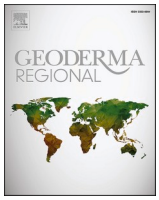
Geoderma Regional 27 (2021) e00424

Contents lists available at ScienceDirect

Geoderma Regional

journal homepage: www.elsevier.com/locate/geodrs



Salinity intrusion prediction using remote sensing and machine learning in data-limited regions: A case study in Vietnam's Mekong Delta

Tien Giang Nguyen [a], Ngoc Anh Tran [a], Phuong Lan Vu [b], Quoc-Huy Nguyen [b,c], Huu Duy Nguyen [b,*], Quang-Thanh Bui [b,c]

[a] Departement of Hydrology and Water Resources, VNU University of Science, Vietnam National University, 334 Nguyen Trai, Thanh Xuan district, Hanoi, Viet Nam
[b] Faculty of Geography, VNU University of Science, Vietnam National University, 334 Nguyen Trai, Thanh Xuan district, Hanoi, Viet Nam
[c] Centre for Applied Research in Remote Sensing and GIS, Faculty of Geography, VNU University of Science, Vietnam National University, 334 Nguyen Trai, Thanh Xuan district, Hanoi, Viet Nam

ARTICLE INFO

Keywords:
Soil salinity
Fluvisols
Mekong
Machine learning
Remote sensing

ABSTRACT

With population growth, the demand for land resources is expected to increase significantly in the coming decades. Maintaining the integrity of soil distribution requires a remarkable amount of work to deal with agricultural extension. Salinity intrusion monitoring is a crucial process, which directly affects sustainable development, especially in areas affected by global warming and in coastal zones. In recent years, various studies have used the soil-water salinity data to evaluate the spatiotemporal increase in salinity intrusion. This study aims to establish a novel framework for monitoring salinity intrusion using remote sensing and machine learning. It focuses on the salinity intrusion in soil, which affects water availability, food security, human health, etc. Numerous algorithms have been implemented to find the best solution for this issue, including Xgboost (XGR), Gaussian processes, support vector regression, deep neural networks, and the grasshopper optimization algorithm (GOA). A total of 143 samples collected from 2016 to 2020 at 39 measurement stations were divided into two sets: 70% training and 30% testing. Thirty-one independent variables were used to develop the model. Vietnam's Mekong Delta, where the salinity intrusion problem is becoming increasingly serious due to global warming and demographics, was selected as the study area. Each of the proposed models was compared and evaluated by applying various statistical indices such as the root mean square error, coefficient of determination (R^2), and mean absolute error. The results show that the prediction model was built successfully by wielding data from the implemented salinity measurement stations, and the XGR-GOA model was better than the other models (R^2 = 0.86, RMSE = 0.076, and MAE = 0.065). This finding demonstrates the feasibility of estimating and monitoring salinity intrusion in data-limited regions by integrating optical satellite images and machine learning, which are easily and cost-effectively obtainable. The proposed conceptual methodology in our study is novel and provides additional useful information for the monitoring and management of salinity intrusion not only in Vietnam's Mekong Delta, but also in other sites that have similar natural and anthropological conditions.

1. Introduction

Soil management is essential to ensure sustainable agricultural development and food production, with high-quality and environmental protection (Manasa et al., 2020). According to the most recent estimates of the UN Food and Agriculture Organization, more than 20% of the cultivated land on Earth is variably degraded by soil salinization (Wu et al., 2018), and this is projected to reach ~50% by 2050 (Bartels and Sunkar, 2005; Wang et al., 2019). Soil salinization affects approximately 10% of the world's food production (Machado and Serralheiro, 2017). It is particularly high in many coastal countries and is said to increase in the future due to climate change (Das et al., 2020). The major part of affected land is in Asia (65%), followed by Africa (19%), and Europe (5%) (Siebert et al., 2013). The situation is becoming increasingly serious in Vietnam, one of the five countries most affected by global warming, with rise in saltwater being the primary concern (Arndt et al., 2015). Although the Mekong Delta in Vietnam contains a large amount of Fluvisols, according to the Ministry of Natural Resources and the

* Corresponding author.
E-mail address: nguyenhuuduy@hus.edu.vn (H.D. Nguyen).

https://doi.org/10.1016/j.geodrs.2021.e00424
Received 29 May 2021; Received in revised form 22 July 2021; Accepted 24 July 2021
Available online 30 July 2021

Environment, 38% of its land could be submerged by 2100, affecting 55% of the population in the area, which would threaten the national food security if effective preventive measures are not taken. Soil salinization leads to serious consequences, especially in the context of population growth, which requires natural resources, especially food, which is directly related to the requirement of more cultivable land (Wang et al., 2020e). Therefore, it is necessary to develop a global strategy to reduce the negative effects of soil salinity.

Soil salinity is the result of a range of phenomena, including irregular rainfall, evaporation, salinity of groundwater, flooding from storm surges, flooding of saline rivers, and the presence of a soluble salt source (Clarke et al., 2015; Salehin et al., 2018). Salinization can occur during soil formation by release of soluble salts either during weathering or by external natural inputs (Ivushkin et al., 2019; Wang et al., 2020e). Inappropriate agricultural practices, such as salt water irrigation in the absence of proper drainage, cause soil salinization (Vermeulen and Van Niekerk, 2017). Therefore, soil salinity should be monitored to better understand its processes and avoid potentially dangerous consequences. According to the literature, soil salinity can be monitored directly or indirectly. Salinity of a system evolves spatiotemporally; therefore, traditional (direct) methods, such as soil sampling followed by laboratory analysis (Allbed et al., 2014; Mulder et al., 2011) prove insufficient and unsuited to meet the evolution speed of this phenomenon, particularly because these methods are very expensive, time-consuming, and difficult to update. A large number of samples are needed to fully assess soil salinity characteristics in large areas, due to their strong spatial change over short distances. Although the number of proximal sensors has increased considerably in recent years, they can only monitor the soil characteristics over a relatively small area (<2 m). Therefore, numerous sensors are required to effectively monitor a large area at the desired spatial resolution (Vermeulen and Van Niekerk, 2017). Thus, faster, cheaper, and more reliable methods need to be explored to monitor soil salinity.

Recently, significant developments have been made in the field of soil salinity assessment. The traditional method has gradually been replaced by remote sensing (indirect methods), which is a more efficient way of monitoring soil salinity in large, data-scarce regions (Davis et al., 2019a; Delavar et al., 2020; Hoa et al., 2019b; Nicolas and Walter, 2006; Wang et al., 2018). This approach uses the interaction between the soil reflectance and salinity indicators (Erkin et al., 2019). With global coverage and a short revisit period (few days), these data are more homogeneous and regular than the field data. In addition, remote sensing provides valuable information from unreachable areas (Wang et al., 2020b; Wang et al., 2020e). Radar and optical technologies supply universal and long-lasting information around the globe, which has been proved from the detection and visualization of the Earth's surface (Farahmand and Sadeghi, 2020; Wu et al., 2020). In the recent years, several studies have successfully monitored soil salinity using the available optical remote sensing data, such as Landsat 8 OLI and Sentinel 2 (Moussa et al., 2020; Nguyen et al., 2018). These approaches focus on a combination of bands. However, optical remote sensing products are difficult to use in the presence of cloud cover and depend on solar radiation. In this context, radar sensors are reliable since they do not depend on weather conditions. Previous studies have utilized radar sensors to assess soil salinity using the microwave P, C, and L bands, which can penetrate to a depth of 150 cm or more from the surface (Yang and Guo, 2019; Zhang et al., 2021). However, it is difficult to distinguish between soil salinity and moisture from the radar backscatter coefficients. In addition, because active systems are conventionally limited to a single polarization and frequency, the amount of information that can be extracted from the radar signal is limited.

To address these limitations, a remote sensing-based machine learning approach, in the form of neural networks, random forests (RF), and decision trees (DT), has been successfully used to study soil salinity (Fathizad et al., 2020; Garajeh et al., 2021; Jiang et al., 2019; Qi et al., 2018). Machine learning analyzes large automatic and semi-automatic datasets to present linear and nonlinear correlations between salinities of the samples and input variables (DEM, normalized difference salinity index, intensity index 1, intensity index 2, enhanced vegetation index…) and uses these relationships to predict soil salinity for regions without any available data (Melesse et al., 2020; Sahour et al., 2020). (Wu et al., 2018) predicted salinity intrusion in the Mussaib area in Central Mesopotamia by integrating optical and radar remote sensing, and machine learning, including support vector regression (SVR) and random forest regression (RFR). (Hoa et al., 2019b) estimated salinity intrusion in the Ben Tre province in southern Vietnam using five machine learning models based on radar data remote sensing, namely multilayer perceptron neural networks (MLPNN), gaussian processes (GPR), RF, radial basis function neural networks (RBFNN), and SVR. (Habibi et al., 2020) integrated machine learning, namely artificial neural networks (ANN), genetic algorithms (GA), DT, partial least square regression (PLSR), and optical remote sensing as a means of predicting soil salinity in the Saveh Plain, Iran. Thus, it is established from previous studies that machine learning can be used successfully to monitor soil salinity. However, these models are data-dependent. As mentioned above, soil sampling in the field is very expensive and time-consuming, especially in large regions. In addition, one of the major issues in the development of realistic models is the problem of overfitting and underfitting when selecting appropriate input variables in the event of lack of in situ data. Therefore, these methods still need to be developed or replaced by robust and automated ones to overcome these limitations.

As observed from the previous studies, water salinity and soil salinity are strongly related; therefore, for the first time, we attempt to develop a comprehensive understanding of the soil salinity in data-scarce regions using a state-of-the-art remote sensing and artificial intelligence technique based on water salinity. Information collected from this method will be used to support decision-makers in land management, especially in sustainable rural development.

2. Material and methods

2.1. Study area

The Vietnamese Mekong Delta is situated downstream of the Mekong River and covers an area of 39,734 km^2. The average altitude in the region is 1–2 m above sea level. The region experiences a tropical climate, with rainy (from May to October) and dry seasons (from November to April). The average rainfall is 1400–2200 mm, of which 90–95% falls during the rainy season. The rivers, streams, and canals are dense, leading to abundant surface water in the Mekong Delta. The average annual flow volume is approximately 500 km^3. The tides in the delta are mixed, diurnal and semi-diurnal, whose magnitude can reach up to 3 m. Usually, there are two troughs and two peaks during the day; however, the relative tide heights change every two weeks (Duy et al., 2021), i.e., when the first troughs decrease day by day, the other troughs increase, and vice versa. Apart from the tidal influence, this region is also influenced by river flooding. The flood season occurs from July to November, which inundates approximately 35–50% (up to 4 m) of the delta's surface (Wassmann et al., 2019). The delta is home to nearly 20 million people (with a density of 500 people/km^2). Four out of five people live in rural areas, and the workforce is mainly involved in agricultural practices. However, this region has been identified as a global "hot spot" that is susceptible to the effects of climate change, particularly rising sea levels, leading to increased salinity in the river systems. Studies have predicted that by 2050, the sea level will increase by 33 cm, and by 2100, it will increase by 1 m, which will lead to the submergence of at least 25% of agricultural land in the Mekong Delta, and ~ 75% of the current cultivated area will be affected by salinity in the dry season. This will cause ~40–50% of the agricultural area to be affected by salt water even during the rainy season, seriously affecting rice crops. Therefore, soil salinity needs to be monitored to form a sustainable agricultural development strategy (Fig. 1).

2.2. Data used

2.2.1. Salinity sample observation collection

A total of 39 automatic measurement stations were set up to measure the salinity in the study site. All of these stations were placed in the transitional area between water and land along the estuaries, rivers, and canals of the Mekong Delta to determine the electrical conductivity (EC), which is directly proportional to salinity. Twelve of the salinity measurement stations were established along estuaries, including Co Chien, Cua Dai, Cua Tieu, Cua Soai Rap, and Ham Luong, while 25 stations were placed along rivers, including Cai Be, Cai Lon, Dong Dien, Dong Nai, Ganh Hao, Hau, Maspero, My Tho, Nhu Gia, Doc, Kien, Tien, Vam Co, Vam Co Dong, and Vam Co Tay; two stations were constructed along canals, including Rach Gia and Phung Hiep. At the location of automatic measurements, the soil had an average pH of 4.9–6.5, average Cl^- index of 0.05–0.25, and average P_2O_5 values of 0.04–0.05. The data were collected daily during 2016–2020, and the noisy and missing data were filtered in the gathering process to avoid bias of machine learning models. All samples were collected in dry weather conditions, without cloud cover at the location of samples, so that satellite imagery data could be incorporated to estimate the operation of the models. These meteorological conditions were chosen for sampling to gather more acute data when freshwater flows from rivers would be narrowed (Habiba et al., 2015). In addition, the Mekong Delta is a large region and all areas cannot be imaged in one day; therefore, the measured salinity values were averaged over four months per year to synchronize the sample dataset and ensure that the entire study area was covered by the satellite orbit. Finally, 70% of the data were used for training, while 30% were applied to validate the models (Fig. 2).

2.2.2. Satellite imagery and preprocessing

Satellite imagery is one of the most common data sources used in many salinity intrusion studies (Tran et al., 2019), where the EC is used to identify potential salinity. Various studies have shown how the relationship between spectral bands and EC value can aid in the prediction of the level of salinity intrusion (Tran et al., 2019). Many studies have also used salinity indices, calculated from spectral band values, to predict the EC (Seifi et al., 2020). In this study, Landsat 8 OLI/TIRS, provided by USGS, was collected and atmospherically corrected by LaSRC using the CFMask algorithm (Vermote et al., 2018). Landsat 8 orbits the Earth in a sun-synchronous manner, close to a polar orbit, with a 16-day repeat cycle for temporal resolution. The Landsat 8 OLI/TIRS image has 11 spectral bands, including eight bands at 30 m spatial resolution, one panchromatic band at 15 m spatial resolution, and two TIRS bands at 100 m spatial resolution, which is sufficient for regional studies (Markham et al., 2018). More than 50 scenes were acquired with a cloud cover of less than 30%, during the period when the water samples were collected, to ensure consistency of the dataset. All sample data without the corresponding satellite imagery were removed.

Fig. 2. The photos of field sampling points.

2.2.3. Salinity intrusion geodatabase

A salinity intrusion database was established to improve the training dataset using indices that have been previously verified in numerous

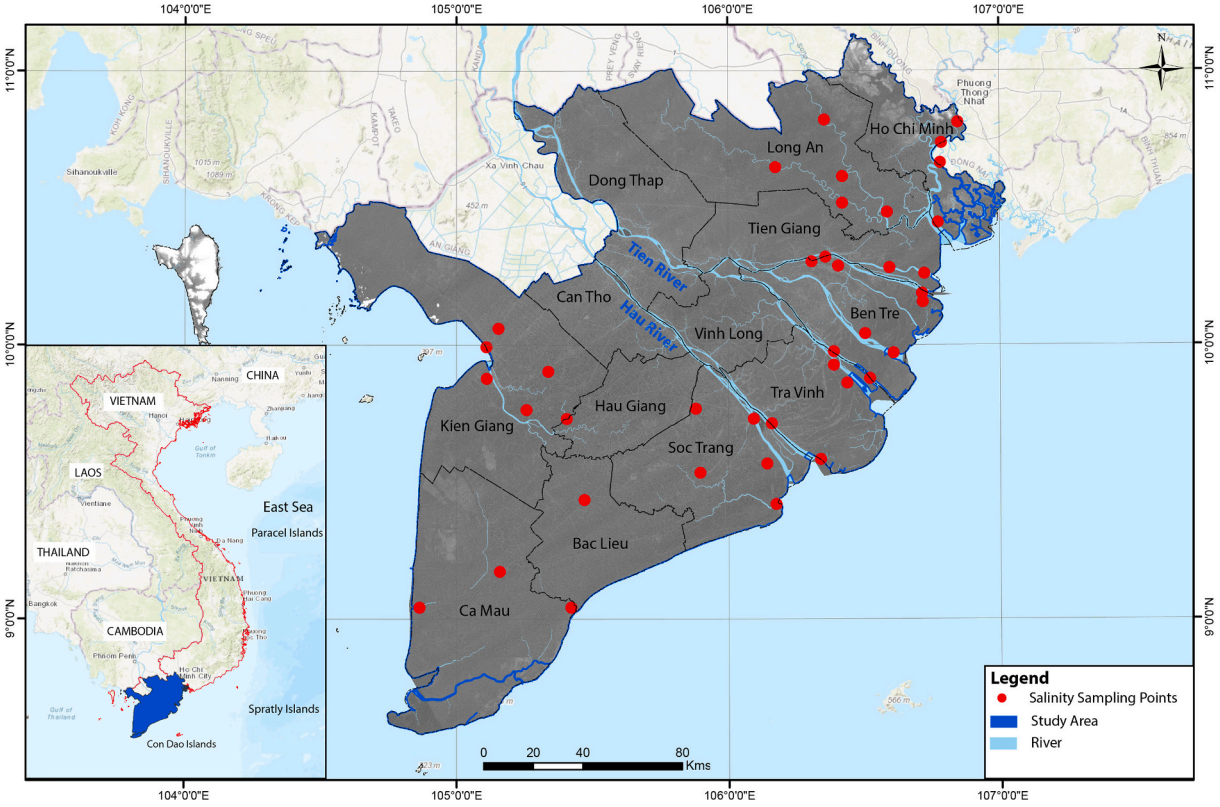


Fig. 1. Location of the Mekong Delta of Vietnam.

studies. Each index demonstrated a constraint on the EC. For example, (Nguyen et al., 2020b) built a regression model from 11 bands of Landsat 8 (B1 to B11) to estimate the EC. (Matinfar et al., 2020) used salinity indices (SI1 to SI11), normalized difference vegetation index, normalized difference salinity index, soil adjusted vegetation index, and vegetation soil salinity index to predict the EC. (Bouaziz et al., 2011) applied intensity indices and enhanced vegetation index to obtain the EC by multiple linear regression and other algorithms. Several specific indices have also been implemented to estimate the EC, such as ND23 and ND47 (Wu, 2019). DEM data were also added to this database. Finally, 31 indices were included in the analysis.

2.3. Machine learning algorithms

2.3.1. XGboost (XGR)

XGR is a collection of powerful gradient boosting algorithms that are both flexible and effective. It is based on the boosting framework proposed by Friedman et al. (Friedman, 2001)(Malik et al., 2020). With a linear model solver and a tree-based learning algorithm, XGR can support a wide range of object functions, including classification, ranking, and regression. Various studies have shown that XGR can overcome the limitation of analyzing power, including the model performance and computational speed for boosted tree algorithms (Torlay et al., 2017). It can work with multi-dimensional data, which is very important when working with large data problems. Even with missing data, XGBoost has normalizing methods to deal with the problem in an effective way, avoiding overfitting or underfitting issues (Jia et al., 2019). Moreover, it includes cross-validation procedures that are not implemented in other algorithms. XGR was built by taking the attributes of different models. In this study, a new model is created by combining different models to reduce errors in the existing models and increase accuracy. This new model can be implemented until no further improvements are necessary, and they return the best prediction model. XGR allows the performance of several major gradient boosting techniques, including gradient, regularized, and stochastic. The use of these techniques depends on the distribution and relationships between the data variables.

2.3.2. Support vector machine (SVM)

Since its introduction (Cortes and Vapnik, 1995), SVM has become one of the most widely used machine learning models. This algorithm has been applied in various fields such as energy (Jindal et al., 2020), medical (Jaramillo et al., 2017), and finance (Jaramillo et al., 2017). It includes two major problem-solving processes: classification (SVC) and regression (SVR). SVR establishes a hyperplane surface to simulate the nonlinear relationship of different variables. This minimizes the observed training error and enhances the generalized regression efficiency. SVR is also considered as a transformation process of inputting variables into a high dimension of vector environment (Guo et al., 2018). The predicted output values can be validated by several cross-validation methods. SVR theory can be represented as follows:

In a dataset P with n training samples (x_i, y_i), and i from 1 to n, $x_i = \{x_{i1}, \ldots, x_{in}\} c \in R^n$ is the input data, and $y_i \in R$ is the corresponding output data. The goal is to build a linear function (g), which is described as a hyperplane closest to the point of the samples, by a non-linear function Δ (Eq. 1).

$$g(x) = w^T{}_* \Delta(x) + b \tag{1}$$

where w^T is the weight and b is the bias coefficient. A conditional function (d) to identify the minimum distance between training samples and the hyperplane is written as follows (Eq. 2):

$$d = |y_i - (w_* x_i + b)| = \begin{cases} 0 \ if \ d < \varepsilon \\ d - \varepsilon \ otherwise \end{cases} \tag{2}$$

where ε is an insensitive loss function. The value of ε will be defined by the user. Finally, the optimization procedure was performed using the trade-off control function with the regularization parameter (C) and two slack parameters (δ and δ^*) (Eq. 3):

$$Min \ \frac{1}{2} w^{T*} w + C \sum_{i=1}^{n} \left(\delta_i + \delta_i^*\right) \tag{3}$$

With $w_* \Delta(x) + b - y_i \leq \varepsilon - \delta_i$, $y_i - w_* \Delta(x) - b \leq \varepsilon + \delta_i^*$, and $\delta_i, \delta_i^* \geq 0$

2.3.3. Deep neural network (DNN)

A DNN is an exalted version of an ANN. The distinction is in the training process, where an initialization phase is performed using the contrastive divergence algorithm (CDA) integrated with a back-propagation algorithm (Guo et al., 2018). CDA was implemented on a restricted Boltzmann machine (RBM) (Wang et al., 2017). Notably, the directions of the arrows were in both ways, except that in the ANN. This means that by using the CDA, the weights are modified by reinforcing the error in each direction.

A neural network is a technology that mimics the behavior of the human mind, especially object detection and flow of data across various layers of artificial connections in the brain (He et al., 2020). Since a neural network is far more comprehensive than logistic regression (LR), it can be used to simulate LR as a specific case of a neural network. In the input layer, a neuron acts as a variable. In this study, the X matrix has 31 variables, i.e., 31 neurons. A neural network model requires numerical values to establish the operating process. If there are l number of numerical variables and m number of categorical variables, each with l category, the total neurons of the input layer will be estimated as $l + m * n$. Each neuron is a linear amalgamation of the values in the previous layer, and receives its inputs from the neurons of the previous layer. In the neural network model, the weights used for training are w_i, with i from 1 to the number of neurons.

The fully connected (FC) layer was implemented as a structure in the regression model. It uses two functions: batch normalization (w/B) and a leaky rectified linear unit (LReLU), which is an activation function. The LReLU has many advantages, such as faster convergence speed, low-cost calculation, and avoidance of the "Dying ReLU" problem. During training, LReLU was used to switch the feature spaces. The number underneath each FC layer box indicates the number of neurons used in that layer (Bui et al., 2020; Costache et al., 2020). Dropout is often used to prevent fitting issues, which occur when neurons are switched off according to their allocated possibilities during the forward step. To resolve fitting problems, a dropout layer was used with a parameter p (default value of $p = 0.2$). Two parameters, α and γ, which are defined as the learning rate and momentum coefficient of the learning gradient incorporation, respectively, are set to tune the neural network model. The batch normalization method has not been not used in the last two layers. The output of the network has only one neuron, which is also an FC layer; however, it does not use the activation function. Finally, a value is returned within the range of the predicted boundary. In the experimental implementation, 0.01 and 0.9 are used as default value of α and γ, respectively (Mehrer et al., 2020; Panda, 2020; Singhania et al., 2017).

2.3.4. Gaussian processes (GPR)

GPR is considered to be one of the most sophisticated machine learning algorithms, and is widely utilized for natural resource and environment management tasks such as biomass estimation, soil moisture monitoring, and landslides (Rohmer and Foerster, 2011; Smith et al., 2014; Stamenkovic et al., 2017). GPR uses a non-parametric Bayes approach to modulate the data (Hoa et al., 2019a). The major advantage of GPR is the ability to adjust parameters to enhance performance models at the highest level.

2.3.5. Grasshopper optimization algorithm (GOA)

GOA was first suggested by (Saremi et al., 2020). It solves

optimization issues by mimicking the action of grasshopper swarms. GOA generates a collection of search agents to construct the predominant artificial swarm. The position of each agent is calculated and compared to select the best candidate as the target. The target starts to attract others in the area, and grasshoppers continue to flock towards the target grasshopper. The grasshopper colony has a peculiar rampaging pattern that can be seen in both the adolescent and tadpole stages (Dwivedi et al., 2020). The grasshoppers run in quick steps at a slow pace during the larval stage of the colony. In adolescence, the colony's primary trait is a sudden movement over a long distance (Dinh, 2021). Another essential characteristic of grasshopper colonies is their quest for food. The search strategy is divided into two patterns by nature-inspired metaheuristics: exploitation and exploration (Gampa et al., 2020). Exploration agents are urged to move quickly, while exploitation agents are more willing to spread locally. The naturalness of grasshoppers, simulates goal finding along with exploration and exploitation.

The movement of grasshoppers is primarily driven by three factors, the social interaction factor is represented by the interplay between agents, the gravitational force factor is represented by the attraction of an agent, and the wind advection factor is represented by the effect of wind on an agent. The Z_i variable is defined as the position of the i^{th} grasshopper. The mathematical formula for Z_i can be represented as follows (Eq. 4):

$$Z_i = I_i + F_i + W_i \tag{4}$$

where I_i denotes social interactions, F_i is the gravitational force, and W_i is the wind advection. The most crucial factor in the movement of grasshoppers is social interaction, which can be determined using the equation below (Eq. 5):

$$I_i = \sum_{j=1;j\neq i}^{n} s*dij*\widehat{dij} \tag{5}$$

$$d_{ij} = |z_j - z_i| \tag{6}$$

$$dij = (z_j - z_i)/d_{ij} \tag{7}$$

$$s(r) = f*e^{-r/l} - e^{-r} \tag{8}$$

where d_{ij} is a unit vector that expresses the interspace in the middle of the i^{th} and j^{th} positions of grasshoppers and can be calculated using Eqs. 6 and 7. The s function describes the social powers and can be calculated using Eq. 8. This function can be altered using f and l parameters. However, the s function cannot assign the level of influence between grasshoppers that are separated by wide distances. The following formula describes the gravity force of the grasshopper (Eq. 9):

$$F_i = -g*\widehat{e_g} \tag{9}$$

where g is defined as a unity vector. This parameter is a gravitational constant and is directed towards the center of the earth's surface. The wind advection of the grasshopper can be simulated as follows (Eq. 10):

$$W_i = u*\widehat{e_w} \tag{10}$$

where u is defined as a unity vector that illustrates the direction of wind, and is in a constant drift. After updating the formulas, the movement function can be described as follows (Eq. 11):

$$Z_i = \sum_{j=1;j\neq i}^{n} s*(|z_j - z_i|)*\frac{z_j - z_i}{d_{ij}} - g*\widehat{e_g} + u*\widehat{e_w} \tag{11}$$

where n is the total number of grasshoppers. In this model, several parameters were added to the calculation to improve the optimization process by altering the exploration and exploitation aptitudes. The effect of gravity can be ignored if it is minimal, and the wind direction is assumed to always converge towards the optimal solution. We have the following equation (Eq. 12):

$$X_i^d = b\left(\sum_{j=1;j\neq i}^{n} b*\frac{ub_d - lb_d}{2}*s*\left(\left|z_j^d - z_i^d\right|\right)*\frac{z_j - z_i}{d_{ij}}\right) + \widehat{T_d} \tag{12}$$

where ub_d is the top confines of the d^{th} dimension, and lb_d is the bottom confines of the d^{th} dimension. Parameter b is updated in each iteration to decrease exploration and increase exploitation. The formula for b is written as follows (Eq. 13):

$$b = bmax - l*\frac{bmax - bmin}{L} \tag{13}$$

where bmax is the maximum value and bmin is the minimum value. Variable l indicates the current iteration. The maximal iterations are defined as L.

2.4. Performance assessment

2.4.1. The coefficient of determination (R^2)

R^2 (or R-squared) is a quantitative tool that helps in identifying the relationship between independent variables and the dependent variable in a regression model (Di Bucchianico, 2008). It indicates the goodness of fit of the data and identifies what percentage of the results are fitted to a regression model. For example, a value of $R^2 = 50\%$ represents a 50% fit of the total data in the model. A high R^2 indicates that the model is highly relevant. However, this does not mean that the model is correct. In reality, the consistency of statistical methodologies could be influenced by the characteristics and units of the variables. The data transformation methodology can also be an impacting factor. R^2 cannot identify whether a regression model is good or bad. In several cases, a high R^2 could be the result of a flawed regression model. We cannot conclude whether this is a bad indication for predicting analytics with a low R^2 value since some good models have low R^2 values. There is no general law for incorporating statistical measurements into model evaluation. The background of the hypothesis or prediction is critical, and the observations of the metric will evolve depending on the situation. R^2 estimates the relationship between movements of the different variables. The calculation of R^2 is as follows:

$$R^2 = 1 - \frac{\sum (y_i - \widehat{y})^2}{\sum (y_i - \bar{y})^2}$$

2.4.2. Root mean squared error (RMSE)

The RMSE is commonly used to calculate the standard deviation of the residuals between the observed and predicted features. RMSE is very useful in cases where large errors are particularly undesirable because of the relatively high weight. Thus, RMSE could not only explain the average error, but also describe other implications, which are more difficult to extract and understand. However, the use of RMSE has significant consequences (Bui et al., 2020). RMSE is useful in cases where large errors are particularly undesirable because of the relatively high weight. Thus, RMSE could not only explain the average error, but also describe other implications, which are more difficult to tease out and understand. The formula for RMSE can be expressed as follows:

$$RMSE = \sqrt{\frac{1}{n}*\sum_{i=1}^{n} (y_i - \widehat{y})^2}$$

2.4.3. The mean absolute error (MAE)

MAE shows the average of the residuals for the observed and predicted features. This coefficient uses the same scale as the input that is being calculated (Willmott and Matsuura, 2005). MAE is commonly used to measure the prediction error in time-series analysis. If the absolute value could not be calculated, the average error will be transformed into the mean bias error (MBE) to determine the average bias of

the model. The formula for MAE is as follows:

$$MAE = \frac{1}{n}^{*} \sum_{i=1}^{n} |y_i - \hat{y}|$$

2.5. Modeling methodology

The methodology used in this study to monitor soil salinity includes four main steps: data collection and preparation, model building, model validation, and construction of the soil salinity map (Fig. 3).

Step 1: Data preparation process plays a very important role in the use of machine learning to monitor soil salinity. In this study, water salinity samples from 2016, 2019, and 2020 were collected under the Vietnamese national ĐTĐL.CN-50/18 project. A total of 31 independent variables were extracted from Landsat 8 OLI images from 2016, 2019, and 2020. These data were used to construct the soil salinity models.

Step 2: Variable selection process plays a very important role in the construction of the model, especially in the context of small data. In the beginning, the 31 independent variables were selected from four groups: terrain attributes, salinity indices, vegetation indices, and remote sensing data. Preliminary analysis process removed the unhelpful variables.

This project used Landsat 8 OLI data with a resolution of 30 m, with many mixed pixels that contain multiple end elements. We have attempted to minimize the independent variables to decrease the complexity of the model and selected 13 independent variables. In this research, eight proposed models (DNN, GPR, SVR, XGR, DNN-GOA, GPR-GOA, SVR-GOA, and XGR-GOA) were trained using 70% of the data, while 30% of the data were used to validate the models.

Step 3: In this research, several statistical indices were used to validate the models, namely RMSE, MAE, and R^2. RMSE was used to measure the difference between the predicted values; this value was smaller, and the model performed better. MAE and R^2 have been used in various studies to validate the salinity intrusion model (Abedi et al., 2021a; Vermeulen and Van Niekerk, 2017).

Step 4: After validating the proposed models, they were utilized to construct a salinity intrusion map of the entire Vietnamese Mekong Delta, which was generated by calculating the salinity of each pixel.

3. Results

3.1. Exploration of input datasets

The importance of the 31 independent variables selected for this study was assessed using the RF algorithm (Table 1). B11 (Thermal Infrared Sensor 2) is the most important variable with FR = 0.3, which is followed by DEM (0.083), B6 (0.077), CRSI (0.066), B3 (0.065), B5 (0.057), B8 (0.045), B7 (0.045), MSI (0.031), Nd47 (0.024), Nd23 (0.024), B4 (0.023) and EVI (0.022). The four factors with FR = 0 (B2, B9, Int 1, NDSI, NDVI) do not influence the models (Fig. 4). Based on the importance diagram of the variables, the first 13 factors, which are the most important, will be used as a new input for the models. The remaining 18 factors were removed because their impact levels were very low and may not improve the performance of the model (Behnamian et al., 2017).

After examining the proposed structures of each model and the characteristics of the input variables, the parameters of each model were identified using the trial-and-error method (Table 2).

3.2. Validation and comparison of the models

After selecting the conditioning factors associated with the EC values of the observation, the soil salinity prediction models were developed using XGR, DNN, SVM, XGR-GOA, DNN-GOA, and SVR-GOA. Five-fold cross-validation was used to obtain the most optimized parameters with the best R^2 and minimum RMSE values.

Table 1
The salinity indices and calculation formulas.

#	Abbreviation	Name	Formula	Source
1	B1	Coastal Aerosol (CA)	B1	(Forkuor et al., 2018)
2	B2	Blue (B)	B2	
3	B3	Green (G)	B3	
4	B4	Red (R)	B4	
5	B5	NIR	B5	
6	B6	SWIR1	B6	
7	B7	SWIR2	B7	
8	B8	Panchromatic (P)	B8	
9	B9	Cirrus (C)	B9	
10	B10	TIRS1	B10	
11	B11	TIRS2	B11	
12	SI1	Salinity Index 1	$SI1 = (B * R)^{1/2}$	(Habibi et al., 2020)
13	SI2	Salinity Index 2	$SI2 = (G * R)^{1/2}$ (Whitney et al., 2018)	
14	SI3	Salinity Index 3	$SI3 = (R * NIR)^{1/2}$ (Wang et al., 2020c)	
15	SI4	Salinity Index 4	$SI4 = (G^2 + R^2 + NIR^2)^{1/2}$	
16	SI5	Salinity Index 5	$SI5 = (G^2 + R^2)^{1/2}$	
17	NDVI	Normalized Difference Vegetation Index	NDVI = (NIR - R)/(NIR + R)	(Whitney et al., 2018)
18	NDSI	Normalized Difference Salinity Index	NDSI = (TIR1 – NIR)/(TIR2 + NIR)	(Solangi et al., 2019)
19	NDII	Normalized Difference Infrared Index	NDII = (NIR - SWIR1)/(NIR – SWIR1)	(Elhag and Bahrawi, 2017)
20	ND23	Normalized Difference Between TM2 and TM3	ND23 = (G - R)/(G + R)	(Wu, 2019)
21	ND47	Normalized Difference Between TM4 and TM7	ND47 = (NIR - SWIR2)/(NIR + SWIR2)	(Wu, 2019)
22	II1	Intensity Index 1	II1 = (G + R) / 2	(Bouaziz et al., 2011)
23	II2	Intensity Index 2	II2 = (G + R + NIR)/2	(Bouaziz et al., 2011)
24	VSSI	Vegetation Soil Salinity Index	VSSI = 2 * G – 5 * (R + NIR)	(Manickam et al., 2021)
25	SAVI	Soil Adjusted Vegetation Index	SAVI = (1 + 0.5) * ((NIR - R)/(NIR + R + 0.5))	(Thiam et al., 2021)
26	CRSI	Canopy Response Salinity Index	$CRSI = (NIR * B – G * R)/(NIR * B – G * R)^{1/2}$	(Ramos et al., 2020)
27	GARI	Green Atmospherically Resistant Vegetation Index	GARI = (NIR – (G + y * (B - R))) / (NIR + (G + y * (B – R)))	(Wang et al., 2020e)
28	EVI	Enhanced Vegetation Index	EVI = 2.5 * ((NIR - R)/(NIR + 6 * R – 7.5 * B + 1))	(Ivushkin et al., 2017)
29	MSI	Moisture Stress Index	MSI = SWIR1/NIR	(Ahmadian et al., 2016)
30	RVI	Ratio Vegetation Index	RVI = NIR/R	(Pouladi et al., 2019)

The results of the individual and hybrid models are presented in Table 3 and Fig. 5. Of the eight models proposed, five showed a good fit for the training data (GPR-GOA, SVR-GOA, XGR-GOA, GPR, and XGR), while three models (DNN-GOA, DNN, and SVR) did not have a satisfactory fit. For the validating dataset, the XGR model and its hybrid have the highest prediction capacity with $R^2 = 0.76$ and 0.86; RMSE = 0.1 and 0.17. The continuation is the GPR-GOA (RMSE = 0.14, MAE = 0.11, and $R^2 = 0.5$) and GPR (RMSE = 0.16, MAE = 0.12, and $R^2 = 0.4$). The four models DNN-GOA (RMSE = 0.3, MAE = 0.25, and $R^2 = -1.2$), DNN (RMSE = 0.17, MAE = 0.18, and $R^2 = 0.28$), SVR-GOA (RMSE = 0.16, MAE = 0.12, and $R^2 = 0.39$), and SVR (RMSE = 0.2, MAE = 0.16, and R^2

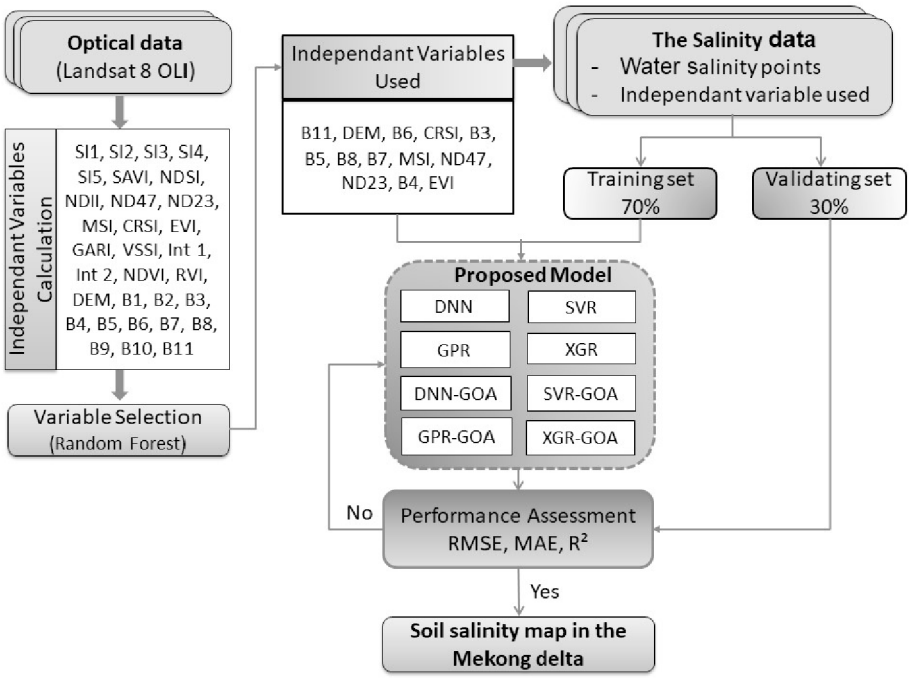


Fig. 3. Flow chart of the soil salinity mapping process.

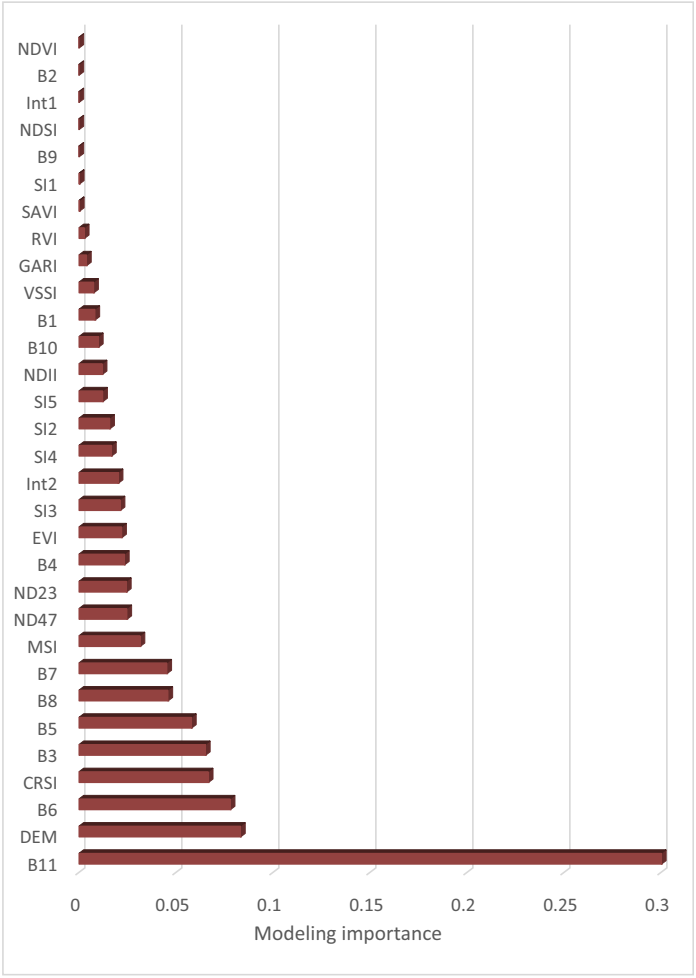


Fig. 4. The importance of the variable conditioning.

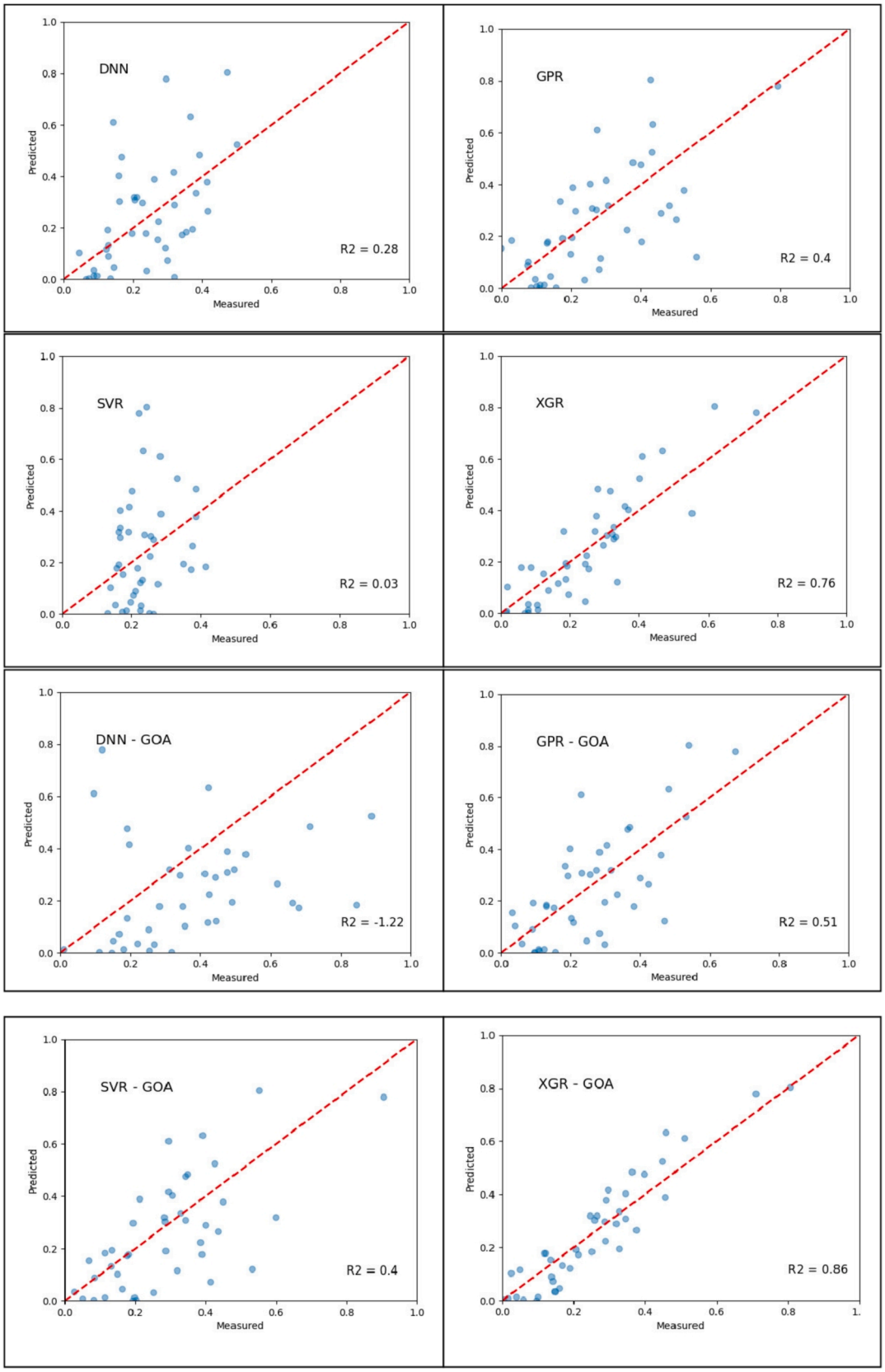


Fig. 5. R^2 value for the validation set.

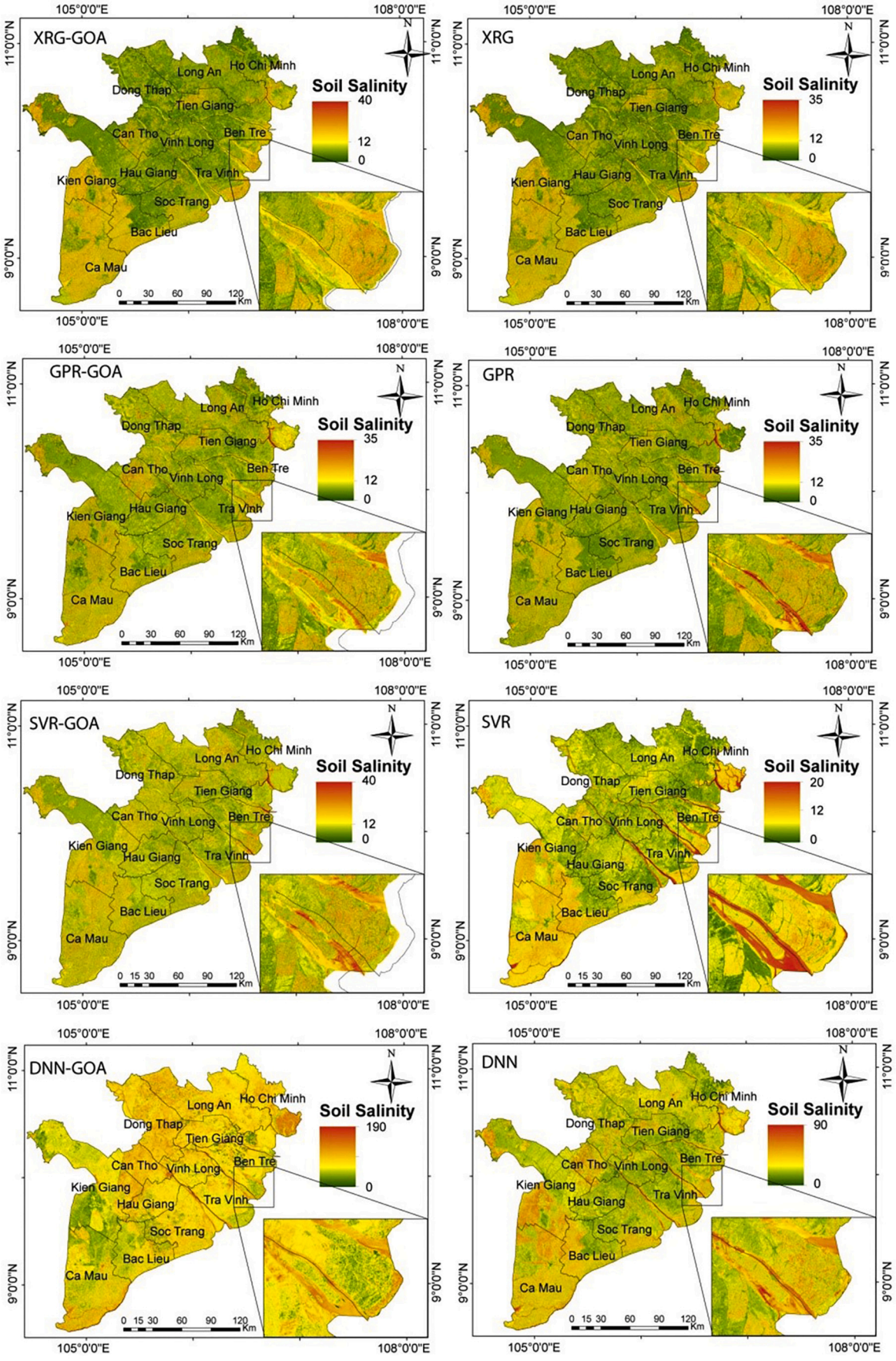


Fig. 6. Soil salinity mapping in the Mekong Delta of Vietnam.

Table 2
The parameters of the proposed models.

Model	Parameters
XG	objective = 'reg:squarederror', n_estimators = 500, learning_rate = 0.3, subsample = 1, colsample_bytree = 1, default = 6
SVM	kernel = 'rbf', gamma = 'scale', C = 1.0, epsilon = 0.1
DNN	1 input layer, 3 hidden layer, 40 neurons, activation function = 'relu', 1 output layer, activation = 'linear', loss = rmse, optimizer = 'adam'
GPR	kernel = 1.0 * Matern(length_scale = 1.0, length_scale_bounds = (1e-1, 10.0),nu = 1.5), random_state = None
XG-GOA	objective = 'reg:squarederror', n_estimators = 500, learning_rate = 0.280872544503152, subsample = 0.964093577693467, colsample_bytree = 1, default = 6
SVM-GOA	kernel = 'rbf', gamma = 99.986133825406, C = 91.9399109732609, epsilon =0.0680353683718021
DNN-GOA	1 input layer, 3 hidden layer, 40 neurons, activation function = 'relu', 1 output layer, activation = 'linear', loss = rmse, batch_size = 25, epoch = 500, pop_size = 30, lower_bound = −1, upper_bound = 1, problem_size = 3881
GPR-GOA	kernel = 94.0836015947774 * ExpSineSquared(length_scale = 64.4006472630534, periodicity = 0.417063266368397, length_scale_bounds = 'fixed', periodicity_bounds = 'fixed') + 85.951675761824 * Matern(length_scale = 99.9999995018722, length_scale_bounds = 'fixed', nu = 0.5), random_state = None

Table 3
Parameters for evaluating models.

	Training dataset			Validating dataset		
	RMSE	MAE	R^2	RMSE	MAE	R^2
DNN-GOA	0.37	0.277	−1.754	0.309	0.255	−1.218
GPR-GOA	4.67E-10	3.41E-10	0.99	0.145	0.119	0.509
SVR-GOA	0.061	0.058	0.923	0.161	0.128	0.398
XGR-GOA	0.0006	0.0004	0.999	0.076	0.065	0.864
DNN	0.165	0.211	0.451	0.176	0.189	0.28
GPR	7.30E-10	4.62E-10	0.99	0.16	0.129	0.404
SVR	0.21	0.165	0.11	0.204	0.163	0.028
XGR	0.00076	0.246	0.999	0.101	0.207	0.76

= 0.02) show lower performance.

3.3. Soil salinity map

According to the above analysis, the XGR-GOA model is better than the other models, so it was used to construct a soil salinity map of the Mekong Delta. This process was carried out by feeding the entire study area (represented by 57 million points), with the 13 independent variables, into the XGR-GOA model, and then generating the soil salinity values (Fig. 4). The provinces of Kien Giang, Ca Mau, Bac Lieu, Soc Trang, Tra Vinh, Ben Tre, and Ho Chi Minh have high salinity values (Fig. 6). The coastal location of these provinces indicates that as the tide rises, seawater enters the land through the rivers, thereby increasing the salinity. At the same time, the provinces of Dong Thap, Long An, Tien Giang, Can Tho, Hau Giang, Vinh Long, eastern Kien Giang have low EC values. These provinces are considered as the "rice bowl" of the country.

4. Discussion

This study develops a comprehensive approach based on the optical imaging and advanced artificial intelligence to identify areas of high salinity. The results are methodologically significant, showing the importance of integrating machine learning and remote sensing to obtain a detailed picture of the distribution of salinity intrusion in the Vietnamese Mekong Delta.

Remote sensing data play an important role in analyzing EC as the effect of salt in the soil results in a specific reflectance, which is the basis for monitoring soil salinity (Wang et al., 2021a). The regions are covered by a crust of white salt, which suggests a high salt level (Wang et al., 2020b). However, in each band of the Landsat 8 OLI data, the spectral reflectance of the samples does not necessarily increase with increasing salinity. This causes difficulties when directly using multispectral bands and their spectral indices for assessment (Davis et al., 2019b). Therefore, it is necessary to identify the main variables that influence soil salinity (Han et al., 2019). To quantify the importance of the independent factors in building the model, we applied an RF model, which filters out factors that minimally affect the exactness of the models. It is very sensitive in detecting the inter-correlation between variables, and helps to avoid the prediction bias (Matin et al., 2018). However, it also increases the computation time, which reduces the performance of the models. This process is very important when we have a dataset with various dependent and independent variables. This assessment is necessary before modeling (Li et al., 2021). In addition, the impact level of selected factors could also be explained and proved by related research and the current status of the study site. The B11 (0.3) band plays a more important role followed by the DEM with an importance level of 0.08. Next in line were B6 with an importance level of 0.077 followed by CRSI (0.066), and B3 (0.065). Similar to the previous studies, salinity index and topography were used to assess the soil salinity. For example, the movement and accumulation of salts is determined by ecological, geological, climatic, and hydrological factors (Wang et al., 2020d), which influence the soil–water balance (Wang et al., 2020c). Therefore, soil salinity differs considerably with differences in geographic location and topography. The water resources in the study area are precipitation, irrigation, and the Mekong River (Wang et al., 2021b). The Mekong Delta is topographically low-lying, and is thus more easily reached by a rising sea level. Moreover, at deeper sections of the riverbed, heavier sea water passes under the layer of fresh water and advances further upstream.

There are interactions between the soil salinity index and vegetation index. In previous studies, areas covered with vegetation have been identified as slightly saline, but the habitat and resilience of different vegetation covers are varied (Ding and Yu, 2014; Peng et al., 2019; Wang et al., 2020c). Regardless of the vegetation index, vegetation zones are generally recognized as soils with no or very low salt content (Ding and Yu, 2014). For this reason, the NDVI index has been removed in several studies (Wang et al., 2020c). This study was not an exception, and five independent variables (B2, B9, Int 1, NDVI, and NDSI) were removed from the models.

The first results were expected because water salinity could successfully predict soil salinity, and has been verified in previous studies. (Salehin et al., 2018) highlighted that water salinity and soil salinity are strongly associated, especially in coastal areas where elevation is low. This is due to the dense river and canal network that carries salty water from the sea to inland areas. (Clarke et al., 2015) assessed soil salinity using a model for soil water balance. In cultivated areas, soil salinity is a result of both the quality and quantity of water, and increases with increase of saline irrigation water. (Rasel et al., 2013) highlighted that soil salinity is responsible for surface water salinity. In the case of the Mekong Delta, the average elevation is ~2 m above sea level, and a large part of the surface in this region has been covered by the rice cultivation, which is often irrigated by river waters. Therefore, soil salinity is influenced by water salinity. Finally, the soil salinity map generated from the observed surface water salinity is in good agreement with maps generated from in situ soil samples (Hoa et al., 2019b; Nguyen et al., 2020b). It consolidates the strong basin-wide interconnection between surface water (i.e., river and canal water) and the shallowest groundwater (Holocene aquifer-qh3) in the study area (Duy et al., 2021; Wagner et al., 2012). Thus, our study provides additional evidence to assess soil salinity using water salinity in the context of climate change.

The final results justified our initial hypothesis that the hybrid models were better than the individual models. This tendency is typical of most of the previous studies that use machine learning to manage environmental problems such as landslides, floods, and groundwater (Bui et al., 2020; Nguyen et al., 2020a). The hybrid model improves

prediction capacity of the base classifications; moreover, instead of being limited to a single classification, this model builds many classifications iteratively to reduce over-addition and under-fit problems (Nhu et al., 2020; Thai Pham et al., 2019). In this study, GOA improved the performance of the individual XGR and GPR models.

The input data to assess the EC value are the same for all models; therefore, the differences in the predictive ability are due to the structural results of each model. The calculation capacity and complexity of each model determines its outcomes (Kisi et al., 2019; Melesse et al., 2020). Because the dataset in our study was not very large (143 water salinity values), XGR-GOA exhibited the best predictive ability. The structure of XGR has a subset of smaller data, which is suitable for trainers. In addition, XGR offers ways to avoid overfitting, such as row down-sampling, column by split levels, and the ability to handle missing data by continuing to train with the model that was built earlier to save time (Abedi et al., 2021b; Wang et al., 2021c). In addition, the XGR algorithm does not contain hidden layers in its structure. Therefore, it works better than data-intelligence algorithms that have hidden layers, such as DNN (Kisi et al., 2019; Sahour et al., 2020). DNNs are suitable for large amounts of data; therefore, this model has no advantage when training with small datasets (Bui et al., 2020; Wang et al., 2020a). Similar to other models (SVR, GPR), which provide the best performance with training data; however, their prediction performance is poor. This shows that these models suffer from over-addition problems due to their small scale. Owing to the non-linear characteristics of various environmental phenomena, including soil salinity, the models work flexibly within a non-linear structure and provide the best results.

Previous studies have applied machine learning models to estimate soil salinity. (Forkuor et al., 2017) used four individual models to estimate the soil properties in southwestern Burkina Faso. The results show that the RFR model with $R^2 = 0.354$ is more efficient. However, this was significantly less successful compared to our models. (Wang et al., 2019) developed five machine learning models to predict salinity intrusion in the Xinjiang Uyghur Autonomous Region, China. The results indicated that the stochastic gradient treeboost (SGT) model with $R^2 = 0.63$ was better. But, even these results show lesser performance than our proposed model. (Wang et al., 2020e) developed the machine learning algorithms, PLSR, convolutional neural network, SVM learning, and RF to analyze salinity intrusion in the Xinjiang region, China. The results show that the RF model with $R^2 = 0.75$ is more successful. (Habibi et al., 2020) applied ANN, PLSR, and DT based on GA to assess soil salinity in central Iran. The results show that the ANN-GA hybrid models with $R^2 = 0.92$ are more accurate than the other models. The results of our study show a general trend compared to the previous studies, and can be used as alternatives to reduce soil salinity problems in the Mekong Delta.

This study is one of the few that were carried out across the Mekong Delta, where soil salinity is rapidly increasing as a result of climate change. The main contribution of this study is the construction of a low-cost novel approach to monitor soil salinity based on the regularly measured water salinity. This can support land-use planning decision-makers, especially in poor countries. However, water salinity depends on the amplitude of the tides, which may influence the performance of the model. The samples in this study showed a strong influence of tides, which were removed to improve the model performance. Additionally, a more detailed analysis is needed to show the weak links in the lower part of earth surface where the water salinity is low, but the soil salinity is high. It also suggests a need for further study on the chemical characteristics of surface water, groundwater, and soil in the area between the river, canal water, and land.

5. Conclusions

This study aimed to examine the predictability of salinity intrusion in the soil based on the value of water salinity using machine learning and remote sensing. The findings are very important and may support decision making for land-use planning, not only for Vietnam, but also for other countries where soil salinity is common. We tested the initial hypothesis that water salinity and soil salinity are strongly related. The second hypothesis was that the hybrid model is more efficient than the individual model for predicting soil salinity. The results show that the water salinity values can significantly classify saline soils by applying machine learning techniques and remote sensing. Among the 31 conditioning factors considered, 13 (B11, DEM, B6, B5, EVI, ND23, B8, MSI, B7, CRSI, B3, ND47, and B7) were found to be the most important for measuring soil salinity. The XGR-GOA model is more efficient than the other models for the tested algorithms, with the highest correlation coefficients (0.86), and lowest RMSE (0.076) and MAE (0.065).

The performance of the models obtained in this study is promising for future low-cost mapping of soil salinity at a regional scale in countries where sea levels are continuously rising and contributing towards the increasing soil salinity, but have scarcity of data. Using water salinity data can reduce soil sampling, especially in large areas. Our findings can help farmers, especially in the developing countries such as Vietnam, by improving their knowledge about the salinity-vulnerable regions so that they may control their practices, adapt relevant cropping systems, and support decision-makers to reduce soil salinity monitoring costs in land use planning and soil management. Although this study was conducted in Vietnam, the results can be used for all the regions where soil salinity is increasing due to climate change.

Declaration of Competing Interest

None.

Acknowledgements

This study was funded by the project “A research on constructing a low-cost pilot model using advanced technology for drinking water supply in the water-scarce zones in the Mekong Delta” (ĐTĐL.CN-50/18). The authors thank all anonymous reviewers for their critical and constructive comments, which have improved the quality of the manuscript.

References

Abedi, F., Amirian-Chakan, A., Faraji, M., Taghizadeh-Mehrjardi, R., Kerry, R., Razmjoue, D., Scholten, T., 2021a. Salt dome related soil salinity in southern Iran: prediction and mapping with averaging machine learning models. Land Degrad. Dev. 32 (3), 1540–1554.

Abedi, F., Amirian Chakan, A., Faraji, M., Taghizadeh-Mehrjardi, R., Kerry, R., Razmjoue, D., Scholten, T., 2021b. Salt dome related soil salinity in southern Iran: prediction and mapping with averaging machine learning models. Land Degrad. Dev. 32, 1540–1554.

Ahmadian, N., Ghasemi, S., Wigneron, J.-P., Zölitz, R., 2016. Comprehensive study of the biophysical parameters of agricultural crops based on assessing Landsat 8 OLI and Landsat 7 ETM+ vegetation indices. GISci. Remote Sens. 53 (3), 337–359.

Allbed, A., Kumar, L., Sinha, P., 2014. Mapping and modelling spatial variation in soil salinity in the Al Hassa oasis based on remote sensing indicators and regression techniques. Remote Sens. 6 (2), 1137–1157.

Arndt, C., Tarp, F., Thurlow, J., 2015. The economic costs of climate change: a multi-sector impact assessment for Vietnam. Sustainability 7 (4), 4131–4145.

Bartels, D., Sunkar, R., 2005. Drought and salt tolerance in plants. Crit. Rev. Plant Sci. 24 (1), 23–58.

Behnamian, A., Millard, K., Banks, S.N., White, L., Richardson, M., Pasher, J., 2017. A systematic approach for variable selection with random forests: achieving stable variable importance values. IEEE Geosci. Remote Sens. Lett. 1–5.

Bouaziz, M., Matschullat, J., Gloaguen, R., 2011. Improved remote sensing detection of soil salinity from a semi-arid climate in Northeast Brazil. Compt. Rendus Geosci. 343 (11), 795–803.

Bui, Q.-T., Nguyen, Q.-H., Nguyen, X.L., Pham, V.D., Nguyen, H.D., Pham, V.-M., 2020. Verification of novel integrations of swarm intelligence algorithms into deep learning neural network for flood susceptibility mapping. J. Hydrol. 581, 124379.

Clarke, D., Williams, S., Jahiruddin, M., Parks, K., Salehin, M., 2015. Projections of on-farm salinity in coastal Bangladesh. Environ Sci Process Impacts 17 (6), 1127–1136.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Costache, R., Ngo, P.T.T., Bui, D.T., 2020. Novel ensembles of deep learning neural network and statistical learning for flash-flood susceptibility mapping. Water 12 (6), 1549.

Das, R.S., Rahman, M., Sufian, N.P., Rahman, S.M.A., Siddique, M.A.M., 2020. Assessment of soil salinity in the accreted and non-accreted land and its implication

on the agricultural aspects of the Noakhali coastal region, Bangladesh. Heliyon 6 (9), e04926.
Davis, E., Wang, C., Dow, K., 2019a. Comparing Sentinel-2 MSI and Landsat 8 OLI in soil salinity detection: a case study of agricultural lands in coastal North Carolina. Int. J. Remote Sens. 40 (16), 6134–6153.
Davis, E., Wang, C., Dow, K., 2019b. Comparing Sentinel-2 MSI and Landsat 8 OLI in soil salinity detection: a case study of agricultural lands in coastal North Carolina. Int. J. Remote Sens. 40, 1–20.
Delavar, M.A., Naderi, A., Ghorbani, Y., Mehrpouyan, A., Bakhshi, A., 2020. Soil salinity mapping by remote sensing south of Urmia Lake, Iran. Geoderma. Reg. 22, e00317.
Di Bucchianico, A., 2008. Coefficient of determination (R 2). Ency. Statis. Qual. Reliab. 1.
Ding, J., Yu, D., 2014. Monitoring and evaluating spatial variability of soil salinity in dry and wet seasons in the Werigan–Kuqa Oasis, China, using remote sensing and electromagnetic induction instruments. Geoderma 235-236, 316–322.
Dinh, P.-H., 2021. A novel approach based on grasshopper optimization algorithm for medical image fusion. Expert Syst. Appl. 171, 114576.
Duy, N.L., Nguyen, T.V.K., Nguyen, D.V., Tran, A.T., Nguyen, H.T., Heidbüchel, I., Merz, B., Apel, H., 2021. Groundwater dynamics in the Vietnamese Mekong Delta: trends, memory effects, and response times. J. Hydrol. 33, 100746.
Dwivedi, S., Vardhan, M., Tripathi, S., 2020. An effect of chaos grasshopper optimization algorithm for protection of network infrastructure. Comput. Netw. 176, 107251.
Elhag, M., Bahrawi, J.A., 2017. Soil salinity mapping and hydrological drought indices assessment in arid environments based on remote sensing techniques. Geosci. Instrum. Meth. Data Syst. 6 (1), 149–158.
Erkin, N., Zhu, L., Gu, H., Tusiyiti, A., 2019. Method for predicting soil salinity concentrations in croplands based on machine learning and remote sensing techniques. J. Appl. Remote. Sens. 13, 1.
Farahmand, N., Sadeghi, V., 2020. Estimating soil salinity in the dried lake bed of Urmia Lake using optical sentinel-2 images and nonlinear regression models. J. Indian Soc. Remote Sens. 1–13.
Fathizad, H., Ardakani, M.A.H., Sodaiezadeh, H., Kerry, R., Taghizadeh-Mehrjardi, R., 2020. Investigation of the spatial and temporal variation of soil salinity using random forests in the central desert of Iran. Geoderma 365, 114233.
Forkuor, G., Hounkpatin, O.K., Welp, G., Thiel, M., 2017. High resolution mapping of soil properties using remote sensing variables in South-Western Burkina Faso: a comparison of machine learning and multiple linear regression models. PLoS One 12 (1), e0170478.
Forkuor, G., Dimobe, K., Serme, I., Tondoh, J.E., 2018. Landsat-8 vs. Sentinel-2: examining the added value of sentinel-2's red-edge bands to land-use and land-cover mapping in Burkina Faso. GISci. Remote Sens. 55 (3), 331–354.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.
Gampa, S.R., Jasthi, K., Goli, P., Das, D., Bansal, R., 2020. Grasshopper optimization algorithm based two stage fuzzy multiobjective approach for optimum sizing and placement of distributed generations, shunt capacitors and electric vehicle charging stations. J. Energy Storage 27, 101117.
Garajeh, M.K., Malakyar, F., Weng, Q., Feizizadeh, B., Blaschke, T., Lakes, T., 2021. An automated deep learning convolutional neural network algorithm applied for soil salinity distribution mapping in Lake Urmia, Iran. Sci. Total Environ. 778, 146253.
Guo, Y., Jia, X., Paull, D., 2018. Effective sequential classifier training for SVM-based multitemporal remote sensing image classification. IEEE Trans. Image Process. 27 (6), 3036–3048.
Habiba, U., Abedin, M.A., Hassan, A.W.R., Shaw, R., 2015. Food Security and Risk Reduction in Bangladesh. Springer.
Habibi, V., Ahmadi, H., Jafari, M., Moeini, A., 2020. Machine learning and multispectral data-based detection of soil salinity in an arid region, Central Iran. Environ. Monit. Assess. 192 (12), 1–13.
Han, L., Liu, D., Cheng, G., Zhang, G., Wang, L., 2019. Spatial distribution and genesis of salt on the saline playa at Qehan Lake, Inner Mongolia, China. CATENA 177, 22–30.
He, T., Kong, R., Holmes, A.J., Nguyen, M., Sabuncu, M.R., Eickhoff, S.B., Bzdok, D., Feng, J., Yeo, B.T.T., 2020. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. NeuroImage 206, 116276.
Hoa, P., Nguyen Vu, G., Nguyen, B., Le, H., Pham, T.D., Hasanlou, M., Bui, D., 2019a. 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).
Hoa, P.V., Giang, N.V., Binh, N.A., Hai, L.V.H., Pham, T.-D., Hasanlou, M., Tien Bui, D., 2019b. 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 Sens. 11 (2), 128.
Ivushkin, K., Bartholomeus, H., Bregt, A.K., Pulatov, A., 2017. Satellite thermography for soil salinity assessment of cropped areas in Uzbekistan. Land Degrad. Dev. 28 (3), 870–877.
Ivushkin, K., Bartholomeus, H., Bregt, A.K., Pulatov, A., Kempen, B., De Sousa, L., 2019. Global mapping of soil salinity change. Remote Sens. Environ. 231, 111260.
Jaramillo, J., Velasquez, J.D., Franco, C.J., 2017. Research in financial time series forecasting with SVM: contributions from literature. IEEE Lat. Am. Trans. 15 (1), 145–153.
Jia, Y., Jin, S., Savi, P., Gao, Y., Tang, J., Chen, Y., Li, W., 2019. GNSS-R soil moisture retrieval based on a XGboost machine learning aided method: performance and validation. Remote Sens. 11 (14), 1655.
Jiang, H., Rusuli, Y., Amuti, T., He, Q., 2019. Quantitative assessment of soil salinity using multi-source remote sensing data based on the support vector machine and artificial neural network. Int. J. Remote Sens. 40 (1), 284–306.
Jindal, A., Kumar, N., Singh, M., 2020. Internet of energy-based demand response management scheme for smart homes and PHEVs using SVM. Futur. Gener. Comput. Syst. 108, 1058–1068.
Kisi, O., Heddam, S., Yaseen, Z.M., 2019. The implementation of univariable scheme-based air temperature for solar radiation prediction: new development of dynamic evolving neural-fuzzy inference system model. Appl. Energy 241, 184–195.
Li, M., Xu, Y., Men, J., Yan, C., Tang, H., Zhang, T., Li, H., 2021. Hybrid variable selection strategy coupled with random forest (RF) for quantitative analysis of methanol in methanol-gasoline via Raman spectroscopy. Spectrochim. Acta A Mol. Biomol. Spectrosc. 251, 119430.
Machado, R.M.A., Serralheiro, R.P., 2017. Soil salinity: effect on vegetable crop growth. Management practices to prevent and mitigate soil salinization. Horticulturae 3 (2), 30.
Malik, S., Harode, R., Kunwar, A., 2020. XGBoost: A Deep Dive into Boosting (Introduction Documentation).
Manasa, M., Katukuri, N.R., Nair, S.S.D., Haojie, Y., Yang, Z., bo Guo, R., 2020. Role of biochar and organic substrates in enhancing the functional characteristics and microbial community in a saline soil. J. Environ. Manag. 269, 110737.
Manickam, L., Subramanian, D., Khandal, S., Hegde, R., 2021. Modeling and mapping of salt-affected soils through spectral indices in Inland Plains of semi-arid agro-ecological region. J. Indian Soc. Remote Sens. 1–7.
Markham, B., Barsi, J., Montanaro, M., McCorkel, J., Gerace, A., Pedelty, J., Hook, S., Raqueno, N., Anderson, C., Haque, M.O., 2018. Landsat-8 on-Orbit and Landsat-9 Pre-Launch Sensor Radiometric Characterization, Earth Observing Missions and Sensors: Development, Implementation, and Characterization V. International Society for Optics and Photonics, p. 1078104.
Matin, S.S., Farahzadi, L., Makaremi, S., Chelgani, S.C., Sattari, G., 2018. Variable selection and prediction of uniaxial compressive strength and modulus of elasticity by random forest. Appl. Soft Comput. 70, 980–987.
Matinfar, H., Fariabi, A., Alavipanah, S., 2020. Evaluating different spectral indices in identification and preparation of soil salinity mapping of arid region of Iran. Desert 25 (1), 77–85.
Mehrer, J., Spoerer, C.J., Kriegeskorte, N., Kietzmann, T.C., 2020. Individual differences among deep neural network models. Nat. Commun. 11 (1), 1–12.
Melesse, A.M., Khosravi, K., Tiefenbacher, J.P., Heddam, S., Kim, S., Mosavi, A., Pham, B. T., 2020. River water salinity prediction using hybrid machine learning models. Water 12 (10), 2951.
Moussa, I., Walter, C., Michot, D., Boukary, I.A., Nicolas, H., Pichelin, P., Guéro, Y., 2020. Soil salinity assessment in irrigated Paddy fields of the Niger Valley using a four-year time series of Sentinel-2 satellite images. Remote Sens. 12 (20), 3399.
Mulder, V., De Bruin, S., Schaepman, M.E., Mayr, T., 2011. The use of remote sensing in soil and terrain mapping—a review. Geoderma 162 (1–2), 1–19.
Nguyen, P.T., Koedsin, W., McNeil, D., Van, T.P., 2018. Remote sensing techniques to predict salinity intrusion: application for a data-poor area of the coastal Mekong Delta, Vietnam. Int. J. Remote Sens. 39 (20), 6676–6691.
Nguyen, H.-D., Pham, V.-D., Nguyen, Q.-H., Pham, V.-M., Pham, M.H., Vu, V.M., Bui, Q.-T., 2020a. An optimal search for neural network parameters using the Salp swarm optimization algorithm: a landslide application. Remote Sens. Lett. 11 (4), 353–362.
Nguyen, K.-A., Liou, Y.-A., Tran, H.-P., Hoang, P.-P., Nguyen, T.-H., 2020b. 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. Prog. Earth Planet Sci. 7 (1), 1–16.
Nhu, V.-H., Shirzadi, A., Shahabi, H., Chen, W., Clague, J., Geertsema, M., Jaafari, A., Avand, M., Miraki, S., Asl, D., Pham, B., Bin, B., Ahmad Lee, S., 2020. Shallow landslide susceptibility mapping by random Forest Base classifier and its ensembles in a semi-arid region of Iran. Forests 11, 421.
Nicolas, H., Walter, C., 2006. Detecting salinity hazards within a semiarid context by means of combining soil and remote-sensing data. Geoderma 134 (1–2), 217–230.
Panda, M., 2020. Elephant search optimization combined with deep neural network for microarray data analysis. J. King Saud Univ. - Comput. Inf. Sci. 32 (8), 940–948.
Peng, J., Biswas, A., Jiang, Q., Zhao, R., Hu, J., Hu, B., Shi, Z., 2019. Estimating soil salinity from remote sensing and terrain data in southern Xinjiang Province, China. Geoderma 337, 1309–1319.
Pouladi, N., Jafarzadeh, A.A., Shahbazi, F., Ghorbani, M.A., 2019. Design and implementation of a hybrid MLP-FFA model for soil salinity prediction. Environ. Earth Sci. 78 (5), 1–10.
Qi, Y., Huo, Z., Feng, S., Adeloye, A.J., Dai, X., 2018. Prediction of consumptive use under different soil moisture content and soil salinity conditions using artificial neural network models. Irrig. Drain. 67 (4), 615–624.
Ramos, T.B., Castanheira, N., Oliveira, A.R., Paz, A.M., Darouich, H., Simionesei, L., Farzamian, M., Gonçalves, M.C., 2020. Soil salinity assessment using vegetation indices derived from Sentinel-2 multispectral data. Application to Lezíria Grande, Portugal. Agric. Water Manag. 241, 106387.
Rasel, H., Hasan, M., Ahmed, B., Miah, M., 2013. Investigation of soil and water salinity, its effect on crop production and adaptation strategy. Int. J. Water Res. Environ. Eng. 5 (8), 475–481.
Rohmer, J., Foerster, E., 2011. Global sensitivity analysis of large-scale numerical landslide models based on Gaussian-process meta-modeling. Comput. Geosci. 37 (7), 917–927.
Sahour, H., Gholami, V., Vazifedan, M., 2020. A comparative analysis of statistical and machine learning techniques for mapping the spatial distribution of groundwater salinity in a coastal aquifer. J. Hydrol. 591, 125321.
Salehin, M., Chowdhury, M.M.A., Clarke, D., Mondal, S., Nowreen, S., Jahiruddin, M., Haque, A., 2018. Mechanisms and Drivers of Soil Salinity in Coastal Bangladesh, Ecosystem Services for Well-Being in Deltas. Palgrave Macmillan, Cham, pp. 333–347.

Saremi, S., Mirjalili, S., Mirjalili, S., Dong, J.S., 2020. Grasshopper optimization algorithm: theory, literature review, and application in hand posture estimation. Nature-Inspired Optimizers 107–122.
Seifi, M., Ahmadi, A., Neyshabouri, M.-R., Taghizadeh-Mehrjardi, R., Bahrami, H.-A., 2020. 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. 20, 100398.
Siebert, S., Henrich, V., Frenken, K., Burke, J., 2013. Update of the digital global map of irrigation areas to version 5. In: Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany and Food and Agriculture Organization of the United Nations, Rome, Italy.
Singhania, S., Fernandez, N., Rao, S., 2017. 3han: A Deep Neural Network for Fake News Detection, International Conference on Neural Information Processing. Springer, pp. 572–581.
Smith, H.K., Clarkson, G.J., Taylor, G., Thompson, A.J., Clarkson, J., Rajpoot, N.M., 2014. Automatic detection of regions in spinach canopies responding to soil moisture deficit using combined visible and thermal imagery. PLoS One 9 (6), e97612.
Solangi, K.A., Siyal, A.A., Wu, Y., Abbasi, B., Solangi, F., Lakhiar, I.A., Zhou, G., 2019. An assessment of the spatial and temporal distribution of soil salinity in combination with field and satellite data: a case study in Sujawal District. Agronomy 9 (12), 869.
Stamenkovic, J., Guerriero, L., Ferrazzoli, P., Notarnicola, C., Greifeneder, F., Thiran, J.-P., 2017. Soil moisture estimation by SAR in alpine fields using Gaussian process regressor trained by model simulations. IEEE Trans. Geosci. Remote Sens. 55 (9), 4899–4912.
Thai Pham, B., Shirzadi, A., Shahabi, H., Omidvar, E., Singh, S.K., Sahana, M., Talebpour Asl, D., Bin Ahmad, B., Kim Quoc, N., Lee, S., 2019. Landslide susceptibility assessment by novel hybrid machine learning algorithms. Sustainability 11 (16), 4386.
Thiam, S., Villamor, G.B., Faye, L.C., Sène, J.H.B., Diwediga, B., Kyei-Baffour, N., 2021. Monitoring land use and soil salinity changes in coastal landscape: a case study from Senegal. Environ. Monit. Assess. 193 (5), 1–18.
Torlay, L., Perrone-Bertolotti, M., Thomas, E., Baciu, M., 2017. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. Brain Inform. 4 (3), 159–169.
Tran, T.V., Tran, D.X., Myint, S.W., Huang, C.-Y., Pham, H.V., Luu, T.H., Vo, T.M., 2019. Examining spatiotemporal salinity dynamics in the Mekong River Delta using Landsat time series imagery and a spatial regression approach. Sci. Total Environ. 687, 1087–1097.
Vermeulen, D., Van Niekerk, A., 2017. Machine learning performance for predicting soil salinity using different combinations of geomorphometric covariates. Geoderma 299, 1–12.
Vermote, E., Roger, J.-C., Franch, B., Skakun, S., 2018. LaSRC (Land Surface Reflectance Code): overview, application and validation using MODIS, VIIRS, LANDSAT and Sentinel 2 data's, IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE 8173–8176.
Wagner, F., Tran, V.B., Renaud, F.G., 2012. Groundwater Resources in the Mekong Delta: Availability, Utilization and Risks, The Mekong Delta System. Springer, pp. 201–220.
Wang, X.-Z., Zhang, T., Wang, R., 2017. Noniterative deep learning: incorporating restricted boltzmann machine into multilayer random weight neural networks. IEEE Trans. Syst. Man, Cybern. Syst. 49 (7), 1299–1308.
Wang, X., Zhang, F., Ding, J., Latif, A., Johnson, V.C., 2018. Estimation of soil salt content (SSC) in the Ebinur Lake wetland National Nature Reserve (ELWNNR), Northwest China, based on a Bootstrap-BP neural network model and optimal spectral indices. Sci. Total Environ. 615, 918–930.
Wang, F., Yang, S., Yang, W., Yang, X., Jianli, D., 2019. Comparison of machine learning algorithms for soil salinity predictions in three dryland oases located in Xinjiang Uyghur autonomous region (XJUAR) of China. Eur. J. Remote Sens. 52 (1), 256–276.
Wang, F., Shi, Z., Biswas, A., Yang, S., Ding, J., 2020a. Multi-algorithm comparison for predicting soil salinity. Geoderma 365, 114211.
Wang, J., Ding, J., Yu, D., Teng, D., He, B., Chen, X., Ge, X., Zhang, Z., Wang, Y., Yang, X., 2020b. Machine learning-based detection of soil salinity in an arid desert region, Northwest China: a comparison between Landsat-8 OLI and Sentinel-2 MSI. Sci. Total Environ. 707, 136092.
Wang, J., Ding, J., Yu, D., Teng, D., He, B., Chen, X., Ge, X., Zhang, Z., Wang, Y., Yang, X., Shi, T., Su, F., 2020c. Machine learning-based detection of soil salinity in an arid desert region, Northwest China: a comparison between Landsat-8 OLI and Sentinel-2 MSI. Sci. Total Environ. 707, 136092.
Wang, N., Xue, J., Peng, J., Biswas, A., He, Y., Shi, Z., 2020d. Integrating remote sensing and landscape characteristics to estimate soil salinity using machine learning methods: a case study from southern Xinjiang, China. Remote Sens. 12, 4118.
Wang, N., Xue, J., Peng, J., Biswas, A., He, Y., Shi, Z., 2020e. Integrating remote sensing and landscape characteristics to estimate soil salinity using machine learning methods: a case study from southern Xinjiang, China. Remote Sens. 12 (24), 4118.
Wang, J., Peng, J., Li, H., Yin, C., Liu, W., Wang, T., Zhang, H., 2021a. Soil salinity mapping using machine learning algorithms with the Sentinel-2 MSI in arid areas, China. Remote Sens. 13 (2), 305.
Wang, J., Peng, J., Li, H., Yin, C., Liu, W., Wang, T., Zhang, H., 2021b. Soil salinity mapping using machine learning algorithms with the Sentinel-2 MSI in arid areas, China. Remote Sens. 13, 305.
Wang, X., Fu, D., Wang, Y., Guo, Y., Ding, Y., 2021c. The XGBoost and the SVM-based prediction models for bioretention cell decontamination effect. Arab. J. Geosci. 14.
Wassmann, R., Phong, N.D., Tho, T.Q., Hoanh, C.T., Khoi, N.H., Hien, N.X., Vo, T.B.T., Tuong, T.P., 2019. High-resolution mapping of flood and salinity risks for rice production in the Vietnamese Mekong Delta. Field Crop Res. 236, 111–120.
Whitney, K., Scudiero, E., El-Askary, H.M., Skaggs, T.H., Allali, M., Corwin, D.L., 2018. Validating the use of MODIS time series for salinity assessment over agricultural soils in California, USA. Ecol. Indic. 93, 889–898.
Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30 (1), 79–82.
Wu, W., 2019. A brief review on soil salinity mapping by optical and radar remote sensing. Res. Dev. Saline Agric. 53–65.
Wu, W., Zucca, C., Muhaimeed, A.S., Al-Shafie, W.M., Fadhil Al-Quraishi, A.M., Nangia, V., Zhu, M., Liu, G., 2018. Soil salinity prediction and mapping by machine learning regression in C entral M esopotamia, I raq. Land Degrad. Dev. 29 (11), 4005–4014.
Wu, W., Muhaimeed, A.S., Al-Shafie, W.M., Al-Quraishi, A.M.F., 2020. Using Radar and Optical Data for Soil Salinity Modeling and Mapping in Central Iraq, Environmental Remote Sensing and GIS in Iraq. Springer, pp. 19–40.
Yang, R.-M., Guo, W.-W., 2019. Using Sentinel-1 imagery for soil salinity prediction under the condition of coastal restoration. IEEE J- STARS 12 (5), 1482–1488.
Zhang, Q., Zhou, Z.-S., Caccetta, P., Simons, J., Li, L., 2021. Sentinel-1 imagery incorporating machine learning for Dryland Salinity Monitoring: a case study in Esperance, Western Australia. IGARSS 2020–2020 IEEE International Geoscience and Remote Sensing Symposium 4914–4917.