PROCEEDINGS, 46th Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 15-17, 2021
SGP-TR-218

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

Rolando Carbonari[1], Dang Ton[2], Alain Bonneville[3], Daniel Bour[4], Trenton Cladouhos[5], Geoffrey Garrison[6], Roland Horne[2], Susan Petty[5], Robert Rallo[3], Adam Schultz[1], Carsten F Sørlie[7], Ingolfur Orn Thorbjornsson[8], Matt Uddenberg[9], Leandra Weydt[10]

[1]Oregon State University; [2]Stanford University; [3]Pacific Northwest National Laboratory; [4]Bour Consulting; [5]Cyrq energy; [6]AltaRock Energy; [7]Equinor; [8]Iceland GeoSurvey (ISOR); [9]Stravan Consulting; [10]Technical University of Darmstadt

E-mail: carbonar@oregonstate.edu, dangton@stanford.edu, alain.bonneville@pnnl.gov, Daniel@bourconsult.com, trenton.Cladouhos@cyrqenergy.com, ggarrison@altarockenergy.com, horne@stanford.edu, susan.petty@cyrqenergy.com, robert.rallo@pnnl.gov, Adam.Schultz@oregonstate.edu, cso@equinor.com, ingolfur.thorbjornsson@isor.is, muddenberg@stravan.co, weydt@geo.tu-darmstadt.de

Keywords: machine learning, data analysis, deep learning, well optimization

ABSTRACT

Drilling optimization may have several definitions, but what they all have in common is the concept of drilling time and well failure reduction, which is of fundamental importance in reducing the overall costs of a geothermal project. To this purpose, an international research coordination network aimed at developing machine learning strategies to improve geothermal drilling efficiency has been established under the support of the EDGE Program of the US Department of Energy (DOE) Geothermal Technologies Office. The EDGE research collaboration involves two US Universities, a DOE National Laboratory and four Geothermal and Oil and Gas companies from several countries (Iceland, Norway, USA).The first year of the project consisted of four major tasks: 1) Data gathering from more than 100 wells from different companies and geothermal fields; 2) Exploratory Data Analysis (EDA) to assess both the quality and the structure of the data, i.e. the presence of gaps, outliers, typos, the correlation between variables and their distribution, and also to define which variables might be more useful for drilling efficiency prediction in such a way that a data format/structure can be defined as a standard for the machine learning procedures; 3) Development of a well data repository for the data products (data, code, analysis workflows and models) developed during the project; and 4) Initial development and testing of machine learning (ML) techniques.

The main findings of the exploratory data analysis and initial machine learning testing can be summarized as follows: *i*) information related to drill bit life cycle and bottom hole assembly are necessary to improve the data clustering as well as to improve the accuracy of machine learning algorithms; *ii*) lithological classifications usually used to describe well cuttings are too specific and idiosyncratic to be useful for machine learning purposes in their raw form; iii) both Random Forest and Deep Learning models were tested. At present, their accuracy in predicting drilling parameters is similar overall, with Deep Learning models slightly outperforming the Random Forest ones as the number of input parameters increases. With regard to idiosyncratic lithological information as appearing in the raw mudlog data, we have tried both dummy encoding and text embedding to encode the lithological information but none of them has resulted in an improvement in the accuracy of the machine learning algorithms in predicting drilling parameters. A new "rock-strength" description needs to be defined for this purpose.

1. INTRODUCTION

Drilling operations account for an overall 30-70 percent of the total costs of a geothermal project (Saleh et al., 2020; Dumas et al., 2013; Finger and Blankenship 2012). Several factors contribute to the driving up of drilling costs such as casing failure, loss circulations, stuck pipe and fast bit wearing, resulting in an increased total drilling time and/or a well failure and abandonment (Saleh et al., 2020; Kruszewski and Wittig, 2018; Teodoriu, 2015; Marbun et al., 2013). Thus, geothermal drilling optimization is of crucial importance to reduce the total cost of geothermal operations and to improve the large-scale deployment of geothermal energy. The variety of the root causes of slow drilling operations has led several authors to tackle this problem from different directions. Some of them focused on the development of new technologies and best practices to reduce the non-productive time (NPT), in particular to prevent casing failures (Salehi et al., 2013, Karimi et al., 2011), mitigate the impact of loss circulations zones and improving the drilling bit's wear resistance (Saleh et al., 2020; Imaizumi et al., 2019; Miyazaki et al., 2019; Raymond et al., 2012). Other authors instead focused on the optimization of the drilling parameters by using both physics-based model and data driven methods (i.e. machine learning and deep learning): in particular, several attempts to optimize the Rate Of Penetration (ROP) and the Mechanical Specific Energy (MSE) have been done (Alali et al., 2020: Sabah et al., 2019; Hedge et al., 2018; Hedge et al., 2017; Basarir et al., 2014). Even though these data-driven approaches yielded promising results in recent years, there are still some limitations on their real-world suitability. Indeed, oftentimes the proposed models are tightly related to small datasets from one or two wells of a single geothermal field, and so they might not be suitable for a different geothermal field or even a new well in the same field. Moreover, these proposed solutions frequently rely on detailed logs data (UCS, pore pressure, gamma-ray), which are not always available when drilling a new well (Alali et al 2020).

The EDGE project aims to build a database of geothermal drilling data coming from a wide range of geologic and operational settings and then to develop a well optimization scheme based on machine learning and deep learning methodologies. The novelty of this project relies on two factors: i) it uses data from different geothermal fields, thus avoiding being context-specific; ii) it creates a continuous

optimization framework for drilling. Thus, by combining active machine-learning methods with incremental learning algorithms, the final optimization scheme will be able to continuously improve as new data are ingested into the database.

This project is structured in two yearlong phases, which in turn consist of several tasks. Here we focus on the first year of the project, which can be summarized in the following tasks: 1) data collection from more than 100 wells from different companies and geothermal fields; 2) development of a well data repository for the data products (data, code, analysis workflows and models) developed during the project; 3) exploratory data analysis (EDA) to assess both the quality and the structure of the data, i.e. the presence of gaps, outliers, typos, the correlation between variables and their distribution; and 4) initial development and testing of machine learning (ML) techniques.

In the following sections, each of these tasks will be discussed separately.

2. DATA COLLECTION AND REPOSITORY

The development of a geothermal well data repository has been the starting point of the first year of the EDGE project. The data gathering relied on both proprietary and public well data. The proprietary wells data have been provided by the EDGE project's industrial partners - AltaRock (USA), Cyrq Energy (USA), Equinor (Norway), Iceland GeoSurvey-ISOR (Iceland)-, while the public data come from the Utah FORGE and Fallon project. At the end of the first year of the EDGE project, data from 113 wells have been collected. The well data have been obtained from different geothermal projects developed over the past thirty years in the USA and Iceland (Figure 1), and thus represent different geological settings.

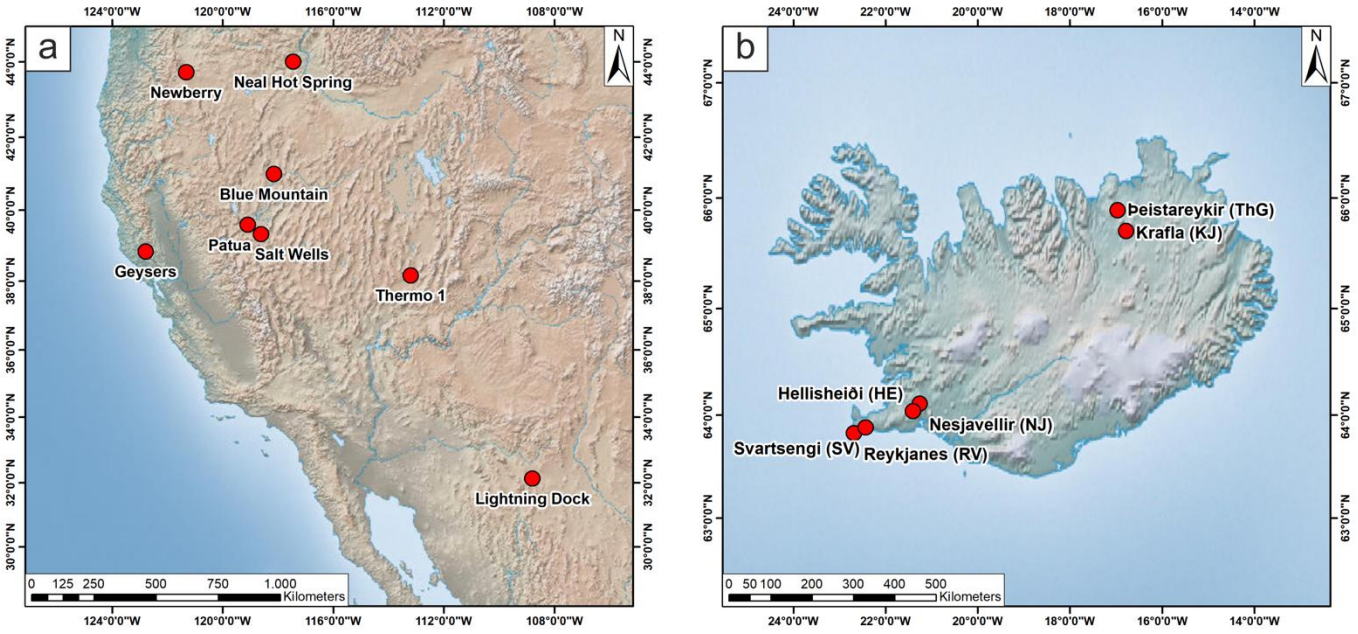

Figure 1: Geothermal fields locations for the US dataset (a) and the Iceland dataset (b).

Once collected, the data have been stored on a virtual machine deployed on Pacific Northwest National Laboratory (PNNL)'s public cloud. The data repository is based on the CKAN platform (Winn 2013) and is structured in different sections. The Raw Data section preserves the original files provided by the project's contributors. Access to this section is read-only to ensure that the original data will not be altered during data analysis tasks. The Processed Data section stores secondary/derived files which are used for analytics and model development and validation. Finally, Analytics and Models sections are used to store other artifacts that are necessary to ensure the reproducibility of the research work. The repository could also allow public access to selected datasets and results. In addition to the web-based UI, the repository implements a data access API that facilitates the direct interaction with analytics workflows.

3. EXPLORATORY DATA ANALYSIS

As mentioned above, (Section 2), at present the developed dataset relies on two main data sources: US data and Icelandic data. These two data groups present different data format, sampling rate and unit of measure. Specifically, the US data, except for the Utah FORGE and Fallon data, are organized into a MySQL database while Icelandic data are contained in Microsoft Excel files. For this reason, the Exploratory Data Analysis (EDA) has been conducted separately on each dataset. In the following sections the main findings of the EDA on each dataset are described.

Table 1: features and units included in the daily bit table.

Feature Name	Feature meaning	Unit
BitNO	Bit Serial Number	Scalar quantity
BitRunNO	Bit runs into the hole	Scalar quantity
BitHrs	Hours of bit usage	Hour [hr]
BitFootage	Drilled footage with one bit	Feet [ft]
BitMudDensity	Mud density	[lbs gal^{-1}]
BitMudFlowAvg	Mudflow	[gals min^{-1}]
BitDrop	Pressure drop	Pound per square inch [psi]
BitPumpPSIAvvg	Pumped pressure	[psi]
BitRPMAvg	Revolution per minute (RPM)	[min^{-1}]
BitROPAvg	Rate of penetration (ROP)	[ft hr^{-1}]
BitTorqAvg	Torque	[lbf ft]
BitWOBAvg	Weight on bit (WOB)	[lbs]
BitHHP	Hydraulic horsepower	[ft lb gal^{-1}]
BitJIF	Jet Impact Force	[lbf]
JetVelocity	Jet velocity	[ft sec^{-1}]
ReportFootage	Daily drilled footage	[ft]
ReportHrs	Daily drilled hours	[hr]
BHANo	Bottom hole assembly ID code	Scalar quantity
BITDiam	Bit diameter	Inch [in]
BHALength	Bottom hole assembly length	[in]
BHAWeight	Bottom hole assembly weight	[lbs]

3.1 EDA US Dataset

The study of the database has shown the presence of two source of drilling data: dailybit and dailydrill tables. The first provides drilling and bit information averaged over a bit life cycle while the second provides only drilling data averaged over one-day of operations. Thus, one record for each day of drilling is provided for the US data, except for those days when a bit has been changed. In this latter case, two records for the same day are present in the dailybit table. The total number of records for the dailydrill and dailybit tables are 4624 and 6608, respectively, both referring to 81 wells and recorded in US customary units. Table 1 shows the features and units of the dailybit table. Several data quality issues such as ambiguities in units and scales, missing data (null data) or invalid data (zero or contradictory data) have been discovered during data evaluation. Thus, extensive pre-processing and filtering for each parameter have been required. For example, by plotting the Kernel Density Estimate to visualize the approximate probability density of the Weight on the Bit (WOB) data, a strong clustering of WOB values is observed around zero (Figure 2a). This skewed distribution was explained by a scaling issue, indeed some WOB values were reported in tons while others in pounds. A more balanced distribution resulted from the re-scaling of the WOB values (Figure 2b). After data cleaning, clustering analysis using Principal Component Analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction techniques have been employed to evaluate whether the provided dailybit or dailydrill dataset are more suitable for machine learning purposes. With these techniques, the dailybit data resulted in a better clustering in the new projection space. This behavior is depicted in Figure 3, showing a comparison between t-SNE clustering analysis results on the dailybit data and the dailydrill data. As it can be observed, the clustering of the data is clearer when the Bit info data are used. These results suggest that the bit information, whenever available, should be incorporated into a standard format with the drilling data.

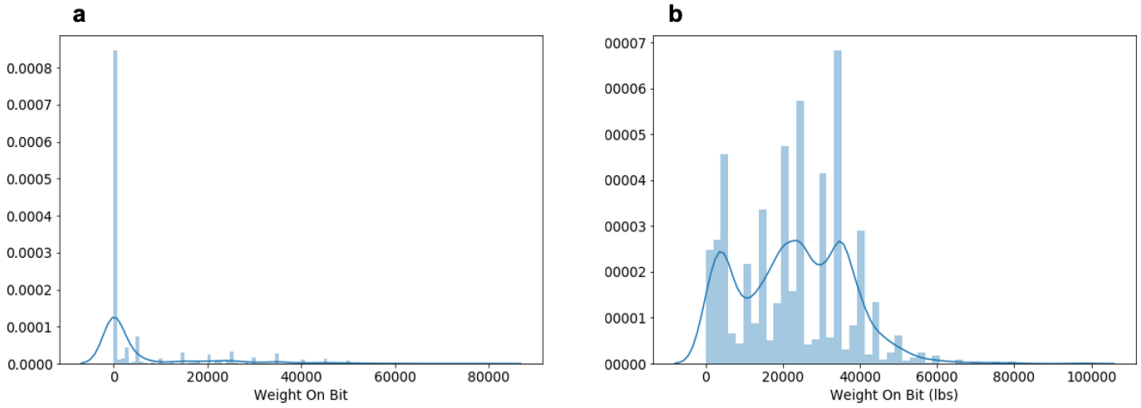

Figure 2: Weight on the Bit (WOB) distribution for the US dataset before (a) and after (b) re-scaling its values.

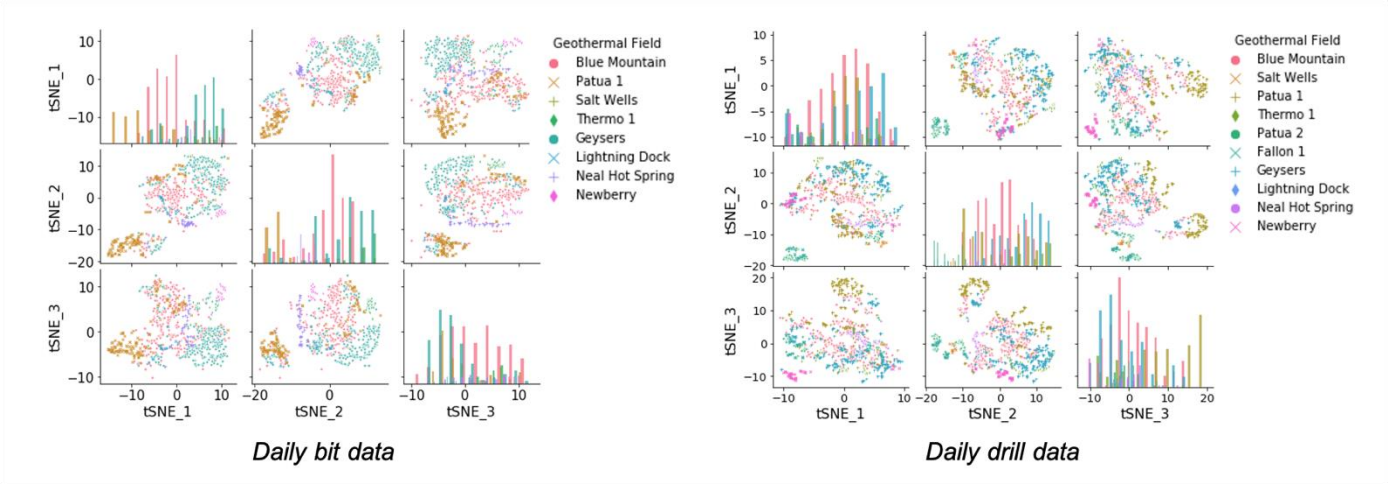

Figure 3: t-SNE plot visualizing the distribution of the data into the new space according. The colors refer to the individual geothermal fields.

3.2 EDA Iceland Dataset

The Icelandic dataset comprises data from 30 wells, each stored in a separate Excel file. In this case, the data have been recorded in the metric system (Table 2) and each record represents an average over 0.5 meter of drilled depth. The total number of records is 126679. The well data quality has also been checked for the Icelandic dataset pointing out the presence of few outliers and anomalous values. In particular, several records with a negative WOB values have been found, which are probably related to reaming or casing operations and not with drilling operations. Thus, these records have not been incorporated when modeling drilling parameters.

Table 2: features and units provided for the Icelandic dataset

Feature Name	Feature meaning	Unit
Depth	Depth	Meters [m]
ROP	Rate of penetration	[m hr^{-1}]
WOB	Weight on bit (WOB)	Tons [t]
TDH	Top drive height	[m]
RPM	Revolution per minute (RPM)	[min^{-1}]
Torque	Torque	[da Nm]
SPP	Stand Pipe Pressure	[bar]
Tot_pump	Pumped fluid flow	[l s^{-1}]
Temp_down	Temperature downhole	[°C]
Temp_ret	Temperature return	[°C]
Diff_temp	Differential temperature	[°C]
Kill_line	High-pressure line	[bar]
Outer_Diam	Outer casing diameter	[in]
Inner_Diam	Inner casing diameter	[in]
Drill_bit_mm	Bit diameter	[mm]

4 PRELIMINARY MACHINE LEARNING MODELS

The last part of the first year of the EDGE project involved the preliminary development and testing of ML models to find out whether or not the collected data were enough to build reliable predictive models. Before starting the development and test of any ML model, the choice of the variables that might be considered as target variables (i.e. features that are related to well drilling performance) as well as the definition of the well performance were thoroughly discussed. The project team has identified four main criteria that should guide the definition of drilling performance: i) minimization of both drilling cost and drilling time; ii) comparison between scheduled and un-scheduled non-productive time; iii) well usability and long-term durability; iv) injectivity or productivity success. At this stage of the study, the project team choose to focus on the Rate of Penetration (ROP) as the target variable. Indeed, among the different drilling parameters, the ROP is strongly related to the efficiency of the drilling process and it is a widely used optimization target in drilling industry (Alali et al., 2020). This is due to both its availability and its connection with working hours and, thus, with the overall drilling cost. However, tying the drilling performance exclusively to the ROP might not be a good choice because, in some geological contexts, its maximization comes at the expense of a faster bit consumption, or increase the chance of stuck pipe in poor circulation conditions, thus driving up the price of the drilling. For this reason, a discussion is currently underway to define other variables to complement the ROP as a measure of drilling efficiency.

Two main approaches have been tested to model the ROP with different level of complexity: decision trees and deep learning. Specifically, a Random Forest model (RF) model and a Dual-branch Dual-path Neural Network (DBDPN) model have been used. Since the US and Icelandic datasets have a different number of features and a different sampling rate, this preliminary analysis has been carried out separately for the two datasets. In the next section, the two algorithms will be described briefly.

4.1 Methodologies and Data Preparation

The original datasets have been preprocessed to eliminate outliers and records unrelated with the drilling ahead of the bit. To this aim, each record with a Null/Zero value for the ROP or WOB has been considered a non-drilling record and discarded before starting the actual ML processes. The Null/Zero values that are not in the ROP or WOB columns are set to 0. Then the dataset is concatenated with a mask. A mask is a matrix that has the same size as the input dataset, filled with 0 and 1: 1 for locations where there is a Null/Zero value and 0 for the rest. This enables the model to work with datasets that originally had Null/Zero values. Furthermore, the DBDPN model also required features' normalization, which ensures a smooth training process. Finally, to avoid overfitting, both the datasets have been divided into a train and test sets in a 1:5 ratio. Thus, 80% of the data are for training purpose and the remaining 20% are for testing. The accuracy of the developed predictive models has been measured through both the root mean squared error (RMSE) and the coefficient of determination (R^2).

4.1.1 Random Forest (RF)

Decision tree methods are a non-parametric supervised learning method based on the recursive partitioning of the data into smaller subsets according to the values of its features. They proved to be really effective in modeling nonlinear data but suffer from over fitting and high variance (Hedge et al., 2017). The Random Forest algorithm, proposed for the first time in the 1990s (Ho, 1995), has proven useful in overcoming these obstacles and has since been used widely in industry. The method relies on an ensemble of decision trees rather than just a single tree. The main characteristic of the RF are bootstrap resampling and random attribute selection. The first allows to build each tree on a partially different training data sample while the second one allows the random selection of a subset of features at each split node of a tree. These two processes result in a diversification and decorrelation of the ensemble trees, allowing a reduction of variance and improvement of the prediction accuracy (Sabah et al., 2019). Three main hyperparameters control the implementation of a RF algorithm: 1) number of trees in the forest; 2) features subset size; 3) bootstrap sample size. In this preliminary work, the RF implemented models have been developed in Python using the scikit-learn package (Pedregosa et al., 2011) with default values of the hyperparameters.

4.1.2. Dual-Branch Dual-Path Neural Network

In recent years, the Dual Path Neural Network (DPN) has been proposed as a novel network architecture to combine the advantages of both Residual Network (ResNet) and Densely Connected Network (DenseNet) – i.e. features re-usage and features re-exploitation (Chen et al., 2017). In DPN, there are two paths that data can flow through: the common data path and the densely connected data path (Figure 4a). The input of each block is the concatenation of both data paths at the block's depth, and the output of each block is composed of the common parts and the densely connected parts. The output's common parts are summed with the common data path, mimicking the behavior of ResNet; while the output's densely connected parts are concatenated with the densely connected data path, mirroring what DenseNet does. This allows DPN to reuse common features with low redundancy while still having the flexibility to learn new features.

Despite its advanced architecture, like most neural networks, DPN performs better if used in conjunction with some form of regularization. For conventional feed-forward neural networks, L1/L2 regularization and Dropout are usually used. However, these conventional techniques do not always work well and can be harmful as they limit the flexibility of the model (Gastaldi, 2017). For this reason, in recent years, the Shake-Shake regularization has been proposed, achieving promising results with ResNet architecture and its derivatives (Gastaldi, 2017).

The Dual-branch Dual-path Neural Network (Figure 4b) is a modification of the DPN architecture that incorporates Shake-Shake regularization and utilizes two branches in each block. By having the strengths from the DPN architecture in conjunction with Shake-Shake regularization, DBDPN should have a superior performance when compared to the plain DPN architecture.

In this study, a 15 blocks DBDPN has been used so far. Each DBDPN block's branch is a conventional two layers feed forward neural network. The network has been trained using Adam optimizer with a learning rate of $3x10^{-4}$. A smoothed L1 loss function (Huber loss) was used with the US dataset, while a L2 loss function with the Icelandic dataset.

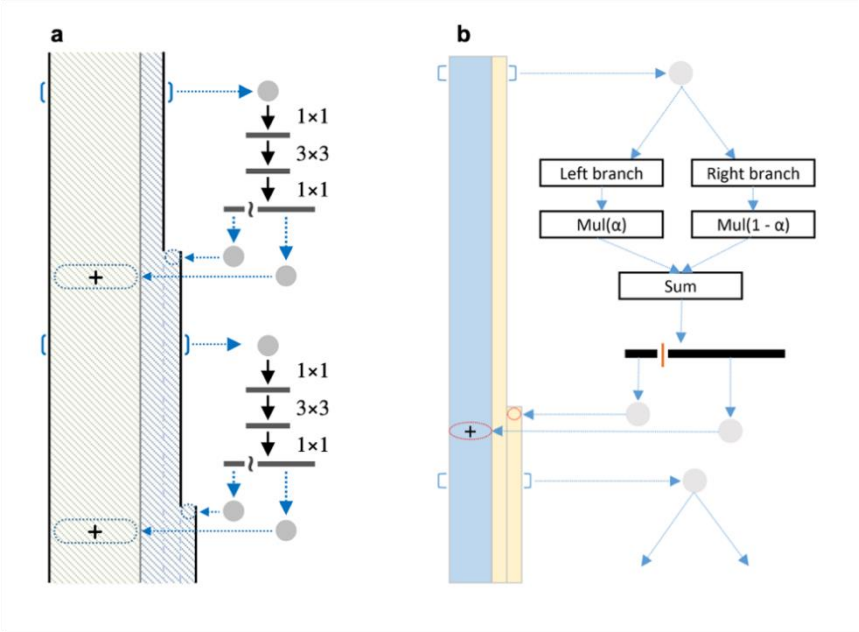

Figure 4: a) Dual-path Network architecture, where the symbol "ʔ" denotes a split operation, and "+" denotes element-wise addition (From Chen et al., 2017); b) Dual-branch Dual-path Neural Network Architecture.

4.2 Results: US Dataset

Both models have been applied to predict the ROP. Starting from the features shown in Table 1, all the features related to drilling time (BitHrs, ReportHrs) and footage (ReportFootage, BitFootage) have not been used as predictors to avoid data leaking because ROP is the ratio between the drilled footage and the corresponding drilling time. Figure 5a shows a cross plot of predicted versus measured values obtained on the test dataset with the RF model. The R^2 score is 0.56 while the RMSE is 6.07 $ft \cdot hr^{-1}$. These two values indicate that the RF model probably is not able to fully grasp the structure of the data. In Figure 5b the feature importance plot is shown. This plot shows which features the RF model is mostly relying upon to predict the ROP. As it can be seen, the BitWOBAvg and BitDiam are actually the most important parameters according to the RF algorithm. Figure 6 shows the cross plot obtained on the test with the DBDPN model. The R^2 score is 0.58 indicating that even this DBDPN is not able to find a clear pattern in the data.

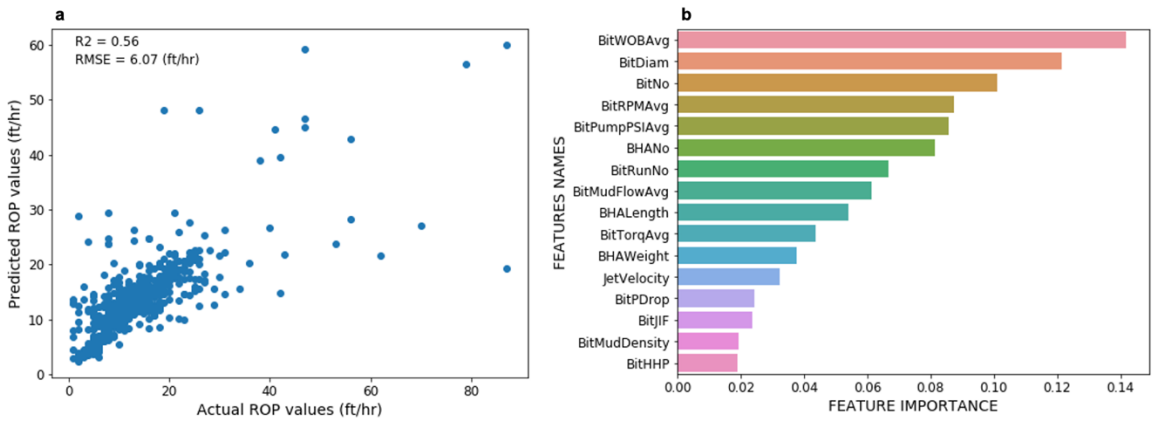

Figure 5: ROP prediction with Random Forest on US dataset. a) cross plot of predicted versus actual ROP values; b) Feature importance plot.

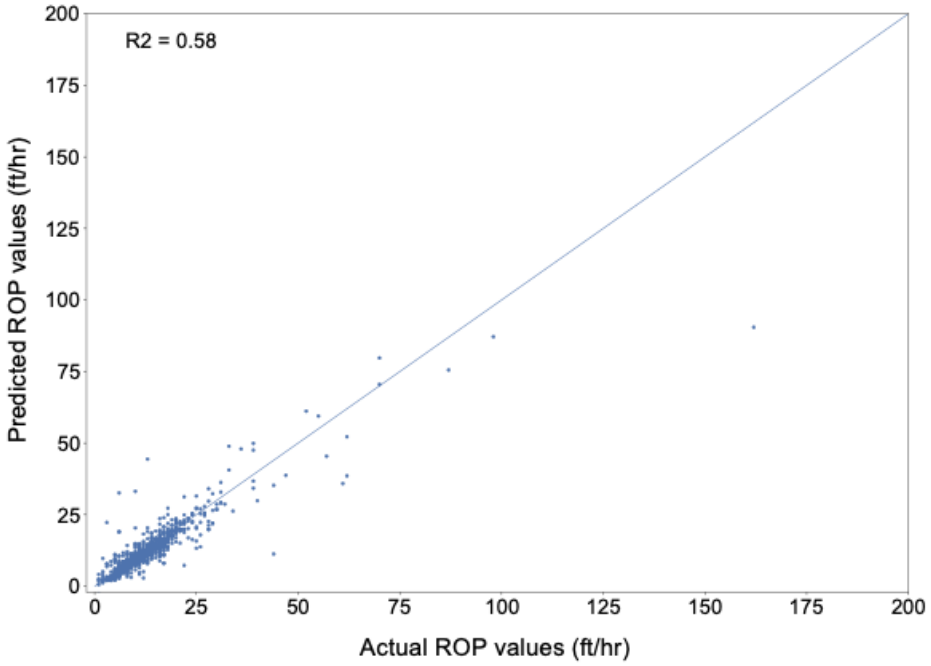

Figure 6: cross plot of predicted versus actual ROP values obtained with DBDPN on US dataset.

The similarity of the two results, obtained with two completely different models, suggests that the actual predictors might not be sufficient to fully characterize the relationship between the drilling parameters and the resulting ROP. The lack of information regarding rock type and rock strength might be crucial, because starting with the same predictors, different ROP can be achieved when drilling rocks with different rock strength. On the other hand, the low prediction accuracy might not be related to the number and kind of features but to the sampling frequency. Indeed, having a daily average data might probably smooth out the complex relationship between the ROP and its predictors.

To address this problem, new features, such as lithological descriptions, were integrated into the ML models and well data with a higher sampling frequency were retrieved. At present, lithological descriptions of the reservoir rocks were retrieved from the original mud logs (cutting descriptions). However, the lithological descriptions were mostly heterogenous including short to very detailed and complex comments, abbreviations, and technical jargon which can't be referred to a specific rock type (Figure 7). Furthermore, the comments do not provide any information on the mechanical behavior of the reservoir rocks. For example, rocks which are similar from a mechanical point of view might have different lithological descriptions or vice versa units of the same rock type might be highly variable with respect to their mechanical behavior due to hydrothermal alteration, brecciation or fracturing. Two different approaches were applied to better interpret those lithological comments for the ML models. The first attempt included manual mapping of individual rock types and classification into boarder groups reflecting their supposed rock strength. For example, after interpreting the lithological comments (e.g. Andesite, Dacite, Tuff) the volcanic rocks were grouped into "soft volcanic deposits" (ash fall deposits, tuff), "volcanoclastic deposits" (pyroclastic deposits, ignimbrites) and lavas, while the latter was subdivided into massive, nonporous and porous lavas (Figure 8a). Additionally, information on brecciation, fracturing or hydrothermal alteration were considered. Subsequently, the mapped information were converted into a binary vector and used as an input feature during the modeling (8b). However, this attempt did not result in a significant improvement of the overall prediction accuracy, most likely due to the complex geologic settings including a high variety of different rock types that can't be depict by using daily average drilling data.

The second approach used Natural Language Processing (NPL) directly applied on the lithological comments. By using a Bidirectional Encoder Representations from Transformers (BERT) NLP model (Devlin et al., 2018), these lithological comments can be transformed into dense vector representation, which can serve as additional input features. The BERT model managed to achieve a F1 score of 0.81 when compared its output vectors with the vectors from the first attempt. This result suggests that an NLP transformer is able to extract meaning from the comments and that it could be probably used to effectively embed lithology comments. The preliminary results obtained with the NLP parsing of the comments however haven't shown a clear improvement in the prediction accuracy.

The embedding of lithology information and the gathering of data with a higher sampling frequency is going to be part of the second year of EDGE project.

WellID	FromDepth	ToDepth	Lithology
14-14	2330	2690	2330-2380 100% Phyllite, 2390 70% Phyllite 30% Quartz Veining, 2400-2600 70-100% Phyllite, 0-30% Quartz Veining, 2610-2690 40-60% Clay 40-50% Clay, 0-10% Quartz Veining
14(11)28 ST1	5124	5190	5124'-5130': 30-70% Granodirorite, 20-30% Felsic/Siliceous Dike, 10-20% Cement, 0-10% Dacite, 0-10% Metal 5130'-5150': 80-100% Granodirorite, 0-10% Dacite, 0-10% Granite 5150'-5190': 20-100% Felsic/Siliceous Dike

Figure 7: sample of lithology comments from the US dataset.

a

Original lithological descriptions	Proposed rock-strength mapping
Tuff, Ash, Ash-Tuff, Lithic-Tuff, Altered Tuff	Soft Volcanic Deposits (SVD)
Pyroclastics, Ignimbrite, Volcanics	Volcanoclastic Deposits (VD)
Basalt, Basaltic, Andesite, andesitic Basalt, Dacite, Rhyolite, Rhyodacite	massive Non-Porous Lava (LNP)
Scoria, vesicular, amygdale, porous	Porous Lava (LP)
Breccia, micro-breccia	Volcanic Breccia (VB)

b

Lithology
100% Andesite,10-60% Dacite; 40–90% Andesite,60-90% Andesite; 10-40% Dacite,80-100% Scoria & Basalt (5800 40% Tuff)

SVD	VD	LNP	LP	VB
1	0	1	1	0

Figure 8: a) Example of the proposed mapping. Different lithological descriptions are mapped into broader categories according to their supposed rock-strength. In this case the mapping only for the volcanic rocks is shown; b) example showing the conversion of a lithology comment into a binary vector to be used into ML models.

4.3 Results: Iceland Dataset

The predictors used with the Icelandic dataset are the one shown in Table 2. Figure 9a shows the cross plot obtained on the test dataset with the RF model. The R^2 score is 0.77 while the RMSE is 4.59 $m \cdot hr^{-1}$. In this case, the RF model is able to provide a good accuracy. Figure 9b shows that TotPump, SPP and WOB are the most important parameters to predict the ROP, which is an expected behavior because these three parameters influence the ROP directly. In order to verify the robustness of both the RMSE and R^2 score achieved with the RF model, a 5-fold cross-validation with 10 repeats has been implemented and the corresponding distributions of these two parameters are depicted in Figure 10. The R^2 distribution (Figure 10a) is centered around 0.77 with all the values ranging from 0.74 to 0.79 while the RMSE (Figure 10b) distribution is centered around 4.7 $m \cdot hr^{-1}$ with all the values ranging from 4.44 to 5 $m \cdot hr^{-1}$. Figure 11 shows the cross plot obtained on the test with the DBDPN model. The R^2 score is 0.75 while the RMSE is 5.03 $m \cdot hr^{-1}$ indicating that the DBDPN model also provides a good accuracy with the Icelandic dataset.

The overall higher accuracy obtained with the Icelandic dataset can be ascribed to both the higher sampling frequency of this dataset (one record each 0.5 meter of drilling versus one record for each drilling day) and the roughly homogeneity in the drilled lithologies (predominantly basaltic lavas). Further analysis will help to better define the source of this difference in the prediction accuracy.

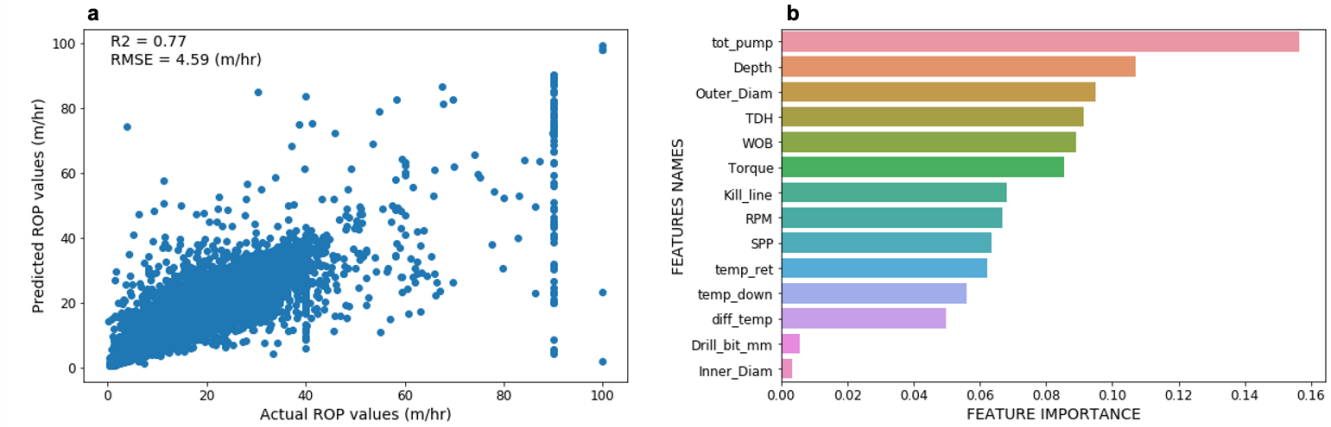

Figure 9: ROP prediction with Random Forest applied on the Icelandic dataset. a) cross plot of predicted versus actual ROP values; b) Feature importance plot.

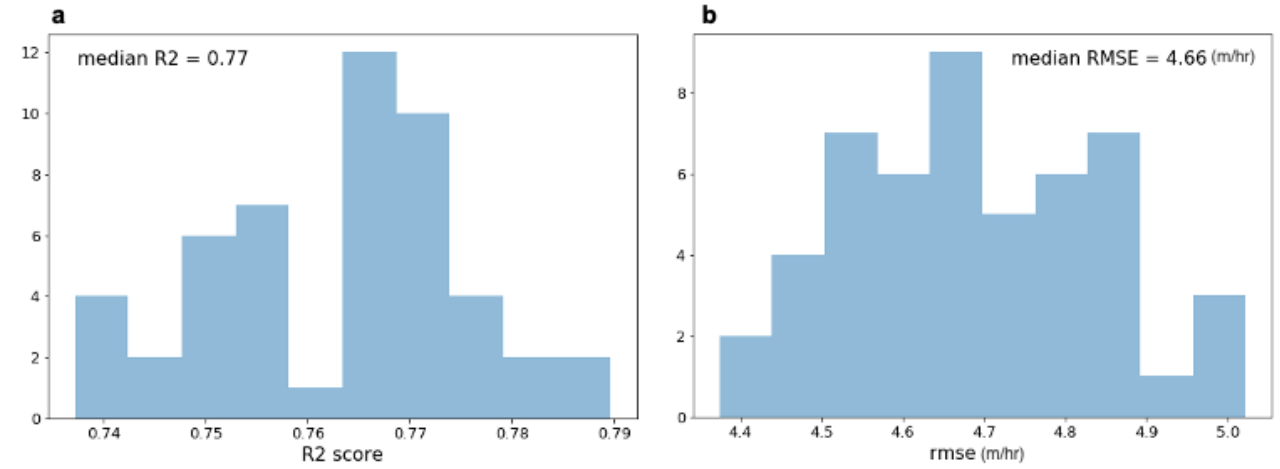

Figure 10: R2 score (a) and RMSE (b) distribution obtained with a 5-fold cross-validation with 10 repeats.

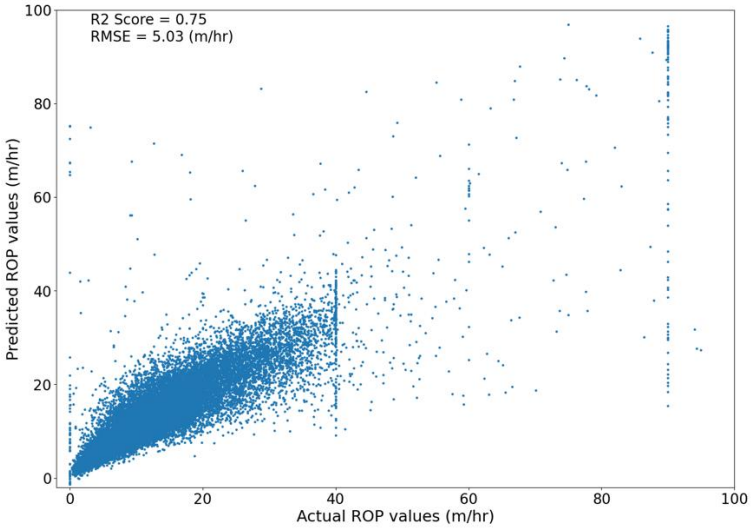

Figure 11: cross plot of predicted versus actual ROP values obtained with DBDPN on US dataset.

5 CONCLUSIONS

The EDGE project aims to build a database of geothermal drilling data and to develop a well optimization scheme based on machine learning and deep learning methodologies. The first part of the project concerned the collection of drilling data from different industrial project contributors. Data retrieved from 113 wells, representing different geological settings, were stored into a data repository created and deployed on the PNNL public cloud. So far, the database includes two data sets, the US and Icelandic dataset, which differ in terms of units, number of features and sampling frequency. After data evaluation and cleaning, preliminary machine learning models were developed to assess the feasibility of building predictive models on the dataset. The first prediction models focused on the rate of penetration (ROP), since this parameter is strictly related to the drilling time and, thus, to the overall drilling cost. Thereby, two main models –the Random Forest and the Dual-branch dual-path neural network – were explored. Both models achieved a good accuracy when trained on the Icelandic dataset (RMSE~4.8; R2~0.76), while being less accurate when applied on the US dataset (RMSE ~6, R2~0.56). This difference in the prediction accuracy can be ascribed to several factors including the different sampling frequency of the two datasets (daily averages versus data provided per 0.5 m) and the higher homogeneity in the drilled lithologies for the Icelandic dataset (mainly basaltic lava). In order to improve the prediction accuracy, lithological information retrieved from mud logs were integrated into the machine learning process. However, the automatic interpretation of lithological information, which are usually stored as long comments or abbreviations, remains challenging. A classification into boarder mechanical categories based on laboratory rock strength data will be tested in the future.

The second year of the project will be devoted to: i) the definition of new target variables to complement the ROP in the optimization of drilling operations; ii) the merging of the two different datasets; iii) the development and testing of machine learning models with the new target variables and with the lithological information; iv) the development of a well failure analysis and v) the deployment of the developed models into the PNNL public cloud.

ACKNOWLEDGEMENTS

This research has been conducted in the framework of the EDGE project supported by the Department of Energy under Award Number DE-EE0008793 – 0000.

REFERENCES

Alali, A. M., Abughaban, M. F., and Aman, B. M.: Hybrid Data Driven Drilling and Rate of Penetration Optimization, Journal of Petroleum Science and Engineering, 108075, (2020).

Basarir, H., Tutluoglu, L., and Karpuz, C.: Penetration rate prediction for diamond bit drilling by adaptive neuro-fuzzy inference system and multiple regressions, *Engineering Geology*, 173, 1-9, (2014).

Chen, Y., Li, J., Xiao H., Jin X., Yan S. and Feng J.: Dual Path Networks, *arXiv:1707.01629 [cs.CV]*, (2017).

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805*, (2018).

Dumas, P., Antics, M., and Ungemach, P.: Report on geothermal drilling. *Co-Funded by the Intelligent Energy Europe Programme of the European Union*, (2013).

Finger, J., and Blankenship, D.: Handbook of best practices for geothermal drilling, *Sandia National Laboratories*, Albuquerque, NM (2010).

Gastaldi X.: Shake-Shake regularization, *arXiv:1705.07485 [cs.LG]*, (2017).

Hegde, C., Daigle, H., and Gray, K. E.: Performance comparison of algorithms for real-time rate-of-penetration optimization in drilling using data-driven models, *SPE Journal*, 23(05), 1-706, (2018).

Hegde, C., Daigle, H., Millwater, H., and Gray, K.: Analysis of rate of penetration (ROP) prediction in drilling using physics-based and data-driven models, *Journal of Petroleum Science and Engineering*, 159, 295-306, (2017).

Ho, T. K.: Random decision forests, *Proceedings of 3rd international conference on document analysis and recognition*, (Vol. 1, pp. 278-282), IEEE, (1995).

Imaizumi, H., Ohno, T., Karasawa, H., Miyazaki, K., Akhmadi, E., Yano, M., Miyashita, Y., Yamada, N., Miyamoto, T., Tsuzuki, M., Kubo, S., and Hishi, Y.: Drilling Performance of PDC bits for Geothermal Well Development in Field Experiments, *Proceedings*, 44th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, CA (2019).

Karimi, M., Moellendick, T. E., and Holt, C.: Plastering effect of casing drilling; a qualitative analysis of pipe size contribution, *SPE Annual Technical Conference and Exhibition*, Society of Petroleum Engineers, (2011).

Kruszewski, M., Wittig, V.: Review of failure modes in supercritical geothermal drilling projects, *Geotherm Energy*, 6, 28, (2018).

Marbun, B., Aristya, R., Pinem, R. H., Ramli, B., Gadi, K. B., and Ganesha, J.: Evaluation of Non Productive Time of Geothermal Drilling Operations–Case Study in Indonesia, *Proceedings*, 38th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, CA (2013).

Miyazaki, K., Ohno, T., Karasawa, H., and Imaizumi, H.: Performance of polycrystalline diamond compact bit based on laboratory tests assuming geothermal well drilling, *Geothermics*, 80, 185-194., (2019).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J.: Scikit-learn: Machine learning in Python, *the Journal of machine Learning research*, 12, pp.2825-2830, (2011).

Raymond, D. W., Knudsen, S. D., Blankenship, D. A., Bjomstad, S., Barbour, J., Drilling, B., Schen, A., and Downhole, N. O. V.: PDC Bits Outperform Conventional Bit in Geothermal Drilling Project (No. SAND2012-7841C), *Sandia National Lab(SNL-NM)*, Albuquerque, NM (2012).

Sabah, M., Talebkeikhah, M., Wood, D. A., Khosravanian, R., Anemangely, M., and Younesi, A.: A machine learning approach to predict drilling rate using petrophysical and mud logging data, *Earth Science Informatics*, 12(3), 319-339, (2019).

Saleh, F. K., Teodoriu, C., Ezeakacha, C. P., and Salehi, S.: Geothermal Drilling: A Review of Drilling Challenges with Mud Design and Lost Circulation Problem, *Proceedings*, 45th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, CA (2020).

Salehi, S., Mgboji, J., Aladasani, A., and Wang, S.: Numerical and Analytical Investigation of Smear Effect in Casing Drilling Technology: Implications for Enhancing Wellbore Integrity and Hole Cleaning, *SPE/IADC Drilling Conference*, Society of Petroleum Engineers, (2013).

Teodoriu, C.: Why and when does casing fail in geothermal wells: a surprising question, *Proceedings*, world geothermal congress, (2015).

Winn, J.: Open data and the academy: an evaluation of CKAN for research data management, (2013).