scientific reports

Check for updates

OPEN

Classification of deep and shallow groundwater wells based on machine learning in the Hebei Plain North China

Hang Zhou[1], Chu Wu[2], Baoqi Li[1], Chuiyu Lu[2]✉, Yong Zhao[2] & Ziyue Zhao[3]✉

Accurately determining the extraction volumes from various aquifers is crucial for effectively managing groundwater overexploitation. A key initial step in quantifying extracted groundwater volumes involves the classification of groundwater wells as either deep or shallow. This study evaluated 881,872 groundwater wells in the Hebei Plain, applying machine learning techniques to classify wells with unknown depths. Through the hydrogeological borehole data, the groundwater wells with known depth are divided into deep wells and shallow wells. Four machine learning algorithms—Random Forest, Support Vector Machine, Logistic Regression, and Naive Bayes—were employed to classify groundwater wells with unknown depths. The accuracy of these models was validated using known-depth well classifications. The results reveal that the Random Forest algorithm exhibited the highest performance among the models, achieving an overall accuracy of 91.23%. According to the Random Forest model, 43.51% of groundwater wells with unknown depths were classified as deep, while 56.49% were classified as shallow. The study also found that wells in areas where salinity exceeds 2 g/L are primarily deep groundwater wells. These findings provide valuable technical insight for groundwater well decommissioning and facilitate the assessment of extracted volumes of deep and shallow groundwater.

Keywords Groundwater wells, Deep and shallow groundwater wells classification, Machine learning, Hebei Plain

The Hebei Plain is among the most severely affected regions in the world for overexploitation of groundwater. This has led to the formation of multiple shallow groundwater decline funnels at the piedmont alluvial plain and a series of composite deep groundwater funnels across the central alluvial and coastal plain. Such overuse has given rise to a range of geological and ecological issues, including land subsidence and seawater intrusion[1–3].

In 2014, comprehensive measures were implemented to manage groundwater overexploitation in Hebei Province. These measures included the establishment of a dual control system targeting both the water level and the volume of groundwater extraction in these overexploited regions. Discerning the precise extraction volumes of both shallow and deep groundwater within the Hebei Plain is imperative. Accurate categorization of groundwater wells into deep or shallow is crucial before setting up detailed measurements of groundwater extraction volumes, especially for water used in agricultural irrigation. This categorization is a key step in evaluating the volumes extracted from different aquifers.

Previous studies of groundwater wells have predominantly concentrated on issues such as groundwater contamination[4–6], stability assessments for groundwater supply systems during the dry season[7,8], effects of lowered aquifer water levels on groundwater well performance[9,10], establishing a quantitative relationship between irrigation energy consumption and water extraction volume[11], and approaches to simplify predictions for the location and depth of new groundwater wells using hydrogeological and construction data[12]. Research is notably deficient regarding the vertical zoning of groundwater wells: that is, distinguishing between deep and shallow groundwater wells. Therefore, there is an urgent need to propose a scientific and rational method to classify the deep and shallow groundwater wells.

[1]Key Laboratory of Roads and Railway Engineering Safety Control (Shijiazhuang Tiedao University), Ministry of Education, Shijiazhuang 050043, China. [2]State Key Laboratory of Water Cycle Simulation and Regulation, China Institute of Water Resources and Hydropower Research, Beijing 100038, China. [3]Hebei Provincial Water Affairs Center, Shijiazhuang 050043, China. ✉email: cylu@iwhr.com; 1173347578@qq.com

Over the years, millions of groundwater wells have been constructed across the Hebei Plain, with completion dates spanning a broad range. The lack of professional competence among those involved in groundwater well construction has frequently resulted in poorly documented depth information. Wells with known depths can be categorized through the analysis of hydrogeological boreholes. Nonetheless, there is a deficiency in effective methods for classifying groundwater wells that do not have documented depths into their correct categories.

Machine learning classification techniques have been widely adopted in a variety of fields[13–16]. The Random Forest (RF) algorithm, known for its exceptional predictive capabilities, has achieved success in numerous industries[17]. Support Vector Machines (SVM), celebrated for their impressive generalization ability and efficiency in deriving optimal solutions for classification tasks, have drawn significant attention[18]. Logistic Regression (LR) serves as a foundational statistical model for binary outcome regression and classification tasks[19]. Furthermore, the Naive Bayes (NB) algorithm, a staple in statistical learning, is often recommended for benchmarking against other techniques[20]. This study employs these four machine learning strategies—RF, SVM, LR, and NB—to classify groundwater wells with unknown depths as either deep or shallow, additionally evaluating the accuracy of each method.

Currently, there is a lack of methods to classify groundwater wells with unknown depths as either deep or shallow wells. Therefore, this paper aims to categorize groundwater wells based on the following objectives: 1. To classify groundwater wells with known depths as deep or shallow by utilizing data from available hydrogeological boreholes.

2. To deploy machine learning techniques to develop classifiers capable of distinguishing deep from shallow groundwater wells, using existing classification outcomes, groundwater well characteristics, and hydrogeological conditions such as salinity of shallow groundwater, abundance of deep and shallow groundwater, and whether the location is within an overexploited zone.

3. To evaluate and compare four machine learning algorithms to both identify the most accurate model for classifying groundwater wells and to categorize wells with unknown depths as deep or shallow.

Materials and methods

Study area

The Hebei Plain experiences a temperate East Asian monsoon climate[21] with average annual precipitation ranging from 500 to 600 mm[22] and a temperature spectrum spanning from 1.8 to 14.2 °C. Owing to significant water resource limitations, Hebei has per capita water resources that are only 11% of the national average in China. Water demand in Hebei Province primarily stems from agricultural and industrial activities, which together account for about 70% of its total water consumption and are heavily dependent on groundwater. The extensive extraction of groundwater within this region has led to the creation of two distinct overexploited areas: a shallow groundwater overexploited zone spanning 36,669.5 km^2 and a deep groundwater overexploited region encompassing 42,157.8 km^2. These two areas intersect, forming an overlapping zone that measures 9134 km^2.

The piedmont alluvial plain features the boundary of the saline water zone along its western edge (Fig. 1). Within this area, the depth to the bottom of the saline layer typically ranges from 40 to 80 m; however, in certain locations, this depth can decrease to below 40 m or increase to around 100 m. Progressing from west to east toward the central parts of the central alluvial and coastal plain, the depth to the bottom of the saline water layer gradually increases, registering between 80 and 120 m. The aquifer grouping under discussion can be subdivided into four distinct hydrogeological layers, designated as Aquifers I–IV[23] (Fig. 2), stratified from the uppermost to the lowermost levels[24]. In regions where the groundwater is entirely fresh, the shallow aquifer system encompasses both Aquifers I and II. However, within saline water zone, Aquifer I alone constitutes the shallow aquifer system. For freshwater zone, the deep aquifer system contains Aquifers III and IV, while in areas with saline waters, this system extends to include Aquifer II in addition to Aquifers III and IV. Generally, the transition zone from freshwater to saline water corresponds with the demarcation between the piedmont alluvial and central alluvial plains; the western side is characterized by fresh groundwater, transitioning to saline groundwater eastward[25].

The abundance of shallow groundwater in the piedmont alluvial plain ranges from 3000 to 5000 m^3/d, whereas it ranges from 100 to 500 m^3/d in the central alluvial and coastal plain. The deep groundwater abundance is high overall, but the deep subsurface aquifer system is missing or discontinuous near the mountain front.

Methods

The dataset used in this study includes records for 881,872 operational groundwater wells on the Hebei Plain as of 2023. Of these, 285,755 groundwater wells have statistical data that includes depth, while 596,117 groundwater wells do not have this depth information available. Additionally, the study integrates data from 1127 hydrogeological boreholes across the Hebei Plain, which offer detailed information on the upper and lower boundaries of each aquifer group.

Regarding the hydrogeological conditions of the Hebei Plain, groundwater wells are classified into deep and shallow categories as follows: Deep wells are those that extract water from deep aquifer systems. Shallow wells are those that extract water from shallow aquifer systems.The specific classification methods of deep and shallow groundwater wells in the Hebei Plain are as follows:

Classification of groundwater wells with known depths

Hydrogeological borehole data are subjected to Kriging interpolation to process the information which is subsequently used for the categorization of groundwater wells with known depths. The classification criteria include the groundwater wells' depth as well as their geographic coordinates, specified by latitude and longitude. The methodology for these calculations consists of the following steps:

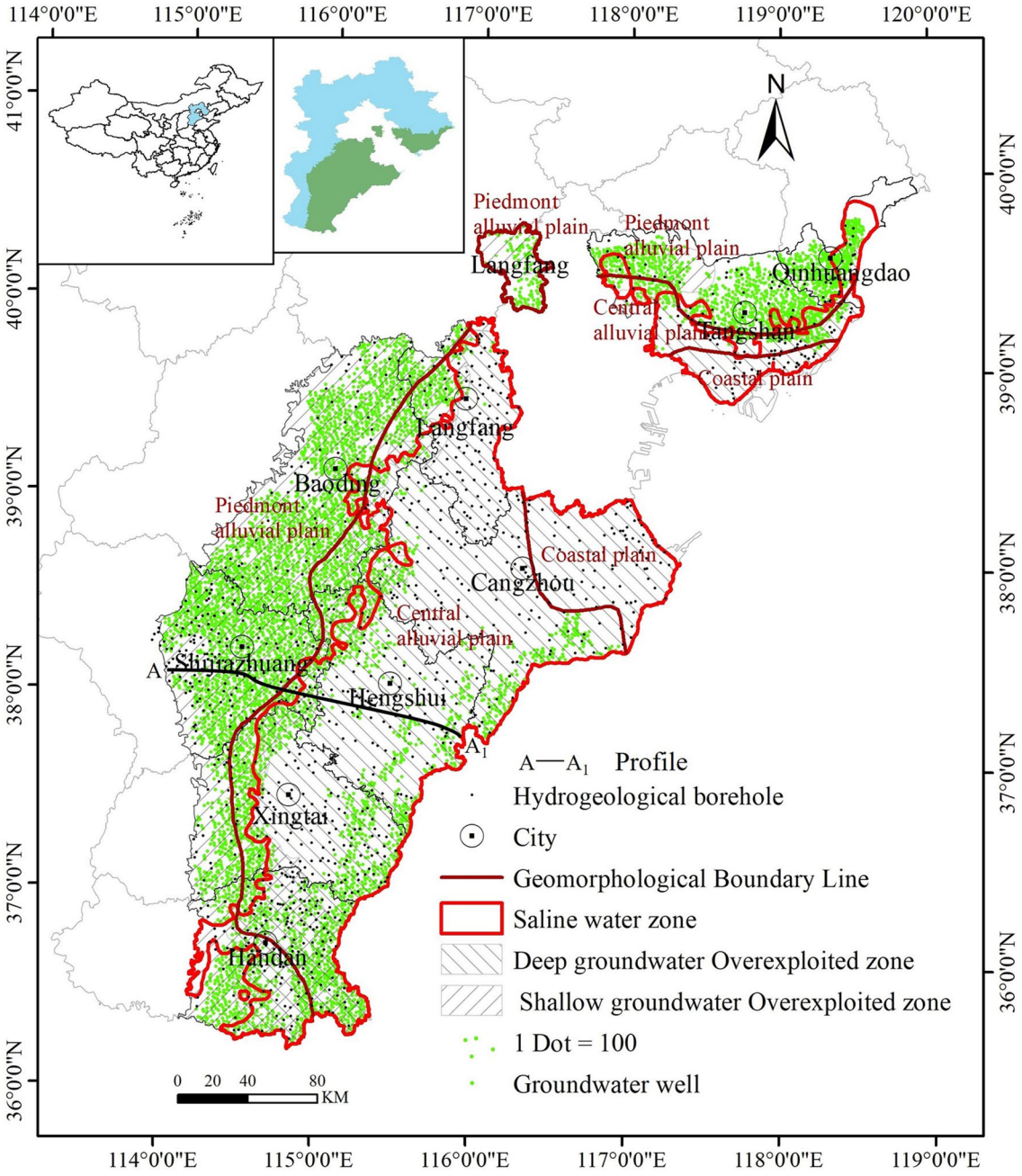

Figure 1. Distribution of groundwater wells and hydrogeological boreholes in Hebei Plain and the range of shallow groundwater saline areas. The figure was created by the author using ArcMap 10.7.

(1) The Kriging interpolation method is applied to the hydrogeological borehole data to determine the bottom depths of the aquifer. Kriging is a widely used spatial interpolation technique in Geographic Information Systems (GIS)[26] that predicts or estimates values at unknown locations based on the spatial correlation between known data points:

$$Z(X_0) = \sum_{i=1}^{n} \lambda_i \cdot Z(X_i) \tag{1}$$

where $Z(X_0)$ is the predicted value at unknown position x, n denotes the number of known points, $Z(X_i)$is the observed value of the ith known point, and λ_i is the interpolation weight, which is calculated by spatial correlation and used for weighted summation of the observed values of known points.

(2) Project the groundwater wells into the aquifer depth map according to latitude and longitude, and set the IF function:

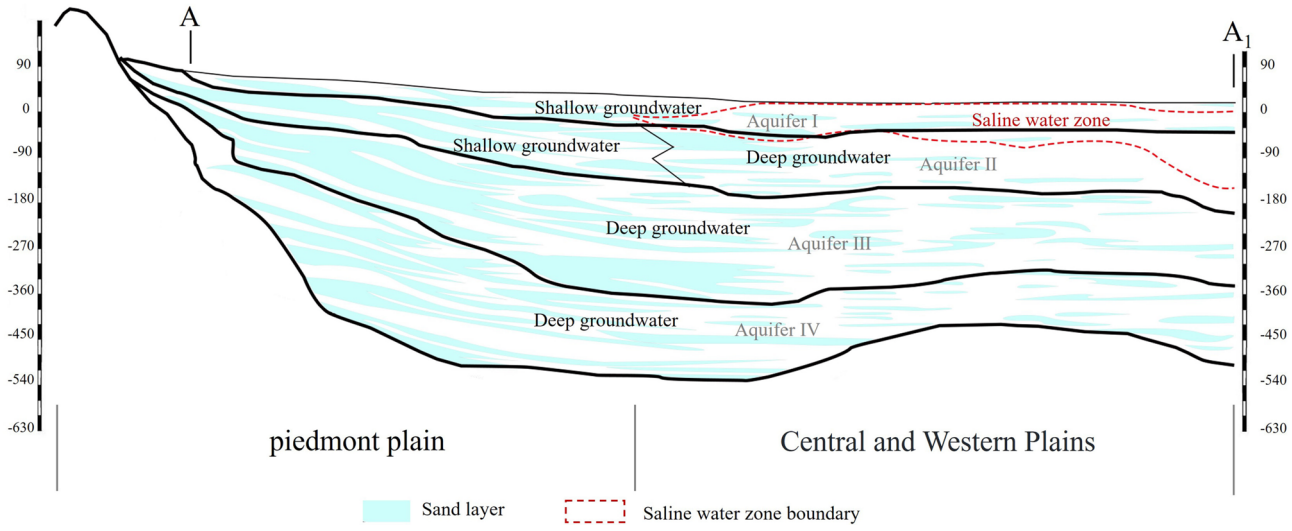

Figure 2. A–A_1 Hydrogeological section.

$$\begin{cases} h_w < h_I, \textit{Groundwater wells belong to aquifer I} \\ h_I < h_w < h_{II}, \textit{Groundwater wells belong to aquifer II} \\ h_{II} < h_w < h_{III}, \textit{Groundwater wells belong to aquifer III} \\ h_{IV} < h_w, \textit{Groundwater wells belong to aquifer IV} \end{cases} \quad (2)$$

Among them, h_I, h_{II}, h_{III}, and h_{IV} are the bottom depths of the four aquifers, while *hw* is the depth of the groundwater wells.

(3) Determine whether the groundwater wells are considered deep or shallow based on the classification of the aquifer to which they belong and the corresponding coordinates.

Classification of groundwater wells with unknown depth

Machine learning techniques are utilized to develop classification models by leveraging existing classification outcomes, groundwater well attributes, and the specific hydrogeological conditions of their respective locales. These conditions encompass factors such as the salinity of shallow groundwater, the abundance of both deep and shallow groundwater, and the determination of whether the site is situated within an overexploited zone. The goal of the models is to differentiate between deep and shallow groundwater wells. Upon determining the most effective model, it is then applied to classify groundwater wells with unknown depths. The machine learning models utilized include:

(1) Support Vector Machine (SVM): SVM[27,28] is a binary linear classification technique in machine learning that separates classes by the largest margin (called the optimal hyperplane) between instances on the boundary lines (known as support vectors). SVM can define nonlinear decision boundaries in high-dimensional variable space by solving a quadratic optimization problem[29,30]. The core idea of the SVM algorithm is to divide the dataset into different categories by finding the maximum-margin hyperplane. During the process of finding this hyperplane, the SVM selects support vectors, which are the data points nearest to the hyperplane. Through the analysis of support vectors, the distribution of data can be inferred, which enables the classification of new data. The SVM has the advantages of strong generalization ability, good model robustness, and applicability to both linear and nonlinear problems, making it widely applicable with broad prospects.

(2) Random Forest (RF): RF was developed by Breiman[31] and is an ensemble algorithm based on multiple decision trees[32]. The classification of an unknown type is determined by voting or averaging the outputs of all the decision trees[33]. Specifically, Random Forest (RF) samples the training data randomly using techniques such as bootstrapping and then trains several different decision trees, for instance, M decision trees. Ultimately, the predictions of the M decision trees are combined through methods like majority voting and averaging to produce the final output[34]. The fundamental concept behind the algorithm is to enhance the model's generalization ability through the introduction of randomness. The RF algorithm is robust with respect to feature selection and can achieve high accuracy without the need for extensive feature selection, making it an efficient supervised machine learning algorithm[17].

(3) Naive Bayes (NB): The NB classification method is based on Bayes' theorem and adopts a probabilistic reasoning approach. NB estimates the conditional probability of a class by 'naively' assuming that the inputs to a given class are independent of each other. This assumption creates a discriminant function that is represented by the product of the joint probabilities, signifying the probability that a particular class is true given the input. NB simplifies the problem of distinguishing classes by calculating the class-conditional

marginal densities, which represent the likelihood that a given sample belongs to one of the potential target classes[20].

(4) Logistic Regression (LR): LR is a multivariate statistical analysis method[35] used for binary classification. It studies the relationship between a dichotomous outcome variable (dependent variable) and one or more predictor variables (independent variables). LR is considered a generalized linear model adapted for dichotomous outcomes. Since the dependent variable takes binary values (usually set to 0 or 1), the logistic function, also known as the Sigmoid function, is introduced into linear regression to transform the output. This effectively converts LR into a normalized linear regression model through the Sigmoid function.

Characterization and feature encoding of groundwater wells for machine learning

The hydrogeological conditions at the location of groundwater wells, such as the abundance of deep groundwater, abundance of shallow groundwater, and salinity of shallow groundwater, are considered as properties of the wells and are encoded using sequential encoding (Table 1). These properties serve as features for machine learning.

Whether the groundwater well is located in an overexploited zone and the groundwater well characteristics do not have a continuous relationship. Therefore, they are encoded separately using binary encoding (Table 2) as features for machine learning.

Finally, deep groundwater wells are encoded as 1 and shallow groundwater wells as 0, serving as the labels for the machine learning model.

Category	Description	Code
abundance of deep groundwater (m^3/d)	0	0
	< 100	1
	100–500	2
	500–1000	3
	1000–3000	4
	3000–5000	5
	> 5000	6
abundance of shallow groundwater (m^3/d)	0	0
	< 100	1
	100–500	2
	500–1000	3
	1000–3000	4
	3000–5000	5
	> 5000	6
salinity of shallow groundwater (g/L)	0–1	0
	1–2	1
	2–3	2
	3–5	3
	< 5	4

Table 1. Sequential encoding of groundwater well characteristics.

Category	Description	Code
Whether located in an overexploited zone	Wells located in a shallow groundwater overexploited zone	0
	Wells located in a deep groundwater overexploited zone	1
	Wells located in a mixed groundwater overexploited zone	10
	Wells not in an overexploited zone	11
Groundwater well characteristics	Agricultural wells	0
	Industrial enterprise wells	1
	Urban centralized water supply wells	10
	Urban domestic wells	11
	Service industry wells	100
	Rural water supply factory wells	101
	Rural domestic wells	110
	Ground source heat pumps wells	111

Table 2. Binary encoding of groundwater well characteristics.

Model performance evaluation

In machine learning, the measurement and evaluation of models is crucial. Utilizing quantitative numerical evaluation metrics and methods enables the swift selection of an optimal model for training and learning from the data, thereby enhancing the effectiveness of the modeling and parameter tuning process. The primary evaluation methods employed in this study include three-fold cross-validation, confusion matrix, kappa coefficient, overall accuracy, precision, recall and F1-score.

(1) Three-fold cross-validation.

Cross-validation is a statistical analysis method used to assess a model's performance. The basic idea is to divide the original data into parts, where one part serves as the training set and another as the validation set. First, the model is trained with the training set, and then the validation set is used to test the trained model, in order to evaluate the model's performance. In k-fold cross-validation (KCV), the original data is divided into k groups. One subset is extracted as the validation set without repetition, and the remaining k−1 groups are combined as the training set. Each time, the k−1 combined subsets serve as the training set, and this process is repeated for all k groups. The three-fold cross-validation method was utilized for the model in this study.

(2) Confusion matrix.

In this study, we carry out binary classification of groundwater wells, which is a problem where instances are classified into two categories: target and non-target. After making predictions with the trained model, the results fall into one of four cases:

True Positive (TP): The true label is the target class, and the prediction is also the target class. True Negative (TN): The true label is the non-target class, and the prediction is also the non-target class. False Positive (FP): The true label is the non-target class, but the prediction is incorrectly made as the target class. False Negative (FN): The true label is the target class, but the prediction is incorrectly made as the non-target class. In this paper, deep wells are designated as the target class, while shallow wells are considered the non-target class.

The confusion matrix is a tool used to illustrate the misclassifications between the predicted outcomes and the actual labels. It tabulates the number of correct and incorrect predictions, which demonstrates the errors made by the model during classification—in terms of both type and quantity of instances. The rows of the confusion matrix correspond to the actual labels, while the columns correspond to the predicted ones, providing a clear visualization of classification accuracy as detailed in Table 3.

(3) Kappa Coefficient.

According to the confusion matrix, we can calculate the Kappa coefficient, which is an important measure for evaluating model performance on unbalanced datasets. The Kappa coefficient penalizes the bias toward the majority class. The greater the imbalance that affects the model classification, the lower the Kappa coefficient will typically be. It is an important measure in the evaluation of multiclass classification when dealing with imbalance issues. The formula for the Kappa coefficient is as follows:

$$\text{Kappa coefficient} = \frac{P_0 - P_e}{1 - P_0} \tag{3}$$

where P_0 is the proportion of instances correctly predicted (sum of the diagonal elements of the confusion matrix) to the total number of instances (sum of all elements in the matrix). *Pe* represents the expected proportion of correct predictions by chance, which is calculated as the sum of the products of the actual and predicted frequencies for each class, divided by the square of the total number of instances.

(4) Overall accuracy.

		Predicted	
		Positive	**Negative**
Actual	Positive	TP	FN
	Negative	FP	TN

Table 3. Confusion matrix charted by the predicted and actual classification.

Overall accuracy refers to the proportion of correctly predicted instances across all categories.

$$\text{Overall accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

(5) Precision.

Precision indicates the proportion of correctly predicted positive observations in the predicted positives for a category, i.e., the probability that a predicted positive is actually positive.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

(6) Recall.

Recall is the proportion of actual positive cases that were correctly identified by the model, and a higher recall signifies that the model is correctly predicting more actual positive cases.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

(7) F1-score.

The F1-score is a statistical measure used to assess a classification model's accuracy, balancing precision and recall, and is particularly useful for imbalanced datasets. It is the harmonic mean of precision and recall, with a value ranging from 0 to 1, where 1 signifies perfect precision and recall, and 0 represents the worst possible performance.

$$F1 - \text{score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{7}$$

Results and discussion

Classification of deep and shallow groundwater wells with known depths

This study employs Kriging interpolation of data from 1127 hydrogeological boreholes distributed across the plain to map the base depths of four distinct aquifers, illustrated in Fig. 3. The key parameters and configuration of Kriging interpolation are presented in the Table 4. The first layer represents the surface, while the 2nd, 3rd, 4th, and 5th layers represent the basement depths of the I, II, III, and IV aquifers, respectively. The depths of four aquifers in the piedmont alluvial plain are similar. However, at the confluence of the piedmont alluvial and central alluvial plains, there is a notable discontinuity in the depths of the confining layers for Aquifers II, III, and IV. Specifically, Aquifer I features a basal depth ranging from 5 to 53 m, Aquifer II from 43 to 217 m, Aquifer III from 63 to 439 m, and Aquifer IV has its base depth ranging from 72 to 619 m.

Subsequently, the study categorizes groundwater wells as either deep or shallow based on their geographical location, depth, and the depths of the aquifer's basal layers. The resulting classifications are displayed in Table 5. It shows that agricultural groundwater wells have the lowest incidence of being classified as deep, at 44.28%, whereas urban centralized water supply wells have the highest, reaching 82.17%. Except for agricultural wells, the majority of other groundwater wells tend to tap into deep groundwater sources.

Figure 4 displays the spatial distribution of deep and shallow groundwater wells. The majority of shallow groundwater wells are located within the piedmont alluvial plain, an area rich in fresh shallow groundwater. Conversely, the shallow groundwater found in the central alluvial and coastal plain is characteristically saline, with their water abundance often falling below 500 m^3/day. The abundance of deep groundwater sources typically surpasses 1000 m^3/day. Consequently, deep groundwater wells are predominantly located in the central alluvial and coastal plain.

To analyze the distribution patterns of deep and shallow groundwater wells, this study examines the depths of groundwater wells. Figure 5 depicts the distribution of well depths across the Hebei Plain. Although the exclusion of groundwater wells with unknown depths may introduce potential errors, obvious trends can still be observed.

In the piedmont alluvial plain, the depths of groundwater wells typically range from 40 to 100 m. However, in some areas, these depths may exceed 100 m. As one moves from west to east into the central part of the central alluvial and coastal plain, the depth of groundwater wells gradually increases, reaching 100–150 m. The central alluvial and coastal plain begin in the central part of Cangzhou City, where the depth of groundwater wells is approximately 200–300 m, and this depth increases gradually as one moves eastward. In the southeastern coastal plain, the depth of some groundwater wells can surpass 400 m, sometimes reaching even greater depths.

In the freshwater zone, the completion depth of groundwater wells is independent of the salinity levels of the shallow groundwater, focusing solely on the proper aquifer selection. Therefore, groundwater wells in these areas often draw water from Aquifers I or II. In contrast, in the saline water zone, to circumvent water quality challenges, groundwater wells are often drilled below the saline water stratum. Given that the base of the saltwater stratum usually coincides with the bottom of Aquifer II, groundwater wells in the central alluvial and coastal plain must be drilled to considerable depths. Consequently, groundwater wells in these areas are commonly deep groundwater wells.

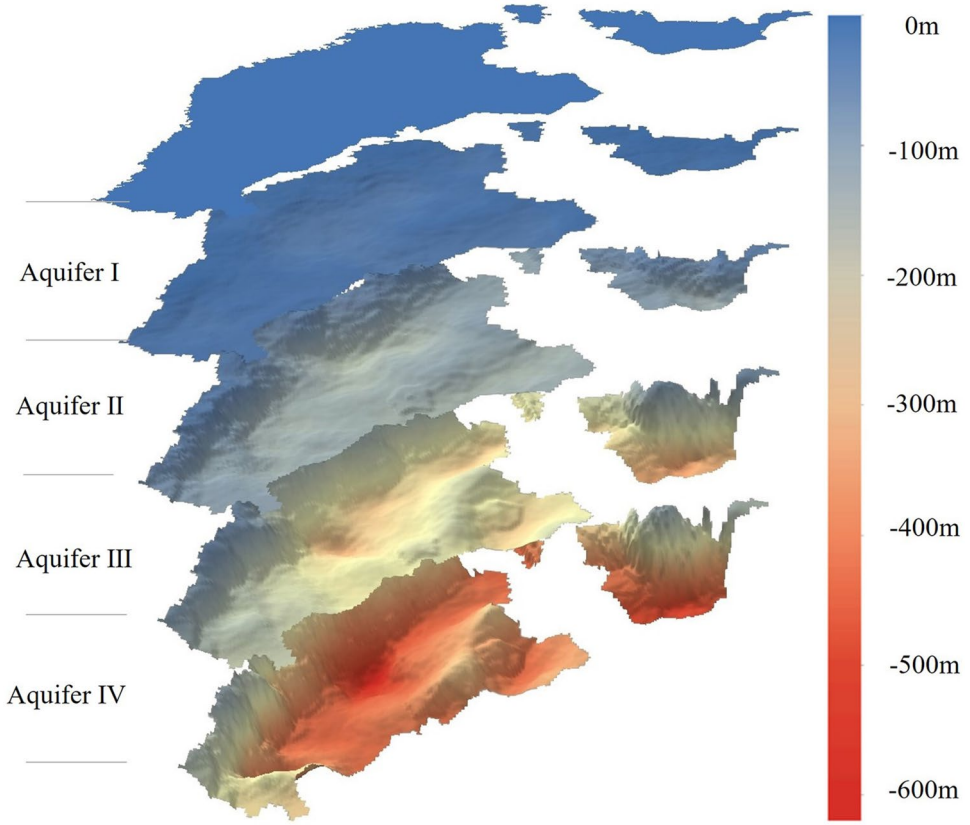

Figure 3. Interpolated aquifer base surfaces using the Kriging geostatistical method.

Parameter name	Parameter value
Interpolation method	Ordinary kriging
Variogram model	Spherical model
Output cell size	1694.51372 m
Search radius setting	12 points

Table 4. The key parameters and configuration of Kriging interpolation.

Category	Deep wells	Shallow wells	Total wells	Proportion of deep wells in total wells (%)
Urban centralized water supply wells	139	60	199	82.17
Urban domestic wells	17	10	27	69.54
Service industry wells	11	8	19	78.87
Industrial enterprise wells	633	114	747	67.75
Rural water supply factory wells	836	436	1272	76.91
Rural domestic wells	2471	970	3441	67.99
Agricultural wells	134,583	145,406	279,989	44.28
Ground source heat pumps wells	49	12	61	79.12
Total wells	138,739	147,016	285,755	48.55

Table 5. Distribution of deep and shallow groundwater wells with known depths across different categories.

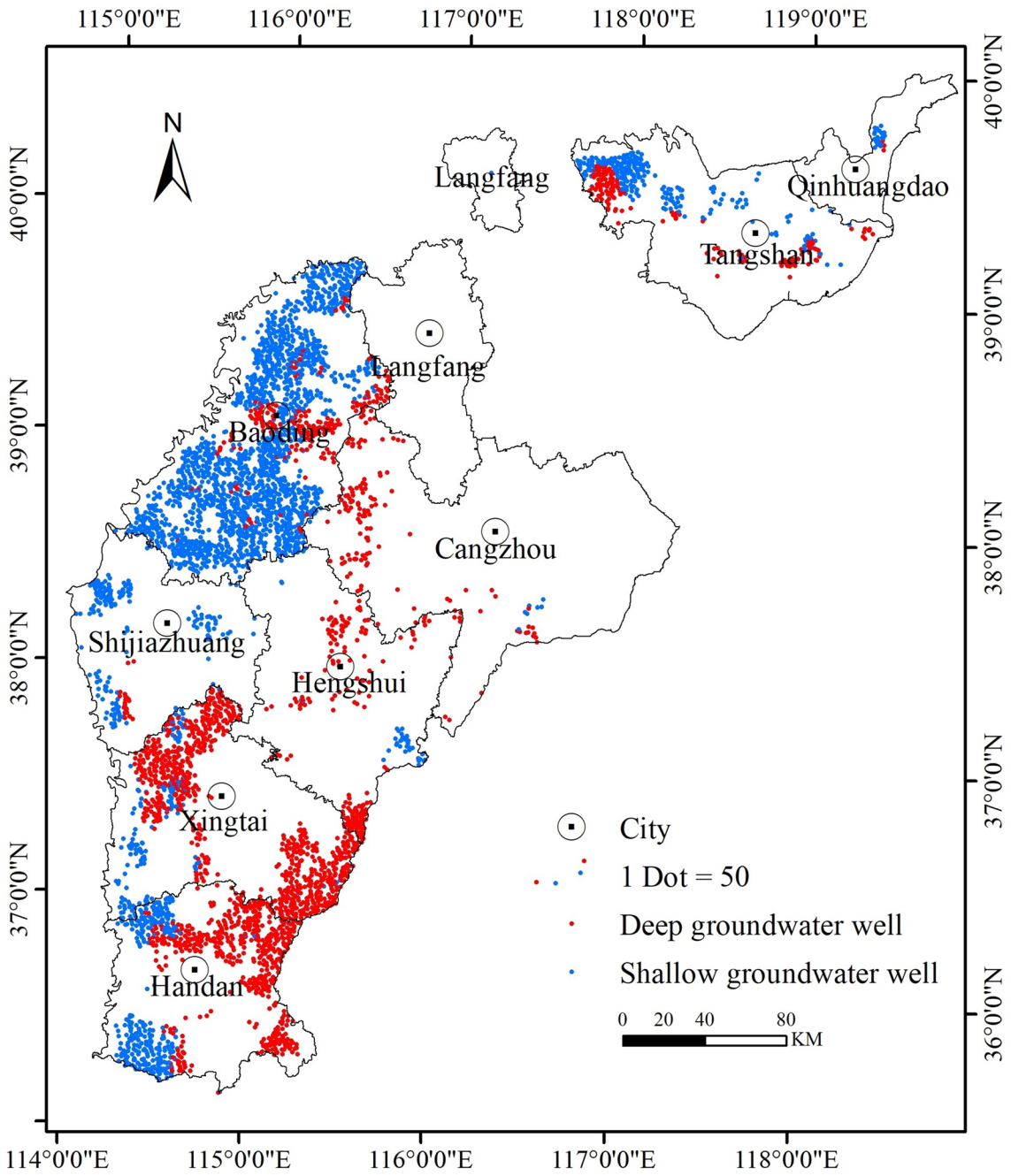

Figure 4. Distribution of deep and shallow groundwater wells with known depths. The figure was created by the author using ArcMap 10.7.

Machine learning classification results

Data from groundwater wells classified in section "Classification of deep and shallow groundwater wells with known depths" are divided into training and test sets at a 7:3 ratio. A threefold cross-validation method is applied within the training phase, with three subsets from the training set used as validation sets for alternately assessing the model's performance.

Table 6 shows that of the four machine learning models, the RF model achieves the best performance with an overall accuracy of 91.23%. It exhibits strong classification capabilities, distinguishing effectively between deep groundwater wells (92.01%) and shallow groundwater wells (90.41%). In contrast, LR, NB, and SVM models exhibit slightly lower overall accuracy rates, Kappa coefficients, recall, and F1-scores. The confusion matrix (Fig. 6) also shows that the RF predictions are more accurate, with errors more "evenly" distributed between the two types of groundwater wells. In contrast, SVM, LOG, and NB tend to classify groundwater wells as shallow wells.

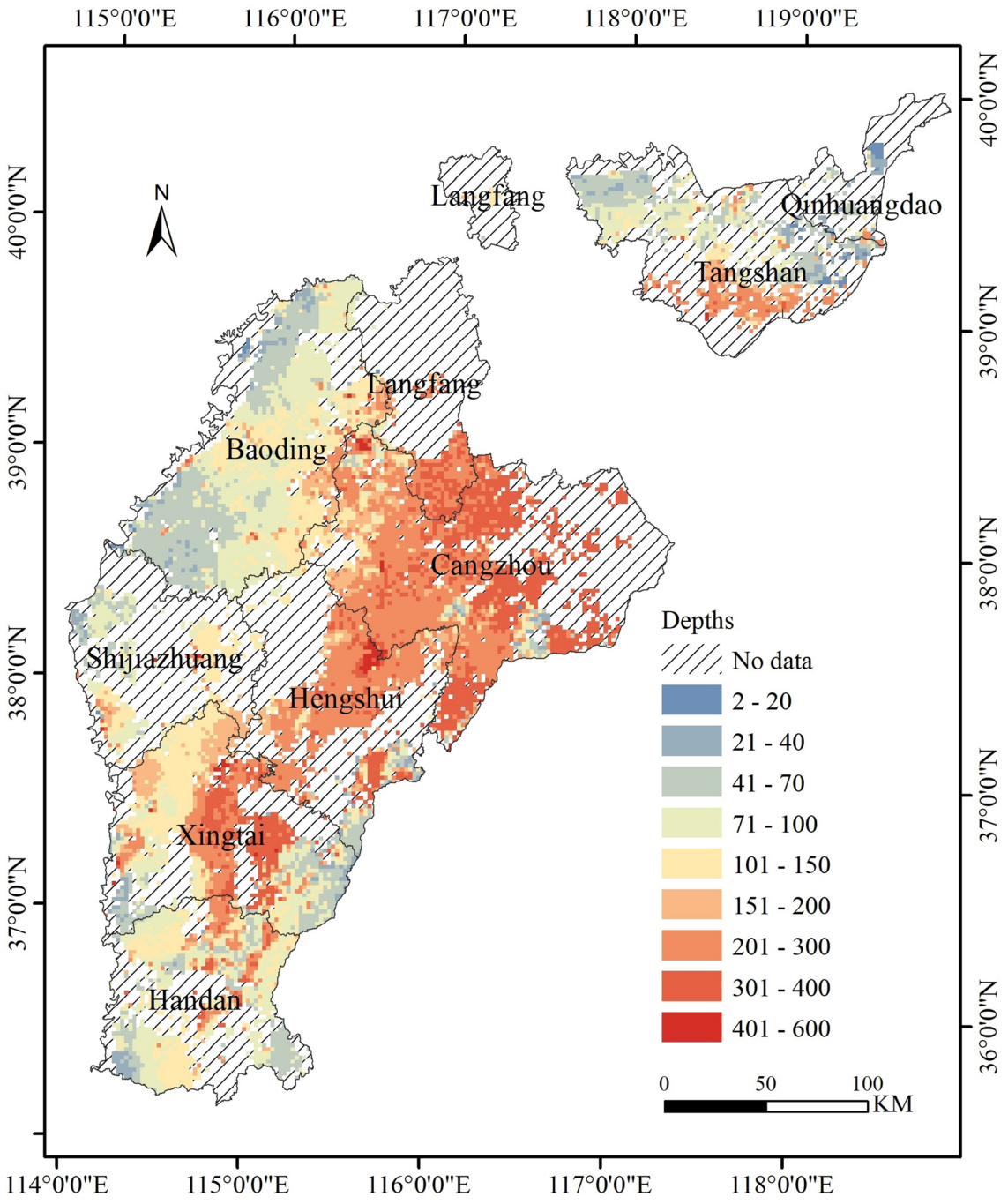

Figure 5. Distribution of well depths in groundwater wells. The figure was created by the author using ArcMap 10.7.

Machine learning model	Overall accuracy (%)	kappa coefficient (%)	Precision (%)	Recall (%)	F1-scores (%)
RF	91.23	82.44	91.43	90.41	90.92
LG	89.88	79.66	95.68	82.89	88.83
SVM	88.35	76.61	92.73	82.48	87.30
NB	89.27	78.46	94.37	82.86	88.24

Table 6. Overall accuracy, kappa coefficient, precision, recall, and F1-score values achieved by different classifiers.

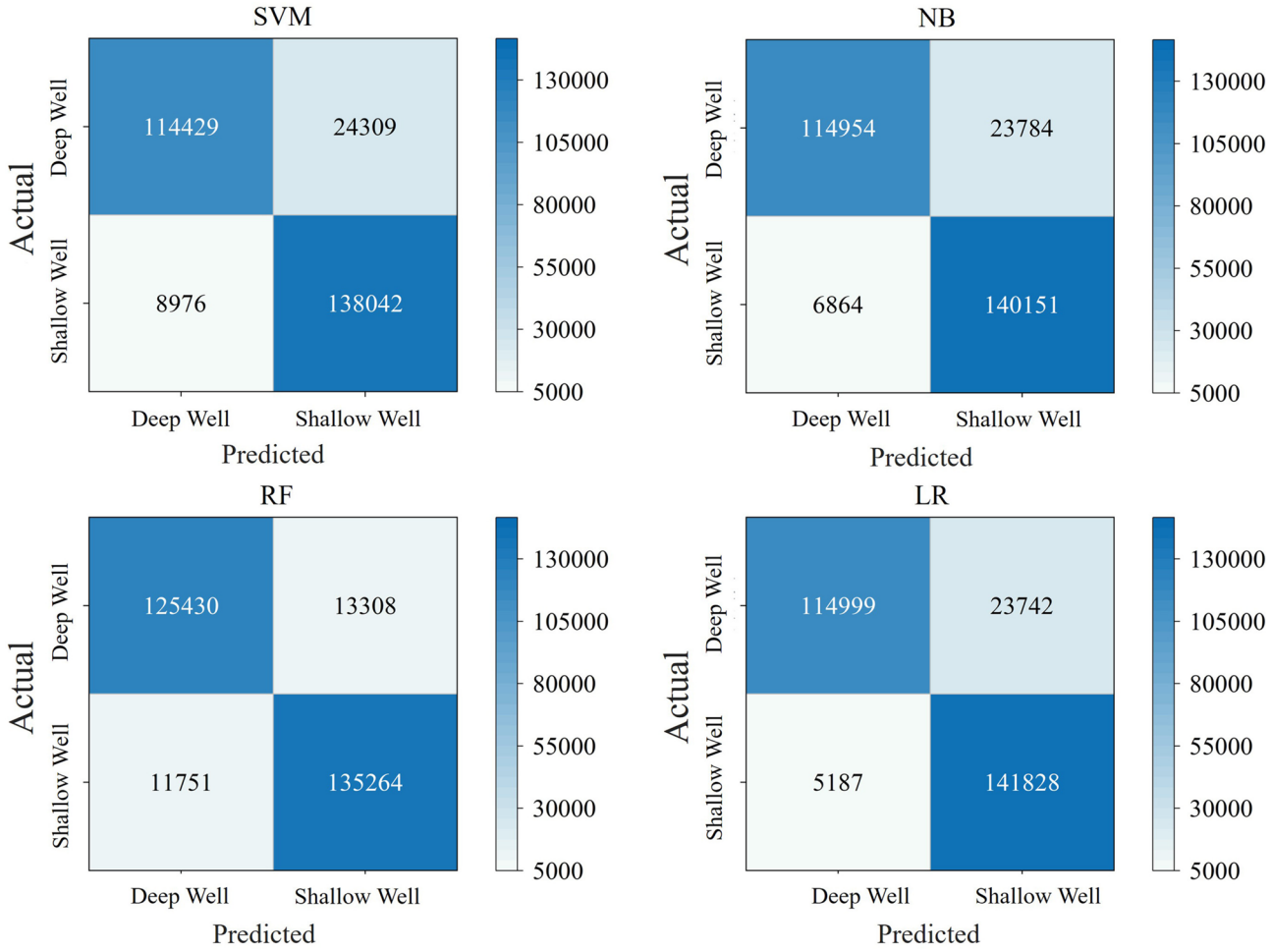

Figure 6. Confusion matrix obtained by the SVM, NB, RF and NB.

While these three models achieve reasonable accuracy, they tend to frequently categorize groundwater wells as shallow. Because the study prioritizes overall prediction accuracy rather than accuracy in specific categories, the Random Forest model, which achieves the highest overall accuracy, is selected as the optimal model. Consequently, the RF model is preferred for deep and shallow groundwater well classification and prediction, and is thus used to classify groundwater wells with unknown depths in Hebei Plain.

Classification of deep and shallow groundwater wells with unknown depths

The random forest model developed in section "Machine learning classification results" is employed for the classification and prediction of groundwater wells with unknown depths. The spatial distribution of these groundwater wells is depicted in Fig. 7, and it closely mirrors the pattern presented in Fig. 4. Shallow groundwater wells are predominantly located in the piedmont alluvial plain, whereas deep groundwater wells are mainly situated across the central alluvial and coastal plains.

As shown in Table 7, deep groundwater wells are predominantly found within the urban centralized water supply wells, rural water supply factory wells, and service industry wells, representing over 80% of the total groundwater wells in each category. In the urban domestic wells and ground source heat pump wells, the usage of deep groundwater wells is also high, exceeding 70%. More than 60% of groundwater wells in both the industrial enterprise wells and rural domestic wells are deep. In contrast, agriculture primarily utilizes shallow groundwater wells, with deep groundwater wells constituting less than 50% of the agricultural total. On the whole, the aggregate number of shallow groundwater wells outnumbers that of deep groundwater wells. Classification results for groundwater wells with unknown depths, obtained using machine learning techniques, show a similarity to the classifications for groundwater wells with known depths presented in section "Classification of deep and shallow groundwater wells with known depths", which corroborates the reliability of the machine learning classification method.

Classification of deep and shallow groundwater wells

The spatial distribution of groundwater level depression funnels correlates strongly with the spatial distribution of groundwater wells. Shallow groundwater funnels are predominantly clustered in the freshwater zone of the piedmont alluvial plain, while deep groundwater funnels are mainly found in the saline water areas of the

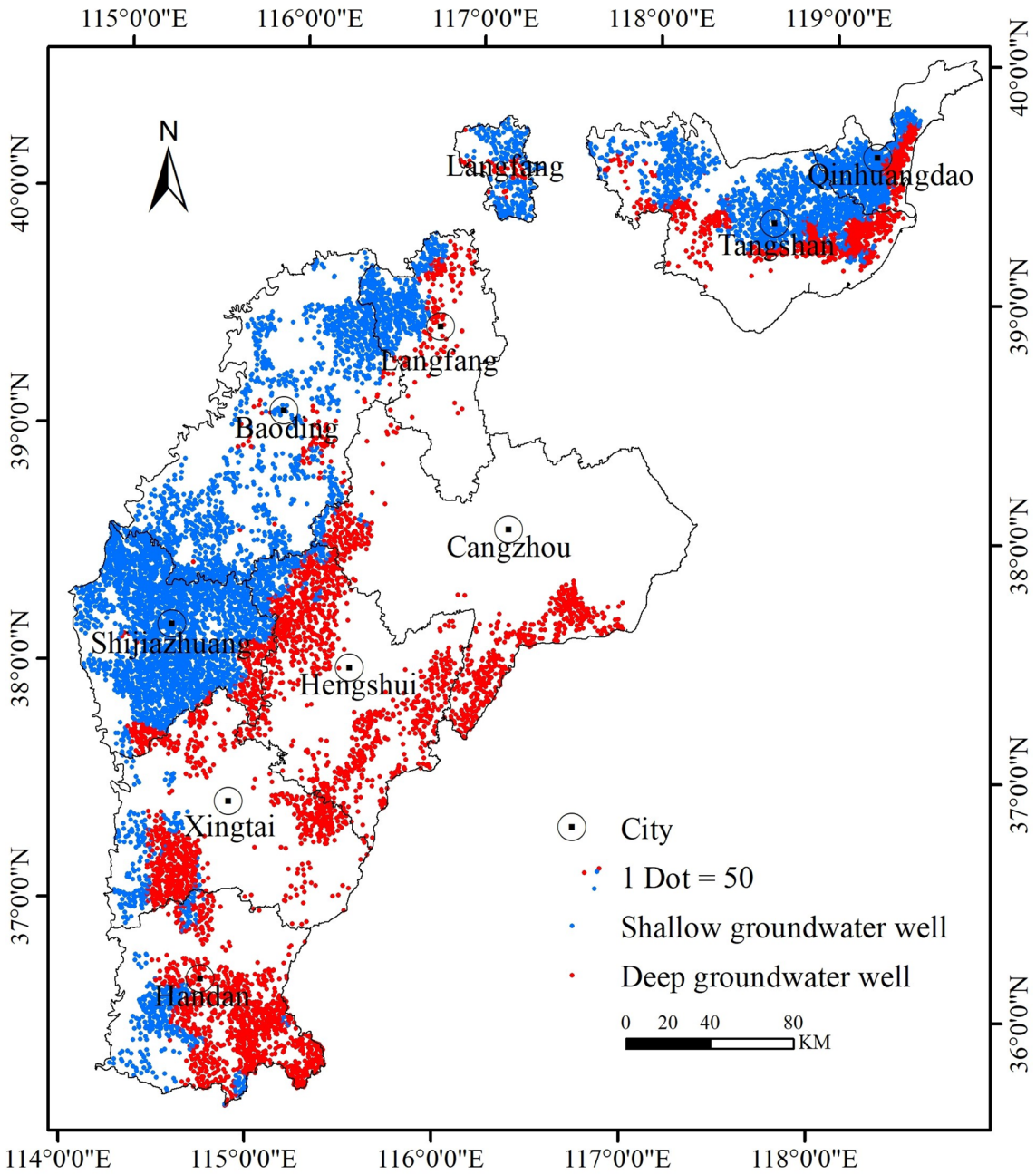

Figure 7. Distribution of deep and shallow groundwater wells with unknown depths. The figure was created by the author using ArcMap 10.7.

Category	Deep wells	Shallow wells	Total wells	Proportion of deep wells in total wells (%)
Urban centralized water supply wells	1078	204	1282	84.09
Urban domestic wells	241	103	344	70.06
Service industry wells	213	52	265	80.38
Industrial enterprise wells	3956	2070	6026	65.65
Rural water supply factory wells	1855	372	2227	83.30
Rural domestic r wells	7859	3894	11,753	66.87
Agricultural wells	242,964	329,751	572,715	42.42
Ground source heat pumps wells	1190	315	1505	79.07
Total wells	259,356	336,761	596,117	43.51

Table 7. Distribution of deep and shallow groundwater wells with unknown depths across different categories.

central alluvial and coastal plain. In a similar vein, shallow groundwater wells are largely concentrated in the freshwater zone of the piedmont alluvial plain, whereas deep groundwater wells are primarily situated in the saline water zones of the central alluvial and coastal plain. The density of groundwater wells is markedly greater in the piedmont alluvial plain than in the central alluvial and coastal plain. This groundwater well distribution exhibits a distinctive pattern: 'west shallow and east deep, west dense and east sparse.'

In the freshwater zone of the piedmont alluvial plain, there are 511,027 groundwater wells, with deep groundwater wells comprising 13.94% and shallow groundwater wells making up 86.06%. By contrast, in the saline water zone of the central alluvial and coastal plain, there are 370,845 groundwater wells. Deep groundwater wells represent 88.13%, while shallow wells account for only 11.87%. This variation in groundwater well distribution between saline and freshwater zones underscores that the salinity of shallow groundwater is a significant factor contributing to the 'west shallow and east deep' groundwater well pattern observed in the Hebei Plain.

The distribution of groundwater wells in the Hebei Plain correlates with groundwater abundance: in the piedmont alluvial plain, characterized by an abundance of shallow groundwater, there is a predominance of shallow groundwater wells; in contrast, in the central alluvial and coastal plain, where shallow groundwater is less abundant, there is a higher proportion of deep groundwater wells. This pattern indicates that the abundance of shallow groundwater significantly influences the choice of well depths. Thus, the abundance of shallow groundwater also plays a pivotal role in the region's characteristic 'west shallow and east deep' groundwater well distribution pattern.

In an effort to curtail the volume of groundwater extraction, Hebei Province initiated an agricultural irrigation water source replacement project from 2014 to 2022. This initiative was implemented extensively across the cities of Cangzhou, Hengshui, and in the eastern regions of Handan and Xingtai cities. As observed in Fig. 8, there is a lower density of groundwater wells in select areas of Cangzhou, Hengshui, the central part of Xingtai, and Handan. The agricultural irrigation water source replacement project in Hebei Province is a significant contributing factor to the 'west dense and east sparse' pattern of groundwater well distribution.

In 2003, the exploitation of saline water with a salinity of 1–2 g/L constituted 21.59% of the total extraction, while water with a salinity of 2–3 g/L made up 8.06%, 3–5 g/L comprised 1.30%, and water with more than 5 g/L accounted for 0.38%. These figures suggest that salinity significantly impacts groundwater extraction.

This paper analyzes the influence of varied salinity levels on the proportion of deep and shallow groundwater wells within three salinity zones: < 1 g/L, 1–2 g/L, and > 2 g/L, for different uses, including urban centralized water supply wells, urban domestic wells, service industry wells, and industrial enterprise wells, among others (Table 8). In the freshwater zone (< 1 g/L), deep groundwater wells are less common. However, they represent about 70% for uses such as urban centralized water supply wells, rural water supply factory wells, and ground source heat pump wells.

When shallow groundwater salinity exceeds 1 g/L, the proportion of deep groundwater wells increases rapidly across all types; overall, deep groundwater wells account for over 91%. Notably, deep agricultural wells increase from 30.94% in the 0–1 g/L zone to 91.35% in the 1–2 g/L zone, a significant jump of 60.41%. For urban centralized water supply wells and service industry wells, the proportion of deep groundwater wells reaches 100%. Agricultural wells and industrial enterprise wells also show high proportions at 91.35% and 92.03%, respectively. When salinity levels exceed 2 g/L, nearly all groundwater wells are deep, with their share exceeding 98%.

The analysis in Table 8 highlights an important finding: a salinity threshold of 1 g/L is crucial for water quality requirements. When the salinity of shallow groundwater is below 1 g/L, it serves as the primary water source due to its low salinity. However, once the salinity of shallow groundwater exceeds 1 g/L, extraction from shallow wells is largely discontinued, and the primary water source shifts to low-salinity deep groundwater.

Agricultural irrigation is the primary driver of excessive groundwater usage in the Hebei Plain, and numerous studies have indicated that saline water, when properly managed, can serve as an alternative to freshwater for irrigation purposes[36–38]. In this study, 38.94% of the deep groundwater wells are located in areas where salinity levels range between 1 and 3 g/L. This suggests that there is considerable potential for the utilization of shallow saline water resources in the Hebei Plain. By fully capitalizing on shallow saline water, the strain on deep groundwater resources can be alleviated while still meeting the demands for agricultural water, thereby aiding in the sustainable management of groundwater supplies and maintaining a balance in the regional water cycle. This approach is crucial for addressing the issue of deep groundwater overexploitation in the Hebei Plain.

Conclusions

In this study, using existing hydrogeological borehole data, groundwater wells with known depths are classified into deep and shallow categories. Four machine learning methods—RF, SVM, LR, and NB—were employed to classify groundwater wells with unknown well depths. The classification was based on five variables: groundwater well attributes, salinity of shallow groundwater, the abundance of both deep and shallow groundwater and the determination of whether the site is situated within an overexploited zone. The model's discriminative ability was evaluated using metrics such as three-fold cross-validation, confusion matrix, kappa coefficient, overall accuracy, precision, recall and F1-score. The results of the study are as follows:

(1) Of the 285,755 groundwater wells with known depths, the proportions of deep and shallow groundwater wells are 48.55% and 51.45%, respectively.
(2) The RF model outperformed the other machine learning methods in classification performance, achieving an overall accuracy of 91.23%, a Kappa coefficient of 82.44%, a recall rate of 91.43%, a precision rate of 90.41%, and an F1-score of 90.92%.
(3) RF was employed to classify wells with unknown depths. Among the 596,117 groundwater wells, the estimated proportions of deep and shallow wells are 43.51% and 56.49%, respectively.

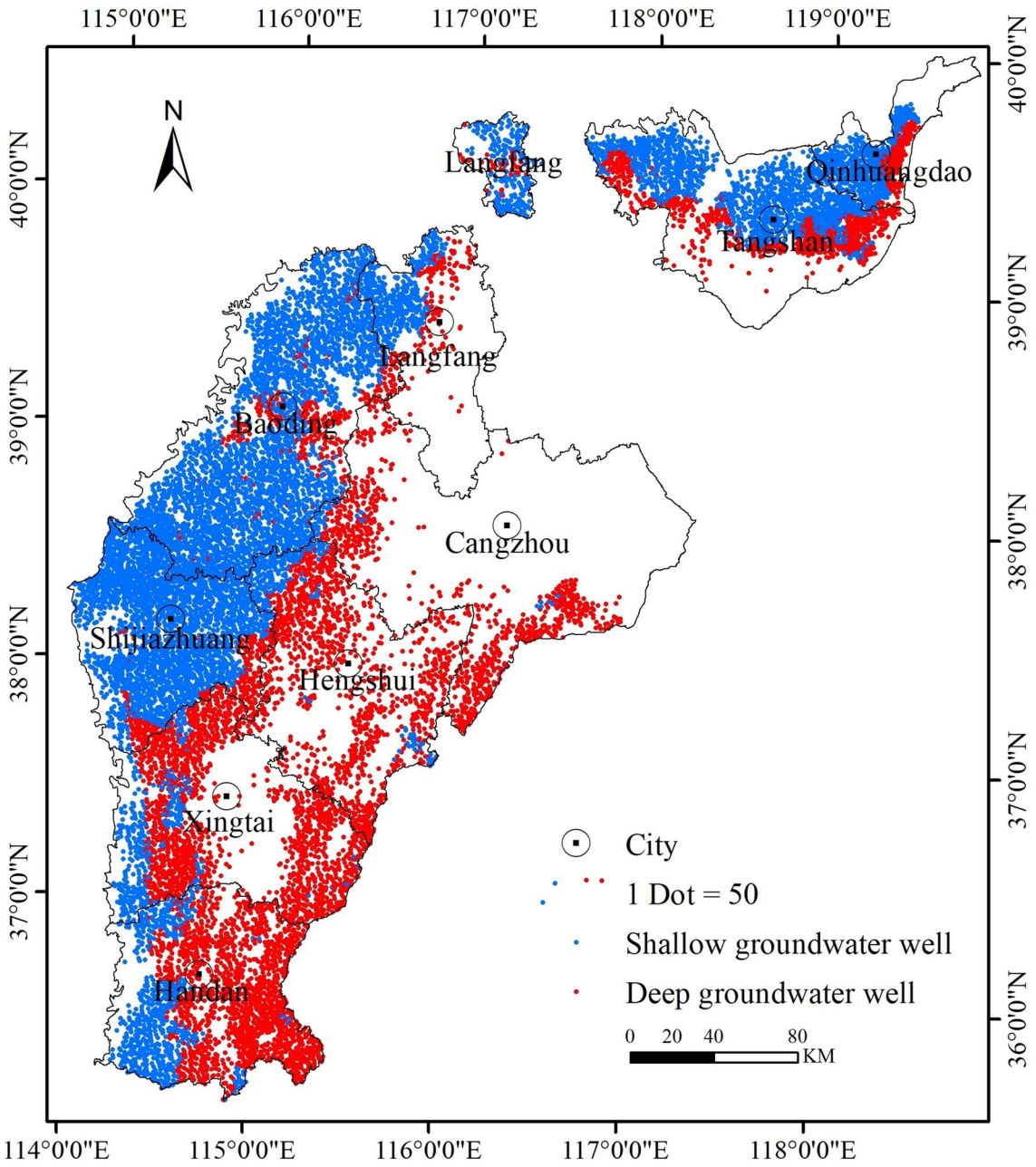

Figure 8. Distribution of deep and shallow groundwater wells. The figure was created by the author using ArcMap 10.7.

Category	Proportion of deep wells in total wells (%)		
Salinity	**0–1 g/L**	**1–2 g/L**	**>2 g/L**
Urban centralized water supply wells	77.55	100.00	100.00
Urban domestic wells	64.35	90.57	100.00
Service industry wells	76.38	100.00	100.00
Industrial enterprise wells	63.82	92.03	99.86
Rural water supply factory wells	74.09	98.04	99.74
Rural domestic wells	59.69	94.00	99.97
Agricultural wells	30.94	91.35	97.71
Ground source heat pumps wells	77.94	99.66	98.86
Total wells	32.11	91.43	97.77

Table 8. Proportion of deep wells within different salinity areas by categories of groundwater wells.

(4) In the freshwater zone of the piedmont alluvial plain, 511,027 groundwater wells were identified. Of these, 13.94% are classified as deep and 86.06% as shallow. In the saline water zone of the central alluvial and coastal plain, among 370,845 groundwater wells, 88.13% and 11.87% are categorized as deep and shallow, respectively.

Using machine learning models, this study successfully classified wells of unknown depths, providing an effective method to distinguish between deep and shallow wells. Other regions can establish groundwater well classification models based on similar hydrogeological parameters and machine learning techniques. In this paper, the Kriging interpolation method was used for data processing, introducing some uncertainty factors.

Data availability

The data that support the findings of this study are available from Department of Water Resources of Hebei Province but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Department of Water Resources of Hebei Province.

Received: 17 March 2024; Accepted: 1 August 2024
Published online: 06 August 2024

References

1. Chen, M., Xu, Y., Pan, H. & Wang, L. Water storage changes in the North China Plain from 2004 to 2019. *Sci. China Technol. Sci.* **66**, 2036–2046. https://doi.org/10.1007/s11431-022-2274-2 (2023).
2. Gong, H. *et al.* Long-term groundwater storage changes and land subsidence development in the North China Plain (1971–2015). *Hydrogeol. J.* **26**, 1417–1427. https://doi.org/10.1007/s10040-018-1768-4 (2018).
3. Han, J. *et al.* Mechanism the land subsidence from multiple spatial scales and hydrogeological conditions—A case study in Beijing-Tianjin-Hebei China. *J. Hydrol. Region. Stud.* **50**, 101531. https://doi.org/10.1016/j.ejrh.2023.101531 (2023).
4. Zanotti, C. *et al.* A cost-effective method for assessing groundwater well vulnerability to anthropogenic and natural pollution in the framework of water safety plans. *J. Hydrol.* **613**, 128473. https://doi.org/10.1016/j.jhydrol.2022.128473 (2022).
5. Andrade, L., Chique, C., Hynds, P., Weatherill, J. & Dwyer, J. The antimicrobial resistance profiles of *Escherichia coli* and *Pseudomonas aeruginosa* isolated from private groundwater wells in the Republic of Ireland. *Environ. Pollut.* **317**, 12. https://doi.org/10.1016/j.envpol.2022.120817 (2023).
6. Lutterodt, G., Gibrilla, A., Andorful, F., Ganyaglo, S. & Oduro-Kwarteng, S. Influence of on-site sanitation on groundwater quality from large diameter wells. *Groundw. Sustain. Dev.* **20**, 12. https://doi.org/10.1016/j.gsd.2022.100862 (2023).
7. Shin, H.-J. *et al.* Vulnerability evaluation of groundwater well efficiency and capacity in drought vulnerable areas. *J. Korean Soc. Agric. Eng.* **61**, 41–53. https://doi.org/10.5389/ksae.2019.61.6.041 (2019).
8. Shin, H. J., Lee, J. Y., Jo, S. M., Sun, C. S. & Chan-Gi, P. Vulnerability assessment of upland public groundwater wells against climate change. *Korean J. Agric. Sci.* **47**, 577–596. https://doi.org/10.7744/kjoas.20200046 (2020).
9. Jasechko, S. & Perrone, D. Global groundwater wells at risk of running dry. *Science* **372**, 418–421. https://doi.org/10.1126/science.abc2755 (2021).
10. Perrone, D. & Jasechko, S. Dry groundwater wells in the western United States. *Environ. Res. Lett.* **12**, 10. https://doi.org/10.1088/1748-9326/aa8ac0 (2017).
11. Li, F. *et al.* Factors influencing electricity-to-water conversion metering method for irrigation water consumption in Hebei Plain. *Chin. J. Eco-Agric.* **30**, 1993–2001 (2022).
12. Sit, M., Langel, R. J., Thompson, D., Cwiertny, D. M. & Demir, I. Web-based data analytics framework for well forecasting and groundwater quality. *Sci. Total Environ.* **761**, 10. https://doi.org/10.1016/j.scitotenv.2020.144121 (2021).
13. Maxwell, A. E., Warner, T. A. & Fang, F. Implementation of machine-learning classification in remote sensing: An applied review. *Int. J. Remote Sens.* **39**, 2784–2817. https://doi.org/10.1080/01431161.2018.1433343 (2018).
14. Haggerty, R., Sun, J. X., Yu, H. F. & Li, Y. S. Application of machine learning in groundwater quality modeling-A comprehensive review. *Water Res.* **233**, 20. https://doi.org/10.1016/j.watres.2023.119745 (2023).
15. Tahmasebi, P., Kamrava, S., Bai, T. & Sahimi, M. Machine learning in geo- and environmental sciences: From small to large scale. *Adv. Water Resour.* **142**, 33. https://doi.org/10.1016/j.advwatres.2020.103619 (2020).
16. Wang, Y. K. *et al.* A comparative study of different machine learning methods for reservoir landslide displacement prediction. *Eng. Geol.* **298**, 12. https://doi.org/10.1016/j.enggeo.2022.106544 (2022).
17. Tyralis, H., Papacharalampous, G. & Langousis, A. A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water* **11**, 37. https://doi.org/10.3390/w11050910 (2019).
18. Raghavendra, N. S. & Deka, P. C. Support vector machine applications in the field of hydrology: A review. *Appl. Soft Comput.* **19**, 372–386. https://doi.org/10.1016/j.asoc.2014.02.002 (2014).
19. LaValley, M. P. Logistic regression. *Circulation* **117**, 2395–2399. https://doi.org/10.1161/circulationaha.106.682658 (2008).
20. Cracknell, M. J. & Reading, A. M. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Comput. Geosci.* **63**, 22–33. https://doi.org/10.1016/j.cageo.2013.10.008 (2014).
21. Tan, Y. H. *et al.* Impact of urbanization on baseflow characteristics in the central catchment of North China Plain. *China. J. Hydrol.-Reg. Stud.* **50**, 18. https://doi.org/10.1016/j.ejrh.2023.101527 (2023).
22. Guo, H. *et al.* Groundwater-derived land subsidence in the North China Plain. *Environ. Earth Sci.* **74**, 1415–1427. https://doi.org/10.1007/s12665-015-4131-2 (2015).
23. Lancia, M., Jing, H., Steed, S. M. & Zheng, C. M. Analysis of hydraulic conductivity characteristics of alluvial sequence in North China Plain. *Environ. Earth Sci.* **80**, 10. https://doi.org/10.1007/s12665-021-09803-3 (2021).
24. Lu, C. P. *et al.* Spatiotemporal variation and long-range correlation of groundwater depth in the Northeast China Plain and North China Plain from 2000~2019. *J. Hydrol.-Reg. Stud.* **37**, 19. https://doi.org/10.1016/j.ejrh.2021.100888 (2021).
25. Su, C., Cheng, Z. S., Wei, W. & Chen, Z. Y. Assessing groundwater availability and the response of the groundwater system to intensive exploitation in the North China Plain by analysis of long-term isotopic tracer data. *Hydrogeol. J.* **26**, 1401–1415. https://doi.org/10.1007/s10040-018-1761-y (2018).
26. Oliver, M. A. & Webster, R. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena* **113**, 56–69. https://doi.org/10.1016/j.catena.2013.09.006 (2014).
27. Boser, B. E., Guyon, I. M. & Vapnik, V. N. in *Proceedings of the fifth annual workshop on Computational learning theory* 144–152 (Association for Computing Machinery, Pittsburgh, Pennsylvania, USA, 1992).
28. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297. https://doi.org/10.1007/BF00994018 (1995).

29. Hsu, C. W. & Lin, C. J. A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Netw.* **13**, 415–425. https://doi.org/10.1109/72.991427 (2002).
30. Karatzoglou, A., Meyer, D. & Hornik, K. Support Vector Machines in R. *J. Stat. Softw.* **15**, 1–28. https://doi.org/10.18637/jss.v015.i09 (2006).
31. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. https://doi.org/10.1023/A:1010933404324 (2001).
32. Sagi, O. & Rokach, L. Ensemble learning: A survey. *WIREs Data Min. Knowl. Discov.* **8**(4), 18. https://doi.org/10.1002/widm.1249 (2018).
33. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140. https://doi.org/10.1023/A:1018054314350 (1996).
34. Krzywinski, M. & Altman, N. Classification and regression trees. *Nat. Methods* **14**, 757–758. https://doi.org/10.1038/nmeth.4370 (2017).
35. Kavzoglu, T., Sahin, E. K. & Colkesen, I. Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides* **11**, 425–439. https://doi.org/10.1007/s10346-013-0391-7 (2014).
36. Guo, K. & Liu, X. J. Reclamation effect of freezing saline water irrigation on heavy saline-alkali soil in the Hetao Irrigation District of North China. *Catena* **204**, 13. https://doi.org/10.1016/j.catena.2021.105420 (2021).
37. Wang, Q. M., Huo, Z. L., Zhang, L. D., Wang, J. H. & Zhao, Y. Impact of saline water irrigation on water use efficiency and soil salt accumulation for spring maize in arid regions of China. *Agric. Water Manag.* **163**, 125–138. https://doi.org/10.1016/j.agwat.2015.09.012 (2016).
38. Wang, H. *et al.* Impacts of long-term saline water irrigation on soil properties and crop yields under maize-wheat crop rotation. *Agric. Water Manag.* **286**, 13. https://doi.org/10.1016/j.agwat.2023.108383 (2023).

Acknowledgements

We acknowledge the reviewers and editors for their valuable advice on improving the quality of this paper. Financial support for this work was provided by the National Key Research and Development Program of China (2023YFC3206501 and 2021YFC3000205), Independent Research Project of the State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin (SKL2022ZD02), and Key R&D Program of Heilongjiang Province (JD22B001).

Author contributions

H. Z.: Conceptualization, Methodology, Supervision, Validation, Writing—review & editing. C.W.: Funding acquisition, Investigation, Project administration, Writing—review & editing. B.L.: Investigation, Methodology, Validation, Writing—review & editing. C.L.: Data curation, Investigation. Y.Z.: Data curation, Investigation. Z.Z.: Data curation, Investigation.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.L. or Z.Z.