Geothermal Energy

RESEARCH Open Access

Exploratory analysis of machine learning methods in predicting subsurface temperature and geothermal gradient of Northeastern United States

Arya Shahdi[1*], Seho Lee[1], Anuj Karpatne[1] and Bahareh Nojabaei[2]

*Correspondence:
aryashahdi@vt.edu
[1] Department of Computer Science at Virginia Tech, Blacksburg, VA, USA
Full list of author information is available at the end of the article

Abstract

Geothermal scientists have used bottom-hole temperature data from extensive oil and gas well datasets to generate heat flow and temperature-at-depth maps to locate potential geothermally active regions. Considering that there are some uncertainties and simplifying assumptions associated with the current state of physics-based models, in this study, the applicability of several machine learning models is evaluated for predicting temperature-at-depth and geothermal gradient parameters. Through our exploratory analysis, it is found that XGBoost and Random Forest result in the highest accuracy for subsurface temperature prediction. Furthermore, we apply our model to regions around the sites to provide 2D continuous temperature maps at three different depths using XGBoost model, which can be used to locate prospective geothermally active regions. We also validate the proposed XGBoost and DNN models using an extra dataset containing measured temperature data along the depth for 58 wells in the state of West Virginia. Accuracy measures show that machine learning models are highly comparable to the physics-based model and can even outperform the thermal conductivity model. Also, a geothermal gradient map is derived for the whole region by fitting linear regression to the XGBoost-predicted temperatures along the depth. Finally, through our analysis, the most favorable geological locations are suggested for potential future geothermal developments.

Keywords: Renewable energy, Geothermal energy, Machine learning, XGBoost, Subsurface temperature, Geothermal gradient

Introduction

Bottom-hole temperature (BHT) measurements have largely been used for mapping subsurface temperatures for geothermal resource analysis across the United States (Blackwell and Richards 2010; Frone and Blackwell 2010; Stutz et al. 2012; Tester et al. 2006). BHT data are predominantly provided by oil and gas wells, where maximum temperature is usually reported at the final drilled depth. In 2010, Blackwell and Richards (2010) incorporated BHT data in northeastern United States with stratigraphic information (Childs 1985), and used a simple thermal conductivity model to generate surface heat

flux and temperature-at-depth maps. Jordan et al. (2016) conducted a thorough analysis to explore the associated risks and potentials of prospective geothermal resources in the states of New York, Pennsylvania and West Virginia. Even though most geothermally active regions are located in the western United States (near Earth's tectonic plate boundaries), Jordan et al. (2016) showed that the stored energy in the low-temperature geothermal regions in the northeast could be utilized for many direct-use applications. Although Snyder et al. (2017) illustrated that myriad industrial and residential direct-use applications of geothermal energy could result in reduction of electricity consumption, there are not many geothermal sites in northeastern states due to a high financial risk. Heat flux and temperature-at-depth are two most important geothermal parameters, which have extensively been investigated through physics-based models.

In the previous geothermal studies, the generalized thermal conductivity model has been adopted to compute the heat flow associated with BHT data points (Blackwell and Richards 2010; Cornell University 2015; Frone and Blackwell 2010; Jordan et al. 2016; Stutz et al. 2012; Tester et al. 2006). To use this model, first the measured bottom-hole temperature is corrected (through various available correlations (Deming 1989)) and is used to calculate the temperature gradient through the following relation:

$$\left(\frac{dT}{dz}\right) = \frac{BHT - T_{surf}}{z}. \tag{1}$$

Next, the geological formation thickness and thermal conductivity values are approximated at each well location's latitude and longitude mainly from Correlation of Stratigraphic Units of North America (COSUNA) (Childs 1985). Then, average thermal conductivity is calculated between surface and the well's depth (Stutz et al. 2012). Finally, the heat flux is calculated through the following equation:

$$Q_s = \bar{k}\left(\frac{dT}{dz}\right). \tag{2}$$

The above formula is oversimplified and only represents the main theoretical framework of the physics-based model, which is used in geothermal energy studies. Despite physics-based model's long-time applicability, they all have some underlying assumptions that could result in uncertainties and, therefore, inaccurate predictions. Some of the assumptions are explained by (Stutz et al. 2012) and (Blackwell and Richards 2010). In particular, there is no easy-to-use method to independently measure the heat flux parameter; it is only approximated through the thermal conductivity model using the BHT data as shown in Eq. (2).

In addition to the geothermal energy industry, subsurface temperature is an extremely important parameter in the oil and gas industry (Bassam et al. 2010; Forrest et al. 2005; Khan and Raza, 1986; Moses, 1961). Characteristics of hydrocarbons are greatly dependent on the temperature and they must be approximated to be used in reservoir and drilling simulations. In practice, it is common to use geothermal gradient maps to obtain the geothermal gradient value at the desired location and then calculate the subsurface temperature at the depth of interest (Forrest et al. 2005; Khan and Raza, 1986).

Machine learning and geostatistics have been used in the variety of applications to help investors make more confident decisions. Due to the inaccessible nature of the

geothermal energy, there is a considerable amount of risk and uncertainty associated with the exploration (Witter et al. 2019), drilling (Lukawski et al. 2016) and production (Bloomquist et al. 2012). There are few comprehensive surveys that focused on analyzing the associated risks to provide insights about the potential of developing geothermal sites (Jordan et al. 2016; Young et al. 2010). Machine learning has been an emerging technology that helped the geothermal energy field in the mentioned stages (Assouline et al. 2019; Beardsmore 2014; Faulds et al. 2020; Rezvanbehbahani et al. 2017; Shi et al. 2021; Tut Haklidir and Haklidir 2020). In the next section, we briefly review the studies which applied machine learning successfully in the fields of geothermal exploration and drilling.

Exploration stage

Recent machine learning advancements in some of the closely related fields of geology and geoscience have tremendously helped the geothermal energy industry in the exploration and drilling stages. For example, applications of machine learning in characterization of geomechanical properties (Keynejad 2018), automated fault detection and interpretation (Ma et al. 2018; Zhang et al. 2014), geophysical data inversion (Araya-Polo et al. 2018) and categorizing different lithofacies (Hall 2016). Perozzi et al. (Perozzi et al. 2019) took it further and proposed machine learning schemes to accelerate geological interpretations (specifically from well-logs) and, consequently, reducing the geothermal exploration costs.

Rezvanbehbahani et al. (2017) proposed a machine learning approach to estimate the geothermal heat flux (GHF) in Greenland using the global GHF data provided by the International Heat Flow Commission (Gosnold and Panda 2002). For modeling, Gradient Boosted Regression Tree method was used with an average 15% relative error, RMSE and r^2 of 0.14 and 0.75, respectively. In that study, even though the authors provided a preliminary map to annotate most favorable locations in Greenland in terms of geothermal potential, however, wellbore bottom-hole temperature data were not utilized. In another effort, machine learning was used to map very shallow geothermal potential (Assouline et al. 2019). In shallow depths, geothermal energy can be a very good source to provide thermal energy for residential areas (Vieira et.al. 2017). Assouline et al. used Radom Forrest to predict three important thermal variables that are crucial in analyzing the geothermal potential of the region. These variables include (1) temperature gradient, (2) thermal conductivity, and 3) thermal diffusivity throughout Switzerland.

Another interesting study was conducted which primarily focused on developing a probabilistic modeling approach to identify the underlying risks in the field of geothermal resource exploration and the application of machine learning in the geothermal energy industry (Beardsmore 2014). An open-source software was developed named "Obsidian" which is capable of joint inversion of numerous geophysical datasets with probabilistic outputs. This study had access to a rich dataset containing formation characteristics, local temperature info and multiple case studies located in different regions of Australia. In addition to 3D temperature-at-depth maps, they were able to generate a 3D probabilistic map where each given point represents the probability of having granite rock type. The combination of the two mentioned maps was intended to directly

help investors choose the right depth, latitude and longitude with the highest success probability.

Drilling stage

After finding the prospective geothermally active regions, geothermal wells are drilled for production. Drilling stage can comprise up to 45% of the total cost of the geothermal project (Muhammad 2019). Machine learning has helped the industry to efficiently design this stage from different aspects. Drilling optimization considerations in geothermal wells can be categorized into (1) reducing drilling time and (2) minimizing operational failures. This subject is shared between geothermal and oil and gas industries where drilling operations are remarkably similar. There are myriad studies where machine learning techniques have successfully addressed the mentioned issues and provided reliable solutions to optimize the drilling stage (Barbosa et al. 2019; Hegde et al. 2020; Hegde and Gray 2017, 2018; Noshi and Schubert 2018). Recently, the Department of Energy has funded a project with the theme of application of deep machine learning to optimize drilling operations (specifically for geothermal wells) which was awarded to Oregon State University with collaboration with one more US university, one DOE National Laboratory, in addition to four geothermal and oil and gas companies from Iceland, US and Norway (DOE, 2019). In the first-year report of this study, the major effort was made around four primary tasks (well data gathering, feature engineering, data repository development, and preliminary machine learning model testing). It was mainly found that more extensive data from bit life cycle and bottom-hole assembly (BHA) are needed to improve the machine learning models. Finally, they compared different machine and deep learning models to predict important drilling parameters and it was found that Random Forrest model outperforms others as number of inputs increases. There was an extra effort to include the lithological information (mainly from mud log data) by dummy encoding and text embedding to, potentially, increase the accuracy (Carbonari et al. 2021).

In this study, we provide an alternative solution of using machine learning methods for predicting subsurface temperature using BHT data from more than 20,750 oil and gas wells in the northeastern United States. Furthermore, the physics-based and machine learning models are compared through an extra dataset containing vertical temperature profile of 58 wells in the state of West Virginia. Finally, we provide the geothermal gradient map using the validated XGBoost model for the northeast region of the United States.

Case study

The Marcellus formation is one of the highest potential hydrocarbon prospects in the United States and is located throughout the northern Appalachian Basin. For several decades, thousands of wells have been drilled in this region which contain, at least one temperature measurement (usually at the final depth). For our analysis, we have used a dataset with raw and corrected BHT, surface temperature, well identification number (API), latitude, longitude, and geological setting information (including layer thickness and conductivity) and many other information from 20,750 oil and gas wells in the northeast. This dataset (Cornell University 2015) has been developed and reported as

part of a DOE funded research grant led by Cornell University. In Fig. 1, we show the geospatial spread of the well locations (of the dataset). In the right plot, the scatter points are referred to 20,750 well locations of the main dataset and the shaded area depicts the region where temperature predictions are provided by our study. The left plot in Fig. 1 is a magnified view of the West Virginia state region where the blue points represent a new set of well locations where we had more than one temperature measurement for each well. In fact, for many wells, subsurface temperature measurements were available along hundreds of meters within the well. We primarily used this dataset for further verification of our geothermal gradient predictions.

Dataset-1 summary

In Table 1, a summary of important parameters (after outlier removal) is provided. We have used 55 features that are included in Table 1. Among the variables, the geological characteristics are included through the multiplication product of each formation conductivity and thickness (6–55). This is consistent with the thermal conductivity theory (Eq. (2)). At each well's latitude and longitude, there are up to 49 formation layers where each layer has specific thickness and conductivity.

Dataset-2 summary

We also exclusively gathered data for additional 58 wells across the West Virginia region (annotated by blue points on Fig. 1). In this dataset, for each well, temperature profile is provided within a depth interval (with the mean and standard deviation of 1167 and 511 m, respectively). We obtained this dataset from West Virginia Geological and Economical Survey (West Virginia Geological and Economical Survey Website n.d.). The digitized data were available in the LAS file format where temperature measurements (along with other geological parameters) were reported at different depths. We primarily used it for comparing our modeling results with those from the physics-based model. We refer to this source as the temperature-profile dataset throughout this paper. Among

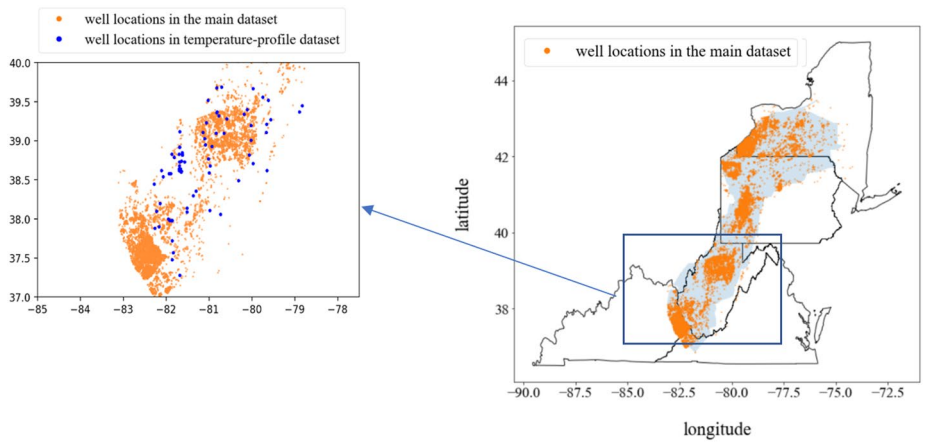

Fig. 1 Right plot represents the spread of oil and gas wells in the first dataset (containing 20,750 BHT data points). In the left plot, the locations of the 58 newly obtained wells (with full temperature profile) are annotated using the blue color

Table 1 Statistical summary of important parameters after outlier removal

	Surface temperature	Depth	Corrected BHT	Heat flow
Unit	°C	m	°C	mW/m^2
Mean	12.4	1154	37	49
std	1.8	459	13.2	13.4
min	8.8	43	10.2	0.2
25%	10.6	868	28.9	41.57
50%	12.1	1129	34.5	47.91
75%	14.3	1358	42.8	55.26
max	15.6	6541	146.9	130.21

Variable number	Name	Unit	Source	Description	Type
1	BHTCorr	°C	Well log report	Corrected bottom-hole temperature	Label
2	LatDegree	–	Well log report	Lat degree of the well's location	Feature
3	LongDegree	–	Well log report	Long degree of the well's location	Feature
4	MeasureDepth	M	Well log report	The depth where BHT is recorded	Feature
5	SurfTemp	°C	Annual average temperature	Surf temperature at the well's location	Feature
6 to 55	KH	W/(°K)	Approximated from the data reported in Correlation of Stratigraphic Units of North America (COSUNA)	Multiplication product of each geological layer's thickness with its corresponding thermal conductivity	Feature

the 58 wells, bottom-hole temperature points of 11 wells already exist in the first dataset (20,750 wells). The rest are new wells which have been used to compare the physics-based model with the machine learning methods.

BHT correction methods

For BHT correction, the authors (Jordan et al. 2016) divided the Appalachian Basin into three regions (West Virginia, Pennsylvania Rome Trough and Allegheny Plateau) and developed exclusive correction correlations based on available information at each of these regions (for example, in Allegheny Plateau region, information about drilling fluids were accessible to the authors in contrast to the West Virginia section where drilling fluid data were not available). For each region, a small set of equilibrium well-log temperature measurements were statistically evaluated and a new set of appropriate BHT corrections were proposed. In West Virginia region, a Generalized Least Square (GLS) regression model was fitted through Eq. (3). For Pennsylvania Rome Trough, no statistically significant relation was found with depth and therefore no adjustment was applied. Fortunately, for Allegheny Plateau, the drilling fluid data were available, and the correlation equations were proposed for different fluids as shown below.

$$\Delta T_{WVA} = -1.99 + 0.00652z,\ 305\ m < z < 2606m, \tag{3}$$

$$\Delta T_{Alle.\ Pt.\ Air} = 0.0104\left(\left(1090^3 + z^3\right)^{0.33} - 1090\right),\ Z < 2500m, \tag{4}$$

$$\Delta T_{Alle.\ Pt.\ Mud} = 0.0155\left(\left(1660^3 + z^3\right)^{0.33} - 1660\right),\ Z < 4000m. \tag{5}$$

Outlier removal approach

For preprocessing, we removed outliers (101 data points) using the common 3σ-rule method where data outside the three standard deviation are considered outliers (Lehmann 2013; Pukelsheim, 1994; Watanabe et al. 2019) using the heat flux parameter (Fig. 2).

The reported temperatures in the temperature-profile dataset are prone to errors and we were required to correct them. Even though there are myriad temperature-correction methods, we decided to use the correction methodology reported by (Jordan et al. 2016) to be consistent with their method. This allowed us to compare our results to those reported by the physics-based model in (Jordan et al. 2016). Since all wells in the temperature-profile dataset are located in the West Virginia region, we decided to use Eq. (3).

Methodology

Machine learning models

In this section, we provide a thorough summary of the machine learning models that have been used in this study to estimate subsurface temperature and geothermal gradient. We decided to use multiple algorithms to train our regression models, including Deep Neural Networks (DNN), Ridge regression (R-reg) models and decision-tree-based models (e.g., XGBoost and Random Forest).

In this paper, we compare the results of four machine learning algorithms. These algorithms are different in nature and it is extremely important to appropriately compare their accuracies and errors. For each algorithm, we primarily focused on developing

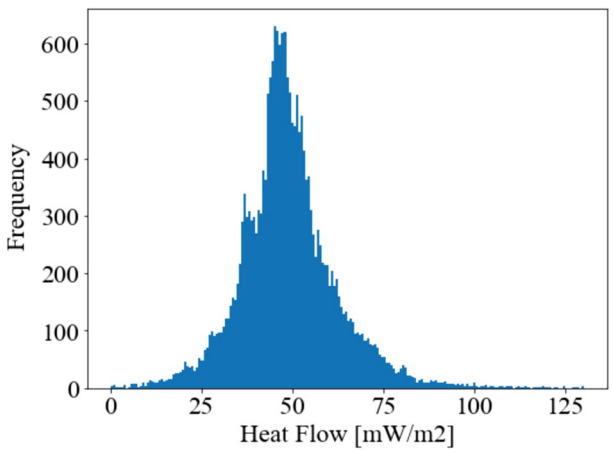

Fig. 2 Heat-flow histogram after outlier removal

the best performing model. This not only applies to hyper-parameter tuning, but also to the data preprocessing. In particular, we standardized the input features for Ridge Regression and DNN. For XGBoost and Random Forest models, we did not observe any improvement after standardizing the features and, therefore, we did not decide to standardize the input features. The tunned hyper-parameters are reported in the GitHub repository (Shahdi and Lee 2021).

Figure 3 illustrates the developed machine learning pipeline which has been used for this study. In the data preprocessing section, outliers are removed, and features are scaled (for R-reg and DNN). Next, hyper-parameters related to each model are tuned using cross-validation. At the end, the final model is also evaluated using cross-validation. This process is repeated for all models.

Ridge regression

In our dataset, there are uncertainties (noise) associated with the BHT data potentially from temperature logging tools, and/or the BHT correction correlations, etc. We used Ridge regression as one of the candidate machine learning models. Despite its simplicity, it is robust to overfitting (regulated by a penalty term known as L2 Regularization) (Hoerl and Kennard 1970). (Wyffels et al. 2008) showed how Ridge Regression is robust to noise and overfitting in reservoir computing and signal processing applications. In another study, it was shown how Ridge Regression can be a superior solution when the multi-collinearity problem between independent variables exists comparing to other complex models (Morgül Tumbaz and İpek 2021). Baruque et al. (Baruque et al. 2019) successfully used Ridge regression for a geothermal application where heat exchanger

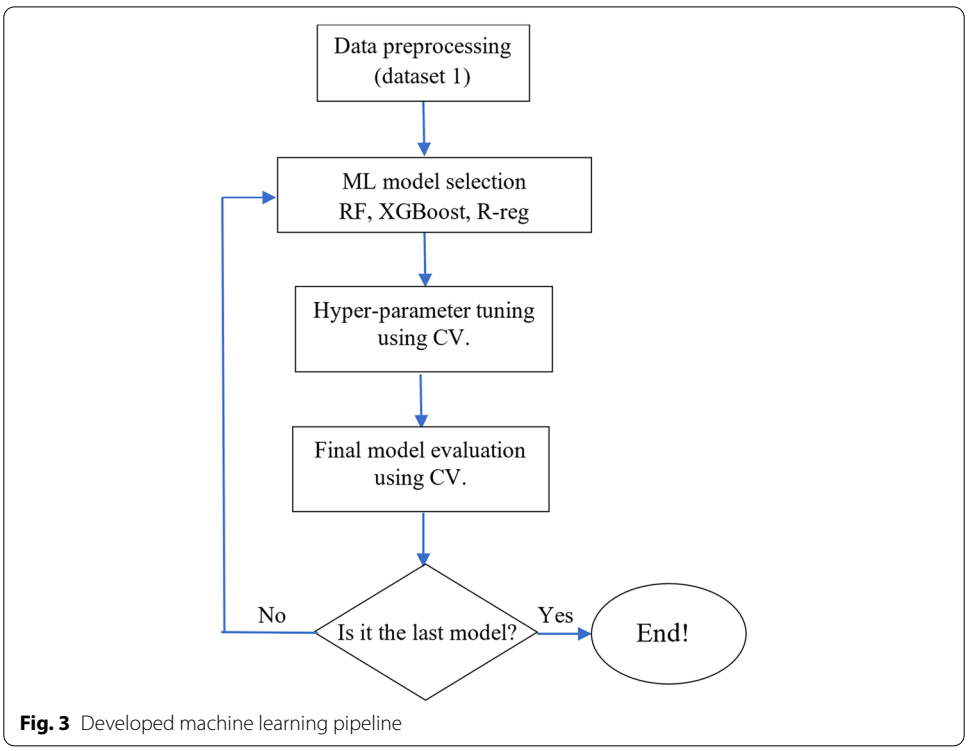

Fig. 3 Developed machine learning pipeline

energy was predicted using time series readings of several sensors. The goal is to find the model's parameters which minimize the objective function.

$$\hat{\theta}^{ridge} = \underset{\theta}{\operatorname{argmin}} \left(y - X\Theta_2^2 + \alpha\Theta_2^2 \right), \tag{6}$$

where hyper-parameter α is a positive number that specifies the trade-off between the ordinary least squares (OLS) and regularization terms. In our implementation, we initially standardized the inputs (with BHT targets) and then fed them into the hyper-parameter tunning section. We used the grid-search method to search for the best alpha (shown in Table 2).

XGBoost and Random Forest

Ensemble modeling approach is a process where numerous base models are generated to estimate an outcome. The base models are independent and diverse and tend to decrease the generalization error of the prediction. This methodology exploits the wisdom of crowds to make an approximation. Even though there are multiple base models associated with an ensemble model, it behaves as a single predictor. Typically, a weighted average of all base models' predictions will be reported as the final outcome (Vijay and Bala 2014). Random forest and XGBoost are both ensemble models which have widely been used for regression and classification problems. Random Forest constructs multiple decision trees at the time of training and provides the average estimation of individual trees (Breiman 2001). Whereas in XGBoost, the estimators (trees) are sequentially added to the ensemble model to improve the accuracy by adding a base learner to correct the shortcomings of the already existing base models. In XGBoost, the shortcomings are determined by gradients (Li 2016). In this study, target imbalance problem is present within our dataset since ninety-sex percent of BHT data correspond to the shallower wells (< 2000 m). On the other hand, the deeper wells contain valuable information with higher temperature values which should not be removed (or be considered as outliers). We mainly used ensemble-based algorithms including Random Forest (Liaw and Wiener 2002) and XGBoost (Chen and Guestrin 2016) because they are believed to work relatively well in a case where target imbalance exists (Moniz et al. 2017). In addition, tree-based models usually improve the accuracy by decreasing the variance in the prediction

Table 2 Information about hyper-parameters related to Ridge-regression, Random Forest and XGBoost models

Model	Hyper-parameter	Range	Optimum
Ridge-Reg	Alpha	[0.001, 100]	0.01
Random Forest	Max_depth	{5,8,10,12,15}	12
Random Forest	N_estimators	{100,500,1000}	500
Random Forest	Min_samples_leaf	{1,2}	2
Random Forest	Min_samples_split	{2,3}	2
XGBoost	Max_depth	{5,8,10,12}	8
XGBoost	N_estimators	{100,500,1000}	500
XGBoost	Learning_rate	{0.01,0.05,0.1,0.2}	0.05
XGBoost	Gamma	{0.1,1,10}	10
XGBoost	Reg_lambda	{0.1,1,10}	10

(Polikar 2012). XGBoost and Random Forest are both tree-based methods which have been successfully applied in geosciences (Gul et al. 2019; Hall 2016; Sun et al. 2020). Single decision tree is often referred to as a weak classifier as it can be susceptible to overfitting (Ho 1998). Random Forest builds an ensemble of multiple decision trees (weak classifiers) in parallel and takes the mean of the predictors for the prediction. Furthermore, during the ensemble construction, random features or columns are dropped while learning every decision tree, so that every tree is de-correlated from other trees as much as possible. XGBoost, on the other hand, builds decision trees in a sequential manner. XGBoost keeps adding decision trees at every step, making a fine separation in space to predict the response variable (Chen and Guestrin 2016). Every new step considers the previous steps which result in accuracy improvement after each iteration. XGBoost is a library that allows XGBoost to be run in parallel in terms of computing.

Deep neural network (DNN)

DNN is a network of connected processing elements (neurons) which are placed in multiple layers and is used to solve classification and regression problems. This is done through a learning process where the model parameters get adjusted in the training phase. In the training stage, the errors are propagated back in the network resulting in updating the model parameters (weights). This process continues till no further improvement is observed in the errors (Maind and Wankar 2014). We developed a deep neural network (DNN) architecture to predict the subsurface temperature. In our features, we include the thermal conductivity and thickness values of up to 55 formation layers for each well. In this relatively large feature dimension, we decided to use DNN to capture the non-linearity between these geological characteristics and bottom-hole temperatures. Bassam et al. (Bassam et al. 2010) was among the first studies that evaluated the application of a shallow artificial neural networks (ANN) in formation temperatures in geothermal wells. In that study, collected BHT logs (during long-shut-in times) have been used for training and validation. Kalogirou et al. (Kalogirou et al. 2012) generated ground temperature map at shallow depths by considering land configuration using ANN.

Deep neural networks attempt to capture the relationships between inputs and outputs using a deep assembly of hidden layers of neurons, where every neuron in a hidden layer receives signals (or activations) from neurons in the previous layer, and transmits activations to all neurons in the subsequent layer. DNN models can capture high amounts of non-linearity using a large (or deep) number of inter-connected hidden layers. We tried different DNN architectures and finally picked a four-layer DNN as illustrated in Fig. 4. In the input layer, the number of nodes is the same as feature numbers followed by two hidden layers where each layer contains 50 nodes. Arrows correspond to connections among nodes and are associated with learnable edge weights. In addition, we selected ReLU activation function in our architecture. For the last neuron at the output layer, the weighted responses from the neurons at the second hidden layer are fed into a linear activation function and the final prediction for temperature is obtained. In Fig. 5, one neuron of the hidden layer is illustrated with the given inputs.

In Table 2, we included the values that are used for hyper-parameter tuning for Ridge-Regression, Random Forest and XGBoost. For DNN, we did not perform

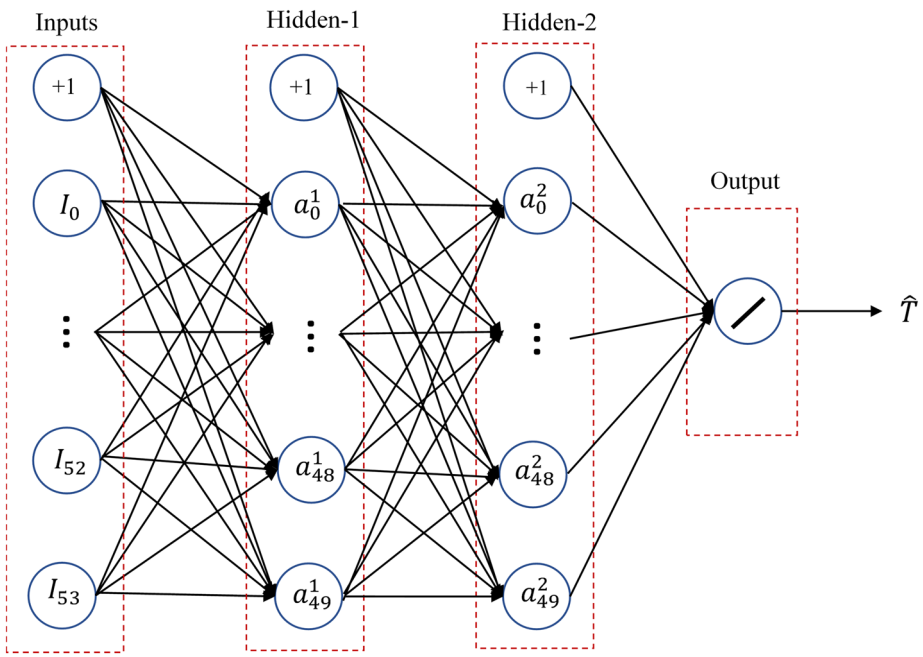

Fig. 4 Deep neural network architecture for subsurface temperature prediction

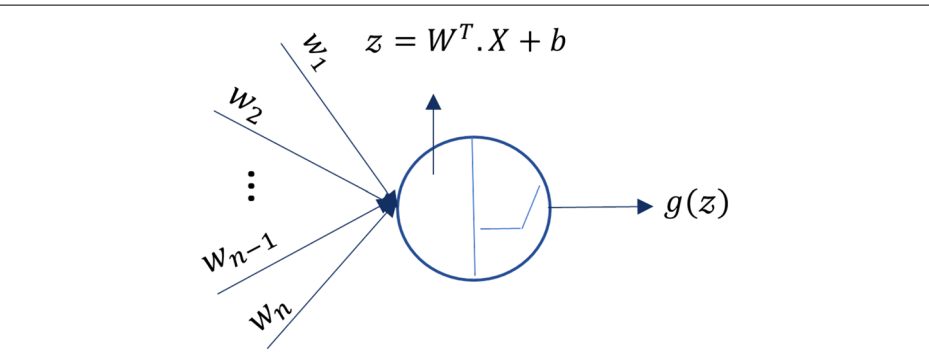

Fig. 5 Single neuron illustration

hyper-parameter tuning in the same fashion (mainly due to the computational time). We examined tens of different architectures and reached to one illustrated above.

Feature space interpolation

Temperature-at-depth maps have extensively been used in geothermal energy studies to illustrate the temperature distribution at a given depth. In this study, we also provide temperature-at-depth maps at different depths in the northeastern United States. This allows investors to have another source of temperature prediction map for any potential future development. In addition, the new machine learning temperature maps can be compared to those from the thermal conductivity model to locate the similarities and differences. A simple concave hull algorithm was used to obtain a tight boundary around the given data points. To avoid sharp edges, we derived average values for the boundary

points and then implemented the algorithm (shaded area in Fig. 1). We initially used an online source code (Dwyer n.d.) and made major modifications to meet our project's needs.

For constructing the subsurface temperature prediction map, the features should be available within different locations (with varying latitude and longitude). Therefore, we interpolated the required features (shown in Table 1) throughout the northeastern region using a Gaussian kernel weighted k-nearest neighbor regression model. These interpolated features are then fed into the trained machine learning models to generate the predicted temperature-at-depth maps. We chose KNN regression method since it is simple and is expected to perform well in our region of interest due to high concentration of wells. We used cross-validation for hyper-parameter tuning of the KNN method (K = 3 and kernel width = 0.037) using 20,750 data points.

Results and discussion

We trained the proposed machine learning models using the main dataset and observed that even though only single temperature measurement points (at each well location) were used for training, the machine learning models successfully predicted underground temperatures. Among the machine learning models, XGBoost and Random Forest outperformed other models and provided more accurate results. For further verifications, we compared the XGBoost, DNN and physics-based model's predictions versus the subsurface temperatures obtained from 58 additional wells in the temperature-profile dataset. This was important because unlike the main dataset, the temperature-profile dataset comprises temperature measurements within depth intervals. This allows us to investigate the machine learning model predictions versus depth. Fortunately, the results show that machine learning models predictions were in close agreement with the measured data.

Temperature-at-depth result analysis

After training and tuning hyper-parameters, we evaluated the accuracy of each model using the test data for using cross-validation. As shown in Fig. 6 and Table 4, XGBoost and Random Forest perform the best among other machine learning models. Statistical hypothesis tests (t tests) were performed. The comparisons of XGBoost with Ridge and DNN suggest that there is sufficient evidence to reject the null hypothesis and the observed differences between XGboost and the other two models in the regression accuracy is likely due to the differences in the models. However, the result of the hypothesis test on XGBoost and Random Forest suggests that there is insufficient evidence to reject the null hypothesis. Table 3 summarizes the p values for the tests.

We then used the trained models to predict subsurface temperature at three different depths (Z = 1000, 2000, 3000 meters)in the northeastern United States. In Fig. 7, temperature predictions are plotted using XGBoost models. For comparison purposes between the physics-based and machine learning subsurface temperature predictions, we used KNN method (k = 8 and width = 1 determined from cross-validation) for temperature interpolation for the physics-based model. To be more elaborate, in the main dataset, at each well's location, the predicted physics-based

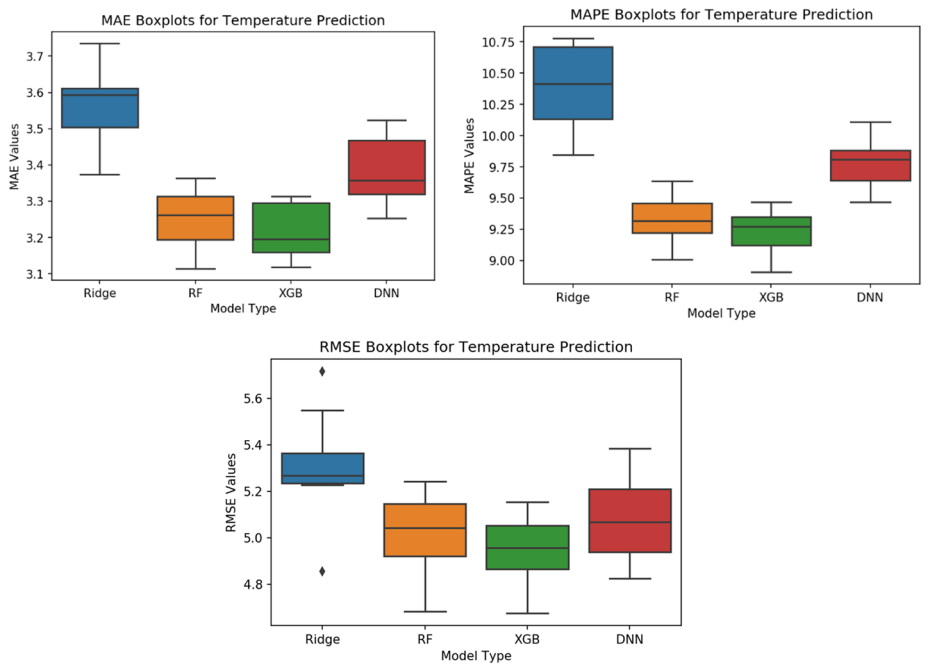

Fig. 6 Accuracy comparison between four machine learning models

Table 3 P-values obtained from statistical hypothesis tests

P-value	Ridge			RF			DNN		
	MAE	MSE	MAPE	MAE	MSE	MAPE	MAE	MSE	MAPE
XGBoost	1.47E−07	0.0019	1.25E−10	0.3693	0.4024	0.2490	0.0004	0.0733	9.28E−05

underground temperatures were provided along the depth. We used this data and KNN interpolation method to approximate the physics-based values at different latitudes, longitudes and depths.

Generalizability analysis

As discussed earlier, the target imbalance problem was present in our dataset since fewer data points were available for depths below 2000 m (or BHT larger than 60 °C). We conducted an experiment to compare XGBoost accuracy for well-represented and underrepresented data points in a test set. In Fig. 8, average percentage error (APE) versus depth is plotted for the test set where well represented and underrepresented data are illustrated by different colors. Furthermore, Fig. 9 shows the target distributions of the same test set (with one-to-one match with data points in Fig. 8). Next, we compared the mean absolute percentage error (MAPE) for well-represented and underrepresented test data and found both values to be remarkably similar (with less than 2% difference). Through this empirical analysis, we confirmed the generalizability of the XGBoost model.

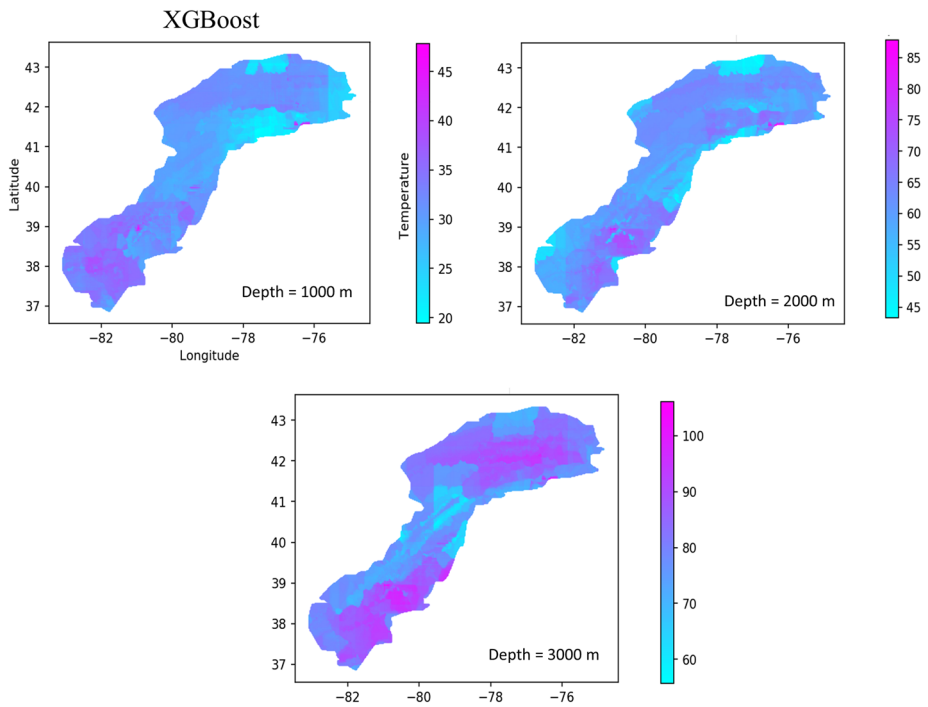

Fig. 7 Temperature map at three different depths using XGBoost model

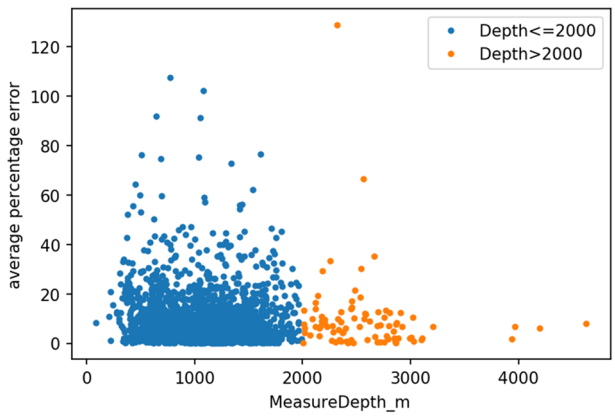

Fig. 8 Average percentage error calculated using XGBoost predictions and true BHT values for well-represented and underrepresented test data. In this instance, MAPE of blue and orange points are 9.17 and 10.05%, respectively

Temperature-profile prediction

In our analysis, we decided to use the corrected temperature-profile dataset (described in "Drilling stage" Section) to evaluate XGBoost and DNN accuracies against the thermal conductivity model. Jordan et al. reported the predicted subsurface temperatures (from the physics-based model) across the depth for each well's latitude and longitude in the main dataset. The size of the available predicted temperature data is 2075*500 where each well had 500 temperature prediction values at different depths. We used KNN regression model

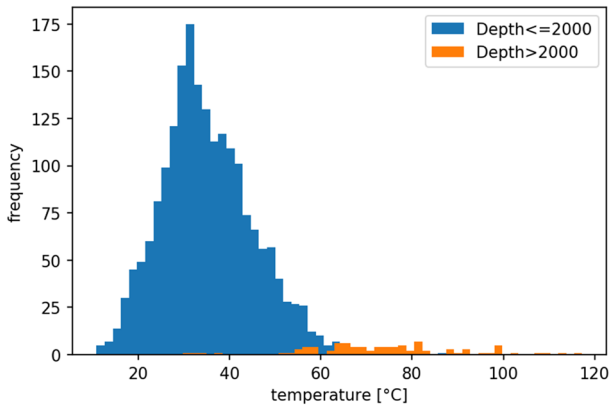

Fig. 9 Target (BHT) distributions for well-represented and underrepresented test data

(using the mentioned data) to interpolate temperature-profile predictions for the physics-based model at the new locations (in the temperature-profile dataset). In the following schematic, we illustrate the procedure that we have used to compare predictions from machine learning and the physics-based models.

After analyzing the results, the mean absolute errors of XGBoost, DNN, and physics-based models were calculated to be 7.3, 7.27, and 8.76, respectively, for the 58 wells. These numbers show that machine learning models can be comparable, in terms of accuracy, to the physics-based thermal conductivity model. It is important to note that we have used multiple interpolations to be able to perform such comparison (Fig. 10). Therefore, there is some level of uncertainties associated with the reported numbers.

For illustration purposes, we include six temperature-profile predictions (in Fig. 11), which are fair representatives of the remaining cases. Among all plots, we could see that the thermal conductivity model performs relatively better in tracking the true temperature data in 11.3 and 11.4. On the other hand, both XGBoost and DNN models provide more accurate results in 11.1 and 11.6. Nevertheless, there are some cases where all models fail to follow the actual data. For example, in plot 11.2, we could see that neither physics-based nor machine learning models predict the temperature profile accurately. Temperature-profile prediction plots of other wells are included in our GitHub repository (Shahdi and Lee 2021). Among machine learning predictions, DNN and XGBoost predictions follow very similar trends even though DNN curves are smoother and have less variation with depth. This is expected because decision-tree-based models tend to show such discrete predictive behavior when used for regression.

In Tables 4 and 5, we include each well's API well identification number with the distance from the closest well in the main dataset. The shown plots are from the wells that are close to at least one of the wells in the main dataset. This is important because it shows that the interpolated temperature values for the physics-based predictions are reliable and close to those reported by the original study (Jordan et al. 2016).

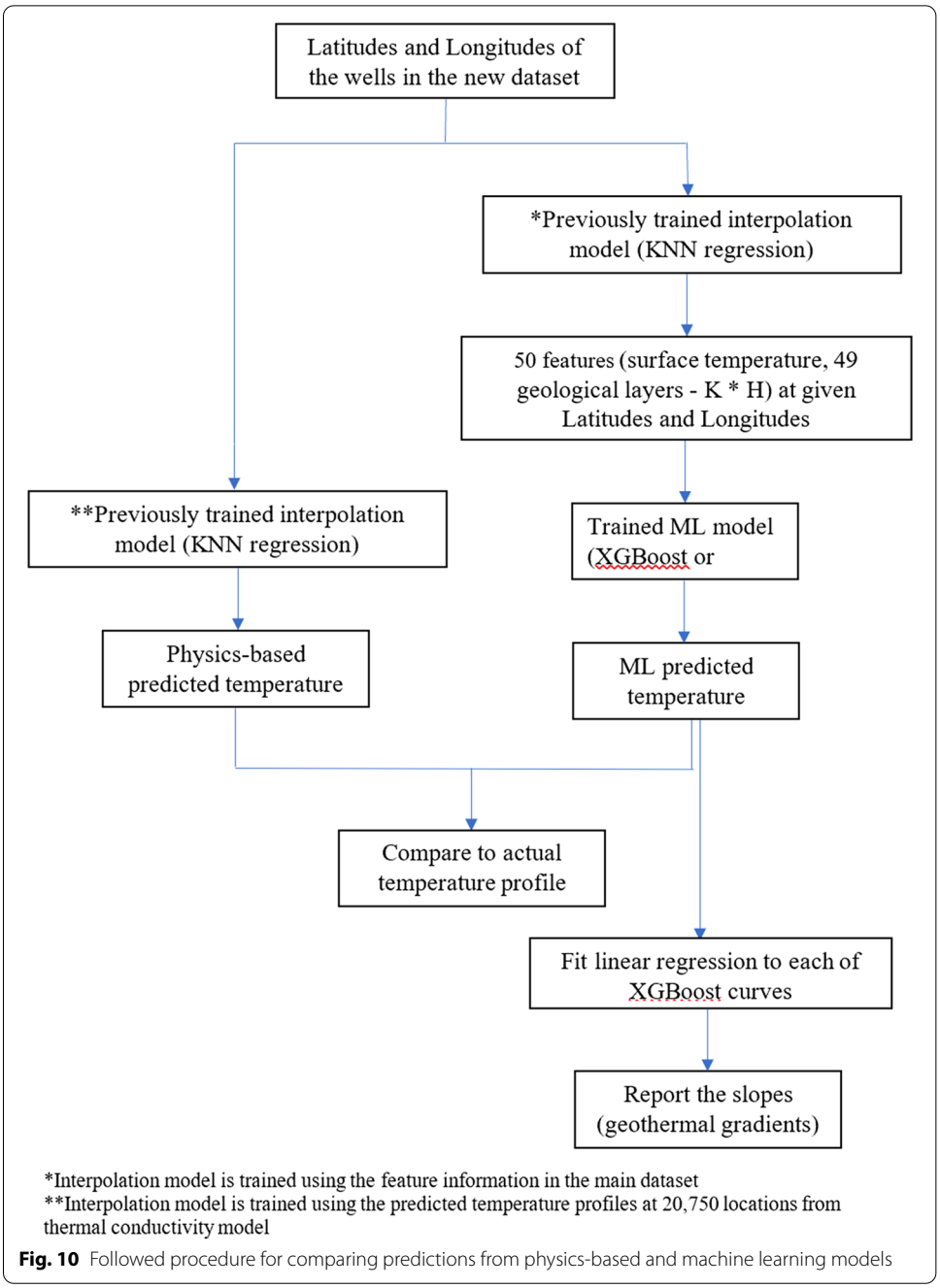

Fig. 10 Followed procedure for comparing predictions from physics-based and machine learning models

Geothermal gradient map

It is very popular to use geothermal gradient maps to predict the subsurface temperature at the desired location. In this study, we provide the geothermal gradient map for the northeastern United States.

Similar to the plots (shown in Fig. 11), we generate temperature-profile predictions for 28,000 locations across the region and then fit a linear regression line to the temperature data for each location. These 28,000 wells are defined symmetrically throughout the region of interest (bounded by the concave hull algorithm which is shown in Fig. 1). This

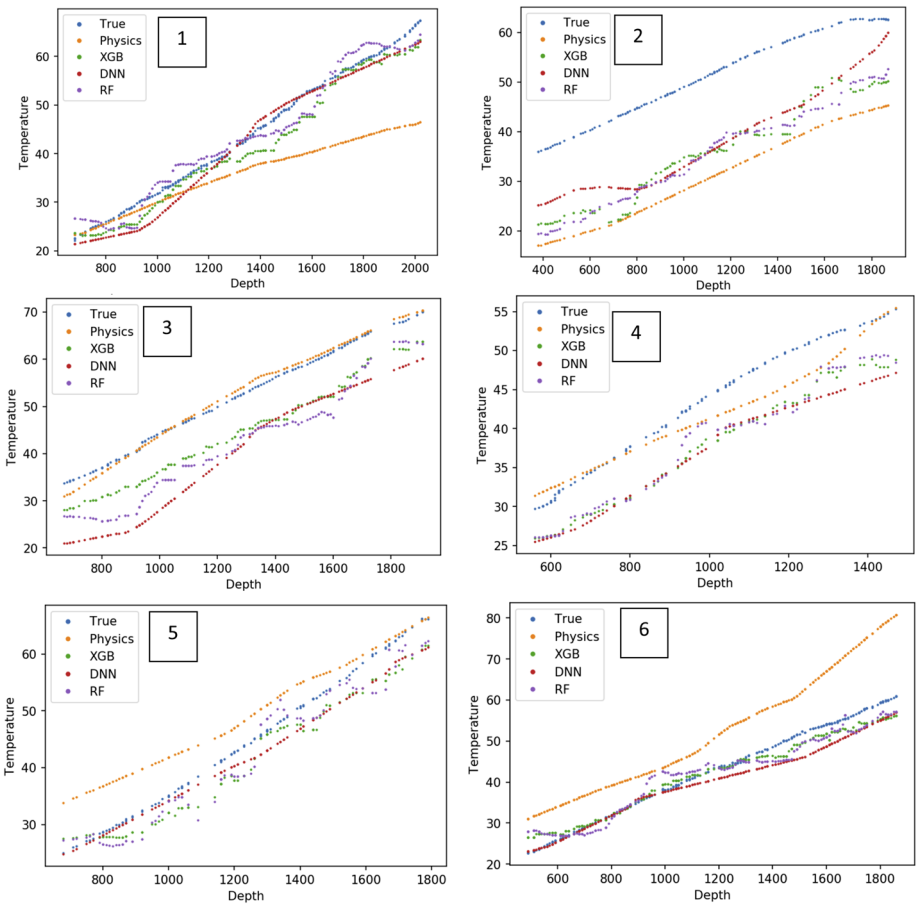

Fig. 11 Temperature-profile predictions using thermal conductivity, XGBoost and DNN models versus measured data. The units are ° C and *m* for temperature and depth, respectively

Table 4 Evaluations of machine learning models using the main dataset

	XGBoost	Random Forest	Deep neural network	Ridge regression
Root mean square error	4.94 ± 0.15	5.01 ± 0.17	5.08 ± 0.18	5.3 ± 0.21
Mean absolute error	3.21 ± 0.07	3.25 ± 0.08	3.39 ± 0.09	3.57 ± 0.1
Mean absolute Percentage error	9.22 ± 0.16	9.32 ± 0.18	9.77 ± 0.33	10.38 ± 0.33

Table 5 Corresponding details about the wells that are shown in Fig. 11. Distance column is referred to the distance from the test well to the closest well in the main dataset

Plot #	API well number	Distance [km]
1	4,710,300,645	0.26
2	4,707,500,050	0.03
3	4,709,501,963	0.22
4	4,700,502,167	0.50
5	4,701,304,647	0.34
6	4,705,900,805	3.27

was necessary for generating a continuous temperature gradient map. Through our analysis, we found that the fitted lines accurately represent the predicted temperatures with average R^2 of 0.97. The reported slopes are equal to the associated geothermal gradients and are illustrated in Fig. 12. The second map in Fig. 12 is a snapshot of an interactive Folium map within our region of interest.

In Fig. 13, areas with predicted geothermal gradient higher than $27\frac{^\circ C}{km}$ (obtained from Random Forest, XGBoost and DNN) are annotated. All three model predictions recommend similar areas in West Virginia and New York states to have high values for temperature gradient. We cautiously suggest these machine learning guided prospective regions for future geothermal developments.

Next, we calculated the mean absolute errors between the geothermal gradients predicted using different models (e.g., physics-based, XGBoost and DNN) and measured temperatures for the temperature-profile dataset (as shown in Table 6).

Conclusion

The goal of this paper is to highlight the importance and applicability of machine learning methods in producing reliable predictions of important geothermal parameters from the rich volumes of data available from geothermal sites. It is critical to understand that this paper does not claim to prove that machine learning models are ubiquitously superior to conventional physics-based models in geothermal energy research. In this study,

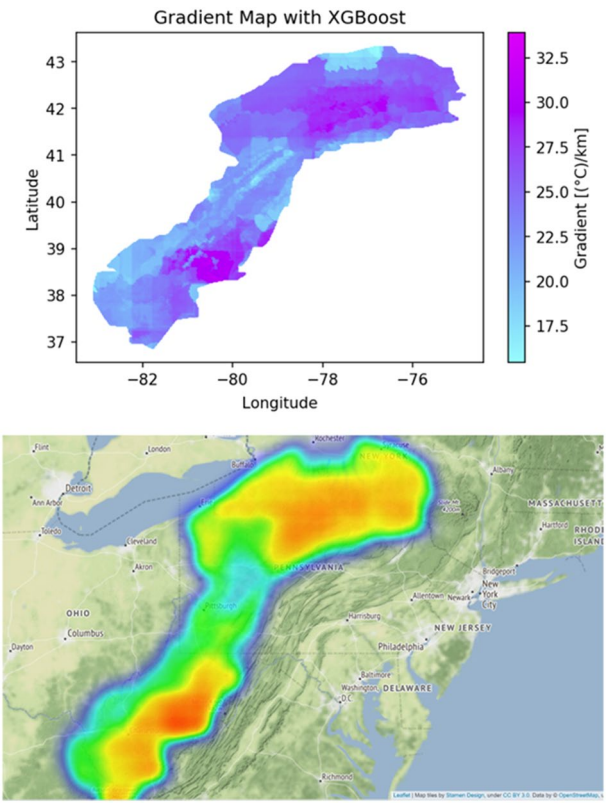

Fig. 12 Geothermal gradient map using XGBoost model. The gradient has the unit of $\frac{^\circ C}{km}$.

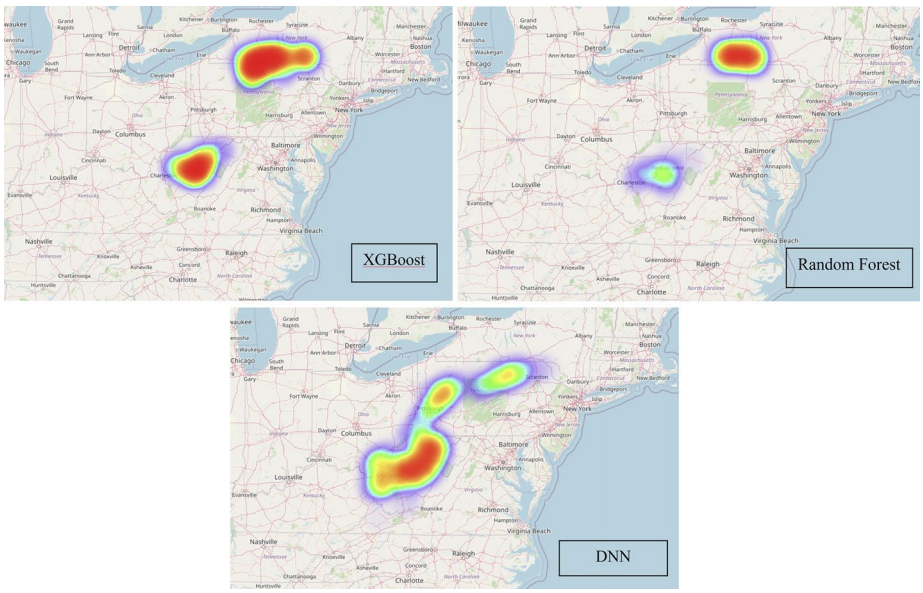

Fig. 13 Regions with subsurface temperature gradient higher than $27\,\frac{^{\circ}C}{km}$ for XGBoost, Random Forest and DNN

Table 6 Average mean absolute errors and standard deviations (with unit of $\frac{^{\circ}C}{km}$ for physics-based, XGBoost and DNN model predictions compared to the measured temperature data

Model	MAE
Physics	6.6
XGBoost	5.6
DNN	7.0

we explored the applicability of four machine learning models in predicting subsurface temperatures in northeastern United States using bottom-hole temperature data and geological information from 20,750 wells. It was shown that XGBoost and Random Forest outperformed all other models, with only 3.21 °C and 3.25 °C mean absolute error. Furthermore, we compared the predictions from machine learning and physics-based models to the measured temperature data obtained from an extra dataset with 58 wells in the state of West Virginia and showed that XGBoost can successfully predict the temperature at different depths. Lastly, we provided a geothermal gradient map for the corresponding region which can be used as a quick tool to calculate the underground temperature at any desired location and depth. In the map, eastern West Virginia along with portions of southwestern New York state show the highest potential.

We believe that this study provides a complementary analysis for geothermal energy exploration for future investments. Furthermore, oil and gas industry can benefit tremendously from this paper too. The presented machine learning models can be incorporated in reservoir and drilling simulators for more accurate subsurface temperature predictions, and consequently, more reliable fluid properties characterization.

Abbreviations
BHT: Bottom-hole temperature; API: Well identification number; DNN: Deep neural network; DOE: Department of energy; KNN: K-nearest neighbors algorithm; ANN: Artificial neural networks; MAE: Mean absolute error; RMSE: Root mean square error; MAPE: Mean absolute percentage error.

Acknowledgements
We thank the departments of Computer Science and Mining and Minerals Engineering at Virginia Tech for their support. This study is partly supported through the Virginia Tech ICTAS JFA award. We ran our codes through the remote server provided by the Physics-Guided Machine Learning (PGML) lab in the Department of Computer Science at Virginia Tech.

Author contributions
AS: Conceptualization, methodology, data curation, writing original draft preparation, software, and validation; SL: software, investigation, visualization, and validation; AK and BN: supervision, writing—reviewing and editing. All authors read and approved the final manuscript.

Funding
This work was funded by the Department of Mining and Minerals Engineering at Virginia Tech with no additional outside funding.

Availability of data and materials
Complete information about the data resources and source codes are provided in a GitHub repository (Shahdi and Lee, 2021). The source codes associated with each of the figures (in the manuscript) and the trained model pickle files are included. We, also, provide the exact locations where we obtained the data which are used in the paper. Finally, we made an instruction video about how to access data and run the models (https://www.youtube.com/watch?v=lc5TMNuvQ-8).

Declarations

Competing interests
We (the authors) declare that there are not competing interests associated with the research.

Author details
[1]Department of Computer Science at Virginia Tech, Blacksburg, VA, USA. [2]Department of Mining and Mineral Engineering at Virginia Tech, Blacksburg, VA, USA.

Received: 27 December 2020 Accepted: 23 June 2021
Published online: 02 July 2021

References

Araya-Polo M, Jennings J, Adler A, Dahlke T. Deep-learning tomography. Leading Edge. 2018;37(1):58–66. https://doi.org/10.1190/tle37010058.1.
Assouline D, Mohajeri N, Gudmundsson A, Scartezzini JL. A machine learning approach for mapping the very shallow theoretical geothermal potential. Geothermal Energy. 2019;7(1):1–50. https://doi.org/10.1186/s40517-019-0135-6.
Barbosa L, Nascimento A, Mathias M, de Carvalho Jr J. Machine learning methods applied to drilling rate of penetration prediction and optimization-a review. J Pet Sci Eng. 2019. https://doi.org/10.1016/j.petrol.2019.106332.
Baruque B, Porras S, Jove E, Calvo-Rolle J. Geothermal heat exchanger energy prediction based on time series and monitoring sensors optimization. Energy. 2019;171:49–60. https://doi.org/10.1016/j.energy.2018.12.207.
Bassam A, Santoyo E, Andaverde J, Herná Ndez JA, Espinoza-Ojeda OM. Estimation of static formation temperatures in geothermal wells by using an artificial neural network approach. Comput Geosci. 2010;36(9):1191–9. https://doi.org/10.1016/j.cageo.2010.01.006.
Beardsmore G. Data fusion and machine learning for geothermal target exploration and characterisation. Technical Report, National ICT Australia Limited (NICTA), Australia; 2014.
Blackwell D, Richards M. New geothermal resource map of the northeastern US and technique for mapping temperature at depth. In Geothermal Resources Council Annual Meeting. 2010. https://www.osti.gov/biblio/1137023. Accessed 27 Dec 2020.
Bloomquist G, Niyongabo P, El-Halabi R, Löschau M. The AUC/KFW Geothermal Risk Mitigation Facility (GRMF)–A Catalyst for East African Geothermal Development. GRC Transactions, 2012; 36(4). https://www.geothermal-library.org/index.php?mode=pubsandaction=viewandrecord=1030213. Accessed 27 Dec 2020.
Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.
Carbonari R, Ton D, Bonneville A, Bour D, Cladouhos T, Garrison G et al. First Year Report of EDGE Project: an International Research Coordination Network for Geothermal Drilling Optimization Supported by Deep Machine Learning and Cloud Based Data Aggregation. Stanford Geothermal Workshop, 3049(7). 2021. https://doi.org/10.1117/12.275844
Chen T, Guestrin C. XGBoost: A scalable tree boosting system. 22nd International Conference on Knowledge Discovery and Data Mining, 785–794. 2016. https://doi.org/10.1145/2939672.2939785
Childs OE. Correlation of stratigraphic units of North America–COSUNA. AAPG Bull. 1985;69(2):173–80.
Cornell University. Appalachian Basin play fairway analysis: thermal quality analysis in low-temperature geothermal play fairway analysis (GPFA-AB). 2015. https://doi.org/10.15121/1261947
Deming D. Application of bottom-hole temperature corrections in geothermal studies. Geothermics. 1989;18(5–6):775–86.

DOE. Toward drilling the perfect geothermal well: an international research coordination network for geothermal drilling optimization supported by deep machine learning and cloud based data aggregation. 2019. https://www.energy.gov/nepa/downloads/cx-101522-toward-drilling-perfect-geothermal-well-international-research-coordination. Accessed 27 Dec 2020.
Dwyer, K. Concave hull—Python code. (n.d.). https://gist.github.com/dwyerk/10561690. Accessed 27 Dec 2020.
Faulds JE, Brown S, Coolbaugh M, Deangelo J, Queen JH, Treitel S, Fehler M, Mlawsky E, Glen JM, Lindsey C, Burns E. Preliminary report on applications of machine learning techniques to the nevada geothermal play fairway analysis. In: 45th workshop on geothermal reservoir engineering. 2020. p. 229–34.
Forrest J, Marcucci E, Scott P. Geothermal gradients and subsurface temperatures in the northern gulf of mexico. GCAGS. 2005;55:233–48.
Frone Z, Blackwell D. Geothermal map of the northeastern United States and the West Virginia thermal anomaly. Geothermal Resources Council, Annual Meeting, 2010, 34, GRC1028668. https://www.osti.gov/biblio/1137024. Accessed 27 Dec 2020.
Gosnold W, Panda B. (2002). The global heat flow database of the international heat flow commission. 2022. https://engineering.und.edu/research/global-heat-flow-database/. Accessed 27 Dec 2020.
Gul S, Aslanoglu V, Tuzen M, Senturk E. Estimation of bottom hole and formation temperature by drilling fluid data: a machine learning approach. 44th Workshop on Geothermal Reservoir Engineering. 2019. https://www.ccs.neu.edu/home/vip/teach/MLcourse/4_boosting/slides. Accessed 27 Dec 2020.
Hall B. Facies classification using machine learning. Lead Edge. 2016;35(10):906–9. https://doi.org/10.1190/tle35100906.1.
Hegde C, Gray K. Use of machine learning and data analytics to increase drilling efficiency for nearby wells. J Nat Gas Sci Eng. 2017;40:327–35. https://doi.org/10.1016/j.jngse.2017.02.019.
Hegde C, Gray K. Evaluation of coupled machine learning models for drilling optimization. J Nat Gas Sci Eng. 2018;56:397–407. https://doi.org/10.1016/j.jngse.2018.06.006.
Hegde C, Pyrcz M, Millwater H, Daigle H, Gray K. Fully coupled end-to-end drilling optimization model using machine learning. J Petrol Sci Eng. 2020. https://doi.org/10.1016/j.energy.2012.06.045.
Ho TK. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell. 1998;20(8):832–44. https://doi.org/10.1109/34.709601.
Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55–67. https://doi.org/10.1080/00401706.1970.10488634.
Jordan T, Richards M, Horowitz F, Camp E. Low Temperature geothermal play fairway analysis for the appalachian basin: phase 1 revised report November 18, 2016. https://doi.org/10.2172/1341349
Kalogirou S, Florides G, Pouloupatis P, Panayides I, Joseph-Stylianou J, Zomeni Z. Artificial neural networks for the generation of geothermal maps of ground temperature at various depths by considering land configuration. Energy. 2012;48(1):233–40. https://doi.org/10.1016/j.energy.2012.06.045.
Keynejad S. Application of machine learning algorithms in hydrocarbon exploration and reservoir characterization. 2018. https://repository.arizona.edu/handle/10150/628470. Accessed 27 Dec 2020.
Khan MA, Raza HA. The role of geothermal gradients in hydrocarbon exploration in Pakistan. J Pet Geol. 1986;9(3):245–58. https://doi.org/10.1111/j.1747-5457.1986.tb00388.x.
Lehmann R. 3σ-rule for outlier detection from the viewpoint of geodetic adjustment. J Surv Eng. 2013;139(4):157–65.
Li C. A gentle introduction to gradient boosting. Boston: Northeastern University; 2016.https://www.ccs.neu.edu/home/vip/teach/MLcourse/4_boosting/slides/gradient_boosting.pdf.
Liaw A, Wiener M. Classification and Regression by RandomForest. R News. 2002;2(3):18–22.
Lukawski M, Silverman R, Tester J. Uncertainty analysis of geothermal well drilling and completion costs. Geothermics. 2016;64:382–91. https://doi.org/10.1016/j.geothermics.2016.06.017.
Ma Y, Ji X, BenHassan N, LuoY. A deep learning method for automatic fault detection. SEG Technical Program Expanded Abstracts 2018. Society of Exploration Geophysicists, 2018; 1941–1945. https://doi.org/10.1190/segam2018-2984932.1
Maind S, Wankar P. Research paper on basic of artificial neural network. IJRITCC. 2014;2(1):96–100.
Moniz N, Branco P, Torgo L. Evaluation of ensemble methods in imbalanced regression tasks. First International Workshop on Learning with Imbalanced Domains: Theory and Applications, 2017; 129–140. http://proceedings.mlr.press/v74/moniz17a.html. Accessed 27 Dec 2020.
Morgül Tumbaz MN, İpek M. Energy demand forecasting: avoiding multi-collinearity. Arab J Sci Eng. 2021;46(2):1663–75. https://doi.org/10.1007/s13369-020-04861-4.
Moses P. Geothermal gradients. Paper Presented at the Drilling and Production Practice, New York, New York, 1961. https://onepetro.org/APIDPP/proceedings-abstract/API61/All-API61/API-61-057/51251. Accessed 27 Dec 2020.
Muhammad AC. Mathematical model of utilization mapping for geothermal energy using machine learning algorithms. 2019. http://103.82.172.44:8080/xmlui/handle/123456789/564. Accessed 27 Dec 2020.
Noshi C, Schubert J. The role of machine learning in drilling operations; a review. SPE/AAPG Eastern Regional Meeting. 2018. https://onepetro.org/conference-paper/SPE-191823-18ERM-MS. Accessed 27 Dec 2020.
Perozzi L, Guglielmetti L, Moscariello, A. Minimizing Geothermal exploration costs using machine learning as a tool to drive deep geothermal exploration. AAPG European Region, 3rd Hydrocarbon Geothermal Cross Over Technology Workshop. 2019. https://www.searchanddiscovery.com/abstracts/html/2019/geneva-90346/abstracts/2019.ER.Geneva.29.html. Accessed 27 Dec 2020.
Polikar R. Ensemble learning. In: Ensemble machine learning (pp. 1–34). 2012. https://doi.org/10.1007/978-1-4419-9326-7_1
Pukelsheim F. The three sigma rule. Am Stat. 1994;48(2):88–91. https://doi.org/10.1080/00031305.1994.10476030.
Rezvanbehbahani S, Stearns LA, Kadivar A, Doug Walker J, Van Der Veen CJ. Predicting the geothermal heat flux in greenland: a machine learning approach. Geophys Res Lett. 2017;44(24):12–271. https://doi.org/10.1002/2017GL075661.
Shahdi A, Lee S. GitHub repository. 2021. https://github.com/seho0808/machine_learning_approach_for_subsurface_temperature_prediction. Accessed 27 Dec 2020.

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. https://doi.org/10.1016/j.apenergy.2020.116046.
Snyder DM, Beckers KF, Young KR. Update on geothermal direct-use installations in the United States. In: Proceedings of forty-second workshop on geothermal reservoir engineering, vol. 42. 2017. p. 1–7.
Stutz GR, Williams M, Frone Z, Reber TJ, Blackwell D, Jordan T, Tester JW. A well by well method for estimating surface heat flow for regional geothermal resource assessment. In: Proceedings of thirty-seventh workshop on geothermal reservoir engineering, Stanford. SGP-TR-194. 2012.
Sun Z, Jiang B, Li X, Li J, Xiao K. A data-driven approach for lithology identification based on parameter-optimized ensemble learning. Energies. 2020;13(15):3903. https://doi.org/10.3390/en13153903.
Tester JW, Anderson BJ, Batchelor AS, Blackwell DD, DiPippo R, Drake EM. The future of geothermal energy—Impact of enhanced geothermal systems (EGS) on the United States in the 21st century: an assessment. Idaho Falls: Idaho National Laboratory; 2006.
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. https://doi.org/10.1007/s11053-019-09596-0.
Vieira A, et al. Characterisation of ground thermal and thermo-mechanical behaviour for shallow geothermal energy applications. Energies. 2017;10(12):2044. https://doi.org/10.3390/en10122044.
Vijay K, Bala D. Predictive analytics and data mining concepts and practice with rapidminer. Amsterdam: Elsevier; 2014.
Watanabe H, Hino H, Akaho S, Murata N. Retrieved Image Refinement by Bootstrap Outlier Test. International Conference on Computer Analysis of Images and Patterns, 11678 LNCS, 505–517. 2019. https://doi.org/10.1007/978-3-030-29888-3_41
West Virginia Geological and Economical Survey Website. (n.d.). https://www.wvgs.wvnet.edu/. Accessed 5 Mar 2020.
Witter J, Trainor-Guitton W, Siler D. Uncertainty and risk evaluation during the exploration stage of geothermal development: a review. Geothermics. 2019;78:233–42. https://doi.org/10.1016/j.geothermics.2018.12.011.
Wyffels F, Schrauwen B, Stroobandt D. Stable output feedback in reservoir computing using ridge regression. International Conference on Artificial Neural Networks, 5163 LNCS(PART 1), 808–817. 2008. https://doi.org/10.1007/978-3-540-87536-9_83
Young KR, Augustine C, Anderson A. Report on the U.S. DOE geothermal technologies program's 2009 risk analysis. 2010. https://digitalscholarship.unlv.edu/renew_pubs/21/. Accessed 27 Dec 2020.
Zhang C, Frogner C, Araya-Polo M, Hohl D. Machine-learning based automated fault detection in seismic traces. 76th European Association of Geoscientists and Engineers Conference and Exhibition 2014: Experience the Energy—Incorporating SPE EUROPEC 2014, 807–811. 2014. https://doi.org/10.3997/2214-4609.20141500

Publisher's Note