

Research

Remote sensing and machine learning algorithms to predict soil salinity in southern Kazakhstan

Yedilkhan Amirgaliyev^{1,3} · Ravil Mukhamediev^{1,2} · Timur Merembayev¹ · Yan Kuchin^{1,2} · Aisulyu Ataniyazova^{1,3} · Perizat Omarova^{1,3}

Received: 11 June 2024 / Accepted: 21 October 2024

Published online: 28 October 2024

© The Author(s) 2024 **OPEN**

Abstract

Salinization and land degradation are significant challenges in the southern regions of Kazakhstan. These issues arise due to climate change, unequal water resource distribution, and human impact. The primary concern revolves around water resources, which are influenced by the area's trans boundary flow of major rivers. The low level of water and food security has pushed the development of new approaches based on remote sensing monitoring and geographic information systems (GIS) to provide solutions for soil salinity. The research aims to focus on utilizing high-resolution radar images. This data type is effective for cloudy weather and can be useful for continued monitoring of some areas. Machine learning methods can solve the problem of automatic mapping of agricultural land salinity in Kazakhstan's southern regions. The precise mapping of the salinity area helps prevent or decrease salinity's impact on agriculture. The experiment realized that complex models such as LightGBM do not have significant accuracy performance over simple models on a small dataset compared with Ridge regression. The results allow us to recommend an approach for further improvement with ground-based measurement data and other deep-learning methods for mapping the salinity of agricultural lands.

Keywords Soil salinity · Machine learning · Decision tree · LightGBM · Ridge linear regression · Synthetic aperture radar

1 Introduction

One of the reasons for agricultural land degradation is increasing salinity, which is caused by climate change and human activity, which can reduce the area of arable land and pastures. Climate factors accelerate soil salinization while increasing agricultural areas, and the increasing use of land and water resources exacerbates the problem [1]. The irrigation salinization of agricultural land in southern Kazakhstan is a significant issue that affects Turkestan, Almaty, Zhambyl, and Kyzylorda regions. The main issue concerns water resources formed by the transboundary flow of large rivers in the area, such as the Syrdarya, Ile, and Chu rivers. Kazakhstan is located in the lower reaches of river basins, making it highly susceptible to water supply issues. In southern Kazakhstan, challenges related to supplying irrigated arable land stem from increased water consumption in neighboring countries and the impact of climate change. Soil salinity and its mapping have become pressing issues due to low water and food security levels in this region, which includes major

✉ Timur Merembayev, t.merembayev@iitu.edu.kz; Yedilkhan Amirgaliyev, amir_ed@mail.ru; Ravil Mukhamediev, ravil.muhamedyev@gmail.com; Yan Kuchin, ykuchin@mail.ru; Aisulyu Ataniyazova, aissulu.12@mail.ru; Perizat Omarova, omarova.peryzat@gmail.com |

¹Institute of Information and Computational Technologies, Almaty, Kazakhstan. ²Institute of Automation and Information Technology, Satbayev University (KazNRTU), Almaty, Kazakhstan. ³Al-Farabi Kazakh National University, Almaty, Kazakhstan.



cities like Almaty, Shymkent, Taraz, and Kyzyl-Orda. Various modern solutions are based on remote sensing monitoring, and GIS has been applied to provide solutions for soil salinity.

In the Syr Darya River basin, situated on the border of Uzbekistan and Kazakhstan, lies a substantial irrigation area known as "Golodnaya Steppe." Kazakhstan possesses approximately 140 thousand hectares of cultivated land within this massif [1]. The agricultural lands in this river basin face secondary salinization due to the region's saline groundwater. Disruptions in the irrigation and drainage systems contribute to salt increase in the upper layers of soil. The adverse impact of soil salinization is evident in reduced crop yields, which in turn adversely affect crop production. Managing water usage and addressing the secondary salinization of fields are crucial issues for the efficient administration of agricultural output and ensuring nutritional security in the southern regions of Kazakhstan.

This study seeks to develop an advanced method for assessing soil salinity by integrating high-resolution radar imagery with field data. The proposed approach involves extracting quantitative textural features from radar images, which are then used to train machine learning models with field measurements serving as ground truth labels. The ultimate goal is to apply this machine learning technique to accurately map soil salinity across arable lands in Kazakhstan.

2 Related works

2.1 Classical salinity estimation methods based on spectral data

For over 40 years, remote sensing data have been used to estimate soil salinity [2]. Numerous similar techniques for assessing the salinity of irrigated agricultural land based on satellite data and GIS technologies have been described in the research papers [3, 4], including the salinity of some areas of irrigated area in Kazakhstan [5, 6]. In recent times, the predominant approach for assessing soil salinity involves leveraging satellite data and GIS technologies. In certain instances, these methods can efficiently rival costly ground-based monitoring techniques. The satellite-based monitoring approaches outlined in various research studies exhibit similar characteristics [7, 8]. Various satellite images or their time series are employed from diverse orbital platforms, including hyperspectral data [9], medium spatial resolution satellite data like MODIS [10], and radar images [11]. To detect soil salinity, researchers analyze the visible and near-infrared spectral ranges [7, 8]. Different salinity indices are developed by combining specific channels and vegetation indices. These indices serve to establish a coherent link between the actual salinity levels (measured on the ground) of arable land and the spectral properties of the underlying surface. Additionally, clustering algorithms for remote sensing data and machine learning techniques can also be employed for data analysis [12, 13].

2.2 Machine learning methods in salinity estimation problems

The field of machine learning (ML) is a part of artificial intelligence (AI) that used to create learning systems based on AI principles [14]. ML methods have been successfully applied to classification, clustering, and regression analysis problems in various scientific disciplines [15–22]. Since around 2012, ML methods have also been used in studies related to soil salinity. The results of combining remote sensing data with machine learning algorithms for large-scale and high-precision spatial prediction of soil salinity highlight the potential of this approach in planning and implementing effective salinity management strategies in semi-arid regions [23, 24].

For example, the assessment and mapping of soil salinity has been demonstrated by [16], in the case of the Aral Sea basin using a combination of input variables, including terrain indices, elevation, remote sensing data, and groundwater levels. The output variable was obtained by measuring the soil profile's electrical conductivity (EC) using an electromagnetic induction instrument at a depth of 1.5 m. The resulting classification algorithm achieved an accuracy metric of about 90% in estimating soil salinity above a threshold of approximately 0.7 dS/m. The authors also analyzed the model's sensitivity and found that soil salinity was influenced mainly by terrain relief, convexity or concavity, and surface water retention. In descending order of importance, the terrain variables significantly influenced the model were curvature, plan curvature, profile curvature, and solar radiation (solar). This research highlights the potential of ML methods to effectively estimate and map soil salinity in different regions.

According to [20], the authors proposed a regression model to monitor groundwater salinization resulting from seawater intrusion into the eastern Nile Delta, Egypt aquifer. The researchers used hydrogeological and hydrogeochemical data from groundwater wells in 1996, 2007, and 2018, as well as the distance from the coastline, aquifer type, and hydraulic pressure, as inputs to create four baseline machine learning models. These models aimed to predict the base

exchange index, groundwater quality index for seawater intrusion, and water quality. Four machine learning models were employed, including logistic regression, Gaussian process regression, feed-forward neural networks, and long short-term memory, one of the deep learning models. The FFNN model outperformed the other models, achieving higher scores for root mean square error (RMSE) and R2 values during testing, with R2 values of 0.9667, 0.9316, and 0.9259 for the base exchange index, groundwater quality index for seawater intrusion, and water quality, respectively.

Extensive research described in study by [15], underscores the correlation between the ground measurements and the Landsat-8 OLI and Sentinel-2 MSI multispectral images. For this, the authors utilized soil samples from dry regions in northwestern China and hyperspectral remote sensing data to estimate soil salinity. The results indicated that the higher resolution Sentinel-2 MSI imagery provided more precise salinity estimations in desert areas.

The approach taken by [14], explores the correlation between the backscattering coefficient and electrical conductivity in Maha-Sarakham province, located in the northeast region of Thailand. The data used in the study was obtained from ALOS-PALSAR in four different polarizations (HH, HV, VH, and VV) at a resolution of 12.5 ms. The ground measurements were taken at a depth of 5 cm with a distance of 20 ms between each measurement, resulting in around 500 data points. GPS was used to identify each measurement accurately with a precision of 4 ms. In the laboratory, electrical conductivity was determined at 25 C (dS/m). Several feedforward neural net models were utilized to establish a regression relationship between radar data and ground measurements. On average, these models demonstrated an R2 value of 0.85 and an RMSE of 5, indicating a strong correlation between the BC and EC.

The problem of salinity mapping in the Mekong Delta has been addressed in [21], where the authors proposed an interesting approach using radar images and machine learning algorithms. They used Sentinel-1 Synthetic Aperture Radar (SAR) C-band radar images to generate features and considered the regression problem to predict soil electrical conductivity. The Gray Level Co-occurrence Matrix (GLCM) method was used to construct features from radar images. The best model was achieved using the Gaussian Process Regressor (GPR) model, which had an RMSE of 2.885, MAE of 1.897, and R2 of 0.808.

The authors from research [25] offer a comprehensive and comparative analysis of machine learning (ML) algorithms for soil salinity modeling, incorporating field-based electrical conductivity (EC) data and Landsat-8 OLI satellite images with derived environmental covariates. As key finding, it can be noted that the SHAP analysis revealed that vegetation indices were more influential in predicting soil salinity in the West Lake Urmia Playas, while visible bands had a greater impact in the Southeast Lake Urmia Playas. This finding underscores the importance of region-specific covariates in soil salinity modeling.

Although remote sensing and radar data have successfully mapped salinity, the studies have focused on areas with humid climates. To the authors' knowledge, similar studies have not been conducted in arid regions. There has been no comparison of methods based on optical remote sensing channels and those based on radar images.

3 Materials and methods

3.1 Study area

As shown in Fig. 1, the first step is to obtain the necessary satellite images from the Sentinel-2 satellite. The preprocessing steps include radiometric and atmospheric corrections to improve image quality and eliminate any distortion caused by factors such as clouds or atmospheric interference. Sentinel-2 captures images in several spectral ranges, from visible to shortwave infrared. Specific ranges sensitive to salinity fluctuations are selected to assess soil salinity. The near infrared (NIR) range is often used because it is sensitive to changes in soil properties associated with salinization.

Spectral indexes are applied to enhanced images to identify areas with high soil salinity. These methods include the analysis of the spectral characteristics of various types of vegetation cover and the identification of those associated with saline soils. The results obtained from the analysis of satellite images are confirmed using ground-based data. Field measurements or other ground-based observations are used to confirm the presence and extent of soil salinity in areas identified by satellite imagery.

Machine learning algorithms can use various data sources, including ground measurements and soil samples, to classify soil salinity. On the other hand, remote sensing provides valuable information about the Earth's surface on a larger scale and can record spatial and temporal changes in soil salinity. Comparing the results of both approaches in

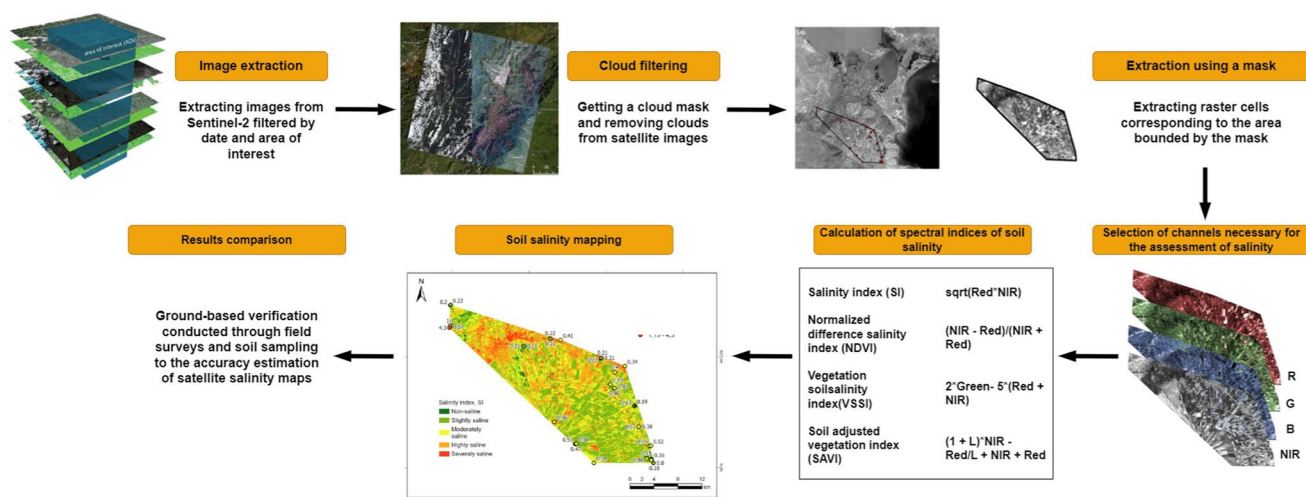


Fig. 1 Stages of soil salinity assessment based on remote sensing

classifying soil salinity provides a comprehensive and accurate understanding of the problem, provides scalability for large-scale monitoring, and confirms the accuracy of classification models.

Three main areas were identified, located in the Zhetisu and Almaty regions, where sampling studies were conducted: near the city of Kunaev, near the village of Shelek and near Lake Alakol (Fig. 2).

Remote sensing methods make it possible to carry out large-scale monitoring of soil salinity in vast territories. This scalability is especially important for understanding the extent and distribution of soil salinity, which can vary significantly



Fig. 2 Extraction of the studied area from a satellite image

Fig. 3 Sentinel-2 bands

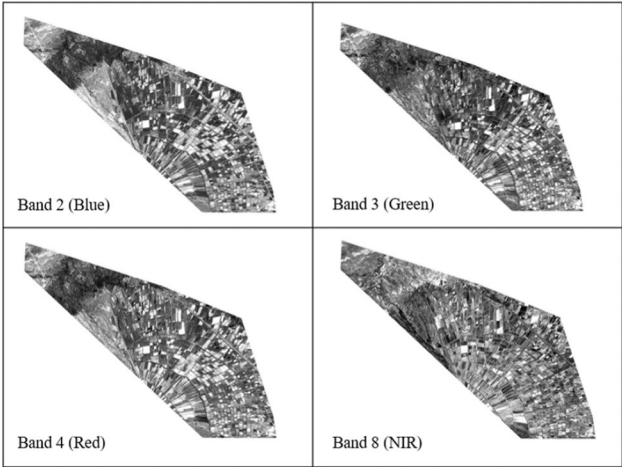


Table 1 Soil salinity classification

Salinity class	EC(1:5) range (dS/m)	Salinity index SI5 (Band 2/ Band 4)
Non-saline	0–0.15	0–0.5
Slightly saline	0.16–0.22	0.5–1
Moderately saline	0.23–0.73	1–1.2
Highly saline	0.74–1.15	1.2–1.5
Severely saline	> 1.16	> 1.6

from region to region. The assessment of soil salinity based on remote sensing data can be carried out by analyzing specific bands sensitive to the presence of salts in the soil. Data was obtained from the Sentinel-2 satellite. It has several spectral bands, each of which is sensitive to different aspects of the Earth’s surface. The blue, green, red, and near-infrared bands are usually used to assess soil salinity (Fig. 3).

3.2 Satellite data

Based on the analysis of Sentinel-2 satellite data, maps were created showing the distribution of soil salinity. Figure 4 shows a soil salinity map of the Shelek region based on the SI (salinity index) spectral index obtained from remote sensing data. The degree of soil salinity in the areas of interest was determined by the electrical conductivity values and the SI5 spectral index, divided into 5 classes (Table 1).

Figure 5 shows a soil salinity map of the Konayev region based on the SI spectral index obtained from remote sensing data [1].

Figure 6 shows a soil salinity map of the Alakol region based on the SI spectral index obtained from remote sensing data [1].

According to Figs. 4, 5, 6, the following conclusion can be drawn: The large-scale territories often demonstrate significant spatial heterogeneity of soil salinity levels (in the case of the Alakol region). Variability in soil properties, land use, and environmental conditions over a large area can make it difficult to accurately collect and model using Sentinel-2 satellite data and LightGBM machine learning algorithm.

However, it is important to note that the accuracy of the results depends on various factors, such as the quality and resolution of the remote sensing data and the availability of ground-based confidence measurements for training algorithms. Additionally, the complexity of the soil salinity model in the area of interest also plays a role. Therefore, while the conclusion about the similarity between the accuracy of assessing the degree of salinity is generally correct, we wanted to confirm the results of ground-based measurements (EC) with the results of remote sensing for a more comprehensive assessment [17, 22].

Fig. 4 Soil salinity map of the Shelek region

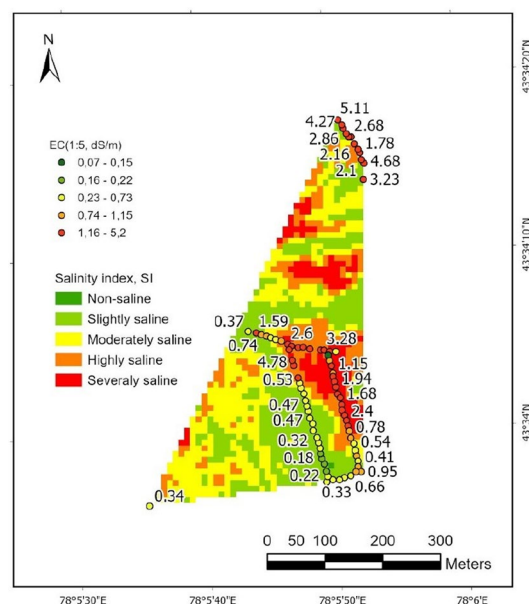


Fig. 5 Soil salinity map of the Konayev region

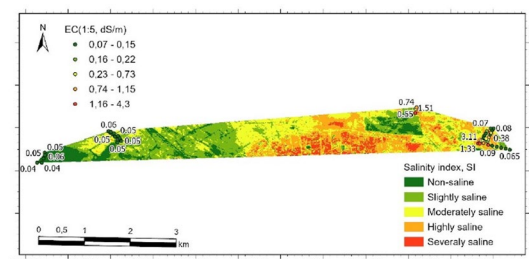
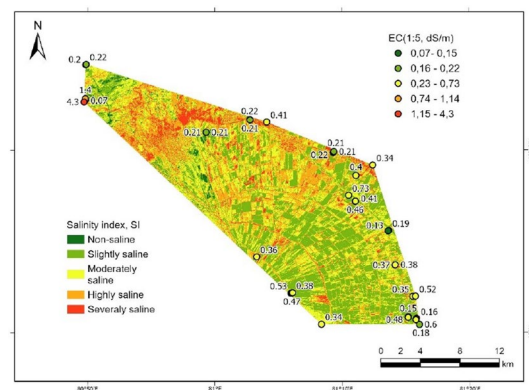


Fig. 6 Soil salinity map of the Alakol region



3.3 Machine learning approach

In the study [21], authors propose several machine learning algorithms, we would like to extend the algorithms and compare their results with other approaches, for example, algorithms based on the extremal gradient boosting approach for constructing various tree solutions. One such algorithm is LightGBM, which was not considered in the study [21]. This algorithm is widely used for classification and regression problems and has proven well for tabular data. Compared to similar algorithms such as XGBoost and CatBoost, LightGBM has an advantage in estimation speed over other algorithms [26, 27]. The second algorithm is ridge regression. The main idea is to apply this algorithm to small datasets and compare them with more discovered algorithms.

The research task is a binary classification problem based on radar data. A similar approach to the model was described in the paper [21, 22]. LightGBM is a gradient boosting framework based on decision trees to increase the model's efficiency and reduce memory usage.

Two innovative methods, Gradient-based One Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), address the constraints of the histogram-based algorithm commonly utilized in all Gradient Boosting Decision Tree (GBDT) frameworks. These techniques, GOSS and EFB, are integral to the LightGBM Algorithm, working in tandem to enhance model efficiency and confer a competitive advantage over other GBDT frameworks.

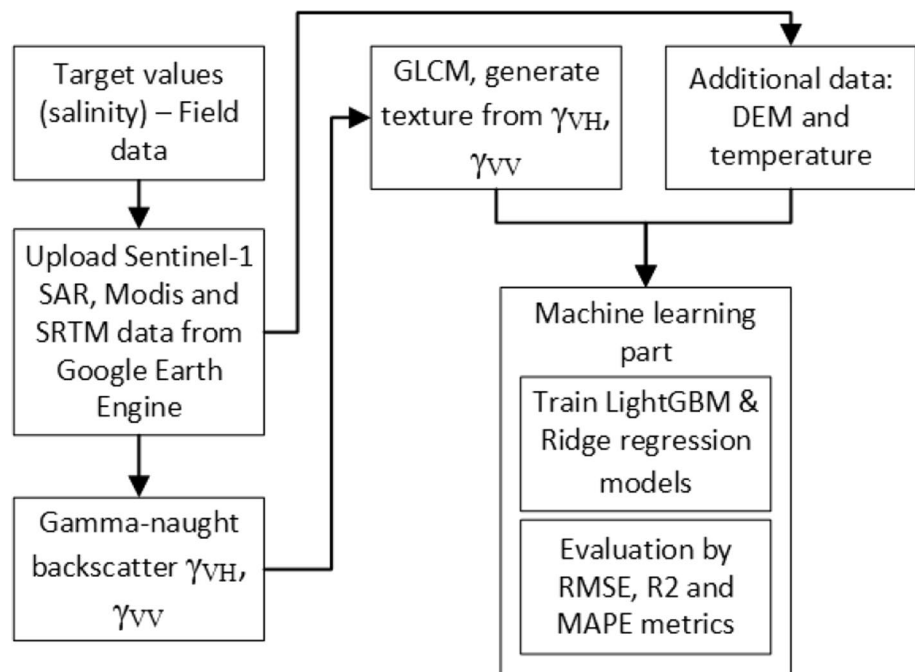
In the Gradient-based One Side Sampling Technique for LightGBM, diverse data instances play distinct roles in determining information gain. Instances with substantial gradients, indicating under-trained models, influence information gain more. GOSS selectively retains instances with significant gradients, such as those surpassing a predetermined threshold or within the upper percentiles, while randomly discarding instances with minor gradients to preserve the accuracy of gain estimation. This approach can result in more precise gain estimation compared to uniform random sampling, particularly when information gain values span a wide range.

Ridge regression, a form of linear regression employed in machine learning to mitigate overfitting in linear models, proves valuable when dealing with highly correlated variables. By penalizing the coefficients of these variables, it helps curb overfitting. This is achieved by introducing a penalty term into the error function, which reduces the magnitude of the coefficients. This approach is advantageous in noisy datasets as it prevents the model from being overly influenced by individual data points. Additionally, ridge regression exhibits lower sensitivity to outliers compared to traditional linear regression methods. However, it comes with the drawback of being computationally intensive and often necessitating larger datasets for accurate results. In ridge regression, the penalty term, known as the L2 norm, is added to the cost function, proportional to the sum of the squares of the coefficients. This adjustment simplifies the optimization problem and results in smaller coefficients. By encouraging a balance between accurate fitting of the training data and model simplicity, ridge regression enhances the generalizability of machine learning models.

The process flowchart for salinity detection is depicted in Fig. 7 [17]. The steps involved in training and deploying the model are as follows:

1. Acquiring salinity data using field studies;
2. Acquiring radar images from Google Earth Engine;
3. Extraction of linear backscattering intensity in VV and VH polarizations;
4. Texture analysis using the Gray Level Co-occurrence Matrix GLCM method;
5. Evaluation of the quality of the trained model.

Fig. 7 Flowchart for training and applying of machine learning model



Field studies of the soil were carried out from May 23 to June 18, 2022. Sampling routes have been designed, considering roads and vehicle access to potential samples. A portable GPS (Garmin 65, positioning accuracy <5 m) was used to record geographic locations. Although the GPS device positioning accuracy was not precise, it was sufficient to provide a suitable location match between samples and remote sensing image pixels (resolution is 10 m for Sentinel 1). Sampling locations are presented and described (Fig. 8).

We defined three main areas located in the Zhetysu and Almaty regions where sampling probes were performed: near Konayev city (dataset is called Konayev), near Shelek village (dataset is called Shelek), and near Alakol lake (dataset is called Alakol). We selected the Konayev and Shelek locations because they are close to the Ili River but on the opposite river bank. Near Konayev, there is less vegetation and sandy soil. The Shelek area is more agricultural, but in summer, alkaline soil appears. The Alakol is close to Alakol Lake, and the alkaline soil differs. In the paper, we used this abbreviation of the dataset to compare the results, see Fig. 9.

A total of 203 soil samples were taken, Fig. 9 shows the samples taken and their sampling routes. According to relief, the sampling process was random, and the distance between samples was as far as possible. All soil specimens were appropriately sealed, labeled, and transported to the laboratory to assess their electrical conductivity. Following transportation, the samples were thoroughly air-dried and sifted through a 2 mm mesh sieve to eliminate impurities. Subsequently, the soil material was combined with water at a ratio of 1:5 to facilitate the analysis of soil electrical conductivity. This analysis used a digital measuring device (Hanna GroLine HI9814) under standard room temperature conditions (25 C).

Sentinel-1 was chosen as the primary source of remote sensing for checking the hypothesis. This type of satellite's main advantages are that it does not depend on weather (no clouds) and has a good resolution (10 ms). Sentinel-1 can obtain multiple images by receiving distinct polarizations at the same time. It can emit a signal in either a horizontal (H) or vertical (V) polarization and receive the signal in both H and V polarizations, allowing it to collect several different images from the same pulse series [28]. Based on the received data VH and VV, we performed preprocessing work to extract data and prepare a dataset for a model.

After preprocessing work on the data from Sentinel 1, we got a dataset of 203 rows with 16 features. Figure 10 shows an RGB image composed of a combination of Red: VH, Green: VV, and Blue: VH/VV on the territory not far from Alakol Lake.

The dataset consists of field soil samples (electrical conductivity) and 16 features generated using a radar image and the Gray Level Co-occurrence Matrix GLCM method [29]. The process of creating a GLCM involves using a

Fig. 8 Characteristics of the sampled areas near Konayev city



(a) Dry sands near the village of Kulantobe (Samples No. 001-005). Moderate salinity.



(b) General landscape (samples No. 006-016). No salinization.



(c) In irrigated fields (samples No. 017-019 were taken from the area of natural vegetation near the field), moderate salinity.



(d) The fallow field (samples No. 020-046). The moderate salinity.

Fig. 9 The map of samples probes for the Almaty and Zhetysu regions

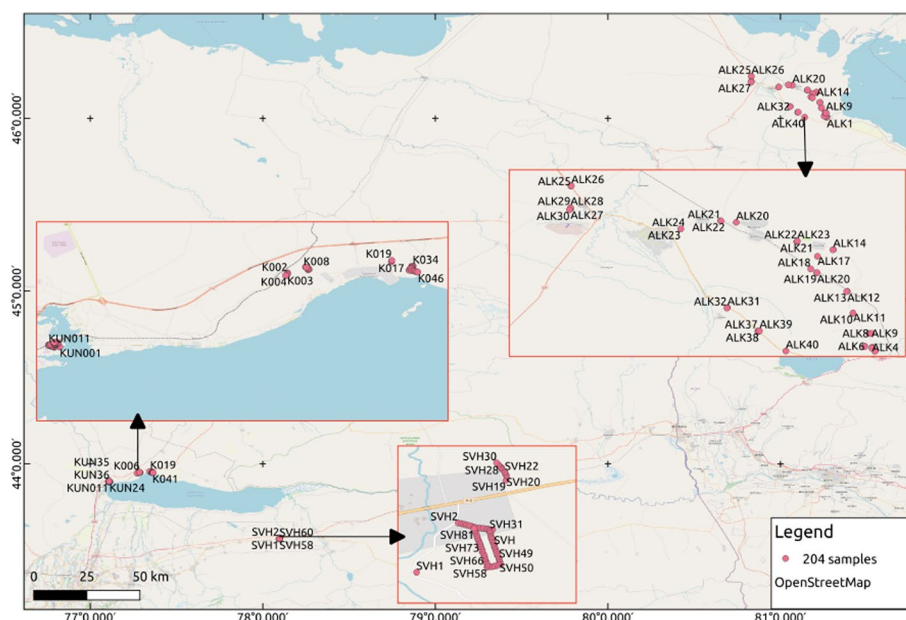
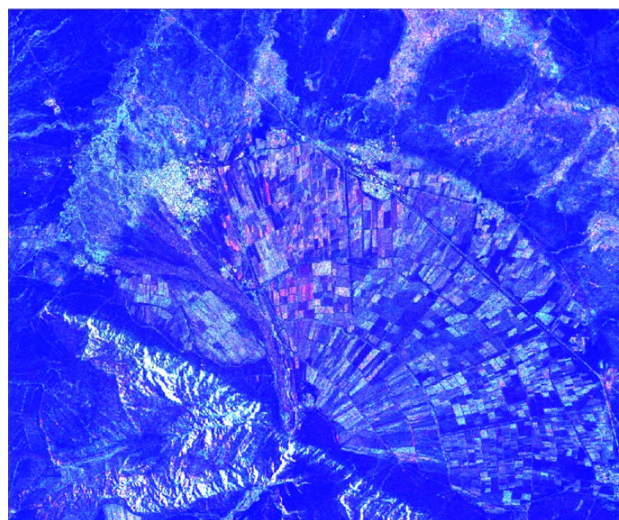


Fig. 10 The composed RGB image, Red: VH, Green: VV, Blue: VH/VV of radar images for the area near Lake Alakol



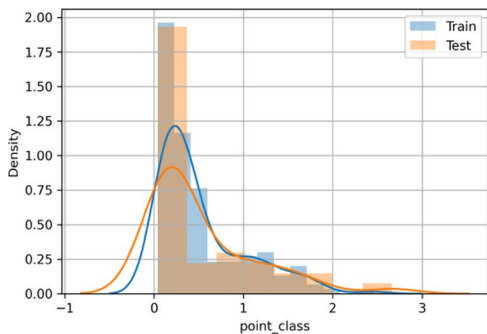
gray-scale image, which is essentially a histogram of grayscale values that co-occur in a given offset over an image. This method and associated texture feature calculations are an image analysis technique. The GLCM is a table that displays the frequency of different combinations of gray levels that co-occur in a specific area of an image. Texture feature calculations use the information in the GLCM to determine the degree of variation in intensity (also known as image texture) at a particular pixel. The GLCM achieves this by calculating how frequently a pixel with a particular gray-level value i appears horizontally, vertically, or diagonally to adjacent pixels with the value j .

Table 2 provides a compilation of abbreviations and corresponding dataset descriptions. Several factors, including GPS-measured altitude, slope derived from a digital elevation model, and temperature data potentially influencing salinity classification, were suggested to enhance radar data. The dataset was expanded using USGS MODIS Earth Surface Temperature (LST) and ground elevation model (ELV) data. These images were acquired at a resolution of 1 km for LST and 30 ms for ELV. Two features were extracted from the ELV data: altitude above sea level and land angle (terrain slope).

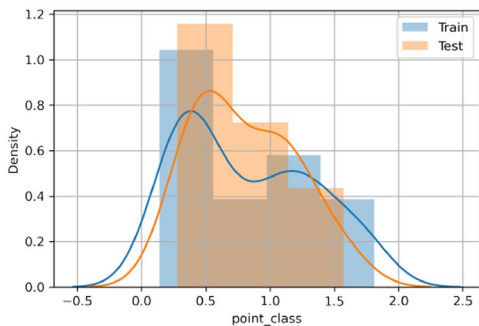
The total data set was randomly divided into training and test sets for experiments. Figure 11 shows the distribution of the electrical conductivity values in training and test datasets for training and validation machine learning models. We tried to maintain the same ratio of electrical conductivity values in the training and test datasets.

Table 2 Names of features and their description

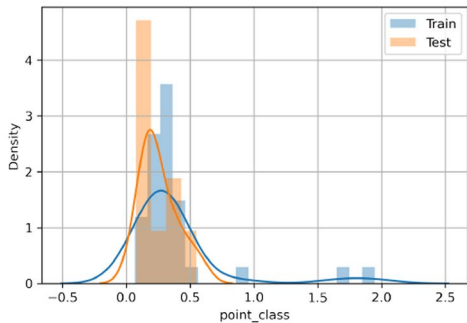
Name	Description
Salinity	Soil salinity, field data
gamma_vv	Polarization VV
dissimilarity_vv	Dissimilarity of gray level co-occurrence matrix for polarization VV
contrast_vv	Contrast of gray level co-occurrence matrix for polarization VV
homogeneity_vv	Homogeneity of gray level co-occurrence matrix for polarization VV
ASM_vv	ASM (homogeneity of an image) of gray level co-occurrence matrix for polarization VV
energy_vv	Energy of gray level co-occurrence matrix for polarization VV
entropy_vv	Entropy of gray level co-occurrence matrix for polarization VV
gamma_vh	Polarization VH
dissimilarity_vh	Dissimilarity of gray level co-occurrence matrix for polarization VH
contrast_vh	Contrast of gray level co-occurrence matrix for polarization VH
homogeneity_vh	Homogeneity of gray level co-occurrence matrix for polarization VH
ASM_vh	ASM (homogeneity of an image) of gray level co-occurrence matrix for polarization VH
energy_vh	Energy of gray level co-occurrence matrix for polarization VH
temp	MODIS land surface temperature
slope	Calculated slope from Digital Elevation Model
altitude	Measured of altitude by GPS



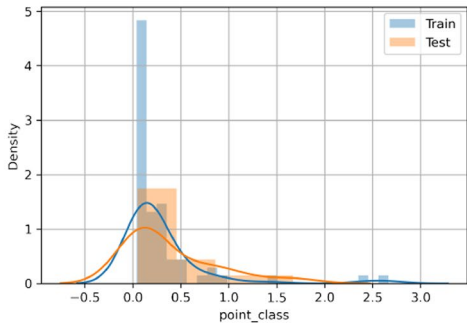
(a) All samples - 203 samples: 162 samples - training, 41 samples - test.



(b) Shelek - 78 samples: 62 samples - training, 16 samples - test.



(c) Alakol - 43 samples: 34 samples - training, 9 samples - test.

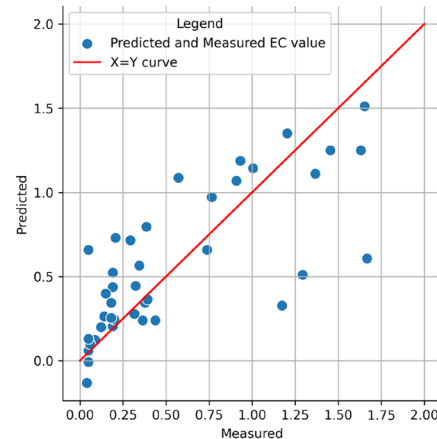


(d) Konayev - 82 samples: 65 samples - training, 17 samples - test.

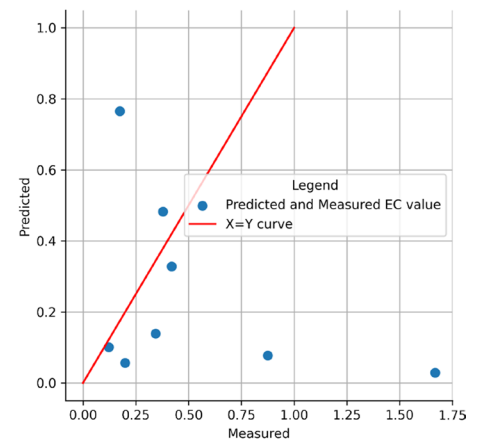
Fig. 11 The distribution of train and test datasets

Table 3 The result of evaluation LightGBM model for the test dataset at 95% confidence interval

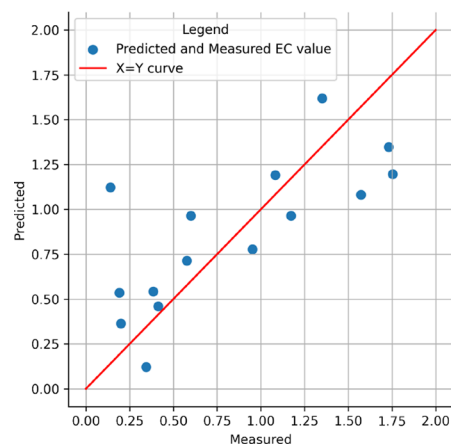
Model	RSME	R2	MAPE	Amount of test dataset
LightGBM (all dataset)	0.4578 ± 0.10	0.403 ± 0.25	114.9 ± 66.5	41
LightGBM (Shelek)	0.3052 ± 0.08	0.3545 ± 0.27	37.45 ± 23.2	16
LightGBM (Alakol)	0.2611 ± 0.09	-2.656 ± 0.43	145.08 ± 24.7	9
LightGBM (Konayev)	0.4459 ± 0.19	0.0496 ± 0.24	201.89 ± 81.8	17

Fig. 12 Scatterplots of soil salinity values predicted by the LightGBM model for the test dataset

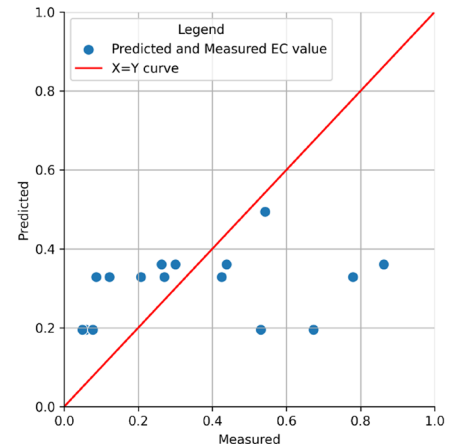
(a) All samples (46).



(b) Alakol area (9 samples).



(c) Shelek area (16 samples).



(d) Konayev area (17 samples).

4 Results

The results of the datasets were used to train two LightGBM and Ridge Linear Regression models. The first experiment with LightGBM was performed on the data from Sentinel-1 on the sampling date. The following metrics were used to evaluate the model: coefficients of determination (R^2), root mean square errors (RMSE) and the average percentage of errors (MAPE). Table 3 shows the results of the model for the test set.

In Table 3, the better result was performed for the Shelek area. Figure 12 shows scatterplots of laboratory measurements (true value) and soil salinity predicted by the LightGBM model. The red line is the 1:1 line (ideal condition). If the accuracy of the regressor is close to 1 value, then the points are closest to the red line. This visualization allows

Table 4 Cross-validation on 5 folds for LightGBM

Model	R2	Standard deviation	Training dataset
LightGBM (all dataset)	0.23	0.12	162
LightGBM (Shelek)	0.46	0.17	62
LightGBM (Alakol)	−1.49	1.45	34
LightGBM (Konayev)	−0.24	0.29	65

Table 5 Comparison of radar data in the sampling date and precipitation date

Model	Size of training dataset	Radar data in the sampling date			Radar data in the precipitation date		
		RSME	R2	MAPE	RSME	R2	MAPE
Ridge regression (all dataset)	162	0.486	0.2113	116.28	0.335	0.224	163.09
Ridge regression (Shelek)	62	0.285	0.677	38.97			
Ridge regression (Alakol)	34	0.22	−2.434	128.58	0.16	−3.622	94.28
Ridge regression (Konayev)	65	0.38	0.101	203.87	0.25	−0.910	171.92

you to identify biases, features (patterns), and anomalies in predicting values. Also, it can be easy to see the potential to improve prediction accuracy.

Cross-validation on 5 folds showed the following result for the LightGBM model (Table 4). The results on the Shelek dataset have a better R2 score.

At the next stage of the experiment, radar data is obtained during precipitation or when the ground is wet. The hypothesis of this experiment was that radar images during the rainy season can positively affect the model's performance. The idea was that wet soil has a different texture than dry soil, and radar waves of wet soil can contain some patterns that should affect machine learning prediction. We used the Ridge Linear Regression model to test this hypothesis.

We chose the dates of Sentinel-1 flight over the area of field sampling after precipitation: Konayev—date of sampling: 26.05.22 and 17.06.22, precipitation was 05.05.22 and 04.06.22, respectively. Alakol—date of sampling: 18.06.22, precipitation was 07.06.22.

Table 5 shows the Ridge Linear Regression model results for the test sample in two periods of satellite data acquisition. The trained Ridge model showed better results on the Shelek dataset and even better than the results of the LightGBM model. The radar data obtained during the rainy season slightly improved the model's accuracy, especially in the RSME metric.

Figure 13 shows a scatterplot for the Ridge model. The density of points close to the red line for the Shelek dataset is explained by the relatively high value of the coefficient of determination $R^2 = 0.677$.

5 Conclusion

SAR data can be used to assess soil salinity in the arid regions of Kazakhstan. At the same time, the numerical experiment showed that soil moisture is a significant factor in improving the results of salinity estimation models using radar images, such as Sentinel-1. In our experiment, the wet ground showed about a 10 percent improvement over dry ground radar data.

The choice of LightGBM and Ridge Regression was guided by a combination of the dataset characteristics, computational efficiency, and the need for interpretable results. The dataset used in this study is relatively small, with 203 samples and 16 features. In such cases, simpler models like Ridge Regression can perform well without the risk of overfitting. LightGBM, while more complex, provides a good balance of performance and speed, making it suitable for handling the features effectively. Experimenting with LightGBM and Ridge regression models on our data shows that complex models such as LightGBM have little accuracy performance over simple models on a small dataset. In this regard, depending on the size of the data set, it is possible to determine which algorithms will be practical in terms of metrics.

In addition, it is important to separate the training and test samples correctly, and it is also necessary to maintain the distribution of values in equal proportions, this will allow the model to be more robust to the values from the test sample.

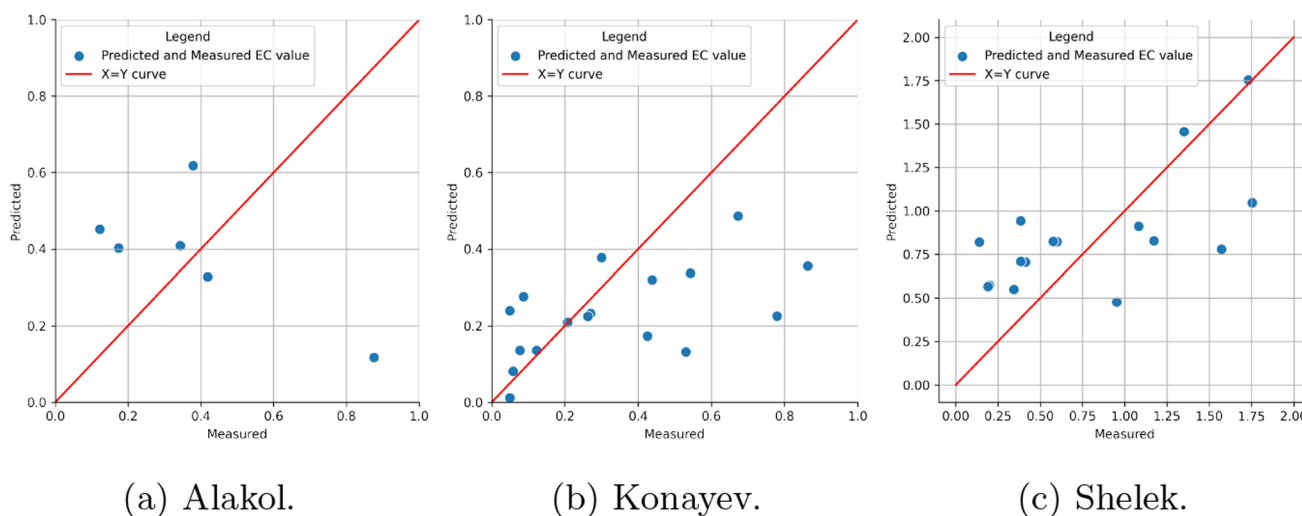


Fig. 13 Scatterplots of soil salinity predicted by the Ridge model

The large-scale territories there's often notable diversity in the distribution of soil salinity levels. The varying soil characteristics, land utilization patterns, and environmental factors across such vast areas pose challenges in effectively capturing and modeling them with remote sensing data and machine learning techniques.

In our further research, we intend to concentrate on detecting salinity by multispectral images based on CNN. Also, we would like to model the salinity in depth with Physics-informed neural networks.

Author contributions Conceptualization, R.M. and Y.A.; methodology, R.M. T.M. and A.A. ; software, T.M. and A.A.; validation, T.M., A.A.Y.K and P.O.; investigation, R.M., Y.A. and P.O.; resources, R.M. and Y.A.; data curation, Y.A., R.M., Y.K., T.M. and A.A.; writing—original draft preparation, R.M., T.M., and Y.K.; writing—review and editing Y.A., A.A. Y.K. and P.O.; visualization, R.M., Y.K. T.M. and A.A.; supervision, Y.A. and R.M.; project administration, Y.A.; funding acquisition, Y.A. All authors have read and agreed to the published version of the manuscript.

Funding This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan Grant BR18574144 “Development of a data mining system for monitoring dams and other engineering structures under the conditions of man-made and natural impacts” and BR21881908 “Complex of urban ecological support (CUES)”.

Data availability Data is presented in GitHub repository: <https://github.com/TimurMZh/QASS>

Declarations

Competing interests The authors declare no Conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. Savin IY, Terekhov A, Amirgaliyev Y, Sagatdinova G. Satellite monitoring of salinization of irrigated soils in southern kazakhstan. *Eurasian Soil Sci.* 2023;56(10):1498–506.
2. Li X, Wang Z, Song K, Zhang B, Liu D, Guo Z. Assessment for salinized wasteland expansion and land use change using gis and remote sensing in the west part of northeast china. *Environ Monit Assess.* 2007;131(1):421–37.
3. Singh A, Dwivedi R. Delineation of salt-affected soils through digital analysis of landsat mss data. *Remote Sens.* 1989;10(1):83–92.
4. Metternicht G, Zinck J. Spatial discrimination of salt-and sodium-affected soil surfaces. *Int J Remote Sens.* 1997;18(12):2571–86.

5. Masoud A, Koike K. Arid land salinization detected by remotely-sensed landcover changes: A case study in the Siwa region, NW Egypt. *J Arid Environ.* 2006;66(1):151–67.
6. Gabdullin B, Zhogolov A, Savin IY, Otarov A, Ibrayeva M, Golovanov D. Application of multi-spectral satellite data for interpretation of soil salinization of the irrigated areas (case study of southern kazakhstan). *Vestnik Moskovskogo universiteta Geografiya.* 2016;5:34–41.
7. Gorji T, Yildirim A, Sertel E, Tanik A. Remote sensing approaches and mapping methods for monitoring soil salinity under different climate regimes. *Int J Environ Geoinf.* 2019;6(1):33–49.
8. Allbed A, Kumar L. Soil salinity mapping and monitoring in arid and semi-arid regions using remote sensing technology: a review. *Adv Remote Sens.* 2013:2013
9. Scudiero E, Skaggs TH, Corwin DL. Regional-scale soil salinity assessment using landsat etm+ canopy reflectance. *Remote Sens Environ.* 2015;169:335–43.
10. Rahmati M, Hamzehpour N. Quantitative remote sensing of soil electrical conductivity using etm+ and ground measured data. *Int J Remote Sens.* 2017;38(1):123–40.
11. Fan X, Weng Y, Tao J. Towards decadal soil salinity mapping using landsat time series data. *Int J Appl Earth Obs Geoinf.* 2016;52:32–41.
12. Qu Y-H, Duan X-L, Gao H-Y, Chen A-P, An Y-Q, Song J-L, Zhou H-M, He T, et al. Quantitative retrieval of soil salinity using hyperspectral data in the region of inner Mongolia hetao irrigation district. *Spectrosc Spect Anal.* 2009;29(5):1362–6.
13. Fallah Shamsi SR, Zare S, Abtahi SA. Soil salinity characteristics using moderate resolution imaging spectroradiometer (modis) images and statistical analysis. *Arch Agronomy Soil Sci.* 2013;59(4):471–89.
14. Phonphan W, Tripathi NK, Tipdecho T, Eiumnoh A. Modelling electrical conductivity of soil from backscattering coefficient of microwave remotely sensed data using artificial neural network. *Geocarto Int.* 2014;29(8):842–59.
15. Zeng W, Zhang D, Fang Y, Wu J, Huang J. Comparison of partial least square regression, support vector machine, and deep-learning techniques for estimating soil salinity from hyperspectral data. *J Appl Remote Sens.* 2018;12(2): 022204.
16. Akramkhanov A, Vlek PL. The assessment of spatial distribution of soil salinity risk using neural network. *Environ Monit Assess.* 2012;184(4):2475–85.
17. Merembayev T, Mukhamediev R, Amirgaliyev Y, Malakhov D, Terekhov A, Kuchin Y, Yakunin K, Symagulov A. The application of machine learning technique to soil salinity mapping in south of Kazakhstan. In: *Asian Conference on Intelligent Information and Database Systems*, 2023:244–253. Springer.
18. Nosair AM, Shams MY, AbouElmagd LM, Hassanein AE, Fryar AE, Abu Salem HS. Predictive model for progressive salinization in a coastal aquifer using artificial intelligence and hydrogeochemical techniques: A case study of the Nile delta aquifer, Egypt. *Environ Sci Pollut Res.* 2022;29(6):9318–40.
19. Omarova P, Amirgaliyev Y, Kozbakova A, Ataniyazova A. Application of physics-informed neural networks to river silting simulation. *Appl Sci.* 2023;13(21):11983.
20. Wang J, Ding J, Yu D, Teng D, He B, Chen X, Ge X, Zhang Z, Wang Y, Yang X, et al. 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.* 2020;707: 136092.
21. Hoa PV, Giang NV, Binh NA, Hai LVH, Pham T-D, Hasanlou M, Tien Bui D. 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.* 2019;11(2):128.
22. Merembayev T, Amirgaliyev Y, Saurov S, Wójcik W. Soil salinity classification using machine learning algorithms and radar data in the case from the south of Kazakhstan. *J Ecol Eng.* 2022;23(10):61–7.
23. Rajath E, Kalambukattu JG, Haldar D, Kumar S. Soil salinity mapping using multisensor data employing machine-learning technique: a case study from indo-gangetic plain. In: *Remote Sensing of Soils*, 2024:439–453. Elsevier.
24. Du R, Xiang Y, Chen J, Lu X, Wu Y, He Y, Xiang R, Zhang Z, Chen Y. Potential of solar-induced chlorophyll fluorescence (sif) to access long-term dynamics of soil salinity using oco-2 satellite data and machine learning method. *Geoderma.* 2024;444: 116855.
25. Aksoy S, Sertel E, Roscher R, Tanik A, Hamzehpour N. Assessment of soil salinity using explainable machine learning methods and landsat 8 images. *Int J Appl Earth Obs Geoinf.* 2024;130: 103879.
26. Al Daoud E. Comparison between xgboost, lightgbm and catboost using a home credit dataset. *Int J Comput Inf Eng.* 2019;13(1):6–10.
27. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. *Artif Intell Rev.* 2021;54:1937–67.
28. Potin P, Rosich B, Grimont P, Miranda N, Shurmer I, O'Connell A, Torres R, Krassenburg M. Sentinel-1 mission status. In: *Proceedings of EUSAR 2016: 11th European Conference on Synthetic Aperture Radar*, 2016:1–6. VDE.
29. Haralick RM, Shanmugam K, Dinstein IH. Textural features for image classification. *IEEE Trans Syst Man Cybern.* 1973;6:610–21.

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