Environmental Science and Pollution Research (2023) 30:74340–74357
https://doi.org/10.1007/s11356-023-27516-x

RESEARCH ARTICLE

Soil salinity prediction using hybrid machine learning and remote sensing in Ben Tre province on Vietnam's Mekong River Delta

Huu Duy Nguyen[1] · Chien Pham Van[2] · Tien Giang Nguyen[3] · Dinh Kha Dang[3] · Thi Thuy Nga Pham[4] · Quoc-Huy Nguyen[1] · Quang-Thanh Bui[1]

Received: 19 December 2022 / Accepted: 4 May 2023 / Published online: 19 May 2023

Abstract

Soil salinization is considered one of the disasters that have significant effects on agricultural activities in many parts of the world, particularly in the context of climate change and sea level rise. This problem has become increasingly essential and severe in the Mekong River Delta of Vietnam. Therefore, soil salinity monitoring and assessment are critical to building appropriate strategies to develop agricultural activities. This study aims to develop a low-cost method based on machine learning and remote sensing to map soil salinity in Ben Tre province, which is located in Vietnam's Mekong River Delta. This objective was achieved by using six machine learning algorithms, including Xgboost (XGR), sparrow search algorithm (SSA), bird swarm algorithm (BSA), moth search algorithm (MSA), Harris hawk optimization (HHO), grasshopper optimization algorithm (GOA), particle swarm optimization algorithm (PSO), and 43 factors extracted from remote sensing images. Various indices were used, namely, root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2) to estimate the efficiency of the prediction models. The results show that six optimization algorithms successfully improved XGR model performance with an R^2 value of more than 0.98. Among the proposed models, the XGR-HHO model was better than the other models with a value of R^2 of 0.99 and a value of RMSE of 0.051, by XGR-GOA (R^2 = 0.931, RMSE = 0.055), XGR-MSA (R^2 = 0.928, RMSE = 0.06), XGR-BSA (R^2 = 0.926, RMSE = 0.062), XGR-SSA (R^2 = 0.917, 0.07), XGR-PSO (R^2 = 0.916, RMSE = 0.08), XGR (R^2 = 0.867, RMSE = 0.1), CatBoost (R^2 = 0.78, RMSE = 0.12), and RF (R^2 = 0.75, RMSE = 0.19), respectively. These proposed models have surpassed the reference models (CatBoost and random forest). The results indicated that the soils in the eastern areas of Ben Tre province are more saline than in the western areas. The results of this study highlighted the effectiveness of using hybrid machine learning and remote sensing in soil salinity monitoring. The finding of this study provides essential tools to support farmers and policymakers in selecting appropriate crop types in the context of climate change to ensure food security.

Keywords Soil salinity · Machine learning · Remote sensing · Ben Tre · Mekong Delta

Responsible Editor: Philippe Garrigues

✉ Tien Giang Nguyen
giangnt@vnu.edu.vn

Huu Duy Nguyen
nguyenhuuduy@hus.edu.vn

Chien Pham Van
Pchientvct_tv@tlu.edu.vn

Dinh Kha Dang
dangdinhkha@hus.edu.vn

Thi Thuy Nga Pham
phamthithuynga_t62@hus.edu.vn

Quoc-Huy Nguyen
huyquoc2311@hus.edu.vn

Quang-Thanh Bui
thanhbq@vnu.edu.vn

[1] Faculty of Geography, VNU University of Science, Vietnam National University, Hanoi, Vietnam

[2] Thuyloi University, 175 Tay Son, Dong Da, Hanoi, Vietnam

[3] Faculty of Hydrology, Meteorology and Oceanography, VNU University of Science, Vietnam National University, 334 Nguyen Trai, Thanh Xuan District, Hanoi, Vietnam

[4] Center for Environmental Fluid Dynamics, VNU University of Science, Vietnam National University, 334 Nguyen Trai, Thanh Xuan District, Hanoi, Vietnam

Introduction

The soil ecosystem is one of the world's most complex and diverse. Furthermore, they provide 98.8% of humanity's food as well as carbon storage, climate regulation, and mitigation of climate change (Nguyen et al. 2021b). However, the intensification of agricultural production, various anthropogenic pressures, and climate change were contributing to soil degradation. The primary forms of this degradation are erosion, loss of organic carbon, pollution, sealing, compaction, salinization, waterlogging, acidification, nutrient imbalance, and loss of biodiversity (Khormali et al. 2009). Among these threats, soil salinity is a significant problem affecting agricultural development worldwide. Salinization is the accumulation of salts in the soil under the influence of salt water supply, the aridity of the climate, or particular hydrological conditions (Naimi et al. 2021; Wang et al. 2020a; Wu et al. 2018). It is estimated that 10 to 30% of the irrigated areas of the world are affected by salinity or alkalinity, i.e., approximately 76 million hectares of irrigated land, of which 69% of affected land is located in Asia, 19% in Africa, and 5% in Europe (Wicke et al. 2011). Salinization affects about 1.5 million hectares of cultivated land annually (Gorji et al. 2017), causing a decrease of about 10% in global food production (Wu et al. 2018). Therefore, assessing the distribution of soil salinity in space is necessary to support decision-makers in building appropriate agriculture development strategies to address food security issues in countries.

The increase in the level of salinity in recent years had negative impacts on the problem of food security in the world. Therefore, various studies have been conducted in the last decades to assess and quantify soil salinity. These methods have been divided into directive (the traditional) and indirect (Nguyen et al. 2021b; Wang et al. 2019a; Wu et al. 2018).

The traditional method is mainly based on measuring the laboratory's electrical conductivity of soil samples (Gorji et al. 2015). Although these methods have been proven effective in assessing and quantifying soil salinity over the past decades, several studies have pointed out that this method needs to be revised and suitable for monitoring the evolution of these phenomena. In particular, this method is complicated to monitor regularly over a large area because it requires many samples to sufficiently evaluate soil salinity characteristics due to their fast spatial modification over short distances (Mulder et al. 2011). Although there has been an increase in the number of the type of proximal sensors including soil-visible and near-infrared (Vis-NIR) and electromagnetic induction (EMI) (Corwin 2021; Hu et al. 2019; Peng et al. 2022), however, these instruments monitor soil salinity in a small area (Vermeulen and Van Niekerk 2017). This requires several sensors to monitor large areas at the required spatial resolutions effectively.

To limit these gaps, in the last years, remote sensing has received attention from researchers as a promising method to monitor soil salinity in a wide area. At different scales and resolutions, remote sensing images are advantageous for earth observation. Currently, remote sensing data is crucial to evaluating and mapping soil salinity both globally and regionally (Taghizadeh-Mehrjardi et al. 2016). Currently, radar and optical data have recently become available. Nevertheless, visual remote sensing products depend on solar radiation and, therefore, are limited to monitoring soil salinity in the presence of cloud cover and rely on solar radiation. Radar sensors provide images with high spatial and temporal resolution. These images offer alternative perspectives to improve the ability to monitor soil salinity. In addition, radar sensors can obtain information regardless of weather conditions (clouds, rain, etc.) and time conditions (day-night). However, one of the significant challenges when using remote sensing imagery is its inability to effectively monitor subsurface processes that do not directly influence the spectral responses of topsoil. Additionally, this method uses radar backscatter coefficients to monitor soil salinity, so in many cases, it is mistaken for soil moisture.

Studies of soil salinity have used geostatistical techniques such as kinging, universal kriging, and regression kriging, especially in interpolating soil salinity from soil sample analysis (Hengl et al. 2007; Li et al. 2015; Shahabi et al. 2017; Tajgardan et al. 2010). Due to their unbiased estimates of the weights surrounding sampling points, these methods have proven to be effective. The potential relationship between remote sensing data and soil characteristics has been investigated using a variety of regression methods, such as multivariate linear regression (Lesch et al. 1995a, b; Vermeulen and Van Niekerk 2017). However, the successful use of this method must satisfy certain assumptions such as the standard normal distribution and the linear relationship. Therefore, this method can lead to significant errors. In the natural environment, soil characteristics are not normally distributed, limiting regression models' quantification capacity. Moreover, the relationships between environmental factors and soil salinity are never linear. Therefore, these methods have been recommended to be replaced by more powerful methods to accurately interpolate soil salinity. Recent research has demonstrated that machine learning-based remote sensing is a useful technique for monitoring soil salinity, including random forest (Fathizad et al. 2020; Wang et al. 2021a), Xgboost (Nguyen et al. 2021b), support vector machine (Jiang et al. 2019), decision tree (Rao et al. 2006; Taghizadeh-Mehrjardi et al. 2014), and artificial neural network (Dai et al. 2011; Jiang et al. 2019). These methods find the linear and non-linear between independent and dependent variables based on the inherent relationships of the data. Therefore, machine learning can effectively solve complex linear patterns of soil salinity phenomena, even in data-limited regions. They also tend to have fewer user-defined parameters, allow estimates of independent variables' importance and can deal with noisy data. However, the issues of overfitting and underfitting are still considered significant challenges when using machine learning. Also, the extrapolation problem, i.e., the model cannot predict the soil salinity if the data is out of range of the training data which needs to be solved to support the decision-makers. Several studies have pointed out that these problems can be solved when integrating appropriate optimization algorithms with individual models (Nguyen 2022a; Tran and Kim 2022). Optimization algorithms can be divided into three main groups: evolutionary based as genetic algorithm (Calixto et al. 2010), swarm based as particle swarm optimization (Jiang and Xue 2022), artificial bee colony (Li et al. 2019), gray wolf optimizer (Cui et al. 2022; Zhou et al. 2022), and physics based as Henry gas solubility optimization (Ding et al. 2021), atom search optimization (Hua et al. 2022). These algorithms were successful in improving the predictive ability of individual models. However, there is no universal conclusion for the best models for soil salinity monitoring. Moreover, according to the non-free-lunch, no general models can solve the problem of soil salinity in all regions (Nguyen 2022a). Therefore, developing new models is essential in practice and in science.

This study aims to develop novel models by integrating the Xgboost (XGB) with optimization algorithms, namely, sparrow search algorithm (SSA), bird swarm algorithm (BSA), moth search algorithm (MSA), Harris hawks optimization (HHO), gray wolf optimizer (GWO), and particle swarm optimization (PSO) in Ben Tre province where often affected by soil salinity problem. The performances of these models are compared with reference models like random forest and CatBoost. The novelty of this study is that the first time the XGB model is integrated with SSA, BSA, MSE, HHO, GWO, and PSO algorithms to estimate soil salinity in the Ben Tre province of Vietnam where the salinity problem is increasingly serious in the context of climate change and sea level rise. The initial hypothesis in this study is that the four optimization algorithms will successfully improve the performance of the XGB model and its accuracy of the hybris models (XGB-SSA, XGB-BSA, XGB-MSA, and XGB-HHO) will outperform the reference models, namely, XGB-GWO and XGB-PSO. This approach can also adapt to the different regions while improving the performance of the prediction models. The results of this study provide the first reference points for the discussion of the soil salinity monitoring method. The finding of this study can help decision-makers to build appropriate strategies to develop agriculture.

Material and methods

Study area

Ben Tre province is one of the 13 provinces of the Vietnamese Mekong Delta, covered by a natural surface of 2359.8 km^2. As of 2021, the province's population is estimated to be around 1.34 million people. The average altitude of the Ben Tre province varies from 1 to 3 m above sea level. The study area is positioned in the tropical monsoon climate, with two main seasons: the rainy season from May to November and the dry season from December to April of the following year. The annual, on average is about 1500 mm, with about 93% of the precipitation concentrated in the rainy season. The drainage network of Ben Tre province is very dense, with four major rivers: My Tho, Ba Lai, Ham Luong, and Co Chien rivers. These rivers are all part of the Mekong system. The tides in Ben Tre province are mixed diurnal and semi-diurnal tides, that is, there are two troughs and two peaks in 1 day.

Ben Tre province has rich soil resources, with many types of soils such as sandy soils, alluvial soils, and alkaline soils. This is considered one of the favorable conditions for the agricultural development of the province. The study area is regarded as the giant breadbasket in the Mekong Delta. However, in recent years, the study area has been severely affected by saline intrusion, affecting agricultural development and food security. According to Vietnam's Ministry of Natural Resources and Environment, over the next 30 years, the average total rainfall in the Mekong Delta in general and Ben Tre province in particular will decrease by 10–20%. The sea level will rise about 33 cm in 2050 and 1 m in 2100. This has a severe effect on the salinity situation of the study area. Therefore, soil salinity assessment plays a vital role in supporting decision-makers to build appropriate strategies for the development of agriculture (Fig. 1).

Data used

Soil sample collection and preparation

A field survey was established to gather soil samples over the Ben Tre province, in which 150 soil samples were collected from 0 to 30 cm depths under the land surface to identify the electric conductivity (EC) value. Various soil-sampling strategies were proposed to enhance the quality of measurement such as zigzag, grid, or transect (Nosrati and Collins 2019). However, these methods were often used in small areas, which were not required high cost (TILSE 2022). In this research, the position of soil

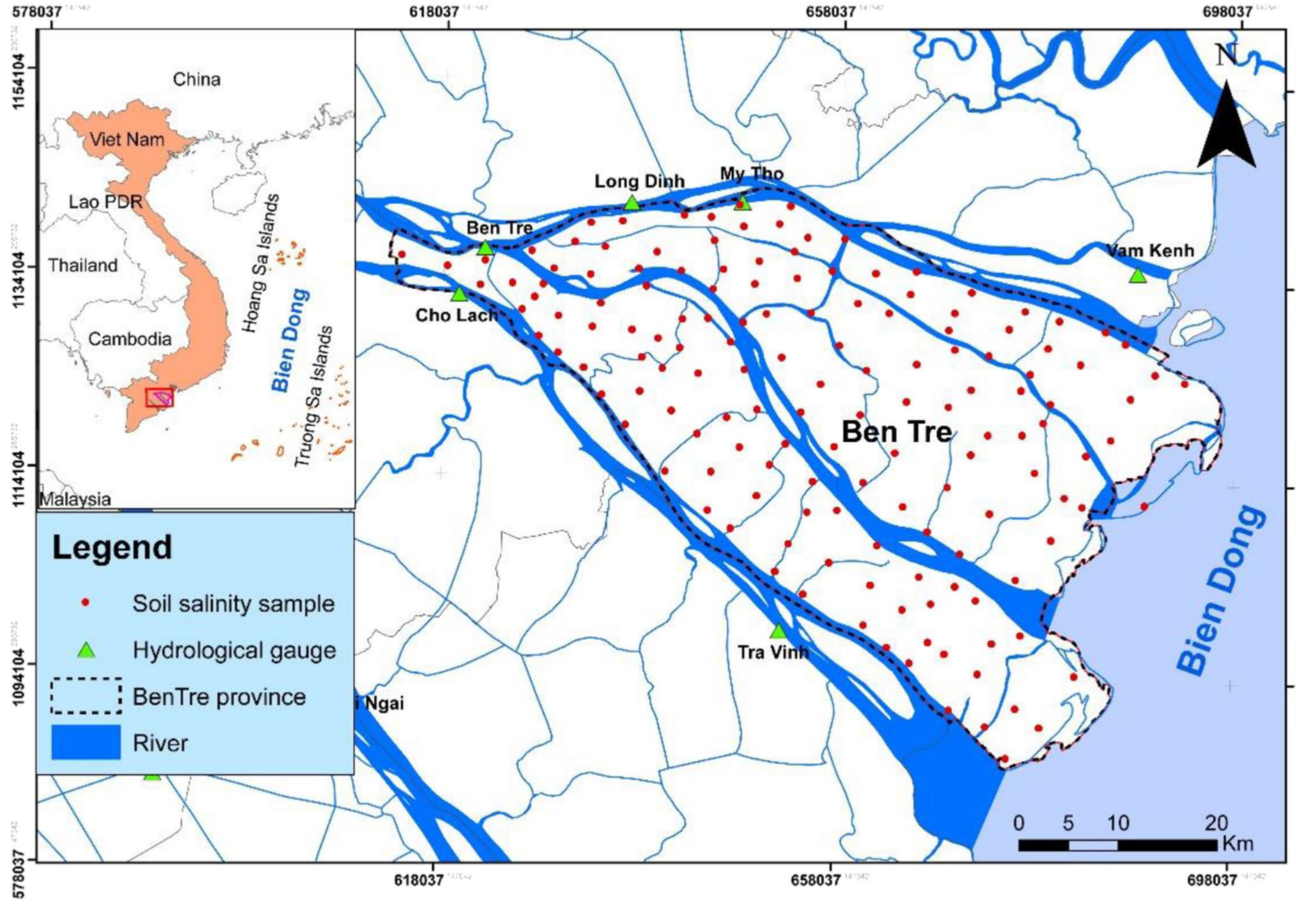

Fig. 1 Location of the Ben Tre province in Vietnam

samples was randomly selected over the entire study area in different soil types to reflect the adequacy of data. The samples were collected according to the Vietnamese Guidance on Sampling Techniques (TCVN 7538-2:2005, ISO 10381-2:2002) and analyzed in the laboratory environment. Besides, the four plus one method was applied for each soil sample to ensure data stability. It means that four subsamples were acquired around one main sample. After taking soil samples, a testing method was used to quantify the EC value by measuring the electrical resistance of one soil and five water (1:5) suspensions (Sándor et al. 2020). The final EC value will be averaged of measured results from the primary and subsamples. In this study, soil samples were accumulated from 2016 to 2021 periods. In addition, all soil samples were collected under dry weather condition without cloud cover at the location of soil samples to combine with satellite imagery data for estimating the process of models.

Soil salinity geodatabase

After collecting the soil sample data, the soil salinity geodatabase was established before training the model. We ordered the Sentinel 2A remote sensing data from January 2020 to June 2021 by using the Google Earth Engine platform (https://earthengine.google.com/). The data was pre-processed (noise removal, radiometric correction, geometric precision correction, cloud removal, and image stitching processes) before extracting 28 indices including BI, GDVI, Int1, Int2, MSI, NDDI, NDI, NDII, NDVI, NDWI, NMDI, RVI, S1, S2, S3, S5, S6, SI, SI1, SI2, SI3, SI4, VSDI, SMMI, CRSI1, EVI1, MSAVI, and SAVI1. These indices are very significant in monitoring the salinized soil by detecting the changes of minerals in the soil via spectral properties (Kılıc et al. 2022). BI (brightness index) works as a representation for observing the changes in soil organic matter, sands, and salinity areas over time (Yahiaoui et al. 2015). The vegetation indices including NDVI, GDVI, RVI, EVI1, SAVI1, and MSAVI allow to identifying of soil-induced influences on the land vegetation based on the moisture differences, roughness variations, shadows, or organic matter differences (Guo et al. 2019; Wei et al. 2021; Zhu et al. 2021). NDWI helps to discriminate waterlogged areas from soil and vegetation areas (Nguyen et al. 2021b). The group of salinity indices including CRSI1, NDSI, SI, SI1, SI2, SI3, and SI4 shows the correlation between soil reflectance at various spectral bands and the salinity parameters, which were used in a lot of previous salinity monitoring studies (Gorji et al. 2020; Halder et al. 2022). On the other hand, Int1, Int2, MSI, NDDI, NDI, NDII, VSDI, and SMMI also allow for the recognition of soil properties, which directly relate to the ability of soil to salinize. Besides, 3 indices including DEM, D2S (distance to sea), and D2R (distance to river) were also obtained and calculated to import into the soil geodatabase. It is related to the process of saltwater intrusion into the soil. The D2S and D2R indices used the Euclid distance algorithm. Every index in the soil geodatabase used the same projection system.

Machine learning approach

Prediction of soil salinity in Ben Tre province was divided into four main steps: (i) data collection and preparation, (ii) construction of soil salinity model; (iii) model validation; and (iv) analysis of the soil salinity map (Fig. 2).

(i) Data collection and preparation: the data in this study were separated into two groups: 150 soil salinity samples collected from several field missions from 2016 to 2021 and 43 independent variables calculated from the Sentinel 2A image from 2021. To increase consistency and reduce data redundancy, 43 independent variables were normalized over the range from 0 to 1 by applying min-max normalization according to Eq. 1.

$$X_{\text{Normalized}=\frac{(x_i-\min)}{(\max-\min)}} \tag{1}$$

Normalization generates new values that maintain the same general distribution and ratios as the source data while applying the same scale to the values of the different numeric columns used in the model. Several normalization techniques, namely, min-max, Z-score, logistic, and logNormal, were used in this. However, the models were more accurate with the min-max normalization technique.

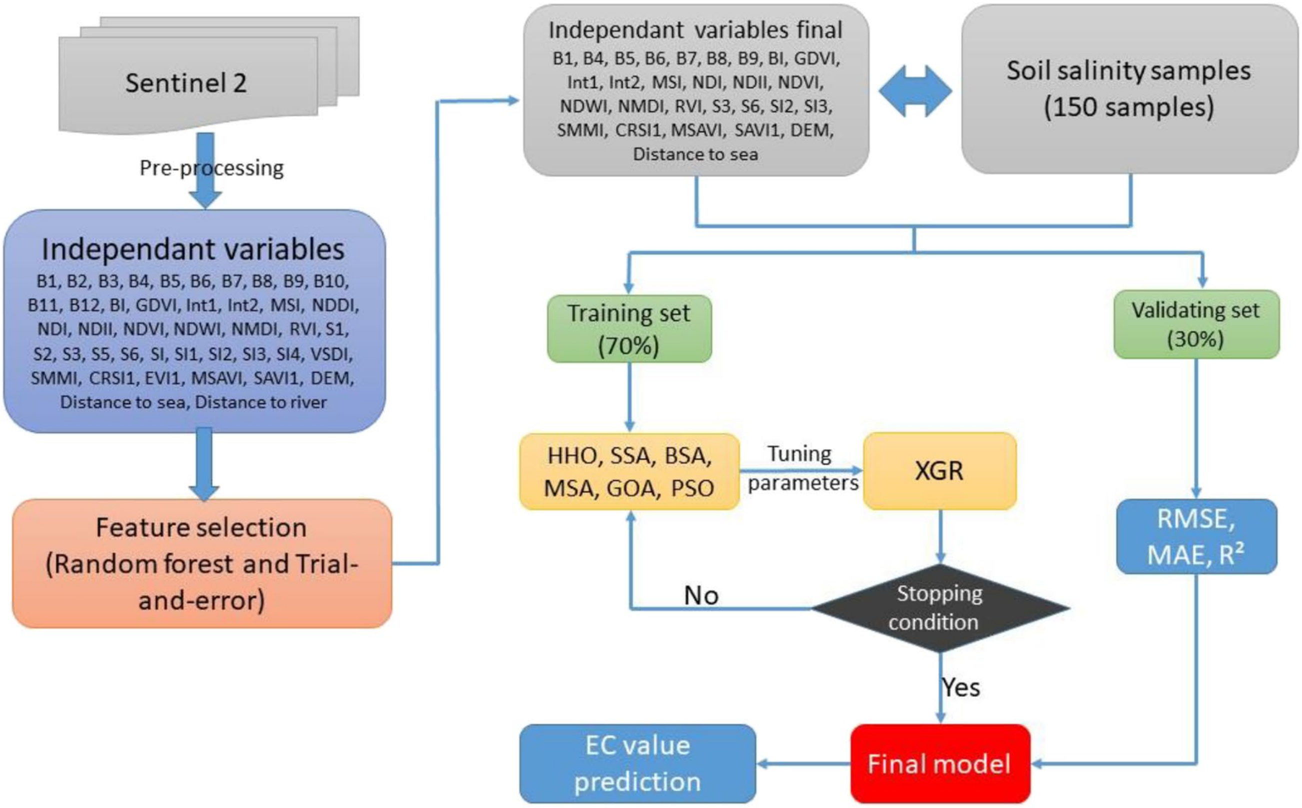
Sentinel 2
Pre-processing
Independant variables
B1, B2, B3, B4, B5, B6, B7, B8, B9, B10, B11, B12, BI, GDVI, Int1, Int2, MSI, NDDI, NDI, NDII, NDVI, NDWI, NMDI, RVI, S1, S2, S3, S5, S6, SI, SI1, SI2, SI3, SI4, VSDI, SMMI, CRSI1, EVI1, MSAVI, SAVI1, DEM, Distance to sea, Distance to river
Feature selection (Random forest and Trial-and-error)
Independant variables final
B1, B4, B5, B6, B7, B8, B9, BI, GDVI, Int1, Int2, MSI, NDI, NDII, NDVI, NDWI, NMDI, RVI, S3, S6, SI2, SI3, SMMI, CRSI1, MSAVI, SAVI1, DEM, Distance to sea
Soil salinity samples (150 samples)
Training set (70%)
Validating set (30%)
HHO, SSA, BSA, MSA, GOA, PSO
Tuning parameters
XGR
RMSE, MAE, R²
Stopping condition
No
Yes
EC value prediction
Final model

Fig. 2 The methodology used for the soil salinity prediction

Finally, 70% of the total soil samples will be used as training samples, while 30% of the remaining will be used for testing and validating the models.

(ii) Model construction: the selection of appropriate independent variables plays an important role in deciding the accuracy of the models because data redundancy can increase model calculation times and reduce the accuracy of regression models. Initially, 43 independent variables were chosen from topography, salinity index, vegetation index, and spectral index. The random forest technique combined with the trial-and-error method was used to select the important variables. In the end, 28 selected independent variables combined with 150 soil salinity samples were used to build nine models (XGB, XGB-SSA, XGB-BSA, XGB-MSA, XGB-HHO, XGB-GOA, XGB-PSO, RF, and CatBoost). Among these data, 70% were used to train the proposed models, while 30% to evaluate the models.

In this study, XGB models were coded using the Python platform with the TensorFlow library. The default parameters of the XGB model were used, except for the parameters n_estimator = 500, learning_rate = 0.3, subsample = 1, and colsample_bytree = 1. The XGB models were integrated with six self-coded optimization algorithms. The parameters for these models are the same (problem_size = 3, batch_size = 25, epoch = 500, pop_size = 50, “fit_func”: fun_avr2), except for the following specific parameters:

For SSA : “lb”: [0] * problem_size, “ub”: [1] * problem_size, ST = 0.8, PD = 0.2, SD = 0.1.

For BSA : “lb”: [0] * problem_size, “ub”: [1] * problem_size, ff = 10, pff = 0.8, c_couples = (1.5, 1.5), a_couples = (1.0, 1.0), fl = 0.5.

For HHO : “lb”: [0] * problem_size, “ub”: [1] * problem_size, n_best = 5, partition = 0.5, max_step_size = 1.0.

For GOA : “lb”: [0] * problem_size, “ub”: [1] * problem_size, c_minmax = (0.00004, 1).

For PSO : “lb”: [0] * problem_size, “ub”: [1] * problem_size, c1 = 2.05, c2 = 2.05, w_min = 0.4, w_max = 0.9.

The values of “ub” and “ib” are the upper and lower bounds of the parameters, which affect the speed of convergence of the models. Usually, these values range from −2 to 2. In this study, trial and error was performed, and the results showed that the values “ib” = 0 and “ub” = 0 give the best results to the model. The “ST,” “PD,” and “SD” parameters correspond respectively to the safety threshold, the number of producers (in percentage), and the number of sparrows that perceive the danger. The parameters “ff,” “pff,” “c_couples,” “a_couples,” and “fl” correspond to the frequency of light, the probability of seeking food, the coefficient of cognitive acceleration, the coefficient of social acceleration, the indirect and direct effect on the alertness behaviors of birds, and the tracking coefficient. The parameters “n_best,” “partition,” and “max_step_size” correspond respectively to the number of best butterflies to keep from one generation to the next, to the proportion of the first partition, and to the maximum step size used in the Levy technique flight. The parameters “c1” and “c2” correspond respectively to the local coefficient and the global coefficient, while “w_min” and “w_max” correspond respectively to the minimum and maximum weight of the bird.

(iii) In this study, several statistical indices, namely, RMSE, MAE, and R^2 were applied to assess the accuracy of soil salinity models.

(iv) After evaluating the proposed models, these models were utilized to generate the soil salinity map in Ben Tre province. This map was created by computing the EC value of each pixel.

Xgboost

Extreme gradient boosting regression (XGR) is a variant of the gradient boosting approach that is primarily used to solve complex regression and classification problems (Sahin 2020). This algorithm was developed based on the boosting framework proposed by Freund and Schapire (1996). XGR is open source and is supported by an end-to-end tree enhancement system that can be extended. XGR generates a new model that can predict the residuals of the previous model, and the results are generated by summing them. This algorithm uses gradient descent to reduce the loss when adding new models. XGR can work well with multi-dimensional data; it is considered one of the important advantages when solving big data problems. The XGR model was used in a variety of studies to estimate susceptibility to natural disasters (Pradhan and Kim 2020; Sahin 2020). The use of the XGBoost model in the previously mentioned research works results in very high accuracy of the final results. XGR strength is that it can reduce processing time by obtaining the optimal number of boosting iterations in a single run (Costache et al. 2022). Furthermore, it is well known that the XGR can improve modeling accuracy by reducing overfitting (Samat et al. 2020), another factor in the decision to use this model for this research.

Sparrow search algorithm (SSA)

SSA is one of the herd optimization algorithms, proposed by Xue and Shen (2020). The foraging and food-holding behaviors of sparrows inspire SSO. The SSA works in 3 main stages (Ouyang et al. 2021; Xue and Shen 2020): the discovery phase, the monitoring phase, and the investigation phase. During the discovery phase, the sparrows discover food and orient the other sparrows in the group. Therefore, about 20% of the sparrows in the flock perform this phase. Meanwhile, the tracking phase is where the sparrow searches for food around its location during the discovery mission. During the survey, sparrows were randomly selected from the flock. They are responsible for sending signals to the swarm to a safe place when predators enter. SSA has the ability for rapid convergence, and it has been successfully applied in several fields. Most details of SSA are present in the study by Xue and Shen (2020). This study used SSA to optimize XGR model parameters including problem size, batch_size, epoch, and pop_size.

Bird swarm algorithm (BSA)

Meng et al. first suggested the BSA (Meng et al. 2016), which was primarily motivated by two behaviors of birds: social behavior and social interaction. It is a new global optimization algorithm inspired by the simulation of birds' foraging, vigilance, and flight behavior, with few adjustable parameters, high convergence accuracy, and strong robustness (Varol Altay and Alatas 2020). The BSA works based on five main principles (Meng et al. 2016):

Principle 1: simulates the alert and foraging behavior of birds. These are considered arbitrary decisions.

Principle 2: during foraging, birds interact with each other and update previous best practices for food and social information.

Principle 3: is the process of vigilance. Birds with high food intake move to the center of the flock. This process is influenced by competition within the flock because all birds want to move to the middle of the community.

Principle 4: Birds with the most significant food reserves become productive birds, while birds with low food reserves become birds of prey. Birds with intermediate food reserves will be selected as productive or abundant birds.

Principle 5: the productive birds are still actively foraging.

The parameters in this situation are the cognitive and social acceleration coefficients. The rules are then altered for three different fuzzy systems, employing them in ascending and descending order and in experiments with trapezoidal and Gaussian membership functions to determine which of these two types produces superior results.

This study used BSA to optimize XGR model parameters including problem_size, batch_size, epoch, and pop_size.

Moth search algorithm (MSA)

The MSA is a bio-inspired metaheuristic developed by Wang (2018). The algorithm simulates phototaxis and Levy butterfly flights (Abd Elaziz et al. 2019). MSA is based on two flight characteristics of moths: levy flight, which is considered an exploitative process, and vertical flight, which is considered exploratory. The entire moth population is separated into two parts: subpopulation 1 and subpopulation 2. This division is based on the physical condition of the moths; however, females are often in the population. Subpopulation 1 and subpopulation 2 will be used to update its position using the Levy and vertical flight transitions (Feng and Wang 2022).

(i) Levy flight: stage 1 individuals will fly around the best individual in the Levy flight, and the positions of the individuals are constantly updated.

(ii) The flight straightly: for each individual in subpopulation 2, they tend to fly towards the light source.

In this study, MSA was used to optimize XGR model parameters including problem_size, batch_size, epoch, and pop_size.

Harris hawks optimization (HHO)

HHO is a meta-heuristic optimization algorithm, proposed by Heidari et al. (2019). This algorithm simulates the hunting behavior of the Harris hawk. Harris hawks hunt by using surprise attacks, and they can execute various hunting strategies. The HHO algorithm works on two main processes: exploration and exploitation. During exploration, Harris's hawk uses his polar eyes to track and detect prey. However, in many cases, hawks cannot catch prey easily. Thus, hawks can wait and observe their target for several hours. In HHO, the Harris hawk is considered the candidate solution and the best candidate solution in the loop is considered the intended prey.

The exploration process is transferred to the exploitation process based on the escape energy of the prey. The energy of the prey is reduced considerably when fleeing. In the exploitation process, a hawk with a surprise pounce attacked the prey identified during the exploration phase. In reality, however, prey tends to get out of dangerous situations. Therefore, different chasing strategies of Harris hawks are employed to encircle prey. Hawks often implement hard and soft siege strategies in practice. That is, Harris hawks will surround their prey gently from all directions or firmly, depending on the prey's energy. More details about the HHO algorithm are described in (Heidari et al. 2019). HHO was successful in several different fields like energy (Houssein et al. 2020) and natural hazard prediction (Tikhamarine et al. 2020).

In this study, HHO was used to optimize XGR model parameters including problem_size, batch_size, epoch, and pop_size.

Grasshopper optimization algorithm (GOA)

GOA is a meta-heuristic, which was developed by Saremi et al. (2017). The foraging behavior of grasshoppers inspires the GOA (Meraihi et al. 2021). Grasshoppers often feed in swarms and are based on two main phases: the exploration phase and the exploitation phase. During the crawling process, they move over a wide range and move flick, while grasshoppers focus on the local searching during the crawling process. These steps aim to reduce operational weaknesses (local optimization) or increase convergence speed (Lv and Peng 2021; Nguyen et al. 2021a). For each move change, the grasshopper connects and interacts with other grasshoppers in the swarm to update the current position, best position, and factors affecting the next move (Moayedi et al. 2021).

In this study, GOA was used to optimize XGR model parameters including problem_size, batch_size, epoch, and pop_size.

Particle swarm optimization (PSO)

PSO is a metaheuristic, that was proposed by Kennedy and Eberhart (1995). PSO is inspired by the movement to optimize food pathways for flocks of birds and fish (Kennedy and Eberhart 1995). In the PSO algorithm, birds and fish are considered as the elements that always find the best solution to solve the problems (optimization of the foraging path) called particles. These problems are solved in an *n*-dimensional space, and *n* represents different algorithm parameters (Band et al. 2020). To optimize foraging pathways, PSO optimizes the position and velocity of each particle. The position and velocity of each particle change after each iteration; they are continuously updated and are represented by an objective function and a direction vector (Bui et al. 2020; Nguyen et al. 2021a). After each loop, the position and velocity of each particle are plotted and calculated according to the following formula:

$x_i^t = \left(x_{i1}^t, x_{i2}^t \ldots x_{in}^t\right)$ and $v_i^t = \left(v_{i1}^t, v_{i2}^t \ldots v_{in}^t\right)$ is the position and velocity of particle *i* at time *t*. Therefore, the position and velocity of particle *i* at $t + 1$ are

$$v_i^{t+1} = w \times v_i^t + c_1 \times r_1 \times \left(p_i^t - x_i^t\right) + c_2 \times r_2 \times \left(g_i^t - x_i^t\right)\text{with}$$

$$- v_{\max} \leq v_i^{t+1} \leq v_{\max} x_i^{t+1} = \left(x_i^t + v_i^{t+1}\right)$$

where *xti* is the *i*-th position of the particle, *pti* is the optimal position found, *gti* is the best position of the particle, $r1$ and $r2$ are the random constants, with values ranging from 0 to 1, w is the weight d inertia, and $c1$ and $c2$ are the social coefficients.

In this study, PSO was used to optimize XGR model parameters including problem_size, batch_size, epoch, and pop_size.

Random forest

RF is considered one of the popular algorithms and is first proposed by Breiman (2001). This algorithm combines between bagging ensemble and random subspace, which bases decision tree prediction to solve classification and regression problems (Quiroz et al. 2018). RF combines all the submodels to obtain results with higher prediction than the individual models. Each submodel was evaluated using majority voting to determine the best model (Nguyen 2022a; Tajik et al. 2020; Zeraatpisheh et al. 2019). RF function to be based on four main steps: (i) resamples from the original dataset using Bootstrap; (ii) the use of subsets to construct the decision tree in the forest; (iii) obtaining the final results by combining the prediction results of all the decision trees; and (iv) select the best result using the majority vote (Chen et al. 2020a). One of the important advantages of RF is that it can solve missing data problems by using the average value of neighboring samples. RF model accuracy can be influenced by the number of decision trees (Ntree) and candidate characteristics in subsets (mtry) (Horning 2010). In this study, several values of Ntree were tried like 100, 200, and 500. The accuracy of RF is higher with Ntree = 500.

CatBoost

CatBoost is considered one of the power machine learning algorithms, first proposed by Dorogush et al. (2018). This algorithm uses the permutation technique to predict classification and regression problems. In the training process, the dataset was randomly permuted and labeled to avoid overfitting and underfitting issues. CatBoost works based on three main steps: (i) dividing the dataset into subsets randomly; (ii) labeling for subsets and converting them to integers; and (iii) encoding the features (Hai Ly et al. 2022). CatBoost implements symmetric trees, which reduce prediction time. It takes advantage of random permutations to have a random parameter. CatBoost handles classification features using concepts such as ordered boosting and response coding (Saber et al. 2021).

Performance assessment

Assessing the model's accuracy is critical for establishing the soil salinity model. Several statistical indices were applied to assess the precision of the proposed models, RMSE, MAE, and R^2.

RMSE is considered a popular index to compute the standard deviation between observation values and prediction values. While MAE calculates the average error between the prediction and observation value. The value of RMSE and MAE varies from 0 to 1. The closer the RMSE and MAE value is to 1, the more accurate the model (Eldeiry and Garcia 2008; Hu et al. 2019; Yan et al. 2007). They are calculated by the following equation:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - Y_j\right)^2}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} | Y_i - Y_j |$$

Yi is the prediction value at sample *i* and *Yj* is the observation value at sample *j*. *N* is the number of samples.

R^2 is an index to quantify the relationships between independent and dependent variables in the regression model. It presents the fitness of the data and identifies the percentages of the fit results in the regression model. For example, the R^2 value of 50% shows that 50% of the data is fit in the model. The higher the R^2 value, the better the fit of the data in the model (Scudiero et al. 2014).

Results

Data pre-processing and exploitation

The selection of factors plays an indispensable role in the process of building a soil salinity model. Because data redundancy confuses the predictive model, which reduces model performance, the importance of 43 independent variables was assessed using the random forest (RF) technique. The results show that Int2 (0.34), BI (0.33), SAVI (0.32), NDVI (0.31), and Band 8 (0.3) are the most important variables on soil salinity in Ben Tre province, followed by MSAVI (0.29), RVI (0.27), NDI (0.26), distance to sea (0.25), GDVI (0.25), SI2 (0.25), SRSI (0.244), NMDI (0.24), NDII (0.23), Band 9 (0.23), SMMI (0.22), Band 1 (0.21), Band 7 (0.21), NDWI (0.21), S6 (0.2), Band 6 (0.2), Band 4 (0.19), S3 (0.19), Band 5 (0.18), SI3 (0.16), and Int1 (0.16). Two factors SI4 (0) and SI1 (0.05) are a minor influence on soil salinity. Moreover, in this study, the trial-and-error method was used to select the independent variables. Based on Figure 3, 20, 21, 22…40 independent variables were tried to use as model input data, and the models are more precise with 28 first independent variables. The 15 variable remainders were eliminated for the soil salinity models.

In general, indices related to soil, vegetation, and brightness rates extracted from remote sensing data significantly affect the proposed models. The reasons associated with the nature of the study area, vegetation, and soil characteristics are the essential environmental variables for the rate of soil salinity. At the same time, the indices related to drought and the distance to the river had a minor influence on soil salinity in the Ben Tre province of Vietnam because the salinity characteristics of the soil are different from the drought characteristics in the study area. In addition, during the dry season, the canal/rift system in the study area does not have enough fresh water upstream, combined with solid tidal propagation leading to increased saline intrusion on the land. This is why the distance to the river is less important than the distance to the variable in the salt intrusion problem in Ben Tre province.

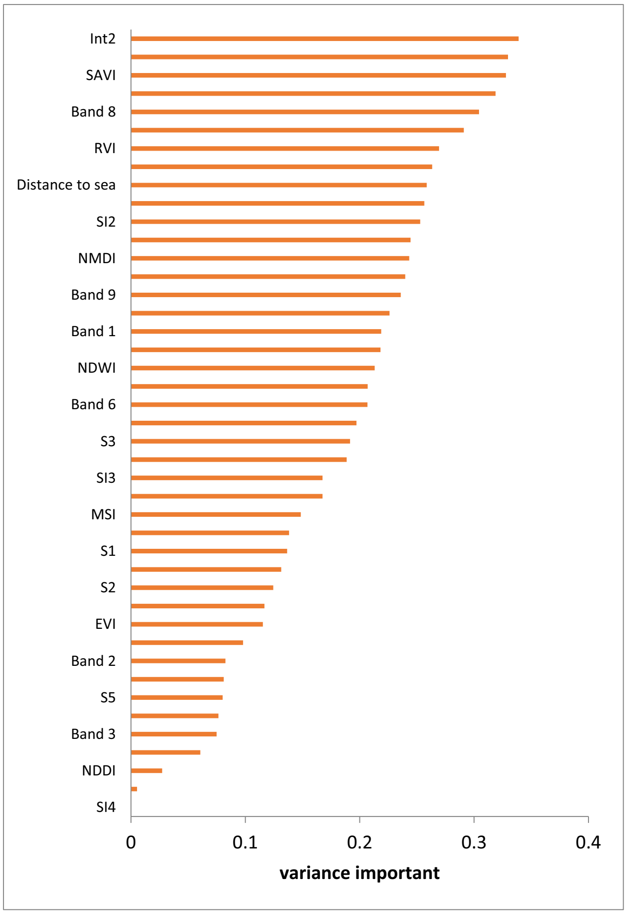

Fig. 3 The importance of the variable conditioning

Model performance comparison

Figure 4 shows the R^2 value for validation data for nine proposed models (XGR, XGR-SSA, XGR-MSA, XGR-BSA, XGR-HHO, XGR-GOA, XGR-PSO, RF, and CatBoost). It was found that all of the proposed models fit well to predict soil salinity. The proposed algorithms improve the precision of the XGR model and surpass the performance of the reference models, namely, RF and CatBoost. Among the proposed models, the XGR-HHO model is better than the other models with an R^2 value of 0.932, followed by XGR-GOA ($R^2 = 0.931$), XGR-MSA ($R^2 = 0.928$), XGR-BSA ($R^2 = 0.926$), XGR-SSA ($R^2 = 0.917$), XGR-PSO ($R^2 = 0.916$), XGR (R^2 = 0.867), CatBoost (R^2 = 078), and RF (R^2 = 0.75), respectively.

Fig. 4 R^2 value for the validation dataset

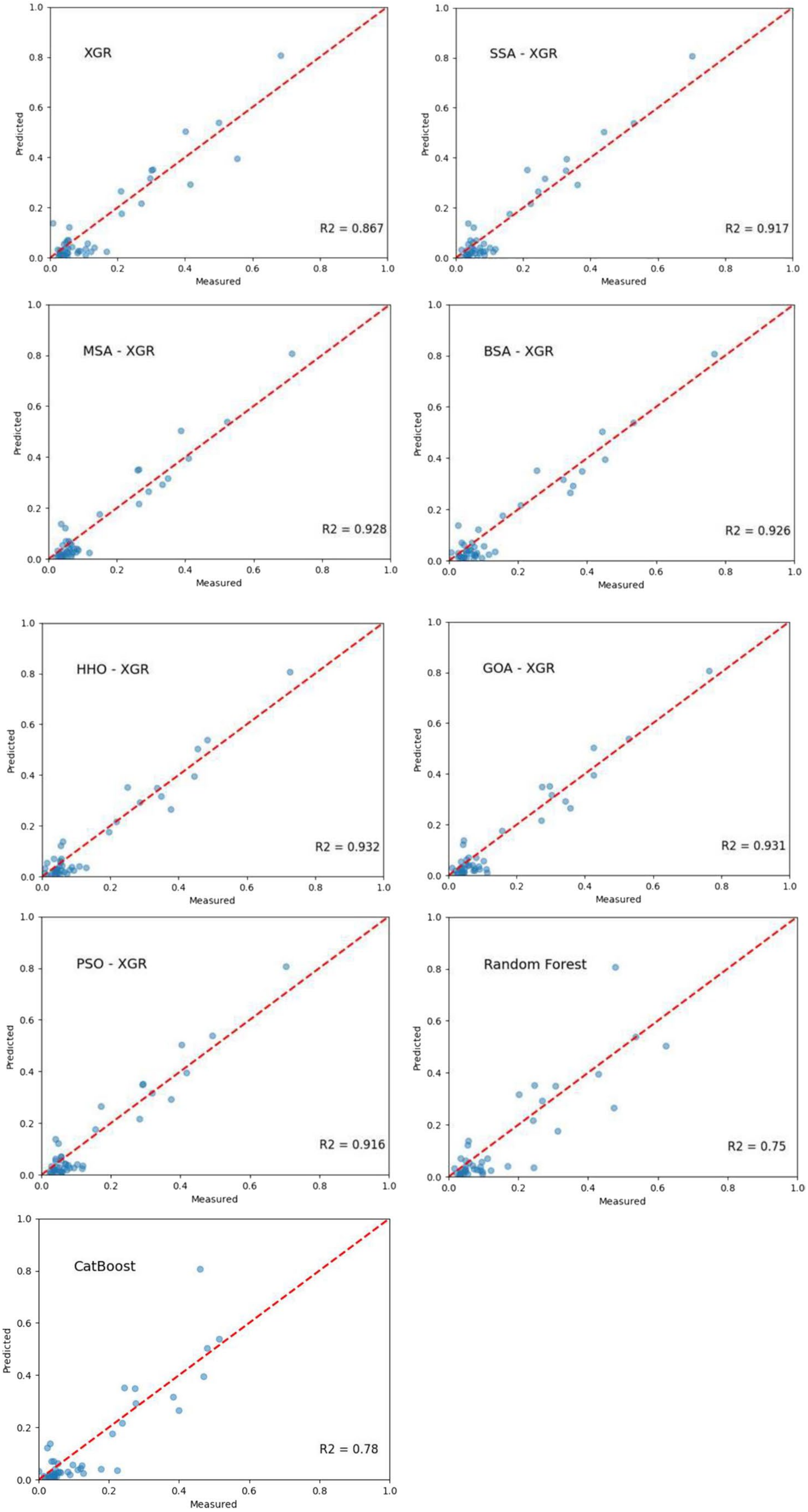

Table 1 Model performance and comparison

	Training dataset			Validating dataset		
	RMSE	MAE	R^2	RMSE	MAE	R^2
XGR	0.12	0.1	0.98	0.1	0.09	0.867
XGR-SSA	0.08	0.07	0.98	0.07	0.06	0.917
XGR-MSA	0.068	0.052	0.99	0.06	0.047	0.928
XGR-BSA	0.072	0.055	0.99	0.062	0.05	0.926
XGR-HHO	0.04	0.03	0.99	0.051	0.043	0.932
XGR-PSO	0.09	0.08	0.99	0.08	0.08	0.916
XGR-GOA	0.05	0.04	0.99	0.055	0.049	0.931
RF	0.21	0.2	0.84	0.19	0.19	0.75
CatBoost	0.15	0.11	0.91	0.12	0.1	0.78

Besides the R^2 index, several indices like RMSE and MAE were utilized to assess the precision of the proposed models. Table 1 shows the value of RMSE and MAE for the proposed studies. The results show that the XGR-HHO model outperformed the other models with the value of RMSE and MAE of 0.051 and 0.042, followed by XGR-GOA (RMSE = 0.055, MAE = 0.049), XGR-MSA (RMSE = 0.06, MAE = 0.047), XGR-BSA (RMSE = 0.062, MAE = 0.05), XGR-SSA (RMSE = 0.07, MAE = 0.06), XGR-PSO (RMSE = 0.08, MAE = 0.08), XGR (RMSE = 0.1, MAE = 0.09), CatBoost (RMSE = 0.12, MAE = 0.1), and RF (RMSE = 0.19, MAE = 0.19), respectively.

In general, six optimization algorithms (SSA, MSA, BSA, HHO, PSO, and GOA) in this study improved the performance of individual XGR models, and the performance of hybrid models outperformed individual models (RF and CatBoost). Because hybrid models can eliminate the weak points of individual models and from that can improve the performance of individual models.

Soil salinity mapping

After validating the proposed models, these models were used to build the soil salinity map in the Ben Tre province. These maps were generated by feeding the models of all the study areas with conditioning factor values. Figure 5 shows the soil salinity maps produced by XGR, XGR-SSA, XGR-MSA, XGR-HHO, XGR-BSA, XGR-GOA, and XGR-PSO. In the field mission in 2021, it can be seen that the western part of the study area is mainly agricultural land and low salinity. However, the closer you get to the sea, the more the salinity phenomenon increases. This zone is primarily aquaculture and partly rice growing. By comparing the results mentioned above with the results of the forecasting models, although there are slight differences between the models, all the models follow a trend that the western regions have a low degree of severity, while the eastern regions are more saline. Among the proposed models, three models XGB-HHO, BSA, and RF have lower salinity values in the east than the remaining models.

Soil salinity in the study area varies over time and space and can be affected by factors such as drought, sea level rise, and endogenous salinization. During the dry season, especially in March and April, when rainfall in the research area decreases, the amount of fresh water from rivers flowing into the soil decreases. This phenomenon causes an increase in seawater intrusion and salinity levels in coastal areas, making them generally more saline than other areas in Ben Tre province. Additionally, in the context of climate change and sea level rise, the canal/ditch system does not receive enough fresh water from upstream, causing strong tidal surges that lead to saltwater intrusion and accumulation. Ben Tre province is considered one of the few provinces in the Mekong Delta region that is less affected by soil salinity, except for areas where aquaculture is practiced (i.e., coastal areas).

Discussion

The region most affected by soil salinity is the Mekong Delta in general and Ben Tre in particular. This natural disaster causes significant damage to the country's economy. For example, in 2020, the Mekong Delta in Vietnam was affected by a major drought which caused soil salinization. During this natural event, salt water penetrated about 110 km inland, surpassing the 2015–2016 record by about 10 km. It was also discovered that seawater was infiltrating the Mekong River Delta from November 2019, which is 2–4 months earlier than in previous years. In February 2020, around 40,000 coastal households lacked fresh water and 20,000 ha of fruit trees as well as 6500 ha of vegetables in Ben Tre were damaged.

The construction of a dam upstream of the Mekong River in recent decades has led to water scarcity downstream, resulting in relative lowering of the riverbed downstream as a result of mining of the riverbed and a rise in sea level. Consequently, the sea level penetrated

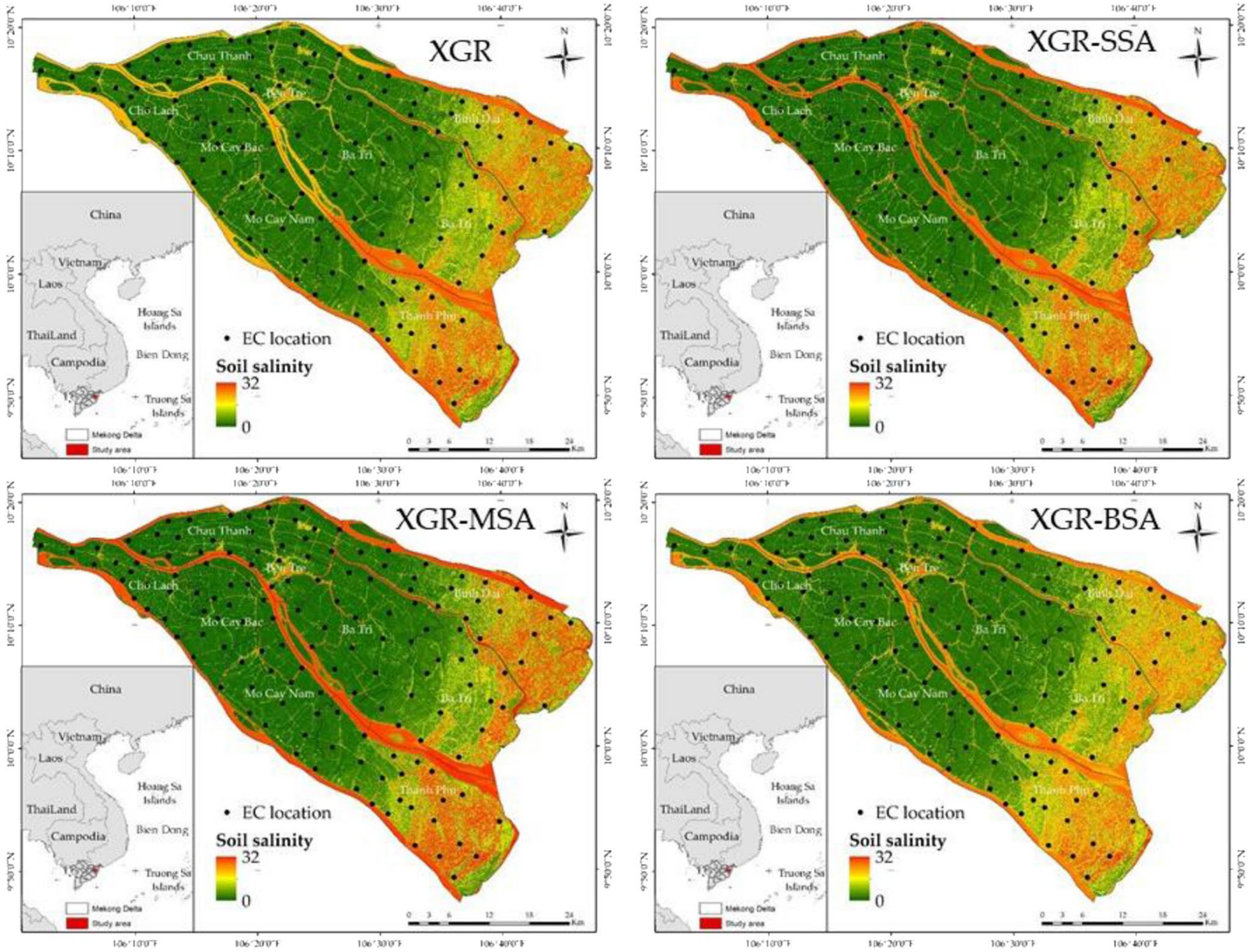

Fig. 5 Soil salinity in Ben Tre province of Vietnam produced by XGR-SSA, XGR-MSA, XGR-BSA, XGR-HHO, XGR-GOA, XGR-PSO, RF, and CatBoost

more and more rapidly into the ground as a result. All this has led to soil salinization in the Mekong Delta. This has been substantiated by previous studies. Hui et al. (2022) pointed out that the change in the hydrological regime is an important factor influencing soil quality and farmers' decision-making. Therefore, developing a state-of-the-art method based on machine learning and remote sensing to monitor soil salinity is very important to help decision-makers plan agricultural development. Remote sensing uses spectral information reflected from objects on the earth to detect helpful information related to soil salinity. Soils with different salinity levels have different spectral characteristics, which are important bases for the surveillance of soil salinity utilizing remote sensing (Aldabaa et al. 2015; Allbed and Kumar 2013). Generally, the areas covered by a white salt crust have a high EC value. However, in several cases, in each band of the Sentinel 2 image, the spectral reflectance value of the soil sample does not increase when the EC value increases. Consequently, it is difficult to map soil salinity directly using multispectral bands and the spectral index (Wang et al. 2021b; Wang et al. 2020b). According to various studies, salinity indices and vegetation indices were applied to monitor and estimate the soil salinity value (Allbed and Kumar 2013). Due to differences in geographical location, topography, and vegetation type, the value of EC under vegetation cover is very different, from low EC value to high EC value. However, in several studies, vegetation areas are considered non-saline or slightly saline areas (Fernandez-Buces et al. 2006; Metternicht and Zinck 2003). Sentinel 2 data with a 10 m resolution can solve these problems. This study did not cover the vegetated area when modeling soil salinity, but we collected additional samples from the vegetated areas to monitor and estimate soil salinity using spectral indices (Taghadosi et al. 2019; Wang et al. 2020a).

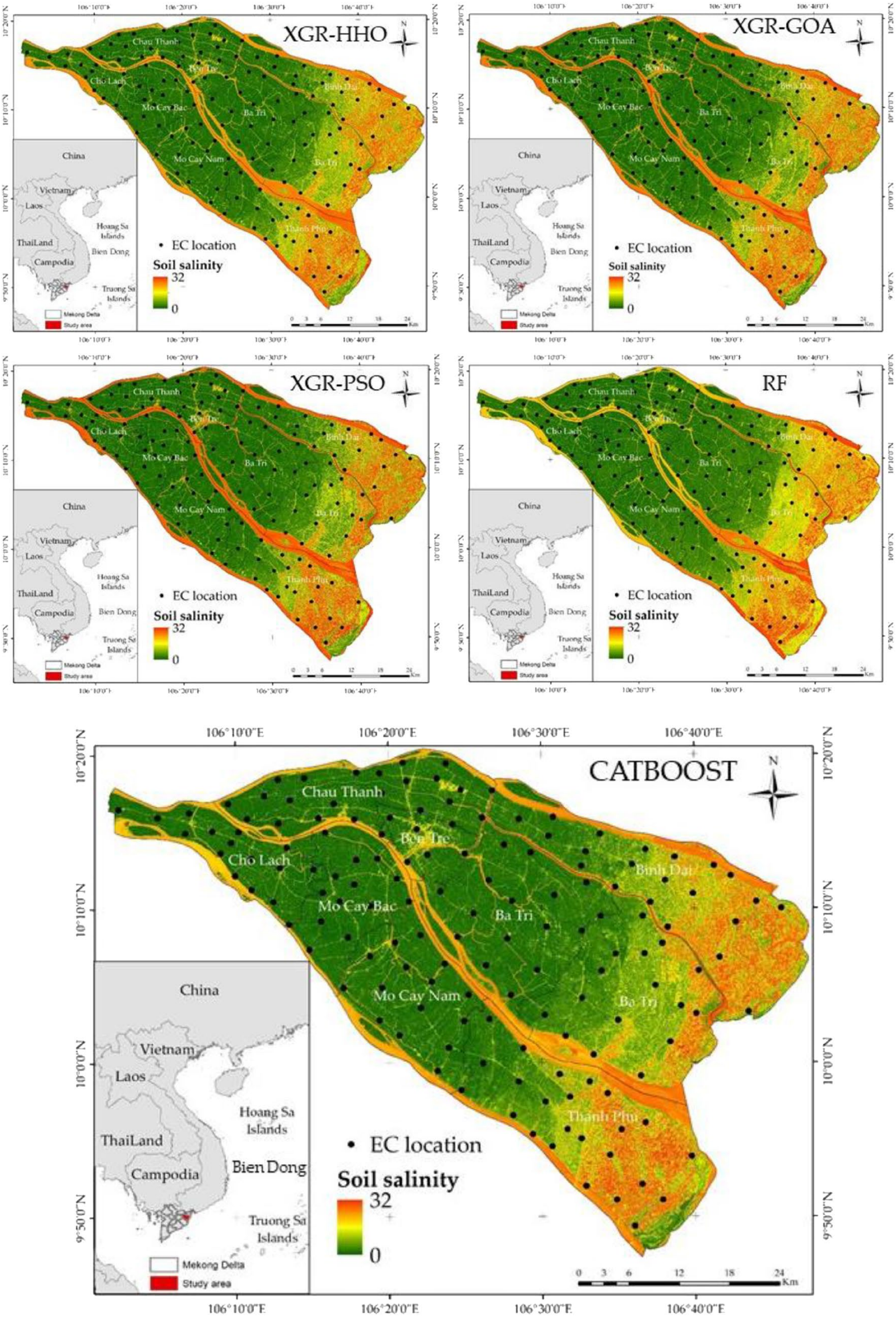

Fig. 5 (continued)

The quantification of variable importance is an important task in building a soil salinity model. Because data redundancy can increase calculation times and sometimes decrease the accuracy of predictive models (Bui et al. 2020; Nguyen 2022b; Nguyen et al. 2021b). In this study, 43 independent variables were selected to build soil salinity models. The RF technique combined with the trial-and-error method was used to determine the appropriate independent variables for the soil salinity model. The results show that the spectral indices, namely, Int2 (0.34), BI (0.33), SAVI (0.32), NDSI (0.32), NDVI (0.31), and Band 8 (0.3) play important roles in monitoring soil salinity. Because this study uses Sentinel 2A images with a 10 m spatial resolution, they contain few mixed pixels, so areas with high spectral reflectance have high EC. In addition, the salinity index and geographical location also play a key role in assessing soil salinity. Due to the characteristics of ecology, hydrology, climate, and human activity that determine the movement and accumulation of salt, therefore, soil salinity varies considerably depending on geographical location (Ge et al. 2022; Wang et al. 2019b; Wang et al. 2021b). In addition, the study area has low relief (< 2 m), often affected by sea level rise. Areas close to the sea have higher EC values than those areas far from the sea (Loc et al. 2021; Tho et al. 2008). Besides using the RF technique to determine the important variables, this study also uses the trial-and-error method to filter out the factors not useful for the soil salinity model. It should be noted that this study applies a data-driven approach to monitor soil salinity. Therefore, besides the dependence on the characteristics of ecology, hydrology, climate, or human activities, the statistical relationships between the independent variables and the value of EC are considered as one of the important elements that determined the importance of variables (Nguyen 2022b).

The results justified the initial assumption in the introduction that the accuracy of hybrid models has surpassed the individual models. In addition to landslides, floods, groundwater salinization, and soil salinization, these results are consistent with results from many different studies using machine learning. Hybrid models can improve the ability of individual model prediction by eliminating weak classifications. In this study, six optimization algorithms, namely, SSA, MSA, BSA, HHO, GOA, and PSO successfully improve the prediction ability of the XGR model. Among these proposed models, the XGR-HHO model was better than the other models because HHO has advantages in terms of ease of implementation and high flexibility. Also, HHO can globally search at different stages (Alabool et al. 2021). The XGR-GOA model was ranked second because the GPA has a fast optimal solution-finding speed with high accuracy and, in particular, GOA is easy to combine with unique algorithms (Wang and Li 2019). Model XGR-MSA was ranked third because, in addition to its flexibility and ease of use, the MSA also has fast convergence speed and high convergence accuracy (Han et al. 2020). The XGR-BSA model was ranked fourth because the BSA has a simple design and is easy to implement. Moreover, this algorithm can solve optimization problems in a short time with high accuracy thanks to its capability to balance the process of exploration and exploitation. The accuracy of the XGR-SSA model was ranked fifth in predicting the value of EC. However, despite this algorithm's success in improving the XGR model's performance, the XGR-HHO, XGR-GWO, XGR-MSE, and XGR-BSA models' performance was superior. This is due to this algorithm's randomness and local optimization difficulty. The XGR-PSO model was ranked sixth since although PSO has the advantage in convergence speed, one of the important disadvantages of PSO is the problem of local optimization.

Several studies have also used machine learning to monitor soil salinity. Wang et al. (2021b) used SVM, RF, and ANN with Sentinel 2 data to construct the soil salinity map. The results indicated that the SVM model was better than the other models with the R^2 value of 0.88. The accuracy of these models is less than in our study. Ge et al. (2022) applied bagging, classification and regression tree, RF, and gradient boosting regression tree (GBRT) with Sentinel 2 data to predict soil salinity in the Ebinur Lake region of China. The results reported that these models had the best accuracy with the R^2 value of 0.88. These models were less efficient than in our study. Nguyen et al. (2021b) integrated GOA with DNN, SVR, XGR, and GPR combined with Landsat 8 data to predict soil salinity in the Vietnamese Mekong Delta. The results reported that the XGR-GOA model had surpassed the other models with the R^2 value of 0.86. The performance of this model works less well than in our current study with the value of R^2 of 0.99. Wei et al. (2022) integrated the particle swarm optimization (PSO) with support vector machine regression (SVR) and back-propagation neural network (BPNN) models to predict soil salinity in Keriya Oasis of China using the PALSAR-2 data and Landsat 8 data. The results indicated that the best accuracy of these models was with the R^2 value of 0.88. So, these models perform worse than the models in our study with the R^2 value of 0.99. Therefore, the results of our research can be used as an alternative solution to support decision-makers in the development of agriculture to ensure food security in the region.

The main contribution of this study is the development of new and less expensive methods to monitor soil salinity. Although this study successfully constructed a technique to monitor soil salinity, however, it also presents general limitations related to the uses of the data. The reliability of the generated soil salinity map is partially dependent on the collection scheme and the number of in situ samples. In this study, 150 samples were obtained in 2016–2020 so the map obtained cannot capture and express the dynamics of salinization processes through time. In addition, shallow groundwater depth at every pixel, which is hardly acquired, was not included as an input variable. So, all of these can increase the uncertainties of the results.

Conclusion

Soil salinization is the most dangerous environmental hazard, significantly affecting the country's food security. Therefore, this study aims to develop a low-cost method to monitor soil salinity in the Ben Tre province of Vietnam, where often affected by soil salinity. The results of this study can support decision-makers in sustainable agricultural development to ensure food security in the country.

In Ben Tre province of Vietnam, factors related to soil characteristics, vegetation, and brightness rates are the most important variables in soil salinity.

The integration of optimization algorithms (HHO, SSA, MSA, BSA, GWO, and PSO) on the XGR model can significantly improve the precision of the XGR model and the performance of the hybrid models has surpassed reference models such as CatBoost and RF. The new models can be used to monitor soil salinity in any region, especially in the data-limited area.

The soil in the study area tends to be highly saline in the eastern regions, while the western regions tend to be slightly salty. The findings in this study can be important tools for building effective soil salinity management strategies in any area of the world. The methods proposed in this study can be developed to assess other natural hazards, such as flooding and groundwater salinization.

The soil salinity in the Mekong Delta and the study area in particular is strongly influenced by sea level rise and climate change. Therefore, in future research, studies on evaluating the effects of these problems on soil salinity are necessary to build appropriate agricultural development strategies to ensure food security.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11356-023-27516-x.

Authors' contribution Huu Duy Nguyen, Tien Giang Nguyen, and Quang Thanh Bui: conceptualization and methodology. Huu Duy Nguyen, Tien Giang Nguyen, Quang Thanh Bui, Dinh Kha Dang, Thi Thuy Nga Pham, Van Chien Nguyen, and Quoc Huy Nguyen: methodology, material preparation, validation, analysis, writing the original draft, reviewing and editing, and writing the review and editing. Huu Duy Nguyen and Quoc Huy Nguyen: data collection. All authors read and approved the final manuscript.

Data availability The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

Abd Elaziz M, Xiong S, Jayasena K, Li L (2019) Task scheduling in cloud computing based on hybrid moth search algorithm and differential evolution. Knowl-Based Syst 169:39–52

Alabool H, Al- Arabiat D, Abualigah L, Heidari AA (2021) Harris hawks optimization: a comprehensive review of recent variants and applications. Neural Comput & Applic 33:8939–8980

Aldabaa AAA, Weindorf DC, Chakraborty S, Sharma A, Li B (2015) Combination of proximal and remote sensing methods for rapid soil salinity quantification. Geoderma 239:34–46

Allbed A, Kumar L (2013) Soil salinity mapping and monitoring in arid and semi-arid regions using remote sensing technology: a review. Adv Remote Sens 2(4):373–385

Band SS, Janizadeh S, Chandra Pal S, Saha A, Chakrabortty R, Shokri M, Mosavi A (2020) Novel ensemble approach of deep learning neural network (DLNN) model and particle swarm optimization (PSO) algorithm for prediction of gully erosion susceptibility. Sensors 20:5609

Breiman L (2001) Random forests. Mach Learn 45:5–32

Bui Q-T, Nguyen Q-H, Nguyen XL, Pham VD, Nguyen HD, 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

Calixto WP, Neto LM, Wu M, Kliemann HJ, de Castro SS, Yamanaka K (2010) Calculation of soil electrical conductivity using a genetic algorithm. Comput Electron Agric 71:1–6

Chen Y, Qiu Y, Zhang Z, Zhang J, Chen C, Han J, Liu D (2020b) Estimating salt content of vegetated soil at different depths with Sentinel-2 data. PeerJ 8:e10585

Corwin DL (2021) Climate change impacts on soil salinity in agricultural areas. Eur J Soil Sci 72:842–862

Costache R, Arabameri A, Moayedi H, Pham QB, Santosh M, Nguyen H, Pandey M, Pham BT (2022) Flash-flood potential index estimation using fuzzy logic combined with deep learning neural network, naïve Bayes, XGBoost and classification and regression tree. Geocarto Int 37:6780–6807

Cui F, Al-Sudani ZA, Hassan GS, Afan HA, Ahammed SJ, Yaseen ZM (2022) Boosted artificial intelligence model using improved alpha-guided grey wolf optimizer for groundwater level prediction: comparative study and insight for federated learning technology. J Hydrol 606:127384

Dai X, Huo Z, Wang H (2011) Simulation for response of crop yield to soil moisture and salinity with artificial neural network. Field Crop Res 121:441–449

Ding W, Nguyen MD, Mohammed AS, Armaghani DJ, Hasanipanah M, Van Bui L, Pham BT (2021) A new development of ANFIS-Based Henry gas solubility optimization technique for prediction of soil shear strength. Transport Geotechnics 29:100579

Dorogush, A.V., Ershov, V., Gulin, A., 2018. CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363.

Eldeiry A, Garcia LA (2008) Detecting soil salinity in alfalfa fields using spatial modeling and remote sensing. Soil Sci Soc Am J 72:201–211

Fathizad H, Ardakani MAH, 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

Feng Y, Wang G-G (2022) A binary moth search algorithm based on self-learning for multidimensional knapsack problems. Futur Gener Comput Syst 126:48–64

Fernandez-Buces N, Siebe C, Cram S, Palacio J (2006) Mapping soil salinity using a combined spectral response index for bare soil and vegetation: a case study in the former lake Texcoco, Mexico. J Arid Environ 65:644–667

Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. Citeseer 96:148–156

Ge X, Ding J, Teng D, Wang J, Huo T, Jin X, Wang J, He B, Han L (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

Gorji T, Sertel E, Tanik A (2017) Monitoring soil salinity via remote sensing technology under data scarce conditions: a case study from Turkey. Ecol Indic 74:384–391

Gorji T, Tanik A, Sertel E (2015) Soil salinity prediction, monitoring and mapping using modern technologies. Procedia Earth and Planetary Science 15:507–512

Gorji T, Yildirim A, Hamzehpour N, Tanik A, Sertel E (2020) Soil salinity analysis of Urmia Lake Basin using Landsat-8 OLI and Sentinel-2A based spectral indices and electrical conductivity measurements. Ecol Indic 112:106173

Guo B, Yang F, Fan Y, Han B, Chen S, Yang W (2019) Dynamic monitoring of soil salinization in Yellow River Delta utilizing MSAVI–SI feature space models with Landsat images. Environ Earth Sci 78:1–10

Hai Ly N, Nguyen HD, Loubiere P, Van Tran T, Șerban G, Zelenakova M, Brețcan P, Laffly D (2022) The composition of time-series images and using the technique SMOTE ENN for balancing datasets in land use/cover mapping. Acta Montan Slovaca 27:2

Halder B, Bandyopadhyay J, Islam M (2022) Climate change impact on soil salinity dynamics at the gosaba cd block in india by integrating geospatial indicators and regression techniques. In: Climate change impacts, mitigation and adaptation in developing countries. Springer, India II, pp 97–125

Han X, Yue L, Dong Y, Xu Q, Xie G, Xu X (2020) Efficient hybrid algorithm based on moth search and fireworks algorithm for solving numerical and constrained engineering optimization problems. J Supercomput 76:9404–9429

Heidari AA, Mirjalili S, Faris H, Aljarah I, Mafarja M, Chen H (2019) Harris hawks optimization: algorithm and applications. Futur Gener Comput Syst 97:849–872

Hengl T, Heuvelink GB, Rossiter DG (2007) About regression-kriging: from equations to case studies. Comput Geosci 33:1301–1315

Horning N (2010) Random Forests: An algorithm for image classification and generation of continuous fields data sets. Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences, Osaka, Japan, pp 1–6

Houssein EH, Hosney ME, Oliva D, Mohamed WM, Hassaballah M (2020) A novel hybrid Harris hawks optimization and support vector machines for drug design and discovery. Comput Chem Eng 133:106656

Hu J, Peng J, Zhou Y, Xu D, Zhao R, Jiang Q, Fu T, Wang F, Shi Z (2019) Quantitative estimation of soil salinity using UAV-borne hyperspectral and satellite multispectral images. Remote Sens 11:736

Hua L, Zhang C, Peng T, Ji C, Nazir MS (2022) Integrated framework of extreme learning machine (ELM) based on improved atom search optimization for short-term wind speed prediction. Energy Convers Manag 252:115102

Hui TR, Park E, Loc HH, Tien PD (2022) Long-term hydrological alterations and the agricultural landscapes in the Mekong Delta: insights from remote sensing and national statistics. Environ Challenges 7:100454

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 40:284–306

Jiang X, Xue X (2022) Comparing Gaofen-5, Ground, and Huanjing-1A spectra for the monitoring of soil salinity with the BP neural network improved by particle swarm optimization. Remote Sens 14:5719

Kennedy J, Eberhart R (1995) Particle swarm optimization, Proceedings of ICNN'95-international conference on neural networks. IEEE, pp 1942–1948

Khormali F, Ajami M, Ayoubi S, Srinivasarao C, Wani SP (2009) Role of deforestation and hillslope position on soil quality attributes of loess-derived soils in Golestan province. Iran Agric, Ecosyst Environ 134:178–189

Kılıc OM, Budak M, Gunal E, Acır N, Halbac-Cotoara-Zamfir R, Alfarraj S, Ansari MJ (2022) Soil salinity assessment of a natural pasture using remote sensing techniques in central Anatolia. Turkey PloS one 17:e0266915

Lesch SM, Strauss DJ, Rhoades JD (1995a) Spatial prediction of soil salinity using electromagnetic induction techniques: 1. Statistical prediction models: a comparison of multiple linear regression and cokriging. Water Resour Res 31:373–386

Lesch SM, Strauss DJ, Rhoades JD (1995b) Spatial prediction of soil salinity using electromagnetic induction techniques: 2. An efficient spatial sampling algorithm suitable for multiple linear regression model identification and estimation. Water Resour Res 31:387–398

Li H, Lu Y, Zheng C, Yang M, Li S (2019) Groundwater level prediction for the arid oasis of Northwest China based on the artificial bee colony algorithm and a back-propagation neural network with double hidden layers. Water 11:860

Li H, Webster R, Shi Z (2015) Mapping soil salinity in the Yangtze delta: REML and universal kriging (E-BLUP) revisited. Geoderma 237:71–77

Loc HH, Lixian ML, Park E, Dung TD, Shrestha S, Yoon Y-J (2021) How the saline water intrusion has reshaped the agricultural landscape of the Vietnamese Mekong Delta, a review. Sci Total Environ 794:148651

Lv Z, Peng R (2021) A novel periodic learning ontology matching model based on interactive grasshopper optimization algorithm. Knowl-Based Syst 228:107239

Meng X-B, Gao XZ, Lu L, Liu Y, Zhang H (2016) A new bio-inspired optimisation algorithm: Bird Swarm Algorithm. Journal of Experimental & Theoretical Artificial Intelligence 28:673–687

Meraihi Y, Gabis AB, Mirjalili S, Ramdane-Cherif A (2021) Grasshopper optimization algorithm: theory, variants, and applications. IEEE Access 9:50001–50024

Metternicht GI, Zinck J (2003) Remote sensing of soil salinity: potentials and constraints. Remote Sens Environ 85:1–20

Moayedi H, Nguyen H, Kok Foong L (2021) Nonlinear evolutionary swarm intelligence of grasshopper optimization algorithm and gray wolf optimization for weight adjustment of neural network. Eng Comput 37:1265–1275

Mulder VL, de Bruin S, Schaepman ME, Mayr TR (2011) The use of remote sensing in soil and terrain mapping — a review. Geoderma 162:1–19

Naimi S, Ayoubi S, Zeraatpisheh M, Dematte JAM (2021) Ground observations and environmental covariates integration for mapping of soil salinity: a machine learning-based approach. Remote Sens 13:4825

Nguyen HD (2022a) Flood susceptibility assessment using hybrid machine learning and remote sensing in Quang Tri province. Vietnam Transactions in GIS

Nguyen HD (2022b) GIS-based hybrid machine learning for flood susceptibility prediction in the Nhat Le–Kien Giang watershed. Vietnam Earth Science Informatics 15:2369–2386

Nguyen HD, Nguyen Q-H, Du QVV, Nguyen THT, Nguyen TG, Bui Q-T (2021a) A novel combination of deep neural network and manta ray foraging optimization for flood susceptibility mapping in Quang Ngai province. Vietnam Geocarto Int:1–25

Nguyen TG, Tran NA, Vu PL, Nguyen Q-H, Nguyen HD, Bui Q-T (2021b) Salinity intrusion prediction using remote sensing and machine learning in data-limited regions: a case study in Vietnam's Mekong Delta. Geoderma Reg 27:e00424

Nosrati K, Collins AL (2019) A soil quality index for evaluation of degradation under land use and soil erosion categories in a small mountainous catchment. Iran J Mountain Sci 16:2577–2590

Ouyang C, Zhu D, Wang F (2021) A learning sparrow search algorithm. Comput Intell Neurosci 2021: 3946958

Peng J, Li S, Makar RS, Li H, Feng C, Luo D, Shen J, Wang Y, Jiang Q, Fang L (2022) Proximal Soil Sensing of Low Salinity in Southern Xinjiang. China Remote Sensing 14:4448

Pradhan AMS, Kim Y-T (2020) Rainfall-induced shallow landslide susceptibility mapping at two adjacent catchments using advanced machine learning algorithms. ISPRS Int J Geo Inf 9:569

Quiroz JC, Mariun N, Mehrjou MR, Izadi M, Misron N, Radzi MAM (2018) Fault detection of broken rotor bar in LS-PMSM using random forests. Measurement 116:273–280

Rao P, Chen S, Sun K (2006) Improved classification of soil salinity by decision tree on remotely sensed images, ICO20: Optical Information Processing. SPIE:911–918

Saber M, Boulmaiz T, Guermoui M, Abdrabo KI, Kantoush SA, Sumi T, Boutaghane H, Nohara D, Mabrouk E (2021) Examining LightGBM and CatBoost models for wadi flash flood susceptibility prediction. Geocarto Int:1–26

Sahin EK (2020) Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. SN Applied Sciences 2:1–17

Samat A, Li E, Wang W, Liu S, Lin C, Abuduwaili J (2020) Meta-XGBoost for hyperspectral image classification using extended MSER-guided morphological profiles. Remote Sens 12:1973

Sándor Z, Tállai M, Kincses I, László Z, Kátai J, Vágó I (2020) Effect of various soil cultivation methods on some microbial soil properties. DRC Sustainable Future 1:14–20

Saremi S, Mirjalili S, Lewis A (2017) Grasshopper optimisation algorithm: theory and application. Adv Eng Softw 105:30–47

Scudiero E, Skaggs TH, Corwin DL (2014) Regional scale soil salinity evaluation using Landsat 7, western San Joaquin Valley, California, USA. Geoderma Reg 2:82–90

Shahabi M, Jafarzadeh AA, Neyshabouri MR, Ghorbani MA, Valizadeh Kamran K (2017) Spatial modeling of soil salinity using multiple linear regression, ordinary kriging and artificial neural network methods. Arch Agron Soil Sci 63:151–160

Taghadosi MM, Hasanlou M, Eftekhari K (2019) Retrieval of soil salinity from Sentinel-2 multispectral imagery. European J Remote Sensing 52:138–154

Taghizadeh-Mehrjardi R, Ayoubi S, Namazi Z, Malone B, Zolfaghari AA, Sadrabadi FR (2016) Prediction of soil surface salinity in arid region of central Iran using auxiliary variables and genetic programming. Arid Land Res Manag 30:49–64

Taghizadeh-Mehrjardi R, Sarmadian F, Minasny B, Triantafilis J, Omid M (2014) Digital mapping of soil classes using decision tree and auxiliary data in the Ardakan region. Iran Arid Land Res Manage 28:147–168

Tajgardan T, Ayoubi S, Shataee S, Sahrawat K (2010) Soil surface salinity prediction using ASTER data: comparing statistical and geostatistical models. Aust J Basic Appl Sci 4:457–467

Tajik S, Ayoubi S, Zeraatpisheh M (2020) Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran. Geoderma Reg 20:e00256

Tho N, Vromant N, Hung NT, Hens L (2008) Soil salinity and sodicity in a shrimp farming coastal area of the Mekong Delta. Vietnam Environ Geol 54:1739–1746

Tikhamarine Y, Souag-Gamane D, Ahmed AN, Sammen SS, Kisi O, Huang YF, El-Shafie A (2020) Rainfall-runoff modelling using improved machine learning methods: Harris hawks optimizer vs. particle swarm optimization. J Hydrol 589:125133

Tilse M (2022) Soil constraint diagnosis and mapping. In: Dang Y, Menzies N, Dalal R (eds) Soil constraints on crop production. Cambridge Scholars Publishing, Newcastle, UK

Tran VN, Kim J (2022) Robust and efficient uncertainty quantification for extreme events that deviate significantly from the training dataset using polynomial chaos-kriging. J Hydrol 609:127716

Varol Altay E, Alatas B (2020) Bird swarm algorithms with chaotic mapping. Artif Intell Rev 53:1373–1414

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

Wang F, Yang S, Wei Y, Shi Q, Ding J (2021a) Characterizing soil salinity at multiple depth using electromagnetic induction and remote sensing data with random forests: a case study in Tarim River Basin of southern Xinjiang. China Sci Total Environ 754:142030

Wang F, Yang S, Yang W, Yang X, Jianli D (2019a) Comparison of machine learning algorithms for soil salinity predictions in three dryland oases located in Xinjiang Uyghur Autonomous Region (XJUAR) of China. European J Remote Sensing 52:256–276

Wang G-G (2018) Moth search algorithm: a bio-inspired metaheuristic algorithm for global optimization problems. Memetic Computing 10:151–164

Wang J-S, Li S-X (2019) An improved grey wolf optimizer based on differential evolution and elimination mechanism. Sci Rep 9:1–21

Wang J, Ding J, Yu D, Ma X, Zhang Z, Ge X, Teng D, Li X, Liang J, Lizaga I (2019b) Capability of Sentinel-2 MSI data for monitoring and mapping of soil salinity in dry and wet seasons in the Ebinur Lake region, Xinjiang, China. Geoderma 353:172–187

Wang J, Ding J, Yu D, Teng D, He B, Chen X, Ge X, Zhang Z, Wang Y, Yang X (2020a) 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, 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 Sensing 13:305

Wang N, Xue J, Peng J, Biswas A, He Y, Shi Z (2020b) Integrating remote sensing and landscape characteristics to estimate soil salinity using machine learning methods: a case study from Southern Xinjiang. China Remote Sensing 12:4118

Wei Q, Nurmemet I, Gao M, Xie B (2022) Inversion of soil salinity using multisource remote sensing data and particle swarm machine learning models in Keriya Oasis. Northwestern China Remote Sensing 14:512

Wei Y, Ding J, Yang S, Wang F, Wang C (2021) Soil salinity prediction based on scale-dependent relationships with environmental variables by discrete wavelet transform in the Tarim Basin. Catena 196:104939

Wicke B, Smeets E, Dornburg V, Vashev B, Gaiser T, Turkenburg W, Faaij A (2011) The global technical and economic potential of bioenergy from salt-affected soils. Energy Environ Sci 4:2669–2681

Wu W, Zucca C, Muhaimeed AS, Al-Shafie WM, Fadhil Al-Quraishi AM, Nangia V, Zhu M, Liu G (2018) Soil salinity prediction and mapping by machine learning regression in Central Mesopotamia, Iraq. Land Degrad Dev 29:4005–4014

Xue J, Shen B (2020) A novel swarm intelligence optimization approach: sparrow search algorithm. Sys Sci Control Eng 8:22–34

Yahiaoui I, Douaoui A, Zhang Q, Ziane A (2015) Soil salinity prediction in the Lower Cheliff plain (Algeria) based on remote sensing and topographic feature analysis. J Arid Land 7:794–805

Yan L, Zhou S, Wu C-F, Li H-Y, Feng L (2007) Improved prediction and reduction of sampling density for soil salinity by different geostatistical methods. Agric Sci China 6:832–841

Zeraatpisheh M, Ayoubi S, Sulieman M, Rodrigo-Comino J (2019) Determining the spatial distribution of soil properties using the environmental covariates and multivariate statistical analysis: a case study in semi-arid regions of Iran. J Arid Land 11:551–566

Zhou J, Huang S, Zhou T, Armaghani DJ, Qiu Y (2022) Employing a genetic algorithm and grey wolf optimizer for optimizing RF models to evaluate soil liquefaction potential. Artif Intell Rev:1–33

Zhu K, Sun Z, Zhao F, Yang T, Tian Z, Lai J, Zhu W, Long B (2021) Relating hyperspectral vegetation indices with soil salinity at different depths for the diagnosis of winter wheat salt stress. Remote Sens 13:250

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.