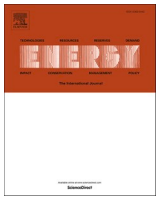
Energy 280 (2023) 128255

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy



Comparative data-driven enhanced geothermal systems forecasting models: A case study of Qiabuqia field in China

Zhenqian Xue [a], Kai Zhang [b,c,d], Chi Zhang [a], Haoming Ma [a], Zhangxin Chen [e,a,*]

[a] Department of Chemical & Petroleum Engineering, University of Calgary, 2500 University Drive NW, Calgary, Alberta, T2N 1N4, Canada
[b] Key Laboratory of Tectonics and Petroleum Resources, Ministry of Education, China University of Geosciences (Wuhan), Wuhan, 430074, China
[c] Key Laboratory of Theory and Technology of Petroleum Exploration and Development in Hubei Province, China
[d] School of Earth Resources, China University of Geosciences (Wuhan), Wuhan, China
[e] Eastern Institute for Advanced Study, Ningbo, China

ARTICLE INFO

Handling Editor: G Iglesias

Keywords:
Geothermal energy
Machine learning
K-nearest neighbors
Support vector machine
Extreme gradient boosting
Artificial neural network

ABSTRACT

Geothermal energy is gaining global attractiveness owing to its abundance and sustainable nature. An in-depth understanding of potential geothermal production provides the energy industry a possibility to diversify the supply portfolio. With the development of artificial intelligence, machine learning offers an efficient alternative to the conventional numerical simulation method in forecasting energy harvesting. However, a comprehensive comparison and an effective algorithm selection are absent from the machine learning applications in forecasting geothermal energy recovery. In this study, four machine learning algorithms based data-driven models are created to determine the optimal choice in predicting geothermal production, including *K*-Nearest Neighbors (KNN), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost) and Artificial Neural Network (ANN). To investigate their application range, two different sizes of data groups are involved to train and test these models, and their performance is comprehensively compared. As the results show, the highest coefficient of determination R^2 of 0.998 is demonstrated in the ANN models showing its promising predictive ability. Besides, the ANN is the most stable with the lowest performance variances between a training set and a validation set. In addition, the ANN is the most adaptable due to its minimal performance differences between different sizes of data groups. By jointly considering the prediction accuracy, stability and adaptability, the ANN is the best choice to substitute numerical simulation for predicting geothermal development. Importantly, the successful implementation of the proposed data-driven model requires 2700 times less computational time compared to numerical simulation, demonstrating a considerable improvement in the prediction efficiency. The results provide a beneficial reference for operators in conducting machine learning to simulate the development of the geothermal system studied, and can be effectively applied in other energy systems.

1. Introduction

Fossil fuels are getting depleted due to the rapidly increasing energy demand, and a significant amount of CO_2 emission is emited by the consumption of fossil fuels, which causing harmful effects on climate change [1,2]. Therefore, most countries have focused on developing sustainable and renewable resources [3]. Geothermal energy can be a substitute for fossil fuels to meet the future energy demand and mitigate climate change owing to its sustainable and low-carbon features. It can be extensively utilized in direct thermal energy harvesting and electricity generation, and the estimated geothermal resources under exploitation reach 30,000 MW, which can avoid the emissions of 46 million tons of CO_2 and save 0.352 billion barrels of oil annually [4]. Hot dry rock (HDR) is one kind of geothermal resource and has a promising potential due to its abundant reserves and wide distributions [5]. The HDR geothermal resources in China are calculated to be 2.6×10^5 times its total annual energy consumption [6]. In addition, it is estimated that the thermal resources in HDR mass on the earth are 100–1000 times larger than fossil energy resources, which can provide global energy needs for approximately 217 million years [7–9].

China has pledged to peak CO_2 emissions by 2030 and achieve carbon neutrality by 2060 [10]. The development of geothermal resources, as a sustainable energy, is determined as an important option in the China National Strategies responding to climate change by The State Council Information Office of the People's Republic of China [10]. In

* Corresponding author. Department of Chemical & Petroleum Engineering, University of Calgary, 2500 University Drive NW, Calgary, Alberta, T2N 1N4, Canada.
E-mail address: zhachen@ucalgary.ca (Z. Chen).

https://doi.org/10.1016/j.energy.2023.128255
Received 17 November 2022; Received in revised form 15 May 2023; Accepted 23 June 2023
Available online 24 June 2023
0360-5442/

Nomenclature

EGS	Enhanced geothermal system
HDR	Hot dry rock
KNN	*K*-Nearest Neighbors
SVM	Support Vector Machine
XGBoost	Extreme Gradient Boosting
ANN	Artificial Neural Network
CMG	Computer Modelling Group
GBDT	Gradient Boosting Decision Tree
SVR	Support vector regression
SLP	Single Layer Perception
MLP	Multilayer Perception
BP	Back-propagation algorithm
RMSE	Root mean squared error
MAE	Mean absolute error
R^2	Coefficient of determination

China, geothermal reservoirs have not yet been developed for long-term power generation [11,12]. The Qiabuqia field is located in the eastern part of the Gonghe basin, northwest China [13]. The total resources in this field have been reported to be probably 200 billion tons of standard coal [12]. The HDR is abundant in its formation from 3200 m to 3705 m in depth with an average temperature of 218 °*C* [11,14]. There are ten available geothermal exploration wells, of which well GR1 is the most promising due to its 236 °*C* peak temperature at 3705 m depth, implying a potential for the construction of an EGS power plant [12,15,16]. Therefore, the Qiabuqia field is selected as the case study in this paper.

However, they are impossible to be directly extracted because of a low permeability of HDR [17]. An enhanced geothermal system (EGS) is, therefore, proposed to develop a highly conductive area in HDR through stimulation methods such as hydraulic fracturing [18]. Energy is captured by a circulating fluid extracting heat from HDR [17,19,20]. The extraction of geothermal energy in HDR is a complex process due to the operation of an EGS. Geothermal production depends on many properties, and its operational factors have significant implications, such as the properties of a circulating fluid, well completion parameters and stimulation method designs [21–26]. These factors exhibit different effects on productivity. Lei et al. (2019) [12] found that a geothermal recovery factor increased as well spacing increased, and decreased as water injection temperature increased. Additionally, Lei et al. (2020) [11] also showed that the influences of well spacing and an injection rate on geothermal development were highly dependent on a fracture half-length. Cai et al. (2022) [27] proposed a novel method to detect fracture driven interaction when pressure gauge is available that could assist the well spacing optimization. Song et al. (2021) [28] revealed that an injection rate had an optimum value for the highest productivity under certain conditions. Chong (2021) [29] demonstrated that a negative relationship between an injection rate and geothermal production occurred in the case of low matrix permeability, while there was a nonlinear correlation between the injection rate and productivity when the matrix permeability is high. Liu et al. (2021) [30] observed that a number of fractures had a significant effect on the production temperature and heat extraction. Zinsalo et al. (2021) [31] observed that the fracture number and productivity presented a nonlinear relationship. Dahi et al. (2020) [32] also showed non-uniform fracture closure could affect the production behavior of the fracture.

Numerical simulation is widely used for forecasting heat extraction performance because of its predictive ability [23]. However, due to complex correlations between these operational conditions and geothermal productivity, production prediction is time-consuming and computationally expensive for numerical simulation [33]. Machine learning provides an alternative way for this task since it can analyze collected data comprehensively and evaluate the underlying relationships without prior knowledge of the data [34]. A critical aspect of a high-performance machine learning algorithm is its data handling, such as feature engineering. It aims to understand the data comprehensively and generate effective features, which can be optimized effectively through the numerical simulation method under domain knowledge [35–39]. Therefore, with a combination of machine learning and domain knowledge, data-driven models can be more effective [34].

Machine learning applications in the geothermal industry have been extensively studied, as shown in Table 1. Different machine learning algorithms were used to predict different geothermal productivities, and their predictive ability has been proven. However, there is no study available to demonstrate an optimal selection among various machine learning algorithms in this area. Different machine learning algorithms can lead to significant variances in their computational time and prediction effectiveness. Optimal data-driven models can forecast geothermal energy production much more effectively than the conventional numerical method in terms of describing complex nonlinear relationships between operational parameters and geothermal electricity

Table 1
Previous machine learning studies in the geothermal industry.

Author & Year	Machine learning algorithm	Research purpose
Keçebaş et al., 2012 [41]	Artificial neural network	Determine energy and exergy efficiencies of the he Afyonkarahisar geothermal district heating system
Porkhial et al., 2015 [42]	Artificial neural network	Determine the reservoir temperature of the Sabalan geothermal field
Tugcua et al., 2017 [43]	Artificial neural network	Predict cooling effect coefficient, exergy efficiency and net present value of geothermal energy aided absorption refrigeration system
Rezvanbehbahani et al., 2017 [44]	Gradient Boosted Regression Tree	Predict the geothermal heat flux in Greenland
Ishitsuka et al., 2018 [45]	Artificial neural network	Estimate the temperature distribution of the Kakkonda geothermal field
Zhou et al., 2019 [22]	Artificial neural network	Predict the production temperature of the Zhacang geothermal field in China
Tut et al., 2020 [46]	Deep neural network	Predict geothermal reservoir temperatures of different geothermal systems
Lösing et al., 2021 [47]	Extreme Gradient Boosting	Predict geothermal heat flow in Antarctica
Hu et al., 2021 [48]	Artificial neural network	Predict the hourly performance of the hybrid system based on an organic Rankine cycle
Senturk et al., 2021 [49]	Artificial neural network	Calculate energy efficiency and exergy efficiency of the geothermal energy powered Kalina cycle
Shahdi et al., 2021 [50]	Random Forest	Predict subsurface temperature and geothermal gradient of Northeastern United States
Bourhis et al., 2021 [51]	Extreme Gradient Boosting	Calculate the undisturbed ground temperature, the ground effective thermal conductivity and the borehole thermal resistance of central and western Switzerland
He et al., 2022 [52]	Support Vector Machine	Predict geothermal heat flow in the Bohai Bay Basin in China
Mehrenjani et al., 2022 [53]	Artificial neural network	Predict hydrogen production, total cost rate and exergy efficiency based on geothermal energy of an organic Rankine cycle
Pei et al., 2022 [54]	Artificial neural network	Predict the long-term performance of energy pile design of six typical thermal load distributions in China
Yang et al., 2022 [55]	Artificial neural network	Predict reservoir temperature of Lindian geothermal field in China
Xiao et al., 2022 [56]	Artificial neural network	Predict outlet temperature of abandoned exploitation wells geothermal power plants

generation. Besides, accuracy was generally selected as the only criterion for evaluating machine learning model performance in previous works, while other performance of a machine learning model, such as the stability to explore if a model has the problems of overfit or underfit, has not been considered. In addition, a fixed size of a dataset was used to train and test their machine learning models. The performance of a data-driven model relies on the sizes of datasets [40]. However, the available real field data varies from field to field. Therefore, it should investigate the model performance on different sizes of datasets to improve the application range of machine learning. Importantly, most of current EGSs has not been produced and there is no available real data that can describe the potential outcomes of different operational conditions under field variabilities. Therefore, machine learning engineer normally uses numerical simulation to collect data samples for training machine learning models. However, numerical simulation is a time-consuming method. Therefore, the exploration of model performance on different sizes of datasets can also help engineers save time in creating data samples by using numerical simulation, which has been ignored in previous research.

This study proposed four data-driven models based on machine learning algorithms to predict the geothermal electricity production of the Qiabuqia geothermal field. The objective of this research is to determine the optimal choice from four kinds of machine learning algorithms by jointly considering their accuracy, stability and adaptability for accurately describing the complex relationships between various working parameters and geothermal energy production in order to substitute numerical models. Besides, two different sizes of datasets are used to evaluate their prediction performance to improve the application range of the proposed data-driven model. The results provide a profitable strategy for predicting geothermal electricity, and also provide a valuable reference for the operators when applying machine learning in the geothermal industry, thereby facilitating the decision on an algorithm selection in different conditions.

2. Methodology

In this study, a hybrid numerical simulation and machine learning workflow is developed for predicting the production of an EGS in the Qiabuqia field. The Computer Modelling Group (CMG) STARS software is used to construct numerical models, which has been proven to provide high performance in the studies of geothermal reservoirs [57–63].

The numerical models are used to determine the input features and generate the data samples. First, a base numerical model is created to characterize the Qiabuqia geothermal field. In order to evaluate the reasonability of this base model, the reservoir properties and operating parameters are captured from the previous research [12], and the results of the average production temperature are compared. Subsequently, a sensitivity analysis is operated on six operating parameters to investigate their influence on geothermal electricity for determining the input features. Second, 2416 numerical models based on the base model are created by CMG CMOST. Different geothermal electricity values are calculated by randomly changing the values of input features within their excepted ranges. These ranges and data samples are filtered by the operational constraints in the Qiabuqia geothermal field. Finally, a series of input features values and their corresponding calculated geothermal electricity values constitute two data groups.

According to a nonlinear relationship between the input features and a target, four machine learning algorithms, including *K*-Nearest Neighbors (KNN), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost) and Artificial neuron network (ANN), are utilized to construct different data-driven models.

Before a model training, the generated data samples in each group are divided into two parts, including the training samples and the testing samples. Besides, a standardization method is utilized to eliminate a dimension difference between the input features, which can promote the model performance [64]. An evaluation matrix is then generated as the criterion for the model performance.

During a training process, four kinds of base data-driven models are first developed. To investigate the stability of these models, the training samples are then split into a training set and a validation set through a *K*-fold cross-validation method. This method can fully utilize the training samples. Each training sample can be used in fitting and validating, which can provide a comprehensive consideration for the model performance [65,66]. In addition, hyperparameter tuning is operated to find the optimal models. Specifically, the value matrices of hyperparameters are tested iteratively in the models until the performance in both the training and validation sets are all the best.

Consequently, an evaluating process is developed. The resulting optimal models are tested by the testing samples. The result of each model on the testing samples represents its prediction accuracy. A difference in model outcomes between the training and validation sets can infer its stability, and that between different groups can deduce its adaptability. Finally, the model performance is jointly analyzed by its prediction accuracy, stability and adaptability. Fig. 1 illustrates the flowchart in this study.

3. Numerical models

In this study, the CMG STARS software is used to create the numerical models and simulate the heat extraction process for the Qiabuqia EGS. This numerical simulator can estimate multi-dimensional fluid flow and heat transfer analyses in multi-phase, multi-component fluids in porous and fractured media [57,67]. Of which, the fluid flow and heat transfer in fractured porous media can be approximately simulated by the dual porosity approach in CMG STARS [68]. For a pure water-based geothermal system, CMG STARS has been proven to provide accurate simulations [57,58,60,61,67,69]. Therefore, this simulator is highly suitable for this study.

3.1. Model description

A 3D model of the Qiabuqia geothermal field is customized as the base model to investigate its in-depth performance and geothermal production. The following reasonable assumptions are made to simplify the numerical model: (1) The reservoir is assumed to be water saturated at the initial state; (2) The chemical reactions in the reservoir are not considered; (3) The aperture of the hydraulic fractures is assumed not to change during the heat extraction; (4) The fracture storage effect is assumed to be low.

The total volume of $2000 \times 1000 \times 3800$ m is created with a number of $40 \times 20 \times 38$ grid blocks. According to the HDR section in this field, the enhanced reservoir in this study is from 3200 m to 3700 m underground, and the grids of the enhanced reservoir are refined five times in the horizontal directions and three times in the vertical directions for accurate calculations. Three vertical wells are generated including one injection well (well GR1) located at the center and two production wells, which are shown in Fig. 2. The well spacing is 500 m each. According to the core analysis, the spacing of natural fractures is set to 10 m [11]. The setting of the reservoir properties in this model is shown in Table 2. Besides, five hydraulic fractures, with a half-length of 483 m, a height of 100 m, a width of 3 mm and an average permeability of 36,000 mD, are developed in the HDR section (3200–3700 m). This stimulation operation has been proven to be reasonable and can be beneficial in the productivity of the Qiabuqia geothermal field [12]. Other operating properties are the same as Lei et al.'s model, where a steady stream of 60 °C circulating water is injected into the reservoir at a rate of 30 kg/s and the production pressure is set to 37 MPa to obtain a stable production rate [12]. The geothermal productivity is calculated after a twenty-year operating cycle, which can represent the potential for the extraction of the Qiabuqia geothermal field [11,12,28]. The operational properties in this model can be found in Table 3.

For boundary conditions, there are no fluid flow and heat transfer

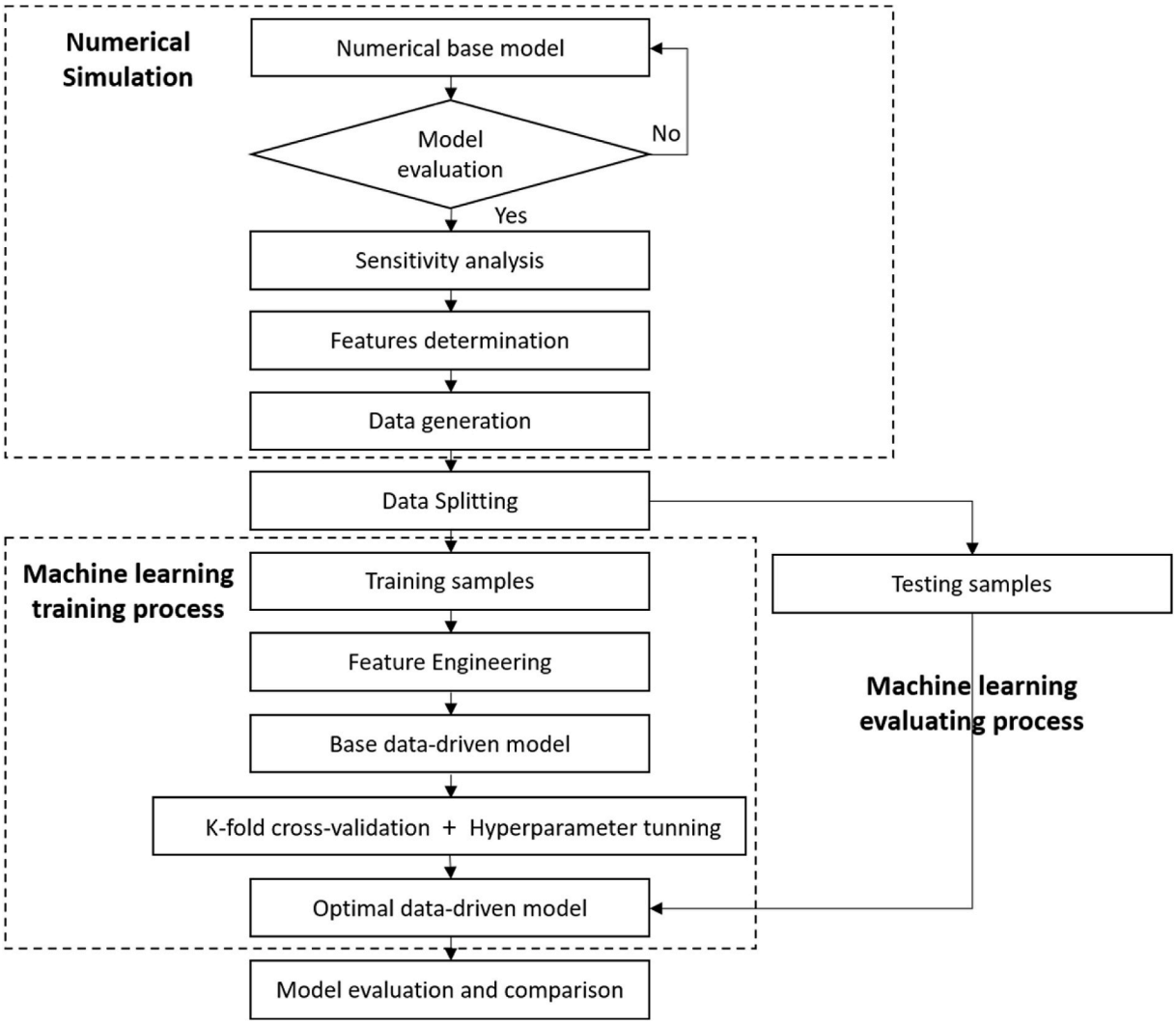


Fig. 1. The flowchart in this study.

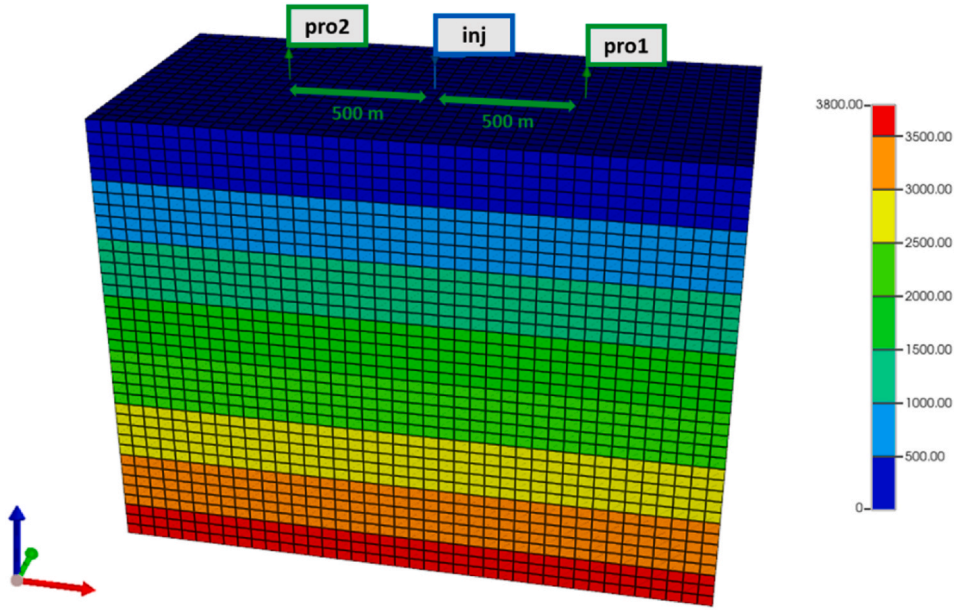


Fig. 2. 3D base model for the Qiabuqia geothermal field.

Table 2
Reservoir properties.

Parameter	Value
Granite density, ρ	2623 kg/m^3
Original porosity, φ	2.49 %
Original permeability, k	0.26 md
Granite heat conductivity, λ	3.0 W/ (m • °C)
Granite specific heat, C_R	980 J/ (kg • °C)
Horizontal natural fracture spacing, L_{n1}	10 m
Vertical natural fracture spacing, L_{n2}	10 m
Initial pressure, p	$P = 1.01 \times 10^5 - 10000z$ (P_a)
Initial temperature, T	$T = 25 - 0.057z$ (°C)

Table 3
Operational properties.

Parameter	Value
Injection fluid temperature, T_i	60 °C
Injection fluid rate, q_i	30 kg/s
Hydraulic fracture half-length, d_f	483 m
Hydraulic fracture number, N_f	5
Well spacing, N_f	500 m
Operating cycle	20 years
Production flow pressure, p_{out}	37 MPa

conditions at the vertical faces, and fluid flow is not allowed at both the top and bottom boundaries. The temperature at the bottom of the model is set according to the temperature distribution of this field. At the top boundary, a heat loss model is used to govern heat losses from the top surface to the atmosphere. At initial conditions, the temperature and pressure distributions are assumed based on the logging data from well GR1, where the initial reservoir temperature is $T = 25 - 0.057z\ (°C)$ and the initial reservoir pressure is $p = 1.01 \times 10^5 - 10000z\ (Pa)$.

The CMG STARS uses Newton's method and an implicit time integrator to solve a problem of fluid flow and heat transfer in porous and fractured media at each time step. The convergence criterion for the solver is set equal to 1×10^{-6} in order to constrain a mass balance error to less than 0.001% of the injected water into the target area. Besides, the grid block dimensions are halved in all directions in a grid refinement study, and results show that a less than 0.2% variance is achieved. Therefore, the original grid dimensions are considered to be sufficient in this study.

3.2. Model evaluation

Because the Qiabuqia geothermal field just started to be developed and has not been processed in producing, the real production data is not available. Therefore, the feasibility of our model is determined by comparing it with Lei's model. The average production temperature of the production wells during a 20-year operating cycle is compared with the result of Lei et al.'s model [12], which is shown in Fig. 3. The parameters including reservoir properties and operating parameters in our model are set the same as in Lei et al.'s model, except for the fracture properties. In Lei et al.'s model, an average permeability of 100 mD was substituted for the five hydraulic fractures, which was computed by the permeability of the original granite and the hydraulic fractures. In our model, the five fractures are generated and the original reservoir permeability is unchanged. The result shows two stages in production temperature and geothermal electricity during an operating cycle, a stable stage and a declining stage, which is consistent with Lei et al.'s result. The errors of production temperature and geothermal electricity between our model and their model are all only about 2%. Accordingly, our model is acceptable for predicting the development of the Qiabuqia geothermal field.

3.3. Feature selection

An accurate determination of the input features could highly improve the performance of a machine learning model [70]. In this study, a sensitivity analysis is operated on the base model and six operational parameters are tested in order to determine the input features. These factors include an injection rate and injection temperature of the circulating water, the number and half-length of hydraulic fractures, the well spacing and the production pressure. The parameters setting for the sensitivity analysis is shown in Table 4. The geothermal electricity generated in twenty years $W_e(W)$ is calculated, and compared as the productivity, which can be expressed by Eq. (4) [71]:

Table 4
Parameters setting for the sensitivity analysis.

Parameter	Value
Well spacing, (m)	200, 300, 400, 500, 600
Injection rate, (kg /s)	30, 40, 50, 60, 70
Injection temperature, (°C)	40, 50, 60, 70, 80
Hydraulic fracture half-length, (m)	50, 150, 250, 350, 450
Hydraulic fracture number	1, 2, 3, 4, 5
Production pressure, (MPa)	35, 36, 37, 38, 39

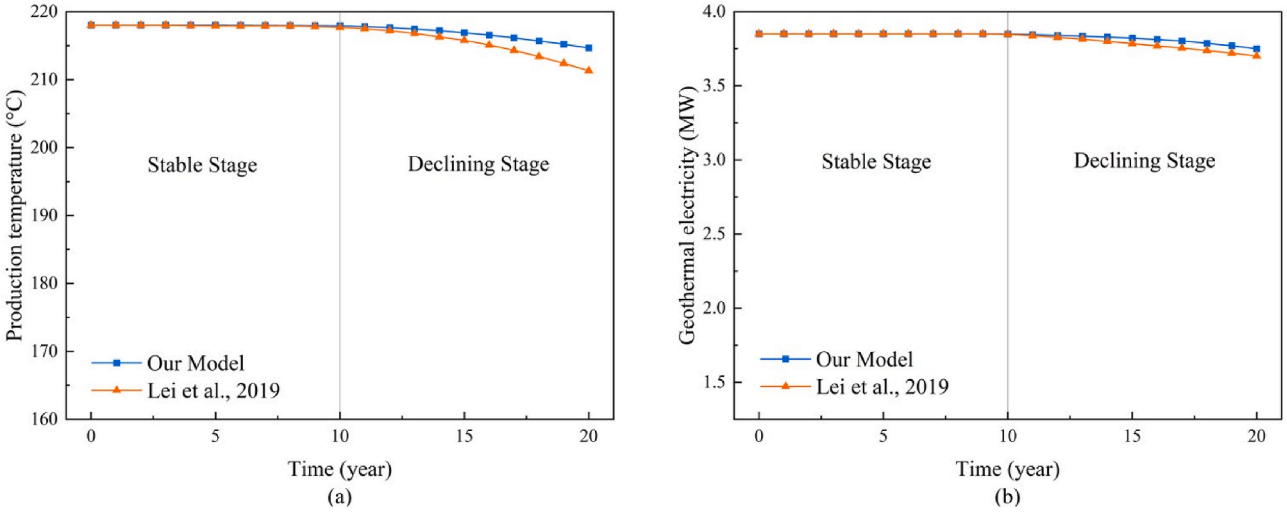


Fig. 3. Model Evaluation: (a) production temperature; (b) geothermal electricity.

$$W_e = 0.45Q\Delta H\left(1 - T_{rej} \,/\, T_{out}\right) \tag{4}$$

where $Q(\mathrm{kg/s})$ is a water production rate in the whole system, $\Delta H(\mathrm{J/kg})$ is an enthalpy change between the injected enthalpy and produced enthalpy, $T_{rej}(\mathrm{K})$ is the rejection temperature that is set at 277.25 *K* due to the mean annual temperature in the Gonghe Basin in 275.55–277.25 K [12], and $T_{out}(\mathrm{K})$ is the average temperature of the production wells.

The results of the sensitivity analysis are given in Fig. 4. This figure indicates that five parameters, the hydraulic fractures number, hydraulic fractures half-length, injection rate, injection temperature, and well spacing, influence the generated geothermal electricity in different ways, while the effect of production pressure on geothermal electricity is insignificant. Higher geothermal electricity occurs in scenarios with a larger well spacing or a lower injection temperature. This is because a difference between the injection temperature and the production temperature of the circulating water is increased, leading to the enhancement of an enthalpy change between the injected enthalpy and produced enthalpy. A longer fracture half-length results in higher geothermal electricity since a bigger area with high conductivity is created in the formation. An optimal injection rate leads to the highest geothermal electricity because of a lower water production rate in the case of a lower injection rate, and the thermal breakthrough and a temperature drop in

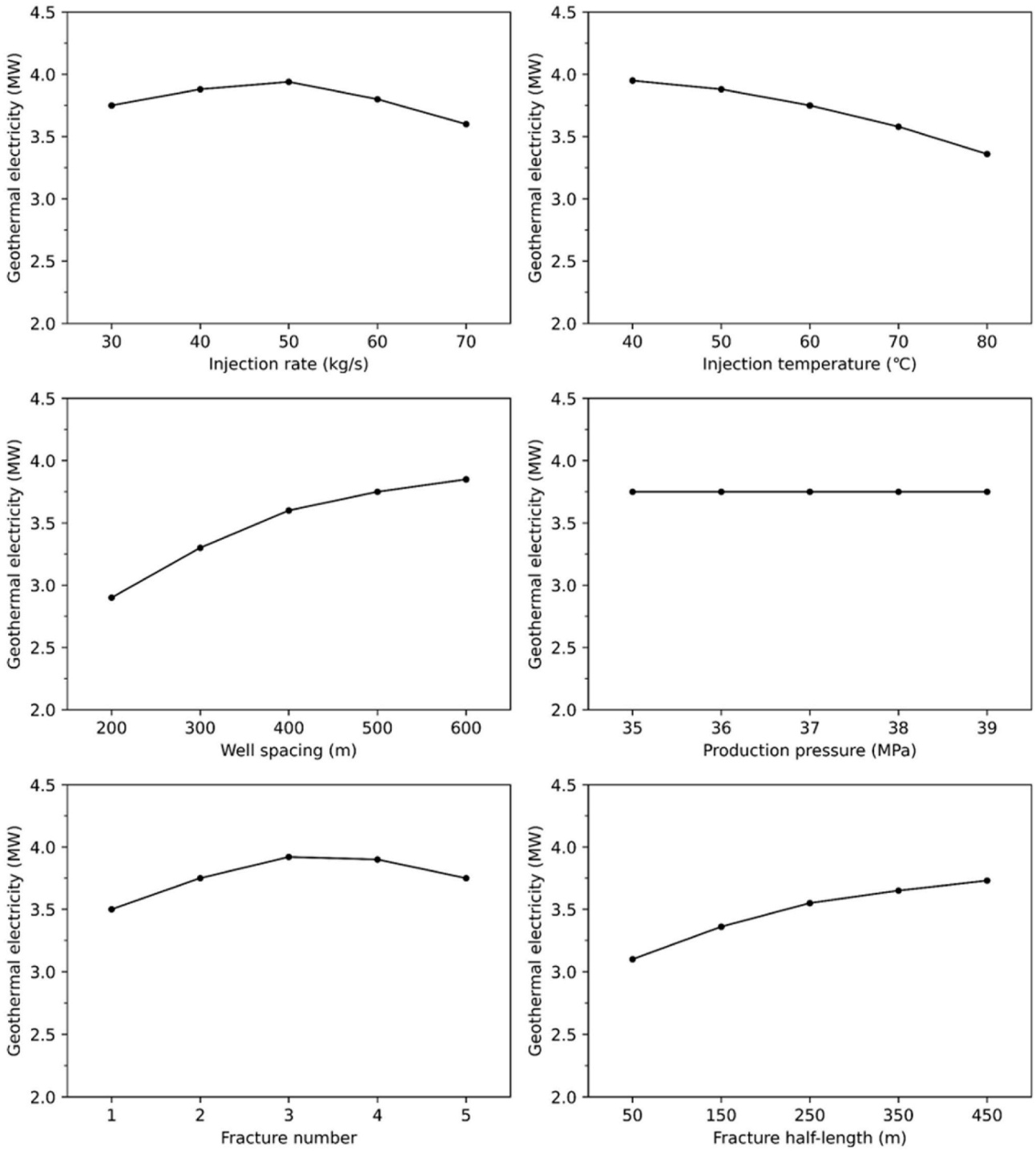


Fig. 4. Effects of operational parameters on geothermal production.

a higher injection rate system. This is also interpreted in fracture numbers. A case with more hydraulic fractures holds a faster flow rate which can provide higher production and thus the electricity production is higher. However, when the number of hydraulic fractures exceeds a critical value, a faster flow rate can encounter thermal breakthrough and the production water temperature can rapidly decline. Consequently, the difference between injection and production temperature is decreased and the resulting electricity production is lower. Therefore, the geothermal electricity increases with the fracture number from 1 to 3 and then decreases with the fracture number from 3 to 5. The little influence of the production pressure can be explained by that the relative pressure difference under a stable water injection rate will not change with different production pressure. Consequently, these parameters, except for the production pressure, are considered the input features for the subsequent data-driven models.

3.4. Data generation

Data samples are collected from simulation models under different configurations of input features. Before developing these numerical models, the ranges of input features should be determined. Two significant criteria should be taken into consideration. The first criterion is that the injection pressure at the bottom of a well should be less than the minimum principal stress to prevent slippage from occurring [12], and the data points with the injection pressure higher than 60 MPa are hence removed since the minimum principal stress in the Qiabuqia geothermal field is 60–72 MPa [12]. The other criterion is that the low-temperature circulating water is undesirable as it may cause scaling and chemical deposition [12]. Therefore, the lowest injection temperature of the circulating water is set to 40 °C. Besides, the hydraulic fracture number is determined to be 5 at most due to the height of a hydraulic fracture (100 m) and the range of the EGS section in this field (3200–3700 m). The hydraulic fracture half-length is not larger than the well spacing. The data section ranges of these parameters are provided in Table 5.

CMG CMOST software is responsible for collecting data samples by developing numerical models. The workflow of creating data samples is shown in Fig. 5. Specifically, based on the base numerical model (Fig. 2), CMOST randomly changes the values of five input features by giving the determined ranges and intervals of these five input features to create new numerical models, and meanwhile, new models are generated with different configurations of these five input features to represent various working schemes. Consequently, 2416 numerical models with 2416 kinds of input features configurations are constructed and their corresponding geothermal electricity is calculated. Therefore, 2416 data samples are collected from these numerical models, where data sample are composed of different sets of input features values and their corresponding geothermal electricity values.

In conclusion, two data groups, Group A with 1510 data points and Group B with 755 data points, are established to investigate the machine learning model performance on different sizes of datasets. In order to evaluate the generalized ability of data-driven models, data samples in each group are randomly split into two parts, with 80% training samples and 20% testing samples. The testing samples are isolated during a training process and can be treated as the new data for evaluation. Their distributions are shown in Table 6.

Table 5
Statistic values of the operational parameters.

	Min Value	Max Value	Interval
Injection rate/q_i, (kg /s)	20	100	2
Injection temperature/T_i, (°C)	40	80	2
Hydraulic fracture half-length/d_f, (m)	50	600	50
Hydraulic fracture number/N_f	1	5	1
Well spacing/D, (m)	200	600	50

4. Data-driven models

4.1. Feature engineering

The dimensions of the input features are crucial to the model performance. A feature with a high dimension is more likely to dominate predictions and decrease their precision [64]. The distributions of different input features in Groups A and B are shown in Fig. 6 and Fig. 7. Each group of the input features has quite different dimensions which vary from single digits up to hundreds. To eliminate the dimensions variance, a standardization process is implemented in the data samples of each group. This method standardizes the input features of training samples by calculating the mean and standard deviation values of the feature variables to center them at 0 and scale 1 [72]. The calculated mean and standard deviation values are stored to be used for transforming the testing samples. The standardization can be written in Eq. (5) [72]:

$$x^* = \frac{x - \mu}{\sigma} \tag{5}$$

where x is the value of a data variable; μ is its mean; σ is its standard deviation.

4.2. Machine learning algorithms

For a regression problem, there are different kinds of machine learning algorithms that can be applied (e.g., linear algorithms, *K*-Nearest Neighbors (KNN), Supper Vector Machine (SVM), Tree-based algorithms and Neural Networks). Generally, these algorithms perform different prediction performance in solving a same problem since different principles are performed. Therefore, the optimal algorithm selection is important for the operators to apply machine learning in predicting geothermal production. Based on our experiments, linear algorithms are hard to study the complex relationships between various operational parameters and geothermal production and cannot demonstrate an accurate prediction eventually. Therefore, we didn't select linear algorithms in this study. In addition, different tree-based algorithms performed similarly in predicting geothermal electricity. Thus, we selected Extreme Gradient Boosting (XGBoost) due to its best estimation performance among all the tree-based algorithms. Consequently, four different kinds of machine learning algorithms are operated to predict geothermal production in this work, including *K*-Nearest Neighbors (KNN), Supper Vector Machine (SVM), Extreme Gradient Boosting (XGBoost) and Artifical Neural Networks.

4.2.1. K-nearest neighbors (KNN)

K-Nearest Neighbors (KNN) is one of the instance-based learning algorithms, and is known as a nonparametric method [73]. The "nonparametric" word means that this method completely relies on data and its internal structure instead of parameters fitting [74]. In a training process, a distance function is used to determine a group of K samples that are nearest to a selected sample and minimizes the error between them. A choice of K is critical for the model performance and an optimum K value should be selected from a group of K values. Weights according to the distance between the selected sample and its neighbors can be also developed to improve the performance. A generalized distance function is the Minkowski distance, which is written in Eq. (6) [75]:

$$d_p(\bar{x}_1, \bar{x}_2) = \left(\sum_{j=1}^{N} \left| x_1^{(j)} - x_2^{(j)} \right|^p \right)^{\frac{1}{p}} \tag{6}$$

where $\bar{x}_1$ and $\bar{x}_2$ are data points in a dataset $X = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n\}, \bar{x}_i \in \mathbb{R}^N$; when $p = 1$, d_p represents the Manhattan distance; when $p = 2$, d_p represents the classical Euclidean distance.

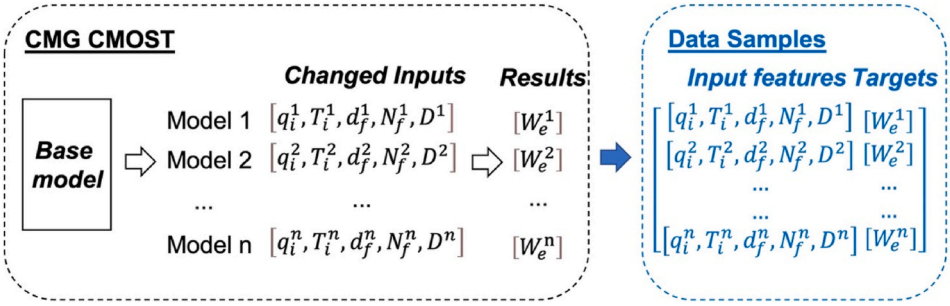


Fig. 5. Workflow of simulation models generation and data collection.

Table 6
Distribution of data points in different groups.

Group	Data number	Training samples number	Testing samples number
A	1510	1208	302
B	755	604	151

4.2.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the supervised learning algorithms, and support vector regression (SVR) based on SVM features is developed by Drucker in 1997 for regression analysis [76]. In SVR, linear regression is formulated according to high-dimensional input data, which can represent a nonlinear relationship between the inputs and the outputs [77]. This linear regression function can be written in Eq. (7) [78,79]:

$$f(x) = \omega^T g(x) + b \tag{7}$$

where $f(x)$ is the predicted output; $g(x)$ represents the nonlinear relationship between the inputs and outputs; ω and b are two adjustable parameters.

The prediction performance of SVR is measured by a loss function, and Vapnik's ε-insensitive loss function is used [80]. It can consider the maximum variance between predictions and actual values, which can be written in Eqs. (8) and (9) [78,79]:

$$\min_{\omega,b,\varepsilon} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{M}(\varepsilon_i + \varepsilon_i^*) \tag{8}$$

$$\text{subject to} \begin{cases} y_i - f(x_i) \le \varepsilon + \varepsilon_i^* \\ f(x_i) - y_i \le \varepsilon + \varepsilon_i \\ \varepsilon_i, \varepsilon_i^* \ge 0, i = 1, \ldots, M \end{cases} \tag{9}$$

where C and ε are two positive parameters that are defined by the user; ε_i and ε_i^* are used to consider soft margins when all data points cannot be mapped through a specific margin.

4.2.3. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an ensemble learning algorithm, which is based on the Gradient Boosting Decision Tree (GBDT) algorithm [81,82]. This algorithm trains data samples through a set of "weak" learners and a "strong" learner to minimize an objection function. During a training process, a "weak" learner is added to a decision tree iteratively for higher accuracy [83], and the results of each tree are accumulated according to weights and a residual, which can be written in Eqs. (10) and (11) [84]. Specifically, the modified weights and the residual of the previous trees are transformed into a new tree for training until a reasonable result is obtained [84].

$$\hat{y}_i^t = \varphi(x_i) = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{t-1} + f_t(x_i), f_k \varepsilon F \tag{10}$$

with

$$F = \{f(x) = \omega_{q(x)}\}, (q : R^m \to T, \omega \varepsilon R^T) \tag{11}$$

where $\hat{y}_i^t$ is a predicted value of a model; f_k is an output value of an independent tree; F is the space of an independent tree; q is the structure of a tree; T is the number of leaf nodes of a tree; ω is a weight of a leaf node, and each f_k corresponds to q and ω..

Compared with GBDT, XGBoost can reduce the overfitting due to the application of a regularization term in an objection function, which can be written in Eqs. (12) and (13) [84]:

$$L^t = \sum_{i}^{n} l(y_i, \hat{y}_i^t) + \sum_{k}^{t} \Omega(f_k) \tag{12}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \tag{13}$$

where l is a loss function used to calculate the difference between the actual target value y_i and the model output $\hat{y}_i^t$; $\Omega(f_k)$ is the regularization term; γ is the minimum loss reduction; λ is a penalty coefficient.

In addition, XGBoost can improve the modification performance of the gradient algorithm because it uses the second partial derivative of the loss function while other gradient boosting algorithms essentially use the partial derivative. Therefore, Eq. (12) can be changed to Eq. (14) [84]:

$$L^t = \sum_{i}^{n}\left[l\left(y_i, \hat{y}_i^{t-1}\right) + \frac{g_i}{1!}f_t(x_i) + \frac{h_i}{2!}f_t(x_i)^2\right] + \Omega(f_i) + \text{constant} \tag{14}$$

where $g_i = \hat{\partial}_{\hat{y}_i^{t-1}} l(y_i, \hat{y}_i^t)$ and $h_i = \hat{\partial}^2_{\hat{y}_i^{t-1}} l(y_i, \hat{y}_i^t)$.

4.2.4. Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) is a computational structure inspired by biological neural networks [85]. In an ANN model, its neurons connect an input layer to an output layer [86]. A structure of the neurons and the connections between the neurons and different layers are critical for the operation of an ANN [74]. A basic neuron unit is connected with the input layer through n input channels, and each of the input features is characterized by a corresponding weight W_i, which is shown in Fig. 8. When an input is transformed into the neurons, they are calculated by their weights and an optional bias (b), and the result will be summed. The final output is produced after the summed result is filtered by an activation function f_a.

According to the number and arrangement of neurons, an ANN model can be divided into Single Layer Perception (SLP) and Multilayer Perception (MLP) [87]. The structure of MLP with a single hidden layer shows in Fig. 9. In the MLP, the results obtained from the input layer are not the final output. They are treated as another input data and split by the corresponding weight (h_{jk}^2) into the next layer for calculations until the operation completes in the final output layer. The flow proceeding from the first layer to the output layer in the same direction is called a

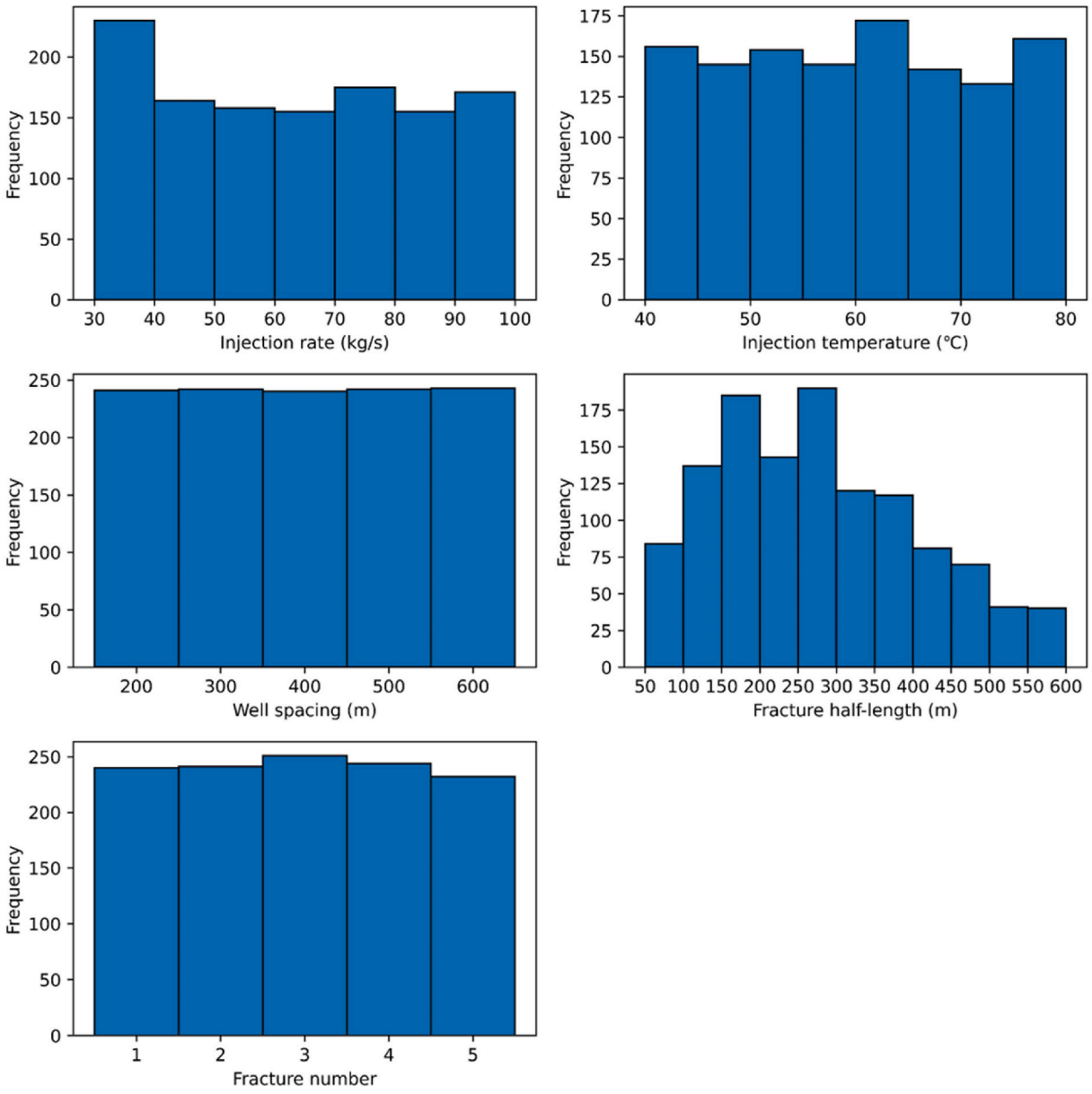


Fig. 6. Distributions of the input features values in Group A.

feed-forward process [73].

The purpose of an ANN is to minimize the error between predicted outputs and actual targets. To achieve it, another flow called Back-propagation (BP) process is operated when a feed-forward process finishes. BP is a gradient-based minimization process that takes a step from the output layer to the input layer [88]. During a BP process, weights are changed by a small amount from the output layer and are used for calculating the weights of previous layers [89,90]. After the weights update to the input layer, another feed-forward process will operate until the minimum overall error is reached.

4.3. Evaluation matrix

To evaluate the performance of data-driven models, three statistical indicators are utilized for optimal configuration. There are a root mean squared error (*RMSE*), a mean absolute error (*MAE*) and a coefficient of determination (R^2). *RMSE* and *MAE* can describe a variance between predictions and true target values. The absolute error is estimated by *MAE* while the error is magnified by *RMSE* since a square function is applied. R^2 can indicate the fit accuracy and measure the prediction performance through the proportion of an explained variance, which varies from 0 to 1. A model with a higher R^2 value and lower *RMSE* and *MAE* values symbolizes better performance. *RMSE*, *MAE* and R^2 are written in Eqs. 15–17 [74]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \tag{15}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \tag{16}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \tag{17}$$

where y_i is the prediction of a sample; $\widehat{y_i}$ is the corresponding actual value; $\overline{y_i}$ is the mean of the actual value of a target; n is the number of the

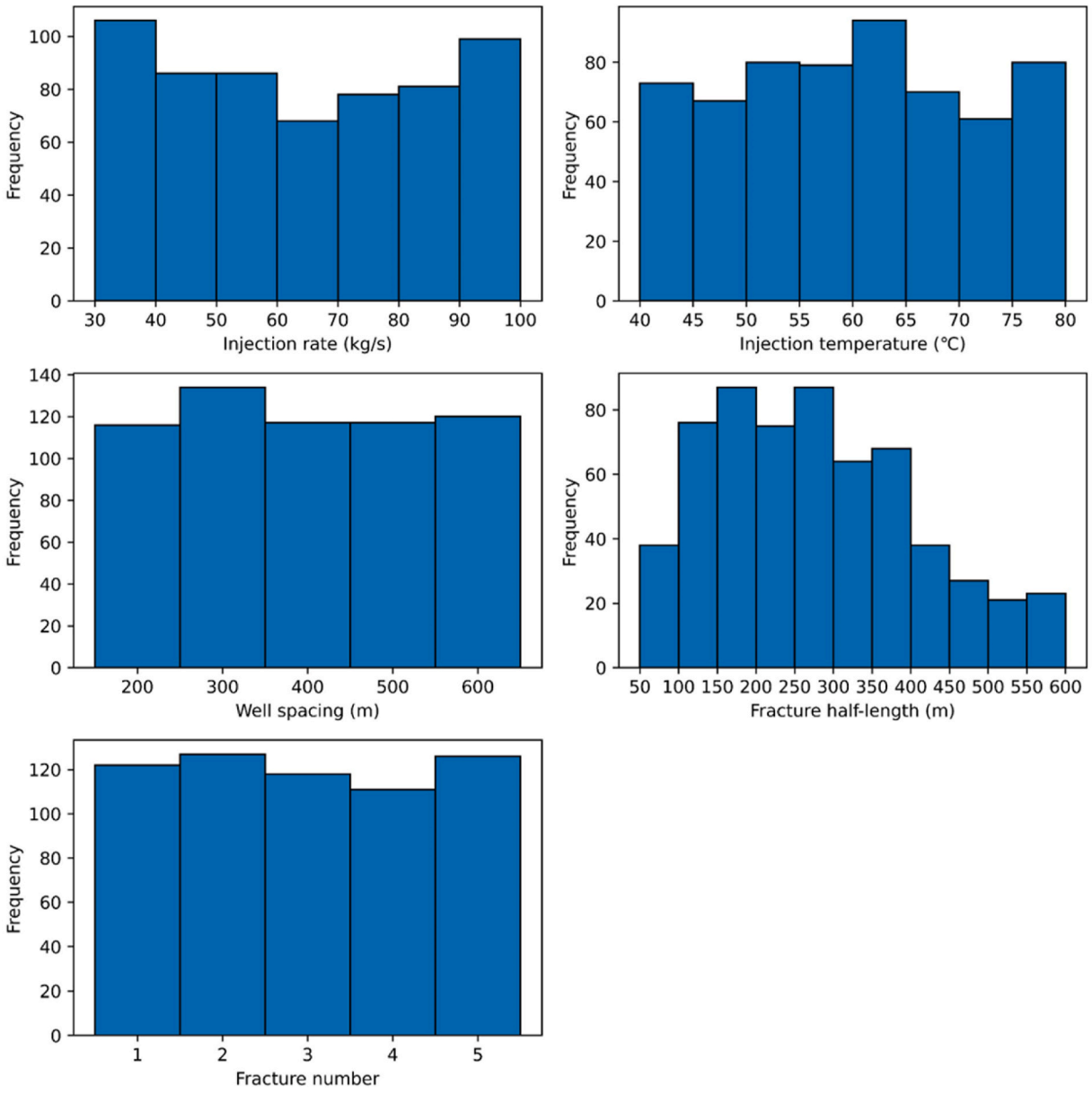


Fig. 7. Distributions of the input features values in Group B.

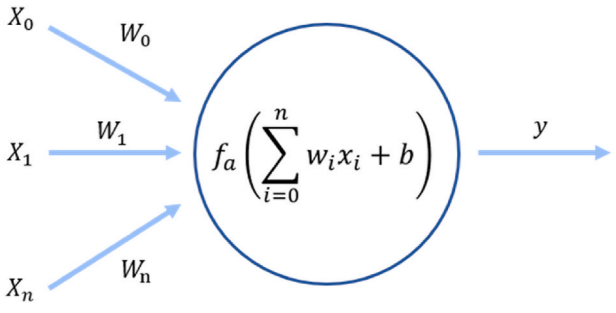


Fig. 8. Structure of a neuron.

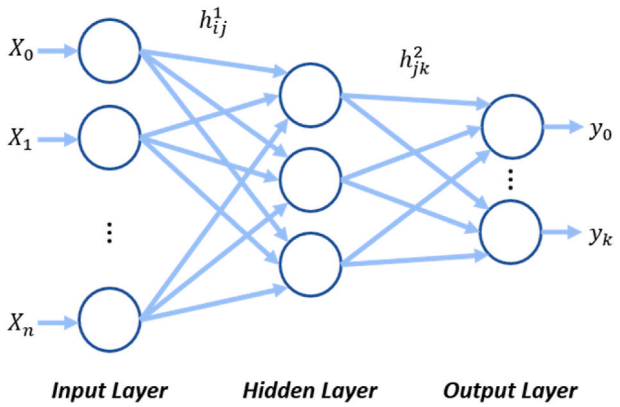


Fig. 9. Structure of an MLP with one hidden layer.

samples.

4.4. Model design and tuning

Before the construction of machine learning models, the hyper-

parameters that dominate the model performance should be determined. Previous studies showed that the performance of a KNN model can be influenced by several hyperparameters including a number of neighbors (*N*), *weights* and the *algorithm* [91–93]. The *algorithm* is to determine the method of computing the nearest neighbors where *'uniform'* means that each neighbor is weighted equally and *'distance'* means that the points are weighted by the inverse of their distance [94]. The performance of the SVM is affected by parameter *C*, *epsilon* and *gamma* [95–98]. There are two options in *gamma*. It is calculated by $1/(no.features)$ under *'auto'* and equal to $1/(no.features \times variance\ of\ samples)$ under *'scale'* [94]. For an XGBoost model, a learning rate (*LR*), the maximum tree depth (*MTD*) that constrains the complexity of the model, the regularization term *alpha*, and the number of estimators (*NE*) can highly affect its accuracy [34,65,95,99–101]. Besides, parameters *subsample* and *colsample_bytree (CST)* can be utilized to prevent overfitting [34,101]. For an ANN model, a value of 1500 is used for the maximum iteration number to save simulation time. The *'ReLU'* function is selected as the activation function and the *'LBFGS'* is determined as the optimizer algorithm. *'LBFGS'* is an optimizer based on Quasi-Newton methods. This method can provide faster calculations and has been proven a successful method in optimization [102–104]. The performance of an ANN model is affected by a learning rate and the structure of hidden layers [105–107]. The *learning rate* is scheduled for the weights updates and its value is set in a range from 0.0001 to 1 in 10 times intervals. The structure of a hidden layer contains a number of hidden layers and neurons in each layer. It is operated from simple hidden layers to complex hidden layers. For example, the hidden layer (4) represents one hidden layer with 4 units, and the hidden layer (4, 8, 4) means three hidden layers with 4, 8 and 4 units in each layer separately. Therefore, these hyperparameters are optimized for determining an optimal model. The details of hyperparameters and their ranges are summarized in Table 7.

After determining the sensitive hyperparameters, four base data-driven models based on the KNN, SVM, XGBoost and ANN are developed by giving the default values of these hyperparameters. Subsequently, a *K*-fold cross-validation method and a grid search method are used to determine the optimal hyperparameter configurations of these data-driven models. The *K*-fold cross-validation method is operated on training samples to fully utilize their data points and split them into a training set and a validation set. The grid search method is to generate different models by changing the hyperparameter values within their corresponding range iterately for training data sampels in the training set, and the model with the best performance in the validating set can be determined as the optimal data-driven model. The workflow of the model tuning is shown in Fig. 10.

Table 7
Hyperparameters setting of KNN, SVM and XGBoost models [94].

Hyperparameter	Description	Range	Optimizd value	
			Group A	Group B
KNN				
N	The number of neighbors	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	6	5
weights	The approach to evaluate the distance of the neighbors	*'uniform'*, *'distance'*	*'distance'*	*'distance'*
algorithm	The method to compute the nearest neighbors	*'ball_tree'*, *'kd_tree'*	*'ball_tree'*	*'ball_tree'*
SVM				
C	The regularization parameter and the strength of the regularization is inversely proportional to *C*	5, 5e2, 5e3, 5e4, 5e5, 5e6	5e6	5e6
gamma	The kernel coefficient	*'scale'*, *'auto'*	*'auto'*	*'auto'*
epsilon	The penalty for the bad predictions	0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1	1	1
XGBoost				
NE	The number of the gradient boosted trees	200, 400, 600, 800, 1000	1000	1000
MTD	The maximum depth for base learners	2, 3, 4, 5, 6, 7, 8, 9, 10	4	4
LR	The step size in updates	0.01, 0.05, 0.1, 0.15, 0.2	0.05	0.05
subsample	The subsample ratio of the training instances	0.3, 0.4, 0.5, 0.6, 0.7, 0.8	0.3	0.3
CST	The subsample ratio of columns for each tree	0.6, 0.7, 0.8,0.9	0.8	0.8
alpha	The L1 regularization term on weights	0, 0.01, 0.05, 0.1, 0.15, 0.2	0.01	0.2
ANN				
learning rate	The step size in updates	0.0001, 0.001, 0.01, 0.1, 1	0.0001	0.0001
hidden layers	The number of hidden layers and units of each layer	(4), (8), …, (4, 4), (4, 8) …, (4, 4, 4), (4, 8, 4) … (10, 10, 10)	(10, 6, 6)	(10, 10, 2)

5. Results and discussion

In this study, four kinds of data-driven models based on KNN, SVM, XGBoost and ANN are conducted to predict the geothermal electricity generated in twenty years of the Qiabuqia field. The input features, containing the injection rate and the injection temperature of the circulating water, the number and half-length of the hydraulic fractures and the well spacing, are determined by the sensitivity analysis in the base numerical model. The data samples are collected through a group of numerical models. To comprehensively investigate the model performance from the prediction accuracy, stability and adaptability, the data samples are assigned to two data groups with 1510 data points and 775 data points separately, and the data points in each group are divided into a training set, a validation set and a testing set. The values of three statistical indicators for all the models are given in Table 8 for both sets and groups.

5.1. The model comparison on the training samples

The training samples of each group are divided into a training set and a validation set. A model with the optimal hyperparameters settings represents that the model performances in both sets are the best. For different data groups, parameters *C*, *epsilon* and *gamma* in the optimal SVM models are the same, which are 5e6, *'auto'* and 1, respectively. In the KNN models, the best *weights* are set to *'distance'* and the *'ball-tree'* algorithm is used in both groups, while *N* is 6 for Group A and 5 for Group B. For the XGBoost, five parameters are set to have the same values in different groups, including *N* with a value of 1,000, *MDT* with 4, *LR* with 0.05, *subsamples* with 0.3 and *CST* with 0.8, while *alpha* is 0.01 and 0.2 in Groups A and B separately. For the ANN models, a *learning rate* of 0.0001 and a total of 22 neurons in hidden layers present the best performance, while the structure of a hidden layer in Group A has three layers with 10, 6 and 6 neurons and three layers with 10, 10 and 2 neurons in Group B, respectively. The details of optimized hyperparameters are shown in Table 7.

Figs. 11–13 depict the results of R^2, *MAE* and *RMSE* for the models of the training set and validation set in different groups. The KNN models demonstrate the worst prediction performance, since their R^2, *MAE* and *RMSE* values in the validation sets of all the groups are the worst

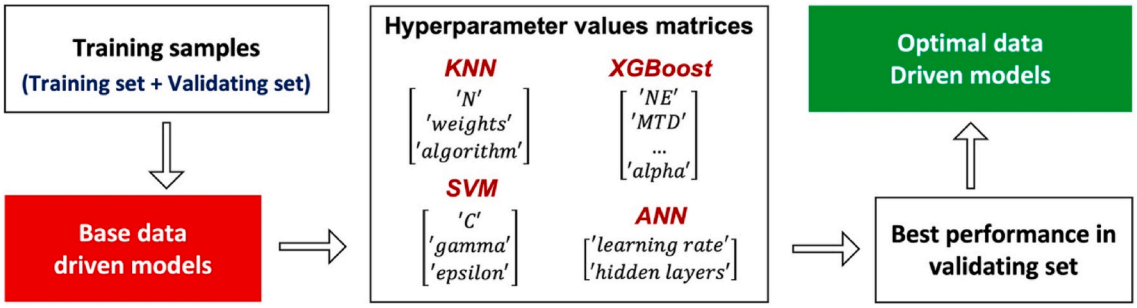


Fig. 10. Workflow of the model tuning.

Table 8
Statistical indicators values of data-driven models for predicting geothermal electricity in different data sets and different data groups.

Model	Training set			Validation set			Testing set		
	R^2	$MAE(W)$	$RMSE(W)$	R^2	$MAE(W)$	$RMSE(W)$	R^2	$MAE(W)$	$RMSE(W)$
Group A									
KNN	1	0	0	0.964	379,967	496,005	0.965	373,695	496,128
XGBoost	0.999	23,338	29,697	0.998	66,959	99,291	0.997	73,982	97,871
SVM	0.998	26,041	60,796	0.998	43,039	77,896	0.998	43,747	71,445
ANN	0.998	50,989	68,684	0.998	56,748	77,528	0.998	56,029	75,079
Group B									
KNN	1	0	0	0.944	472,766	605,408	0.948	479,207	629,523
XGBoost	0.999	17,737	22,470	0.996	98,595	134,392	0.996	95,481	128,228
SVM	0.998	34,318	74,865	0.997	79,236	128,902	0.996	83,921	145,890
ANN	0.998	48,178	63,803	0.998	64,726	85,626	0.998	62,089	86,175

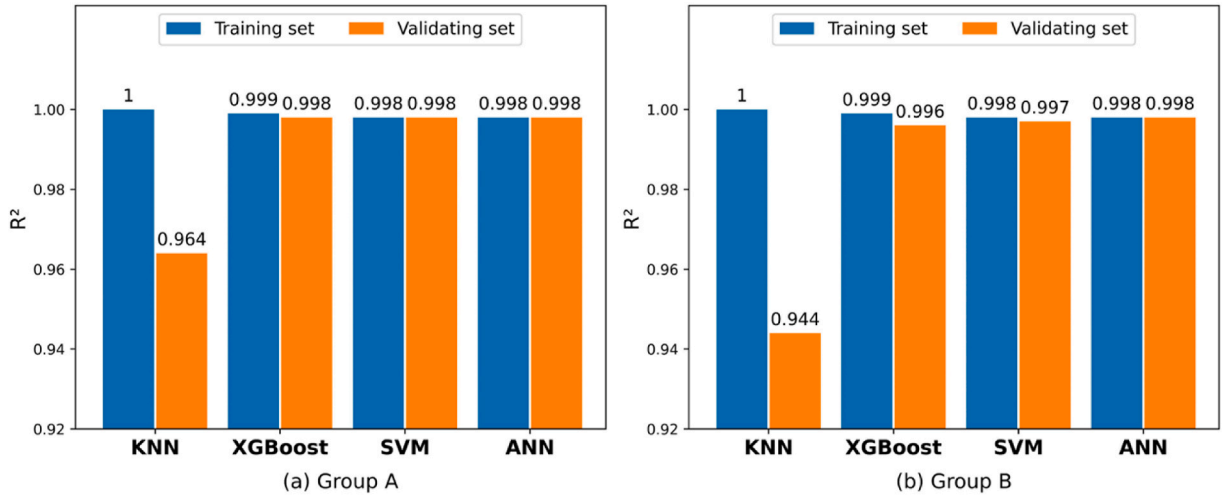


Fig. 11. R^2 values of different models on the training samples in different data groups (a) Group A with 1510 data points and (b) Group B with 755 data points.

although those values are the best in a training set. Compared with the SVM and ANN models, the XGBoost models demonstrate a better performance in the training sets of both groups with a higher R^2 and lower *MAE* and *RMSE*. However, these indicators are worse in the validation sets. Therefore, the performance of SVM and ANN are better than the XGBoost. The SVM and ANN models show similar performance in Group A. Specifically, they provide the same R^2 values of 0.998 in both the training and validation sets. Besides, their *RMSE* values are similar where the SVM model holds 60,796 W in the training set and 77,896 W in the validation set while the ANN model holds 68,684 W and 77,528 W separately. In Group B, the ANN performs better than the SVM, although their R^2 values are similar and the *MAE* values of the SVM are lower than the ANN in both sets. It is because the ANN model shows a much greater extent lower *RMSE* values than the SVM model.

On the other hand, the performance variance between the training set and validating set can show the stability of a machine learning model to investigate if an overfit or underfit problem happens. From the results, the contrasts of average increases in *MAE* and *RMSE* between the training sets and the validation sets of the two groups reveal that the ANN algorithm is the most stable. Specifically, an average of 25.6% increase in *MAE* and 18.5% increase in *RMSE* are the minimum among the algorithms studied, followed by the SVM (48.1% in *MAE* and 32.0% in *RMSE*), the XGBoost (73.6% in *MAE* and 76.7% in *RMSE*) and the KNN (100% in *MAE* and *RMSE*). It can also infer that the overfitting problem occurs in the KNN model and the stability is bad in the XGBoost model. However, it cannot be determined that the overfitting appears in the XGBoost model since a higher than 0.996 R^2 is calculated in both groups. Generally, the stability of different models can be ranked as: ANN > SVM > XGBoost > KNN.

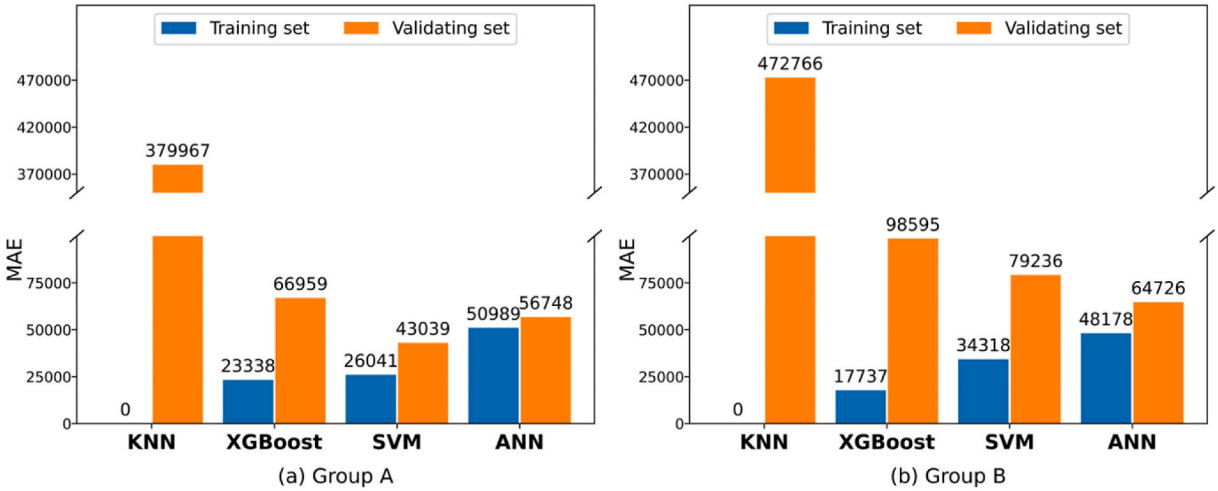


Fig. 12. *MAE* values of different models on the training samples in different data groups (a) Group A with 1510 data points and (b) Group B with 755 data points.

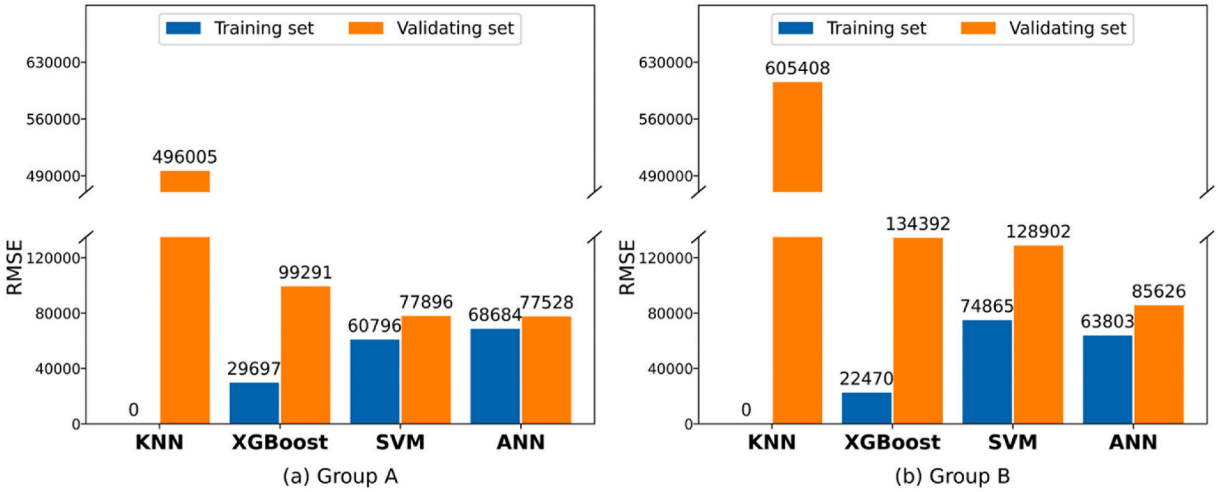


Fig. 13. *RMSE* values of different models on the training samples in different data groups (a) Group A with 1510 data points and (b) Group B with 755 data points.

5.2. The model comparisons on the testing samples

The results of three statistical indicators on the testing samples and the scatter diagrams of the predicted values and true target values in different groups are shown in Figs. 14 and 15. The dot properties in these figures are designed according to the errors between the predictions and true targets. Specifically, the size of each scatter increases and the color deepens with the enlargement of the corresponding error. Therefore, the dots with smaller sizes and lighter colors represent better performance in their corresponding models and vice versa.

For Group A, the SVM and ANN models are better in prediction accuracy than the XGBoost and KNN models due to their smaller *MAE* and *RMSE* values. Compared with the SVM model, the ANN model is comparable but has slightly weaker overall performance according to its higher *RMSE* and *MAE* (ANN: $MAE = 56{,}029$ W, $RMSE = 75{,}079$ W; SVM: $MAE = 43{,}747$ W, $RMSE = 71{,}445$ W). However, the predicted value with a high variance is less likely to occur in the ANN model. It is concluded from the scatter plots. There are fewer large dots exhibited in the ANN model but more can be observed in the SVM model. This also explains that the variance between the *MAE* and *RMSE* in the SVM model is larger than that in the ANN model. The relatively large errors of these outliers increase *RMSE* through the error square function. Therefore, it can be deduced that the most of predictions of the SVM model are close to the true targets but some prediction outliers are unavoidable. According to the scores of the statistical indicators and the scatter plots, the generalization ability in Group A can be ranked as: SVM $\geq$ ANN $>$ XGBoost $>$ KNN.

For Group B, the ANN has the best generalization ability according to the lowest *MAE* and *RMSE* and the highest R^2. Besides, the number of deep-coloured and big-sized dots in the ANN plot is the least. On the other hand, the KNN is the opposite. The predictive ability of the SVM and XGBoost models is comparable due to the same R^2 values and similar *MAE* and *RMSE* values. However, some larger dots are illustrated in the SVM plot compared to the XGBoost plot, demonstrating that the prediction outliers of the former are worse. Therefore, the generalization ability in Group B can be ranked as: ANN $>$ XGBoost $\geq$ SVM $>$ KNN.

Comparing the performance in different groups, the adaptability of a machine learning model can be determined. In the results, all the models produce a better predictive ability when the data amounts increase. The variances in *MAE* and *RMSE* between different data groups can prove that the ANN has better adaptability than other algorithms. In the ANN, the variance in *MAE* is 9.8% and that in *RMSE* is 12.9%, while they are 47.9% and 51.0% in the SVM, 22.5% and 23.7% in the XGBoost, and 22.0% and 21.2% in the KNN, respectively. Therefore, the adaptability can be ranked as: ANN $>$ KNN $>$ XGBoost $>$ SVM.

Previous studies chose a machine learning algorithm to predict geothermal development but did not demonstrate the reason for its selection [52–56]. In this study, by combining the ranks of prediction

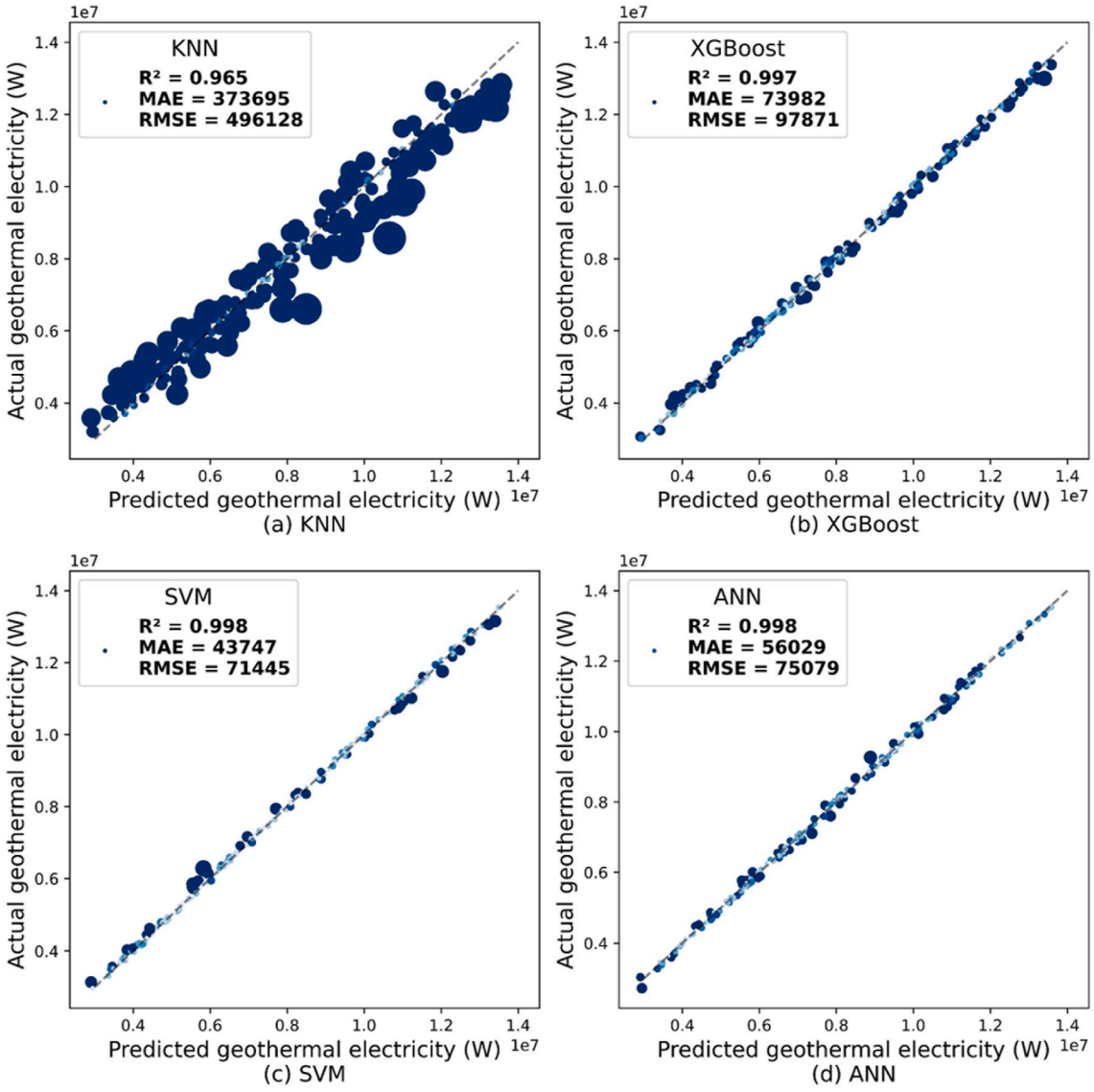


Fig. 14. Generalization ability of different models in Group A (1510 data points) (a) the KNN model, (b) the XGBoost model, (c) the SVM model, and (d) the ANN model.

ability, stability and adaptability, the ANN algorithm is the best choice for predicting geothermal electricity due to its best predictive ability from the highest R^2 value of 0.998, the best stability from the minimum performance variance between different sets, and the best adaptability from the comparison between various groups. Importantly, the time to predict over 150 data points by the ANN algorithm is less than 20s, which is around 54,000s by numerical simulation, illustrating the enormous time-saving in prediction efficiency. Therefore, the proposed ANN model shows the high potential for substituting numerical simulation to accurately simulate the in-depth performance of a geothermal system, and show its feasibility to facilitate its optimization development for generating a sustainable and economical geothermal system.

6. Conclusions

This paper presents an application of data-driven models that combine numerical simulations and machine learning algorithms for estimating geothermal electricity in the Qiabuqia field in China. The numerical simulations determine the input features and generate data samples. This is followed by the comparisons of four different models (i.e., KNN, SVM, XGBoost and ANN) in three data sets (i.e., training, validation and testing sets) and two data groups (i.e., Group A with 1510 data samples and Group B with 755 data samples).

The results of numerical simulations show that the injection rate and temperature of the circulating water, the number and half-length of fractures, and the well spacing affect the geothermal electricity and are selected as the input features for data-driven models, while the production pressure has little impact on production. In the data-driven models studied, the rank of prediction accuracy in Group A is SVM $\geq$ ANN $>$ XGBoost $>$ KNN, and that in Group B is: ANN $>$ XGBoost $\geq$ SVM $>$ KNN. The stability is ranked as ANN $>$ SVM $>$ XGBoost $>$ KNN, and the adaptability is ranked as ANN $>$ KNN $>$ XGBoost $>$ SVM. Therefore, the KNN is impracticable in predicting geothermal development due to the occurrence of overfitting. The XGBoost and SVM are not highly acceptable since they present undesirable stability and adaptability separately. The ANN is highly recommended in predicting geothermal development because it provides the best stability with the minimum performance variance between a training set and a validating set. Besides, it demonstrates the slightest variance between different sizes of data groups with a promising predictive accuracy, where the R^2 values are all 0.998. Importantly, the ANN-based data-driven model also significantly improves the efficiency of prediction. Less than 20 s of

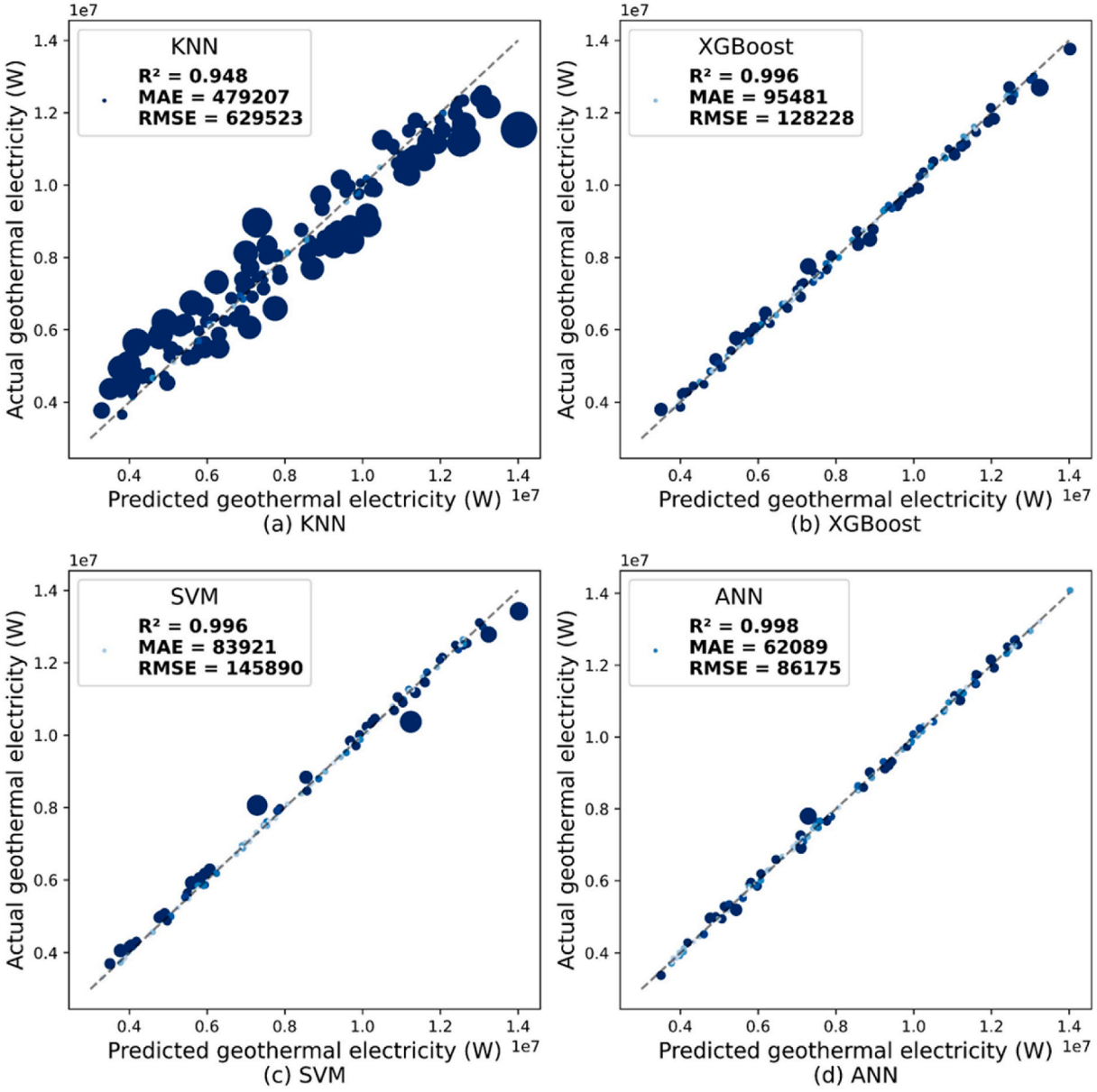


Fig. 15. Generalization ability of different models in Group B (755 data points) (a) the KNN model, (b) the XGBoost model, (c) the SVM model, and (d) the ANN model.

simulation time requires 2700 times less compared to the numerical simulation method which takes about 54,000 s.

The results provide a valuable reference for machine learning applications of studying geothermal systems and can be effectively applied in other renewable energy systems, which can facilitate the prediction of their development. Besides, the proposed data-driven method can be directly used to optimize an energy system or conduct economic analysis, and the optimization of the Qiabuqia geothermal field based on the generated data-driven models can be considered in future research. Although the proposed data-driven method shows great advantages and potential, it could be further improved by considering more reservoir properties for the input features to enhance its application range. Besides, if the Qiabuqia geothermal field starts to produce and the actual production field is available, we will further evaluate and improve the proposed model. In addition, our model is based on one geothermal field, if the field data of other geothermal fields can be utilized, we believe that our model can be improved and accurately apply in predicting geothermal production of various fields.

Credit author statement

Zhenqian Xue: Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing; **Kai Zhang**: Data curation, Writing – review & editing; **Chi Zhang**: Validation, Writing – review & editing; **Haoming Ma**: Investigation, Writing – review & editing; **Zhangxin Chen**: Conceptualization, Writing – review & editing, Supervision.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This research has been made possible by contributions from the Natural Sciences and Engineering Research Council (NSERC)/Energi Simulation Industrial Research Chair in Reservoir Simulation, the Alberta Innovates (iCore) Chair in Reservoir Modeling, and the Energi Simulation/Frank and Sarah Meyer Collaboration Centre.

References

[1] Xue Z, et al. Thermo-economic optimization of an enhanced geothermal system (EGS) based on machine learning and differential evolution algorithms. Fuel 2023:340.
[2] Ma H, et al. Optimized schemes of enhanced shale gas recovery by CO2-N2 mixtures associated with CO2 sequestration. Energy Convers Manag 2022;268:116062.
[3] Zhang S, et al. Well placement optimization for large-scale geothermal energy exploitation considering nature hydro-thermal processes in the Gonghe Basin, China. J Clean Prod 2021;317:128391.
[4] Wang Z, et al. Hydrate deposition prediction model for deep-water gas wells under shut-in conditions. Fuel 2020;275:117944.
[5] Lu S-M. A global review of enhanced geothermal system (EGS). Renew Sustain Energy Rev 2018;81:2902–21.
[6] Wang G, et al. Assessment of geothermal resources in China. In: Proceedings, thirty–eighth workshop on geothermal reservoir engineering. Stanford, California, Febuary: Stanford University; 2013.
[7] Dudley BJBsr. London, UK, accessed Aug. BP statistical review of world energy, vol. 6; 2018, 00116. 2018.
[8] Lund JW, et al. Characteristics, development and utilization of geothermal resources-a Nordic perspective 2008;31(1):140–7.
[9] Duchane DJCG. Status of the United States hot dry rock geothermal technology development program 1994;19(1).
[10] China, T.S.C.I.O.o.t.P.s.R.o. Responding to climate change: China's policies and actions. 2021. Available from: http://www.scio.gov.cn/zfbps/32832/Document/1715506/1715506.htm#:~:text=In%202015%2C%20China%20set%20its,new%20NDC%20targets%20and%20measures.
[11] Lei Z, et al. Electricity generation from a three-horizontal-well enhanced geothermal system in the Qiabuqia geothermal field, China: slickwater fracturing treatments for different reservoir scenarios. Renew Energy 2020;145:65–83.
[12] Lei Z, et al. Exploratory research into the enhanced geothermal system power generation project: the Qiabuqia geothermal field, Northwest China. Renew Energy 2019;139:52–70.
[13] Xu T, et al. Prospects of power generation from an enhanced geothermal system by water circulation through two horizontal wells: a case study in the Gonghe Basin, Qinghai Province, China. Energy 2018;148:196–207.
[14] Gao K, et al. Numerical simulation study of a novel horizontally layered enhanced geothermal system: a case study of the Qiabuqia geothermal area, qinghai province, China. J Therm Sci 2021;30(4):1328–40.
[15] Lee K. Classification of geothermal resources-an engineering approach. Auckland, NZ: Geothermal Institute, The University of Auckland; 1996.
[16] Zhang C, et al. Parametric study of the production performance of an enhanced geothermal system: a case study at the Qiabuqia geothermal area, northeast Tibetan plateau. Renew Energy 2019;132:959–78.
[17] Genter A, et al. Contribution of the exploration of deep crystalline fractured reservoir of Soultz to the knowledge of enhanced geothermal systems (EGS). Compt Rendus Geosci 2010;342(7–8):502–16.
[18] Brown DW, Duchane DV. Scientific progress on the fenton hill HDR project since 1983. Geothermics 1999;28(4):591–601.
[19] Gerber L, Maréchal F. Environomic optimal configurations of geothermal energy conversion systems: application to the future construction of Enhanced Geothermal Systems in Switzerland. Energy 2012;45(1):908–23.
[20] Breede K, et al. A systematic review of enhanced (or engineered) geothermal systems: past, present and future. Geoth Energy 2013;1(1):4.
[21] Zhang W, et al. Study of the enhanced geothermal system (EGS) heat mining from variably fractured hot dry rock under thermal stress. Renew Energy 2019;143:855–71.
[22] Zhou L, et al. Analysis of influencing factors of the production performance of an enhanced geothermal system (EGS) with numerical simulation and artificial neural network (ANN). Energy Build 2019;200:31–46.
[23] Gong F, et al. Evaluation of geothermal energy extraction in Enhanced Geothermal System (EGS) with multiple fracturing horizontal wells (MFHW). Renew Energy 2020;151:1339–51.
[24] Zhang K, Lau HC. Utilization of a high-temperature depleted gas condensate reservoir for CO2 storage and geothermal heat mining: a case study of the Arun gas reservoir in Indonesia. J Clean Prod 2022:343.
[25] Zhang K, Lau HC. Sequestering CO2 as CO2 hydrate in an offshore saline aquifer by reservoir pressure management. Energy 2022;239:122231.
[26] Zhang Y, et al. Reservoir stimulation design and evaluation of heat exploitation of a two-horizontal-well enhanced geothermal system (EGS) in the Zhacang geothermal field, Northwest China. Renew Energy 2022;183:330–50.
[27] Cai Y, Dahi Taleghani A. Using pressure changes in offset wells for interpreting fracture driven interactions (FDI). J Petrol Sci Eng 2022:219.
[28] Song G, et al. Multi-objective optimization of geothermal extraction from the enhanced geothermal system in Qiabuqia geothermal field, Gonghe Basin. Acta Geologica Sinica - English Edition 2021;95(6):1844–56.
[29] Chong Q. On geothermal heat extraction from the basal cambrian sandstone unit in central Alberta, Canada. Schulich School of Engineering; 2021.
[30] Liu G, et al. Impacts of fracture network geometries on numerical simulation and performance prediction of enhanced geothermal systems. Renew Energy 2021;171:492–504.
[31] Zinsalo JM, Lamarche L, Raymond J. Design and optimization of multiple wells layout for electricity generation in a multi-fracture enhanced geothermal system. Sustain Energy Technol Assessments 2021;47.
[32] Dahi Taleghani A, Cai Y, Pouya A. Fracture closure modes during flowback from hydraulic fractures. Int J Numer Anal Methods GeoMech 2020;44(12):1695–704.
[33] Shi Y, Song X, Song G. Productivity prediction of a multilateral-well geothermal system based on a long short-term memory and multi-layer perceptron combinational neural network. Appl Energy 2021:282.
[34] Huang Z, Chen Z. Comparison of different machine learning algorithms for predicting the SAGD production performance. J Petrol Sci Eng 2021:202.
[35] MacInnes J. In: Atkinson P, et al., editors. Exploratory data analysis. London: SAGE Publications Ltd; 2020.
[36] Dong G, Liu H. Feature engineering for machine learning and data analytics. Boca Raton, FL: CRC Press/Taylor & Francis Group First edition; 2018.
[37] Chen J, Jiang F. Designing multi-well layout for enhanced geothermal system to better exploit hot dry rock geothermal energy. Renew Energy 2015;74:37–48.
[38] Sun F, et al. Geothermal energy development by circulating CO2 in a U-shaped closed loop geothermal system. Energy Convers Manag 2018;174:971–82.
[39] Xue Z, Chen Z. Deep learning based production prediction for an enhanced geothermal system (EGS). Day 2 Thu 2023;March 16:2023.
[40] Alpaydin E. Introduction to machine learning. MIT press; 2020.
[41] Keçebaş A, Yabanova İ. Thermal monitoring and optimization of geothermal district heating systems using artificial neural network: a case study. Energy Build 2012;50:339–46.
[42] Porkhial S, et al. Modeling and prediction of geothermal reservoir temperature behavior using evolutionary design of neural networks. Geothermics 2015;53:320–7.
[43] Tugcu A, Arslan O. Optimization of geothermal energy aided absorption refrigeration system—GAARS: a novel ANN-based approach. Geothermics 2017;65:210–21.
[44] Rezvanbehbahani S, et al. Predicting the geothermal heat flux in Greenland: a machine learning approach. Geophys Res Lett 2017;44(24).
[45] Ishitsuka K, et al. Resistivity-based temperature estimation of the kakkonda geothermal field, Japan, using a neural network and neural kriging. Geosci Rem Sens Lett IEEE 2018;15(8):1154–8.
[46] Tut Haklidir FS, Haklidir M. Prediction of reservoir temperatures using hydrogeochemical data, western anatolia geothermal systems (Turkey): a machine learning approach. Nat Resour Res 2020;29(4):2333–46.
[47] Lösing M, Ebbing J. Predicting geothermal heat flow in Antarctica with a machine learning approach. J Geophys Res Solid Earth 2021;126(6).
[48] Hu S, et al. Thermo-economic optimization of the hybrid geothermal-solar power system: a data-driven method based on lifetime off-design operation. Energy Convers Manag 2021:229.
[49] Senturk Acar M. Multi-stage artificial neural network structure-based optimization of geothermal energy powered Kalina cycle. J Therm Anal Calorim 2020;145(3):829–49.
[50] Shahdi A, et al. Exploratory analysis of machine learning methods in predicting subsurface temperature and geothermal gradient of Northeastern United States. Geoth Energy 2021;9(1).
[51] Bourhis P, et al. Machine learning enhancement of thermal response tests for geothermal potential evaluations at site and regional scales. Geothermics 2021;95.
[52] He J, et al. A machine learning methodology for predicting geothermal heat flow in the bohai bay basin, China. Nat Resour Res 2022;31(1):237–60.
[53] Mehrenjani JR, Gharehghani A, Sangesaraki AG. Machine learning optimization of a novel geothermal driven system with LNG heat sink for hydrogen production and liquefaction. Energy Convers Manag 2022;254:115266.
[54] Pei H, et al. Long-term thermomechanical displacement prediction of energy piles using machine learning techniques. Renew Energy 2022;195:620–36.
[55] Yang F, et al. Artificial neural network based prediction of reservoir temperature: a case study of Lindian geothermal field, Songliao Basin, NE China. Geothermics 2022;106.
[56] Xiao F, et al. Performance enhancement of horizontal extension and thermal energy storage to an abandoned exploitation well and satellite LNG station integrated ORC system. Appl Therm Eng 2022;214:118736.
[57] Asai P, et al. Effect of different flow schemes on heat recovery from Enhanced Geothermal Systems (EGS). Energy 2019;175:667–76.
[58] Asai P, et al. Performance evaluation of enhanced geothermal system (EGS): surrogate models, sensitivity study and ranking key parameters. Renew Energy 2018;122:184–95.
[59] Asai P, et al. Efficient workflow for simulation of multifractured enhanced geothermal systems (EGS). Renew Energy 2019;131:763–77.
[60] Hofmann H, et al. Potential for enhanced geothermal systems in Alberta, Canada. Energy 2014;69:578–91.
[61] Ansari E, Hughes R, White CD. Modeling a new design for extracting energy from geopressured geothermal reservoirs. Geothermics 2018;71:339–56.

[62] Zhang L, et al. Potential assessment of CO2 injection for heat mining and geological storage in geothermal reservoirs of China. Appl Energy 2014;122:237–46.
[63] Araújo TP, Leite MGP. Flow simulation with reactive transport applied to carbonate rock diagenesis. Mar Petrol Geol 2017;88:94–106.
[64] Duboue P. The art of feature engineering: essentials for machine learning. Cambridge University Press; 2020.
[65] Pedregosa F, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;(12):2825–30.
[66] Zhang J, et al. A supervised learning approach for accurate modeling of CO2–brine interfacial tension with application in identifying the optimum sequestration depth in saline aquifers. Energy Fuel 2020;34(6):7353–62.
[67] Chong Q, Wang J, Gates ID. Evaluation of energy extraction from a geothermal resource in central Alberta, Canada using different well configurations. Geothermics 2021;96:102222.
[68] Warren J, Root PJ. The behavior of naturally fractured reservoirs. Soc Petrol Eng J 1963;3(3):245–55.
[69] Ansari E, Hughes R, White CD. Statistical modeling of geopressured geothermal reservoirs. Comput Geosci 2017;103:36–50.
[70] Zhao Z, Liu H. Spectral feature selection for data mining. Boca Raton, FL: Chapman & Hall/CRC data mining and knowledge discovery series; 2012 [CRC Press].
[71] Zeng Y-C, et al. Numerical simulation of heat production potential from hot dry rock by water circulating through a novel single vertical fracture at Desert Peak geothermal field. Energy 2013;63:268–82.
[72] Galli S. Python feature engineering cookbook: over 70 recipes for creating, engineering, and transforming features to build machine learning models. Birmingham: Birmingham: Packt Publishing, Limited; 2020.
[73] Bonaccorso G. Mastering machine learning algorithms : expert techniques to implement popular machine learning algorithms and fine-tune your models. first ed. Birmingham: Packt; 2018.
[74] Giuseppe B. Mastering machine learning algorithms. second ed. Packt Publishing; 2020.
[75] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Statistician 1992;46(3):175–85.
[76] Hashemizadeh A, et al. Experimental measurement and modeling of water-based drilling mud density using adaptive boosting decision tree, support vector machine, and K-nearest neighbors: a case study from the South Pars gas field. J Petrol Sci Eng 2021;207:109132.
[77] Huang Y, et al. Campus building energy usage analysis and prediction: a SVR approach based on multi-scale RBF Kernels. Cham. Springer International Publishing; 2015. p. 441–52. Cham.
[78] Drucker H, et al. Support vector regression machines. Adv Neural Inf Process Syst 1996;9.
[79] Bishop CM, Nasrabadi NM. Pattern recognition and machine learning, vol. 4. Springer; 2006.
[80] Chen Y, et al. Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings. Appl Energy 2017;195:659–70.
[81] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: International conference on knowledge discovery and data mining. ACM; 2016.
[82] Dong Y, et al. A data-driven model for predicting initial productivity of offshore directional well based on the physical constrained eXtreme gradient boosting (XGBoost) trees. J Petrol Sci Eng 2022:211.
[83] Tang J, et al. A new ensemble machine-learning framework for searching sweet spots in shale reservoirs. SPE J 2020;26(1):482–97. 1996.
[84] Chen T. Introduction to boosted trees. University of Washington Computer Science 2014;22(115):14–40.
[85] Kattan A, Abdullah R, Geem ZW. Artificial neural network training and software implementation techniques. Hauppauge, UNITED STATES: Nova Science Publishers; 2011 [Incorporated].
[86] Mohammed M. In: Khan MB, Bashier EBM, ebrary I, editors. Machine learning : algorithms and applications. Boca Raton: CRC Press; 2017.
[87] Amirian E, Z.J.J.G.J.o.T. Chen. Optimization. Cognitive data-driven proxy modeling for performance forecasting of waterflooding process 2017;8(2):1–8.
[88] Lippmann R. An introduction to computing with neural nets. IEEE ASSP Mag 1987;4(2):4–22.
[89] Foroud T, Seifi A, AminShahidi B. Assisted history matching using artificial neural network based global optimization method – applications to Brugge field and a fractured Iranian reservoir. J Petrol Sci Eng 2014;123:46–61.
[90] Ayyadevara, V.K., Pro machine learning algorithms: a hands-on approach to implementing algorithms in Python and R. 2018, Berkeley, CA: Berkeley, CA: Apress L. P.
[91] Ortiz-Bejar J, et al. K-nearest neighbor regressors optimized by using random search. IEEE; 2018.
[92] Silva DJ, Ventura J, Araújo JP. Predicting the performance of magnetocaloric systems using machine learning regressors. Energy and AI 2020;2:100030.
[93] Johannesen NJ, Kolhe M, Goodwin M. Relative evaluation of regression tools for urban area electrical energy demand forecasting. J Clean Prod 2019;218:555–64.
[94] Pedregosa F, et al. Scikit-learn: Machine learning in Python 2011;12:2825–30.
[95] Fan J, et al. Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: a case study in China. Energy Convers Manag 2018;164:102–11.
[96] Hong W-C. Electric load forecasting by seasonal recurrent SVR (support vector regression) with chaotic artificial bee colony algorithm. Energy 2011;36(9):5568–78.
[97] Ko C-N, Lee C-M. Short-term load forecasting using SVR (support vector regression)-based radial basis function neural network with dual extended Kalman filter. Energy 2013;49(1):413–22.
[98] Zhong Z, Carr TR. Application of mixed kernels function (MKF) based support vector regression model (SVR) for CO2 – reservoir oil minimum miscibility pressure prediction. Fuel 2016;184(C):590–603.
[99] Chung Y-S. Factor complexity of crash occurrence: an empirical demonstration using boosted regression trees. Accid Anal Prev 2013;61:107–18.
[100] Liu W, Liu WD, Gu J. Predictive model for water absorption in sublayers using a Joint Distribution Adaption based XGBoost transfer learning method. J Petrol Sci Eng 2020;188:106937.
[101] Mo H, et al. Developing window behavior models for residential buildings using XGBoost algorithm. Energy Build 2019;205:109564.
[102] Badem H, et al. A new efficient training strategy for deep neural networks by hybridization of artificial bee colony and limited–memory BFGS optimization algorithms. Neurocomputing 2017;266:506–26.
[103] Biglari F. Dynamic scaling on the limited memory BFGS method. Eur J Oper Res 2015;243(3):697–702.
[104] Shi Z, Yang G, Xiao Y. A limited memory BFGS algorithm for non-convex minimization with applications in matrix largest eigenvalue problem. Math Methods Oper Res 2015;83(2):243–64.
[105] Berihun Mamo N, Atta Dennis YAW. Artificial neural network based production forecasting for a hydrocarbon reservoir under water injection. Petrol Explor Dev 2020;47(2):383–92.
[106] He W, et al. Using of artificial neural networks (ANNs) to predict the thermal conductivity of zinc oxide–silver (50%–50%)/water hybrid Newtonian nanofluid. Int Commun Heat Mass Tran 2020;116:104645.
[107] Naqvi SR, et al. Pyrolysis of high-ash sewage sludge: thermo-kinetic study using TGA and artificial neural networks. Fuel 2018;233:529–38.