scientific reports

Check for updates

OPEN

# Reservoir temperature prediction based on characterization of water chemistry data—case study of western Anatolia, Turkey

Haoxin Shi, Yanjun Zhang✉, Ziwang Yu & Yunxing Yang

**Reservoir temperature estimation is crucial for geothermal studies, but traditional methods are complex and uncertain. To address this, we collected 83 sets of water chemistry and reservoir temperature data and applied four machine learning algorithms. These models considered various input factors and underwent data preprocessing steps like null value imputation, normalization, and Pearson coefficient calculation. Cross-validation addressed data volume issues, and performance metrics were used for model evaluation. The results revealed that our machine learning models outperformed traditional fluid geothermometers. All machine learning models surpassed traditional methods. The XGBoost model, based on the F-3 combination, demonstrated the best prediction accuracy with an $R^2$ of 0.9732, while the Bayesian ridge regression model using the F-4 combination had the lowest performance with an $R^2$ of 0.8302. This study highlights the potential of machine learning for accurate reservoir temperature prediction, offering geothermal professionals a reliable tool for model selection and advancing our understanding of geothermal resources.**

**Keywords** Reservoir temperature, Hydrogeochemistry, Geothermometer, Machine learning

Geothermal energy is a renewable energy source that is rich in reserves, widely distributed, stable and reliable. Vigorous development and utilization of geothermal energy is of great significance to the implementation of carbon peak and carbon neutral goals. The determination of reservoir temperature is an important parameter indispensable to the study of geothermal resources exploration, geothermal potential assessment and geothermal development and utilization[1–5]. In order to make full use of geothermal resources, accurate prediction of reservoir temperature has become an active research topic[6].

Generally speaking, the prediction methods of reservoir temperature can be divided into two categories: direct measurement method and indirect calculation method. The direct measurement method utilizes the site drilling to directly measure the temperature, and determines the thermal reservoir temperature by calculating the average of the thermal reservoir top plate and bottom plate temperatures. However, in most cases, there is no drilling hole in the site or the depth of the drilling hole does not reach the geothermal reservoir, and at the same time, after the drilling of the final hole, the accuracy of the measured temperature varies greatly according to the time when the drilling hole has been stationary, so that direct measurement of the reservoir temperature by on-site drilling is a very expensive and time-consuming work[7,8]. Indirect calculation method, i.e. fluid geothermometer method, which utilizes the relationship between the content of chemical components in underground hot water and gases and the temperature of the reservoir for thermal storage estimation, the basic principle of which is that, after chemical equilibrium is reached between minerals and fluids or different fluids in deep thermal reservoirs, the temperature decreases during the rise of hot water to the surface, but the content of the chemical components remains unchanged, and the temperature in the geothermal reservoirs can therefore be estimated based on the equilibrium temperatures of the chemical reactions[9,10]. Due to their low cost, methods based on geothermometers for predicting reservoir temperatures in geothermal systems have been widely popularized and rapidly developed, among which, there are cation-based methods: Na–K[11–18], Na–K-Ca[19], K-Mg[20]; K-Ca[13]; Silica-based methods for geothermometers[21–23]; gas chemistry-based methods for geothermometers, etc.[24–26]. Although some progress has been made in reservoir temperature prediction based on geothermometers, the computational results of the geothermometer method still have large differences when compared with the direct temperature measurement method. In response to the uncertainty of the results, when evaluating the geothermometer results, it is usually

College of Construction Engineering, Jilin University, Changchun 130026, China. ✉email: zhangyanj@jlu.edu.cn

necessary to rely on a variety of calculation methods and combine them with the actual characteristics of the site to make a comprehensive judgment, which greatly increases the complexity and workload of the work.

With the popularization of computers and the development of machine learning algorithms, the consideration of bringing hydrogeochemical data into machine learning algorithms for reservoir temperature prediction has become a new approach to explore the prediction of reservoir temperature[27]. Díaz-González et al. (2008) improved three Na–K geothermometer equations using artificial neural networks and linear regression to improve geothermal temperature prediction in geothermal systems[18]; Porkhial et al. (2015) Modeling and prediction of geothermal reservoir temperatures were attempted through a neural network model[28]; Perez-Zarate et al. (2019) employed a three-layer artificial neural network, taking $CO_2$, $H_2S$, $CH_4$, and $H_2$ as inputs and bottomhole temperature as output, to perform multivariate analysis on fluid gas composition and predict geothermal reservoir temperature[27]; Tut Haklidir and Haklidir (2020) Reservoir temperatures in western Anatolia, Turkey, were predicted using hydrogeochemical data through linear regression, linear support vector machine and deep neural network methods[29]; Varol Altay et al. (2022) further utilized hydrogeochemical data from different geothermal areas in western Anatolia, Turkey, to propose a hybrid artificial neural network model based on heuristic optimization algorithms for predicting reservoir temperatures[30]. In the same year, Afandi et al. (2022) used Artificial Neural Network (ANN) model to predict probe temperature[31]. Davoodi and Vo Thanh (2023) proposed the LSSVM machine learning model for predicting the residual captive index of carbon dioxide solubility at global geologic sequestration sites, hydrogen uptake values of porous carbon materials, and combined machine learning with optimization to propose the LSSVM-COA model to improve the prediction accuracy while reducing the Associated uncertainties[32–34]; Davoodi and Mehrad (2023) proposed hybrid machine learning for rapid prediction of rheological and filtration properties of water-based drilling fluids, achieving accurate and reliable prediction of filtration properties of drilling fluids and applying hybrid machine learning to assist in prediction of uniaxial compressive strength using drilling variables[35,36].

Overall, there is a general lack of sufficient training data when scholars adopt machine learning to predict reservoir temperature, while previous studies tend to directly take the collected hydrochemical data as inputs without considering the relationship between different combinations of inputs (hydrochemical data) and outputs (reservoir temperature), which results in the proposed methods having their own scope of applicability without strong generalizability.

The main objective of this paper is to investigate the performance of machine learning models that take into account data characterization and to determine the applicability of using machine learning for reservoir temperature prediction. By searching the literature, 83 sets of hydrogeochemical data and reservoir temperature data were collected, and the characteristics of the dataset were carefully analyzed using normalization, box plots, and mutual correlation analysis; Build a total of four machine learning regression models, Bayesian Ridge Regression, Decision Tree Regressor, eXtremeGradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM); Solve the model accuracy problem caused by a small amount of data through fivefold cross-validation; construct a prediction model considering multiple data combinations through multiple data combination forms. Performance evaluation metrics are used to evaluate the model to demonstrate the method's performance in predicting reservoir temperatures and to identify the optimal algorithms, while various model predictions are compared with traditional geothermometer methods to determine the applicability and accuracy of the prediction models.

## Data preparation and preprocessing

### Data preparation

The 83 data sets in the paper are from the research dataset of Tut Haklidir and Haklidir (2020), which consists of hydrogeochemical data and reservoir temperatures obtained by different researchers in various regions of Western Anatolia, Turkey (Fig. 1). Each group contains the following data parameters: Pondus Hydrogenii (PH), electric conductivity (EC), $K^+$, $Na^+$, boron, silicon dioxide ($SiO_2$), $Cl^-$, temperature (T). PH represents

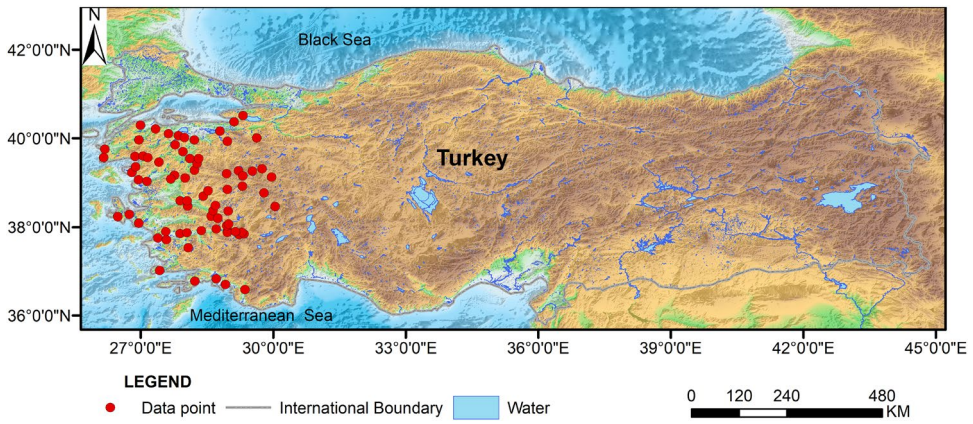


**Figure 1.** Relief map of western Anatolia, Turkey, showing the location of the study area. Study data from (Tut Haklidir and Haklidir[29]). Base map from Grid Extract (noaa.gov https://www.ncei.noaa.gov/maps/grid-extract/).

the alkalinity of the water, and can also indicate groundwater mixing, geochemical process information; EC is an important parameter for dissolved solids in geothermal fluids, and has been used to predict reservoir temperatures; $K^+$ and $Na^+$ are the main cations, and can indicate the interactions between the hot water and rocks; Boron is a trace element, which represents the circulation of groundwater, and a high concentration of boron indicates a high-temperature reservoir in the deep subsurface; $SiO_2$ is an important indicator for predicting the geothermal temperature, and the content of silica in geothermal fluids depends on the solubility of quartz in water at different temperatures and pressures. Silica solubility increases with increasing temperature, and dissolved silica in natural water is generally unaffected by other ions, the formation of complexes and volatilization and dissipation, and the rate of precipitation slows down with decreasing temperature, so that the concentration of silica in surface water is a good indicator of the temperature of subsurface thermal reservoirs; $Cl^-$ is the main anion, representing the salinity of subsurface hot water; the temperature is the geothermometer-calculated hot spring reservoir temperatures and measured reservoir temperatures from geothermal wells.

Information on data characteristics such as maximum, median, and minimum values are listed in Table 1, where the minimum and maximum values indicate that the dataset attributes vary over a very biao wide range, e.g., conductivity EC in microS/cm ranges from 300 to 10,330, while $K^+$ is a concentration (mg/l) ranging from 0.8 to 191. Detailed data are shown in Table 1.

The accuracy of the data will affect the completeness and accuracy of the research results. The data concentration in deep reservoirs can be affected by the vapor fraction, resulting in higher or lower data concentration[37,38]. Since this study focuses on how to choose the best input parameters and use reasonable machine learning algorithms to obtain more accurate temperatures, the effect of vapor fraction on the data is not considered, and whether the data are from a single well or multiple wells is not considered. In this paper, we will use the normalization method to reduce the data dimensionality and improve the interpretability of the data, as well as to improve the accuracy of the study results through data significance analysis and cross-validation.

## Data preprocessing

The datasets have different scale features (Table 1), which usually have different dimensions, and in most models in machine learning, the different dimensions of the different features cause a large range of values to be calculated. Therefore, the raw data is usually normalized in order to improve the interpretability of the data, reduce noise and redundancy, and ensure better results from the model[39]. In this study, normalization will be used to scale the raw data to between [−1, 1] and then brought into the model for the study. The normalization equation is shown in Eq. (1):

$$x^* = \frac{x - x_{mean}}{x_{max} - x_{min}} \tag{1}$$

where: $x^*$ is the normalized data, $x_{mean}$ is the mean of the original data, $x_{max}$ and $x_{min}$ denote the maximum and minimum values of the original data.

Box plots in Fig. 2 display the distribution of normalized feature data, where the small boxes represent data means, the horizontal lines across indicate medians, and black diamond-shaped boxes denote data outliers. The figure reveals that there are three sets of data containing outliers, namely electrical conductivity (EC), silicon dioxide ($SiO_2$), and $Cl^-$. The medians of almost all feature data groups do not align with the centerline of the boxes, implying that the distributions of each data group are asymmetric.

## Data characterization

In machine learning models, the advantages and disadvantages of model training are not only related to the data dimensions, but also the selection of data features is the key to decide whether the model is good or bad. The selection of features not only needs to comprehensively reveal the problem, but also should not increase the computational burden by selecting a large number of features. Reasonable feature selection can simplify the model, speed up the model training speed, make the model have better interpretation, and at the same time can reduce overfitting to improve the generalization ability of the model[40].

Since the feature data may have different degrees of influence on reservoir temperature, in order to determine the most reasonable combination of features, feature selection was performed before model training using the SelectKBest class, along with the f_regression function which computes numerical correlations, was utilized for feature selection. The f_regression function employs F-tests to calculate coefficients between each feature and the target variable.

The intercorrelation equation is shown in Eq. (2):

| | PH | EC (μS/cm) | $K^+$ (mg/l) | $Na^+$ (mg/l) | Boron (mg/l) | $SiO_2$ (mg/l) | $Cl^-$ (mg/l) | T (°C) |
|---|---|---|---|---|---|---|---|---|
| Max | 9.1 | 10,300 | 191 | 1810 | 38 | 650 | 945 | 245 |
| Median | 7.5 | 2590 | 50 | 540 | 8.6 | 96 | 85 | 131 |
| Min | 5.8 | 300 | 0.8 | 2.6 | 0 | 11 | 3 | 50 |
| Mean | 7.55 | 2976.58 | 65.29 | 632.24 | 10.06 | 165.89 | 94.36 | 144.80 |
| Standard deviation | 0.78 | 1913.41 | 57.10 | 513.08 | 9.91 | 152.87 | 352.38 | 56.21 |

**Table 1.** Summary statistics for the dataset.

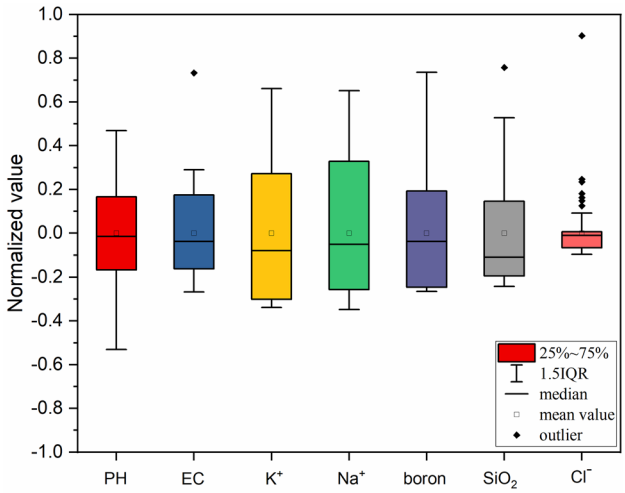


**Figure 2.** Data feature box diagram.

$$F(X,Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{2}$$

where: $p(x,y)$ is the joint probability distribution function of $X$ and $Y$, and $p(x)$, $p(x)$ are the marginal probability distribution functions of $X$ and $Y$ respectively. When $X$ and $Y$ are independent random variables, $F(X,Y) = 0$; when X and Y are the same variables, $F(X,Y) = 1$. Therefore, the $F$ value takes the range between [0, 1], the larger the $F$ value, the stronger the correlation between the features and variables. According to the calculation results $SiO_2$ correlation is the strongest, its correlation is 0.86944, followed by $Na^+$, $K^+$ and EC, in order to make the results more intuitive, the vertical coordinate is used logarithmic coordinates, and the importance of the features is plotted (Fig. 3).

## Performance measure

In order to select the best machine learning model for reservoir temperature prediction, it is necessary to measure the predictive ability of the model with the help of evaluation metrics. In the process of model evaluation, it is often necessary to use different indicators for different problems, and among the many evaluation indicators, most of them can only reflect part of the model's performance in a one-sided way, so if they are not used reasonably, not only can they not find out the problems of the model itself, but also they will come to a wrong conclusion. In this paper, we mainly study the regression problem for reservoir temperature prediction, so we choose three indicators, root mean square error (RMSE), mean absolute error (MAE) and decidable coefficient R-Square, for evaluation.

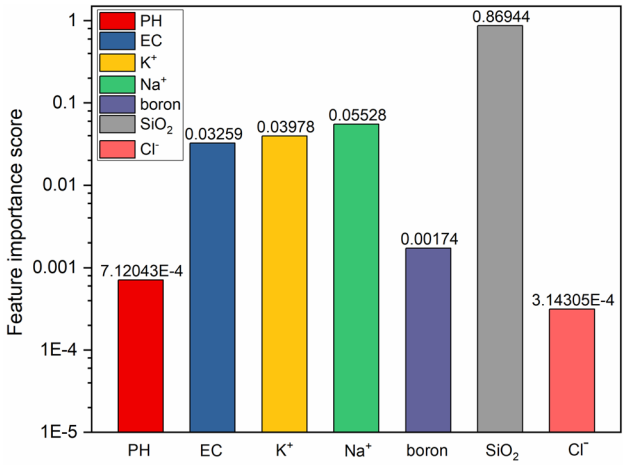


**Figure 3.** Characteristic importance map.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\widehat{y}_i - y_i)^2} \tag{3}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|\widehat{y}_i - y_i\right| \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}(\widehat{y}_i - y_i)^2}{\sum_{i=1}(\widehat{\overline{y}_i} - y_i)^2{}_i} \tag{5}$$

where: root mean square error (RMSE) is the square root of the mean square error. Mean Absolute Error (MAE) is the average of the absolute errors, that is, the average of the errors between the measured value and the true value, which can better reflect the actual situation of the prediction value errors. R-Squared is the value of R-squared.

## Reservoir temperature prediction model

In this paper, based on the Scikit-learn open source algorithm package for machine learning in Python[41], a total of four machine learning models, namely Bayesian Ridge Regression, Decision Tree Regressor, XGBoost and LightGBM, were used. 80% of the data were used to train the models and 20% of the data were used to validate the models for model building, hyper-parameter optimization and result prediction. The mathematical equations and theoretical methods of the above machine learning can be deeply understood through references[14,42,43], and only short definitions and applications are given in the paper.

### Bayesian ridge regression

Bayesian Ridge Regression evolved from Bayesian linear regression[44], which combines the ideas of Ridge Regression and Bayesian statistics. That is, an L2 regularization term (penalty term) is added to the loss function of Bayesian Ridge Regression to control the complexity of the model, and the core idea is to introduce a prior distribution (usually Gaussian) into the loss function to describe the uncertainty of the parameters, and then estimate the model parameters by Maximum A Posteriori Estimation (MAP), which is equivalent to minimizing the loss function while considering the prior distribution of the parameters.

$$p(w|\lambda) = N\left(w|\alpha, \lambda^{-1}I_p\right) \tag{6}$$

where $I_p$ is the order unit square, $N$ is the Gaussian distribution, $\alpha$ is the hyperparameter mean, and $\lambda$ is the standard deviation.

The regularization term in the algorithm can enhance the robustness of the model, and it is not easy to overfit when the sample size of this study is small. At the same time, the algorithm provides a probabilistic framework, allowing the uncertainty of the prediction to be quantified, taking into account the needs of data fitting and the stability of parameters, and has the advantages of robustness and high precision. However, assuming that there is a linear relationship between the predictor variable and the response variable, it may not be able to capture the complex nonlinear pattern in the data, and the model has the problems of large amount of data calculation and long calculation time.

### Decision tree regression

Decision tree regression is a regression analysis method based on decision trees[45], based on the powerful algorithms of decision trees, which are able to fit complex datasets better even when faced with some complex problems. A decision tree is a tree structure in which each internal node represents a test of a feature attribute, each branch represents a test output, and each leaf node represents a predicted value. Starting from the root node, based on the test results of each internal node, the samples are assigned to different child nodes until the leaf node is reached, the average target value of the training samples in the leaf node is the predicted value, the regression tree predicts at each node for a specific value, and when splitting the training set the goal is to find a split that minimizes the MSE.

Each regression tree corresponds to a division of the input space and an output value on the redivision cell. Assuming that the input space is partitioned into M cells $R1, R2, \ldots, Rm$, and there is a fixed output value $Cm$ on each cell $Rm$, the regression tree model can be expressed as:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \tag{7}$$

In this study, the data contains nonlinear features, and the decision tree can flexibly adapt to this complex relationship. Simultaneous decision trees are able to efficiently process multiple features, including multiple variables involved in the prediction, such as PH, conductivity, ion concentration, etc. But decision trees can perform poorly when faced with highly complex relationships.

### XGBoost

eXtremeGradient Boosting (XGBoost) is an integrated learning algorithm based on Gradient Boosting Tree[46,47], and the basic idea is to build a more powerful predictive model by combining multiple weak learners (usually decision trees). The model is iterated and optimized with each round of gradient boosting to provide superior predictive performance. The optimization objective function of the XGBoost model is:

$$\sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)] + \Omega(f_t) \quad (8)$$

The core of XGBoost is the gradient boosting tree algorithm, which continuously adds different trees to the model and grows the tree model through feature splitting, each time a tree is added it is equivalent to learning a new function, which gradually improves the performance of the model by training a series of weak learners iteratively, where each new weak learner corrects the prediction error of the previous round of weak learners. Meanwhile, XGBoost supports parallel processing, which enables it to effectively utilize multi-core processors to accelerate the model training process and improve the training efficiency. XGBoost uses L1 and L2 regularization, similar to ridge regression and LASSO, to control the complexity of the model, to prevent overfitting, and to improve the generalization ability of the model. Since the model contains multiple tunable parameters, careful parameter tuning is required and too many parameter choices can lead to overfitting. Due to the sensitivity to outliers, it is often necessary to handle outliers in the preprocessing stage to avoid their negative impact on the model.

### LightGBM

Light Gradient Boosting Machine (LightGBM) is a high-performance gradient boosting tree algorithm[48,49] similar to XGBoost. LightGBM is optimized in terms of data structure and algorithms to make it perform well in large-scale data and high-dimensional feature cases to perform well.

.LightGBM is suitable for a wide range of classification and regression problems, but proper tuning of the hyper-parameters is required to achieve optimal performance. The algorithm is based on a histogram-based learning approach and is applicable to multiple geological and chemical variables that may be involved in reservoir temperature prediction. A tree-based learning algorithm is employed to support parallel computing with high training and prediction speeds. The ability to flexibly capture nonlinear relationships in reservoir temperature data provides more accurate predictions. Meanwhile, the algorithm has a relatively low memory footprint and is suitable for use in resource-limited environments. However, it is more sensitive to outliers, which need to be dealt with in the preprocessing stage to avoid their negative impact on the model.

### Hyperparameters tuning

Previous studies usually directly divide the data into training set and testing set proportionally, when the data set is not large enough, different division methods, different models are obtained, and when the division method is not good enough, it is difficult to select a good model and parameters, cross-validation is an effective way to solve this problem[50]. N-fold cross-validation is performed by randomly dividing the data into n groups, with n − 1 group for training and 1 group for validation, each subset of the group is used as validation data, loop n times to train n model results, and average the results of all the groups to produce the individual accuracy of the model as the final result (Fig. 4).This method can effectively prevent model overfitting, and at the same time, it is better able to find more appropriate model parameters.

For the small dataset problem in this paper, GridSearchCV in sklearn is used to search for the optimal parameters and re-fit the model.GridSearchCV has two functions, grid search and cross-validation, which ensures that the parameter with the highest accuracy can be found within the specified parameter range, and auto-tuning, so that as long as the parameter is inputted, the optimal result and the parameters will be given.

## Results and discussion

### Comparison of different machine learning algorithms

The learning and training of four algorithms are implemented based on the sklearn machine learning package in the python open source library, and the hyperparameters of each algorithm are set as follows (Table 2), and the optimal parameters are selected by the GridSearchCV search in "Hyperparameters tuning".

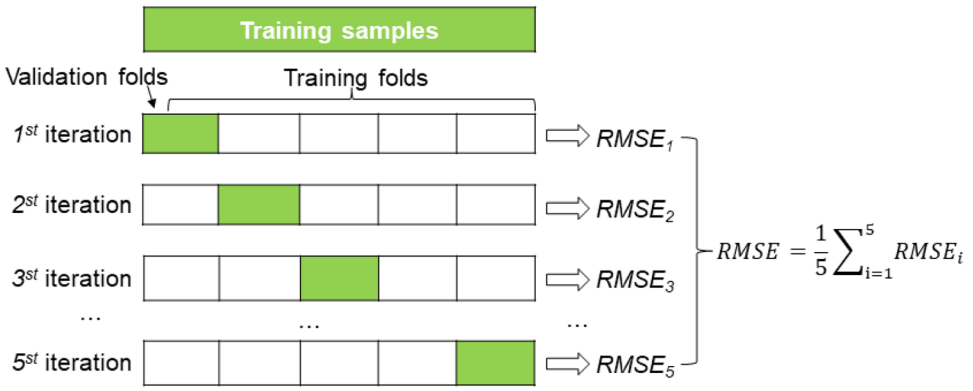


**Figure 4.** 5-fold cross-validation.

| | | |
|---|---|---|
| BayesianRidge Regression | lpha_1 | (−8, 0, 10) |
| | lpha_2 | (−8, 0, 10) |
| | ambda_1 | (−8, 0, 10) |
| | ambda_4 | (−8, 0, 10] |
| Decision Tree Regression | best_tree | [3, 5, 6, 7, 8, 9] |
| XGBoost | max_depth | [3, 5, 7] |
| | learning_rate | [0.01, 0.05, 0.1] |
| | n_estimators | [100, 500, 1000] |
| | subsample | [0.5, 0.7, 1.0] |
| | colsample_bytree | [0.5, 0.7, 1.0] |
| | reg_alpha | [0.01, 0.1, 1.0] |
| | reg_lambda | [0.01, 0.1, 1.0] |
| LightGBM | num_leaves | [20, 30, 40] |
| | learning_rate | [0.01, 0.05, 0.1] |
| | max_depth | [4, 5, 6] |
| | min_child_weight | [0.1, 1, 5] |

**Table 2.** Structure parameters.

*BayesianRidge regression*
Four control parameters, lpha_1 controls the normal prior, lpha_2 controls the observation error, ambda_1 controls the strength of all regression coefficients gradually approaching 0, and ambda_4 controls the strength of all regression coefficients gradually approaching a common value. The optimized values of the control parameters were obtained by applying a grid search. Table 2 shows the range of control parameter values chosen to obtain the optimal parameters, which were obtained from the search: lpha_1: 1.0, lpha_2: 1e−08, ambda_1: 1e−08, ambda_2: 1.

*Decision tree regression*
Decision tree regression mainly controlled by the maximum depth of the tree, in order to prevent overfitting in Table 2 set up for obtaining the optimal parameters of the control range, through the grid search to obtain the optimal decision tree model depth of 6.

*XGBoost*
XGBoost consists of seven control parameters, max_depth controls the maximum depth of the tree; learning_rate indicates the learning rate, which controls the iteration rate and prevents overfitting; n_estimators indicates the number of integrated weak estimators, the larger the n estimators are, the stronger the learning ability of the model is, and the easier the model is to overfitting; subsample controls the proportion of sampling from the sample; colsample_bytree controls the proportion of all features randomly sampled when constructing each tree; reg_alpha:L1 regularization coefficient; reg_lambda: :L2 regularization coefficient. Table 2 shows the range of control parameter values chosen to obtain the optimal parameters. The optimal parameters obtained from the search are: max_depth: 3, learning_rate: 0.1, n_estimators: 1000, subsample: 0.7, colsample_bytree: 1.0, reg_alpha. 0.01, reg_lambda: 1.0.

*LightGBM*
LightGBM controlled by four parameters, num_leaves is the number of leaf nodes on a tree, and max_depth to control the shape of the tree, the parameter has a great impact on the performance of the model, need to focus on regulating the parameters. learning_rate indicates the learning rate, choose a relatively small learning rate can obtain stable and better model performance, max_depth controls the maximum depth of the tree, too large a value of overfitting will be more serious, min_child_weight is the sum of all the samples in the smallest child node, the parameter is too large the model will be underfitting, too small will lead to overfitting, need to be adjusted according to the data. The model will be underfitted if the parameter is too large and overfitted if it is too small. The optimal parameters obtained through grid search are: num_leaves: 20, learning_rate: 0.05, max_depth: 4, min_child_weight: 0.1.

The prediction performance of the four machine learning algorithms was analyzed by optimal parameter selection, and Fig. 5 shows the prediction errors of the training dataset and the test dataset. From the figure, it can be concluded that the prediction error of the test dataset is basically similar to that of the training dataset. Among the four methods, the MAE of the XGBoost training dataset and the test dataset are 0.002 and 7.08, and the RMSE is 0.003 and 10.24, indicating that the algorithms have relatively low generalization. The LightGBM training dataset and the test dataset have the highest MAE and RMSE errors are the smallest and the generalizability is the highest. all four algorithms have good MAE and RMSE, indicating that the above machine learning algorithms can be used to predict reservoir temperature.

The prediction results of the training set and test set were plotted (Fig. 6), all the models had good prediction results in the training set. Migrating the trained models to the test set, the results showed that, the $R^2$ of Bayesian

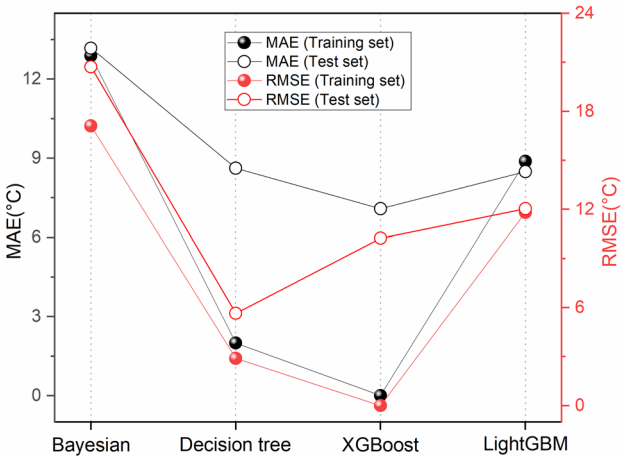


**Figure 5.** MAE and RMSE for the training and test sets of the 4 models (the solid center of the figure shows the training set and the hollow center shows the results of the test set; MAE is shown in black and RMSE is shown in red).

Ridge Regression Algorithm = 0.8302, Decision Tree Algorithm $R^2 = 0.96$, XGBoost Algorithm $R^2 = 0.9657$ and LightGBM Algorithm $R^2 = 0.9493$, Except for the Bayesian ridge regression algorithm, the predicted values of the three tree-based algorithms match well with the true values in the test set (all of them reach more than 94%). This shows that the tree-based machine learning algorithms can accurately predict the underground reservoir temperature. After the comprehensive evaluation of MAE, RMSE and $R^2$ indexes, it can be seen that the machine learning algorithm using XGBoost has the best accuracy in predicting the results, and LightGBM and decision tree algorithms are the second best.

## Comparison of different feature combinations

In this section, based on the correlation of features (Fig. 3) and combining the significance of different features, different forms of feature combinations are constructed and listed in Table 3, which are compared with the prediction performance of F-4 to explore the effect of feature selection on the temperature prediction performance.

Four algorithms were used to train the model for each of the above forms of feature combinations, and the prediction results of the training and test sets were plotted (Fig. 7). The results show that Bayesian Ridge regression has a better predictive ability of the model when choosing reasonable input parameters (F-1, F-2), and the predictive effect of the model decreases with the input of lower influencing factors (Fig. 7, red F-3, F-4); Decision Tree regression and XGBoost model have a small difference in the prediction error of the reservoir temperature under the conditions of different feature combinations (green and yellow dotted lines); LightGBM model has higher predictive ability of the model at F-2 and F-3, and the predictive effect of the model decreases at F-1 and F-4 (blue dotted line).

The test results of the four algorithms with the four combination forms are further compared (Fig. 7 and Table 4). The results show that the model accuracies of different algorithms with different feature combinations range from 0.8302 to 0.9732, and XGBoost performs the best with an average accuracy of 0.9703, followed by Decision Tree algorithm with an average accuracy of 0.9625, and LightGBM with an average accuracy of 0.9373. Among them, XGBoost and Decision Tree algorithms do not depend on the selection of features, and XGBoost does not depend on the selection of features with different feature combinations. XGBoost and Decision Tree algorithms are not strongly dependent on the selection of features, and the accuracy of XGBoost is in the range of 0.9657 to 0.9732 for different combinations of features; Decision Tree algorithm is the next best, and its accuracy is in the range of 0.96 to 0.97 for different combinations of features.The average value of the accuracy of Bayesian Ridge The accuracy of Bayesian Ridge Regression and LightGBM algorithms is unstable and sensitive to the selection of features, Bayesian Ridge Regression algorithm has the largest fluctuation in accuracy, with a minimum of 0.8302 when using the feature combination F-4, and a maximum of 0.9537 when using the feature combination F-1; LightGBM algorithm has a maximum of 0.9537 when using the feature combination F-1. The accuracy of LightGBM algorithm is slightly lower than 0.9 when only feature combination F-1 is used, and the average accuracy of different algorithms is maximum 0.9577 when feature combination F-2 is used, followed by feature combination F-1 (the average accuracy of the algorithms is 0.9482), and the average accuracy of the algorithms is minimum 0.9263 when feature combination F-4 is used.

By comparing the prediction performance of the four modeling algorithms with the F-4 combination, it is shown that a reasonable selection of input features can improve the prediction results of the model.

Furthermore, the evaluation results of different algorithms using various feature combinations based on the MAE and RMSE metrics are presented in Table 5 and Fig. 8. The results indicate that, for different algorithms using different feature combinations, the Mean Absolute Error (MAE) ranges from 6.0287 to 14.4431, and the Root Mean Squared Error (RMSE) ranges from 4.78 to 20.7185. XGBoost and Decision Tree algorithms exhibit the smallest prediction errors. XGBoost achieves an MAE ranging from 6.0287 to 7.08 across different feature

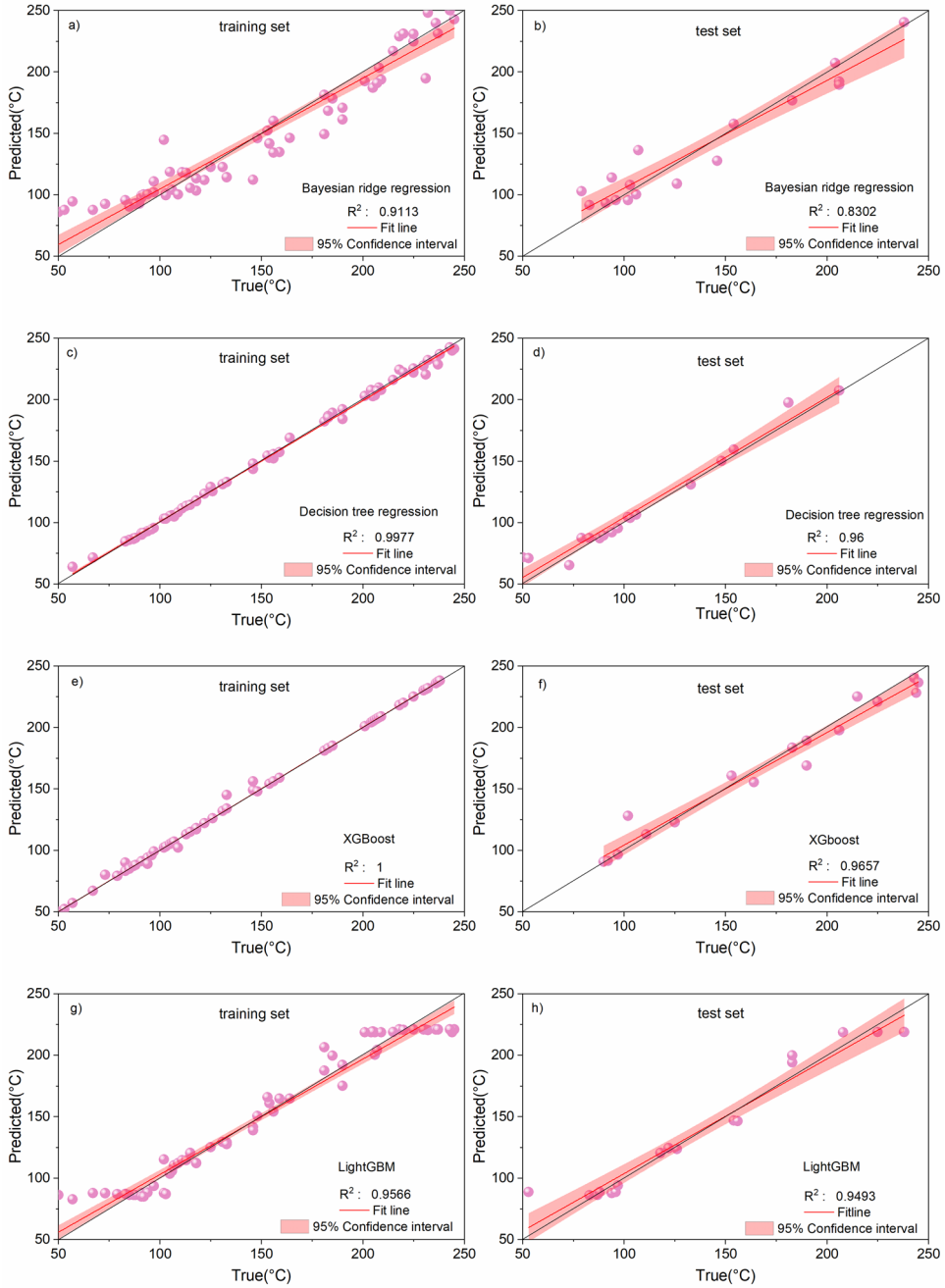


**Figure 6.** Prediction results and $R^2$ for the training and test sets of the 4 models.

| Feature Combinations | PH | EC (μS/cm) | $K^+$ (mg/l) | $Na^+$ (mg/l) | Boron (mg/l) | $SiO_2$ (mg/l) | Cl– (mg/l) |
|---|---|---|---|---|---|---|---|
| F-1 | | √ | √ | √ | | √ | |
| F-2 | | √ | √ | √ | √ | √ | |
| F-3 | √ | √ | √ | √ | √ | √ | |
| F-4 | √ | √ | √ | √ | √ | √ | √ |

**Table 3.** Four different feature combinations.

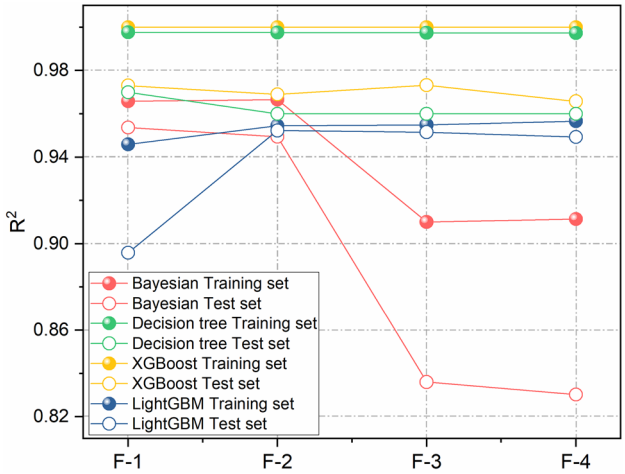


**Figure 7.** $R^2$ of different algorithms using different feature combinations.

| Feature combinations | Bayesian Ridge Regression | Decision Tree Regressor | XGBoost | LightGBM | Mean |
|---|---|---|---|---|---|
| F-1 | 0.9537 | 0.9700 | 0.9731 | 0.8958 | 0.9482 |
| F-2 | 0.9494 | 0.9600 | 0.9690 | 0.9523 | 0.9577 |
| F-3 | 0.8360 | 0.96 | 0.9732 | 0.9516 | 0.9302 |
| F-4 | 0.8302 | 0.96 | 0.9657 | 0.9493 | 0.9263 |
| Mean | 0.8923 | 0.9625 | 0.9703 | 0.9373 | |

**Table 4.** $R^2$ of different algorithms using different feature combinations.

| Feature combinations | Criterion | Bayesian ridge regression | Decision tree regression | XGBoost | LightGBM | Mean |
|---|---|---|---|---|---|---|
| F-1 | MAE | 8.5781 | 7.25 | 6.0287 | 14.4431 | 8.4575 |
| | RMSE | 10.8248 | 4.78 | 9.0757 | 17.8509 | 11.2503 |
| F-2 | MAE | 9.1957 | 8.06 | 6.6483 | 8.3928 | 7.3392 |
| | RMSE | 11.3144 | 5.12 | 9.7433 | 11.6747 | 10.1981 |
| F-3 | MAE | 11.9209 | 8.29 | 5.74627 | 8.3489 | 7.9040 |
| | RMSE | 20.3617 | 5.6 | 9.0491 | 11.7625 | 12.3658 |
| F-4 | MAE | 13.1750 | 8.62 | 7.08 | 8.4929 | 8.5995 |
| | RMSE | 20.7185 | 5.65 | 10.2402 | 12.0344 | 12.9033 |

**Table 5.** MAE and RMSE of different algorithms using different feature combinations.

combinations, followed by the Decision Tree algorithm with an MAE range of 7.25 to 8.62. However, Bayesian Ridge Regression and LightGBM algorithms exhibit instability and sensitivity to feature selection. When using only feature combination F-1, the LightGBM algorithm shows larger MAE and RMSE values. The Bayesian Ridge Regression algorithm demonstrates the greatest fluctuation in error, with RMSE exceeding 20 when using feature combinations F-3 and F-4. With feature combination F-2, the algorithms exhibit the smallest average errors, with an MAE of 7.3392 and RMSE of 10.1981. Next is feature combination F-3, with an average MAE of 7.9094 and RMSE of 12.3658. Feature combination F-4 results in the largest average errors for the algorithms, with an MAE of 8.5995 and RMSE of 12.9033.

We further calculated the running time of each model under various combinations (Fig. 9), which reflects the computational performance of each model by recording the execution time of each model's learning curve, and the sum of the validation time for the validation sample size. Usually, an optimal model requires less running time, while a bad model will be very time-consuming, as can be seen in Fig. 9, the running efficiencies are LightGBM, Decision Tree Regression, XGBoost, and Bayesian Ridge Regression in descending order.

Taking into consideration the evaluation results of the four metrics mentioned above, the optimal combination of reservoir temperature prediction is as follows: when the feature combination F-3 is adopted and XGBoost algorithm is selected, the model error is minimum and the accuracy is maximum of 0.9732.

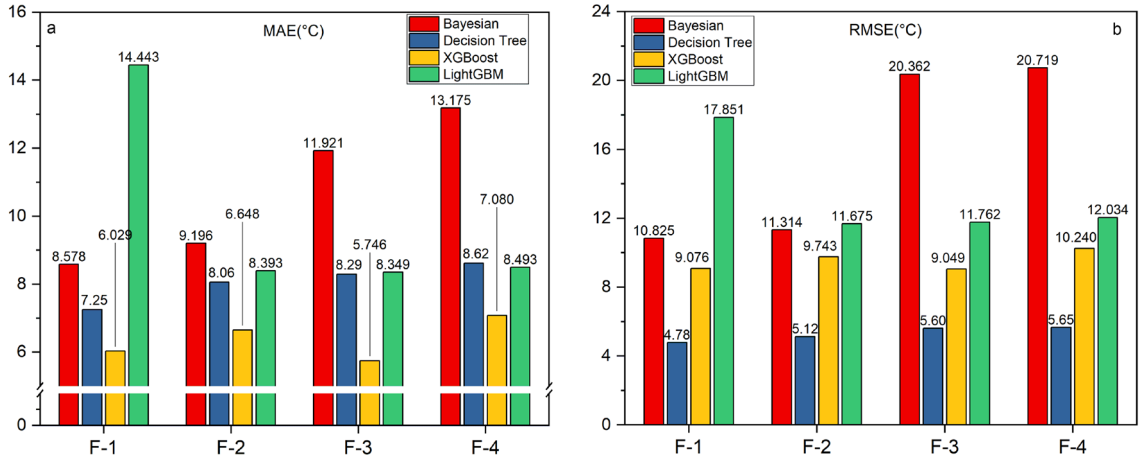


**Figure 8.** MAE and RMSE of different algorithms using different combinations of features.

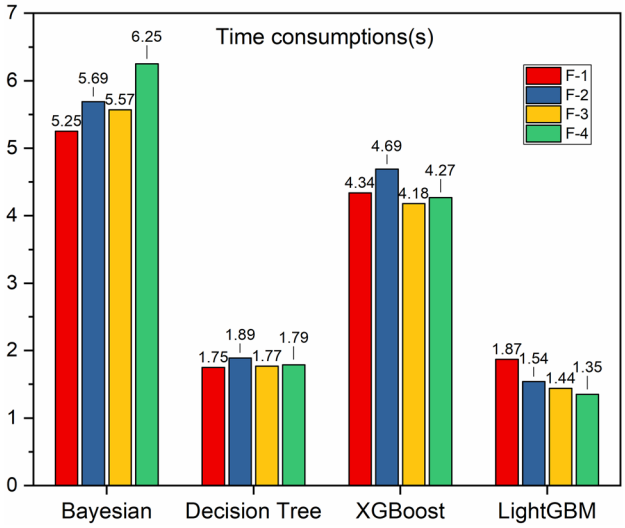


**Figure 9.** Time consumptions for learning curves of machine learning models.

## Comparison with traditional geothermometer methods

Based on the sodium–potassium cation and $SiO_2$-based geothermometers (Suppl. Appendix B), a comparison between the predicted reservoir temperatures using the geochemical geothermometer formula and the measured temperatures is presented (Fig. 10). and The sodium–potassium geothermal temperature scale formula is based on cation-exchange reactions and does not apply to hot water where mixing of hot water of different origins occurs, nor does it apply to acidic water with a pH much less than 7[11–13]. The silica geothermometer method is based on the solubility of silica minerals and is applicable in the interval of 20 to 250°C[51]. Therefore, only the alkaline (pH greater than 7) dataset was selected for the calculations when sodium–potassium geothermal temperature scaling method was used.

The results show that the $Na/K^1$ geothermometer has good prediction results mainly for reservoirs above 150 °C and is not applicable to this dataset; the results of the geothermometers by $Na/K^2$, $Na/K^3$, $Na/K^4$, $Na/K^5$ $Na/K^6$ and $Na/K^7$ are much larger than the actual temperatures, which are not in accordance with the actual situation. For the $SiO_2$ ground thermometer, the estimated temperature of $SiO_2{}^1$ is slightly larger than the actual temperature; the estimated temperatures of $SiO_2{}^4$, $SiO_2{}^5$, $SiO_2{}^6$, and $SiO_2{}^7$ are lower than the actual temperature; the estimated temperatures of $SiO_2{}^2$ and $SiO_2{}^3$ are close to the actual temperature and provide reasonable predictions for the geothermal field. The $SiO_2{}^2$ and $SiO_2{}^3$ methods yield Mean Absolute Errors (MAE) of 30.80 and 51.16, respectively, and Root Mean Square Errors (RMSE) of 57.56 and 71.65, respectively.

The $SiO_2{}^2$ and $SiO_2{}^3$ geothermometer estimation results are compared with the machine learning based reservoir temperature prediction. All the prediction results based on machine learning are superior to the $SiO_2{}^2$ and $SiO_2{}^3$ geothermometer method. The difference in the error distribution of the two types of geothermometers is large, which indicates that the machine learning algorithms have a certain degree of superiority.

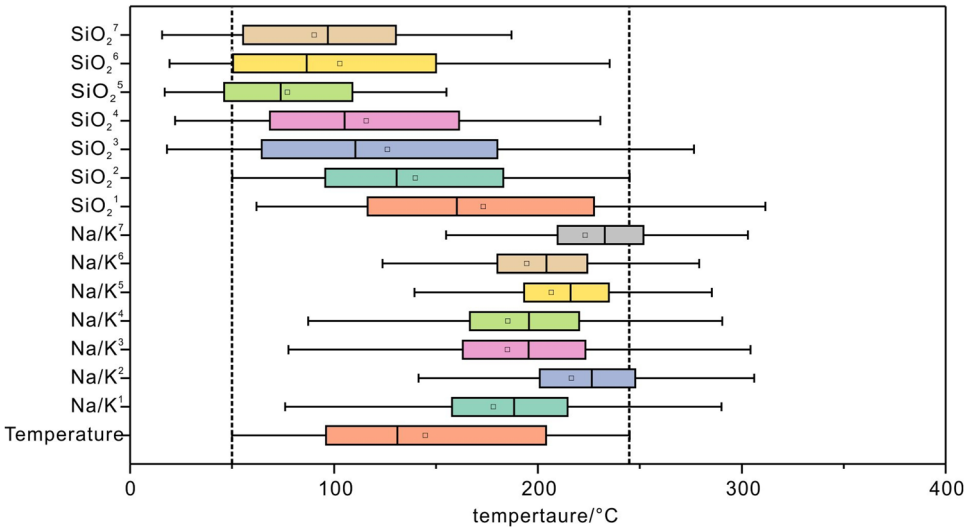


**Figure 10.** Comparison of predicted reservoir temperature with measured temperature based on chemical geothermometer.

## Generalizability analysis of the model

In the most ideal case, we expect the model to perform without substantial performance bias when applied to different datasets. That is, a model with good generalization is able to successfully apply the patterns or laws learned during training to new and different datasets, rather than just performing well on the training data. In this section, we validate the generalizability of the model by applying data from previous published results [Shadfar Davoodi, Hung Vo Thanh (2023), Shadfar Davoodi, Mohammad Mehrad (2023)], brought into our trained model for prediction. Table 6 compares the RMSE, MAE and $R^2$ values achieved using the modeling algorithms of this study for predictions presented in published studies. The results in Table 6 show that placing new data into the model is still able to make good predictions and the most accurate predictions are for the XGBoost model of this study with RMSE ,MAE and $R^2$ of 0..328, 0.228 and 0.997 respectively.

## Conclusions

In this paper, reservoir prediction models with different machine learning algorithms were trained using the same dataset based on the geothermal dataset of western Turkey, and the prediction performance of Bayesian Ridge Regression, Decision Tree Regression, XGBoost, and LightGBM was compared, and the effect of different feature combinations on the prediction performance of reservoir temperature was investigated, and the results were compared with those of the estimates from traditional geothermometers. Based on the above studies, the main findings are as follows:

Without considering the data features, among the four algorithms, the machine learning algorithm of XGBoost has the best accuracy of $R^2 = 0.9657$, followed by LightGBM and Decision Tree algorithms.
By comparing the predictive results of different feature combinations, a reasonable selection of input features can improve the prediction results and prediction efficiency of the model.
When the optimal combination for reservoir temperature prediction is feature combination F-3 and the XGBoost algorithm is chosen, the model error is minimized, achieving the highest accuracy of 0.9732.

| ML model | RMSE | MAE | $R^2$ | Data sources |
|---|---|---|---|---|
| Bayesian ridge regression | 17.105 | 12.884 | 0.911 | Davoodi and Vo Thanh (2023)[34] |
| Decision tree regression | 12..719 | 9.917 | 0.927 | |
| XGBoost | 10.240 | 7.081 | 0.968 | |
| LightGBM | 11.805 | 8.876 | 0.956 | |
| Bayesian ridge regression | 8.167 | 6.495 | 0.827 | Davoodi and Mehrad (2023)[36] |
| Decision tree regression | 3.467 | 2.161 | 0962 | |
| XGBoost | 0.328 | 0.228 | 0.997 | |
| LightGBM | 12.596 | 8.565 | 0.882 | |

**Table 6.** Model performance in previously studied data.

The prediction accuracy and stability of the machine learning method are obviously better than that of the traditional geothermometer method, and the results of this study can help the application of machine learning in reservoir temperature prediction and extend it to the related fields of engineering geology.

## Outlook

In the future, the effect of deep reservoir data characteristics on reservoir temperature will be further explored, and the effect of steam fractions on element concentration will be considered in the model and whether the data is from a single well or multiple Wells.

## Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request, and the underlying data are given in the Appendix of this paper.

Received: 14 November 2023; Accepted: 10 April 2024
Published online: 06 May 2024

## References

1. Lund, J. W., Freeston, D. H. & Boyd, T. L. Direct application of geothermal energy; 2005 worldwide review. *Geothermics* **34**, 691–727. https://doi.org/10.1016/j.geothermics.2005.09.003 (2005).
2. Neupane, G. *et al.* Geothermometric evaluation of geothermal resources in southeastern Idaho. *Geotherm. Energy Sci.* **4**, 11–22 (2016).
3. Guo, Q., Pang, Z., Wang, Y. & Tian, J. Fluid geochemistry and geothermometry applications of the Kangding high-temperature geothermal system in eastern Himalayas. *Appl. Geochem.* **81**, 63–75. https://doi.org/10.1016/j.apgeochem.2017.03.007 (2017).
4. Hou, Z. *et al.* Reconstruction of different original water chemical compositions and estimation of reservoir temperature from mixed geothermal water using the method of integrated multicomponent geothermometry; A case study of the Gonghe Basin, northeastern Tibetan Plateau, China. *Appl. Geochem.* **108**, 104389. https://doi.org/10.1016/j.apgeochem.2019.104389 (2019).
5. Okoroafor, E. R. *et al.* Machine learning in subsurface geothermal energy: Two decades in review. *Geothermics* **102**, 102401. https://doi.org/10.1016/j.geothermics.2022.102401 (2022).
6. Afandi, A., Lusi, N., Subono & Ayu Febriani, S. D. Prediction of the distribution of geothermal sources based on the geothermal temperature gradient in the Blawan Bondowoso. *Case Stud. Therm. Eng.* **25**, 100931 https://doi.org/10.1016/j.csite.2021.100931 (2021).
7. Kaeshkov, I. S., Kremenetskiy, M. I. & Buyanov, A. V. *SPE Russian Oil and Gas Exploration & Production Technical Conference and Exhibition*.
8. Hashish, R. G. & Zeidouni, M. Injection profiling in horizontal wells using temperature warmback analysis. *Comput. Geosci.* **25**, 215–232. https://doi.org/10.1007/s10596-020-10000-7 (2021).
9. Acevedo-Anicasio, A. *et al.* GaS_GeoT: A computer program for an effective use of newly improved gas geothermometers in predicting reliable geothermal reservoir temperatures. *Geotherm. Energy (Heidelberg)* **9**, 1–41. https://doi.org/10.1186/s40517-020-00182-9 (2021).
10. Jiexiang, L. *et al.* Estimates of reservoir temperatures for non-magmatic convective geothermal systems; insights from the Ranwu and Rekeng geothermal fields, western Sichuan Province, China. *J. Hydrol.* (*Amsterdam*) **609**, 127668 https://doi.org/10.1016/j.jhydrol.2022.127668 (2022).
11. Truesdell A, H. Summary of section III-Geochemical techniques in exploration. In *Proceedings of the 2nd U. N. Symposium on the Development and Use of Geothermal Resources*. Vol. 1. 1iii–1xxix (1976).
12. Fournier, R. O. Revised equation for the Na/K geothermometer. *Geotherm. Resour. Council* **3**, 221–224 (1979).
13. Tonani, F. B. Some remarks on the application of geochemical techniques in geothermal exploration. In *Commission of the European Communities* (*Report*) *EUR*. 428–445. https://doi.org/10.1007/978-94-009-9059-3_38 (1980).
14. Arnórsson, S., Gunnlaugsson, E. & Svavarsson, H. The chemistry of geothermal waters in Iceland. III. Chemical geothermometry in geothermal investigations. *Geochim. Cosmochim. Acta* **47**, 567–577 https://doi.org/10.1016/0016-7037(83)90278-8 (1983).
15. Nieva, D. & Nieva, R. Developments in geothermal energy in Mexico—Part twelve. A cationic geothermometer for prospecting of geothermal resources. *Heat Recov. Syst. CHP* **7**, 243–258 https://doi.org/10.1016/0890-4332(87)90138-4 (1987).
16. Verma, S. P. & Santoyo, E. New improved equations for NaK, NaLi and $SiO_2$ geothermometers by outlier detection and rejection. *J. Volcanol. Geotherm. Res.* **79**, 9–23. https://doi.org/10.1016/S0377-0273(97)00024-3 (1997).
17. Can, I. A new improved Na/K geothermometer by artificial neural networks. *Geothermics* **31**, 751–760. https://doi.org/10.1016/S0375-6505(02)00044-5 (2002).
18. Díaz-González, L., Santoyo, E. & Reyes-Reyes, J. Three new improved Na/K geothermometers using computational and geochemiometrical tools: Application to the temperature prediction of geothermal systems. *Rev. Mex. Cienc. Geol.* **25**, 465–482 (2008).
19. Fournier, R. O. & Truesdell, A. H. An empirical Na–K–Ca geothermometer for natural waters. *Geochim. Cosmochim. Acta* **37**, 1255–1275. https://doi.org/10.1016/0016-7037(73)90060-4 (1973).
20. Giggenbach, W. F. Geothermal solute equilibria. Derivation of Na–K–Mg–Ca geoindicators. *Geochim. Cosmochim. Acta* **52**, 2749–2765. https://doi.org/10.1016/0016-7037(88)90143-3 (1988).
21. Fournier, R. O. Chemical geothermometers and mixing models for geothermal systems. *Geothermics* **5**, 41–50. https://doi.org/10.1016/0375-6505(77)90007-4 (1977).
22. Arnórsson, S., Gunnlaugsson, E. & Svavarsson, H. The chemistry of geothermal waters in Iceland. II. Mineral equilibria and independent variables controlling water compositions. *Geochim. Cosmochim. Acta* **47**, 547–566. https://doi.org/10.1016/0016-7037(83)90277-6 (1983).
23. Dulanya, Z., Morales-Simfors, N. & Sivertun, Å. Comparative study of the silica and cation geothermometry of the Malawi hot springs: Potential alternative energy source. *J. Afr. Earth Sci.* **57**, 321–327 (2010).
24. Arnórsson, S. Gas chemistry of the Krísuvík geothermal field, Iceland, with special reference to evaluation of steam condensation in upflow zones (1987).
25. Blamey, N. J. F. Composition and evolution of crustal, geothermal and hydrothermal fluids interpreted using quantitative fluid inclusion gas analysis. *J. Geochem. Explor.* **116**, 17–27. https://doi.org/10.1016/j.gexplo.2012.03.001 (2012).
26. Barragan, R. M., Nunez, J., Arellano, V. M. & Nieva, D. EQUILGAS; program to estimate temperatures and in situ two phase conditions in geothermal reservoirs using three combined FT-HSH gas equilibria models. *Comput. Geosci.* **88**, 1–8. https://doi.org/10.1016/j.cageo.2015.12.009 (2016).
27. Perez-Zarate, D., Santoyo, E., Acevedo-Anicasio, A., Diaz-Gonzalez, L. & Garcia-Lopez, C. Evaluation of artificial neural networks for the prediction of deep reservoir temperatures using the gas-phase composition of geothermal fluids. *Comput. Geosci.* **129**, 49–68. https://doi.org/10.1016/j.cageo.2019.05.004 (2019).

28. Porkhial, S., Salehpour, M., Ashraf, H. & Jamali, A. Modeling and prediction of geothermal reservoir temperature behavior using evolutionary design of neural networks. *Geothermics* **53**, 320–327. https://doi.org/10.1016/j.geothermics.2014.07.003 (2015).
29. Tut Haklidir, F. S. & Haklidir, M. Prediction of reservoir temperatures using hydrogeochemical data, western Anatolia geothermal systems (Turkey): A machine learning approach. *Nat. Resour. Res.* **29**, 2333–2346. https://doi.org/10.1007/s11053-019-09596-0 (2020).
30. Varol Altay, E., Gurgenc, E., Altay, O. & Dikici, A. Hybrid artificial neural network based on a metaheuristic optimization algorithm for the prediction of reservoir temperature using hydrogeochemical data of different geothermal areas in Anatolia (Turkey). *Geothermics* **104**, 102476. https://doi.org/10.1016/j.geothermics.2022.102476 (2022).
31. Afandi, A., Lusi, N., Catrawedarma, I. G., Subono, N. B. & Rudiyanto, B. Prediction of temperature in 2 meters temperature probe survey in Blawan geothermal field using artificial neural network (ANN) method. *Case Stud. Therm. Eng.* **38**, 102309. https://doi.org/10.1016/j.csite.2022.102309 (2022).
32. Davoodi, S. *et al.* Machine-learning predictions of solubility and residual trapping indexes of carbon dioxide from global geological storage sites. *Exp. Syst. Appl.* **222**, 119796. https://doi.org/10.1016/j.eswa.2023.119796 (2023).
33. Davoodi, S. *et al.* Machine-learning models to predict hydrogen uptake of porous carbon materials from influential variables. *Sep. Purif. Technol.* **316**, 123807. https://doi.org/10.1016/j.seppur.2023.123807 (2023).
34. Davoodi, S., Vo Thanh, H., Wood, D. A., Mehrad, M. & Rukavishnikov, V. S. Combined machine-learning and optimization models for predicting carbon dioxide trapping indexes in deep geological formations. *Appl. Soft Comput.* **143**, 110408. https://doi.org/10.1016/j.asoc.2023.110408 (2023).
35. Davoodi, S., Mehrad, M., Wood, D. A., Ghorbani, H. & Rukavishnikov, V. S. Hybridized machine-learning for prompt prediction of rheology and filtration properties of water-based drilling fluids. *Eng. Appl. Artif. Intell.* **123**, 106459. https://doi.org/10.1016/j.engappai.2023.106459 (2023).
36. Davoodi, S., Mehrad, M., Wood, D. A., Rukavishnikov, V. S. & Bajolvand, M. Predicting uniaxial compressive strength from drilling variables aided by hybrid machine learning. *Int. J. Rock Mech. Min. Sci.* **170**, 105546. https://doi.org/10.1016/j.ijrmms.2023.105546 (2023).
37. Li, Y. *et al.* Temperature changes the dynamics of trace element accumulation in *Solanum tuberosum* L.. *Clim. Change* **112**, 655–672. https://doi.org/10.1007/s10584-011-0251-1 (2012).
38. Mann, U., Frost, D. J., Rubie, D. C., Becker, H. & Audétat, A. Partitioning of Ru, Rh, Pd, Re, Ir and Pt between liquid metal and silicate at high pressures and high temperatures—Implications for the origin of highly siderophile element concentrations in the Earth's mantle. *Geochim. Cosmochim. Acta* **84**, 593–613. https://doi.org/10.1016/j.gca.2012.01.026 (2012).
39. Ioffe, S. & Szegedy, C. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. arXiv:1502.03167. https://ui.adsabs.harvard.edu/abs/2015arXiv150203167I (2015).
40. Bermingham, M. L. *et al.* Application of high-dimensional feature selection: Evaluation for genomic prediction in man. *Sci. Rep.* **5**, 10312–10312. https://doi.org/10.1038/srep10312 (2015).
41. Pedregosa, F. *et al. Scikit-learn: Machine Learning in Python*. Vol. 12. 2825–2830 (2011).
42. Luo, J., Gan, Y., Vong, C.-M., Wong, C.-M. & Chen, C. Scalable and memory-efficient sparse learning for classification with approximate Bayesian regularization priors. *Neurocomputing (Amsterdam)* **457**, 106–116. https://doi.org/10.1016/j.neucom.2021.06.025 (2021).
43. Magris, M. & Iosifidis, A. Bayesian learning for neural networks: An algorithmic survey. *Artif. Intell. Rev.* **56**, 11773–11823. https://doi.org/10.1007/s10462-023-10443-1 (2023).
44. Tipping, M. E. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **1**, 211–244. https://doi.org/10.1162/15324430152748236 (2001).
45. Nordhausen, K. The elements of statistical learning: Data mining, inference, and prediction, second edition, Trevor Hastie, Robert Tibshirani, Jerome Friedman. *Int. Stat. Rev./Rev. Int. Stat.* **77**, 482–482 (2009).
46. Chen, T. & Guestrin, C. *XGBoost: A Scalable Tree Boosting System*. arXiv:1603.02754. https://ui.adsabs.harvard.edu/abs/2016arXiv160302754C (2016).
47. Zhong, R., Johnson, R. & Chen, Z. Generating pseudo density log from drilling and logging-while-drilling data using extreme gradient boosting (XGBoost). *Int. J. Coal Geol.* **220**, 103416. https://doi.org/10.1016/j.coal.2020.103416 (2020).
48. Meng, Q. *et al. A Communication-Efficient Parallel Algorithm for Decision Tree*. arXiv:1611.01276. https://ui.adsabs.harvard.edu/abs/2016arXiv161101276M (2016).
49. Zhang, H., Si, S. & Hsieh, C.-J. *GPU-Acceleration for Large-scale Tree Boosting*. arXiv:1706.08359. https://ui.adsabs.harvard.edu/abs/2017arXiv170608359Z (2017).
50. Pu, Y., Apel, D. B. & Hall, R. Using machine learning approach for microseismic events recognition in underground excavations: Comparison of ten frequently-used models. *Eng. Geol.* **268**, 105519. https://doi.org/10.1016/j.enggeo.2020.105519 (2020).
51. Verma, M. P. Chemical thermodynamics of silica: A critique on its geothermometer. *Geothermics* **29**, 323–346. https://doi.org/10.1016/S0375-6505(99)00064-4 (2000).


## Acknowledgements

This study was supported by the National Natural Science Foundation of China (project ID: 42172274).

## Author contributions

Haoxin Shi: Data curation, Formal analysis, Methodology, Validation, Visualization, Writing-original draft. Yanjun Zhang: Conceptualization, Supervision, Writing-review & editing. Ziwang Yu: Conceptualization, Funding acquisition, Methodology, Validation. Yunxing Yang: Visualization, Writing-review & editing.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-59409-5.

**Correspondence** and requests for materials should be addressed to Y.Z.