PROCEEDINGS, 45th Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 10-12, 2020
SGP-TR-216

Assessing the Relation between Petrophysical and Operational Parameters in Geothermal Wells: A Machine Learning Approach

Raj Kiran, Saeed Salehi[1]

Well Construction Technology Center, The University of Oklahoma, Norman, OK, USA

[1]salehi@ou.edu

Keywords: drilling engineering, deep learning, filtering, machine learning, real-time monitoring

ABSTRACT

Drilling is a critical operation in the geothermal wells to unlock the potential resources. However, considering the amount of data generated, it is very difficult to track and detect anomalies in the operational domain. In this paper, we have investigated the different data mining algorithms that can capture the pattern in the operational parameters considering the other petrophysical properties. In addition, a suitable framework is proposed for assessing the patterns in the operational system.

We used the FORGE well log data of already drilled wells and synthesize the evolution of dynamic data with a 5-second interval. Now, on each segment of the data, the suitable algorithms were implemented to identify and remove the effect of operational parameters using a series of digital filtering techniques. Then, the filtered version of well logs was used as input for unsupervised machine learning algorithms such as k-nearest neighbors, decision tree classification, and deep learning models with hidden layers. Finally, hazardous zones are classified using the classifications, which can improve the confidence in the operation. Such classifications can be an invaluable tool as it is difficult to identify and classify the anomalies visually from the raw operational data. Overall the proposed framework can significantly improve the drilling operation in geothermal wells and can be further extended for real-time monitoring systems, which are highly exhaustive jobs.

1. INTRODUCTION

During the last decades, the pursuit of unlocking the geothermal reservoirs is on the uprise. The geothermal reservoirs are the source of natural energy, which has the potential to mitigate the environmental concerns synonymous with oil and gas exploration. However, the technologies implemented in making these geothermal reservoir fields suitable for producing the energy are still not fully developed and has a vast scope of improvement. One of the vital operations before establishing any geothermal plant is to drill through the high temperature zones. These zones accumulate the thermal energy, which is transferred from the underground level to the surface to generate electricity. However, the drilling in these reservoirs encounters severe problems such as lost circulation. Studies suggest that more than 10 percent of the average drilling costs are accrued due to lost circulation problems. Snyder et al. (2019) indicated that 89 percent of lost circulation in SURGE well drilling accounted for the non-productive time in the longest section of geothermal drilling. Hence, one of the key points to make drilling efficient in such formations is to identify the problematic zones in the real-time and suggest the conducive parameters for drilling operations. With the advancement of technology and computational prowess, data mining is a significant avenue to deal with classifying these zones based on parameters in real-time. Here, we used real field data from FORGE well (21-31) drilled in Nevada to develop a workflow that can identify these problematic zones.

The geothermal industry uses oil and gas technology for exploration, drilling, and completion. However, a typical geothermal well costs significantly more than a conventional oil and gas well, which might be attributed to the size of the well. However, there is more than just increased size driving the higher costs associated with geothermal wells. These formations comprise of hard rock and are highly fractured. The highly fractured formation can result in high permeable formation. Due to the fractured and highly permeable formations encountered during the geothermal drilling, lost circulation is a more significant problem than in the sedimentary regions typical of oil and gas drilling. The fluid loss may be slow, one to two barrels per hour, or it may be severe, no mud returns to the surface regardless of the pumping rate. When fluid losses become large enough, the mud can no longer adequately perform its intended functions, resulting in severe consequences up to and including loss of the well. With massive losses, the cost of make-up fluids can become prohibitive. Even if the value of make-up fluids can be justified, the unsolved, lost circulation problem can prevent other drilling and completion activities from being performed. Finally, these problems are further aggravated and accelerated by the high temperatures in the formation.

1.1. The FORGE Geothermal Field

The Fallon FORGE (Frontier Observatory for Research in Geothermal Energy) EGS site is in Churchill County, Nevada, which has been used for testing and improving technologies under the US Department of Energy initiative for research in geothermal energy. A well was constructed in this field to study and testing of geothermal exploration reservoirs, as shown in Figure 1. The well site is located in west-central Nevada, which is leased to Ormat Nevada, Inc. This Forge well (21-31) was drilled to approximately 6100 ft depth. Several lithologies were encountered during the drilling. These lithologies included quaternary sediments, quaternary-tertiary basin-fill sediments, Miocene mafic volcanics, altered basalt, quartz monzonite intrusions, quartzite, and altered-rhyolite (Kraal, 2018). Furthermore, the Fallon formation shows the signature of various minerals. Some intervals show the presence of the Illite/mica mineral-rich in phengitic/Fe, while others show the Fe, Mg, and Fe-Mg based chlorite compositions. Apart from these minerals, calcites and Kaolinite are observed at this location (Kraal, 2018). The minerals showed a trend from low-temperature alteration to high temperature alteration, which can also be crucial for the identification of lost circulation zones. However, temperature logs were not accounted for in

our machine learning workflow.

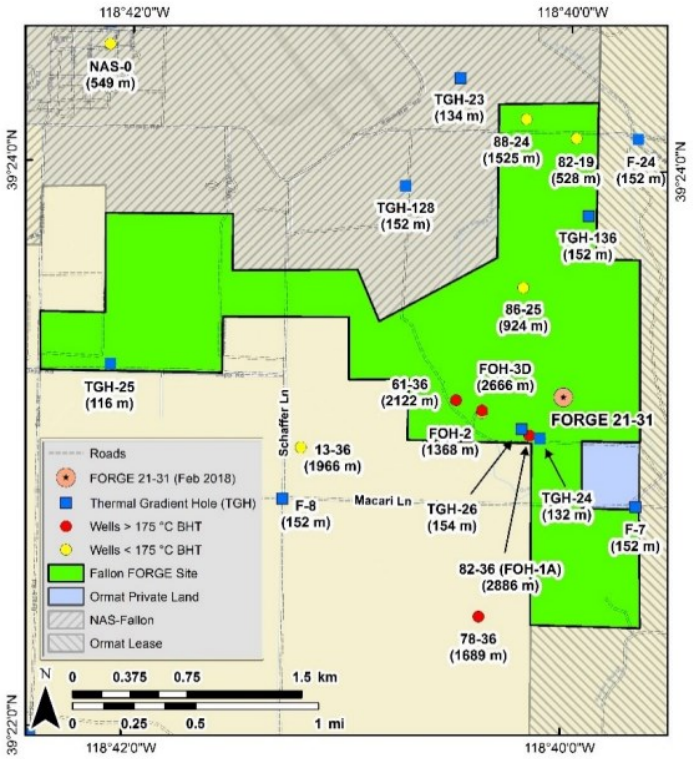

Figure 1. FORGE Well Site Nevada (After Delwiche et al., 2018).

2. METHODOLOGY

2.1. Existing data

Data collected during the drilling contained several logs and geological properties and is available at the geothermal data repository. The data used in this study included the real-time drilling data, including the rate of penetration, rotary, speed, weight on bit, pump pressure and torque for depth up to 6058 ft. The initial data of 130 ft is discarded in this study as the logging data were not recorded up to this height and the default values were reported for several of these parameters. The active pit gain was reported in the well log. This dynamic pit gain was transformed in lost circulation based on the definition of loss of fluid to the formation. The fluid loss of more than 50 barrels were considered as lost circulation zones, which are depicted with a purple line in the log data shown in Figure 2.

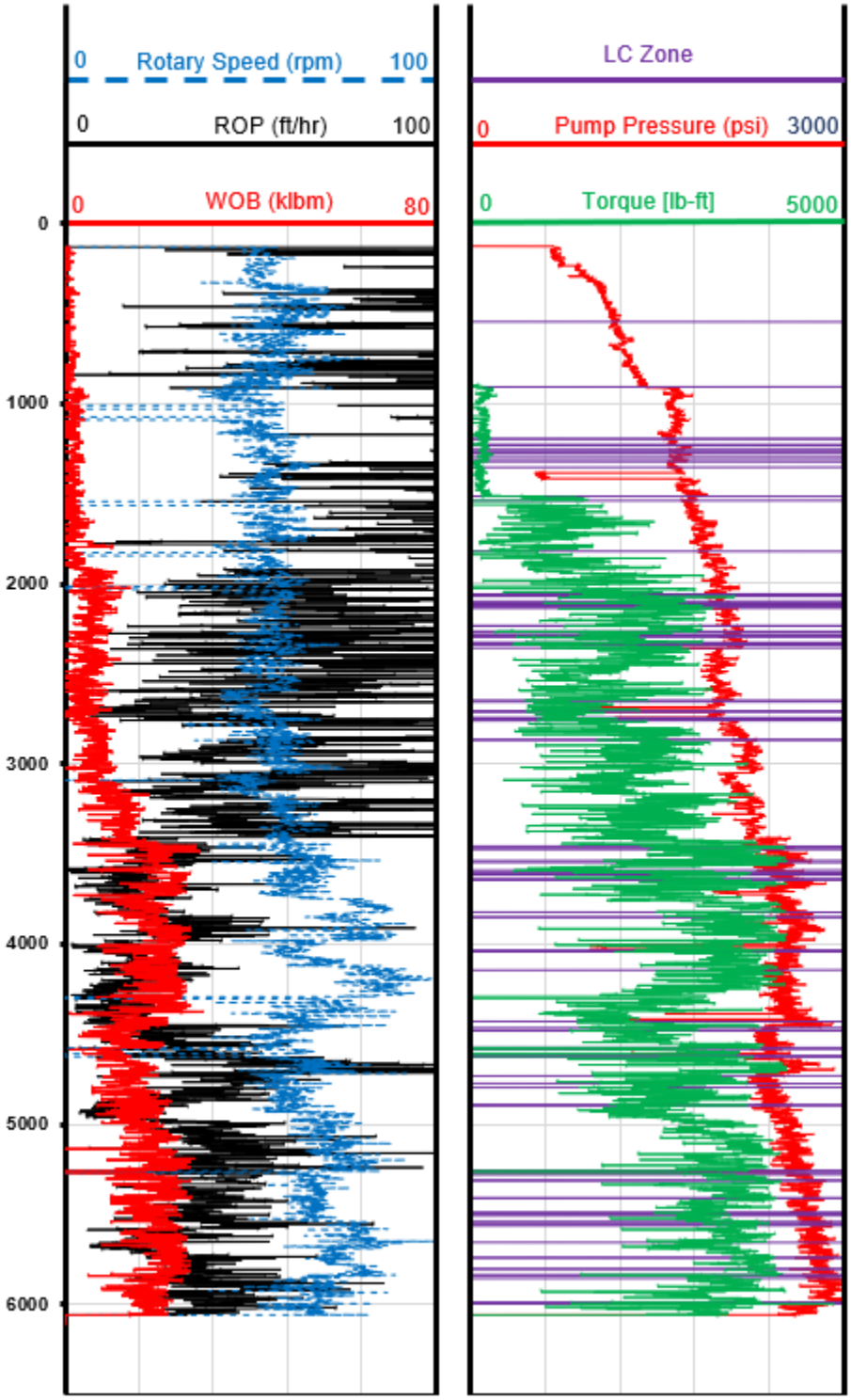

Figure 2. Drilling data for FORGE 21-31 Well (after Blankenship et al., 2018).

2.2. Data Preprocessing

The FORGE well 21-31 logging data was used in this analysis. To start with, the data was first classified into lost circulation and non-lost circulation zones based on the pit gain from well log. This classification was assigned the Boolean value (0 and 1) based on this problem. Now, with this classification, the data corresponding to different logging parameters was normalized based on the full dataset. Finally, this normalized data were divided into random 70:30 split with their classes associated with it. Seventy percent of this data was used in the training of the model, whereas 30 percent of the total data was for evaluation of the model. The log data was first treated with a windows-based filter, and this filtered log was subjected to machine learning and deep learning algorithms and was used to assess the accuracy of these models. The workflow for this classification is explained in the following flow chart:

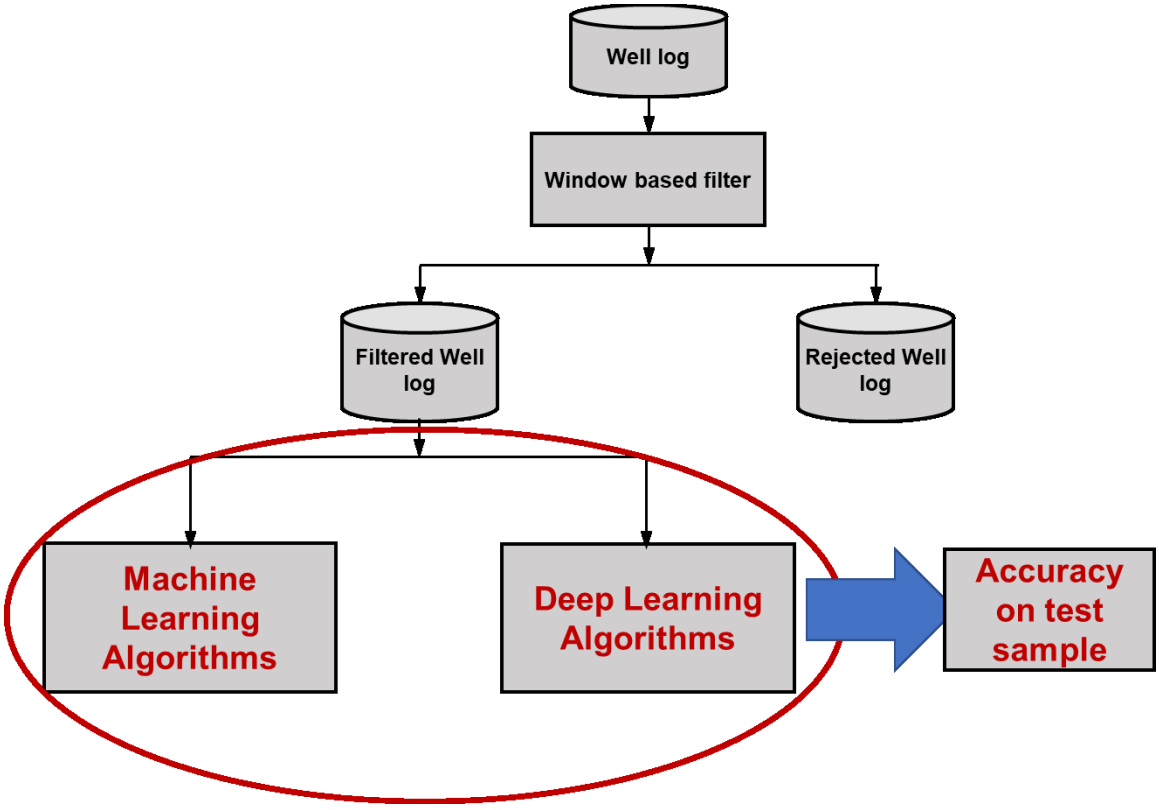

Figure 3: Workflow followed in this study. We feed the well log into a preprocessing window-based filter. The filtered log is then used for machine learning and deep learning. The accuracy of these models is then tested on the randomly selected test data.

The well log traces were used for pre-processing to remove the noise from the data. In this study, we adopted the sliding window-based filter. The main reason behind using this filter is to retain the spikes and maximum flooding surfaces, which is not maintained in traditional despiking. The maximum flooding surfaces have the finite thickness and are used by geologists to guide their interpretation. Hence, these attributes are necessary to keep the geological attributes related to log traces (Saurabh et al., 2019). These attributes can be a valuable feature for identifying the lost circulation zone, as geology is crucial to such problematic zones. Zapata et al. (2016 and 2017) demonstrated the reservoir performance in a conventional simulator. Sinha et al. (2016 and 2018) demonstrated the importance of multivariate and conventional techniques to quantify the significance of multiple variables on well performance.

3. MACHINE LEARNING ALGORITHMS

Machine learning algorithms are widely used for the classification of patterns embedded in the datasets. In this study, we have used four types of supervised learning-based approaches to assess the accuracy of each algorithm that can quantify the lost circulation zones from FORGE 21-31 well data. These models include K-nearest neighbor, decision tree classification, random forest classification, and gradient boosting classifications.

3.1. K-nearest Neighbor (kNN) Algorithm

The k-nearest neighbors (kNN), one of the oldest and simplest methods for classification when implemented with prior domain knowledge in which the rule classifies each unlabeled set by a majority among its k-nearest neighbors in training set. The robustness of the classifier depends on the metric distance used to identify the nearest neighbors (Cover and Hart, 1967). kNN identifies the k nearest neighbors of the test instance where the label sets of its neighboring instances are obtained. Each classified sample neighbor is considered as a piece of evidence corresponding to the class of that pattern. Denoeux (1995) used k-nearest neighbor classification in combination with Dempster's rule of combination to test the effectiveness of the classification strategy. In this study, a simple approach has been adopted to train the data set and then predict the accuracy of the remaining data. The accuracy of the training set and the testing set was found out to be 87 and 81 percent, respectively.

Furthermore, based on the number of nearest neighbors, the accuracy of the testing set is shown in figure 4. The highest accuracy can be observed corresponding to 5 closest neighbors.

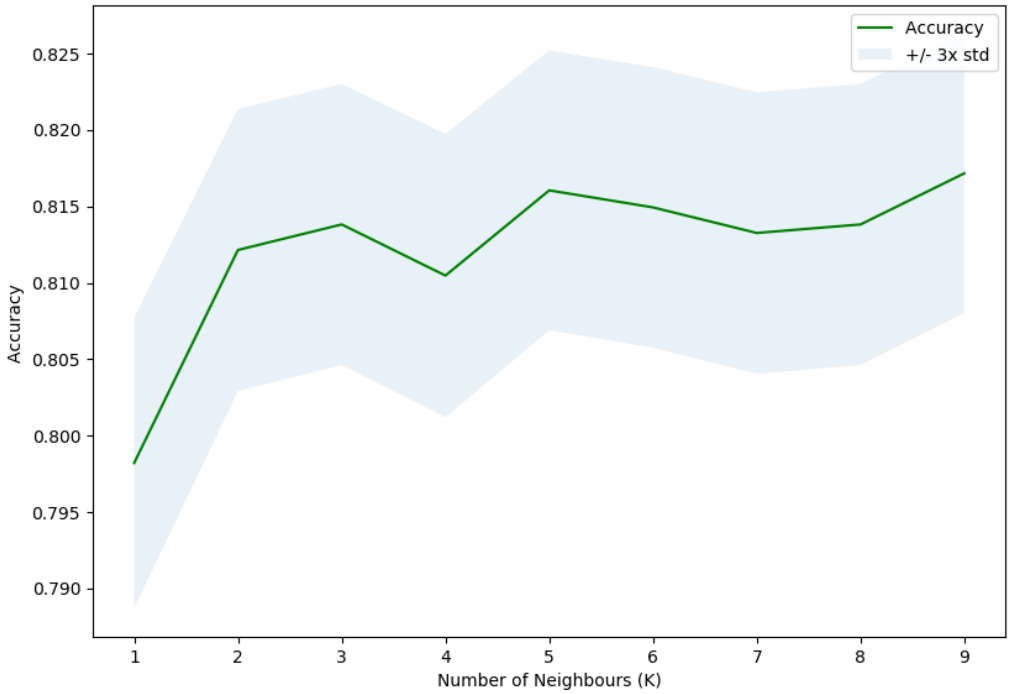

Figure 4. K-nearest neighbor algorithm accuracy vs. number of neighbors

3.2. Decision Tree Classifier

Machine Learning techniques have been an attractive avenue to capture the patterns using the training of the input data and generate classifiers to detect similar patterns in the new data set. Stein et al. (2007) used genetic algorithms to select a subset of input features for decision tree classifiers for detection in intrusion rates in a network. The model showed that the resulting decision trees could have better performance than those built with all available features. The use of decision tree classifiers goes back to the past several decades for a multistage classification strategy (Swain and Hauska, 1977). Recently, such classifiers have found its relevance in detecting anomalies in EEG signals in combination with other algorithms such as fast Fourier transformation. Polat and Gunes (2007) used FFT to extract features from EEG signals and used decision tree classifiers in conjunction with k-fold cross-validation, classification accuracy, and sensitivity values for the design of the intelligent diagnostic system. In this study, we implemented a similar strategy of ensemble decision tree classifiers with 10-fold cross-validation to classify the lost circulation zone. The accuracy of the test set was found to be 0.75. In comparison to other methods, the accuracy of this method is on the lower side.

3.3. Random Forest Classification

The random forest classification is a type of ensemble learning algorithm which is developed on the premise of better performance by a set of classifiers than by an individual classifier (Breiman, 1996; Dietterich, 2000). Breiman (2001) proposed the construction of a set of classifiers in which each classifier casts a single vote for the assignation of the most frequent class to the input vector. Such a system of classifier shows better efficiency on larger datasets with a larger set of input variables, which generates an internal unbiased estimate of generalization error and computes the distances between pairs of cases useful for locating outliers (Rodriguez-Galiano et al., 2012). It uses the randomly selected features or combination of features to grow the tree at each node. Then, it uses the attribute selection measures such as Gini Index, gain-ratio, and Chi-square to design the decision tree. In this study, we have implemented the Gini Index, which can be defined by the following equation:

$$G = \sum \sum_{j \neq i} (f(C_i, T)/|T|)\left(f\left(C_j, T\right)/|T|\right)$$

Where T is the training set and $(f(C_i, T)/|T|)$ is the probability that a selected case belongs to class C_i. After using a combination of features, the decision tree is grown up to its maximum depth. RF can be used to assess the relative importance of different features. However, we do not have a large number of features in this dataset. Hence, we focus on the accuracy of our test data based on the maximum depth and number of trees in the forest. Figures 5 show the accuracy of the classifier, which increases with the increase in the number of trees. Pal (2005) used RF for classifying the types of land covers and found out that the classification is independent of the number of trees. However, in this case, we observed out that the accuracy increases with an increase in the number of trees and approaches a steady state after the number of trees becomes 40. It can be inferred from figure 5 that the maximum accuracy attained using the RF classification is 85 percent, which is better than that of the decision tree classifier.

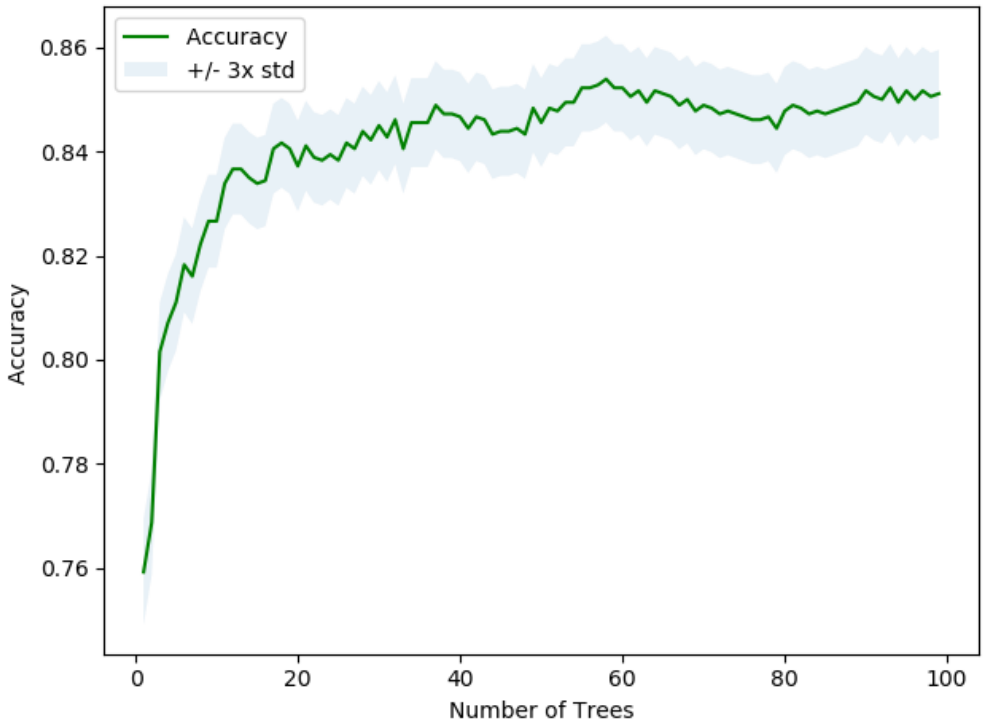

Figure 5. Random Forest Accuracy based on number of trees

3.4. Gradient Boosting Classification

Gradient boosting is an effective classifier constructed using an additive model in a forward stage-wise fashion, which predicts in an ensemble of weak models based on decision trees (Friedman, 2002). It combines individual weak complementary classifiers sequentially where a new weak learner is constructed to provide maximum correlation with the negative gradient of loss functions at each stage of iterations (Natekin and Knoll, 2013). Usually, logistical regression is used as the loss function. Then, a decision tree is used to make a prediction based on a series of rules which consist of different nodes. The extreme gradient boosting speeds up tree construction and uses a new algorithm for tree searching (Torlay et al., 2017). The accuracy of such models is highly dependent on the number of boosting stages, which is resistant to the over-fitting phenomenon.

Several authors have reported that the larger number of boosting stages results in better performance of the model. Hence, we plotted the number of boosting stages to the accuracy of the test data set for lost circulation prediction zones, as depicted in Figure 6. Figure 6 suggests that the accuracy increases with an increase in the boosting stages, and it becomes asymptotic at approximately 400 iterations. The maximum obtained accuracy for this algorithm was 83 percent.

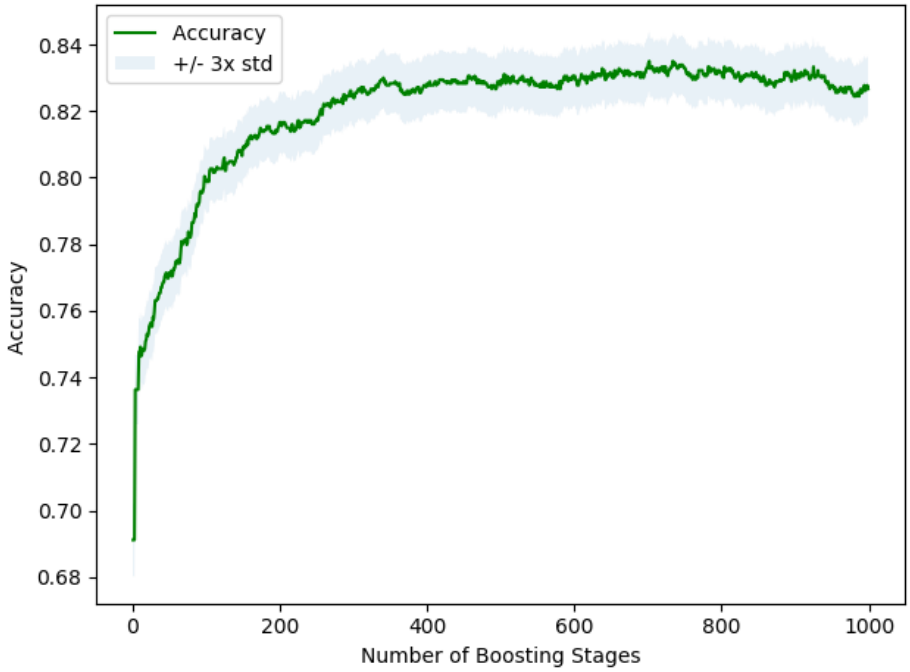

Figure 6. The test set accuracy vs. the number of boosting stages for lost circulation prediction using Gradient Boosting Algorithm

3.5. Deep Learning Algorithm

A number of studies have used deep learning as the mainstream technique for state-of-the-art performances in many machine learning tasks such as text classification, image classification, machine translation, and speech recognition (Krizhevsky et al., 2012; Zhang et al., 2015; Deselaers et al., 2009; Chorowski et al., 2015). There are various types of neural networks, such as recurrent, feedforward, and backpropagation, which are used in deep learning models. These deep learning models usually deploy different functions such as rectified linear units (ReLU) and sigmoid for activation of the convolutional neural network (Agostinelli et al., 2014). ReLU is an activation function that works by thresholding values 0. This implies that it outputs 0 for any input data when the value is less than 0 and behaves linearly for positive values of input data, as shown in Figure 7 (a). Similarly, the sigmoid function is based on the exponential value, which is shown in Figure 7(b).

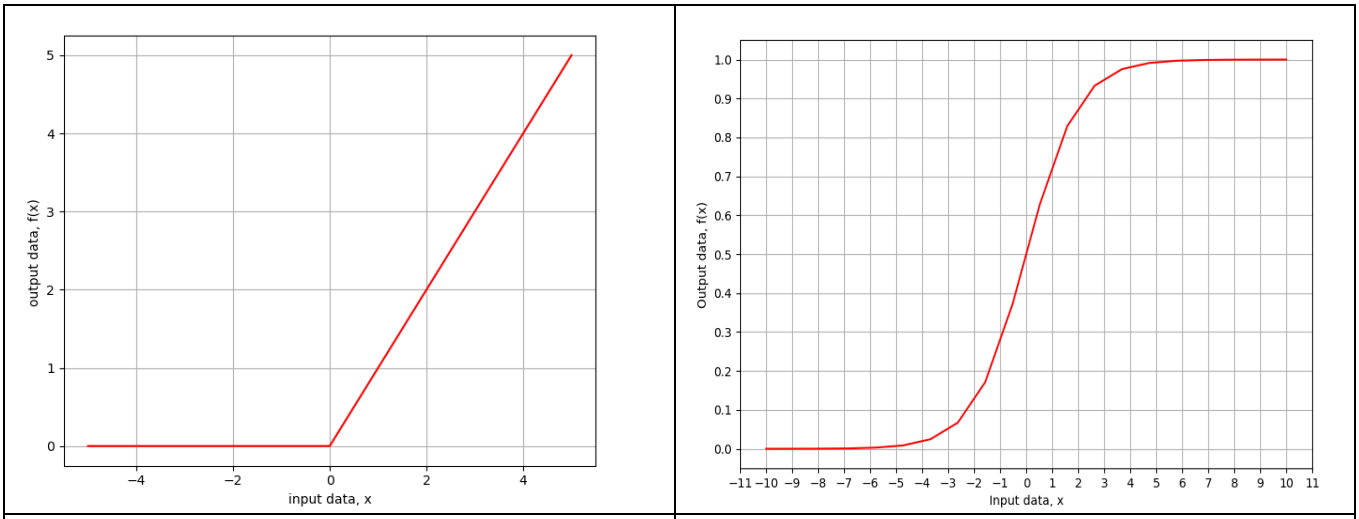

Figure 7. (a) The Rectified Linear Unit (ReLU) activation function and (b) Sigmoid function

One of the tricky aspects of deep learning is the optimization of the training criterion in which the size of the neural network and non-convexity of the training objectives poses difficulty. On the one hand, the first-order methods impact the speed in case of the ill-posed objective function, while the second-order structure is still not exploited for training the deep networks. The learning rate specific to parameters is viable alternatives to improve the accuracy of the deep networks.

In this study, we use the combination of rectified linear units (ReLU) and sigmoid functions for the activation function. The binary cross-entropy loss function is used in conjunction with a dropout value of 0.1 to avoid overfitting (Srivastava et al., 2014). The optimizations are an essential part of convolutional neural networks. We used Adam optimizer, which is an algorithm for first-order gradient-based optimization of stochastic objective functions based on adaptive estimates of lower-order moments (Kingma and Ba, 2015). The hypothesis in this study is that an optimized workflow can be constructed to achieve maximum accuracy from deep learning models for the classification of lost circulation zones. This model can be further implemented in real-time to predict the lost circulation zone.

Figure 8 shows the accuracy of the test dataset with a single vanilla hidden layer based on the ReLU activation function and squash it with several layers of the sigmoid function. The combination of single ReLU and sigmoid activation function for the hidden layer and sigmoid activation function for the output layer shows the best accuracy. However, the double hidden layer of the ReLU activation function with the sigmoid activated output layer shows an increased accuracy of 84 percent, as shown in Figure 9.

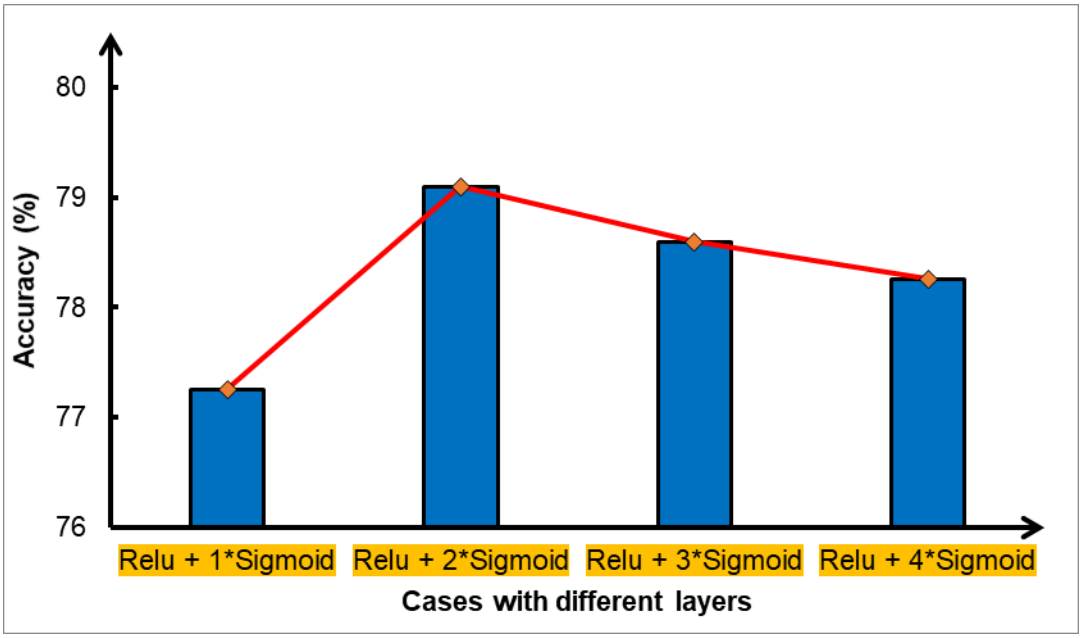

Figure 8. The test set accuracy of lost circulation prediction using single Rectified Linear Units layer activation and multilayered Sigmoid activated layers of deep network

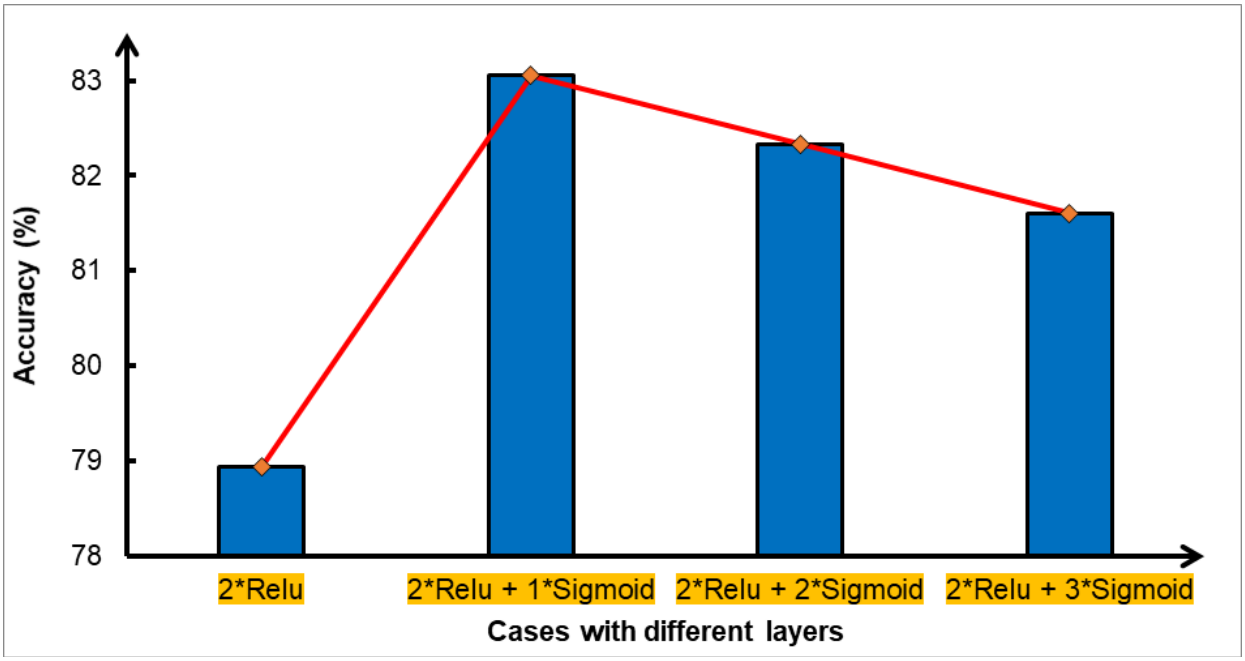

Figure 9. The test set accuracy of lost circulation prediction using double Rectified Linear Units layer activation in combination with Sigmoid activated layers of deep network

The accuracy of test set prediction is dependent on the number of epochs which achieves steady state. It can be inferred from Figure 10 that 200 number of epochs makes the highest accuracy for classification.

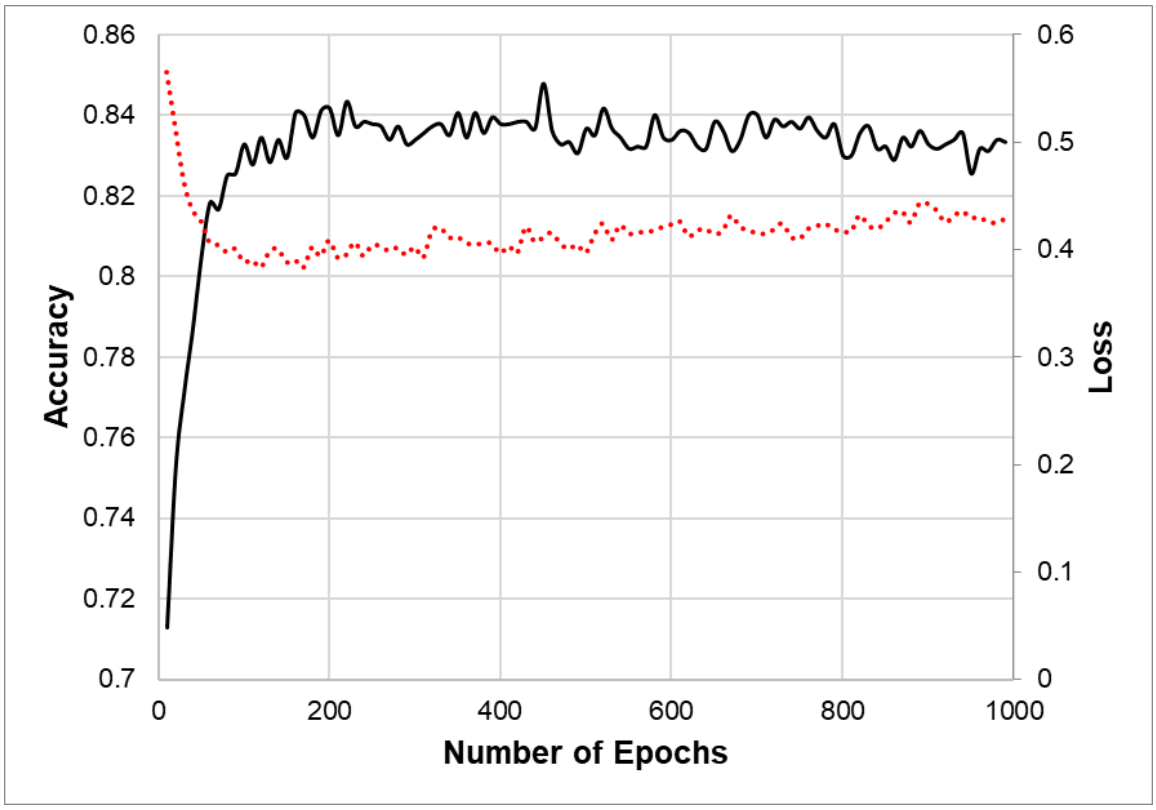

Figure 10. The test set accuracy and loss for the number of epochs in the deep network

Overall, the deep learning model suggests that the optimum efficiency is achieved with a double layer of ReLU activation function as a hidden layer, and the sigmoid activation function for the output layer with 200 epochs.

4. CONCLUSION

Lost circulation is a common problem in geothermal fields. A beforehand knowledge of zones prone to lost circulation can hugely impact the operational drilling costs. This study targets to establish a workflow that can be used to predict the lost circulation zones. Machine learning programs have been used to predict different embedded patterns in the data set. With this motivation, in this paper, we used the geothermal drilling operational data from FORGE 21-31 well data to test the workflow. In this workflow, we first implemented the window-based filter for despiking the full data set. On the filtered data, we test different machine learning algorithms such as k-nearest neighbor, decision tree, gradient boosting, and random forest classifiers to predict the lost circulation zone. The full dataset was randomly split in which 70 percent of data from the same well was used as the training data, whereas 30 percent of the dataset was used for testing purposes. Out of these four classifiers, random forest showed the highest accuracy, while decision tree-based classifiers showed the lowest accuracy. Besides the machine learning algorithms, we also implemented several combinations of deep learning models. We observed that deep learning modules were highly stable, and the accuracy is highly sensitive to activation functions. The two hidden layers with rectified linear units (ReLU) based activation functions, in combination with the sigmoid activation function-based output layer, showed the highest accuracy. The deep learning models showed an increase in the accuracy with the increase in the number of epochs.

ACKNOWLEDGMENTS

Authors of this paper would like to thank the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Geothermal Program Office Award Number DE-EE0008602 for providing travel support to attend 45th Workshop on Geothermal Reservoir and present this work.

REFERENCES

Agostinelli, Forest, Matthew Hoffman, Peter Sadowski, and Pierre Baldi., 2014. Learning activation functions to improve deep neural networks. arXiv preprint arXiv:1412.6830.

Blankenship, D., Kennedy, M., Faulds, J., Sabin, A., Akerly, J., Robertson-Tait, A., Blake, K., Siler, D.L., Hinz, N., Tiedman, A., Lazaro,

M., Glen, J., Hickman, S., Williams, C., Pettit, W., 2017. An Update on the Proposed Frontier Observatory for Research in Geothermal Energy (FORGE) at Fallon, NV. Proceedings of the Forty-Second Workshop on Geothermal Reservoir Engineering, Stanford University.

Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y., 2015. Attention-based models for speech recognition. In Advances in neural information processing systems, pp. 577-585.

Cover, T. and Hart, P., 1967. Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), pp.21-27.

Deselaers, T., Hasan, S., Bender, O. and Ney, H., 2009, March. A deep learning approach to machine transliteration. In Proceedings of the Fourth Workshop on Statistical Machine Translation, pp. 233-241.

Denoeux, T., 1995. A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE transactions on systems, man, and cybernetics, 25(5), pp.804-813.

Dietterich, T.G., 2000. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine learning, 40(2), pp.139-157.

Friedman, J.H., 2002. Stochastic gradient boosting. Computational statistics & data analysis, 38(4), pp.367-378.

Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kraal, K.O., Ayling, B. and Calvin, W., 2020. Applications of Hyperspectral Imaging to Geothermal Drill Core and Cuttings: Case Studies from Nevada, Western USA.

Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

Natekin, A. and Knoll, A., 2013. Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, p.21.

Pal, M., 2005. Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 26(1), pp.217-222.

Polat, K. and Gunes, S., 2007. Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform. Applied Mathematics and Computation, 187(2), pp.1017-1026.

Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M. and Rigol-Sanchez, J.P., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67, pp.93-104.

Siler, D.L., Hinz, N.H., Faulds, J.E., Ayling, B., Blake, K., Tiedeman, A., Sabin, A., Blankenship, D., Kennedy, M., Rhodes, G. and Sophy, M.J., 2018. The geologic and structural framework of the Fallon FORGE site. In Proceedings of the 43rd Workshop on Geothermal Reservoir Engineering, Stanford University.Breiman, L., 1996. Bagging predictors. Machine learning, 24(2), pp.123-140.

Sinha, S., Kiran, R., Tellez, J. and Marfurt, K., 2019, October. Identification and Quantification of Parasequences Using Expectation Maximization Filter: Defining Well Log Attributes for Reservoir Characterization. In Unconventional Resources Technology Conference, Denver, Colorado, 22-24 July 2019 (pp. 62-73). Unconventional Resources Technology Conference (URTeC); Society of Exploration Geophysicists.

Sinha, S., Devegowda, D., and Deka, B., 2016. Multivariate Statistical Analysis for Resource Estimation in Unconventional Plays Application to Eagle Ford Shales. Society of Petroleum Engineers. doi:10.2118/184050-MS

Sinha , S., Lima , R., Qi, J., Paez , L., and Marfurt, K., 2018. Well-log attributes to map upward-fining and upward-coarsening parasequences, Society of Exploration Geophysicists,10.1190/segam2018-2998559.1Snyder, N.K., Visser, C.F., Alfred, E.I., Baker, W., Tucker, J., Quick, R., Nagle, T., Bell, J., Bell, S., Bolton, D., and Nagandran, U., 2019. Geothermal Drilling and Completions: Petroleum Practices Technology Transfer (No. NREL/TP-6A20-72277). National Renewable Energy Lab. (NREL), Golden, CO.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), pp.1929-1958.

Stein, G., Chen, B., Wu, A.S., and Hua, K.A., 2005, March. Decision tree classifier for network intrusion detection with GA-based feature selection. In Proceedings of the 43rd annual Southeast regional conference-Volume 2 (pp. 136-141). ACM.

Swain, P.H. and Hauska, H., 1977. The decision tree classifier: Design and potential. IEEE Transactions on Geoscience Electronics, 15(3), pp.142-147.

Torlay, L., Perrone-Bertolotti, M., Thomas, E., and Baciu, M., 2017. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. Brain informatics, 4(3), p.159.

Zapata, Y. and Sakhaee-Pour, A., 2016. Modeling adsorption-desorption hysteresis in shales: Acyclic pore model. Fuel, 181, pp.557-565. https://doi.org/10.1016/j.fuel.2016.05.002.

Zapata, Y. and Sakhaee-Pour, A., 2017. Pore-body and-throat size distributions of The Geysers. Geothermics, 65, pp.313-321. https://doi.org/10.1016/j.geothermics.2016.10.008.

Zhang, X., Zhao, J., and LeCun, Y., 2015. Character-level convolutional networks for text classification. In Advances in neural information processing systems (pp. 649-657).