

Article

An Explainable AutoML Framework for Soil Salinity Mapping Using Multi-Source Data in an Arid Irrigated District, China

Hong Guan ^{1,2}, Qidong Ding ³, Junhua Zhang ^{3,*}  and Lei Zhu ^{1,*}

¹ School of Civil and Hydraulic Engineering, Ningxia University, Yinchuan 750021, China

² Department of Environmental and Resource Engineering, Ningxia Technical College of Wine and Desertification Prevention, Yinchuan 750021, China

³ School of Ecology and Environment, Ningxia University, Yinchuan 750021, China

* Correspondence: jhzhang@nxu.edu.cn (J.Z.); nxuzhulei@163.com (L.Z.)

Abstract

Soil salinization limits agricultural production and the management of soil and water resources in arid irrigated regions. Regional prediction remains difficult, while soil salinity is influenced by hydrology, topography, and land cover. This study developed an explainable automated machine learning (AutoML) framework to map surface soil salinity in the Qingtongxia Irrigation District, Ningxia, China. The analysis combined 108 soil samples (0–10 cm), collected in March–April 2024, with 36 candidate input features from Sentinel-2 spectral indices, topography, soil texture, groundwater depth, and climate. Pearson correlation analysis and recursive feature elimination with cross-validation (RFECV) selected 10 key input features. Among the tested models, AutoML achieved the best validation performance, with $R^2 = 0.78$, MAE = 1.94 g/kg, and RMSE = 2.65 g/kg. The resulting 30 m prediction map captured broad regional patterns of soil salinity, with predicted values from 0.34 to 22.91 g/kg and higher salinity in northern and northeastern areas. Shapley additive explanations (SHAP) analysis highlighted brightness index, groundwater depth, clay content, elevation, and salinity index I as influential features. These findings suggest that explainable AutoML can help identify regional salinity risk and guide future sampling and model refinement.

Keywords: soil salinity prediction; AutoML; multi-source data; SHAP-based interpretability; Qingtongxia Irrigation District

1. Introduction

Soil salinization is a major form of land degradation worldwide. The Food and Agriculture Organization of the United Nations estimates that approximately 1.381 billion hectares of soils are affected by salinity, covering 10.7% of the total global land area [1,2]. Excessive salt accumulation can impair soil structure and reduce soil fertility, consequently suppressing crop growth and causing yield losses. The adverse impacts of soil salinity pose a substantial threat to regional ecological security and agricultural sustainability. Soil salinity exhibits pronounced regional variability and spatial heterogeneity [3,4]. Accurate identification of the spatial distribution patterns of soil salinity is essential for clarifying salinization mechanisms, supporting targeted management, and improving land resource utilization.

Soil salt content (SSC) is commonly assessed by field soil sampling followed by laboratory analysis, whereas portable electromagnetic induction instruments can provide rapid in situ estimates when properly calibrated [5,6]. Laboratory analysis of field-collected



Academic Editor: Yuki Yoshi Iwata

Received: 16 May 2026

Revised: 29 June 2026

Accepted: 6 July 2026

Published: 10 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

samples provides direct and reliable SSC estimates at sampled locations. At regional scales, point-level accuracy alone does not ensure spatial representativeness. Capturing the spatial heterogeneity of SSC across large areas requires dense and representative sampling. The cost and time required for repeated field sampling therefore make sparse point-based measurements insufficient, on their own, to resolve continuous spatial variability or support repeated monitoring over time.

Satellite imagery has increasingly been used to retrieve SSC [7]. Satellite missions and sensors, including Sentinel-2, Landsat-8 and MODIS, provide multi-source spectral observations and enable large-scale salinity retrieval and monitoring. Compared with in situ point measurements, satellite remote sensing offers broad spatial coverage, efficient data acquisition, and strong temporal monitoring capacity, thereby partly overcoming the limited spatial continuity of point-based observations [8,9]. However, SSC retrieval from satellite imagery remains constrained by cloud cover, atmospheric conditions, vegetation interference, and the spatial and temporal resolutions of sensors [10].

Advances in spectral measurement technologies have rendered multi-source soil salinity observations increasingly available. Integrating multi-source data can improve model predictive accuracy. However, the resulting increase in data dimensionality may introduce redundant information, increase the risk of model overfitting, and reduce computational efficiency and generalizability [11,12]. Moreover, soil salinization is governed by a range of environmental drivers, such as climate, topography, soil properties, hydrological conditions, and irrigation practices. The dominant drivers and underlying mechanisms of soil salinization vary substantially among regions. This limits the applicability of fixed sets of features or a single modeling approach across regions [13].

Machine learning (ML) has been widely used for soil salinity mapping because it can capture nonlinear patterns between environmental covariates and soil salinity [14]. RF, XGBoost, and other ensemble methods are commonly used for soil salinity prediction, while many applications still rely on manually configured workflows for feature selection, model selection, and hyperparameter tuning, potentially reducing reproducibility and prediction stability [15]. In the Qingtongxia Irrigation District, salt accumulation is influenced by Yellow River irrigation, intense evaporation, shallow groundwater, drainage conditions, and topographic variation, making manually configured modeling workflows less stable and potentially suboptimal for salinity prediction. An automated machine learning (AutoML) framework can automate model comparison, hyperparameter tuning, and ensemble construction, allowing the modeling workflow to be selected through systematic evaluation [16,17]. With feature selection and SHAP analysis, this framework may support more reproducible salinity mapping and help interpret the relative importance of selected features.

Primary and secondary salinization are widespread in arid and semiarid irrigated regions, where salt accumulation reflects both natural processes and human activities [18]. However, high-resolution salinity mapping in these systems remains constrained by the difficulty of integrating heterogeneous geospatial predictors with interpretable modelling. Here, we developed an AutoML workflow, coupled with SHAP analysis, to map soil salinity at 30 m resolution in the Qingtongxia Irrigation District, northern Ningxia, China. By linking prediction with interpretation, this workflow provides a basis for mapping soil salinity and examining its associations with geospatial covariates in irrigated drylands.

2. Study Area and Datasets

2.1. Study Area

The Qingtongxia Irrigation District is located in the northern Ningxia Hui Autonomous Region, China (37°49′–39°23′ N, 105°37′–106°39′ E). The study area represents one of the

major irrigated agricultural regions in the Yellow River Basin. It covers a regional area of approximately $2.12 \times 10^4 \text{ km}^2$, with a designed irrigated area of approximately $3.37 \times 10^3 \text{ km}^2$ (Figure 1). This area lies in the upper reaches of the Yellow River Basin and forms part of the Ningxia Plain. It borders the Helan Mountains to the west, the Yellow River to the east, and the Inner Mongolia Autonomous Region to the north, with margins extending east of the Yellow River. The terrain is flat and the climate is arid temperate continental, with low and seasonally uneven precipitation. Mean annual precipitation is only 180–220 mm, much lower than annual evaporation of approximately 2000 mm. Consequently, local agricultural production depends heavily on irrigation water diverted from the Yellow River. In addition to natural constraints, prolonged irrigation periods, inefficient irrigation practices, inadequate drainage, and the resulting rise in the groundwater table further accelerate salt accumulation in the soil profile and at the land surface.

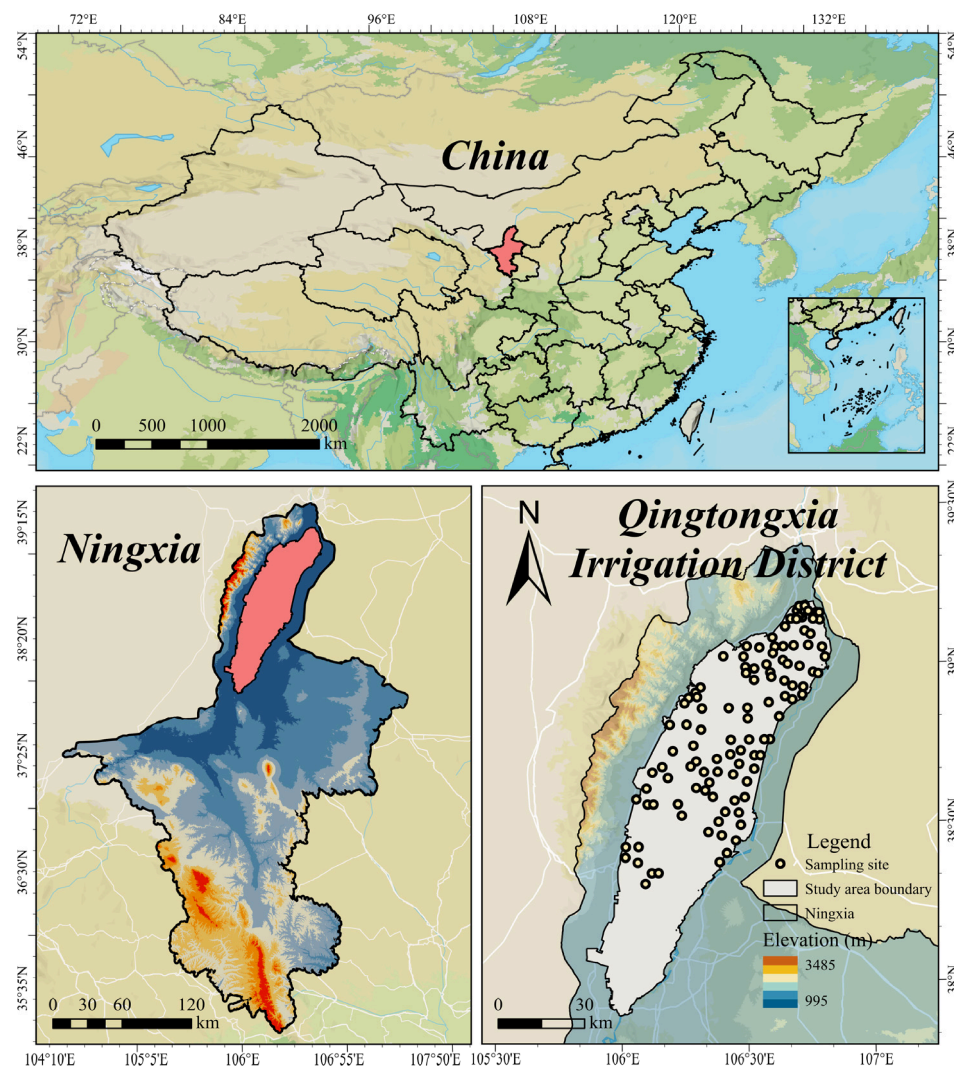


Figure 1. Location of the study area and sampling sites.

2.2. Field Soil Sampling and Laboratory Analysis

A total of 108 sampling sites were selected for surface soil collection across the Qingtongxia Irrigation District. Site selection accounted for land-use type, known salinization patterns, and spatial coverage within the study area. Based on field reconnaissance and existing information on salt-affected land, more sites were placed in the central and northern parts of the district, where salinization was expected to be more severe. Of the 108 sites, 74 were located in cultivated land and 34 in uncultivated land. Of these sites, 66 were

located in Shizuishan City and 42 in Yinchuan City. Areas strongly affected by roads, settlements, ditches, or other local disturbances were avoided.

Field sampling was conducted from March to April 2024 from the 0–10 cm surface soil layer. This period was selected to match the acquisition period of the remote-sensing imagery and to minimize canopy-related spectral interference before the main growing season. At each site, a 10 m × 10 m quadrat was established. Five soil subsamples were collected from the center and four corners using a five-point composite sampling scheme and mixed to obtain one composite sample per site (Figure 2). In total, 540 subsamples were combined into 108 composite soil samples. The geographic coordinates of each site were recorded using a handheld GPS device.

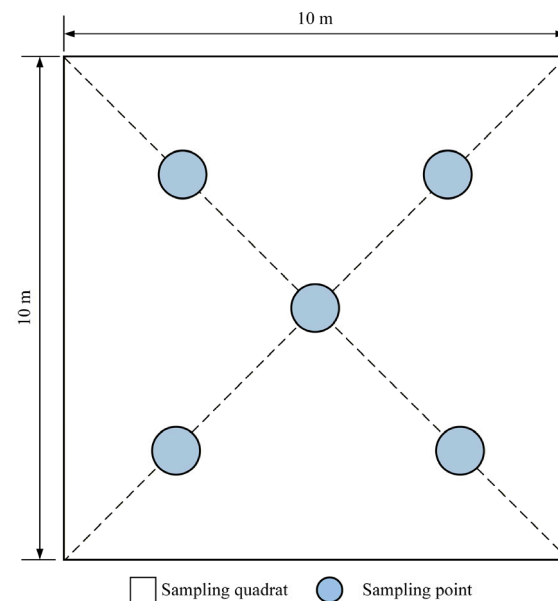


Figure 2. Layout of five-point composite soil sampling.

After collection, the soil samples were transported to the laboratory, air-dried, gently crushed, and passed through a sieve. Soil extracts were prepared using deionized water at a soil-to-water ratio of 1:5 (*w/v*). SSC was determined gravimetrically by oven-drying the extract at 105 °C to constant weight, following established procedures for determining total water-soluble salts in soil [19]. The soluble-ion composition indicated low proportions of Ca^{2+} , HCO_3^- , and CO_3^{2-} , suggesting that gypsum and bicarbonate salts had limited influence on gravimetric SSC determination.

3. Methodology

3.1. Overview

A multi-source data-driven framework was developed for soil salinity prediction and model interpretability analysis. The overall workflow comprises four steps: data preprocessing; feature construction and selection; model development and evaluation; and spatial prediction and model interpretation (Figure 3). (1) We integrated multi-source data (Sentinel-2 remote sensing imagery, topographic and geomorphological factors, soil properties, and environmental covariates) to construct the input features required for soil salinity prediction. (2) All input features were normalized, and key features were then identified using Pearson correlation analysis coupled with RFECV, thereby reducing information redundancy and optimizing the feature combination. (3) Five ML models were developed based on support vector regression (SVR), random forest (RF), gradient boosting decision tree (GBDT), extreme gradient boosting (XGBoost), and AutoML. The

model performance for soil salinity prediction was systematically evaluated using the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE). (4) We generated a spatial distribution map of soil salinity in the study area based on the optimal model. The effects and contributions of key factors driving spatial variability in soil salinity were revealed by SHAP analysis to enhance the interpretability of the model outputs.

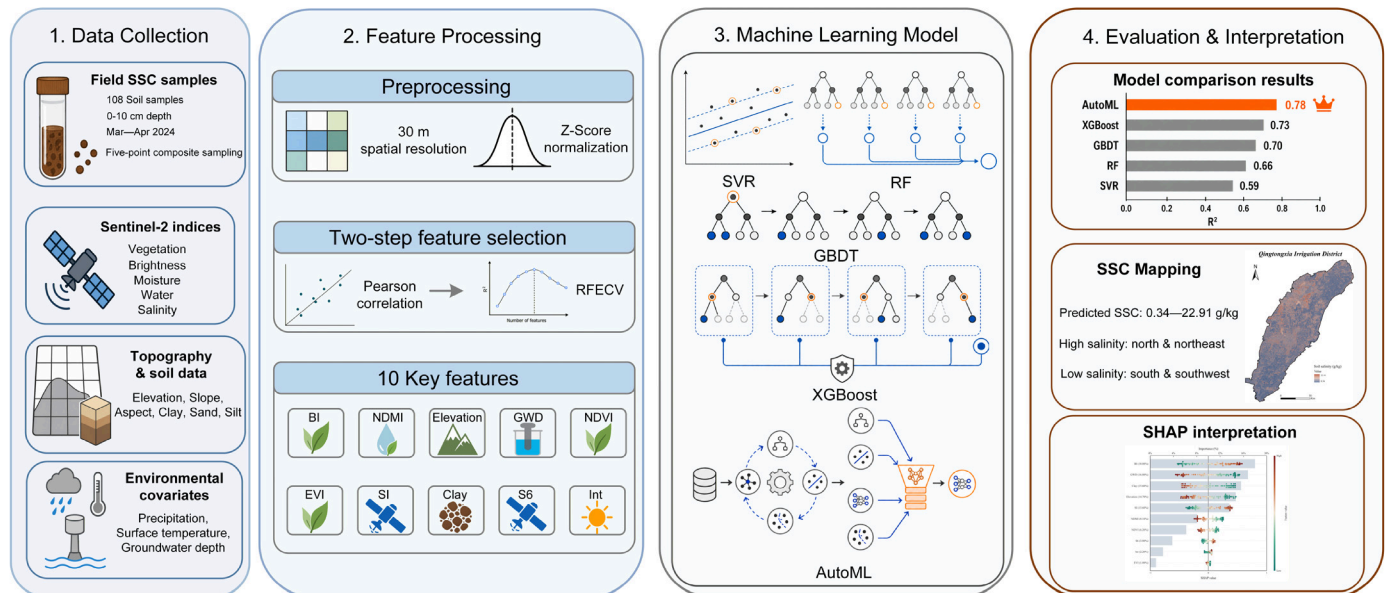


Figure 3. Flowchart of the study.

3.2. Input Feature Construction

The candidate input features were assembled to represent spectral, soil, topographic, hydrological, and climatic controls on soil salinity (Table 1). In total, 36 candidate features were used for feature selection and model development.

Guided by previous salinity mapping studies and recent reviews, Sentinel-2 spectral bands were used to calculate 12 vegetation indices, 2 brightness indices, 2 water indices, and 11 salinity indices that characterize vegetation status, surface brightness, moisture conditions, and salinity-related spectral responses [20–23]. The blue (B), green (G), red (R), and near-infrared (NIR) bands have native spatial resolutions of 10 m, whereas the short-wave infrared (SWIR) band has a native spatial resolution of 20 m. The vegetation indices characterize vegetation cover, indirectly reflecting vegetation responses to salt stress. The brightness indices describe the spectral properties of bare soil and high-reflectance surfaces, which help to identify enhanced surface reflectance in salinized areas. The water indices quantify variation in surface water and moisture conditions. The salinity indices capture the characteristic spectral responses of salinized soils in visible and near-infrared bands. Sentinel-2 imagery was downloaded from the Google Earth Engine (GEE) cloud platform (<https://developers.google.com/earth-engine/datasets> (accessed on 15 January 2026)). To ensure temporal consistency between remote sensing observations and field sampling, we selected images acquired during the soil sampling period. All images had cloud cover below 10%, which minimized the influence of cloud contamination on image quality. The images were mean-composited for subsequent analysis. Sentinel-2 Level 2A surface reflectance products were generated using the Sen2Cor atmospheric correction algorithm [24].

Table 1. Details of multi-source variables used in this study.

Category	Variable	Abbreviation
Soil	Soil clay content	<i>Clay</i>
	Soil sand content	<i>Sand</i>
	Soil silt content	<i>Silt</i>
Topography	Elevation	<i>Elevation</i>
	Slope	<i>Slope</i>
	Aspect	<i>Aspect</i>
Vegetation	Ratio vegetation index	$RVI = NIR/R$
	Difference vegetation index	$DVI = NIR - R$
	Green difference vegetation index	$GDVI = NIR - G$
	Normalized difference vegetation index	$NDVI = (NIR - R)/(NIR + R)$
	Transformed normalized difference vegetation index	$TNDVI = \sqrt{NDVI + 0.5}$
	Green normalized difference vegetation index	$GNDVI = (NIR - G)/(NIR + G)$
	Weighted difference vegetation index	$WDVI = NIR - 0.1R$
	Enhanced vegetation index	$EVI = 2.5 \cdot \frac{NIR - R}{NIR + 6R - 7.5B + 1}$
	Soil adjusted vegetation index	$SAVI = \frac{1.5(NIR - R)}{NIR + R + 0.5}$
	Modified Soil adjusted vegetation index	$MSAVI = 0.5 \left[2NIR + 1 - \sqrt{(2NIR + 1)^2 - 8(NIR - R)} \right]$
Brightness	Infrared percentage vegetation index	$IPVI = \frac{NIR(NDVI + 1)}{2(NIR + R)}$
	Greenness index	$GI = G/R$
	Brightness index	$BI = \sqrt{R^2 + G^2 + NIR^2}/2$
Water	Intensity index	$Int = \sqrt{G + R + NIR}/2$
	Normalized difference moisture index	$NDMI = (NIR - SWIR)/(NIR + SWIR)$
Salinity	Normalized difference water index	$NDWI = (G - NIR)/(G + NIR)$
	Salinity index 1	$S1 = B/R$
	Salinity index 2	$S2 = (B - R)/(B + R)$
	Salinity index 3	$S3 = (G \times R)/B$
	Salinity index 5	$S5 = (B \times R)/G$
	Salinity index 6	$S6 = (R \times NIR)/G$
	Salinity index I_T	$SI_T = (R/NIR) \times 100$
	Salinity index I	$SI = \sqrt{B \times R}$
	Salinity index I1	$SI1 = \sqrt{G \times R}$
	Salinity index I2	$SI2 = \sqrt{G^2 + R^2 + NIR^2}$
	Salinity index I3	$SI3 = \sqrt{G^2 + R^2}$
	Normalized difference salinity index	$NDSI = (R - NIR)/(R + NIR)$
Others	Mean annual precipitation	<i>Rain</i>
	Surface temperature	<i>ST</i>
	Groundwater depth	<i>GWD</i>

Clay, *sand*, and *silt* contents were selected as soil variables to capture the potential effect of soil texture on salt transport processes. Soil data were obtained from the global gridded soil dataset released by the International Soil Reference and Information Centre (<https://soilgrids.org/> (accessed on 15 January 2026)) [25]. This dataset provides estimates of soil properties at multiple standard soil depths, with a spatial resolution of 250 m. *Elevation*, *slope*, and *aspect* were used to characterize the potential role of topographic conditions in shaping spatial variability in soil salinity. These variables were derived from the Shuttle Radar Topography Mission (SRTM) 1-arc-second global digital elevation model [26], which has a spatial resolution of 30 m (<https://developers.google.com/earth-engine/datasets> (ac-

cessed on 15 January 2026)). Mean annual precipitation, surface temperature, and groundwater depth were selected as environmental covariates to indicate the regional environmental context for the development and evolution of soil salinization. Precipitation and temperature data were obtained from monthly datasets with a spatial resolution of 1 km, released by the National Tibetan Plateau/Third Pole Environment Data Center (<https://data.tpdac.ac.cn/en/data> (accessed on 15 January 2026)). Groundwater depth data were obtained from automatic groundwater monitoring stations of the Ningxia water resources system. Observations from March to April 2024 were averaged at each station and interpolated using kriging to generate a continuous groundwater depth layer.

To ensure consistency in their spatial resolution, multi-source variables were preprocessed using a uniform workflow. We implemented bilinear interpolation in ArcGIS Pro 3.4.3 (Esri, Redlands, CA, USA) to resample all raster variables at a spatial resolution of 30 m. The geographic coordinate system was set to WGS 1984. All input features were standardized using z-score transformation, and SSC was log-transformed before model training. Predictions were transformed back for model evaluation and spatial mapping.

3.3. Feature Selection

Linear correlations among input features can lead to redundant information, and some features may be weakly associated with soil salinity. Directly using all input features in a soil salinity prediction model would increase computational complexity. Additionally, the introduction of uninformative variables could reduce predictive accuracy and generalizability.

Pearson correlation analysis quantifies linear correlations between variables [27]. The correlation coefficient r ranges from -1 to 1 , with values > 0 indicating positive correlations and values < 0 indicating negative correlations. $|r|$ reflects the strength of the correlation. We applied Pearson correlation analysis to evaluate the relationships between input features and soil salinity, and statistical significance was determined by $p < 0.05$.

RFECV is a wrapper feature selection method that evaluates the relative contribution of input features to model predictive performance [28]. This method progressively removes the features that contribute least to model performance. It uses cross-validation to evaluate different feature subsets, thereby identifying the subset with the best cross-validated performance. Using this method, we further selected input features strongly associated with soil salinity and removed features with low predictive contribution.

3.4. Model Development

To systematically compare the applicability and predictive performance of different ML methods for soil salinity prediction, five models were developed using 108 soil samples collected through field observations. The samples were divided into a calibration set and a validation set at an 8:2 ratio by stratified random sampling. The calibration set was used for model development, and the validation set was used for performance evaluation.

SVR is a regression method based on statistical learning theory [29]. It minimizes the overall prediction error while allowing most sample prediction errors to remain within a specified tolerance threshold:

$$\min_{\omega, b} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^n \max \left\{ 0, \left| y_i - (\omega^T(x_i) + b) \right| - \varepsilon \right\} \quad (1)$$

where x_i denotes the input feature vector, y_i denotes the output value, n denotes the number of samples, $\|\omega\|^2$ measures model complexity, ε denotes the error tolerance threshold, and C denotes the penalty parameter.

RF is an ensemble regression algorithm [30], in which multiple decision trees are trained on random subsamples of the original data, and predictions are generated by averaging the outputs of individual trees:

$$\{\omega_b, \hat{\Theta}_b\}_{b=1}^B = \arg \min_{\{\omega_b, \Theta_b\}} \frac{1}{n} \sum_{i=1}^n (y_i - \frac{1}{B} \sum_{b=1}^B \omega_b T(x_i; \Theta_b))^2, \omega_b \geq 0, \sum_{b=1}^B \omega_b = 1 \quad (2)$$

$$Y = \sum_{b=1}^B \omega_b T(X; \hat{\Theta}_b) \quad (3)$$

where Y denotes the final prediction, ω_b and Θ_b denote the weight and parameters of the b th decision tree, respectively, and $T(x_i, \Theta_b)$ denotes the prediction produced by that tree.

GBDT sequentially builds new decision trees to fit the residuals from the previous iteration, thereby iteratively minimizing the corresponding objective function [31]:

$$\Gamma = \sum_{i=1}^n l(\hat{Y}, Y) \quad (4)$$

where Γ denotes the objective function, $\sum_{i=1}^n l(\hat{Y}, Y)$ denotes the total loss over all samples, $\hat{Y}$ denotes the observed value, and Y denotes the predicted value.

Building on GBDT, XGBoost incorporates gradient and Hessian information from a second order Taylor expansion [32] and regularizes the tree structure and leaf weights:

$$\Gamma = \sum_{i=1}^n l(\hat{Y}, Y) + \sum_{k=1}^N \Omega(Y_k) \quad (5)$$

where $\sum_{i=1}^n l(\hat{Y}, Y)$ denotes the total loss over the n samples and $\sum_{k=1}^N \Omega(Y_k)$ denotes the regularization term, which controls model complexity.

AutoML is an automated modeling approach that searches for the optimal modeling pipeline within a predefined set of candidate models and their associated hyperparameter spaces using a common evaluation metric and validation strategy [33]. In this study, AutoML searched the candidate models and their hyperparameter combinations using the calibration set; selected multiple top-performing models as base learners; and automatically constructed a stacking ensemble model to integrate complementary information across models, thereby improving predictive accuracy and generalizability. The set of candidate models, $\mathcal{M}$ is defined as follows:

$$\mathcal{M} = \{m_1, m_2, \dots, m_j\} \quad (6)$$

where m_j denotes the j th candidate model. The cross-validation loss of m_j under the hyperparameter setting θ_j can be written as follows:

$$\mathcal{J}(m_j; \theta_j) = \frac{1}{K} \sum_{k=1}^K \mathcal{L}_k(m_j, \theta_j) \quad (7)$$

where K denotes the number of cross validation folds, and $\mathcal{L}_k(m_j, \theta)$ denotes the loss of model m_j under the hyperparameter setting θ on the validation subset in the k th fold.

After evaluating the set of candidate models, AutoML applies a stacking strategy to learn from the predictions generated by multiple base models. The prediction for an input feature vector X can be written as follows:

$$\begin{cases} Y_b = f_b(X; \theta_b^*) \\ Y = H(Y_1, Y_2, \dots, Y_b) \end{cases} \quad (8)$$

where Y denotes the final prediction, X denotes the input feature vector, Y_b denotes the prediction generated by the b th base model, $f_b(\cdot)$ denotes the prediction function of that model, θ_b^* denotes its optimized hyperparameter setting, and $H(\cdot)$ denotes the stacking ensemble function.

To improve predictive performance, we optimized the hyperparameters of SVR, RF, GBDT, and XGBoost using the calibration set. The hyperparameter search was conducted with the Tree-structured Parzen Estimator (TPE) algorithm implemented in Optuna, and the optimal hyperparameter combination was selected based on the R^2 value on the calibration set. For AutoML, AutoGluon automatically handled model selection, hyperparameter optimization, and model ensembling using the Good_quality preset. The final stacking ensemble used RF, XGBoost, CatBoost, and NeuralNetFastAI as base learners, which were integrated by a weighted ensemble meta-learner with learned weights of 0.40, 0.25, 0.20, and 0.15, respectively. The key hyperparameter values for each model are listed in Table 2.

Table 2. Optimized hyperparameter settings and descriptions for different models.

Model	Hyperparameter	Value/Type	Description
SVR	C	0.2	Penalty parameter controlling the trade-off between model complexity and training errors
	Kernel	RBF	Type of kernel function
	ϵ	0.1	Width of the ϵ -insensitive tube controlling the tolerance to prediction errors
RF	N_estimators	53	Number of trees
	Max_depth	2	Maximum depth of the decision tree
	Min_samples_leaf	3	Minimum number of samples per leaf node
GBDT	Learning_rate	0.01	Learning rate for weight updates
	N_estimators	36	Number of boosting trees
	Max_depth	3	Maximum depth of the decision tree
XGBoost	Learning_rate	0.03	Learning rate for weight updates
	N_estimators	43	Number of boosting trees
	Max_depth	2	Maximum depth of the decision tree
AutoML	presets	Good_quality	Preset configuration controlling the overall model selection and training quality

3.5. Model Evaluation

To evaluate the predictive performance of the ML models, R^2 , MAE, and RMSE were calculated using the validation set. R^2 quantifies the extent to which a model explains the variability of observed values. R^2 values closer to 1 indicate better model fit, whereas lower R^2 values indicate weaker explanatory power. MAE represents the mean absolute deviation between predicted and observed values. RMSE measures the overall magnitude of prediction errors and is more sensitive to large errors. Regarding MAE and RMSE, lower values indicate smaller prediction errors and greater predictive accuracy, whereas higher

values indicate inferior predictive performance. The formulas for these statistical metrics are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (O_i - S_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} \quad (9)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |O_i - S_i| \quad (10)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (O_i - S_i)^2} \quad (11)$$

where O_i and S_i denote the observed and predicted values for sample i , respectively; $\bar{O}$ denotes the mean of the observed values; and n denotes the total number of samples.

3.6. SHAP-Based Model Interpretation

SHAP is a post hoc interpretability method based on Shapley values from cooperative game theory [34]. This analysis decomposes the model prediction into additive contributions from individual input features. It explains which features drive a given prediction and how much each feature contributes to predictive accuracy at the level of individual samples. Under an additive feature attribution framework, SHAP presents the model output $f(x)$ for any sample x as the sum of a baseline value and the contributions of all features:

$$f(x) = \phi_0 + \sum_{j=1}^p \phi_j \quad (12)$$

where ϕ_0 denotes the baseline output, computed as the expected model output over the background data; and ϕ_j denotes the contribution of feature j to the prediction for sample x . ϕ_j is defined as the Shapley weighted average of the marginal contribution of feature j over all feature subsets that do not include feature j :

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(p - |S| - 1)!}{p!} [\nu(S \cup \{j\}) - \nu(S)] \quad (13)$$

where F denotes the set of all features; S denotes any subset of features that does not contain feature j ; and $|S|$ denotes the size of subset S . The term $\frac{|S|!(p - |S| - 1)!}{p!}$ is the Shapley weight, and $\nu(S \cup \{j\}) - \nu(S)$ represents the marginal contribution of feature j to the model output when it is added to subset S .

3.7. Experimental Setting

Five ML models were used to predict soil salinity. All model construction, training, result analysis, and regional mapping were performed on the same computing platform, which was equipped with an Intel Core i7-12700KF CPU (3.60 GHz), 64 GB RAM, and an NVIDIA GeForce RTX 3070 Ti GPU and configured with Windows 11, Python 3.12.12, scikit-learn 1.7, Optuna 4.8.0, AutoGluon 1.5.0, and SHAP 0.50.

4. Results

4.1. Descriptive Statistics of SSC

The SSC in the study area ranged from 0.20 to 24.56 g/kg, with a mean of 5.14 g/kg and a median of 2.87 g/kg (Table 3). In the full dataset, the coefficient of variation for SSC was 1.09, indicating high dispersion and variability among the samples. The skewness and kurtosis of SSC were 1.78 and 2.57, respectively, suggesting a right skewed and relatively peaked distribution. All datasets contained outliers mainly in the high-value range (Figure 4), showing a pronounced right-tailed distribution of SSC. With respect to soil

salinity level, the samples were predominantly classified as severely salinized soils (25.00%) and moderately salinized soils (23.15%). Non-saline soils and slightly salinized soils accounted for moderate proportions (19.44% each), with extremely salinized soils making up a relatively small proportion (12.96%). The calibration and validation sets showed broadly consistent means, medians, and distribution patterns, implying that the dataset partitioning was representative.

Table 3. Descriptive statistics of soil salt content in the study area (g/kg).

Dataset	Sample Number	Minimum	Maximum	Mean	Median	Interquartile Range	Standard Deviation	Coefficient of Variation	Skewness	Kurtosis
Full dataset	108	0.20	24.56	5.14	2.87	5.41	5.63	1.09	1.78	2.57
Calibration set	86	0.24	24.56	5.17	2.94	5.64	5.67	1.10	1.77	2.58
Validation set	22	0.20	21.50	5.04	2.77	5.11	5.58	1.11	1.83	2.67

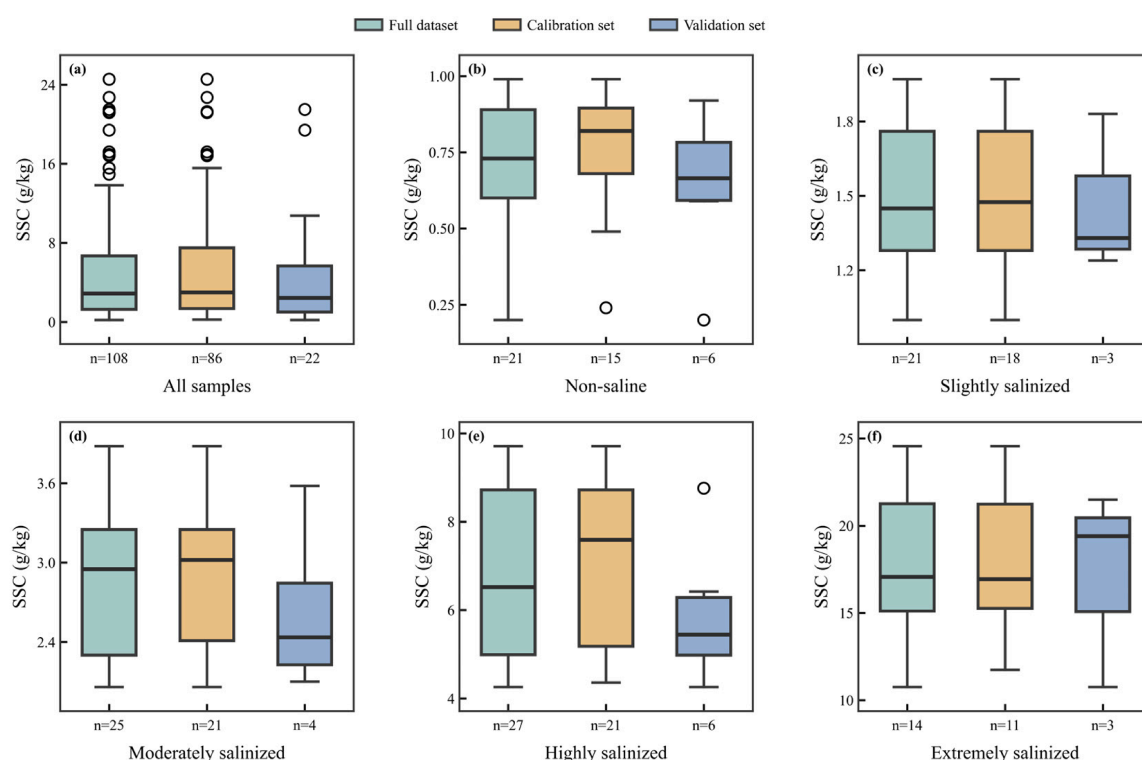


Figure 4. Soil salt content (SSC) distributions across salinity classes for the whole dataset, calibration set, and validation set: (a) all samples; (b) non-saline soils; (c) slightly salinized soils; (d) moderately salinized soils; (e) highly salinized soils; and (f) extremely salinized soils.

4.2. Results of Feature Selection

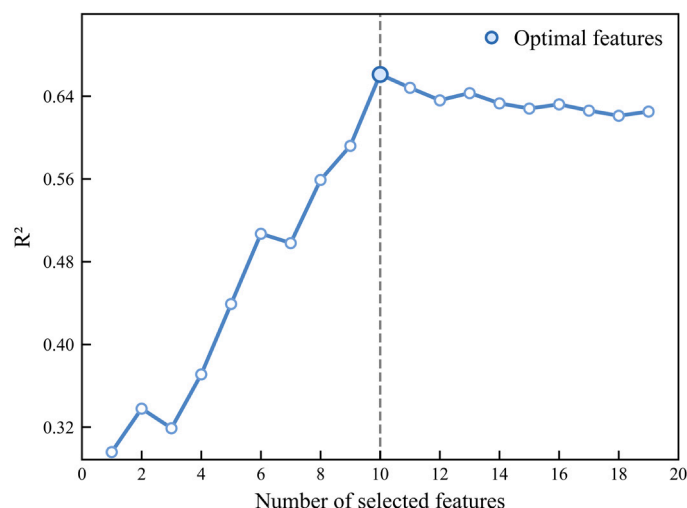
Pearson correlation analysis revealed that 19 of the 36 input features were significantly correlated with SSC ($p < 0.05$; Table 4). The SSC-related features comprised spectral indices for vegetation, water, brightness, and salinity, in addition to *Elevation* and *GWD*. Among them, *BI* showed the strongest correlation with SSC ($p < 0.001$, $r = 0.365$). Ten features were negatively correlated with SSC at the 0.01 level, whereas the remaining eight features showed either negative or positive correlations with SSC at the 0.05 level. Some vegetation and salinity indices had identical or highly similar r values with SSC. For example, *RVI*, *GDVI*, *NDVI*, *TNDVI*, *SAVI*, and *MSAVI* all had an r value of -0.279 with SSC, suggesting possible redundancy or collinearity among these features. Therefore, RFECV was used for further feature selection.

Table 4. Pearson correlation analysis of soil salt content with 36 input variables.

Variable	<i>r</i>	Variable	<i>r</i>
<i>BI</i>	0.365 ***	<i>Int</i>	0.235 *
<i>NDMI</i>	−0.325 **	<i>NDSI</i>	0.174
<i>Elevation</i>	−0.321 **	<i>NDWI</i>	−0.174
<i>GWD</i>	−0.295 **	<i>IPVI</i>	−0.173
<i>WDVI</i>	−0.282 **	<i>GI</i>	−0.168
<i>RVI</i>	−0.279 **	<i>ST</i>	0.148
<i>GDVI</i>	−0.279 **	<i>S5</i>	0.147
<i>NDVI</i>	−0.279 **	<i>S1</i>	−0.140
<i>TNDVI</i>	−0.279 **	<i>S2</i>	−0.140
<i>SAVI</i>	−0.279 **	<i>S3</i>	0.117
<i>MSAVI</i>	−0.279 **	<i>SI3</i>	0.106
<i>DVI</i>	−0.276 *	<i>SI1</i>	0.092
<i>GNDVI</i>	−0.276 *	<i>Slope</i>	0.070
<i>EVI</i>	−0.261 *	<i>Silt</i>	0.047
<i>SI_T</i>	0.260 *	<i>Sand</i>	0.034
<i>SI</i>	0.260 *	<i>SI2</i>	0.034
<i>Clay</i>	−0.243 *	<i>Rain</i>	0.010
<i>S6</i>	0.240 *	<i>Aspect</i>	0.004

Note: * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

The RFECV results showed that the model generally produced higher R^2 values with an increasing number of input features. A peak value of 0.661 was reached when 10 features were included (Figure 5). Adding more features thereafter produced minor variation in the R^2 value, which exhibited a slightly decreasing trend. This means that additional features did not improve model performance and might have introduced redundancy. Therefore, the final feature set encompassed 10 optimal features: *BI*, *NDMI*, *Elevation*, *GWD*, *NDVI*, *EVI*, *SI*, *Clay*, *S6*, and *Int*.

**Figure 5.** RFECV curve for optimal feature selection.

4.3. Performance Comparison of Different Models for Ssc Prediction

The five ML models showed considerably different predictive performance for SSC retrieval (Table 5). Compared with other models, the AutoML model performed best in both the calibration and validation sets. In the calibration set, the AutoML model achieved an R^2 value of 0.84, with MAE of 1.62 g/kg and RMSE of 2.26 g/kg. In the validation set, the AutoML model yielded an R^2 value of 0.78, with MAE of 1.94 g/kg and RMSE of 2.65 g/kg. These results are indicative of good model fits and low prediction errors. XGBoost was the second-best model, with R^2 of 0.70, MAE of 2.29 g/kg, and RMSE of 3.10 g/kg in the validation set. The RF model slightly outperformed the GBDT model, both

of which showed intermediate performances. The SVR model had the lowest R^2 value and the largest errors, suggesting its relatively low predictive accuracy.

Table 5. Comparative performance of different models on the calibration and validation sets.

Model	Calibration Set			Validation Set		
	R^2	MAE	RMSE	R^2	MAE	RMSE
SVR	0.46	3.08	4.14	0.33	3.52	4.63
RF	0.74	2.09	2.88	0.67	2.42	3.26
GBDT	0.72	2.15	2.99	0.62	2.60	3.48
XGBoost	0.77	1.95	2.70	0.70	2.29	3.10
AutoML	0.84	1.62	2.26	0.78	1.94	2.65

The measured SSC values were compared with the predicted values from each ML model in the calibration and validation sets (Figure 6). The proximity of the fitted regression lines to the 1:1 line was used to evaluate model fit and prediction bias, and possible saturation at high SSC. While all models captured variation in SSC to some extent, they differed markedly in model fit, performance stability, and their ability to capture high SSC values. The AutoML model performed best, producing the tightest clustering of data points. Most points lay close to the 1:1 line, indicating strong agreement between measured and predicted values. This pattern also indicated weak compression in the high SSC range, with no obvious saturation effect in the final model. The R^2 values for the calibration and validation sets were 0.84 and 0.78, respectively. The regression line showed the smallest deviation from the 1:1 line, with $\alpha = 9.99^\circ$. These results suggest high predictive accuracy and stable performance of this model across the two datasets.

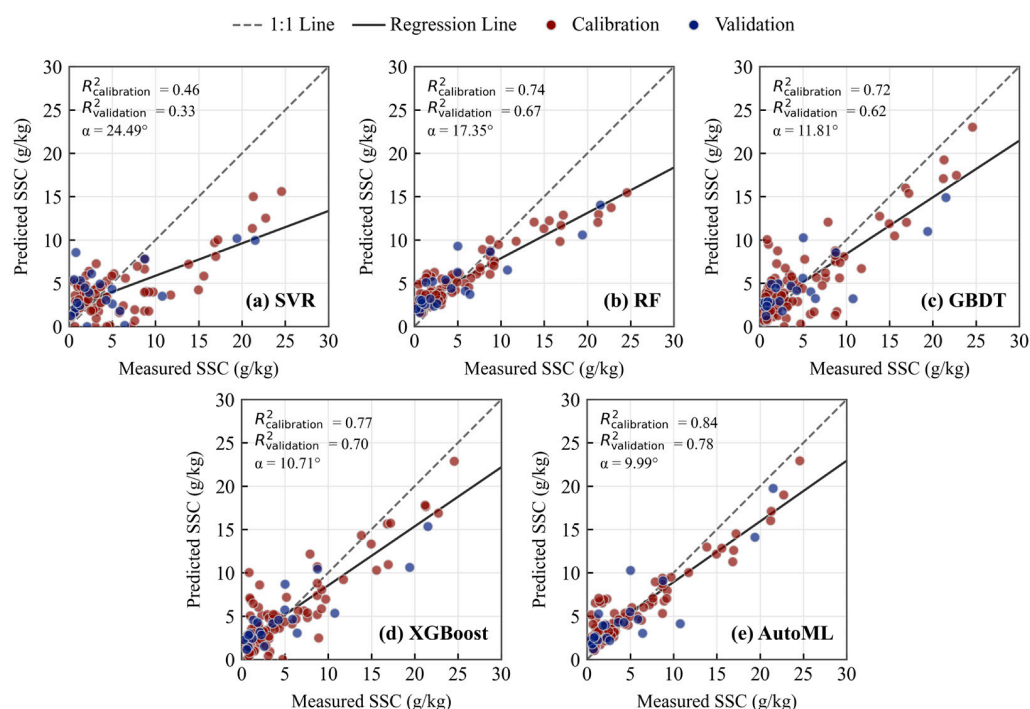


Figure 6. Comparison of measured and predicted SSC across five machine learning models.

The XGBoost model performed well, although it was inferior to the AutoML model. The data points were relatively well clustered, particularly in the low-value and moderate-value ranges. The calibration-set and validation-set R^2 values were 0.77 and 0.70, respectively, indicating relatively strong predictive performance. However, the model still showed

certain bias in the high-value range, suggesting that it was less capable of capturing extreme SSC values.

The RF and GBDT models both showed moderate predictive performance. Their predictions agreed reasonably well with the observations in the low-value range. For higher SSC values, however, the predictions departed increasingly from the 1:1 line. This departure was most evident in the high-value range, where the predictions became more dispersed and tended to underestimate SSC. The validation-set R^2 values for the RF and GBDT models were 0.67 and 0.62, respectively, lower than those for the AutoML and XGBoost models, indicating weaker generalization.

The SVR model showed the weakest performance, with the widest scatter and the clearest overall departure of data points from the 1:1 line. This model systematically underestimated SSC in the medium-to-high-value range. The calibration-set and validation-set R^2 values were only 0.46 and 0.33, respectively. The regression line showed the largest angle, with $\alpha = 24.49^\circ$. The results are indicative of large prediction errors and poor performance stability.

4.4. Soil Salinity Mapping

The spatial distribution of soil salinity in the study area was predicted by the AutoML model, with values ranging from 0.34 to 22.91 g/kg (Figure 7). High-salinity areas were mainly located in the northern and northeastern parts of the study area, with several scattered patches of high salinity in the central part. Low-salinity areas were concentrated in the southern and southwestern parts of the study area. The distribution of soil salinity showed an upward trend from south to north, with local clustering and patchiness. The resulting prediction map shows how SSC varies across the study area, providing a spatially continuous representation of the model results. This model demonstrated great potential for regional soil salinization monitoring and mapping.

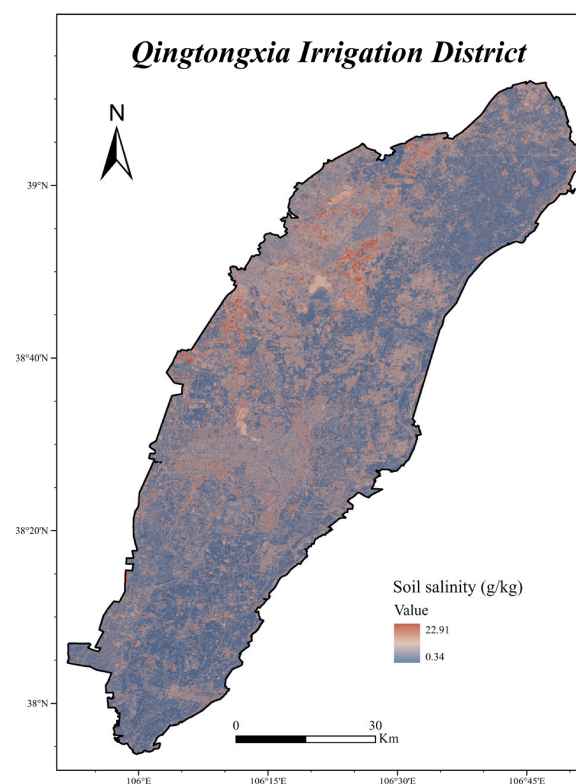


Figure 7. Spatial distribution of soil salinity predicted by the AutoML model in the Qingtongxia Irrigation District.

4.5. SHAP-Based Interpretation of the AutoML Model

SHAP analysis for the AutoML model revealed that *BI* was the largest contributor to the model output, with feature importance of 18.0% (Figure 8). Other important variables included *GWD* (16.8%), *Clay* (15.6%), *Elevation* (14.7%), and *SI* (13.6%). Together, these five variables accounted for 78.70% of the cumulative contribution, representing the dominant drivers of the model predictions. By contrast, *NDMI*, *NDVI*, *S6*, *Int*, and *EVI* showed lower importance values (1.0–8.1%), indicating their minor contributions. Considering the distribution of SHAP values, samples with high *BI* and *SI* values were mainly associated with positive SHAP values. This means that increases in these two variables generally raised the predicted values. Conversely, samples with high *GWD*, *Clay*, *Elevation*, *NDMI*, and *NDVI* values were distributed in the negative range of SHAP values. This implies that increases in these five variables had a predominantly negative effect on the model predictions.

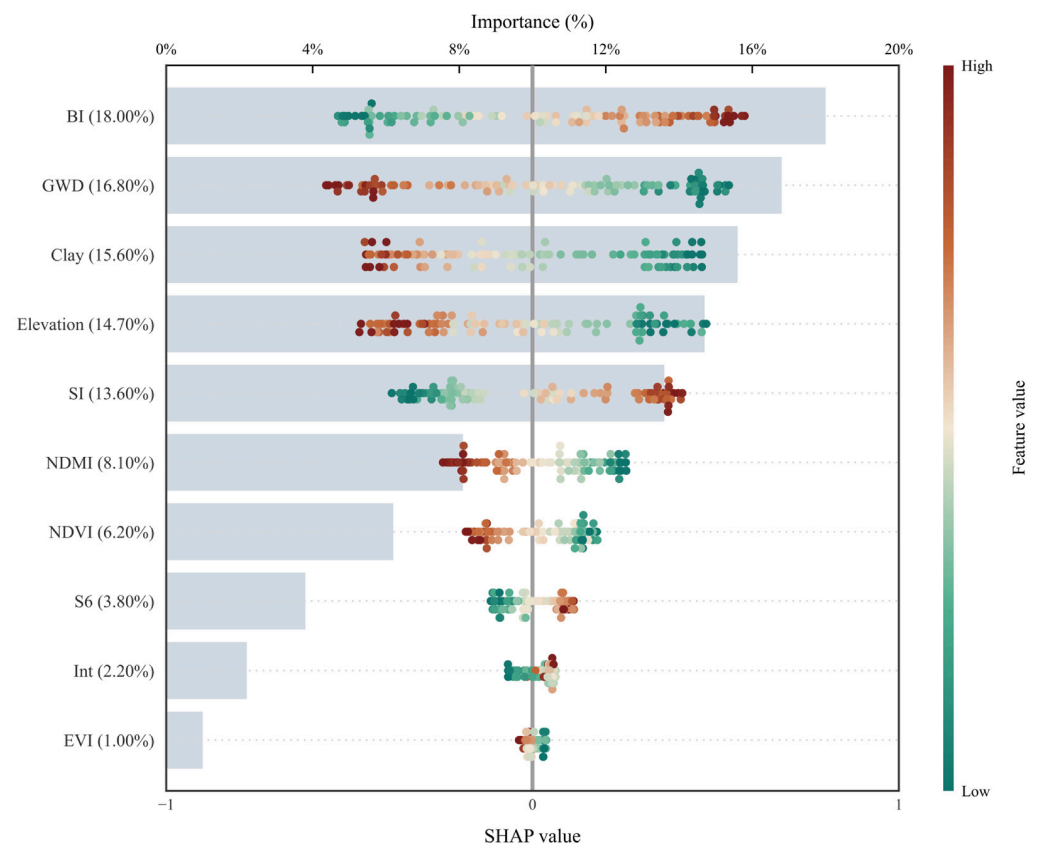


Figure 8. SHAP feature importance analysis results of the AutoML model.

5. Discussion

5.1. Controls on Surface Soil Salinity

We developed an explainable AutoML framework that integrates remote sensing, topographic, soil, and environmental data for soil salinity mapping. Pearson correlation analysis coupled with RFECV identified 10 key features from 36 candidate features: *BI*, *NDMI*, *Elevation*, *GWD*, *NDVI*, *EVI*, *SI*, *Clay*, *S6*, and *Int*. This result indicates that surface soil salinity in the study area cannot be explained by single remote sensing signals. Soil salinity is influenced by a combination of surface spectral responses, topography, groundwater, and soil texture, in agreement with previous studies [35,36].

From a remote sensing perspective, the roles of *BI* and *Int* should be interpreted as spectral responses, rather than direct causal factors of soil salt accumulation. In arid irrigated regions like the Qingtongxia Irrigation District, strong evaporation drives up-

ward capillary flow in the soil profile. As such, salts are transported from groundwater and deep soil water and concentrated in the surface layer. The resulting salt crusts, exposed soil, and low vegetation cover increase visible reflectance, thereby elevating the values of brightness indices [37]. *SI* and *S6* capture the spectral signatures of saline soils in the visible and near-infrared bands. Therefore, these two salinity indices show positive associations with predicted SSC values. *NDVI*, *EVI*, and *NDMI* mainly reflect vegetation cover and canopy water status. Elevated soil salinity inhibits the growth of crops and natural vegetation while reducing canopy water storage. This explains the overall negative contributions of these vegetation and water indices to model predictions in this study.

Furthermore, *Elevation* and *GWD* reflect the influence of topographic and groundwater conditions on soil salt transport. Low-lying areas are more likely to accumulate runoff and salts, where shallow groundwater can enhance the upward movement of water and salts under strong evaporation [38,39]. Accordingly, high *Elevation* and greater *GWD* were generally associated with lower predicted SSC values in this study. Higher *Clay* also showed a negative SHAP contribution to the model output. The negative SHAP contribution of *Clay* therefore warrants a cautious interpretation. In this study area, this pattern may reflect spatial covariation between higher clay content and lower-salinity cropland, together with associated irrigation, drainage, groundwater, and land-use conditions. Soil texture is therefore best considered as one component of the local environmental setting, and additional field evidence would help clarify whether it has an independent influence on salt accumulation.

5.2. Advantages of AutoML for Salinity Prediction

Previous studies on soil salinity mapping in arid and irrigated regions have reported validation R^2 values of approximately 0.63 to 0.91, depending on sensor type, sample size, predictor set, and validation strategy [40–42]. Against this background, the AutoML model in this study showed comparable performance, with $R^2 = 0.78$, $MAE = 1.94$ g/kg, and $RMSE = 2.65$ g/kg on the validation set. This result indicates that AutoML can support robust SSC prediction by integrating spectral, topographic, hydrological, and soil texture input features.

The improved performance of AutoML was mainly related to model screening and stacked ensemble learning. Rather than relying on a single optimized learner, AutoML first evaluated multiple candidate models and then integrated the better-performing models into a weighted ensemble. In this study, the final ensemble selected RF, XGBoost, CatBoost, and NeuralNetFastAI as base learners, with learned weights of 0.40, 0.25, 0.20, and 0.15, respectively. The ensemble was largely driven by RF, while XGBoost, CatBoost, and NeuralNetFastAI contributed complementary predictive information.

This ensemble strategy helped reduce the dependence on any single algorithm. RF is generally stable for noisy and nonlinear tabular data, whereas XGBoost and CatBoost can capture different gradient boosting patterns through distinct regularization and feature handling strategies. NeuralNetFastAI provided an additional representation of relationships among the selected predictors. By combining these learners, AutoML reduced bias associated with individual models and improved prediction stability across low, moderate, and high SSC ranges. This may explain why the predicted versus observed SSC values were closer to the 1:1 line than those from the individual optimized models.

5.3. Future Improvements

The present study provides a regional-scale assessment of surface soil salinity by integrating field observations, remote sensing variables, topographic factors, groundwater depth, and soil texture information. The 108 soil samples supported model development

and validation across the main salinity-affected zones of the Qingtongxia Irrigation District. Nevertheless, soil salinity in irrigated farmland can vary substantially over short distances because of local differences in terrain, drainage, groundwater conditions, irrigation management, and land use. A denser and more spatially balanced sampling network would therefore be helpful in future studies. This would further improve the characterization of local salinity heterogeneity and strengthen the calibration and validation of salinity prediction models.

The interpretation of soil texture variables also requires consideration of data scale. In this study, *Clay*, *Sand*, and *Silt* were obtained from the 250 m ISRIC dataset and resampled to the 30 m modelling grid to match the other predictors. This procedure ensured spatial consistency among input variables, but the soil texture data should still be regarded as broad background information rather than detailed 30 m field observations. Therefore, the SHAP contribution of *Clay* is best interpreted together with local hydrological, topographic, drainage, and land use conditions. It may reflect the role of soil texture in water retention, permeability, and salt transport, but it may also contain spatial association with broader landscape units. Future studies could incorporate higher resolution soil texture datasets or field-measured soil texture observations when available. This would help clarify the specific contribution of soil texture to salt accumulation and further improve the reliability of regional salinity mapping.

6. Conclusions

This study established an explainable AutoML framework driven by multi-source data for spatial mapping of soil salinity in an arid irrigated region—the Qingtongxia Irrigation District. Data integration across multiple sources and feature selection via a two-step procedure yielded a more parsimonious and complementary set of input variables. The selected features indicated that soil salt content (SSC) in the study area was mainly represented by spectral responses to surface salinity, groundwater depth, topography, soil texture, vegetation status, and water conditions.

Compared with individual models, including SVR, RF, GBDT, and XGBoost, AutoML achieved higher predictive performance and more stable generalization. This finding suggests that automated model search, hyperparameter optimization, and stacking ensembles can effectively address soil salinity prediction with limited sample sizes, heterogeneous predictors, and nonlinear relationships. The proposed framework reduces subjective choices in manual model development and improves the reproducibility of regional salinity mapping.

The spatial map generated by the AutoML model showed that SSC in the study area generally increased from south to north. The northern and northeastern zones were the primary risk areas for soil salinization, while localized patches of elevated SSC in the central area warrant further field verification. SHAP analysis showed that high surface brightness and pronounced salinity-related spectral responses were associated with higher predicted SSC values. Greater groundwater depth, elevation, vegetation cover, and water availability generally corresponded to lower salinity levels. These findings pave the way for identifying soil salinization risk by zone and developing differentiated irrigation and drainage management strategies.

The explainable AutoML framework improves the accuracy of soil salinity prediction while enhancing the interpretability of model outputs. This approach provides a robust tool for soil salinity mapping, salinization risk identification, and agricultural water and land resource management in arid and semiarid irrigated regions.

Author Contributions: Methodology, H.G.; Validation, Q.D.; Formal analysis, Q.D.; Investigation, H.G.; Writing—original draft, H.G.; Writing—review & editing, J.Z. and L.Z.; Supervision, L.Z.; Funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (No. 2023YFD1900103), the National Natural Science Foundation of China (NO. 42467036), the Leading Talents in the Technological Innovation Program of Ningxia (NO. 2022GKLRLX02), and the Natural Science Foundation of Ningxia (NO. 2024AAC03363).

Data Availability Statement: Data will be made available on request.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Badavath, H.; Saski, C.A. Plant Adaptation to Salinity: Physiological Pathways and Prospects for Crop Improvement—A Review. *Genes* **2026**, *17*, 253. [\[CrossRef\]](#) [\[PubMed\]](#)
2. FAO. *Global Status of Salt-Affected Soils—Main Report*; FAO: Rome, Italy, 2024.
3. Wang, T.; Dong, J.; Lyu, B.; Gao, X.; Wang, N.; Shi, Z. High-resolution global soil salinity and sodicity mapping (1980–2024): Box-Cox-based sample optimization, multi-source remote sensing features, and uncertainty quantification. *Remote Sens. Environ.* **2025**, *330*, 114991. [\[CrossRef\]](#)
4. Mzid, N.; Boussadia, O.; Albrizio, R.; Stellacci, A.M.; Braham, M.; Todorovic, M. Salinity properties retrieval from Sentinel-2 satellite data and machine learning algorithms. *Agronomy* **2023**, *13*, 716. [\[CrossRef\]](#)
5. Zhao, J.; Fan, Y.; Xuan, J.; Shi, M.; Wang, D.; Wu, H.; Bi, Y.; Li, Y. Monitoring of soil salinization and analysis of driving factors in the oasis zone of South Xinjiang. *Land* **2025**, *14*, 803. [\[CrossRef\]](#)
6. Salem, O.H.; Jia, Z. Evaluation of different soil salinity indices using remote sensing techniques in Siwa Oasis, Egypt. *Agronomy* **2024**, *14*, 723. [\[CrossRef\]](#)
7. Gao, Z.; Li, X.; Zuo, L.; Zou, B.; Wang, B.; Wang, W.J. Unveiling soil salinity patterns in soda saline-alkali regions using Sentinel-2 and SDGSAT-1 thermal infrared data. *Remote Sens. Environ.* **2025**, *322*, 114708. [\[CrossRef\]](#)
8. Sun, J.; Song, W.; Sun, X.; Cui, J.; Ma, H. Analysis of Soil Salinization Characteristics in Coastal Area of Panjin City, Liaoning Province. *Water* **2025**, *17*, 2666. [\[CrossRef\]](#)
9. Zhou, Q.; Zhang, Y.; Liu, Z.; Wang, D.; Chen, H.; Liu, P. Integrating both driving and response environmental variables to enhance soil salinity inversion. *Agronomy* **2025**, *15*, 1995. [\[CrossRef\]](#)
10. Mao, Y.; Splinter, K.D. Application of SAR-Optical fusion to extract shoreline position from Cloud-Contaminated satellite images. *ISPRS J. Photogramm. Remote Sens.* **2025**, *220*, 563–579. [\[CrossRef\]](#)
11. Lv, B.; Luo, Y.; Ni, J.; Zhao, S.; Liang, J.; Liu, Y.; Zhang, Q. MMFF: Multiview and multi-level feature fusion method within limited sample conditions for SAR image target recognition. *ISPRS J. Photogramm. Remote Sens.* **2025**, *224*, 302–316. [\[CrossRef\]](#)
12. Xu, H.; Sun, H.; Xu, Z.; Wang, Y.; Zhang, T.; Wu, D.; Gao, J. kNDMI: A kernel normalized difference moisture index for remote sensing of soil and vegetation moisture. *Remote Sens. Environ.* **2025**, *319*, 114621. [\[CrossRef\]](#)
13. Sahbeni, G.; Ngabire, M.; Musyimi, P.K.; Székely, B. Challenges and opportunities in remote sensing for soil salinization mapping and monitoring: A review. *Remote Sens.* **2023**, *15*, 2540. [\[CrossRef\]](#)
14. Li, Z.; Li, X. Global trends and research hotspots on the improvement of saline soil: Insights from bibliometric analysis. *Agric. Water Manag.* **2025**, *321*, 109923. [\[CrossRef\]](#)
15. Mukhamediev, R.I.; Terekhov, A.; Amirgaliyev, Y.; Popova, Y.; Malakhov, D.; Kuchin, Y.; Sagatdinova, G.; Symagulov, A.; Muhamedijeva, E.; Gricenko, P. Using pseudo-color maps and machine learning methods to estimate long-term salinity of soils. *Agronomy* **2024**, *14*, 2103. [\[CrossRef\]](#)
16. Khatiriolyaee, M.; Yang, L.; Pazzi, R.W. A comprehensive review of traffic prediction: From traditional machine learning to AutoML. *Neurocomputing* **2025**, *666*, 132148.
17. Liu, X.; Hu, Y.; Li, X.; Du, R.; Xiang, Y.; Zhang, F. An interpretable model for salinity inversion assessment of the South bank of the Yellow River based on optuna hyperparameter optimization and XGBoost. *Agronomy* **2024**, *15*, 18. [\[CrossRef\]](#)
18. Sun, Y.; Wang, L.; Yang, S.; Qu, Z.; Zhang, D. Optimizing Irrigation and Drainage Practices to Control Soil Salinity in Arid Agroecosystems: A Scenario-Based Modeling Approach Using SaltMod. *Agronomy* **2025**, *15*, 2239. [\[CrossRef\]](#)
19. Zhang, Q.; Gong, M.; Luo, W.; He, J.; Ding, J.; Jiang, P. A dataset on topsoil salinization characteristics in the Tailan River Irrigation District on the northern margin of the Tarim Basin in Xinjiang. *Sci. Data* **2026**, *13*, 604. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Abuelgasim, A.; Ammad, R. Mapping soil salinity in arid and semi-arid regions using Landsat 8 OLI satellite data. *Remote Sens. Appl. Soc. Environ.* **2019**, *13*, 415–425. [\[CrossRef\]](#)
21. Allbed, A.; Kumar, L.; Sinha, P. Mapping and modelling spatial variation in soil salinity in the Al Hassa Oasis based on remote sensing indicators and regression techniques. *Remote Sens.* **2014**, *6*, 1137–1157. [\[CrossRef\]](#)
22. Abbas, A.; Khan, S. Using remote sensing techniques for appraisal of irrigated soil salinity. In Proceedings of the International Congress on Modelling and Simulation (MODSIM), Christchurch, New Zealand, 10–13 December 2007; pp. 2632–2638.

23. Yahiaoui, I.; Douaoui, A.; Zhang, Q.; Ziane, A. Soil salinity prediction in the Lower Cheliff plain (Algeria) based on remote sensing and topographic feature analysis. *J. Arid. Land* **2015**, *7*, 794–805. [[CrossRef](#)]
24. Phiri, D.; Simwanda, M.; Salekin, S.; Nyirenda, V.R.; Murayama, Y.; Ranagalage, M. Sentinel-2 data for land cover/use mapping: A review. *Remote Sens.* **2020**, *12*, 2291. [[CrossRef](#)]
25. Hengl, T.; Mendes de Jesus, J.; Heuvelink, G.B.; Ruiperez Gonzalez, M.; Kilibarda, M.; Blagotić, A.; Shangguan, W.; Wright, M.N.; Geng, X.; Bauer-Marschallinger, B. SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE* **2017**, *12*, e0169748. [[CrossRef](#)] [[PubMed](#)]
26. Hayakawa, Y.S.; Oguchi, T.; Lin, Z. Comparison of new and existing global digital elevation models: ASTER G-DEM and SRTM-3. *Geophys. Res. Lett.* **2008**, *35*, L17404. [[CrossRef](#)]
27. Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson correlation coefficient. In *Noise Reduction in Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–4.
28. Ma, W.; Zhang, X.; Shen, Y.; Xie, J.; Zuo, G.; Zhang, X.; Jin, T. Incorporating recursive feature elimination and decomposed ensemble modeling for monthly runoff prediction. *Water* **2024**, *16*, 3102. [[CrossRef](#)]
29. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [[CrossRef](#)] [[PubMed](#)]
30. Belgiu, M.; Drăguț, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [[CrossRef](#)]
31. Liang, W.; Luo, S.; Zhao, G.; Wu, H. Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms. *Mathematics* **2020**, *8*, 765. [[CrossRef](#)]
32. Niazkar, M.; Menapace, A.; Brentan, B.; Piraei, R.; Jimenez, D.; Dhawan, P.; Righetti, M. Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023). *Environ. Model. Softw.* **2024**, *174*, 105971. [[CrossRef](#)]
33. He, X.; Zhao, K.; Chu, X. AutoML: A survey of the state-of-the-art. *Knowl.-Based Syst.* **2021**, *212*, 106622. [[CrossRef](#)]
34. Li, Z. Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Comput. Environ. Urban Syst.* **2022**, *96*, 101845. [[CrossRef](#)]
35. Xie, J.; Shi, C.; Liu, Y.; Wang, Q.; Zhong, Z.; He, S.; Wang, X. Soil salinization prediction through feature selection and machine learning at the irrigation district scale. *Front. Earth Sci.* **2025**, *12*, 1488504. [[CrossRef](#)]
36. Jiang, Z.; Hao, Z.; Ding, J.; Miao, Z.; Zhang, Y.; Alimu, A.; Jin, X.; Cheng, H.; Ma, W. Weighted variable optimization-based method for estimating soil salinity using multi-source remote sensing data: A case study in the Weiku Oasis, Xinjiang, China. *Remote Sens.* **2024**, *16*, 3145. [[CrossRef](#)]
37. Chen, Q.; Vaudour, E.; Richer-de-Forges, A.C.; Arrouays, D. Spectral indices in remote sensing of soil: Definition, popularity, and issues. A critical overview. *Remote Sens. Environ.* **2025**, *329*, 114918. [[CrossRef](#)]
38. Li, W.; Kang, S.; Du, T.; Ding, R.; Zou, M. Optimal groundwater depth and irrigation amount can mitigate secondary salinization in water-saving irrigated areas in arid regions. *Agric. Water Manag.* **2024**, *302*, 109007. [[CrossRef](#)]
39. Lv, T.; Kang, S.; Liu, Y.; Bian, M.; Tong, L.; Li, W. Analysis of driving factors of soil salinity in Southern Xinjiang irrigation areas under dry-sowing and wet-emerging conditions. *Agric. Water Manag.* **2025**, *322*, 109962. [[CrossRef](#)]
40. Chaaou, A.; Ait-Ichou, H.; Hachemy, S.E.; Chikhaoui, M.; Naimi, M.; Hssaisoune, M.; El Hafyani, M.; Ait Brahim, Y.; Bouchaou, L. Mapping soil salinity using machine learning and remote sensing data in semi-arid croplands. *Front. Soil Sci.* **2025**, *5*, 1653400. [[CrossRef](#)]
41. Wang, N.; Xue, J.; Peng, J.; Biswas, A.; He, Y.; Shi, Z. Integrating remote sensing and landscape characteristics to estimate soil salinity using machine learning methods: A case study from Southern Xinjiang, China. *Remote Sens.* **2020**, *12*, 4118. [[CrossRef](#)]
42. Wang, J.; Peng, J.; Li, H.; Yin, C.; Liu, W.; Wang, T.; Zhang, H. Soil salinity mapping using machine learning algorithms with the Sentinel-2 MSI in arid areas, China. *Remote Sens.* **2021**, *13*, 305. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.