scientific reports

Check for updates

OPEN

Investigation of the spatial and temporal variation of soil salinity using Google Earth Engine: a case study at Werigan–Kuqa Oasis, West China

Shilong Ma[1,2,3], Baozhong He[1,2,3]✉, Boqiang Xie[1,2,3], Xiangyu Ge[1,2,3] & Lijing Han[1,2,3]

Large-scale soil salinity surveys are time-costly and labor-intensive, and it is also more difficult to investigate historical salinity, while in arid and semi-arid regions, the investigation of the spatial and temporal characteristics of salinity can provide a scientific basis for the scientific prevention of salinity, With this objective, this study uses multi-source data combined with ensemble learning and Google Earth Engine to build a monitoring model to observe the evolution of salinization in the Werigan–Kuqa River Oasis from 1996 to 2021 and to analyze the driving factors. In this experiment, three ensemble learning models, Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM), were established using data collected in the field for different years and some environmental variables, After the accuracy validation of the model, XGBoost had the highest accuracy of salinity prediction in this study area, with RMSE of 17.62 dS m^{-1}, R^2 of 0.73 and RPIQ of 2.45 in the test set. In this experiment, after Spearman correlation analysis of soil Electrical Conductivity (EC) with environmental variables, we found that the near-infrared band in the original band, the DEM in the topographic factor, the vegetation index based on remote sensing, and the salinity index soil EC had a strong correlation. The spatial distribution of salinization is generally characterized by good in the west and north and severe in the east and south. Non-salinization, light salinization, and moderate salinization gradually expanded southward and eastward from the interior of the western oasis over 25 years. Severe and very severe salinization gradually shifted from the northern edge of the oasis to the eastern and southeastern desert areas during the 25 years. The saline soils with the highest salinity class were distributed in most of the desert areas in the eastern part of the Werigan–Kuqa Oasis study area as well as in smaller areas in the west in 1996, shrinking in size and characterized by a discontinuous distribution by 2021. In terms of area change, the non-salinized area increased from 198.25 in 1996 to 1682.47 km^2 in 2021. The area of saline soil with the highest salinization level decreased from 5708.77 in 1996 to 2246.87 km^2 in 2021. overall, the overall salinization of the Werigan–Kuqa Oasis improved.

Soil salinization has become one of the threats to global agricultural systems[1], and it is expected that with climate change, the impact of salinization will be wider and the degree of harm will increase, in addition, the formation mechanism of salinization is complex[2]. For regulating salinization and preventing soil degradation, it is crucial to comprehend the characteristics of salinization's spatial and temporal distribution and its evolutionary patterns[3].

Traditional laboratory analysis for soil salinity monitoring is time-consuming and labor-intensive, and because salinity changes widely across space and time, it is challenging to precisely characterize the geographical distribution of salinity and its evolutionary patterns[4]. Digital mapping has made a splash in the field of soil science, thanks to the advancement of computer hardware and software, as well as the creation of geographic information systems, global positioning systems, remote or proximity sensors, and digital elevation models that

[1]College of Geography and Remote Sensing Sciences, Xinjiang University, No. 777 Huarui Street, Xinjiang 830017 Urumqi, China. [2]Xinjiang Key Laboratory of Oasis Ecology, Xinjiang University, 830017 Urumqi, China. [3]Key Laboratory of Smart City and Environment Modelling of Higher Education Institute, Xinjiang University, 830017 Urumqi, China. ✉email: sunnyhe@xju.edu.cn

have generated huge volumes of data[5], The use of remote sensing techniques to detect salinity has increased in importance with the emergence of remote sensing satellites. Microwave and multitemporal optical remote sensing are efficient methods for identifying surface salinity parameters[6].

Various salinity indices have been constructed for modeling and prediction using the rich waveband information of optical satellites[7,8]. As in the instance of Khan et al.[9] who utilized salinity indices (SI) to categorize and analyze salinity-prone terrain, remote sensing-based salinity indices can instantly respond to the salinity status of the surface in places where it is barren or sparsely vegetated. Due to the influence of other elements including soil moisture, vegetation cover, and data collection time, it is extremely challenging to obtain pure saline spectral information in natural situations. Because salt-tolerant plants thrive in arid and semi-arid climates, vegetation index is employed as an Indirect indicator for salinity[10]. Many salinity prediction studies, such as Ramos, et al.[7] used the Canopy Response Salinity Index (CRSI), Enhanced Vegetation Index (EVI), and Normalized Difference Vegetation Index (NDVI) to assess salinity in the field; other indices widely used for salinity monitoring are Soil Adjust Vegetation Index (SAVI), Ratio Vegetation Index (RVI), and Divergence Vegetation Index (DVI), and Green Vegetation Index (GVI)[11,12].

The formation of soil salinity is highly nonlinearly related to many environmental factors, and machine learning algorithms are popular in the field of salinity research using their efficient data mining capabilities[13,14]. It has been difficult to choose the optimal model for a specific area when digitally mapping soils, but machine learning has been demonstrated to perform better than conventional statistical models at accurately predicting salinity[15,16]. The performance of various machine learning algorithms has also been compared with linear regression models and among machine learning algorithms for salting inversion analysis, including Multi-Layer Perceptron-Artificial Neural Network (MLP-ANN), Multivariate Adaptive Regression Splines (MARS), Classification and Regression Tree (CART), support vector regression (SVR), and RF. With the maturation of the ensemble learning method, it is frequently employed in picture classification research[17], nevertheless, it is not commonly used in soil salinity prediction studies. To assess the geographical variability of soil salinity and alkalinity in agricultural regions impacted by salinity, several researchers have employed random forests, with satisfactory results[18]. Recent studies that forecast salinity have employed XGBoost[19,20], while other ensemble learning techniques, including light gradient boosting machine, have seldom ever been published in the field of salinity research LightGBM[21]. Therefore, in this study, three ensemble learning models were applied to the prediction and mapping of salinity to evaluate their potential application in salinity monitoring efforts. Long-term salinity monitoring in arid and semi-arid areas is essential because it can adequately address local human-land linkages and serve as a guide for salinity control. The enormous volume of data makes the information extraction procedure in multi-temporal remote sensing challenging. Advantageously, Google Earth Engine offers a powerful data processing platform that includes a variety of geographical data, including various types of remote sensing data[22]. The spatial and spectral resolution of multispectral remote sensing is well suited for salinity monitoring due to its large coverage and ease of acquisition[6,23]. In this study, Landsat5 TM and Landsat OLI satellites were selected as the remote sensing data sources for this study because of the need to predict the salinity distribution in the inversion epoch and because of the good performance of Landsat satellites in salinity monitoring[24,25].

In this study, four years of experimental data were aggregated to make the prediction model more stable and to produce more accurate information on the spatial distribution of salinization. The specific objectives of this study were: (1) Evaluating the predictive power of RF, XGBoost, and LightGBM in ensemble learning for soil conductivity (2) Digital mapping of salinity distribution in 1996, 2006, 2017, and 2021 based on remote sensing data using an optimal prediction model; (3) The spatial and temporal variable features of salinization in Werigan–Kuqa Oasis during the last 25 years; (4) Discuss the effects of arable land expansion and land remediation on salinity.

Materials and methods

Study area. The area of study is the Werigan–Kuqa River Oasis (also known as the Werigan–Kuqa Oasis), which is situated at an altitude of 901–1069 m above sea level in the north-central Tarim Basin of the Xinjiang Uygur Autonomous Region. It has an area of around 9769.76 km^2. The Werigan–Kuqa Oasis features a typical warm-temperate continental dry climate due to its deep interior location and distance from the sea, with average annual precipitation and evaporation of 70 and 1100 mm, respectively, and a high evapotranspiration ratio of 16:1. The research region mostly consists of desert, agriculture, grassland, and woodlands, with salt- and drought-tolerant plants flourishing in the desert. Werigan–Kuqa Oasis is generally flat, with a high water table, a long dry season, and strong evaporation. In this context, salts can easily accumulate on the surface, so the area chosen as the study area is representative and has great significance for the improvement of the ecological environment and the development of agricultural production (Fig. 1).

Sample collection and survey. Field sampling and surveys of the Werigan–Kuqa Oasis are conducted annually, with most of the sampling taking place in July each year. The location of sampling points as well as the number of sampling points were determined by combining existing digital soil maps (salinity maps, soil type, soil texture) and land use/cover types, while sampling strategies were changed based on field observations from the previous year to take into account changes from year to year (Fig. 1). The location of each sampling point is recorded using a portable GPS, and the soil samples are packed in (approximately 500 g) transparent sealed bags for the next step of laboratory analysis. In this study, 4 years of soil surface (0–10 cm) electrical conductivity (EC) data were summarized and screened. The sampling times in the field were July 2006, with 36 samples; July 2017, with 84 samples; July 2018, with 75 samples, and June 2021, with 63 samples. All samples underwent air drying, grinding, homogenization, and sieving at a 0.15 mm size. For every 20 g of soil, add 100 ml of distilled

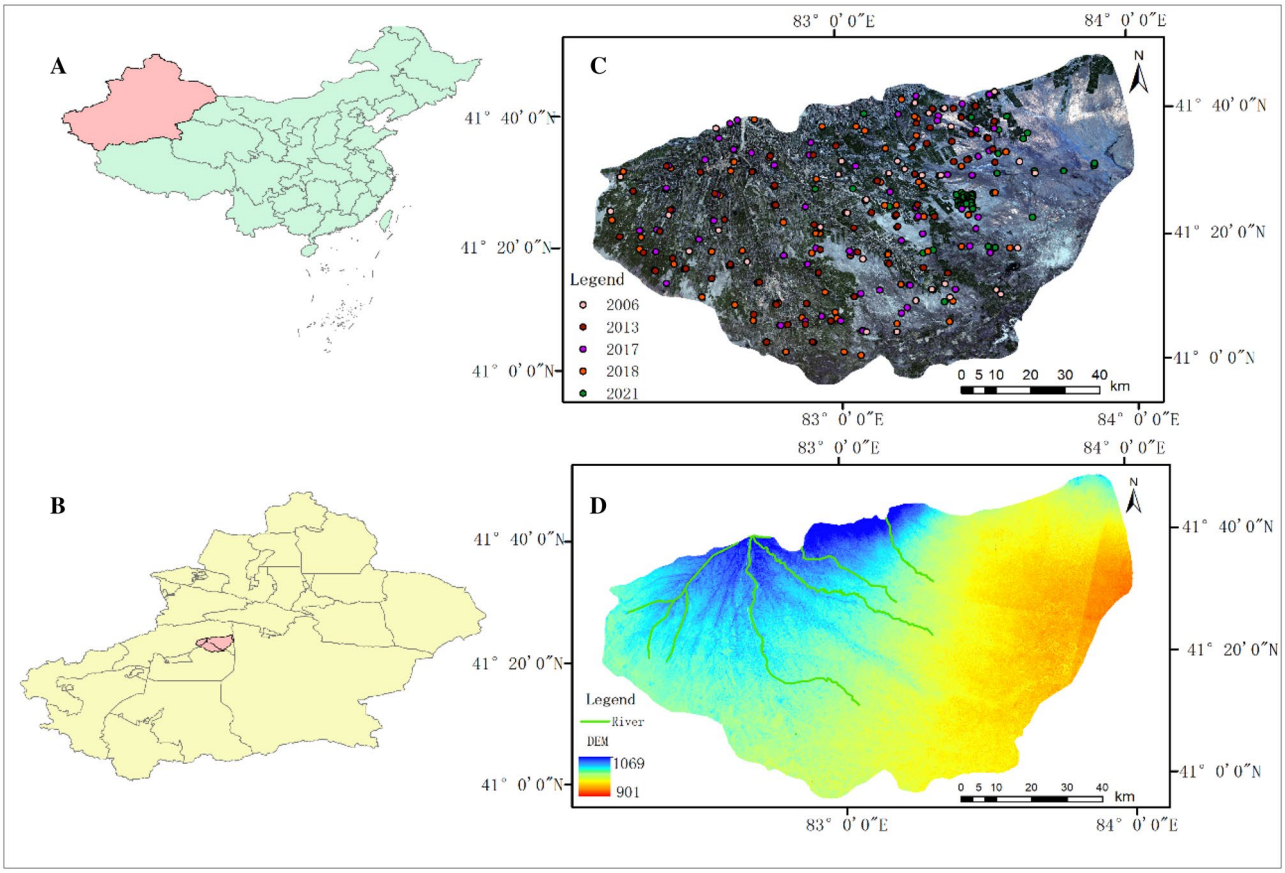

Figure 1. Figure (**A**) shows the location of Xinjiang, Figure (**B**) shows the location of the study area in Xinjiang, Figure (**C**) shows the distribution of sampling sites in the study area in different years, and figure (**D**) is the elevation of the study area.

water, mix thoroughly for 30 min, and then leave for 24 h. At room temperature of 25 °C, the soil conductivity was measured using a digital multiparameter measuring system (Multi 3420 Set B, WTW GmbH, Germany) fitted with a composite electrode (TetraCon 925)[26].

Environmental variables. The key to the selection of environmental variables is that the covariates must respond to the nature of soil formation, climate, biology and landscape type, etc. According to the SCORPAN framework (S is for soil, C is for climate, O is for organisms, R is for relief, P is for parent material, A is for age, and N is for space.)[5], a series of environmental factors were selected, including each of the original bands of Landsat5 TM and Landsat8 OLI, various indices derived from remote sensing (vegetation index, salinity index), elevation data and their derived indices (e.g. terrain moisture index, TWI).

Remote sensing-based environment variables. In this study, the remote sensing-based index extraction was done in the Google Earth Engine cloud platform. The Landsat5 TM image of July 22, 2006, and Landsat8 OLI images of July 4, 2017, July 23, 2018, and July 15, 2021, are selected, which matched the sampling time, were selected to have less than 10% cloudiness. The remote sensing-based environmental variables include 6 raw bands, 12 vegetation indices, 9 salinity indices, 1 carbonate index, and 1 brightness index (Table 1).

Terrain attributes. In this study, 11 topographic indices were generated using 30 m resolution DEM data from the Geospatial Data Cloud (http://www.gscloud.cn/), clipped, and stitched together using SAGA GIS software (Table 2). The results of Vermeulen and Van Niekerk[41] showed that the use of elevation data and its derived topographic indices as geostatistical and machine learning input variables have a great potential for salinity prediction to monitoring salt accumulation in irrigated areas.

Model framework. *Random forest.* Random Forest, developed by Breiman[42], is a popular ensemble learning algorithm based on tree-based bagging (bootstrap aggregation)[43], which has the advantage of having nonlinear mining capabilities, data distribution that does not need to conform to any assumptions, handling both rank and continuous variables, preventing overfitting, fast training, and quantitative description of the contribution of variables. RF is a bagging improvement that enhances variable selection[44], Instead of selecting the optimal split among all characteristics at each node, RF randomly picks a subset of features to decide the

Auxiliary	Index	Acronym	Formula	Reference
Vegetation indices	Normalized difference vegetation index	NDVI	(NIR−R)/(NIR + R)	27
	Generalized difference vegetation index	GDVI	(NI R^2−R^2)/(NI R^2 + R^2)	27
	Normalized difference vegetation index	GNDVI	(NIR−G)/(NIR + G)	28
	Green ratio vegetation index	GRVI	NIR/(G−1)	29
	Optimizied soil adjusted vegetation index	OSAVI	(NIR−R)/(NIR + R + θ)	30
	Ratio vegetation index	RVI	NIR/R	31
	Soil adjusted vegetation index	SAVI	((NIR−R)/(NIR + R + L))*(1 + L)	32
	Brightness index	BRI	$(G^2 + R^2)^{0.5}$	33
	Carbonate index	CAEX	B/G	33
	Canopy response salinity	CRSI	$(((NIR*R)-(G*B))/((NIR*R)+(G*B)))^{0.5}$	34
	Difference vegetation index	DVI	NIR−R	35
	Enhanced vegetation index	EVI	2.5*(NIR−R)/(NIR + 6R−7.5B + 1)	36
	Green atmospherically resistant vegetation index	GARI	(NIR−(G + y*(B-R)))/(NIR + (G + y*(B + R)))	37
	extented EVI	EEVI	2.5*(NIR + SWIR1−R)/ (NIR + SWIR1 + 6*R−7.5*B + 1)	38
Soil-related indices	Salinity index	SIT	(R/NIR)*100	39
	Salinity index	SI	(R−NIR)/(R + NIR)	39
	Salinity index	SI1	$(R*G)^{0.5}$	39
	Salinity index	SI2	$(NIR^2 + R^2 + G2)^{0.5}$	39
	Salinity index	SI3	$(R^2 + G2)^{0.5}$	39
	Salinity index	SI4	(R*NIR)/G	40
	Salinity Ratio index	SAIO	(R−NIR)/(G + NIR)	33
	Salinity index	SIA	(B/R)	39
	Salinity index	SIB	(B−R)/(B + R)	33

Table 1. Auxiliary data based on remote sensing.

Auxiliary	Index	Acronym	Reference
Dem derivatives	Aspect	ASP	SAGA GIS
	Convergence index	CI	
	LS-factor	LSF	
	Relative slope position	RSP	
	Slope	Slope	
	Topographic wetness index	TWI	
	Valley depth	VD	
	DEM	DEM	
	Channel network distance	CND	

Table 2. Terrain attributes.

split, this makes RF more resilient to noise and less prone to overfitting. In addition, RF can handle outliers very well[45]. The number of trees and predictor variables that the random forest model allows the decision tree to grow as large as it can without being trimmed is its critical factor. The primary hyperparameters modified in this study are the number of trees in the forest and the number of features thought to divide at each leaf node[46]. In this work, we used the open-source machine learning package Scikit-learn to create an RF mode[47].

Extreme gradient boosting. Extreme Gradient Boosting (XGBoost) is a popular boosting-based ensemble machine learning algorithm[48], this algorithm was used in the Kaggle signal recognition competition and has attracted a lot of attention for its outstanding efficiency and high prediction accuracy[49]. Boosting, in contrast to bagging, is an iterative method that successively adds new trees to the integration, and samples erroneously predicted by the prior tree are given higher weights in the succeeding trees. Thanks to numerous significant systematic and algorithmic enhancements, the gradient boosting framework is implemented effectively and flexibly in XGBoost[49,50]. The number of gradients boosting trees (n_ estimators), learning rate (eta), maximum depth of the tree (max_depth), and column per level of the subsample ratio are some of the important hyperparameters that are tuned by XGBoost. To train XGBoost models, the open-source Scikit-Learn software is utilized.

Light gradient boosting machine. Light Gradient Boosting Machine (LightGBM) is a framework that implements the idea of GBDT (Gradient Boosting Decision Tree) algorithm[51], a boosting decision tree tool open-sourced by the Microsoft DMTK team, which has fast training speed and less memory usage, which greatly speeds up the training and also has better model accuracy. LightGBM performs the following optimizations on the traditional GBDT algorithm: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bunding (EFB)[51]. GOSS is a subsampling technique used to create training sets to build the base tree in the integration, select data with larger gradients from the sample to increase their contribution to the computed Information gain, and EFB merges certain data features to reduce the data dimensionality[52]. Generally, the prediction accuracy is significantly influenced by the hyperparameters[53]. So, before employing LightGBM, we need the first figure out how many and how widely its hyperparameters may vary. The number of Leaves, Learning Rate, and Maximum Depth is the important factors.

For this experiment, the above three models were done in the Spyder platform based on the Python 3.9.7 programming language.

Model parameter optimization. The efficacy of the model application depends on the choice of model parameters. In the fields of statistical analysis and machine learning, the K-Fold cross-validation method is frequently used to assess the generalizability of models. The grid search method is an exhaustive search method that specifies the values of the parameters, it is carried out by Scikit-GridSearchCV, learn's which arranges and combines the possible values of each parameter, lists all combinations that could exist, and performs cross-validation to optimize the estimation function's parameters in order to obtain the best learning algorithm[54]. The minimum value of Root Mean Square Error (RMSE) is used as the criterion for the selection of model parameters, In this experiment, it is assumed that the value of K is 5, as follows:

1. Divide the dataset into the training set, test set, and K-fold division of the training set data.
2. Determine the range of each parameter of the model, taking a random forest as an example, and determine the number of decision trees m as well as the depth h. The combination of parameters is the cross nodes of a two-dimensional grid with m and has horizontal and vertical axes.
3. Choose any K-1 data from the training set, choose a set of cross-node parameters, create one decision tree using a sample of all the K-1 data, forecast the final 1 data, and compute the average root mean square error of all trees on the final 1 training sample.
4. Repeat the above two steps until you have traversed K-1 copies of the data.
5. Iterate through the parameter combinations of all crossover nodes of the grid. 6.
6. Steps 3 to 5 are repeated, using cross-validation to calculate the performance of the model in the test dataset. (Table 3) shows the combination of model parameters optimized by grid search.

Evaluation of prediction accuracy. In this research, the coefficient of determination R^2, the root mean square error(RMSE), and the performance to interquartile distance(RPIQ) are used to assess the performance of RF, XGBoost, and Lightgbm. The closely R^2 is to 1, the more accurate models are fitted. The closer the number is to 0, the smaller the difference between the measured value and the predicted value of the model, and the greater the model's ability to forecast the future. The value of RMSE is inversely related to the accuracy of the model. RPIQ is the interquartile range to RMSE ratio, and the interquartile range is the difference between 75 and 25% of the sample values. It is commonly accepted that RPIQ < 1.7 implies low model prediction dependability, $1.7 \le \text{RPIQ} \le 2.2$ suggests somewhat balanced prediction ability, and RPIQ ≥ 2.2 indicates highly strong prediction ability. RPIQ is a more reasonable and objective measure when compared to the Ratio of Performance to Deviation (RPD), especially for soil samples with an unusual distribution[55,56]. Equations (1)–(3) show the expression of these model evaluation metrics:

$$R^2 = \frac{\sum_{i=1}^{n} \left(X_i^* - Y_i^*\right)^2}{\sum_{i=1}^{n} (X_i - Y_i)^2} \tag{1}$$

	n_estimators	max_depth	learning_rate	Subsample	colsample_bytree
RF	10	10	Null	Null	Null
XGBoost	43	4	0.1	0.5	0.9
LightGBM	22	4	0.2	0.5	0.9

Table 3. Combination of parameters for different models.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (2)$$

$$RPIQ = \frac{\Delta Q}{RMSE} \quad (3)$$

where N is the number of samples, X_i is the measured EC value, Y_i is the calculated value, X_i^* is the mean measured EC value, Y_i^* is the estimated soil EC value, SD represents the standard deviation, and ΔQ is the interquartile distance (IQR), which is the difference between the upper quartile ($Q3$) and the lower quartile ($Q1$).

Soil EC prediction and mapping for different years. The flow of this experiment is shown in (Fig. 2). The Google Earth Engine cloud platform was used to calculate and obtain the remote sensing-based environmental variables corresponding to the sampling time to establish a soil EC prediction model. Since the sampling time is mainly concentrated in July, based on the optimal model, the spatial distribution maps of soil EC in July of each year in 1996, 2006, 2017, and 2021 are obtained (the remote sensing data of June 24 is chosen because the remote sensing Image of July 1996 Is too cloudy to meet the mapping requirements), and this step is done by using the Spyder development environment with the help of GDAL, Pandas and other libraries to complete the mapping.

Results

Soil EC descriptive statistics. In this experiment, the final data of 258 soil EC samples were obtained after the outliers were removed from the sample data. Following statistical analysis, the soil's electrical conductivity (EC) minimum, maximum, mean, standard deviation, coefficient of variation, kurtosis, and skewness were determined (Table 4).

Soil EC values in the Werigan–Kuqa Oasis ranged from 0.079 dS m^{-1} to 143.4 dS m^{-1}, showing that the samples had a high span. The skewness of 1.37 is much higher than 0, which indicates that the sample data do not obey a normal distribution. The standard deviation was 33.2 dS m^{-1} and the coefficient of variation was 1.19, which is greater than 1, thus belonging to strong variability, which is consistent with the study of Wang, et al.[40], showing the high spatial variability of soil EC values in the Werigan–Kuqa Oasis area.

Correaltion analysis. In modeling soil salinity monitoring, not all environmental variables are involved in modeling and there are differences in their contribution to EC prediction[40], therefore, it is necessary to screen the environmental variables. Based on the statistical analysis of the sample EC values, the skewness was 1.47

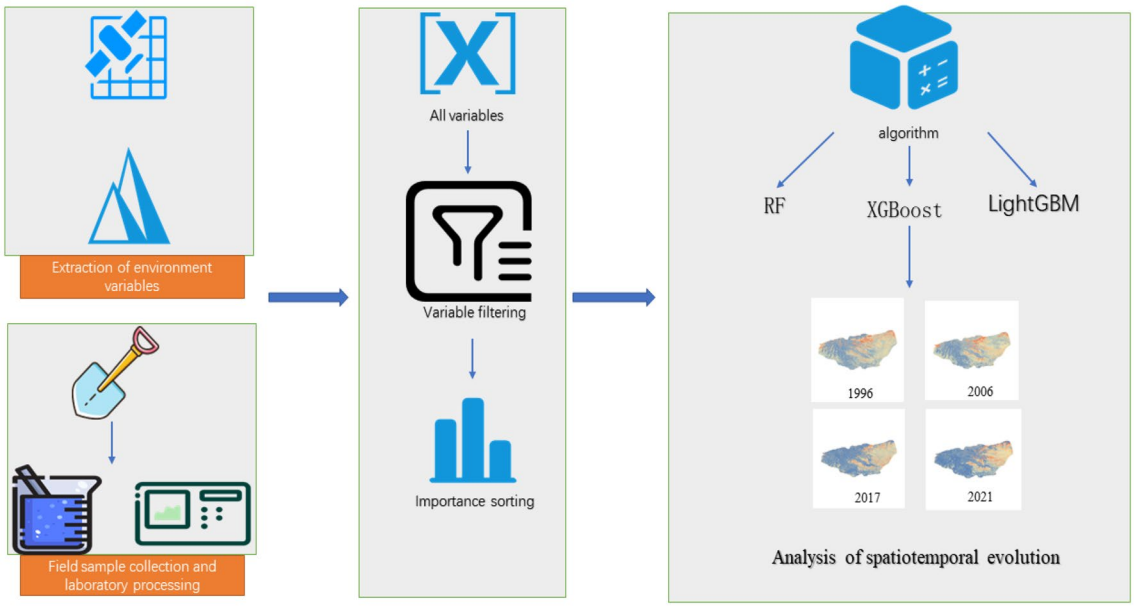

Figure 2. Flow chart.

EC sample data	Max(dS m^{-1})	Min(dS m^{-1})	Mean(dS m^{-1})	Std.D(dS m^{-1})	CV (%)	Kurtosis	Skewness
Whole data(n = 258)	143.4	0.079	27.9	33.2	1.19	1.15	1.37

Table 4. Soil EC descriptive statistics.

(Table 4), so Spearman correlation analysis was used in the analysis of the relationship between environmental variables and soil EC values. In this study, 38 environmental variables (original band, vegetation index, salinity index, topography index, etc.) were initially selected, and after Spearman correlation analysis, 31 environmental variables were selected and the remaining relevant variables were not significantly correlated (Table 5).

Among the raw bands of remote sensing, the correlations with soil EC were NIR (R = − 0.610), SWIR2 (R = 0.423), Red (R = 0.372), SWIR1 (R = 0.3), and Green (R = 0.246) in descending order. Salinity indices, as direct indicators in salinity monitoring[57], showed good correlation with soil EC, and all nine selected salinity indices were significantly correlated with EC values, with correlation coefficients up to 0.531(SIA, SIB, SIT, SAIO are all salinity indices, which are different combinations of different waveforms), The correlation between vegetation index and soil EC values in descending order is, GARI (R = − 0.626), EVI (R = − 0.596), DVI (R = − 0.572), GDVI (R = − 0.541), OSAVI (R = − 0.541),RVI (R = − 0.534), NDVI (R = − 0.533), SAVI (R = − 0.550), CRSI (R = − 0.506), GRVI (− 0.469), GNDVI (R = − 0.468), it can be seen that the vegetation index is a good Indicator as an Indirect Indicator of salinity monitoring. Compared to NDVI, SAVI increases the vegetation signal and decreases the soil background, therefore, there is a strong correlation with soil EC (R = − 0.55), in addition, OSAVI has the same correlation as SAVI, but OSAVI avoids the complex calculation of soil baseline parameters. Among the topographic correlation factors, the higher correlation is with DEM (R = − 0.463), followed by CND (R = − 0.175), and finally RSP (R = − 0.174). The lower correlation between topography and Its Indices with EC Is explained by the overall flatness of the Werigan–Kuqa Oasis. Finally, the carbonate index CAEX correlated significantly (R = 0.612) with soil EC values, which were determined by the soil properties of the study area.

Importance of selected environmental covariates. Different environmental factors have different predictive contributions to soil EC in predictive models, and not all environmental factors are significant variables in the modeling[58], so it is necessary to rank the importance of environmental variables, and this study will rank the importance of features using each of the three models themselves and observe the differences in the contribution of variables in the three models.

Figures 3, 4, 5 show the results of the three models for feature selection, the degree of contribution of the variables differed, but individual variables showed high contribution in all three models, and among the vegetation indices, most of them generally contributed well, with CRSI being the most stable and showing high contribution in all three models, in agreement with Scudiero et al.[34] and Wu et al.[59], GARI performed best among all environmental variables involved in RF. Remote sensing primitive bands are pivotal in the participation in modeling, in the study of related scholars, the relationship between each band and saline soils was analyzed in detail, the greater the salt in the soil, the higher the reflectance of all TM spectral bands[59] and the spectral reflectance of $CaCO_3$, $CaSO_4 \cdot 2H_2O$, and gypsum sand were analyzed in the laboratory, they concluded that salt minerals can be detected when they are the main soil component[60], among the primitive bands involved in modeling, the NIR band stands out, especially in the participation in the random forest modeling process, the contribution is second only to GARI. The salinity index stands out as a direct indicator in sparsely vegetated areas, and the SIA performed consistently in this study in terms of contribution across the three prediction models. The salinity index integrates most of the soil properties affected by salinity, and the salinity index is also very cost-effective for possible large-scale surveys to prevent soil salinity at the landscape scale[57].

Prediction accuracy. In this experiment, two approaches are used for model validation, the validation approach of slicing the dataset into training and test sets, and the cross-validation approach (Table 6, Fig. 6), and it was found that the R^2 value of XGBoost was the highest among the three models in both the training and test sets, 0.84, 0.73, respectively, and the RMSE value was also the lowest in the training and test sets, 13.57 dS m^{-1}, 17.62 dS m^{-1}, respectively. The RPIQ value is also the highest, 3.32 in the training set and 2.45 in the test set. When RPIQ ≥ 2.2, it means that the model achieves excellent prediction, and compared with the performance of RF and LightGBM models in the test set (2.39 and 2.32, respectively), XGBoost has excellent prediction ability. Similarly, XGBoost has the lowest RMSE value of 19.9 dS m^{-1} for the three models after tenfold cross-validation.

Factors	R	Factors	R	Factors	R
EEVI	− .219**	SI	.531**	BRI	.315**
Green	.246**	SI1	.320**	CAEX	.612**
Nir	− .610**	SI2	− .308**	CRSI	− .506**
Red	.372**	SI3	.324**	DVI	− .572**
Swir1	.300**	GDVI	− .541**	EVI	− .596**
Swir2	.423**	GNDVI	− .468**	GARI	− .626**
SI4	− .315**	GRVI	− .469**	RSP	− .174**
SIA	− .531**	NDVI	− .533**	DEM	− .463**
SIB	− .531**	OSAVI	− .541**	CND	− .175**
SIT	.531**	RVI	− .534**		
SAIO	.525**	SAVI	− .550**		

Table 5. The correlation between the variables and soil EC(0–10 cm). **Significant $p < 0.01$; *Significant $p < 0.05$.

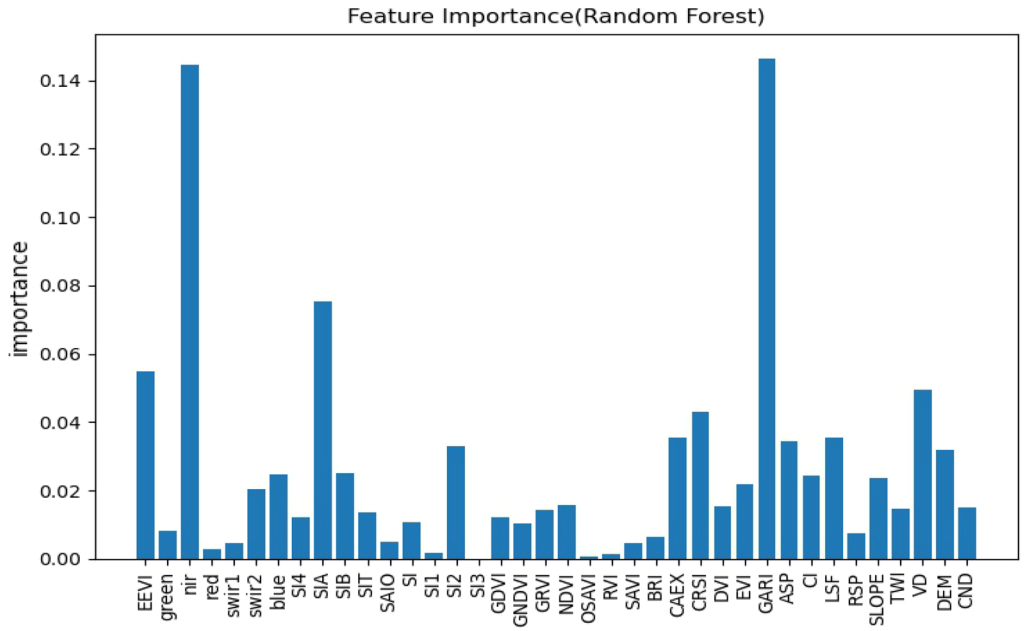

Figure 3. Characteristic importance diagram of RF.

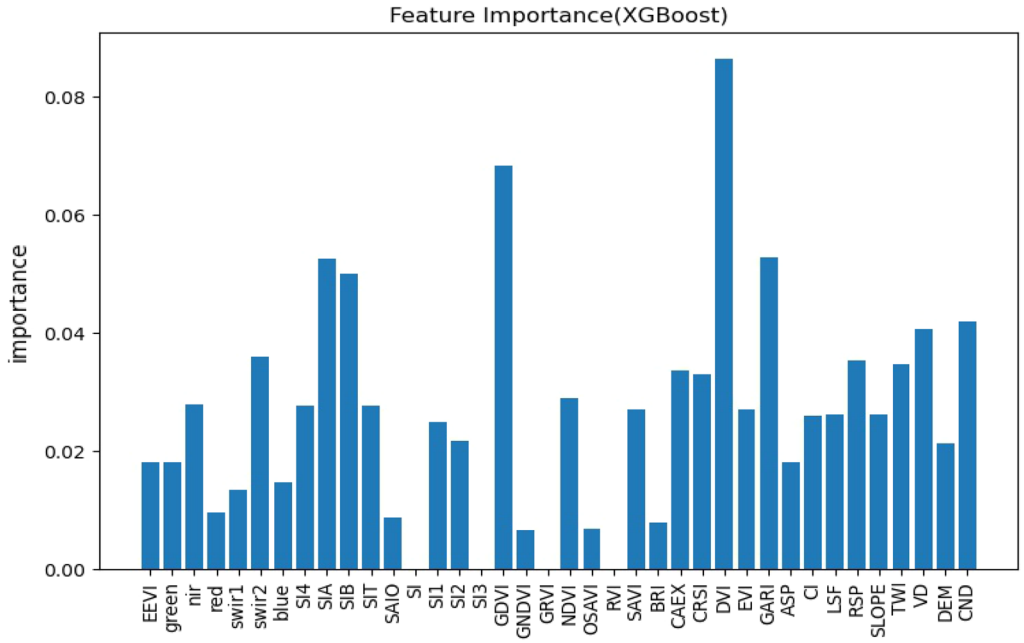

Figure 4. Characteristic importance diagram of XGBoost.

Therefore, XGBoost will be used as the optimal model for the digital mapping of the spatial distribution of salinity.

Spatial and temporal distribution characteristics and evolutionary trends of Salinization in 1996, 2006, 2017, and 2021. In the research region, all soil samples were divided into six groups by the frequently used soil salinity classification method for further analysis and visualization[61] (Table 7), and the spatial distribution of soil salinization in the Werigan–Kuqa Oasis on August 11, 1996, July 22, 2006, July 4, 2017, and July 15, 2021, were inverted using the selected optimal model and the corresponding optimal variables (Fig. 7). To further verify the accuracy of the salinity spatial distribution map after reclassification, this experiment used the 2017 and 2021 sample points as the validation set, and the accuracy was verified using the confusion matrix

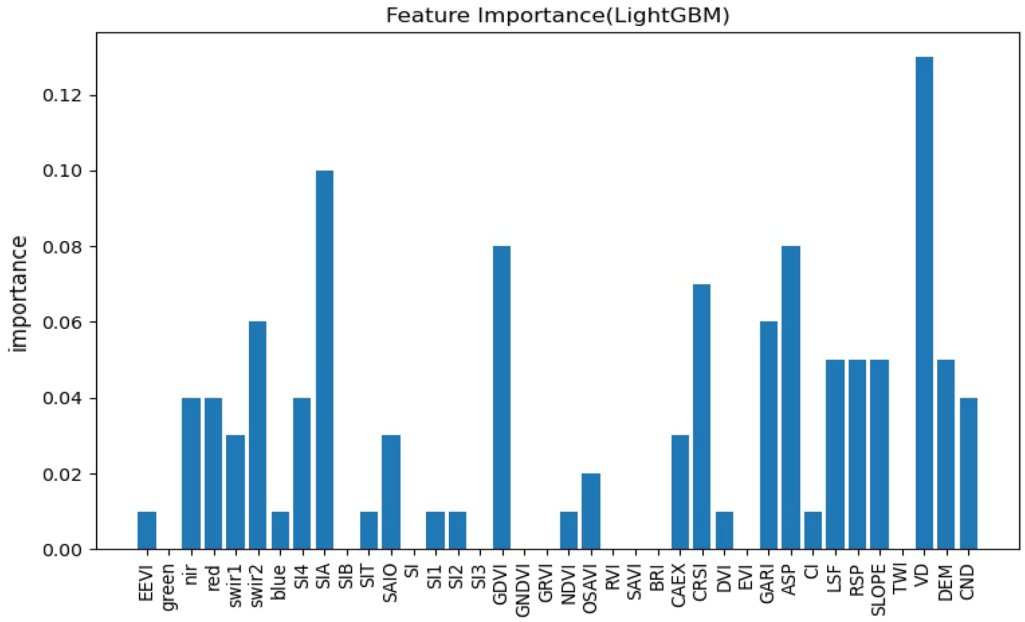

Figure 5. Characteristic importance diagram of LightGBM.

Algorithm	Calibration			Validation			Cross-validation	
	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ	R^2	RMSE
RF	0.82	14.3	3.15	0.69	18.19	2.39	0.6	20.29
XGBoost	0.84	13.57	3.32	0.73	17.62	2.45	0.6	19.9
LightGBM	0.71	18.57	2.43	0.64	18.58	2.32	0.57	20.1

Table 6. The performance of each of the three models in the validation set and training set.

and kappa coefficient (Fig. 7), and the kappa coefficient was obtained as 0.71, which indicates that the salinity map has a high degree of consistency.

According to (Fig. 8), the spatial distribution of salinization in the Werigan–Kuqa Oasis shows a distribution characteristic of good in the west and north and severe in the east and south. The moderate and below salinization in the Werigan–Kuqa Oasis is distributed in the west and north of the Werigan–Kuqa Oasis, an oasis area with good irrigation conditions (Fig. 1), where the main feature type is arable land, the terrain is relatively high, not easily waterlogged, and the vegetation cover is relatively high. With the expansion of the spatial extent of arable land, light salinization and below also show a corresponding radial change to the south, southwest, and southeast, and become more continuous spatially. By 2021, on the western and southern edges of the Werigan–Kuqa Oasis, very heavy salinization has been transformed into light salinization, in the eastern and northeastern regions, spatially discontinuous new arable land emerged, so that mild salinization also took the form of sporadic spatial distribution.

Severe and very severe salinization was mainly distributed in the northern part of the Werigan–Kuqa Oasis in 1996, and by 2006, salinization in the region improved and gradually shifted to the east and south, developing to the southeast by 2021. The development trend of severe and very severe salinization over 25 years is closely related to the low southeast and high northwest topography of the Werigan–Kuqa Oasis (Fig. 1).

The most pronounced spatial distribution and evolutionary characteristics of saline soils with the highest degree of salinization were mainly distributed in the southwestern edge of the Werigan–Kuqa Oasis and most of the desert areas in the east in 1996, shifting to classes such as severe and very severe in 2017, and improving significantly by 2021, especially in the eastern desert areas. Relying on years of field surveys, it was found that sparse salt vegetation grows in the eastern part of the Werigan–Kuqa Oasis, while the southeastern part of the area is sparsely forested. As a result of enhanced vegetation protection efforts in the eastern area, the vegetation cover has increased significantly and, therefore, the evaporation of surface water has decreased accordingly, reducing the rate of salt accumulation on the surface.

Change in area of salinization at different levels. As shown in (Fig. 9), the non-salinized area of the Werigan–Kuqa Oasis is 198.25 km^2 in 1996 and 1682.47 km^2 in 2021, an increase of 748.6%; Mild salinization

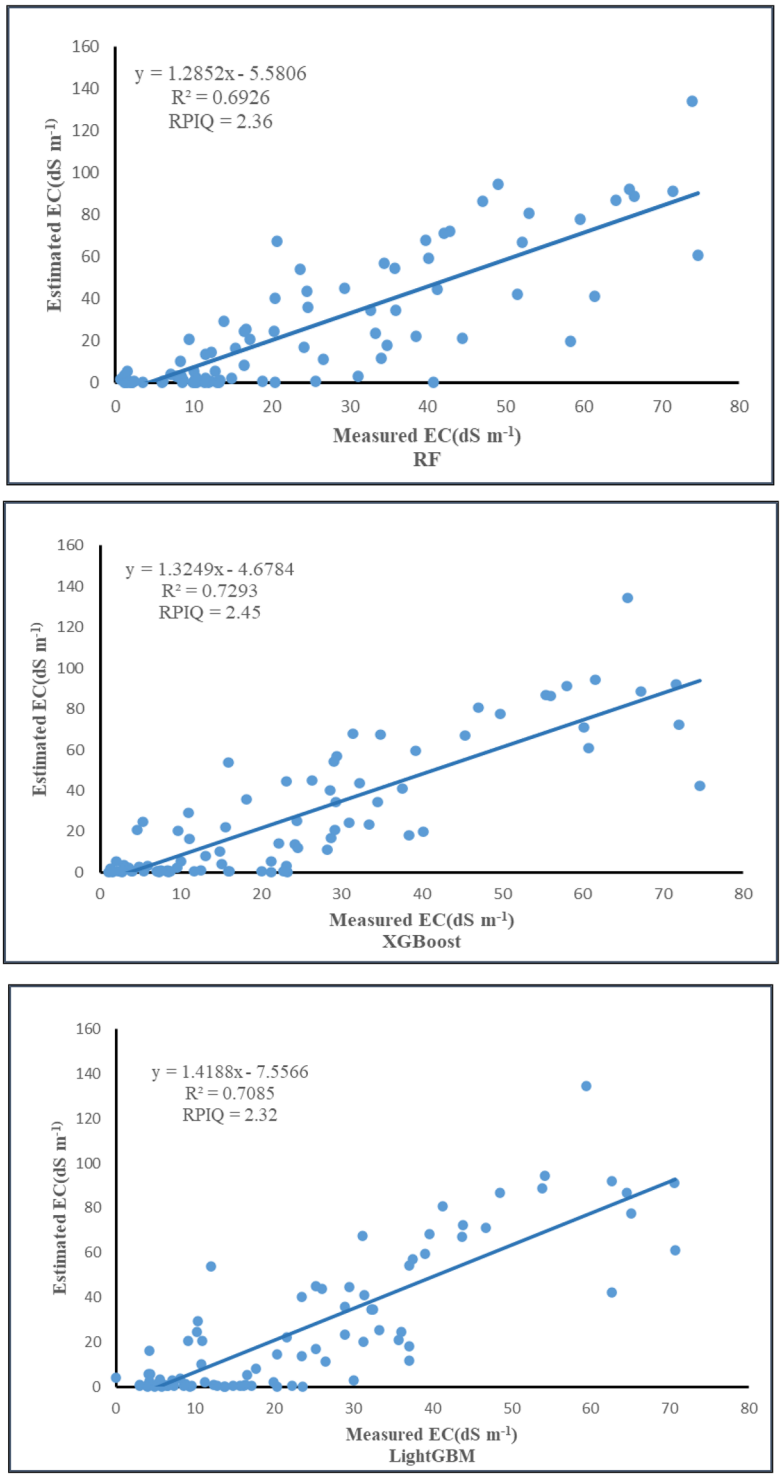

Figure 6. Measured and predicted regression analysis of the three models.

was 346.78 km^2 in 1996 and increased year by year since then to 1441.29 km^2 in 2021, an increase of 315.6% compared to 1996; Moderate salinization remained stable from 1996 to 2006 and increased substantially by 2017 to 1062.26 km^2 by 2021, an increase of 134.8% compared to 1996; Heavy salinization was 431.26 km^2 in 1996 and 838.132 km^2 in 2021; Very heavy salinization remains relatively stable from 1996 to 2021, with an area of 2498.74 km^2 by 2021; The area of saline soil was 5708.77 km^2 in 1996, then declined to 5168.7 km^2 in 2006, followed by a greater decline to 794.48 km^2 in 2017 and 2246.87 km^2 in 2021, a decrease of 60.6% compared to 1996. Based on the results of the above statistical analysis: during the last 25 years, the non-salinized, lightly salinized, and moderately salinized areas increased more, the saline soil area decreased more, and the heavy and

Salinity constraint	EC(dS m^{-1})
Non-salinization	4 >
Mild salinization	4–8
Moderate	8–12
Heavy salinization	12–16
Extremely high	16–32
Saline soil	32 <

Table 7. Grades or classes of soil salinity.

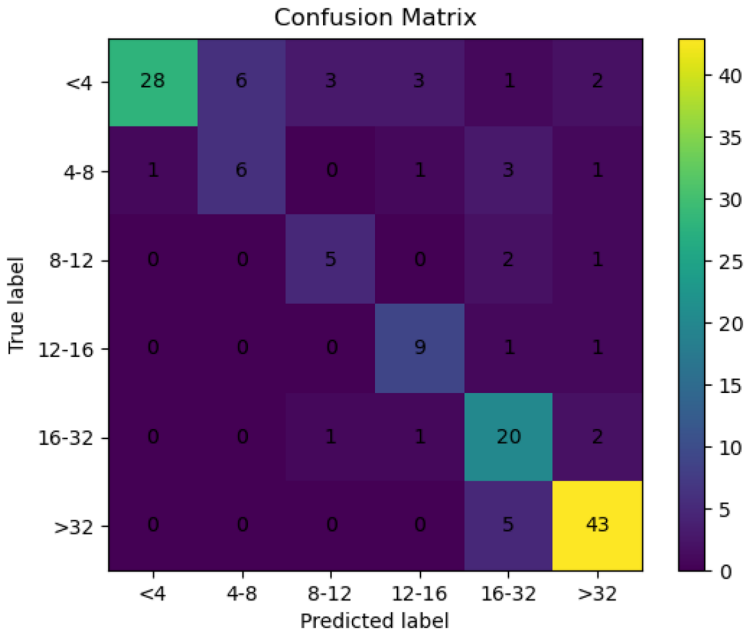

Figure 7. Confusion matrix verification.

very heavy salinization changed less and remained stable, so there was an improvement of soil salinization in the Werigan–Kuqa Oasis.

Discussion

Long-time series of salinity monitoring. Various multispectral sensors rely on the spectral reflectance properties of the ground for ground monitoring[62], and the spectral reflectance varies for different levels of salinity, often with a white salt crust attached to the ground surface in highly saline areas. The higher the salinization, the higher the spectral reflectance of each band will increase accordingly[13], therefore, it becomes possible to monitor salinization using raw bands or derived spectral indices of remote sensing. In previous studies on salinity monitoring, the choice of environmental variables varied, such as direct use of salinity indices for estimating soil salinity[63], indirect estimation of soil salinity using vegetation indices[64], or combining multiple environmental variables and grouping them to predict comparisons[58].

The objective of this study is to map the spatial distribution of salinization in the Werigan–Kuqa Oasis in different years and analyze the changing trend of salinization area in different grades. Therefore, remote sensing data that can match the sampling time in different years are selected, and a stable soil EC prediction model is established based on the extraction of environmental variables from remote sensing images, which makes it possible to accomplish the goal of salinization spatial distribution mapping realistically and accurately and provide data reference for salinization management and water resources management. The earliest data collection in this study area began in 2006, so in this modeling, sample data from 2006, 2017, 2018, and 2021 were ensemble for modeling, making full use of the available laboratory data. This study utilizes the Google Earth Engine platform for fast online computational processing. Therefore, the remote sensing cloud platform presented by Google Earth Engine is an excellent option for environmental monitoring research that uses lengthy time series of remote sensing data.

Spatial and temporal evolutionary characteristics of salinization. The distribution of saline salinization in the Werigan–Kuqa Oasis shows distinct regional characteristics. In the southeast and east of

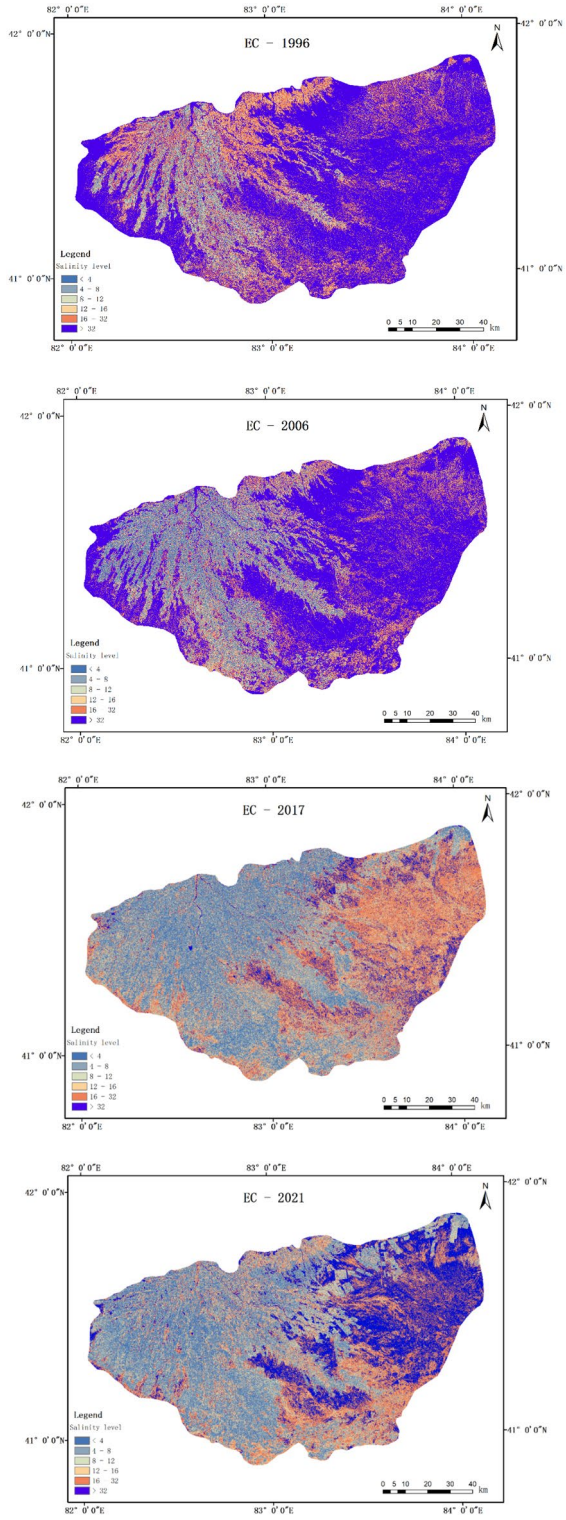

Figure 8. Spatial distribution of soil salinization in 1996, 2006, 2017 and 2021.

the Werigan–Kuqa Oasis, which is the most affected area by salinization, salinization of very severe and higher grades is distributed, and the spatial and temporal evolution characteristics are obvious. The low elevation compared to other areas of the Werigan–Kuqa Oasis (Fig. 1D) makes it possible to distribute high concentrations of salts in this area[40]. After years of field investigation and sampling, seasonal floods often gather in this area, and according to Ding and Yu[4], it was found that the salts accumulated on the surface of the area do not drain outward, which makes it more difficult to manage salinization. In addition, the area is dominated by sandy soils, and during the dry season, salts are easily deposited on the surface after water evaporation[4]. During the

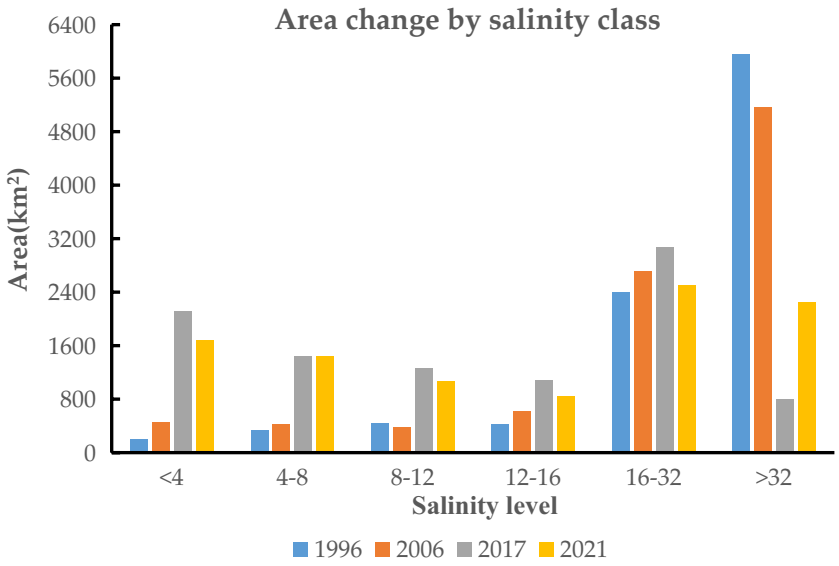

Figure 9. Trends in the area of different levels of salinization.

25 years, salinization in the eastern part of the Werigan–Kuqa Oasis has improved significantly because the local government has strengthened the vegetation protection of the desert, and built alkali drainage canals in the sparsely vegetated areas of the desert to reduce seasonal waterlogging to a certain extent, and strictly monitored overgrazing practices, so that the vegetation coverage and the area covered by the area have gradually increased, and therefore the area of very heavy salinization in the area has decreased in recent years.

In the southeast of the Werigan–Kuqa Oasis fringe area, salinization of severe and higher grades is distributed and has not improved significantly in individual areas during the last 25 years, which is since the economy of the study area is dominated by irrigated agriculture and surface irrigation is a common irrigation method, and the salts in the soil inside the Werigan–Kuqa Oasis are transported to the downstream through surface irrigation water, which deposits salts on the downstream surface and eventually intensifies the formation of salinization This is the reason why salinization is higher at the edge of the oasis than in the interior of the oasis[4].

The salinization of moderate and lower grades is distributed in the interior of the oasis. Since the economy of the study area is based on irrigated agriculture, especially in the western and southwestern regions of the study area, which are more dependent on this economic activity, the formation of mild salinization in the region is strongly related to agricultural irrigation, while the irrigation of the regional arable land is gradually changing from the previous surface irrigation to drip irrigation, which may aggravate salinization in the region. The spatial and temporal evolution of salinization within the oasis is also more pronounced during the 25 years, due to the expansion of the arable land area, which increases significantly by 2021 compared to 1996, especially in the southwest and northeast of the study area, and therefore, the salinization grade changes accordingly, from severe and above grade to moderate and below, and to ensure healthy crop survival, before planting The land is drained of alkali to ensure healthy crop survival. In addition, the salinization of arable land areas tends to be consistent, and the area of salinization of heavy and above grades is reduced and fragmented, because the local government has been carrying out comprehensive land improvement work, leveling dry land and barren land; renovating and reinforcing branch canals and field branch; building rural field roads less than 4.5 m, serving production and travel, especially since 2018, the local government has carried out the construction of high-standard farmland, making the land more flat and contiguous, with better agricultural facilities, more fertile land and better disaster resistance. The results of the study show that human activities are the key factors affecting the aggravation and management of salinization[58], and the key lies in whether humans destroy or protect land and water resources, and as the core area of the Belt and Road, it should focus on the protection of the ecological environment, and its starting point should be the management of salinization in arid areas. The irrational use of water resources is related to the salinity of the soil[65], so in the future, we should discuss the planting pattern of the Werigan–Kuqa Oasis and a more economical and efficient irrigation method. It is gratifying to note that the government has in recent years become more disciplined in water resources management, such as the implementation of the river chief system, which strictly regulates the reckless diversion of rivers; the implementation of the water station chief system in irrigation areas, which provides more precise and efficient control of irrigation water resources; and the implementation of the forest chief system, which increases the protection of forest land. Through these measures, the salinization of the Werigan–Kuqa Oasis has been improved.

Conclusions

This study uses multi-year field collection data and multi-source data with the help of the ensemble learning method and Google Earth Engine cloud platform to complete the digital mapping of salinity spatial distribution in 1996, 2006, 2017, and 2021, analyze the spatial and temporal evolution characteristics and driving factors of salinity in Werigan–Kuqa Oasis, and draw the following conclusions:

(1) Among the three ensemble learning models, RF, XGBoost, and LightGBM, XGBoost had an RMSE of 17.62 dS m^{-1}, R^2 of 0.73, and RPIQ of 2.45 in the test set, which had higher prediction accuracy compared with the other two models, and more accurate salinization distribution maps were obtained using XGBoost.
(2) The salinization in the study area generally shows the distribution characteristics of good in the west and north and severe in the east and south. The moderate and below salinization is distributed in the oasis areas with good irrigation conditions and smooth drainage. And severe and above salinization is mainly distributed in the desert areas in the east and southeast.
(3) The spatial and temporal variation of salinization in the study area has changed significantly in the last 25 years, with non-salinization and light salinization expanding in the east and southwest spatial distribution with the increase of arable land area and effective remediation planning of arable land. The distribution area of salinization of severe and above grades has shrunk more significantly.

Data availability

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Received: 13 November 2022; Accepted: 6 January 2023
Published online: 16 February 2023

References

1. Singh, A. Soil salinization management for sustainable development: a review. *J. Environ. Manag.* **277**, 111383 (2021).
2. Hassani, A., Azapagic, A. & Shokri, N. Global predictions of primary soil salinization under changing climate in the twenty first century. *Nat. Commun.* **12**, 6663. https://doi.org/10.1038/s41467-021-26907-3 (2021).
3. Hassani, A., Azapagic, A. & Shokri, N. Predicting long-term dynamics of soil salinity and sodicity on a global scale. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 33017–33027. https://doi.org/10.1073/pnas.2013771117 (2020).
4. Ding, J. & Yu, D. Monitoring and evaluating spatial variability of soil salinity in dry and wet seasons in the Werigan–Kuqa Oasis, China, using remote sensing and electromagnetic induction instruments. *Geoderma* **235–236**, 316–322. https://doi.org/10.1016/j.geoderma.2014.07.028 (2014).
5. McBratney, A. B., Mendonça Santos, M. L. & Minasny, B. On digital soil mapping. *Geoderma* **117**, 3–52. https://doi.org/10.1016/s0016-7061(03)00223-4 (2003).
6. Metternicht, G. I. & Zinck, J. A. Remote sensing of soil salinity: potentials and constraints. *Remote Sens. Environ.* **85**, 1–20. https://doi.org/10.1016/s0034-4257(02)00188-8 (2003).
7. Ramos, T. B. *et al.* Soil salinity assessment using vegetation indices derived from Sentinel-2 multispectral data. Application to Lezíria Grande Portugal. *Agricultural Water Management* **241**, 106387. https://doi.org/10.1016/j.agwat.2020.106387 (2020).
8. Wang, J. *et al.* Soil salinity mapping using machine learning algorithms with the Sentinel-2 MSI in Arid areas, China. *Remote Sens.* **13**(2), 305. https://doi.org/10.3390/rs13020305 (2021).
9. Khan, N. M., Rastoskuev, V. V., Sato, Y. & Shiozawa, S. Assessment of hydrosaline land degradation by using a simple approach of remote sensing indicators. *Agric. Water Manag.* **77**, 96–109 (2005).
10. Zhao, W., Zhou, C., Zhou, C., Ma, H. & Wang, Z. Soil salinity inversion model of oasis in arid area based on UAV multispectral remote sensing. *Remote Sens.* **14**, 1804 (2022).
11. Allbed, A. & Kumar, L. Soil salinity mapping and monitoring in arid and semi-arid regions using remote sensing technology: a review. *Adv. Remote Sens.* **02**, 373–385. https://doi.org/10.4236/ars.2013.24040 (2013).
12. Peng, J. *et al.* Estimating soil salinity from remote sensing and terrain data in southern Xinjiang Province China. *Geoderma* **337**, 1309–1319. https://doi.org/10.1016/j.geoderma.2018.08.006 (2019).
13. Wang, J. *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* **707**, 136092. https://doi.org/10.1016/j.scitotenv.2019.136092 (2020).
14. Hoa, P. V. *et al.* Soil salinity mapping using SAR Sentinel-1 data and advanced machine learning algorithms: a case study at ben Tre province of the Mekong river delta (Vietnam). *Remote Sens.* **11**, 128 (2019).
15. Zhou, T., Geng, Y., Chen, J., Pan, J. & Lausch, A. High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms. *Sci. Total Environ.* **729**, 138244 (2020).
16. Lu, H., Yang, L., Fan, Y., Qian, X. & Liu, T. Novel simulation of aqueous total nitrogen and phosphorus concentrations in Taihu Lake with machine learning. *Environ. Res.* **204**, 111940 (2022).
17. Zhang, E., Zhang, X., Jiao, L., Li, L. & Hou, B. Spectral–spatial hyperspectral image ensemble classification via joint sparse representation. *Pattern Recogn.* **59**, 42–54 (2016).
18. Nabiollahi, K. *et al.* Assessing agricultural salt-affected land using digital soil mapping and hybridized random forests. *Geoderma* **385**, 114858. https://doi.org/10.1016/j.geoderma.2020.114858 (2021).
19. Abedi, F. *et al.* Salt dome related soil salinity in southern Iran: prediction and mapping with averaging machine learning models. *Land Degrad. Dev.* **32**, 1540–1554. https://doi.org/10.1002/ldr.3811 (2020).
20. Qi, G., Chang, C., Yang, W. & Zhao, G. Soil salinity inversion in coastal cotton growing areas: a integration method of satellite-ground spectral fusion and satellite-UAV collaboration. *Land Degrad. Dev.* https://doi.org/10.1002/ldr.4287 (2022).
21. Jafarzadeh, H., Mahdianpari, M., Gill, E., Mohammadimanesh, F. & Homayouni, S. Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and PolSAR data: a comparative evaluation. *Remote Sens.* **13**, 4405. https://doi.org/10.3390/rs13214405 (2021).
22. Ivushkin, K. *et al.* Global mapping of soil salinity change. *Remote Sens. Environ.* **231**, 111260. https://doi.org/10.1016/j.rse.2019.111260 (2019).
23. Moreira, L. C. J., Teixeira, A. D. S. & Galvão, L. S. Potential of multispectral and hyperspectral data to detect saline-exposed soils in Brazil. *GISci. Remote Sens.* **52**, 416–436. https://doi.org/10.1080/15481603.2015.1040227 (2015).

24. Gorji, T., Yildirim, A., Hamzehpour, N., Tanik, A. & Sertel, E. Soil salinity analysis of Urmia Lake Basin using Landsat-8 OLI and Sentinel-2A based spectral indices and electrical conductivity measurements. *Ecol. Indicat.* **112**, 106173. https://doi.org/10.1016/j.ecolind.2020.106173 (2020).
25. Masoud, A. A., Koike, K., Atwia, M. G., El-Horiny, M. M. & Gemail, K. S. Mapping soil salinity using spectral mixture analysis of landsat 8 OLI images to identify factors influencing salinization in an arid region. *Int. J. Appl. Earth Observ. Geoinform.* **83**, 101944. https://doi.org/10.1016/j.jag.2019.101944 (2019).
26. Wang, J. *et al.* 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. https://doi.org/10.1016/j.geoderma.2019.06.040 (2019).
27. Carlson, T. N. & Ripley, D. A. On the relation between NDVI, fractional vegetation cover, and leaf area index. *Remote Sens. Environ.* **62**, 241–252 (1997).
28. Gitelson, A. A. & Merzlyak, M. N. Remote sensing of chlorophyll concentration in higher plant leaves. *Adv. Space Res.* **22**, 689–692 (1998).
29. Tucker, C. J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. Environ.* **8**, 127–150 (1979).
30. Rondeaux, G., Steven, M. & Baret, F. Optimization of soil-adjusted vegetation indices. *Remote Sens. Environ.* **55**, 95–107 (1996).
31. Birth, G. S. & McVey, G. R. Measuring the color of growing turf with a reflectance spectrophotometer 1. *Agron. J.* **60**, 640–643 (1968).
32. Huete, A. R. A soil-adjusted vegetation index (SAVI). *Remote Sens. Environ.* **25**, 295–309 (1988).
33. Taghizadeh-Mehrjardi, R., Minasny, B., Sarmadian, F. & Malone, B. P. Digital mapping of soil salinity in Ardakan region, central Iran. *Geoderma* **213**, 15–28. https://doi.org/10.1016/j.geoderma.2013.07.020 (2014).
34. Scudiero, E., Skaggs, T. H. & Corwin, D. L. Regional-scale soil salinity assessment using Landsat ETM + canopy reflectance. *Remote Sens. Environ.* **169**, 335–343. https://doi.org/10.1016/j.rse.2015.08.026 (2015).
35. Jordan, C. F. Derivation of leaf-area index from quality of light on the forest floor. *Ecology* **50**, 663–666 (1969).
36. Jiang, Z., Huete, A. R., Didan, K. & Miura, T. Development of a two-band enhanced vegetation index without a blue band. *Remote Sens. Environ.* **112**, 3833–3845 (2008).
37. Gitelson, A. A., Kaufman, Y. J. & Merzlyak, M. N. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sens. Environ.* **58**, 289–298 (1996).
38. Shi, C. *et al.* Quantitative inversion of soil salinity and analysis of its spatial pattern in agricultural area in Shihezi of Xinjiang. *Geogr. Res.* **33**, 2135–2144 (2015).
39. Allbed, A., Kumar, L. & Aldakheel, Y. Y. Assessing soil salinity using soil salinity and vegetation indices derived from IKONOS high-spatial resolution imageries: applications in a date palm dominated region. *Geoderma* **230–231**, 1–8. https://doi.org/10.1016/j.geoderma.2014.03.025 (2014).
40. Wang, F., Shi, Z., Biswas, A., Yang, S. & Ding, J. Multi-algorithm comparison for predicting soil salinity. *Geoderma* **365**, 114211. https://doi.org/10.1016/j.geoderma.2020.114211 (2020).
41. Vermeulen, D. & Van Niekerk, A. Machine learning performance for predicting soil salinity using different combinations of geomorphometric covariates. *Geoderma* **299**, 1–12. https://doi.org/10.1016/j.geoderma.2017.03.013 (2017).
42. Breiman, L. Machine learning. *Mach. Learn.* **45**, 5–32 (2001).
43. Chen, S. *et al.* A high-resolution map of soil pH in China made by hybrid modelling of sparse soil data and environmental covariates and its implications for pollution. *Sci. Total Environ.* **655**, 273–283 (2019).
44. Altman, N. & Krzywinski, M. Ensemble methods: bagging and random forests. *Nat. Methods* **14**, 933–935 (2017).
45. Guan, Y., Grote, K., Schott, J. & Leverett, K. Prediction of soil water content and electrical conductivity using random forest methods with UAV multispectral and ground-coupled geophysical data. *Remote Sens.* **14**, 1023 (2022).
46. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
47. Pedregosa, F. *et al.* Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
48. Zhang, Y., Liang, S., Zhu, Z., Ma, H. & He, T. Soil moisture content retrieval from Landsat 8 data using ensemble learning. *ISPRS J. Photogram. Remote. Sens.* **185**, 32–47. https://doi.org/10.1016/j.isprsjprs.2022.01.005 (2022).
49. Chen, T. & Guestrin, C. in *the 22nd ACM SIGKDD International Conference.*
50. Friedman, J. H. Greedy function approximation: a gradient boosting machine.. *Ann. Statist.* https://doi.org/10.1214/aos/1013203451 (2001).
51. Ke, G. *et al.* Lightgbm: a highly efficient gradient boosting decision tree. *Adv. Neural Inform. Process. Syst.* **30** (2017).
52. Su, H. *et al.* Super-resolution of subsurface temperature field from remote sensing observations based on machine learning. *Int. J. Appl. Earth Observ. Geoinform.* **102**, 102440. https://doi.org/10.1016/j.jag.2021.102440 (2021).
53. Sun, X., Liu, M. & Sima, Z. A novel cryptocurrency price trend forecasting model based on LightGBM. *Financ. Res. Lett.* **32**, 101084 (2020).
54. Kennedy, J. & Eberhart, R. in *Proceedings of ICNN'95-International Conference on Neural Networks*. 1942–1948 (IEEE).
55. Bellon-Maurel, V. & McBratney, A. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils–Critical review and research perspectives. *Soil Biol. Biochem.* **43**, 1398–1410 (2011).
56. Nocita, M. *et al.* Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. *Soil Biol. Biochem.* **68**, 337–347 (2014).
57. Zovko, M. *et al.* A geostatistical Vis-NIR spectroscopy index to assess the incipient soil salinization in the Neretva River valley Croatia. *Geoderma* **332**, 60–72. https://doi.org/10.1016/j.geoderma.2018.07.005 (2018).
58. Ge, X. *et al.* 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. https://doi.org/10.1016/j.catena.2022.106054 (2022).
59. Wu, D., Jia, K., Zhang, X., Zhang, J. & Abd El-Hamid, H. T. Remote sensing inversion for simulation of soil salinization based on hyperspectral data and ground analysis in Yinchuan China. *Nat. Resour. Res.* **30**, 4641–4656 (2021).
60. Madani, A. A. Soil salinity detection and monitoring using landsat data: a case study from Siwa Oasis, Egypt. *GISci. Remote Sens.* **42**, 171–181. https://doi.org/10.2747/1548-1603.42.2.171 (2013).
61. Richards, L. A. *Diagnosis and improvement of saline and alkali soils* (Scientific Publishers, 2012).
62. Zeraatpisheh, M., Ayoubi, S., Jafari, A., Tajik, S. & Finke, P. Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. *Geoderma* **338**, 445–452 (2019).
63. Han, L., Liu, D., Cheng, G., Zhang, G. & Wang, L. Spatial distribution and genesis of salt on the saline playa at Qehan Lake, Inner Mongolia, China. *Catena* **177**, 22–30 (2019).
64. Zhang, T.-T. *et al.* Detecting soil salinity with MODIS time series VI data. *Ecol. Ind.* **52**, 480–489. https://doi.org/10.1016/j.ecolind.2015.01.004 (2015).
65. Wichelns, D. & Qadir, M. Achieving sustainable irrigation requires effective management of salts, soil salinity, and shallow groundwater. *Agric. Water Manag.* **157**, 31–38 (2015).

Acknowledgements

We greatly appreciate the anonymous reviewers and editors who evaluated our article and provided insightful feedback. This study was supported by the project of Natural Science Foundation of Xinjiang Uygur Autonomous

Region (2019D01C024), the Xinjiang Uygur Autonomous Region Education Department Tianchi Doctoral Research Project (tcbs201816), and the Xinjiang University Doctoral Research Initiation Grant Program (BS180239).

Author contributions

S.M.: Conceptualization, Methodology, Investigation, Software, Formal, Writing-orginal draft, Formal analysis. B.H.: Investigation, Supervision, Writing - Review & Editing, Funding acquisition. B.X.: Investigation, Software. X.G.: Investiagtion, Formal analysis. L.H.: Investigation, Software.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.H.

Reprints and permissions information is available at www.nature.com/reprints.

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