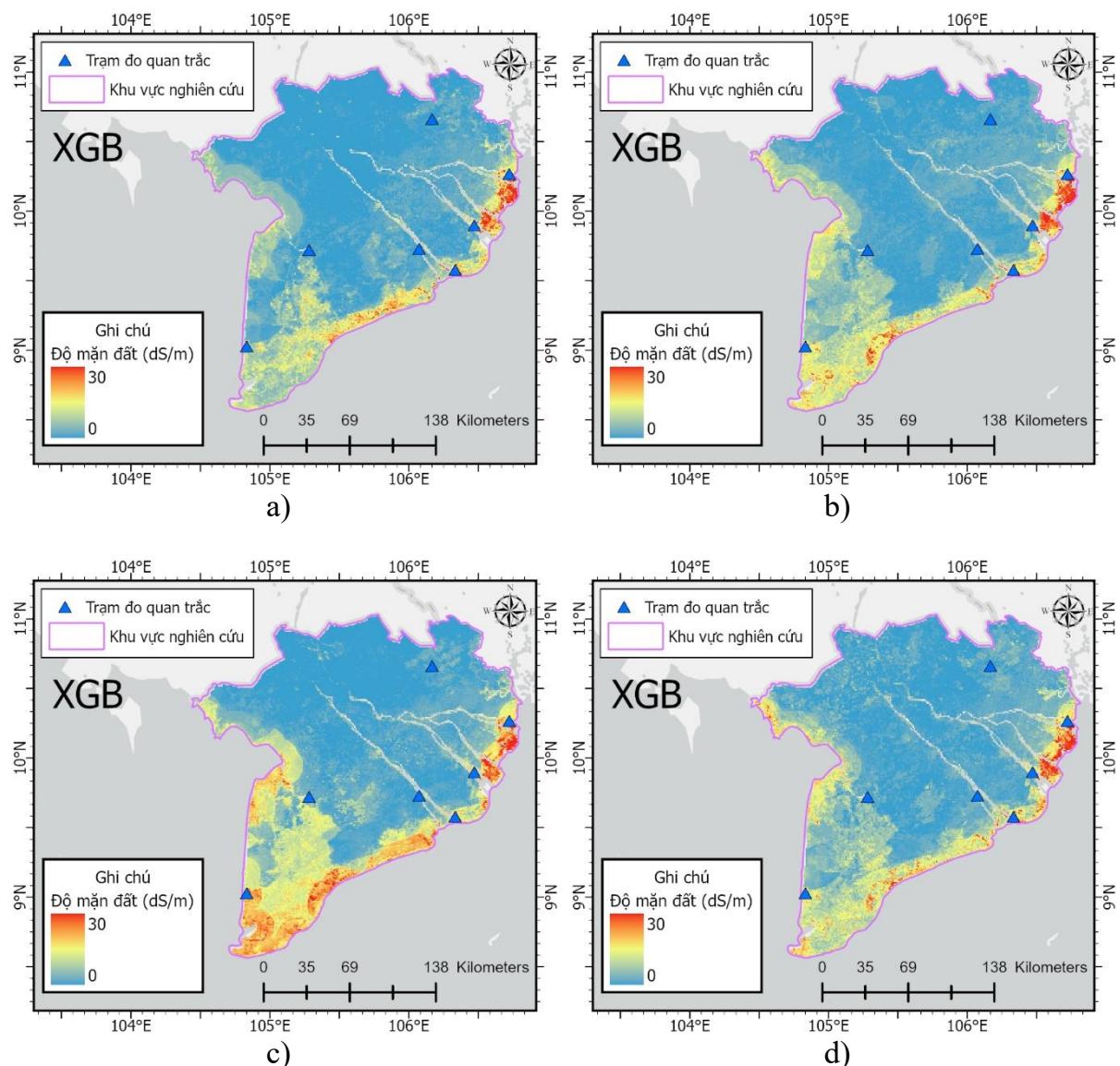
Tháng 3 là giai đoạn ghi nhận mức độ mặn mạnh nhất trong cả mùa khô. Bản đồ mô phỏng thể hiện rõ các mảng màu đỏ cam đậm liên tục chạy dọc ven biển phía Nam, dải duyên hải phía Đông Nam và các vùng cửa sông chính. Đây cũng là thời kỳ lưu lượng thượng nguồn xuống mức thấp nhất, kết hợp với nhiệt độ cao và bốc hơi mạnh, khiến quá trình tích tụ muối trong đất diễn ra rõ rệt. Các vùng ven biển chuyển sang trạng thái nhiễm mặn nặng, trong khi các khu vực nội đồng dù vẫn duy trì mức mặn thấp nhưng đã xuất hiện những dải ảnh hưởng nhẹ theo hướng lan sâu của mặn.

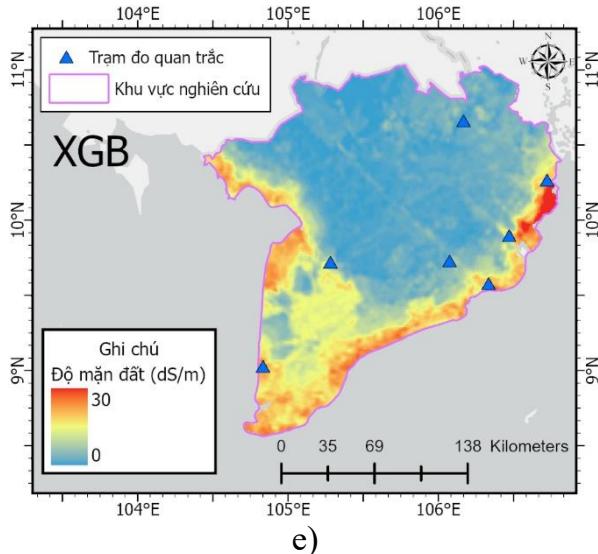
Từ tháng 4, mức độ mặn bắt đầu suy giảm, thể hiện qua sự thu hẹp của các vùng màu đỏ và cam. Mặn vẫn tập trung chủ yếu ở dải ven biển, nhưng cường độ giảm rõ rệt so với tháng 3. Nhiều vùng nội đồng đã phục hồi mức ảnh hưởng thấp hơn. Đến tháng 5, xu thế giảm tiếp tục được khẳng định khi mưa đầu mùa bắt đầu xuất hiện, giúp bổ sung nước ngọt và làm loãng lượng muối tồn đọng. Tuy nhiên, khu vực bán đảo phía Nam vẫn còn duy trì mức mặn khá cao, chưa giảm sâu như các vùng cửa sông khác.

Khi đối chiếu với số liệu thực đo tại các trạm quan trắc, xu thế biến động theo thời gian giữa mô hình và thực tế có sự tương đồng cao. Phần lớn các trạm ven biển chịu ảnh hưởng từ biển Đông đều ghi nhận mặn tăng từ tháng 1, đạt đỉnh trong khoảng tháng 2 đến tháng 3 và giảm dần trong tháng 4 và 5. Riêng tại trạm ven biển Tây (như trạm Sông Đốc), số liệu thực đo cho thấy độ mặn duy trì ở mức rất cao và tiếp tục tăng nhẹ đến tháng 5, điều này hoàn toàn trùng khớp với kết quả mô hình tại bản đồ tháng 5

khi khu vực này vẫn hiển thị các dải màu đậm, phản ánh đúng đặc thù khó thoát mặn và chịu ảnh hưởng triều cường mạnh từ cả hai phía biển.

Việc mô hình Random Forest mô phỏng đúng quy luật tăng, cực đại và giảm của mùa khô, đồng thời nắm bắt được cả những ngoại lệ tại các vùng sinh thái đặc thù, cho thấy mô hình có khả năng phản ánh hợp lý điều kiện thực tế của khu vực nghiên cứu. Sự tương đồng giữa bản đồ mô phỏng và số liệu quan trắc cũng khẳng định mức độ tin cậy của mô hình, đặc biệt trong việc xác định chính xác các khu vực ven biển chịu rủi ro mặn cao và vùng nội đồng có khả năng duy trì điều kiện đất ngọt ổn định.





Hình 15: Bản đồ xâm nhập mặn được xây dựng từ mô hình XGB; a) Tháng 1; b) Tháng 2; c) Tháng 3; d) Tháng 4; e) Tháng 5 năm 2025

Kết quả mô phỏng bằng mô hình XGBoost cho giai đoạn từ tháng 1 đến tháng 5 thể hiện rõ nét quy luật biến động của xâm nhập mặn trong mùa khô tại khu vực Đồng bằng sông Cửu Long. Đây cũng là mô hình có hiệu suất dự báo cao nhất trong nghiên cứu, do đó các bản đồ mô phỏng thể hiện độ chi tiết và tính liên tục về mặt không gian tốt hơn so với các mô hình khác.

Trong tháng 1, phạm vi nhiễm mặn vẫn còn hạn chế và chủ yếu tập trung dọc theo các vùng ven biển, đặc biệt là dải ven biển bán đảo Cà Mau và các khu vực cửa sông. Nội đồng các tỉnh như Đồng Tháp, An Giang và phần lớn Cần Thơ vẫn duy trì mức độ mặn rất thấp, thể hiện đặc điểm giai đoạn đầu mùa khô khi lưu lượng nước từ thượng nguồn vẫn đủ lớn để duy trì áp lực đẩy mặn.

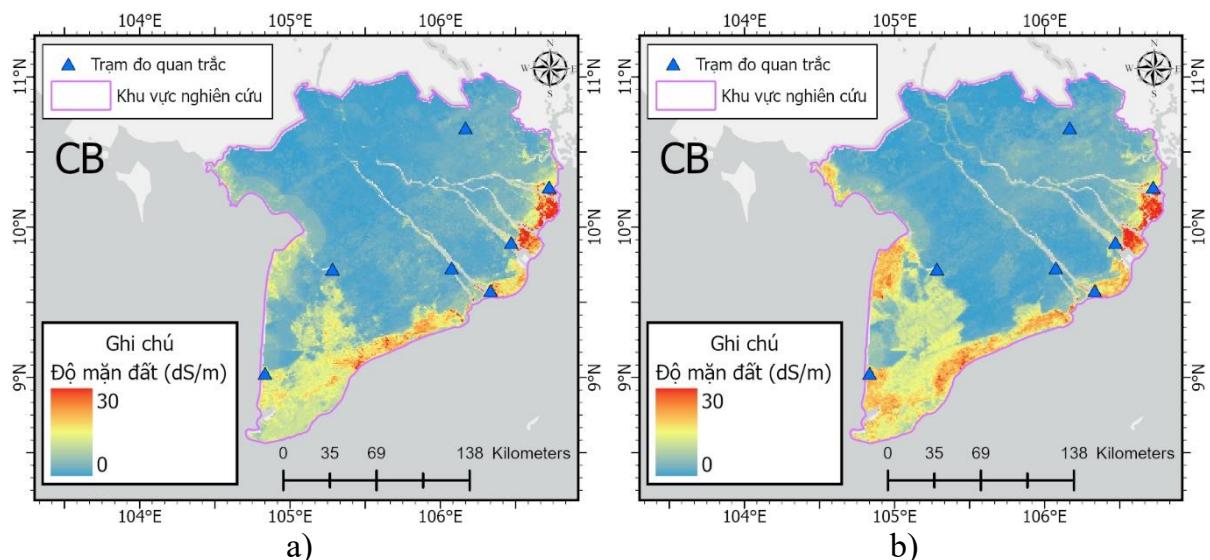
Sang tháng 2, mặn bắt đầu mở rộng phạm vi và lan sâu hơn vào các hệ thống kênh rạch. Vùng ven biển Cà Mau thể hiện sự gia tăng mạnh về cường độ mặn, trong khi các dải mặn ven theo sông Cổ Chiên, sông Hậu và sông Tiền bắt đầu kéo dài và xâm nhập sâu hơn vào nội đồng. Đây là thời điểm dòng chảy thượng nguồn suy giảm, tạo điều kiện thuận lợi cho triều cường đẩy mặn vào sâu trong đất liền.

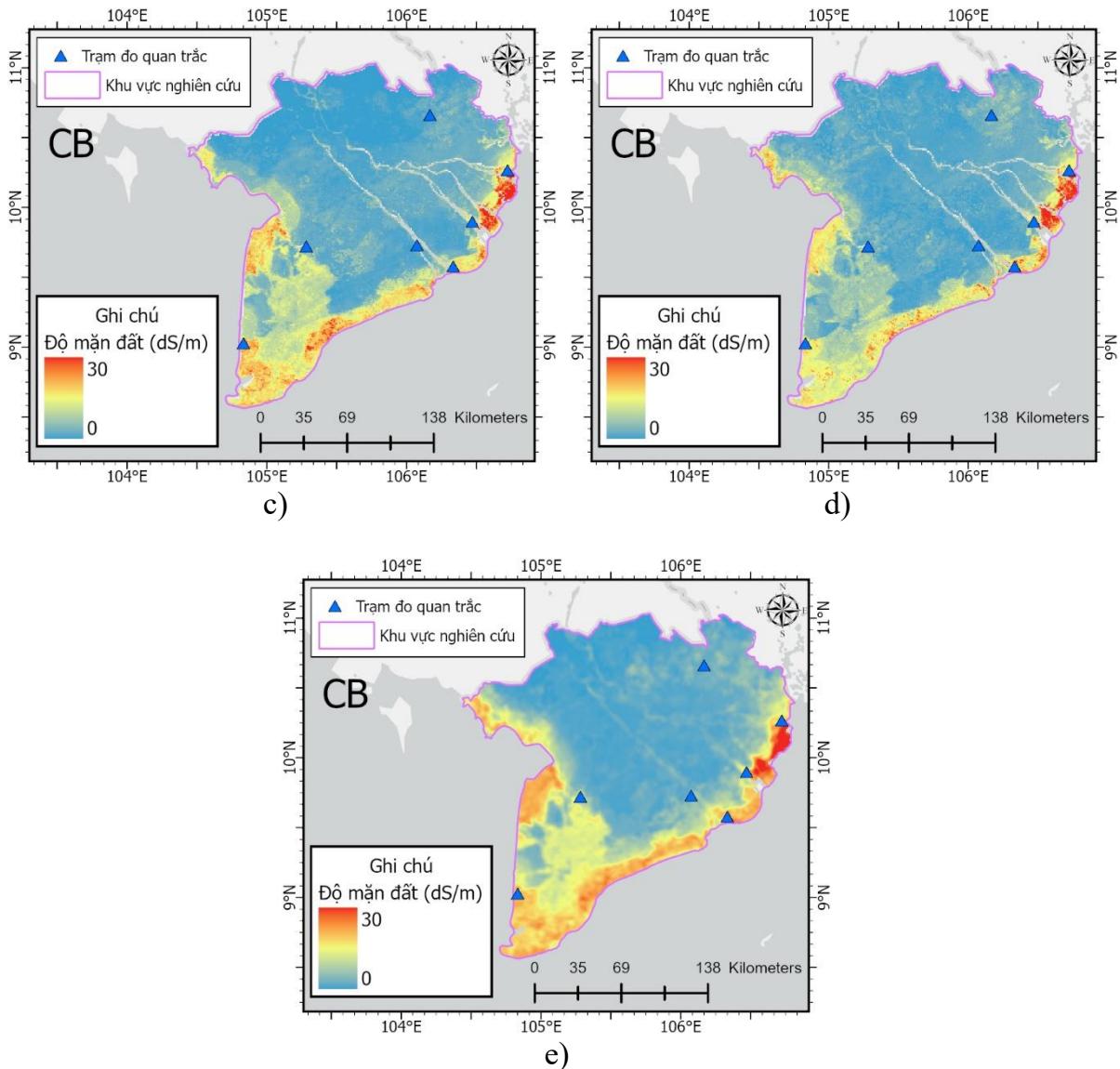
Tháng 3 tiếp tục là thời kỳ ghi nhận xâm nhập mặn mạnh nhất, đúng với chu kỳ cao điểm của mùa khô. Bản đồ mô hình XGB mô phỏng rất rõ các dải mặn màu đỏ và cam đậm liên tục dọc ven biển Cà Mau, khu vực chịu ảnh hưởng kép từ cả biển Đông và biển Tây. Các dải mặn này lan sâu vào các tuyến kênh trực chính, phản ánh chính xác đặc điểm thủy văn trong giai đoạn mực nước sông xuống thấp nhất trong năm kết hợp với điều kiện bốc hơi mạnh.

Bước sang tháng 4, mức độ mặn bắt đầu có xu hướng suy giảm nhưng vẫn duy trì cường độ cao tại các khu vực sát biển. Màu sắc trên bản đồ tại các vùng cửa sông chuyển dần sang các tông màu nhạt hơn, thể hiện quá trình pha loãng khi bắt đầu có những cơn mưa chuyển mùa. Đến tháng 5, xu thế giảm mặn diễn ra rõ rệt hơn trên diện rộng, tuy nhiên khu vực bán đảo Cà Mau vẫn duy trì các dải màu đậm, cho thấy lượng muối tích tụ tại đây chưa thể rửa trôi hoàn toàn.

Khi so sánh xu thế mô phỏng với số liệu đo thực tế tại các trạm quan trắc, mô hình XGB tái hiện rất sát quy luật biến động theo thời gian. Phần lớn các trạm như Tuyên Nhơn, Bến Tre, Đại Ngãi, Vàm Kênh và Trà Kha đều tuân theo quy luật tăng dần từ đầu mùa, đạt cực đại vào tháng 3 và giảm dần trong tháng 4 và 5. Riêng tại trạm Sông Đốc, số liệu thực đo cho thấy độ mặn vẫn duy trì ở mức rất cao và tiếp tục tăng nhẹ vào tháng 5, điều này hoàn toàn đồng nhất với kết quả hiển thị trên bản đồ XGB tháng 5 khi khu vực này vẫn giữ gam màu đỏ sẫm.

Xét về tổng thể, mô hình XGB cho thấy khả năng mô phỏng ưu việt với việc tái hiện chi tiết cấu trúc không gian của xâm nhập mặn, đặc biệt là sự chuyển tiếp mượt mà giữa các vùng mặn và ngọt. Sự tương đồng cao giữa bản đồ mô phỏng và số liệu quan trắc, bao gồm cả việc nắm bắt đúng xu hướng đặc biệt tại trạm Sông Đốc, đã khẳng định độ tin cậy của mô hình trong việc dự báo diễn biến xâm nhập mặn mùa khô.





Hình 16: Bản đồ xâm nhập mặn được xây dựng từ mô hình CB; a) Tháng 1; b) Tháng 2;  
c) Tháng 3; d) Tháng 4; e) Tháng 5 năm 2025

Kết quả mô phỏng theo mô hình CatBoost cho giai đoạn từ tháng 1 đến tháng 5 tại khu vực nghiên cứu diễn ra rõ rệt theo chu kỳ mùa khô, phù hợp với xu thế chung của toàn vùng Đồng bằng sông Cửu Long. Trong tháng 1, mặn chỉ mới xuất hiện rải rác tại các khu vực ven biển thuộc phần phía nam bán đảo Cà Mau và dải ven theo bờ biển Đông Nam. Các tỉnh nội đồng vẫn duy trì mức độ mặn rất thấp, phản ánh giai đoạn đầu mùa khô khi dòng chảy sông Mê Công vẫn còn tương đối dồi dào, đủ sức đẩy lùi ranh giới mặn ra xa cửa sông.

Sang tháng 2, mặn bắt đầu lan rộng hơn dọc theo các tuyến kênh và sông chính hướng ra biển, đặc biệt là khu vực ven biển phía đông và đông nam của vùng nghiên cứu. Diện tích nhiễm mặn tăng lên đáng kể, nhất là tại Cà Mau và các dải ven biển kéo

dài lên phía bắc. Xu thế này phản ánh sự suy giảm dòng chảy từ thượng nguồn vào giai đoạn giữa mùa khô, tạo điều kiện thuận lợi để mặn tiến sâu vào hệ thống cửa sông và nội đồng.

Tháng 3 tiếp tục là giai đoạn cao điểm của xâm nhập mặn, tương tự như ghi nhận thực tế và kết quả của các mô hình khác. Bản đồ mô phỏng cho thấy các vùng ven biển phía đông và đông nam bị bao phủ bởi các dải màu đậm, kéo dài liên tục, trong khi khu vực bán đảo Cà Mau xuất hiện mức độ mặn cao và phân bố trên diện rộng. Tuy cường độ có phần thấp hơn so với ven biển, nhưng một số vùng nội đồng đã xuất hiện dấu hiệu gia tăng nhẹ về nồng độ mặn, phù hợp với điều kiện thời tiết khô hạn kéo dài và lượng nước ngọt bổ sung giảm mạnh trong tháng này.

Từ tháng 4, mô hình bắt đầu mô phỏng xu thế suy giảm khi phạm vi và cường độ mặn dần thu hẹp. Các vùng ven biển vẫn chịu ảnh hưởng mặn nhưng các dải màu đậm không còn lan sâu vào trong nội đồng như tháng 3. Đến tháng 5, mức độ mặn giảm rõ rệt trên phần lớn khu vực, đặc biệt tại các vùng cửa sông Tiền và sông Hậu. Sự suy giảm này hoàn toàn phù hợp với đặc điểm khí hậu khi mưa chuyển mùa bắt đầu xuất hiện, bổ sung lượng nước ngọt và đẩy lùi mặn dần ra phía biển.

Khi so sánh với số liệu thực đo tại các trạm quan trắc, xu thế biến động theo thời gian được mô phỏng bởi mô hình CatBoost cho thấy sự tương đồng khá tốt. Các trạm chịu ảnh hưởng trực tiếp từ biển Đông như Vành Kênh, Trà Kha, Bến Tre đều ghi nhận mức độ mặn tăng mạnh từ tháng 1, đạt đỉnh vào tháng 3 rồi giảm dần từ tháng 4. Riêng đối với trạm Sông Đốc, số liệu thực đo cho thấy độ mặn vẫn duy trì ở mức cao và tăng nhẹ vào tháng 5, điều này cũng được mô hình phản ánh một phần qua việc duy trì các gam màu nóng tại khu vực bán đảo Cà Mau trên bản đồ tháng 5, khác biệt với sự nhạt màu nhanh chóng tại các cửa sông khác.

Mặc dù kết quả hiển thị của CatBoost có độ chi tiết về mặt không gian chưa cao bằng mô hình XGBoost, mô hình vẫn thể hiện tốt xu hướng chung của quá trình xâm nhập mặn và xác định đúng vị trí các khu vực trọng điểm bị ảnh hưởng. Sự ổn định trong diễn biến giữa các tháng cho thấy CatBoost là một công cụ đáng tin cậy để mô phỏng quy luật tăng giảm của mặn theo chu kỳ mùa khô, phục vụ tốt cho công tác đánh giá và quy hoạch tổng thể.

Về mặt tổng thể, mô hình Random Forest cho thấy khả năng mô phỏng phù hợp với quy luật biến động chung của xâm nhập mặn theo mùa. Tuy nhiên, so với hai mô

hình còn lại, kết quả từ RF có độ phân giải chi tiết thấp hơn, ranh giới giữa các vùng mặn – ngọt chưa thực sự sắc nét và độ nhạy đối với các điểm cực trị cục bộ còn hạn chế.

Mô hình CatBoost cho kết quả ở mức trung gian giữa RF và XGBoost. CB vận hành khá ổn định và thể hiện rõ ràng các vùng mặn cao, đặc biệt vào giai đoạn đỉnh điểm tháng 3. Mặc dù vậy, mô hình có xu hướng làm mượt dữ liệu hơn so với thực tế, khiến ranh giới mặn tại một số khu vực ven biển bị mở rộng nhẹ so với quan trắc. Dù vậy, xu thế thời gian và cấu trúc không gian chung của CB vẫn phù hợp với diễn biến thủy văn của vùng.

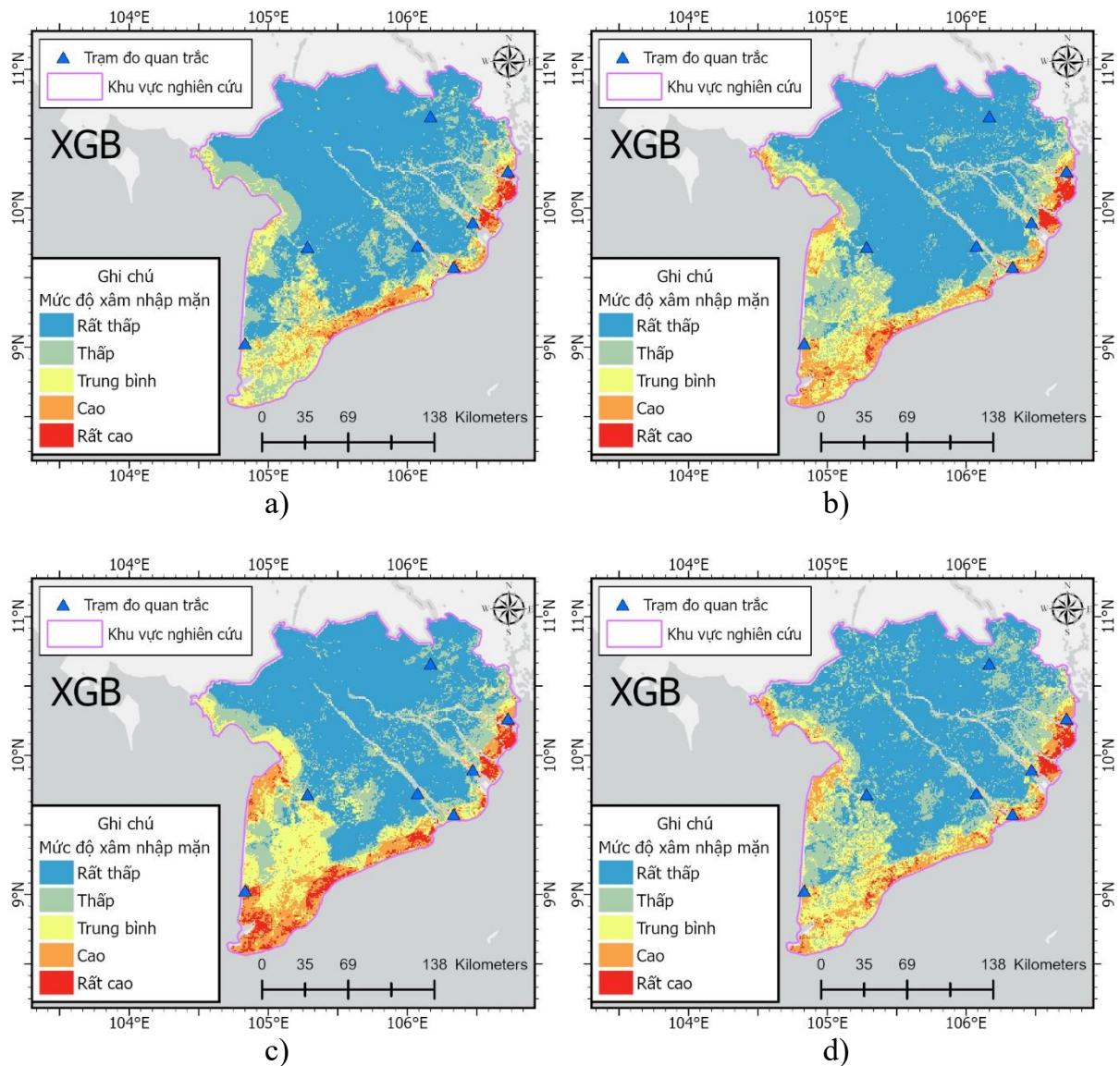
Trong khi đó, XGBoost là mô hình thể hiện hiệu suất vượt trội nhất trong nghiên cứu. Mô hình tái hiện sắc nét các dải mặn ven biển và phân tách tốt sự khác biệt giữa vùng nhiễm mặn nặng và vùng ảnh hưởng nhẹ. XGB đặc biệt nổi bật trong giai đoạn tháng 3 khi mô phỏng chính xác các tâm mặn cao tại Bến Tre, Trà Vinh, Bạc Liêu và Cà Mau. Xu hướng suy giảm mặn trong tháng 4 và 5 cũng được biểu diễn hợp lý: phần lớn diện tích giảm mặn nhanh, riêng khu vực bán đảo Cà Mau vẫn duy trì cảnh báo mặn cao đúng với thực tế khó thoát mặn của vùng này. Đây là mô hình đạt sự cân bằng tốt nhất giữa độ chính xác, độ ổn định và khả năng mô phỏng chi tiết.

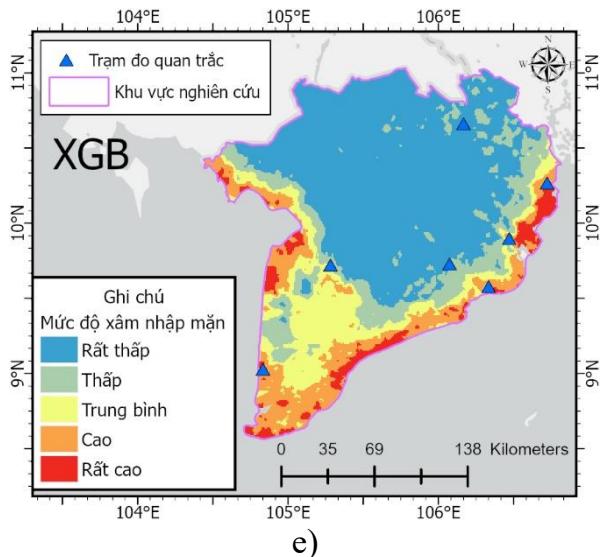
Kết luận, cả ba mô hình đều mô phỏng thành công chu kỳ tăng – đạt đỉnh – giảm của xâm nhập mặn mùa khô. Trong đó, XGB thể hiện chất lượng dự báo tốt nhất và chi tiết nhất; CB mang lại kết quả ổn định và thuyết phục, trong khi RF cho kết quả ở mức độ chấp nhận được nhưng hạn chế về chi tiết không gian. Sự thống nhất cao giữa kết quả mô phỏng của XGBoost và dữ liệu quan trắc tại các trạm là cơ sở vững chắc để đề xuất sử dụng XGBoost làm mô hình chính trong nghiên cứu và các ứng dụng dự báo thực tế.

Sau khi có được mô hình tốt nhất, ở đây là mô hình XGBoost, bản đồ độ mặn đát mô phỏng được tiếp tục đưa vào bước phân lớp để xây dựng bản đồ phân vùng nguy cơ xâm nhập mặn. Trong nghiên cứu này, phương pháp phân lớp Jenks[49] được sử dụng nhằm đảm bảo việc chia ngưỡng phản ánh đúng nhất đặc trưng phân bố của dữ liệu. Jenks là phương pháp phân loại tối ưu hóa giới hạn lớp dựa trên cấu trúc “tự nhiên” của tập dữ liệu, bằng cách giảm thiểu phương sai trong từng lớp và đồng thời tối đa hóa phương sai giữa các lớp. Điều này cho phép các giá trị có đặc điểm tương đồng được nhóm lại với nhau, trong khi các giá trị khác biệt được tách rõ thành lớp khác, giúp bản đồ thể hiện đúng sự phân hóa mức độ mặn trên không gian[49].

Áp dụng phương pháp Jenks đối với bản đồ mặn từ mô hình XGBoost giúp xác định các ngưỡng phân lớp phù hợp với phân bố thực tế của độ mặn trong khu vực nghiên cứu. Các giá trị mô phỏng liên tục được phân chia thành những nhóm có ý nghĩa, thường bao gồm các mức: rất thấp, thấp, trung bình, cao và rất cao. Nhờ đó, bản đồ trở nên trực quan, phản ánh đúng cấu trúc không gian của hiện tượng xâm nhập mặn, đồng thời dễ dàng sử dụng trong phân tích và ra quyết định.

Dưới đây là bản đồ phân bố mức độ xâm nhập mặn trên mô hình XGB





Hình 17: Bản đồ mức độ xâm nhập mặn được xây dựng từ mô hình XGB; a) Tháng 1; b) Tháng 2; c) Tháng 3; d) Tháng 4; e) Tháng 5 năm 2025

Phân bố không gian sau phân lớp Jenks thể hiện rõ xu thế xâm nhập mặn đặc trưng mùa khô ở khu vực nghiên cứu. Các vùng ven biển, đặc biệt là khu vực cửa sông phía Đông, Đông Nam và toàn bộ dài ven biển bán đảo Cà Mau, xuất hiện diện tích lớn thuộc nhóm mặn cao và rất cao. Điều này phù hợp với vị trí địa lý giáp biển, nơi chịu tác động trực tiếp của thủy triều và sự suy giảm dòng chảy từ thượng nguồn. Ngược lại, các khu vực nội đồng nằm sâu trong đất liền chủ yếu nằm trong nhóm rất thấp đến thấp, phản ánh đặc điểm địa hình và khoảng cách xa biển giúp các khu vực này được bảo vệ tốt hơn trước sự xâm nhập của mặn. Việc phân lớp bằng Jenks góp phần làm cho bản đồ trở nên trực quan, dễ quan sát và thể hiện rõ sự phân bậc của độ mặn, hỗ trợ hiệu quả cho công tác đánh giá mức độ rủi ro và lập kế hoạch quản lý đất đai trong điều kiện biến đổi khí hậu.

Bên cạnh biểu diễn không gian, việc thống kê diện tích theo từng mức phân lớp cho từng tháng là bước quan trọng nhằm đánh giá định lượng mức độ ảnh hưởng của xâm nhập mặn trong mùa khô. Thông qua việc tính toán diện tích của năm nhóm phân lớp từ tháng 1 đến tháng 5, nghiên cứu có thể theo dõi sự biến động của phạm vi mặn theo thời gian, xác định thời điểm mặn đạt cực đại cũng như giai đoạn bắt đầu suy giảm khi chuyển sang mùa mưa. Các giá trị diện tích này đóng vai trò minh chứng quan trọng cho việc nhận diện xu thế mở rộng hay thu hẹp của từng mức độ rủi ro, đồng thời giúp đánh giá mức độ tích tụ muối trong đất theo từng giai đoạn của mùa khô. Kết quả thống kê chi tiết của năm mức phân lớp cho từng tháng được trình bày trong bảng dưới đây,

cho phép quan sát rõ các giai đoạn mặn mạnh, thời kỳ ổn định cũng như thời điểm suy giảm khi mùa mưa bắt đầu xuất hiện. Bảng thống kê được trình bày dưới đây:

Mô hình XGB	Rất thấp ( $\text{km}^2$ )	Thấp( $\text{km}^2$ )	Trung bình( $\text{km}^2$ )	Cao( $\text{km}^2$ )	Rất cao( $\text{km}^2$ )
Tháng 1	23737	7127.187	4279.996	2046.771	504.8982
Tháng 2	23134.64	6515.174	4212.425	3112.321	741.9372
Tháng 3	20977.46	5994.596	5694.167	3587.948	1472.843
Tháng 4	20460.84	8660.262	4896.387	2962.145	739.0601
Tháng 5	21387.01	5918.852	5511.855	4843.809	1737.724

Bảng 6: Bảng thống kê diện tích xâm nhập mặn

Số liệu thống kê diện tích theo từng mức phân lớp của mô hình XGBoost trong giai đoạn từ tháng 1 đến tháng 5 đã làm rõ xu hướng biến động cụ thể tại từng khu vực. Đối với nhóm diện tích có độ mặn rất thấp, đại diện cho các vùng sinh thái ngọt hóa tại An Giang, Đồng Tháp và Cần Thơ, số liệu cho thấy sự thu hẹp liên tục và đáng kể trong giai đoạn đầu mùa khô. Diện tích này giảm từ mức  $23737 \text{ km}^2$  vào tháng 1 xuống mức thấp nhất là  $20460 \text{ km}^2$  vào tháng 4. Xu hướng sụt giảm này minh chứng cho quá trình lấn dần của nước mặn từ các cửa sông thuộc Vĩnh Long và Cà Mau vào sâu trong nội đồng khi dòng chảy nước ngọt từ thượng nguồn suy giảm. Phải đến tháng 5, diện tích vùng ngọt tại các tỉnh thượng nguồn này mới bắt đầu có dấu hiệu hồi phục và tăng nhẹ trở lại lên mức  $21387 \text{ km}^2$  nhờ những cơn mưa chuyển mùa đầu tiên giúp đẩy lùi ranh giới mặn.

Trái ngược hoàn toàn với xu hướng tại vùng nội đồng, nhóm diện tích nhiễm mặn cao và rất cao phân bố tập trung tại vùng Cà Mau và dải ven biển Vĩnh Long lại thể hiện sự gia tăng khắc nghiệt. Diện tích đất nhiễm mặn ở mức rất cao đã tăng gấp gần ba lần chỉ sau hai tháng đầu năm, từ  $504 \text{ km}^2$  vào tháng 1 lên  $1472 \text{ km}^2$  trong tháng 3. Đây là giai đoạn đỉnh điểm của xâm nhập mặn khi các yếu tố bất lợi như mực nước sông xuống thấp và bốc hơi mạnh khiến muối tích tụ nhanh chóng trên diện rộng tại các vùng cửa sông và bãi bồi ven biển.

Đặc biệt, xu hướng biến động vào cuối mùa khô tại khu vực Cà Mau diễn ra rất phức tạp và khác biệt so với các vùng khác. Mặc dù tháng 4 ghi nhận sự sụt giảm tạm thời của diện tích nhiễm mặn rất cao xuống còn  $739 \text{ km}^2$  do tác động rửa trôi bề mặt của các đợt mưa dông cục bộ, nhưng sang tháng 5 diện tích này lại bất ngờ tăng vọt và thiết lập mức đỉnh mới cao nhất trong cả chuỗi khảo sát là  $1737 \text{ km}^2$ . Hiện tượng này phản ánh cơ chế tích tụ muối đặc thù tại vùng bán đảo Cà Mau, nơi nền nhiệt độ cao vào tháng 5 làm gia tăng quá trình bốc hơi, kéo theo nước ngầm chứa muối từ dưới sâu thâm thấu

ngược lên bề mặt. Đồng thời, do đặc thù địa hình trũng thấp và hệ thống thủy lợi khép kín, độ trễ của quá trình rửa mặn khiến lượng muối tồn đọng trong đất tại đây chưa thể được giải phóng ngay lập tức.

Tổng hợp lại, kết quả phân tích cho thấy sự phân hóa sâu sắc về xu hướng giữa các vùng địa lý theo ranh giới mới. Trong khi nhóm các tỉnh nội đồng như An Giang, Đồng Tháp bắt đầu phục hồi độ ngọt và mở rộng diện tích canh tác an toàn vào tháng 5 thì vùng Cà Mau và ven biển Vĩnh Long vẫn đối mặt với rủi ro mặn gia tăng cực đại. Việc xác định rõ vị trí và thời điểm diện tích nhiễm mặn nặng đạt đỉnh vào tháng 5 là cơ sở quan trọng để các nhà quản lý khuyến cáo thận trọng trong việc bố trí mùa vụ tại các vùng ven biển nhạy cảm, tránh xuống giống khi đất chưa thực sự được ngọt hóa hoàn toàn.

## CHƯƠNG 4: KẾT LUẬN VÀ ĐỀ XUẤT

Nghiên cứu đã xây dựng được mô hình dự báo độ mặn đất cho khu vực Đồng bằng sông Cửu Long dựa trên bộ dữ liệu viễn thám đa nguồn kết hợp với các thuật toán học máy. Trong ba mô hình được kiểm định, mô hình XGBoost cho thấy hiệu quả cao nhất và được lựa chọn để xây dựng bản đồ phân vùng nguy cơ xâm nhập mặn cho toàn vùng nghiên cứu. Kết quả đánh giá theo kiểm định chéo với các chỉ số RMSE, MAE và hệ số tương quan cho thấy mô hình đạt độ chính xác cao, mô phỏng tốt sự biến động theo không gian và thời gian của độ mặn đất trong mùa khô.

Kết quả phân tích tầm quan trọng của các biến cho thấy hai chỉ số quang học NDSI và NDVI, cùng với biến khoảng cách đến biển, là ba yếu tố có ảnh hưởng mạnh nhất tới khả năng dự báo độ mặn. Đây đều là các biến phản ánh trực tiếp cấu trúc bề mặt và mức độ tác động của nguồn mặn biển vào đất liền. Đáng chú ý, hệ số phản xạ bề mặt trích xuất từ dữ liệu CYGNSS cũng nằm trong nhóm các biến có mức đóng góp lớn, đứng thứ tư trong số toàn bộ các biến được sử dụng. Điều này chứng tỏ tín hiệu GNSS có khả năng phản ánh tốt sự thay đổi điện môi liên quan đến độ ẩm và độ mặn bề mặt đất. So với dữ liệu quang học vốn dễ bị suy giảm chất lượng trong điều kiện mây che phủ, việc hệ số phản xạ thể hiện mức ảnh hưởng ổn định là minh chứng rõ ràng cho tiềm năng của dữ liệu GNSS trong giám sát xâm nhập mặn, đặc biệt tại các vùng ven biển nhiệt đới.

Bản đồ mô phỏng độ mặn theo tháng cho thấy sự phân hóa không gian rất rõ trong mùa khô. Các khu vực ven biển phía Đông và Đông Nam của vùng nghiên cứu – đặc biệt là dải ven biển Cà Mau và khu vực cửa sông thuộc Vĩnh Long – duy trì mức độ mặn cao đến rất cao do chịu tác động trực tiếp của thủy triều và đặc điểm địa mạo thấp. Trong khi đó, vùng nội đồng phía Tây gồm An Giang, Đồng Tháp và phần lớn Cần Thơ hầu như luôn duy trì mức mặn thấp, phù hợp với điều kiện phù sa ngọt, mạng lưới kênh cấp nước ổn định và khoảng cách xa biển. Sự phân hóa này đồng thời phản ánh đúng cơ cấu sản xuất trong khu vực: vùng ven biển tập trung vào nuôi trồng thủy sản, còn vùng nội đồng là khu vực chuyên canh cây ăn trái và sản xuất lúa chất lượng cao. Kết quả mô hình cũng cho thấy hiện tượng trễ giữa đỉnh mặn trong nước (thường xảy ra vào tháng 3) và đỉnh mặn tích tụ trong đất (rõ nhất vào tháng 4–5). Điều này phù hợp với quy luật thủy văn mùa khô, khi độ mặn nước sông suy giảm trước do mưa đầu mùa, trong khi lượng muối đã thẩm thấu vào đất cần thời gian lâu hơn để được rửa trôi.

Bên cạnh các kết quả tích cực, nghiên cứu vẫn còn tồn tại một số hạn chế. Một số khu vực nội đồng được mô hình dự báo chưa thật sự ổn định do mật độ điểm đo thực địa còn thấp, đặc biệt ở xa khu vực cửa sông. Độ phân giải không gian của các lớp dữ

liệu vệ tinh như CYGNSS và MODIS chưa đủ chi tiết để mô tả các biến động mặn quy mô nhỏ. Ngoài ra, ảnh hưởng của hệ thống thủy lợi, công đập và các hoạt động điều tiết nước không được mô tả trong dữ liệu đầu vào, gây ra sai khác cục bộ ở một số thời điểm. Các yếu tố khí hậu như El Niño hoặc La Niña cũng có thể làm thay đổi mạnh độ mặn nhưng chưa được phân tích theo chuỗi nhiều năm.

Trong thời gian tới, nghiên cứu có thể được cải thiện bằng cách tăng cường thu thập số liệu thực địa, đặc biệt tại các vùng nội đồng; tích hợp thêm dữ liệu radar để mô tả tốt hơn cấu trúc bề mặt; đưa vào các biến thủy văn ; nâng cao độ phân giải dữ liệu viễn thám thông qua các kỹ thuật hạ quy mô; và thử nghiệm các mô hình học sâu để khai thác tốt hơn quan hệ phi tuyến giữa độ mặn, độ ẩm và phản xạ bề mặt. Những hướng phát triển này sẽ góp phần nâng cao độ chính xác của mô hình, đồng thời tăng giá trị ứng dụng trong công tác giám sát, dự báo và quản lý xâm nhập mặn tại Đồng bằng sông Cửu Long trong bối cảnh biến đổi khí hậu ngày càng gia tăng.

Việc lập bản đồ giám sát xâm nhập mặn với độ chính xác cao giữ vai trò đặc biệt quan trọng trong công tác quản lý tài nguyên nước, quy hoạch sử dụng đất và đảm bảo sinh kế bền vững cho cộng đồng, nhất là trong bối cảnh biến đổi khí hậu và suy giảm dòng chảy thượng nguồn ngày càng nghiêm trọng. Mục tiêu phát triển một phương pháp tiếp cận hiện đại dựa trên dữ liệu viễn thám đa nguồn và các thuật toán học máy như Random Forest, CatBoost và XGBoost để dự báo và phân vùng nguy cơ xâm nhập mặn cho toàn vùng Đồng bằng sông Cửu Long đã được hoàn thành. Những kết quả thu được khẳng định tiềm năng lớn của học máy và dữ liệu vệ tinh trong việc mô phỏng diễn biến mặn theo không gian và thời gian, cho phép xác định chính xác các khu vực chịu rủi ro cao và các vùng được bảo vệ tốt hơn.

Các phát hiện của nghiên cứu không chỉ đóng góp thiết thực cho Việt Nam, nơi xâm nhập mặn đang gây ra nhiều thách thức đối với nông nghiệp, thủy sản và nguồn nước sinh hoạt, mà còn có giá trị tham khảo cho các khu vực ven biển khác trên thế giới đang chịu tác động tương tự. Bản đồ được xây dựng từ mô hình cho phép nhà quản lý nhận diện rõ các khu vực cần ưu tiên các giải pháp công trình và phi công trình, điều chỉnh cơ cấu cây trồng, bố trí lại không gian sản xuất và nâng cao khả năng thích ứng của cộng đồng. Hy vọng rằng kết quả nghiên cứu có thể trở thành cơ sở quan trọng hỗ trợ các nhà hoạch định chính sách, cơ quan quản lý và địa phương triển khai các biện pháp phù hợp nhằm giảm thiểu thiệt hại do xâm nhập mặn, hướng tới quản lý tài nguyên nước hiệu quả và phát triển bền vững cho toàn vùng Đồng bằng sông Cửu Long.

## TÀI LIỆU THAM KHẢO

- [1] T. G. Nguyen, N. A. Tran, P. L. Vu, Q.-H. Nguyen, H. D. Nguyen, and 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 Reg.*, vol. 27, p. e00424, Dec. 2021, doi: 10.1016/j.geodrs.2021.e00424.
- [2] F. Khormali, M. Ajami, S. Ayoubi, Ch. Srinivasarao, and S. P. Wani, “Role of deforestation and hillslope position on soil quality attributes of loess-derived soils in Golestan province, Iran,” *Agric. Ecosyst. Environ.*, vol. 134, no. 3, pp. 178–189, Dec. 2009, doi: 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 Environ. Sci.*, Jan. 2011, doi: 10.1039/c1ee01029h.
- [4] T. Gorji, E. Sertel, and A. Tanik, “Monitoring soil salinity via remote sensing technology under data scarce conditions: A case study from Turkey,” *Ecol. Indic.*, vol. 74, pp. 384–391, Mar. 2017, doi: 10.1016/j.ecolind.2016.11.043.
- [5] “Bộ TT Truyền Thông - Tài nguyên và Môi trường.”
- [6] “Electrical Conductivity Methods for Measuring and Mapping Soil Salinity,” in *Advances in Agronomy*, vol. 49, Academic Press, 1993, pp. 201–251. doi: 10.1016/S0065-2113(08)60795-6.
- [7] H. T. Nguyen, A. K. Nguyen, Y.-A. Liou, P. P. Hoang, and H. T. Nguyen, “Estimation of Salinity Intrusion by Using Landsat 8 OLI Data in The Mekong Delta, Vietnam,” Aug. 17, 2018, *Preprints*: 2018080301. doi: 10.20944/preprints201808.0301.v1.
- [8] Y. U. Haq, M. Shahbaz, S. Asif, K. Ouahada, and 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, p. 8121, Jan. 2023, doi: 10.3390/s23198121.
- [9] V. V. Salomonson, “Remote Sensing, Historical Perspective,” in *Encyclopedia of Remote Sensing*, Springer, New York, NY, 2014, pp. 684–691. doi: 10.1007/978-0-387-36699-9\_158.
- [10] Michael N. Demers and Teri Library, *Fundamentals Of Geographic Information Systems Michael N. Demers*. Teri Publication, 1997
- [11] A. M. Michael, *Irrigation Theory And Practice - 2Nd Edn: Theory and Practice*. Vikas Publishing House, 2009.
- [12] M. Seifi, A. Ahmadi, M.-R. Neyshabouri, R. Taghizadeh-Mehrjardi, and H.-A. Bahrami, “Remote and Vis-NIR spectra sensing potential for soil salinization estimation in the

eastern coast of Urmia hyper saline lake, Iran,” *Remote Sens. Appl. Soc. Environ.*, vol. 20, p. 100398, Nov. 2020, doi: 10.1016/j.rsase.2020.100398.

- [13] Decock, C., Lee, J., Necpalova, M., Pereira, E. I. P., Tendall, D. M., and Six, J.: Mitigating N<sub>2</sub>O emissions from soil: from patching leaks to transformative action, *SOIL*, 1, 687–694, <https://doi.org/10.5194/soil-1-687-2015>, 2015.
- [14] P. Rengasamy, “World salinization with emphasis on Australia,” *J. Exp. Bot.*, vol. 57, no. 5, pp. 1017–1023, Mar. 2006, doi: 10.1093/jxb/erj108.
- [15] E. C. Brevik *et al.*, “The interdisciplinary nature of *SOIL*,” *SOIL*, vol. 1, no. 1, pp. 117–129, Jan. 2015, doi: 10.5194/soil-1-117-2015.
- [16] A. Peña, L. Delgado-Moreno, and J. A. Rodríguez-Liébana, “A review of the impact of wastewater on the fate of pesticides in soils: Effect of some soil and solution properties,” *Sci. Total Environ.*, vol. 718, p. 134468, May 2020, doi: 10.1016/j.scitotenv.2019.134468.
- [17] M. R. K. Manasa, N. R. Katukuri, S. S. Darveekaran Nair, Y. Haojie, Z. Yang, and R. bo Guo, “Role of biochar and organic substrates in enhancing the functional characteristics and microbial community in a saline soil,” *J. Environ. Manage.*, vol. 269, p. 110737, Sept. 2020, doi: 10.1016/j.jenvman.2020.110737.
- [18] D. G. Metternicht and D. A. Zinck, Eds., *Remote Sensing of Soil Salinization: Impact on Land Management*. Boca Raton: CRC Press, 2008. doi: 10.1201/9781420065039.
- [19] “Bộ TT Truyền Thông - Tài nguyên và Môi trường.”
- [20] I. El Naqa and M. J. Murphy, “What Is Machine Learning?,” in *Machine Learning in Radiation Oncology: Theory and Applications*, I. El Naqa, R. Li, and M. J. Murphy, Eds., Cham: Springer International Publishing, 2015, pp. 3–11. doi: 10.1007/978-3-319-18305-3\_1.
- [21] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
- [22] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD ’16. New York, NY, USA: Association for Computing Machinery, Tháng Tám 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
- [23] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” Jan. 20, 2019, *arXiv*: arXiv:1706.09516. doi: 10.48550/arXiv.1706.09516.

- [24] “The Reflected Global Navigation Satellite System (GNSS-R): from Theory to Practice,” in *Microwave Remote Sensing of Land Surface*, Elsevier, 2016, pp. 303–355. doi: 10.1016/B978-1-78548-159-8.50007-4.
- [25] C. Ruf, S. Gleason, A. Ridley, R. Rose and J. Scherrer, "The nasa cygnss mission: Overview and status update," *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Fort Worth, TX, USA, 2017, pp. 2641-2643, doi: 10.1109/IGARSS.2017.8127537.
- [26] S. Aksoy, E. Sertel, R. Roscher, A. Tanik, and N. Hamzehpour, “Assessment of soil salinity using explainable machine learning methods and Landsat 8 images,” *Int. J. Appl. Earth Obs. Geoinformation*, vol. 130, p. 103879, June 2024, doi: 10.1016/j.jag.2024.103879.
- [27] G. Ma, J. Ding, L. Han, Z. Zhang, and S. Ran, “Digital mapping of soil salinization based on Sentinel-1 and Sentinel-2 data combined with machine learning algorithms,” *Reg. Sustain.*, vol. 2, no. 2, pp. 177–188, Apr. 2021, doi: 10.1016/j.regsus.2021.06.001.
- [28] V. H. Nguyen, J. Germer, V. N. Duong, and F. Asch, “Soil resistivity measurements to evaluate subsoil salinity in rice production systems in the Vietnam Mekong Delta,” *Surf. Geophys.*, vol. 21, no. 4, pp. 288–299, July 2023, doi: 10.1002/nsg.12260.
- [29] Hoa, P. V., Giang, N. V., Binh, N. A., Hai, L. V. H., Pham, T.-D., Hasanlou, M., & Tien Bui, D. (2019). Soil Salinity Mapping Using SAR Sentinel-1 Data and Advanced Machine Learning Algorithms: A Case Study at Ben Tre Province of the Mekong River Delta (Vietnam). *Remote Sensing*, 11(2), 128. <https://doi.org/10.3390/rs11020128>.
- [30] C. C. Chew and E. E. Small, “Soil Moisture Sensing Using Spaceborne GNSS Reflections: Comparison of CYGNSS Reflectivity to SMAP Soil Moisture,” *Geophys. Res. Lett.*, vol. 45, no. 9, pp. 4049–4057, 2018, doi: 10.1029/2018GL077905.
- [31] S. Zhang *et al.*, “Use of reflected GNSS SNR data to retrieve either soil moisture or vegetation height from a wheat crop,” *Hydrol. Earth Syst. Sci.*, vol. 21, no. 9, pp. 4767–4784, Sept. 2017, doi: 10.5194/hess-21-4767-2017.
- [32] M. C. Ha, “Evolution of soil moisture and analysis of fluvial altimetry using GNSS-R,” Theses, Université Paul Sabatier - Toulouse III, 2018.
- [33] V. P. Lan *et al.*, “Application of GNSS Reflectometry in Water Level Monitoring using Low-cost GNSS Antenna: A Case Study in Tam Giang Lagoon, Thua Thien Hue Province,” *VNU J. Sci. Earth Environ. Sci.*, vol. 38, no. 4, Dec. 2022, doi: 10.25073/2588-1094/vnuees.4878.
- [34] P. L. Vu *et al.*, “Demonstrating the Potential of Low-Cost GNSS Receiver for tidal monitoring, storms, and flood detecting: example of 2022 Noru Storm in Thua Thien

Hue province, Vietnam,” *Acta Montan. Slovaca*, vol. 28, no. 4, pp. 1034–1046, 2023, doi: 10.46544/ams.v28i4.19.

[35] H. D. Nguyen *et al.*, “Flood susceptibility mapping using advanced hybrid machine learning and CyGNSS: a case study of Nghe An province, Vietnam,” *Acta Geophys.*, vol. 70, no. 6, pp. 2785–2803, Dec. 2022, doi: 10.1007/s11600-022-00940-2.

[36] J. Wang *et al.*, “A novel retrieval model for soil salinity from CYGNSS: Algorithm and test in the Yellow River Delta,” *Geoderma*, vol. 432, p. 116417, Apr. 2023, doi: 10.1016/j.geoderma.2023.116417.

[37] Shao, Z., Ahmad, M. N., & Javed, A. (2024). Comparison of Random Forest and XGBoost Classifiers Using Integrated Optical and SAR Features for Mapping Urban Impervious Surface. *Remote Sensing*, 16(4), 665. <https://doi.org/10.3390/rs16040665>

[38] A. Calabia, I. Molina, and S. Jin, “Soil Moisture Content from GNSS Reflectometry Using Dielectric Permittivity from Fresnel Reflection Coefficients,” *Remote Sens.*, vol. 12, no. 1, p. 122, Jan. 2020, doi: 10.3390/rs12010122.

[39] Wang J D, Sun Z G, Yang T, Zhu K Y, Shao C X, Peng J B, Li S J, Wang W Y, Gao Y N and Yue H Y. 2023. A remote sensing method for retrieving soil salinity based on CYGNSS : Taking the Yellow River Delta as an example. National Remote Sensing Bulletin, 27 (2) : 351-362 DOI : 10.11834/jrs.20210466.

[40] P. Zeiger, F. Frappart, J. Darrozes, C. Prigent, and C. Jiménez, “Analysis of CYGNSS coherent reflectivity over land for the characterization of pan-tropical inundation dynamics,” *Remote Sens. Environ.*, vol. 282, p. 113278, Dec. 2022, doi: 10.1016/j.rse.2022.113278.

[41] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto, and S. Ridella, “The ‘K’ in K-fold Cross Validation,” *Comput. Intell.*, 2012.

[42] T. O. Hodson, “Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not,” *Geosci. Model Dev.*, vol. 15, no. 14, pp. 5481–5487, July 2022, doi: 10.5194/gmd-15-5481-2022.

[43] P. Sedgwick, “Pearson’s correlation coefficient,” *BMJ*, vol. 345, p. e4483, July 2012, doi: 10.1136/bmj.e4483.

[44] BỘ XÂY DỰNG TƯ VẤN LẬP QUY HOẠCH : VIỆN QUY HOẠCH XÂY DỰNG MIỀN NAM VÀ TƯ VẤN QUỐC TẾ RUA”.

[45] Eroglu, O., Kurum, M., Boyd, D., & Gurbuz, A. C. (2019). High Spatio-Temporal Resolution CYGNSS Soil Moisture Estimates Using Artificial Neural Networks. *Remote Sensing*, 11(19), 2272. <https://doi.org/10.3390/rs11192272>.

- [46] 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,” *Environ. Sci. Pollut. Res.*, vol. 30, no. 29, pp. 74340–74357, June 2023, doi: 10.1007/s11356-023-27516-x.
- [47] H. D. Nguyen, V. T. Pham, Q.-H. Nguyen, and Q.-T. Bui, “Soil salinity prediction using satellite-based variables and machine learning: Case study in Tra Vinh province, Mekong Delta, Vietnam,” *Vietnam J. Earth Sci.*, vol. 47, no. 2, pp. 201–219, Feb. 2025, doi: 10.15625/2615-9783/22438.
- [48] “Trung Tâm KTTV Nam Bộ.”
- [49] J. Chen, S. T. Yang, H. W. Li, B. Zhang, and J. R. Lv, “Research on Geographical Environment Unit Division Based on the Method of Natural Breaks (Jenks),” *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. XL-4-W3, pp. 47–50, Nov. 2013, doi: 10.5194/isprsarchives-XL-4-W3-47-2013.