Vietnam Journal of Earth Sciences, 47(2), 216-234, https://doi.org/ 10.15625/2615-9783/22438

**Vietnam Journal of Earth Sciences**
https://vjs.ac.vn/index.php/jse

# Soil Salinity Prediction Using Satellite-Based Variables and Machine Learning: Case study in Tra Vinh province, Mekong Delta, Vietnam

Huu Duy Nguyen, Viet Thanh Pham, Quoc-Huy Nguyen*, Quang-Thanh Bui

*Faculty of Geography, University of Science, Vietnam National University, Ha Noi, 334 Nguyen Trai, Thanh Xuan district, Hanoi, Vietnam*

Received 22 January 2025; Received in revised form 04 February 2025; Accepted 21 February 2025

ABSTRACT

The precision of estimating soil salinity is considered a key task in solving soil salinity problems and irrigation management of agriculture. This problem is increasingly important in the Mekong Delta, where it is severely affected by this phenomenon in the context of climate variability. Therefore, this paper aims to construct a soil salinity map with high accuracy using machine learning and Sentinel 2A, namely Xgboost (XGB) and Random Forest (RF). The province of Tra Vinh in the Mekong Delta has been selected as the case study. 68 soil salinity samples were collected in August 2024, and 25 conditioning factors extracted from the Sentinel 2A image were used as input data for the machine-learning model. Three statistical indices, namely root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination ($R^2$), were used to evaluate the effectiveness of machine learning models. The results showed that with an $R^2$ value of 0.86, the XGB model was superior to the RF model with an R2 value of 0.67.

Furthermore, Tra Vinh province, the coastal region, and along the Mekong River are more severely affected by soil salinity with an electrical conductivity (EC) value of more than 10. This region, more affected by soil salinity, is related to rising tides and sea levels in the context of climate variability. This study plays an important role and can support farmers in regions affected by soil salinity in building investment measures to reduce the impacts of soil salinity on the development of agriculture.

*Keywords:* Soil salinity, Tra Vinh province, Mekong delta, machine learning.

## 1. Introduction

Soil salinity is considered the most serious environmental problem, causing significant effects on ecosystem health, soil property, and cultivation growth (Vermeulen and Van Niekerk 2017, Wang and Sun 2024). According to FAO, about 20% of irrigated farmlands and agricultural ecosystems worldwide have been affected by soil salinity (Xiao, Ji et al., 2023). Soil salinity results from very complex processes related to hydrological, climatic regimes, groundwater exploitation, and human activities. Anthropoid activities such as plowing in a natural element characterized by low rainfall, high evaporation, and high groundwater level make

*Corresponding author, Email: huyquoc2311@hus.edu.vn

croplands more affected by the soil salinity problem (Vermeulen and Van Niekerk 2017, Wang and Sun 2024). The Mekong Delta is the delta situation most affected by soil salinity in the world. According to data from the Ministry of Agriculture and Rural Development, soil salinity affected approximately 42.5% of natural areas and 430,000 people in the dry season of 2019-2020 (Tran, Tsujimura et al. 2021). Among them, the province of Tra Vinh was the most affected by soil salinity. According to the Department of Agriculture and Rural Development of Tra Vinh province, saltwater intrusion during the dry season of 2019-2020 caused about 1,000 billion VND of damage; among them, rice suffered the heaviest damage, with 919 billion VND. In addition, dozens of hectares of crops and more than 271 hectares of fruit trees in the province were also damaged, accounting for more than 30% of the area. In the context of population growth and urbanization, the request for natural resources and food is growing; it requires more land to develop agriculture. Therefore, careful monitoring, quantitative evaluation, and construction of the spatial distribution map of soil salinity are considered an important task to support decision makers or local authorities in managing land resources to develop agriculture to ensure nutrition security in the country (Nguyen, Liou, et al. 2020, Nguyen, Germer, et al. 2024).

Acquiring timely and accurate soil salinity information plays an essential role in preventing and controlling the soil salinity situation to ensure food security in the region. Soil salinity monitoring is considered the first step to reveal soil salinity occurrence (Sarkar, Rudra et al. 2023, Wang and Sun 2024). The literature review shows several methods to assess soil salinity from traditional to modern, including physical analysis-based, physically-based, remote sensing, and data-driven models. Physical analysis-based methods provide knowledge about the saltwater movement by monitoring solute concentrations in a specific period (Ren, Wei et al. 2019). This method is often combined with the physically-based model to understand the saltwater movement process better. Although the methods based on the physical analysis are beneficial for understanding the process of saltwater intrusion in a short time, this method has been applied only in the case of a simple saltwater intrusion process. In reality, this process is very complex. Therefore, several studies have used physically-based models such as MODFLOW, SEAWAT, and MIKE… which are considered the alternative method to examine saltwater intrusion into coastal multi-aquifer systems in the context of climate change and human activities (Mirlas 2012, Keilholz, Disse et al. 2015, Dunlop, Palanichamy et al. 2019). In recent years, physically based models have been widely used to model the dynamics of groundwater and saltwater intrusion into coastal multi-aquifer systems. These models are based on detailed descriptions and provide a mechanistic understanding of the physical processes not only at the small scale but also at a large scale. However, constructing these models requires detailed data and, therefore, is not always possible.

Remote sensing has been widely utilized in soil salinity spatial distribution map construction at local and regional levels. Sentinel 2 satellite images with temporal resolution, multiple frequency bands, and high spatial resolution have been widely used in many previous studies (Metternicht and Zinck 2003; Wu, Mhaimeed et al. 2014; Gorji, Sertel et al. 2017). Due to its higher spatial and temporal resolution, the multispectral instrument (MSI) outperforms the operational land imager (OLI) in soil salinity monitoring. Furthermore, OLI tends to overestimate the

area of salinized land and saline areas with vegetation. However, many studies have compared the accuracy of these two sensors in monitoring soil salinity, and the results show similar accuracy (Chaaou, Chikhaoui, et al. 2024, Sirpa-Poma, Satgé, et al. 2024). However, the results also demonstrate that the different salinity levels in the electrical conductivity (EC) are considered using regression models. Soil salinity monitoring using remote sensing and GIS data can be limited by atmosphere, clouds, vegetation, and temporal resolution. However, these limitations can be overcome using more sensor data and field data. However, along with the rapid development of remote sensing data, research and development of new methods are needed to process these data accurately and efficiently.

In recent years, several researchers have developed the data drive model for soil salinity studies (Khanh, Ngoc et al. 2024, Wang and Sun 2024). The data drive model is based on the correlations between covariates and dependent variables to predict the spatial distributions of soil salinity. The models of soil salinity prediction can be separated into linear and non-linear regression models. Linear regression models include partial least squares regression (PLSR) (Zeng, Zhang et al. 2018) and inverse density weighted regression (IDW) (Zhao, Cao et al. 2019), which have been applied to predict soil salinity. However, most linear models present poor precision in areas with high spatial salinity variability. Recently, machine learning has been developed to build a soil salinity map in different regions. Commonly used algorithms include support vector machine (Jiang, Rusuli, et al. 2019), random forest (Fathizad, Ardakani, et al. 2020), Xgboost (Aksoy, Sertel, et al. 2024), Catboost (Mantena, Mahammood, et al. 2023), convolutional neural network (CNN) (Garajeh, Malakyar, et al. 2021). Salinity intrusion is very complex and has been controlled by several factors. Therefore, linear models are complicated to simulate in real situation, while nonlinear models can better fit the contributions of various factors of soil salinity. However, selecting appropriate models in some models is considered one of the difficult tasks. In addition, researchers have received attention to the challenges of predicting soil salinity in regions with high spatial variability using the machine learning model. A comprehensive assessment of the distributions of several factors causing soil salinity using machine learning remains insufficient. Therefore, the purpose of this paper is to soil salinity estimation using machine learning and remote sensing in the Mekong Delta, for example, in Tra Vinh province. Specifically, this study a) explores the appropriate conditioning factors to predict soil salinity in the Tra Vinh province; b) estimates soil salinity using machine learning (XGB, RF) and remote sensing (Sentinel 2A).

This study has the possibility of contributing to the literature by proposing a theoretical framework that fills the gaps in previous studies to develop a soil salinity map. This study aims to achieve precision and efficiency in the construction of soil salinity maps using machine learning. The results of this research play a significant role in improving the knowledge of soil salinity in the Mekong Delta in general and Tra Vinh Province in particular. It can support decision-makers or farmers in proposing practical measures to reduce the effects of soil salinity.

## 2. Description of the study area

Tra Vinh is located in the southeast of the Mekong Delta, with a natural area of 2,391 $km^2$, representing 5.77% of the Mekong Delta area. Most Tra Vinh province has a low relief, fluctuating from 0.6 to 1.0 m. The soil in the study area is fertile and composed of five main soil groups: sandy soil, saline soil, alum soil, alluvial soil, and ridge soil. Among them, the saline soil group has the largest

area, 47,362 hectares, accounting for 19.81%; the Lip land group has an area of 35,838 hectares; the Alluvial soil group has 34,180 hectares (14.30%); the Aluminous soil group has 32,910 hectares, accounting for 13.77% of the area; Sandy soil group with 8,250 hectares equivalent to 3.45% of the natural area. Tra Vinh province is located in the hot and humid tropical subequatorial monsoon climate zone. The climate is split into two distinct seasons: the rainy season starts from May to October, accounting for more than 80% of the total annual precipitation, and the dry season from November to April. The precipitation in Tra Vinh is at a low average level and tends to decrease (in 2020, only 68.1% compared to 2016), and the distribution is unstable. The upstream flows and tidal regimes of the East and West Seas strongly influence the hydrological regime of the Mekong Delta. The tides of the East Sea have a semi-diurnal regime. The high tide lasts about 6 hours, and the low tide lasts about 7 hours. The average tidal amplitude is about 3 to 4 m. At the same time, the tidal regime in the West Sea is very complex and generally belongs to diurnal tides. Although there are also 2 peaks and 2 troughs during the day, the tidal amplitude is much smaller than that of the East Sea, only about 0.8 to 1.2 m (Fig. 1).

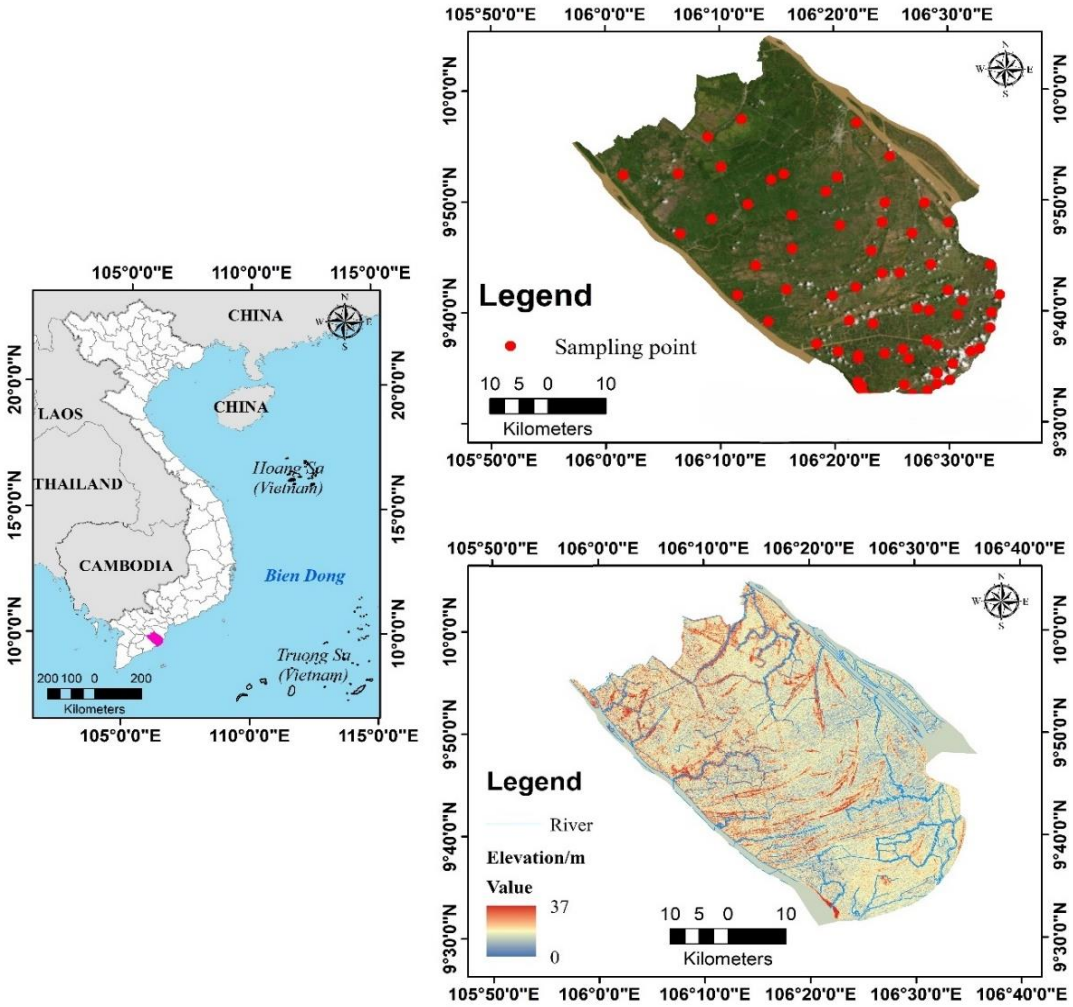


*Figure 1*. Location of Tra Vinh province in the Mekong Delta of Vietnam

Tra Vinh province's agricultural crop season is divided into three main crops: Winter spring, summer autumn, and autumn winter. The winter-spring crop usually lasts from November to March, the main crop with the highest yield of the year due to favorable weather conditions and abundant water resources. The summer-autumn crop lasts from April to August, which occurs during the dry season, so production in this crop often encounters difficulties, primarily due to saltwater intrusion in coastal areas. The Autumn-Winter crop lasts from September to December and is considered a secondary crop due to lower yields. Saltwater intrusion is a significant challenge for agricultural production in Tra Vinh, especially during the dry season, seriously affecting crop yields, especially in coastal areas.

## 3. Materials and Methodology

### *3.1. Material*

#### *3.3.1. Soil Salinity Sample Collecting and Laboratory Analysis*

In this study, 68 soil salinity samples with a 0 to 30 cm depth were collected from the field mission in August 2024 in the viticulture regions. August was chosen for the salinity sampling because August is the break time between two crops in Tra Vinh, helping minimize the impact of farming activities on the salinity of soil samples. Each sample was collected from four corners and one in the center of a 3×3 m sampling area. 500 g of soil was collected and sealed in a plastic bag at each location. Furthermore, the portable GPS on the field mission recorded the geography information for each point. After the soil salinity samples were transported to the laboratory for analysis. Before analysis, soil samples were air dried, and impurities such as gravel, tree branches, etc., were removed. The soil samples were mixed with pure water solution in the same ratio of 1:5, which was 15 g of soil mixed with 75 ml of distilled water. EC values were measured in soil and water solutions using a multiparameter measuring device WTW inoLab ® Multi 3420 Set B (Wissenschaftlich-Technische Werkstätten GmbH, Germany). At each point, the EC value is the mean of a sample.

In the end, each sample has been assigned the value of the conditioning variable to reach the entire data set. These data were divided into two parties, with a rate of 70/30. That is to say that this study uses 70% (48 samples) to train machine learning models and 30% (20 samples) to justify the effectiveness of the models.

#### *3.3.2. Conditioning variable*

Selection is considered one of the essential tasks when using machine learning to estimate environmental problems such as soil salinity because they are the factors that control salinity levels in a region (Erkin, Zhu, et al. 2019, Wang, Xue, et al. 2020). In this study, 25 conditioning factors were utilized to build the soil salinity model in Tra Vinh province, Mekong Delta, including 12 Sentinel 2A images (B1, B2, B3, B4, B5, B6, B7, B8, B8A, B9, B11 and B12) and four topographic factors (Elevation, Aspect, Curvature, and Slope), one hydrological factor (Distance to the river), one climatic factor (rainfall), one vegetation factor (NDVI) and six salinity index factors (SI, S1, S2, S3, S5, S6) (Fig. 2 and Table 1). All variables were transformed with a resolution of 10 m using the Resample tool in ArcGIS.

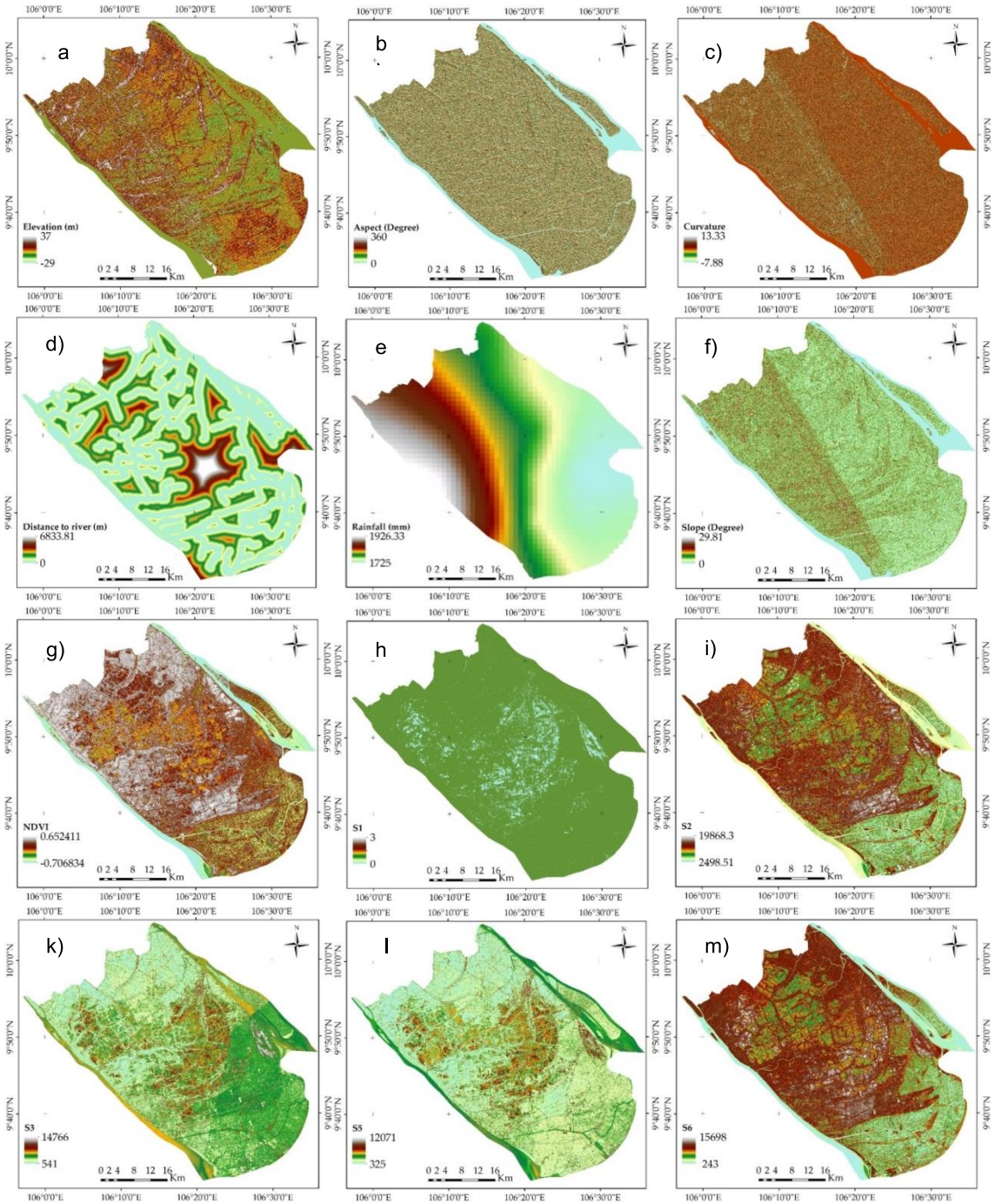


*Figure 2*. Example of conditioning variable for soil salinity model: a) Elevation, b) Aspect, c) Curvature, d) Distance to river, e) Rainfall, f) Slope, g) NDVI, h) S1 (Salinity Index), i) S2 (Salinity Index), k) S3 (Salinity Index), l) S5 (Salinity Index), m) S6 (Salinity Index)

*Table 1*. The calculation equations of the conditioning variable

| Abbreviation | Conditioning variable | Equation |
|---|---|---|
| B1 | Coastal earosol | Band 1 |
| B2 | Blue | Band 2 |
| B3 | Green | Band 3 |
| B4 | Red | Band 4 |
| B5 | Vegetation red edge1 | Band 5 |
| B6 | Vegetation red edge2 | Band 6 |
| B7 | Vegetation red edge3 | Band 7 |
| B8 | Nir | Band 8 |
| B8A | Vegetation Red Edge | Band 8A |
| B9 | Water vapour | Band 9 |
| B11 | SWIR1 | Band 11 |
| B12 | SWIR2 | Band 12 |
| NDVI | Normalised Difference Vegetation Index | (Band 8 – Band 4)/(Band 8 + Band 4) |
| SI | Salinity Index | (Band 2 + Band 4)0.5 |
| S1 | | Band 2/ Band 4 |
| S2 | | (Band 2 – Band 4)/( Band 2 + Band 4) |
| S3 | | (Band 3 × Band 4)/ Band 2 |
| S5 | | (Band 2 × Band 4)/ Band 3 |
| S6 | | (Band 4 × Band 8)/ Band 3 |
| Elevation<br>Aspect<br>Slope<br>Curvature<br>Rainfall<br>Distance to river | | |

In this study, three Sentinel-2 images were collected between July and September 2024 to ensure coverage of the entire study area. These images were stitched together using the Mosaicking tool in the ENVI 5.4 software. Before use, the Sentinel-2 images are cropped to the study area, and atmospheric correction is performed. It should be noted that the Sentinel image channels have different resolutions, so these image channels are adjusted to the exact spatial resolution of 10 m. Although the satellite images in the study were selected with minimal cloud cover, they still had to be filtered for clouds before use. Even though four topographic factors were extracted from DEM, which were built using a topographic map with a scale of 1:50,000 (available from the Ministry of Natural Resources and Environment). Rainfall was constructed from 10 climate stations in Tra Vinh province. NDVI was computed from Band 5, and Band 4 of the Sentinel 2A image, and salinity indices were calculated from Band 2, Band 3, Band 4, and band 8 of the Sentinel 2A image.

Band 1 and Band 9 were generally used to study atmospheric correction and water transparency. Although this band has a less direct influence on soil salinity, it can be used to eliminate atmospheric influences when analyzing other bands on soil salinity. Band 2 was used to analyze the reflectance of the bare soil. Soils influenced by salinity often have high reflectance in this region. So, it is considered one of the essential bands in soil salinity analysis. Band 3 was used to evaluate the reflectance of vegetation. It is imperative to distinguish bare soil from soil covered with vegetation. Because in salinity-influenced regions, vegetation has been influenced and is very difficult to develop. Band 4 is considered an essential factor to analyze the effects of

salinity on vegetation. Because salinity-influenced regions often have low reflectance due to vegetation degradation. Band 5 is a very sensitive factor to changes in vegetation stress. It is very effective in assessing the impacts of salinity on the photosynthesis process and vegetation health. So, it presents the relationships between salinity and vegetation development. Bands 6 and 7 are used to analyze vegetation structures. This is very important in distinguishing the influences of salinity on vegetation. Band 8 is an important factor in analyzing soil moisture and structure. Soil salinity often has different reflectance due to soil structure and moisture. Band 8A is essential to distinguish transition regions between salinity and non-salinity regions. Band 11 is used to analyze changes in soil moisture and mineral. It is very effective in distinguishing high-salinity regions. Band 12 is very effective in constructing a bare map due to the spectral signature of the soil. In general, in regions with high salinity, vegetation is tough to develop (Yahiaoui, Bradaï et al. 2021; Gerardo and de Lima 2022; Yimer, Sodango, et al. 2022; Kaplan, Gašparović, et al. 2023).

Topographic factors are vital in soil salinity monitoring, particularly in deltas such as the Mekong Delta and the Red River Delta. Elevation is considered an essential factor in analyzing the local distribution of salinity. In low-lying regions such as the Mekong Delta, saltwater intrusion originates from the sea or estuary and is stagnant. This leads to increased salt accumulation in the soil. Therefore, low-lying regions and proximity to the river are strongly affected by salinity (Cramer, Hobbs et al. 2004).

Furthermore, the tide has strongly affected these regions, which drives increased salinity (Nguyen, Liou, et al. 2020). This aspect is critical in evaporation and salt accumulation (Loc, Lixian, et al. 2021). Therefore, the southern and southwestern slopes are heavily influenced by solar radiation, which leads to high evaporation, which can focus more on soil salinity (Pessarakli and Szabolcs 2019). The curvature presents the convex shape of the surfaces, which affects the drainage capacity of the water and the distribution of salts. However, slope impacts the velocity and volume of flow in a region with a low slope that drives salt stagnation in soils (Triki Fourati, Bouaziz, et al. 2017). In Tra Vinh province, the coastal area has low elevation, and the difference between regions is insignificant, but topographic factors are essential in affecting soil salinity. The direction directly affects soil humidity, with areas facing south and southwest often having higher evaporation rates, leading to higher salt concentrations.

In contrast, areas facing north and northeast tend to maintain humidity better, diminishing soil salt concentrations. Although the slope in the area is small, it still affects the surface runoff and drainage rates. Low-slope areas often accumulate water, especially in coastal lowlands, increasing salinity levels. Furthermore, in Tra Vinh, the combination of topographic factors, tidal effects, and dense river systems has created favorable conditions for saltwater intrusion, significantly increasing soil salinity.

In the Mekong Delta, soil salinity is generally the consequence of the sea and the river. Therefore, regions near the river or sea tend to have more salinity (Yang, Huang et al. 2015). Although precipitation also plays an essential role in providing fresh water and increasing the river's flow, it reduces the soil's salinity level. When rainfall is abundant, especially in the rainy season, it increases the level of the river, which carries salt water away from the shore, thus reducing the level of saltwater intrusion. In contrast, during the dry season, when the river level reduces sharply, this facilitates the penetration of salt water deep into the land (Isidoro and Grattan

2011, Eswar, Karuppusamy et al. 2021). In this study, the average annual rainfall is used to calculate soil salinity using a machine learning model.

NDVI is considered one of the key factors in soil salinity evaluation because soil salinity directly influences the development of agriculture. Previous studies have widely used this index (Mehla, Kumar et al. 2024). The salinity frequently negatively affects plant growth, particularly reducing leaf area, as reflected in the NDVI index. Furthermore, NDVI has been demonstrated to be an effective tool for assessing plant health on a large scale and over a long period (Nguyen, Tran, et al. 2021). This study collected samples in August 2024, when rice plants were flowering. This is the rice growing season. At the same time, Sentinel-2 images were also collected during this period, ensuring that the NDVI values accurately reflect the growth situation of rice plants. Therefore, NDVI can be used as a reliable indicator to monitor and evaluate the impact of salinity on crops in the study area.

Six salinity indices (SI, S1, S2, S3, S5, S6) play an essential role in evaluating or predicting the salinity of any region globally. It provides the process of salt accumulation on surfaces. Combining these indicators in the analysis, accurately assessing salinity intrusion and understanding the overall salinity intrusion process to develop agriculture sustainably (Wang, Peng et al. 2021).

### *3.2. Methodology*

This study applied two machine learning models, namely XGB and RF, to monitor soil salinity in the Tra Vinh province. The soil salinity monitoring was divided into four main steps: (i) collect soil salinity samples and conditioning factors; (ii) establish machine learning models; (iii) evaluate the effectiveness of the proposed models; (iv) analyze soil salinity map. Figure 3 shows the soil salinity prediction methodology in the province of Tra Vinh.

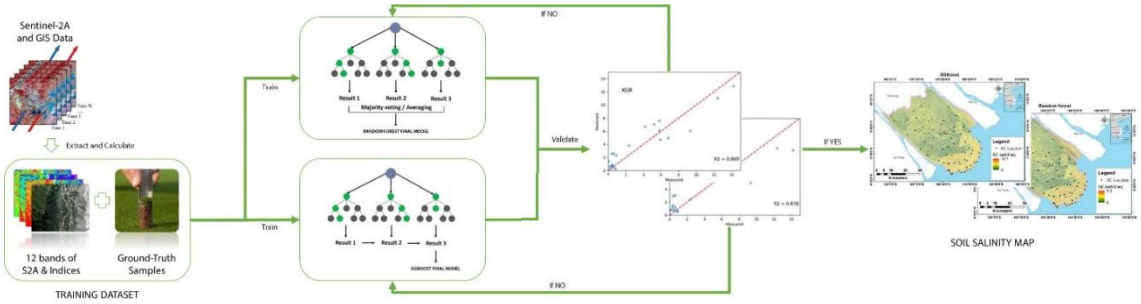


*Figure 3*. Methods used for soil salinity prediction in Tra Vinh province using XGB and RF

#### *3.2.1. Xgboost (XGB)*

XGB is the gradient-boosting family algorithm proposed by (Friedman 2001). This model is based on sequential ensemble learning and weak learners. This boosting algorithm converts several weak learners into a single strong learning model (Zarei, Hasanlou et al. 2021). It starts by building a first model on the data and then builds a second one, focusing on accurately predicting the observations that the first model predicted poorly. Combining these two models is supposed to be better than the ones taken individually. This boosting process is repeated several times, each successive model trying to correct the flaws of the previous models. The XGB algorithm is capable of processing a large volume of data sets. This makes it particularly useful for Big Data applications such as soil salinity prediction (Nguyen, Tran

et al. 2021, Aksoy, Sertel et al. 2024). The performance of XGB depends on adjusting parameters like learning rate, n_estimators, max depth, and gamma. In this study, the parameters of the XGB model were adjusted using the trial-and-error method. Finally, the parameter values were learning rate = 0.05, n_estimators = 500, max_depth = 10, and gamma = 1.

*3.2.2. Random forest (RF)*

RF is a machine-learning algorithm designed and proposed by (Breiman 2001). This algorithm is mainly based on the assembly of decision trees. That is to say, it combines the results to obtain the best results. The RF can consist of some trees, and the number of trees is a parameter that the crossover has validated. Each tree is trained on a subset of the data set and gives a result. The final results are the average value of all trees (Lee, Kim, et al. 2017). The RF function is based on three main steps: (i) randomly selecting a sample from the entire data set. (ii) Generate a tree in the forest for each sample. (iii) calculating the average value from the value of each tree (Habibi, Delavar, et al. 2023). In the training process, RF can reduce bias and increase variance to avoid the overfitting problem. Additionally, the RF model can solve missing data using the voice value. The accuracy of RF depends on three primary parameters: the number of nodes, the number of trees, and the number of sampled features (Islam, Talukdar et al. 2021). In this study, we used the trial-and-errors method to adjust the RF. In the end parameters, the RF model was the number of nodes = 10, the number of trees = 500, and the number of features sampled features = none.

*3.2.3. The assessment model proposed*

In this study, various statistical indices, namely RMSE, MAE, and R2, were used to evaluate the performance of the proposed model. Previous studies have widely used these indices (Ge, Ding et al. 2022, Kaplan, Gašparović et al. 2023).

RMSE and MAE present the differences between the observation and prediction values. While $R^2$ is a crucial statistical measure used to evaluate the effectiveness of a linear regression model in describing the relationship between variables, $R^2$ is a crucial statistical measure. Quantifies the proportion of the variance of the dependent variable that is predictable from the independent variables.

## 4. Results

### *4.1. Selection of independent variables*

The selection of appropriate conditioning factors is crucial for the machine learning model to learn the relationships between the location of saline areas and their causes, enabling better future estimation of soil salinity. The importance of conditioning factors was assessed using the RF model, which assigns a value to each factor based on the relationships between sample locations and conditioning factors. The higher the value of a factor, the more important that factor is. Among the 25 independent factors in this study, B8A, B7, and S6 were the most pertinent (Fig. 4). B8A is essential for analyzing soil moisture. Although the Mekong Delta, in general, and Tra Vinh province, in particular, are at a low altitude, salt water penetrates the soil. B8A can evaluate this process. B7 is essential in measuring the health of vegetation.

In Tra Vinh, soil salinity directly influences rice cultivation, so B7 was ranked second most important of all the factors. S6 was third, thanks to its ability to build a soil salinity model in a water-saturated environment. This is necessary in the province, where soil salinity is affected by saltwater originating from rivers and the sea, often mixed with stagnant water. B11, B9, B6, B5, elevation, S2, B8, and distance to the river took positions 4 to 11 respectively. B11 is

very sensitive to changes in humidity and mineral salts; in Tra Vinh, soil salinity is becoming more and more severe due to high evaporation and humidity. This causes salts to accumulate on land surfaces, especially in areas where there are stagnant waters. B9 improves the accuracy of soil salinity estimation by eliminating errors related to atmospheric humidity, which is particularly important in humid regions. B6 and B5 analyze the effects of the environment on vegetation, such as soil salinity. Vegetation in Tra Vinh province, such as rice farms, is frequently impacted by soil salinity. B8 measures soil moisture. The soil salinity in the area affected regions with less humidity and proximity to the sea. The study area is in a region with low altitude and proximity to the sea and rivers and is, therefore, more affected by soil salinity. This is why elevation and distance to the river were selected as essential factors.

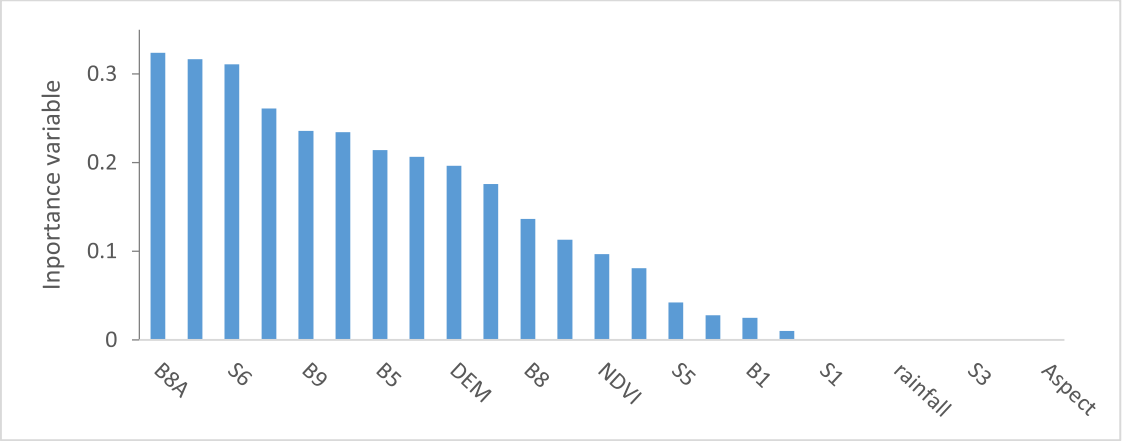


*Figure 4*. Variable Important Using a Random Forest

The factors S1, B3, rainfall, slope, S3, B2, and aspect did not influence soil salinity in Tra Vinh province. Therefore, they were eliminated from the machine learning model. Soil salinity was most influenced by B8A, B7, S6, elevation, and distance to the river; it was not influenced by aspect, slope, or rainfall (Fig. 4).

In the context of climate change and sea level rise, soil salinity is affected by many different factors (or variables). These factors are bound together by a linear relationship and linked through a nonlinear relationship. Each factor has a different level of interaction with soil salinity. However, how do we determine these values? XGB has solved this problem by quantifying and evaluating the level of impact of each factor in the model by determining the total number of decision tree splits to reduce the noise of the forecast results of each factor. From there, XGB helps build the model by focusing only on the factors with the most significant impact, eliminating factors that have little or no effect on the research subject, such as soil salinity. This is very important to ensure that the model operates effectively, meeting the requirements of analysis and forecasting when the impact factors have continuous and complex changes. At the same time, this process also helps to enhance the model's performance. In addition, during the model setup process, XGB allows us to improve the model by correcting and calibrating the model parameters after each training loop. Thanks to that, the built model has higher accuracy and reliability.

### *4.3. Model Comparison and Performance Evaluation*

This study uses $R^2$ to evaluate the machine learning model's performance in monitoring soil salinity. Among the two proposed models, with an $R^2$ value of 0.86, the XGB model has higher predicted precision than the RF model, with an R2 value of 0.67. Additionally, to ensure the reliability of the two proposed models, this study utilized RMSE and MAE to evaluate the model performance. For the training data set process, with RMSE and MAE values of 0.25 and 0.22, respectively, the XGB model performed more than the RF model with RMSE and MAE values of 0.38 and 0.35. For the validation of the data set process, with the RMSE and MAE values of 0.32 and 0.3, respectively, the XGB model continues to have more precision than the RF model with the RMSE and MAE value of 0.42 and 0.41, respectively (Fig. 5 and Table 2).

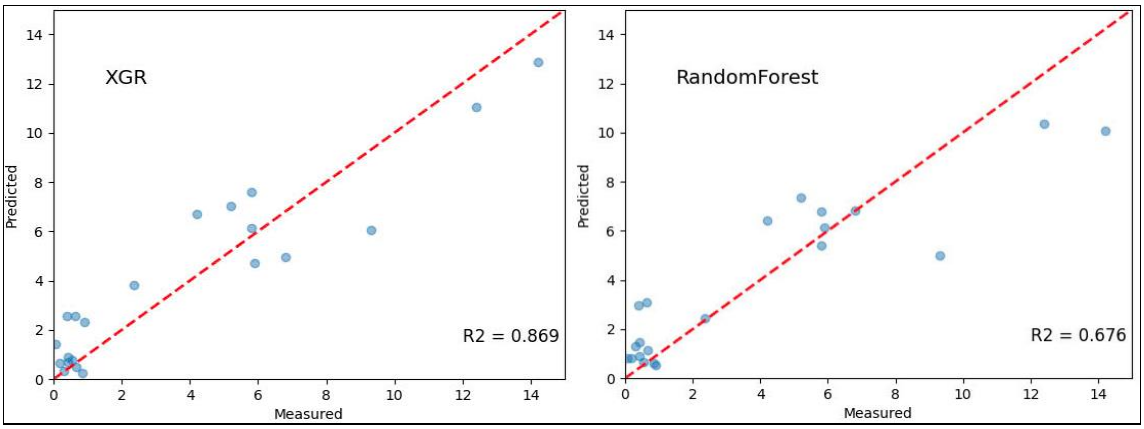


*Figure 5*. $R^2$ value for the XGB and RF models

*Table 2*. Model Performance for Soil Salinity Mapping

| | Training dataset | | | Validating dataset | | |
|---|---|---|---|---|---|---|
| | RMSE | MAE | $R^2$ | RMSE | MAE | $R^2$ |
| XGB | 0.25 | 0.22 | 0.92 | 0.32 | 0.3 | 0.86 |
| RF | 0.38 | 0.35 | 0.72 | 0.42 | 0.41 | 0.67 |

### *4.4. Soil Salinity Map*

Figure 6 shows the soil salinity map in Tra Vinh province using two proposed models. The results showed that the areas with the most considerable saline intrusion have the highest EC values of about 9.5 mS/cm, concentrated in the southeast region and along the Tien and Hau rivers. Meanwhile, the areas with less salinity intrusion are concentrated further inland. The spatial distribution of regions affected by salinity intrusion reflects the role of topography, distance, and river density.

According to the Department of Agriculture and Rural Development of Tra Vinh province, during the dry seasons of 2019 and 2020, drought and soil salinity caused more than VND 1 trillion, of which rice suffered the most damage at VND 919 billion. In addition, more than 271 hectares of fruit trees were also affected by salinity.

In the book "*Diagnosis and Improvement of Saline and Alkali Soils*" by Regional Salinity Laboratory (US) of J. K. Brown and his colleagues in 1954, the level of soil salinity intrusion based on soil electrical conductivity values was divided into five different levels including nonsaline (EC ~ 0-2 mS/m), slightly saline (EC ~ 2-4 mS/m), saline (EC ~ 4-8 mS/m), strongly saline (EC ~ 8-16 mS/m) and extremely saline (EC > 16 mS/m) (Richards

1954). Based on the soil salinity maps, we could see that the coastal areas in the southeast region of Tra Vinh province are experiencing severe salinity intrusion and tend to cause increased soil salinity in the central mainland area from nonsaline or slightly saline to saline level through the river and canal systems spread throughout the province, even if there are minor differences between the two models.

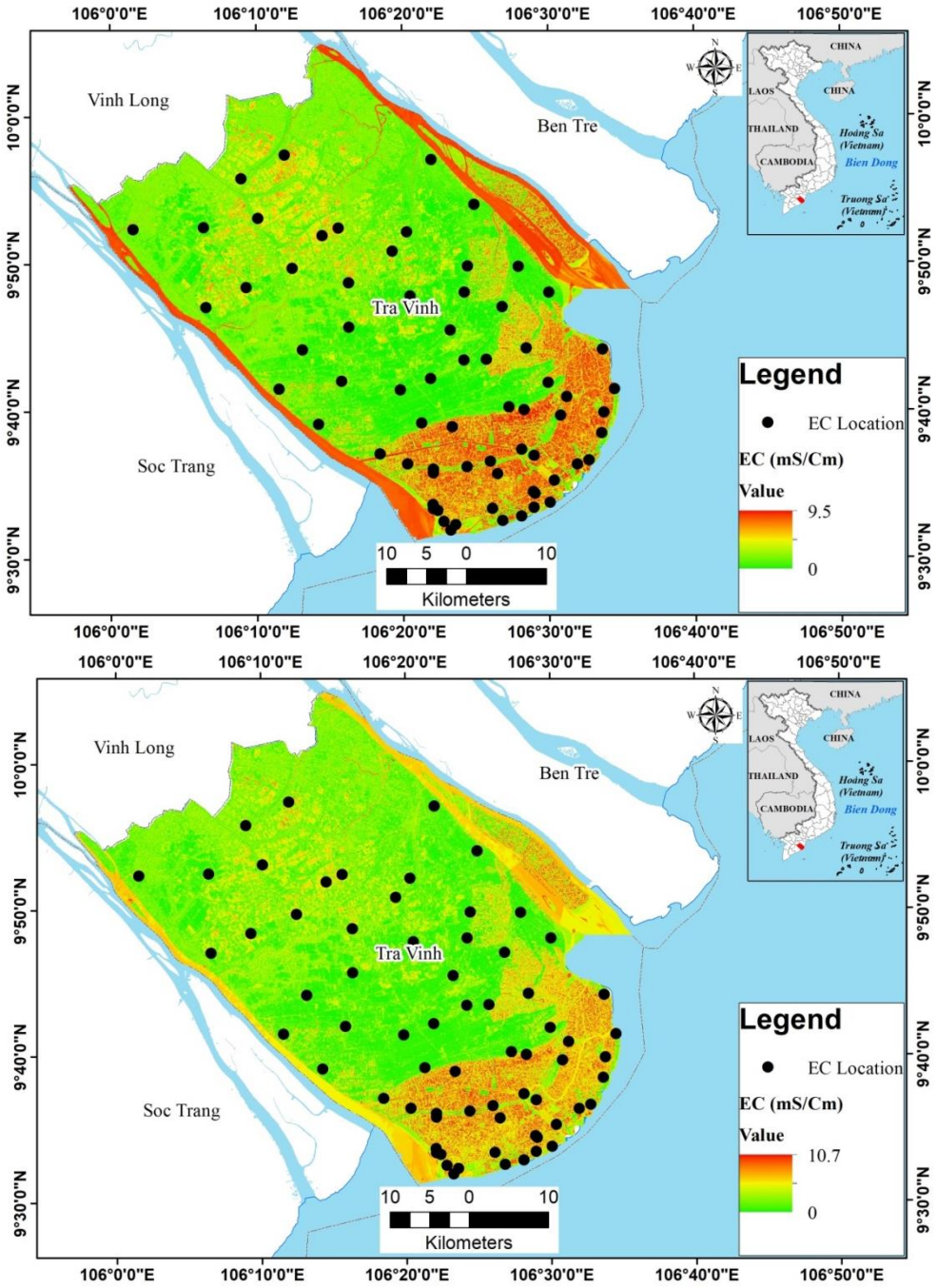


*Figure 6*. Soil salinity map in the Tra Vinh province using XGB (above) and RF (below)

The soil salinity map is considered an important tool to support decision makers or farmers in understanding the distribution of soil salinity and propose effective intervention 2measures to reduce the effects of soil salinity on agricultural development.

## 5. Discussions

Soil salinity is considered one of the serious environmental problems, causing significant damage to the development of agriculture in the country. The soil salinity phenomenon is increasingly serious in deltas where there is low altitude and is influenced frequently by rising sea levels (Vermeulen and Van Niekerk 2017, Xiao, Ji et al. 2023). The altitude of the Mekong Delta varies from 0.5 to 2 m above sea level, making this region more affected by tides and sea level rise (Wassmann, Hien, et al. 2004, Kim-Anh, Liou et al. 2020). The flat terrain creates favorable conditions for deep penetration of salt water into the continent through the river canal, especially in the dry season when the river level is very low (Ngoc 2017, Van Binh, Kantoush, et al. 2020). In addition, although the dense river density is favorable for agricultural development, this is also the main route for seawater to penetrate the mainland, making saline intrusion increasingly severe, especially with climate change.

Meanwhile, the Mekong Delta is also home to nearly 20 million people, located along the Tien and Hau rivers, of which 80% of the population lives in rural areas and works in the agricultural sector (Nguyen, Van et al. 2023). Therefore, soil salinity significantly affects this region's food security and livelihoods. The Tra Vinh is one of the provinces most seriously affected by soil salinity due to low altitude, tide, and sea level rise. So, seawater enters the mainland through the river system, as the river level is increasingly lowered due to the construction of dams and reservoirs upstream, combined with sea level rise in the context of climate change (Nguyen, Liou et al. 2020). The aim of this study is to construct the spatial distribution map of soil salinity, which can support farmers in proposing investment measures to reduce the effects of soil salinity on agriculture.

This study evaluates the feasibility of using Sentinel 2A optical images to assess soil salinity. The use of optical imagery to build soil salinity maps in this study indicates that optical spectral reflectance quickly affects salinity and surface characteristics, such as plant health and soil moisture. Additionally, Sentinel 2A images with 12 spectral bands, including visible, near-infrared, and short-wave infrared regions, enable them to monitor changes in the reflectance characteristics of soils and vegetation affected by salinity, particularly for band 5, band 6m, band 7, and band 11, band 12. Furthermore, Sentinel 2A images have different resolutions from 10m to 60m, allowing soil salinity at small and medium scales. The 5-day temporal resolution allows Sentinel 2A images to provide continuous salinity data over time (Taghadosi, Hasanlou, et al. 2019; Gerardo and de Lima 2022).

In this study, machine learning combines with remote sensing to monitor soil salinity. Machine learning can solve massive volumes of data from different sources, such as satellite images and ground sensors. This allows for comprehensive and detailed integration and analysis of the causes of salinity. Furthermore, soil salinity is influenced by multiple factors and is often nonlinear. However, machine learning is considered an effective tool to present complex non-linear relationships, which is very difficult to achieve with traditional models. Finally, machine learning can be applied in similar regions, which allows them to be used in different environmental contexts (Wang, Shi, et al. 2020, Wang, Peng et al. 2021). However, the

accuracy of machine learning models strongly depends on the quality and representativeness of the training data.

Furthermore, selecting appropriate algorithms in some algorithms is challenging because there are no universal guides to choosing them (Nguyen, Tran et al. 2021). This study uses two popular algorithms, namely XGB and RF, to monitor soil salinity in Tra Vinh province. The XGB model was more efficient among the two proposed models, with an $R^2$ value of 0.86. Because XGB uses numerous weak learners in a sequential method that repeatedly enhances observations, this method reduces the high biases that can sometimes be recurrent in machine learning models. Indeed, XGB offers a considerable number of hyperparameters. Thanks to this diversity of parameters, it is possible to have total control over the implementation of gradient boosting. It is also possible to add different regularizations in the loss function, limiting an Overfitting phenomenon often occurring when using Gradient Boosting algorithms (Qiu and Zhou 2023, Kiriakidou, Livieris et al. 2024). With an $R^2$ value of 0.67, the RF model was less precise than the XGB. RF uses many trees in the model or trains the model on large data sets, which can require high computational costs. Although RF models are often more efficient during the training phase, creating predictions can be more time-consuming than other algorithms, especially when dealing with large datasets or models with a large number of trees.

Additionally, because RF incorporates many decision trees, it becomes difficult to clearly understand the logic behind each prediction, which sometimes leads to the model being viewed as a "black box" that is difficult to interpret (Langsetmo, Schousboe et al. 2023). Finally, although this study only uses 68 soil salinity samples as input data for the XGB and RF model. However, the precision of these models was more than 0.7 for training data and 0.67 for validation data. These results are acceptable because the XGB and RF models can handle small data sets through parameter and ensemble optimization. In addition, both models have the advantage of avoiding the overfitting problem by using parameters such as max_depth or min_samples. Although the soil salinity samples used in this study were small, this study uses the RF model to evaluate the importance of conditioning factors; therefore, the XGB and RF model can achieve high precision by evaluating the complex relationships between input variables.

Although this study successfully built the soil salinity map using machine learning with high accuracy. However, this study also has limitations related to data use. This study collected 68 soil salinity samples evenly distributed throughout the Tra Vinh province to use as input data for the machine learning model. The number of samples was similar to previous studies that used machine learning to estimate the soil salinity of different regions worldwide. However, 68 samples can be considered insufficient to have high precision. However, collecting soil salinity samples is very difficult, especially in developing countries, due to the limited funding and time. In future research, we will try to add samples to improve the model performance. The second is related to the machine learning model, and the third is related to the nonlinear relationship between soil salinity locations in the past and the conditioning factors. This study collected soil samples at depths of 0 to 30 cm. However, the big question is whether the salinity data collected from Sentinel 2A satellite images can reflect soil characteristics at 0 to 30 m depths. Sentinel 2A data can be compared to soil samples collected at 0 to 5 m or 0 to 10 m depths. In addition, many environmental factors, such as soil quality, humidity, or human activity, can affect soil

salinity, such as vegetation and the salinity index.

Furthermore, the machine learning models in this study also have uncertainties related to optimizing model parameters. This study uses a trial-and-error method to select the optimal parameters. However, this method stops only when a seemingly suitable solution is obtained but does not guarantee that the solution is optimal. In many cases, better solutions may be missed due to a lack of information or because not all possibilities have been examined. In the future, our goal is to integrate machine learning algorithms with optimization algorithms to provide more accurate models in soil salinity monitoring. Finally, this study uses Sentinel-2 images to monitor and analyze soil salinity. Although Sentinel-2 images have significant advantages in high spatial (10-20 m) and temporal (revisit every 5 days) resolution, they are suitable for small-scale monitoring with high accuracy.

Furthermore, multispectral bands can effectively capture salinity-related features. However, this study's difference between soil sample collection and satellite image acquisition could influence field and remote sensing data consistency. However, several studies point out that soil salinity varies mainly with seasons, especially when the river and sea levels are low (Nguyen, Tran, et al. 2021). Therefore, this temporal difference did not influence the results of this study.

Finally, one of the limitations of this study is related to the use and selection of conditioning factors. Although the study used 25 conditioning factors to monitor and control soil salinity, some critical factors were not collected due to finance and time constraints. In the future, we will try to add more factors directly related to soil salinity, such as soil moisture, river density, and high tide level. These factors can provide more detailed and accurate information, helping to clarify the causes and mechanisms that lead to soil salinity, thus improving the effectiveness of analysis and forecasting models.

The results of this study can be used to predict or estimate soil salinity in the coastal region, even for prediction with high accuracy. These results are undoubtedly helpful in supporting decision-makers or farmers in managing land resources to develop agriculture to ensure security in the region or country.

## 6. Conclusions

Soil salinity is considered one of the most dangerous environmental problems, and it has significantly affected the development of agriculture and food security in the region, particularly the Mekong Delta. Therefore, soil salinity monitoring is essential in intervening ineffective measures. This study aims to construct the soil salinity map with high precision using machine learning and Sentinel 2A, namely XGB and RF, in the Tra Vinh province of the Mekong Delta. The conclusion of this study is as follows.

(i) This study justifies the ability of the XGB and RF models to monitor soil salinity. This result can support decision-makers or farmers in constructing intervention measures to reduce the effects related to soil salinity.

(ii) Of the two proposed models, the XGB model had a higher performance than the RF model, with an $R^2$ value of 0.86. Therefore, this study recommends using the XGB model to monitor soil salinity in the province of Tra Vinh.

(iii) The soil salinity map highlights that areas near the sea and along the Tien and Hau rivers are more affected by soil salinity due to the low altitude and density of the river. These maps can support decision-makers or farmers in managing land resources to develop agriculture.

This study successfully constructed a high-accuracy soil salinity map to support sustainable land resource management.

Furthermore, in policy-making, governments can use soil resources to implement remote sensing technologies and machine learning models to predict soil salinity, providing practical strategies and policies for agricultural development. By applying this method, decision-makers or planners can significantly improve food security in the region or at the national level.

## Reference


Aksoy S., et al., 2024. Assessment of soil salinity using explainable machine learning methods and Landsat 8 images. International Journal of Applied Earth Observation and Geoinformation, 130, 103879.

Breiman L., 2001. Random forests. Machine learning, 45, 5–32.

Chaaou A., et al., 2024. Potential of land degradation index for soil salinity mapping in irrigated agricultural land in a semi-arid region using Landsat-OLI and Sentinel-MSI data. Environmental Monitoring and Assessment, 196(9), 843.

Cramer V.A., et al., 2004. The influence of local elevation on soil properties and tree health in remnant eucalypt woodlands affected by secondary salinity. Plant and Soil, 265(1), 175–188.

Dunlop G., et al., 2019. Simulation of saltwater intrusion into coastal aquifer of Nagapattinam in the lower cauvery basin using SEAWAT. Groundwater for Sustainable Development, 8, 294–301.

Erkin N., et al., 2019. Method for predicting soil salinity concentrations in croplands based on machine learning and remote sensing techniques. Journal of Applied Remote Sensing, 13(3), 034520–034520.

Eswar D., et al., 2021. Drivers of soil salinity and their correlation with climate change. Current Opinion in Environmental Sustainability, 50, 310–318.

Fathizad H., et al., 2020. Investigation of the spatial and temporal variation of soil salinity using random forests in the central desert of Iran. Geoderma, 365, 114233.

Friedman J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics, 1189–1232.

Garajeh M.K., et al., 2021. An automated deep learning convolutional neural network algorithm applied for soil salinity distribution mapping in Lake Urmia, Iran. Science of the Total Environment, 778, 146253.

Ge X., et al., 2022. Updated soil salinity with fine spatial resolution and high accuracy: The synergy of Sentinel-2 MSI, environmental covariates and hybrid machine learning approaches. Catena, 212, 106054.

Gerardo R., I.P. de Lima, 2022. Sentinel-2 Satellite Imagery-Based Assessment of Soil Salinity in Irrigated Rice Fields in Portugal. Agriculture, 12(9), 1490.

Gorji T., et al., 2017. Monitoring soil salinity via remote sensing technology under data scarce conditions: A case study from Turkey. Ecological Indicators, 74, 384–391.

Habibi A., et al., 2023. Flood susceptibility mapping and assessment using regularized random forest and naïve bayes algorithms. ISPRS annals of the photogrammetry. Remote Sensing and Spatial Information Sciences, 10, 241–248.

Isidoro D., S. Grattan, 2011. Predicting soil salinity in response to different irrigation practices, soil types and rainfall scenarios. Irrigation Science, 29, 197–211.

Islam A.R.M.T., et al., 2021. Flood susceptibility modelling using advanced ensemble machine learning models. Geoscience Frontiers, 12(3), 101075.

Jiang H., et al., 2019. Quantitative assessment of soil salinity using multi-source remote sensing data based on the support vector machine and artificial neural network. International Journal of Remote Sensing, 40(1), 284–306.

Kaplan G., et al., 2023. Soil salinity prediction using Machine Learning and Sentinel-2 Remote Sensing Data in Hyper-Arid areas. Physics and Chemistry of the Earth, Parts a/b/c, 130, 103400.

Keilholz P., et al., 2015. Effects of land use and climate change on groundwater and ecosystems at the middle reaches of the Tarim River using the MIKE SHE integrated hydrological model. Water, 7(6), 3040–3056.

Khanh P.T., et al., 2024. Evaluation of machine learning models for mapping soil salinity in Ben Tre

province, Vietnam. Multimedia Tools and Applications, 1–20.

Kim-Anh N., et al., 2020. Soil salinity assessment by using near-infrared channel and Vegetation Soil Salinity Index derived from Landsat 8 OLI data: a case study in the Tra Vinh Province, Mekong Delta, Vietnam. Progress in Earth and Planetary Science, 7(1), 1–16.

Kiriakidou N., et al., 2024. C-XGBoost: A tree boosting model for causal effect estimation. IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, 58–70. Doi://doi.org/10.1007/978-3-031-63219-85.

Langsetmo L., et al., 2023. Advantages and disadvantages of random forest models for prediction of hip fracture risk versus mortality risk in the oldest old. Journal of Bone and Mineral Research Plus, 7(8), e10757.

Lee S., et al., 2017. Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. Geomatics, Natural Hazards and Risk, 8(2), 1185–1203.

Loc H.H., et al., 2021. How the saline water intrusion has reshaped the agricultural landscape of the Vietnamese Mekong Delta, a review. Science of the Total Environment, 794, 148651.

Mantena S., et al., 2023. Prediction of soil salinity in the Upputeru river estuary catchment, India, using machine learning techniques. Environmental Monitoring and Asessment, 195(8), 1006.

Mehla M.K., et al., 2024. Soil salinity assessment and mapping using spectral indices and geostatistical techniques-concepts and reviews. Remote Sensing of Soils, Elsevier, 429–437.

Metternicht G.I., J. Zinck, 2003. Remote sensing of soil salinity: potentials and constraints. Remote Sensing of Environment, 85(1), 1–20.

Mirlas V., 2012. Assessing soil salinity hazard in cultivated areas using MODFLOW model and GIS tools: a case study from the Jezre'el Valley, Israel. Agricultural Water Management, 109, 144–154.

Ngoc T.A., 2017. Assessing the effects of upstream dam developments on sediment distribution in the Lower Mekong Delta, Vietnam. Journal of Water Resource and Protection, 9(7), 822–840.

Nguyen H.D., et al., 2023. Soil salinity prediction using hybrid machine learning and remote sensing in Ben Tre province on Vietnam's Mekong River Delta. Environmental Science and Pollution Research, 30(29), 74340–74357.

Nguyen K.-A., et al., 2020. Soil salinity assessment by using near-infrared channel and Vegetation Soil Salinity Index derived from Landsat 8 OLI data: a case study in the Tra Vinh Province, Mekong Delta, Vietnam. Progress in Earth and Planetary Science, 7(1), 1–16.

Nguyen T.G., et al., 2021. Salinity intrusion prediction using remote sensing and machine learning in data-limited regions: A case study in Vietnam's Mekong Delta. Geoderma Regional, 27, e00424.

Nguyen V.H., et al., 2024. Evaluating topsoil salinity via geophysical methods in rice production systems in the Vietnam Mekong Delta. Journal of Agronomy and Crop Science, 210(1), e12676.

Pessarakli M., I. Szabolcs, 2019. Soil salinity and sodicity as particular plant/crop stress factors. Handbook of Plant and Crop Stress. Fourth Edition, CRC Press, 3–21.

Qiu Y., J. Zhou, 2023. Short-term rockburst damage assessment in burst-prone mines: an explainable XGBOOST hybrid model with SCSO algorithm. Rock Mechanics and Rock Engineering, 56(12), 8745–8770.

Ren D., et al., 2019. Analyzing spatiotemporal characteristics of soil salinity in arid irrigated agro-ecosystems using integrated approaches. Geoderma, 356, 113935.

Richards L.A., 1954. Diagnosis and improvement of saline and alkali soils, US Government Printing Office.

Sarkar S.K., et al., 2023. Coupling of machine learning and remote sensing for soil salinity mapping in coastal area of Bangladesh. Scientific Reports, 13(1), 17056.

Sirpa-Poma J., et al., 2024. Complementarity of Sentinel-1 and Sentinel-2 data for soil salinity monitoring to support sustainable agriculture practices in the Central Bolivian Altiplano. Sustainability, 16(14), 6200.

Taghadosi M.M., et al., 2019. Retrieval of soil salinity from Sentinel-2 multispectral imagery. European Journal of Remote Sensing, 52(1), 138–154.

Tran D.A., et al., 2021. Evaluating the predictive power of different machine learning algorithms for groundwater salinity prediction of multi-layer coastal aquifers in the Mekong Delta, Vietnam. Ecological Indicators, 127, 107790.

Triki Fourati H., et al., 2017. Detection of terrain indices related to soil salinity and mapping salt-affected soils using remote sensing and geostatistical techniques. Environmental Monitoring and Assessment, 189, 1–11.

Van Binh D., et al., 2020. Long-term alterations of flow regimes of the Mekong River and adaptation strategies for the Vietnamese Mekong Delta. Journal of Hydrology: Regional Studies, 32, 100742.

Vermeulen D., A. Van Niekerk, 2017. Machine learning performance for predicting soil salinity using different combinations of geomorphometric covariates. Geoderma, 299, 1–12.

Wang F., et al., 2020. Multi-algorithm comparison for predicting soil salinity. Geoderma, 365, 114211.

Wang J., et al., 2021. Soil salinity mapping using machine learning algorithms with the Sentinel-2 MSI in arid areas, China. Remote Sensing, 13(2), 305.

Wang N., et al., 2020. Integrating remote sensing and landscape characteristics to estimate soil salinity using machine learning methods: A case study from Southern Xinjiang, China. Remote Sensing, 12(24), 4118.

Wang W., J. Sun, 2024. Estimation of soil salinity using satellite-based variables and machine learning methods. Earth Science Informatics, 1–13.

Wassmann R., et al., 2004. Sea level rise affecting the Vietnamese Mekong Delta: water elevation in the flood season and implications for rice production. Climatic Change, 66, 89–107.

Wu W., et al., 2014. Mapping soil salinity changes using remote sensing in Central Iraq. Geoderma Regional, 2, 21–31.

Xiao C., et al., 2023. Prediction of soil salinity parameters using machine learning models in an arid region of northwest China. Computers and Electronics in Agriculture, 204, 107512.

Yahiaoui I., et al., 2021. Performance of random forest and buffer analysis of Sentinel-2 data for modelling soil salinity in the Lower-Cheliff plain (Algeria). International Journal of Remote Sensing, 42(1), 148–171.

Yang L., et al., 2015. Mapping soil salinity using a similarity-based prediction approach: a case study in Huanghe River Delta, China. Chinese Geographical Science, 25, 283–294.

Yimer A.M., et al., 2022. Analysis and Modeling of Soil Salinity Using Sentinel-2A and LANDSAT-8 images in the Afambo Irrigated Area, Afar Region, Ethiopia. Doi: 10.20944/preprints202204.0250.v1.

Zarei A., et al., 2021. A comparison of machine learning models for soil salinity estimation using multispectral earth observation data. ISPRS annals of the photogrammetry. Remote Sensing and Spatial Information Sciences, 3, 257–263.

Zeng W., et al., 2018. Comparison of partial least square regression, support vector machine, and deep-learning techniques for estimating soil salinity from hyperspectral data. Journal of Applied Remote Sensing, 12(2), 022204–022204.

Zhao W., et al., 2019. Comparison of IDW, cokriging and ARMA for predicting spatiotemporal variability of soil salinity in a gravel-sand mulched jujube orchard. Environmental Monitoring and Assessment, 191, 1–15.