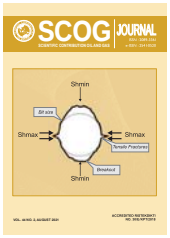
Scientific Contributions Oil & Gas, Vol. 44. No. 2, August 2021: 141 - 152

SCIENTIFIC CONTRIBUTIONS OIL AND GAS
Reserach and Development Centre for Oil & Gas Technology
"LEMIGAS"

Journal Homepage:http://www.journal.lemigas.esdm.go.id
ISSN: 2089-3361, e-ISSN: 2541-0520



Prediction of Hydraulic Fractured Well Performance using Empirical Correlation and Machine Learning

Kamal Hamzah[1,2)], Amega Yasutra[2)], and Dedy Irawan[2)]
[1]PT. Medco E&P Indonesia
The Energy Building 36th Floor, SCBD Lot 11A
Jl. Jendral Sudirman, Kav. 52-53, Jakarta 12190, Indonesia
[2]Institut Teknologi Bandung
Jl. Ganesha 10, Lb. Siliwangi, Kec. Coblong, Bandung, Jawa Barat 40132 Indonesia
Corresponding author: humas@itb.ac.id

Manuscript received: May, 8th 2021; Revised: July, 20th 2021
Approved: August, 30th 2021; Available online: September, 2nd 2021

ABSTRACT - Hydraulic fracturing has been established as one of production enhancement methods in the petroleum industry. This method is proven to increase productivity and reserves in low permeability reservoirs, while in medium permeability, it accelerates production without affecting well reserves. However, production result looks scattered and appears to have no direct correlation to individual parameters. It also tend to have a decreasing trend, hence the success ratio needs to be increased. Hydraulic fracturing in the South Sumatra area has been implemented since 2002 and there is plenty of data that can be analyzed to resolve the relationship between actual production with reservoir parameters and fracturing treatment. Empirical correlation approach and machine learning (ML) methods are both used to evaluate this relationship. Concept of Darcy's equation is utilized as basis for the empirical correlation on the actual data. The ML method is then applied to provide better predictions both for production rate and water cut. This method has also been developed to solve data limitations so that the prediction method can be used for all wells. Empirical correlation can gives an R^2 of 0.67, while ML can give a better R^2 that is close to 0.80. Furthermore, this prediction method can be used for well candidate selection means.

Keywords: Hydraulic Fracturing, Well Performance, Empirical Correlation, Machine Learning.

How to cite this article:

Kamal Hamzah, Amega Yasutra, and Dedy Irawan, 2021, Prediction of Hydraulic Fractured Well Performance using Empirical Correlation and Machine Learning, Scientific Contributions Oil and Gas, 44 (2) pp., 141-152.

INTRODUCTION

Hydraulic fracturing is a stimulation treatment to enhance well productivity and improves the economic value of well reserves. It have been widely applied to both low and moderate permeability reservoirs. In low permeability reservoir, it greatly contributes both to well productivity and to well reserves, while in moderate permeability reservoirs, it accelerates production without impacting the well reserves (Holditch & Ma, 2016).The production improvement and additional reserve (if any) have to be justified economically because hydraulic fracturing jobs requires high cost and involve large scale equipment.

Meanwhile, hydraulic fracturing in some fields in South Sumatra has been implemented on more than 200 wells since 2002. Year by year, it becomes more challenging due to increasingly limited well candidates with good reservoir properties and decreasing trend in production results. Thus, hydraulic fracturing optimization both in planning

that include well candidate selection/design and job execution has to be applied. The objective is to increase fracturing job success ratio and therefore contribute to more oil production.

However, the production results from hydraulically fractured wells looks scattered and has no clear correlation to individual well parameters (Azhari, 2015). It gets even more challenging due to the decreasing trend in production result of hydraulic fracturing job that make the job economics becoming marginal. There are hundreds of data sets that are potentially useful for the evaluation. The data covers the reservoir parameters from primary log data, petrophysical analysis, and dynamic parameters from well-testing analysis. Hydraulic fracturing parameters both the treatment data and fracture geometry result are also utilized.

To ensure that hydraulic fracturing job can give additional value for the field, there is a need to estimate and quantify conclusively the result of hydraulic fracturing job in order to minimize unsuccessful job. Thus, the prediction tool has to be developed to determine well candidate based on reservoir and well properties. Two (2) methods to develop the prediction tools are presented in this paper.

The first method is empirical correlation equation. It contains a mathematical equation based on some parameters from a given set of empirical data that will be used for predicting other data (Ribarič & Šušteršič, 2017). In this paper, the empirical correlation equation is used to predict the hydraulic fracturing result based on combined parameters of reservoir pressure, transmissibility data and dimensionless productivity index from hydraulic fracturing treatment.

The second method is machine learning approach. Machine learning is the study of computer algorithms that can improve automatically through experience by discovering general rules in large data sets to meet the user's interest (Mitchell 1997). Temizel, *et al.* (2021) has described the applications of machine learning in oil & gas industry and provide its capability and limitations in unconventional reservoir engineering and well completion calculations. In many cases, machine learning was proven able to predict the output that has problem in data limitation and data quality (Makhotin, *et al.*, 2019).

Finally, both two (2) approaches are expected to yield a robust prediction tool that can be easily applied to all wells using common primary data as input parameters. The prediction tool is then utilized in well candidate selection to determine the good well candidates and eliminate non-potential well candidates in the future hydraulic fracturing job campaign. It's also expected to ensure and guide the fracturing treatment optimization to maximize the production result.

DATA AND METHODS

The workflow in developing this study consist of 5 major phases that represents the whole process and details of working procedure. It is shown on Fig.1 that consists of data preparation, empirical correlation approach, machine learning approach to predict production and water cut, and its application on well candidate selection. Actual data field will be incorporated in the discussion of each step to ensure that this methodology is applicable.

A. Data Preparation

Data from actual hydraulic fracturing job were collected and tabulated that including input and output data. The output data is the production performance whereas input data covers well data including well completion data such as perforating length and wellbore size, and reservoir data including several parameters such as reservoir pressure, open-hole log and petrophysical data. Hydraulic fracturing parameters also covers actual treatment data and calculated fracture geometry.

The data was evaluated to become a ready-to-use data set. It consists of several steps that includes data conversion, ignoring & filling missing values and outlier data elimination. The data set is tested by using Pearson correlation coefficient (Pearson, 1920) to measure the direction and strength of the linear relationship between input parameters and output parameter.

B. Empirical Correlation Approach

The empirical correlation approach utilizes the Darcy equation in radial condition as shown in equation (1). This equation need reservoir properties that are represented by transmissibility and reservoir pressure and fracture parameters. Transmissibility can be defined from petrophysical analysis and well testing. The latter method is preferred because it represents the reservoir quality at certain radius. This method utilizes the pre-frac data obtained in mini-fall-off test and analyzed using short

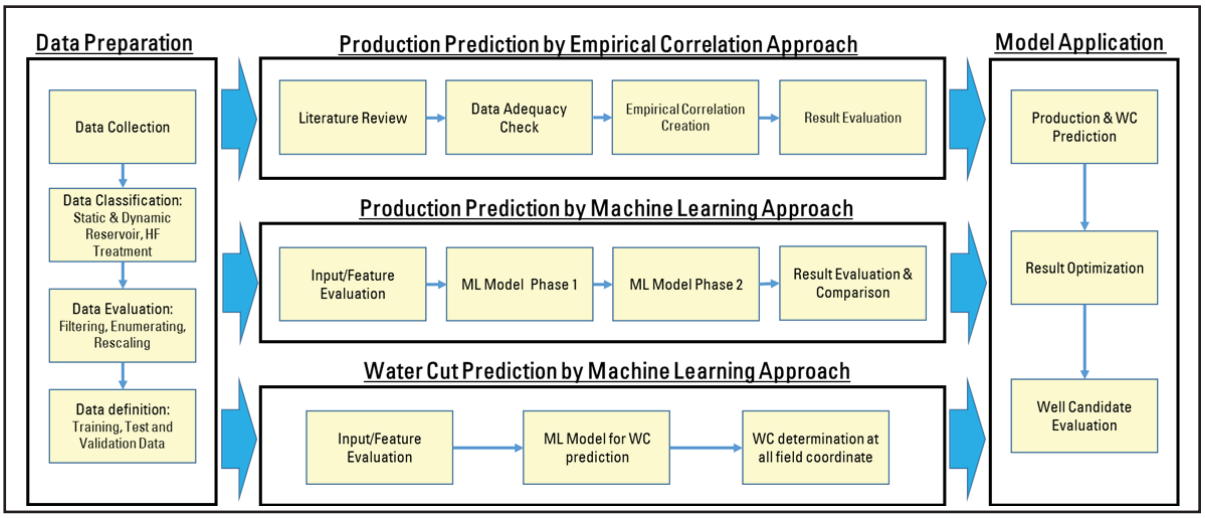


Figure 1
Workflow for performance prediction of hydraulic fractured well.

impulse injection test to obtain reservoir permeability (Abousleiman *et al*., 1994). Through this method, we can also estiamte the reservoir pressure.

$$q = \frac{kh(P_r - P_{wf})}{141.2\ \mu B \ln(\frac{r_e}{r_w} + S)} \quad (1)$$

Fracture parameter is represented by dimensionless productivity index (PI) or J_D as shown in equation (2). This parameter can be calculated based on a simulated result of fracture geometry using equivalent wellbore radius concept and unified fracture design method thru proppant number.

$$J_D = \frac{1}{\ln(\frac{r_e}{r_w} + S)} = \frac{1}{\ln(\frac{r_e}{r_w'})} \quad (2)$$

Prats (1961) first introduce the idea of the effective wellbore radius ($r_{w'}$). It based on a simple balancing of flow areas between a wellbore and a fracture gives the equivalent value of $r_{w'}$ for a propped fracture. Cinco-Ley *et al*. (1978) later integrated this into a full description of reservoir response, including transient flow. For pseudoradial flow, rw′ is expressed as a function of fracture length (X_f) and dimensionless fracture conductivity (C_{fD}).

The unified fracture design methodology provided by Economides, Oligney, and Valkó (2002), expands the above approach to include fracture and well drainage area dimensions that will not reach pseudoradial flow before the onset of pseudosteady state. The key of this approach is the idea that for a given proppant volume and well drainage area, there is a fracture half length, width, and conductivity that maximize the well productivity. For a given proppant volume, square well drainage area, and values for both proppant and reservoir permeability, the dimensionless proppant number, Np, is defined as

$$N_p = \frac{4k_f x_f w}{k{x_e}^2} = \frac{4k_f x_f wh}{k{x_e}^2 h} = \frac{2k_f V_{prop}}{kV_{res}} \quad (3)$$

All above parameters are combined based on equation (1) and then plotted versus production result (in BFPD) at pseudosteady state regime. The correlation of this cross plot can be utilized in a form of empirical correlation to predict the result of next hydraulic fracturing job.

C. Machine Learning Approach

Machine learning (ML) is a broad subfield of artificial intelligence aimed to enable machines to extract patterns from data set. It is based on mathematical statistics, numerical methods, optimization, probability theory, discrete analysis, geometry and etc (Smola & Vishwanathan, 2008). There are three main components in ML that consists of data, features or parameters and algorithms or method.

Nowadays there are four main directions in machine learning that consist of classical ML, ensemble methods, reinforcement learning and neural networks & deep learning (www.vas3k.com). Classical ML utilizes pure statistics method and consist of supervised (e.g. linear regression) and unsupervised (e.g. clustering). Ensemble methods construct a set of classifiers and then classify new data points by taking a (weighted) vote of their

predictions (Dietterich, 2000). Some wellknown ensemble method are Random Forest (Breiman, 2001), Gradient Boosting (Friedman, 2001) and Ada-Boost (Dietterich, 2000). Reinforcement learning is used in case no data input but it have an environment to live in. Neural Networks and Deep Leaning is used for replacement of all previous algorithms. It often used in object identification, speech recognition and synthesis, image processing and etc.

In this paper, ML is used to create relationship between production result as output and reservoir/ hydraulic fracturing parameters as input. The ML approach is divided into two phases. In phase 1, the input data is limited to pressure, transmissibility and PI dimensionless data, similar to the empirical correlation approach. It aims to compare the result of empirical correlation and ML methods, and to evaluate the importance of each parameter in machine learning. In phase 2, the input used in phase 1 is replaced by primary data such as open-hole log, petrophysical parameters, and fracturing treatment. This phase is proposed to make the ML model usable practically by using parameters that almost all wells have.

The ML model that used in this phase is supervised machine learning (linear regression) and ensemble method that consist of Random Forest, Gradient Boosting and Adaboost. The best model is chosen based on mean absolute error (MAE), coefficient of determination (R^2) and Pearson correlation coefficient (R) on actual production rate and prediction based on ML model.

Beside production rate prediction, the ML is also used to predict the water cut (WC) and net oil production. In this case, the k-Nearest Neighbor (k-NN) model is utilized due to the similarity of its base concept with the WC prediction based on geographical coordinates. The main concept of kNN is to predict the label of a query instance based on the labels of k closest instances in the stored data (Kang, 2021). The stored data in our case is the WC data of each producer wells.

D. Application on Well Candidate Selection

The best ML model on production rate and water cut predictions are implemented in well candidate selection for future hydraulic fracturing job. Economic value is determined for each well based on the net oil prediction. This method can eliminate wells with low oil production potential that yields low economic value.

Moreover, the ML model can be utilized to optimize the fracturing treatment plan in order to have more oil production. The optimization can be conducted by tuning the proppant type, proppant volume and other parameters.

RESULTS AND DISCUSSION

This section presents results of empirical correlation and ML approaches to establish the relationships between actual production and reservoir/fracturing parameters.

A. Empirical Correlation Approach

There are 45 hydraulic fractured wells that have complete data as required by the Darcy's equation. Products of pressure, transmissibility data, and dimensionless productivity index is plotted against the actual production rate to get the trend and empirical correlation as well.

Two (2) correlations have been developed based on two methods to define the dimensionless PI. The first method is the Cinco-Ley correlation using

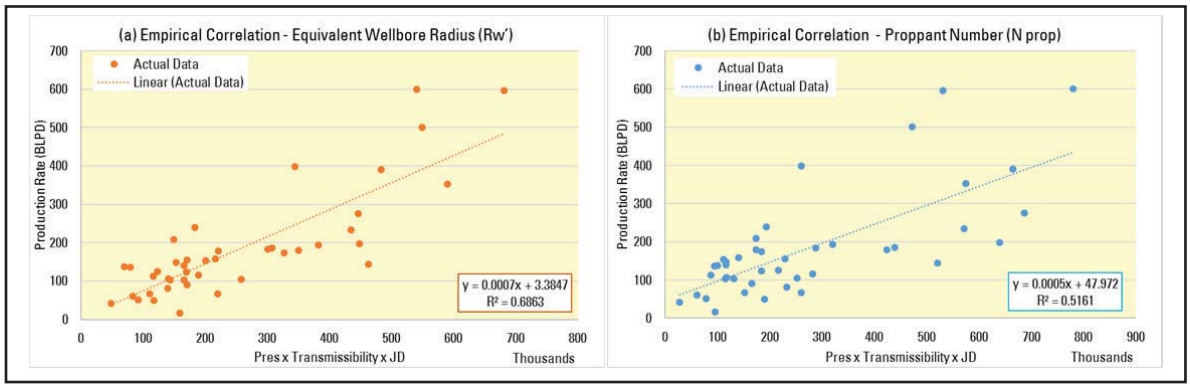


Figure 2
Empirical Correlation based on (a) Cinco Ley Correlation using equivalent wellbore radius and (b) Unified Fracture Design using proppant number.

the equivalent wellbore radius whereas the second method is unified fracture design using proppant number. The empirical correlation results for each method are shown on Figure 2.

The empirical correlation shows that there is a linear correlation between the product of pressure, transmissibility, and dimensionless PI to the production rates. It complies basic Darcy's equation. Empirical correlation based on dimensionless PI using wellbore equivalent radius gives better result than empirical correlation based on proppant number. The R^2 of the first method is 0.67 which is higher than R^2 from the second method of 0.57. So is the Pearson correlation coefficient, it shows that the first method give +0.82 which is higher number than the second method of only +0.72. The cross plot between actual production and production result based on empirical correlation is shown in Figure 3. At production rate of more than 400 BLPD, the plot starts to deviate from line R^2=1, resulting in R^2 of lower than 0.7.

B. Machine Learning Approach

Machine learning phase 1 was carrying out using input as same input as for the empirical correlation approach. The objective is to assess and ensure that ML can be utilized for the production prediction of hydraulic fractured well. The result then will be compared to the empirical correlation approach.

Input evaluation has been applied by assessing the input using feature engineering and feature importance for each ML model. As described on Figure 4, reservoir transmissibility appears to be the most influential parameters to the production rate result. It is represented by the Pearson correlation coefficient value of +0.72, which is the highest among the parameters. Then it is followed by reservoir pressure and dimensionless PI.

Consistent with the feature importance evaluation, reservoir transmissibility becomes the most important for all ML models with values between 0.5 and 0.6. Reservoir pressure and J_D have ranging values within 0.1 – 0.3 for all ML models.

The 45 wells data is then divided into two groups of 80% data for training and 20% data for testing. As previously mentioned, the ML models used to predict production results is Linear Regression, Random Forest, Gradient Boosting and AdaBoost.

Training and testing result evaluations for each ML model are shown in Table 1. The ML model that has R^2 consistent above 0.7 are Random Forest and Gradient Boosting. AdaBoost has a tendency for overfitting in the training, having testing R^2 of only 0.67. Linear regression shows not too robust as the testing R^2 is larger than the training value. As described on Fig. 5, the cross plot between actual production and prediction from ML model that follow the R^2=1 are Random Forest and Gradient Boosting. Linear regression shows scatter and AdaBoost clearly shows overfitting in the training data set but scatter in the testing data set.

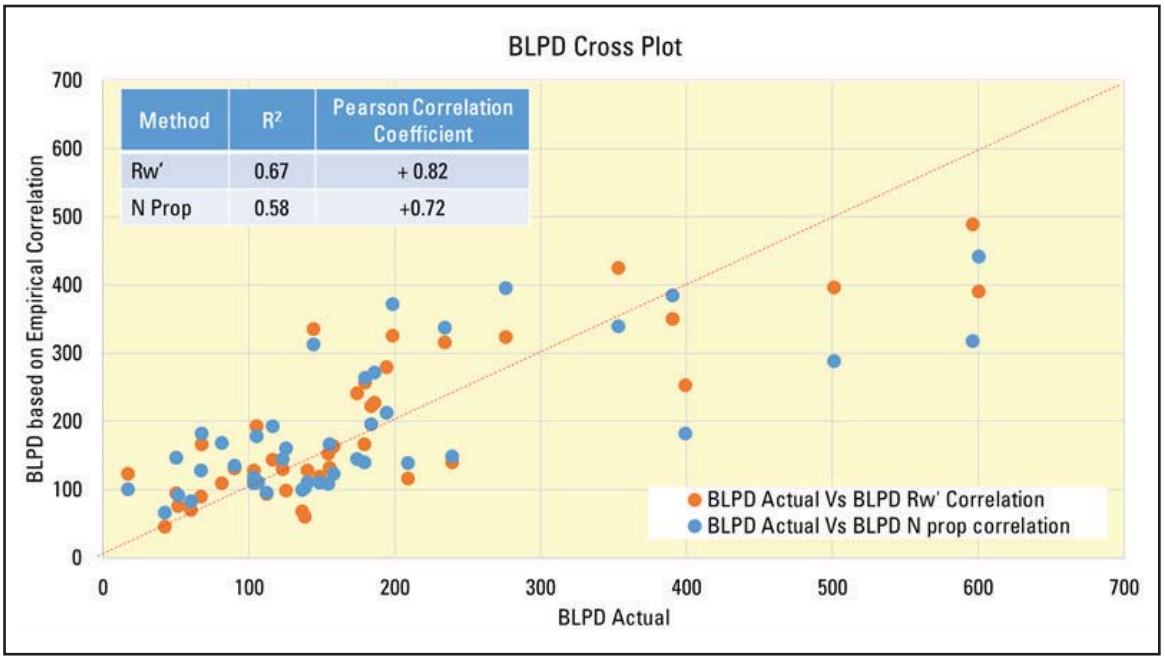


Figure 3
Cross Plot between production actual Vs production from empirical correlation.

Through this approach, it can be concluded that ML can be used to estimate the production result of hydraulic fractured well. Transmissibility from mini-fall-off test becomes the most important parameter in machine learning. However, the main concern is that the ML model cannot be applied to other well because of:

- The availability of transmissibility data from mini-fall-off test are limited to only 45 wells. It required high effort/cost to obtain this data.
- The correlation between actual transmissibility and petrophysical analysis is low at $R^2 = 0.597$. The utilization of transmissibility of petrophysical analysis will lead to lower R^2 in the correlation and produce more error. Thus, it cannot replace the actual transmissibility.

C. Machine Learning with Primary Data Set

In order to have ML model that is practically able to be used for well candidate selection, ML model based on the primary data set is then developed. This ML model will utilize common parameters that almost all of wells have. The input parameters will consist of basic well data, open-hole log, petrophysical analysis and fracturing treatment. Fracturing treatment data are preferred to be used because lack of validation on fracture geometry data. Only limited number of wells that have temperature log to validate the simulated fracture geometry.

Feature engineering is applied to both reservoir properties and fracturing treatment to select the parameters that can be input as feature in ML model. The value of Pearson correlation coefficient to production data that higher than +1 and lower than -1 are selected to feature in ML. Some important parameters were also selected to be input feature in ML. Finally, the parameters that were selected to be the input feature in ML model are shown in Table 2.

There are 140 wells data set to be analyzed by ML. They are divided into two groups consisting of 80% data for training and 20% data for testing. Same as previous, ML model used to predict the production result are Linear Regression, Random Forest, Gradient Boosting and AdaBoost.

The training and testing results of each ML model are shown in Table 3. The ML models that show R^2

Table 1
ML phase 1 result evaluation

ML model	Training : 80%			Testing: 20%		
	MAE	R^2	R	MAE	R^2	R
Random Forest	32	0.88	0.94	29	0.76	0.95
Linear Regression	64	0.57	0.75	19	0.87	0.96
Gradient Boosting	2.1	1	1	26	0.83	0.91
AdaBoost	0.1	1	1	33	0.67	0.95

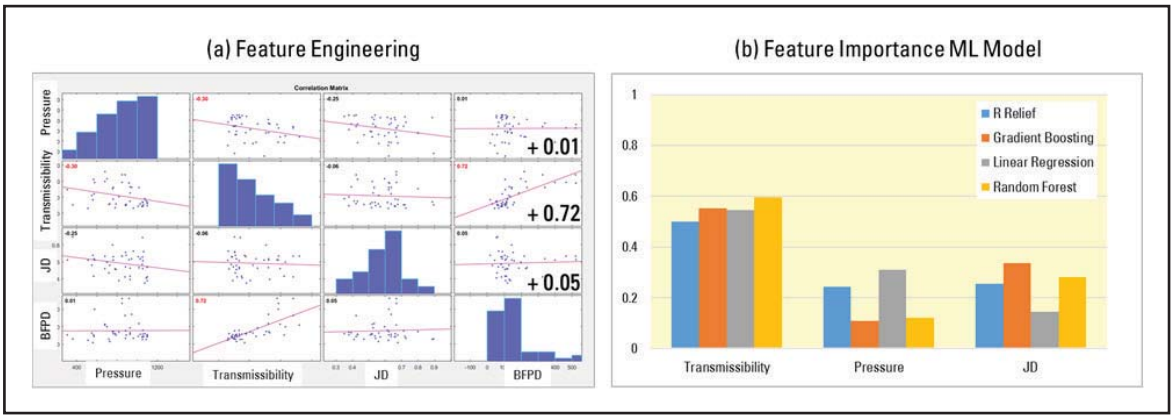


Figure 4
Input evaluation using (a) feature engineering and (b) feature importance for each ML Model.

consistent above 0.7 are Random Forest, AdaBoost and Gradient Boosting. Random Forest and Gradient Boosting give satisfying results based on R^2 and Pearson Correlation coefficient. AdaBoost tend to have over-fitting in the training but also give good result in the testing, whereas Linear Regression's R^2 value is the lowest one.

Figure 6 also describes the cross plot between actual production and prediction result from each ML model. Training and testing data that follow the R^2=1 was shown in gradient boosting and random forest. However, both Random Forest and Gradient Boosting exhibit deviation from line R^2=1 at rate higher than 500 BLPD. This might be caused by the small data population in this production rate range. So the ML model does not have adequate data for having accurate prediction at this range production rate.

However, Linear Regression model shows scattered correlation between actual and prediction result both for traninig and testing data set. Linear Regression does not seem fit with the typical data with many input and have an issue in the quality. AdaBoost tend to have over fitting in training data set so that looks scatter in testing data set.

Finally, we can inferred that Gradient Boosting is the most reliable ML model to predict the production rate of hydraulic fractured well. Gradient Boosting has stable R^2 both in training and testing data set that indicates the robustness of this model. This model is then recommended to be applied in the production rate prediction.

D. Water Cut (WC) Prediction

ML can also be utilized to predict the water cut of each well by position. For this purpose, k-Nearest Neighbors (k-NN) is used for the water cut prediction based on XY coordinate as well as vertical position of the bottom zone. The dataset is taken from all producer wells with the updated WC and is also divided into two sets, training and testing. This partition to ensure that the model has good robustness and consistency.

Table 2
Input evaluation result for ML phase 2

Reservoir Parameters	R	Remarks
Field	0.43	
Perforation Length	0.13	
Reservoir Pressure	0.06	Important parameter
Thickness	0.21	
GR	-0.21	
Resistivity	0.12	
Density	-0.1	
V clay	-0.31	
Perm	0.43	
kh	0.46	
Fracturing Parameters		
Proppant Type	0.26	
Proppant Volume	-0.37	Important parameter
Fluid Type	0.42	
Fluid Volume	-0.25	Important parameter
Frac Gradient	0.13	

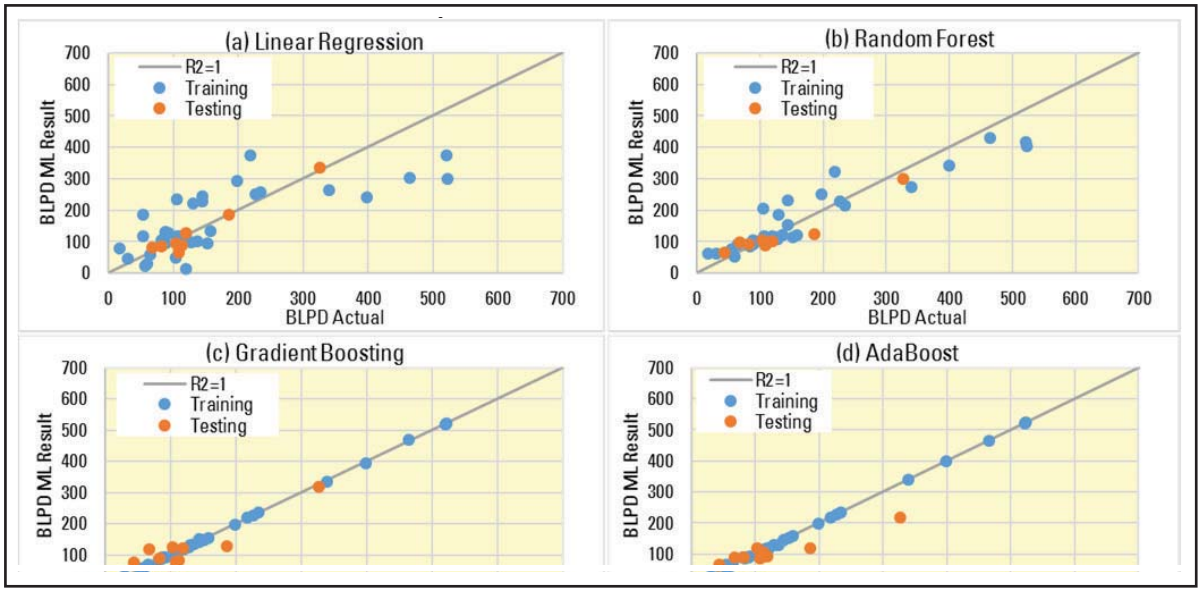


Figure 5
Cross plot between actual production Vs production result from each ML model in phase 1.

The k-NN manipulates the training data and classifies the new test data based on distance metrics. There are some parameters that need to be tuned to improve the performance such as K value, distance metric and weights. Scaling/normalizing the data set can help to improve the k-NN performance. K value indicates the count of the nearest neighbors. Some method to define distance metric can be adjusted to have reliable model such as Euclidean, Manhattan and Chebyshev distance (Cantrell, 2000).

For the WC prediction in this case, the optimum model is obtained using k=5 and Chebyshev distance. The result of kNN on WC prediction in two fields are shown by Figure 7. Figure 7.a shows the position of well that currently producing. The WC is measured from liquid sample by laboratory testing. Figure 7.b represents the result of k-NN model in WC mapping on every coordinate on those two fields. Field 1 shows that in any well location, the wells mostly have high WC. This field has been drained since 2002 and is experiencing pressure depletion. Water injection was then applied to maintain the pressure.

On the other hand, Field 2 shows better WC at any location. This field has very low permeability thus the recovery factor is still low. The reservoir in this field has been massively developed since 2015.

E. Application on Well Candidate Selection

The most reliable of the ML models for fluid production and water cut prediction is then applied to predict the performance of well candidate after hydraulic fracturing job. There are total eigth well candidates for next hydraulic fracturing campaign. These well candidates have already had water cut from the existing zone and produce oil rate below

Table 3
ML phase 2 result evaluation

ML model	Training : 80%			Testing: 20%		
	MAE	R^2	R	MAE	R^2	R
Random Forest	39	0.87	0.95	47	0.77	0.91
Linear Regression	77	0.45	0.67	75	0.5	0.81
Gradient Boosting	37	0.89	0.96	52	0.78	0.9
AdaBoost	1.9	0.99	0.99	53	0.72	0.89

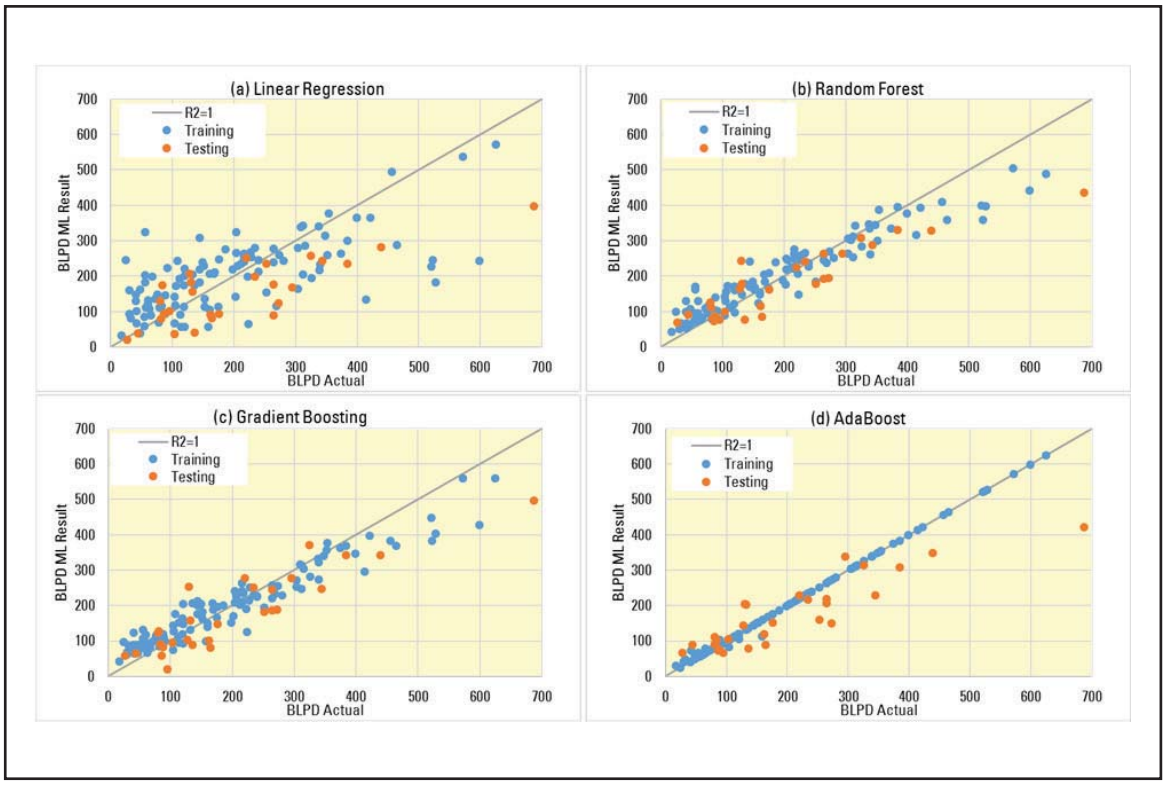


Figure 6
Cross plot between actual production Vs production result from each ML model in phase 2.

5 BOPD. Workover change layer and hydraulic fracturing as stimulation treatment will be proposed to these wells.

The prediction tool based on Gradient Boosting and k-NN is shown in Table 4. The economic cut off for hydraulic fracturing job is equivalent to the initial production 40 BOPD. Based on the prediction, there are four wells that will have initial oil production higher than 40 BOPD and three wells that lower than 40 BOPD.

Further assessment is then applied to those three wells. Preliminary optimization to the hydraulic fracturing treatment is assessed and the result is predicted. After optimization of proppant type and volume, then we can conclude that only one well can afford to get higher than 40 BOPD, while the remaining wells is still under economic cut off. Thus, the other two wells as shown in Table 5 will be excluded from hydraculic fracturing job plan.

By this result, it can be confirmed that ML model through Gradient Boosting and k-NN model can be applied for both production rate and water cut predictions. The ML approach can be applied using primiary data that almost all wells have. Empirical correlation cannot be applied in this case because of the data limitation such as unavailability of actual transmissibility data.

The objective to develop the prediction tool that can be applied for any wells can be achieved. It enable quantitative comparison on estimated hydraulic fracturing result so that potential well candidate can be easily selected. This tool also can be utilized to ensure and guide the fracturing treatment optimization to maximize the production result.

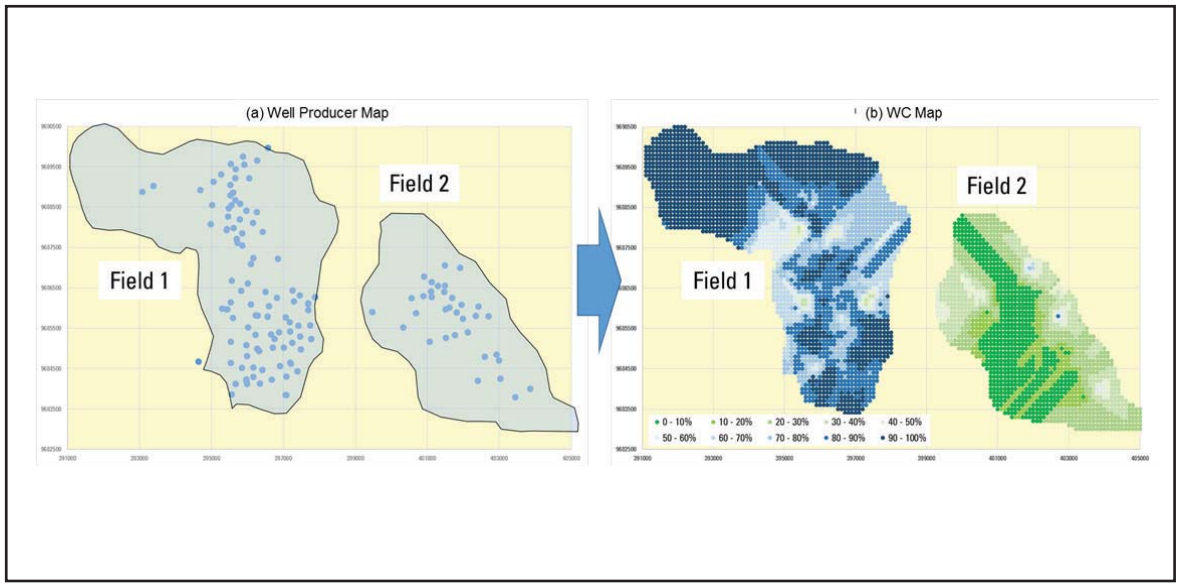


Figure 7
Water cut prediction map using k-NN model.

Table 4
Prediction result based on ML model for hydraulic fracturing well candidates

Well	Res pressure (psig)	KH (md.ft)	Proppant type	Proppant volume (klbs)	BLPD prediction	WC prediction	BOPD prediction	Remarks
W-0103	698	112	3	60	166	0.78	37	Below economic cut off
W-0313	553	266	3	60	175	0.85	26	Below economic cut off
W-0329	612	663	3	60	274	0.83	48	
W-0343	624	928	3	60	243	0.81	47	
W-0331	849	16.4	3	90	134	0.25	100	
W-0335	972	16.4	3	90	82	0.39	51	
W-0367	500	508	3	60	207	0.89	22	Below economic cut off

Table 5
Prediction result based on ML model for hydraulic fracturing well candidates after optimization

Well	Res pressure (psig)	KH (md.ft)	Proppant type	Proppant volume (klbs)	BLPD prediction	WC prediction	BOPD prediction	Remarks
W-0103	698	112	3	80	184	0.78	41	
W-0313	553	266	3	80	193	0.85	28	Below economic cut off
W-0329	612	663	3	60	274	0.83	48	
W-0343	624	928	3	60	243	0.81	47	
W-0331	849	16.4	3	90	134	0.25	100	
W-0335	972	16.4	3	90	82	0.39	51	
W-0367	500	508	3	80	225	0.89	24	Below economic cut off

CONCLUSIONS

Some conclusions can be inferred from the above discussion about the prediction of hydraulic fracturing job result using empirical correlation and machine learning model. Several points that can be inferred is below:

Empirical correlation based on Darcy equation gives maximum R^2 of 0.67 and Pearson correlation coefficient +0.82 between actual and prediction production. It can be applied as prediction tool but require input parameters which is not all wells have such as actual transmissibility and dimensionless PI.

Machine learning phase 1 (Random Forest & Gradient Boosting) using same input as empirical correlation, give better result with $R^2 > 0.75$ and Pearson correlation coefficient > 0.9 between actual and prediction production. It describes that ML approach can be utilized in prediction of hydraulic fractured well performance.

Machine learning phase 2 (Gradient Boosting) by utilizing primary data such as open-hole-log, petrophysical analysis and frac treatment data yield $R^2 > 0.8$ and Pearson correlation coefficient > 0.9 between actual and prediction production. It confirms that this method is the most reliable prediction tool with the lowest error.

The most important parameter that majorly contributed to hydraulic fracturing results is the reservoir quality that can be expressed by permeability or transmissibility data.

K-NN model can be utilized to predict the WC using coordinate and altitude well. It can give a satisfying result with R^2 value of 0.845 in the testing data set.

ML model through Gradient Boosting for production prediction and k-NN for WC prediction can be used as prediction tool for well candidate selection implementation and frac treatment optimization.

Some recommendations for further improvement in the prediction of hydraulic fracturing result using machine learning are below:

Mini-fall-off analysis should be applied and collected from all hydraulic fracturied job to create the correlation between open-hole data & petrophysycal analysis with the actual transmissibility data. It will help to create a reliable empirical correlation as prediction tools. However, it will extend the overall execution time and lead to additional extra cost.

Some parameters that also influence the fracture geometry such as shale barrier and stress contrast based on sonic log should be considered as additional feature in the ML model. This paper excludes these parameters because of the lack of sonic log data availability in the data set.

Individual ML model to determine each of reservoir properties factor and hydraulic fracturing factor may help to improve R^2 in the production prediction and enable additional optimization methods in hydraulic fracturing. Additional number of data is required for this purpose.

Additional hydraulic fracturing job data from other area (outside South-Sumatra) will improve the data diversity thus yield more reliable prediction tool. The data confidentiality and company's discrecy is the major challenge to overcome.

ACKNOWLEDGEMENTS

The authors would like to thank PT Medco E&P management for permission to use the field data on this paper. We are also thankful to ITB lecturer for the guidance and direction to publish this scientific paper.

GLOSSARY OF TERMS

Symbol	Definition	Unit
q	Production Rate	BLPD
k	Reservoir permeability	md
Pr	Reservoir Pressure	psig
Pwf	Flowing bottom hole pressure	psig
μ	Viscosity	cp
B	Formation Volume factor	-
re	Reservoir radius	ft
rw	Wellbore radius	ft
rw’	Equivalent Wellbore radius	Ft
S	Skin	-
$\frac{kh}{\mu}$	Transmissibility, the product of permeability, reservoir thickness and fluid viscosity that represent the productivity of reservoir.	md.ft/cp
JD	Dimensionless Productivity Index	-
Np	Proppant Number	-
kf	Fracture permeability	md
xf	Fracture length	ft
xe	Reservoir length	ft
w	Fracture width	ft
V prop	Volume proppant	ft3
V res	Volume reservoir	ft3
Hydarulic Fracturing	A type of well stimulation treatment designed to bypass near-wellbore damage and improve the fluid flow path from the formation to the well.	
Impulse Testing	A specialized well testing procedure that enables analysis of the reservoir response following a relatively short duration of fluid injection or production (Ayoub et al., 1988)	
Mini-fall-off Test	An injection-falloff diagnostic test performed before a main fracture stimulation treatment. The intent is to break down the formation to create a short fracture during the injection period, and then to observe closure of the fracture system and the reservoir response after closure	
Pearson Correlation Coeffcient (R)	The linear correlation coefficient developed by Karl Pearson that measures the strength and the direction of a linear relationship	
Machine Learning	A broad subfield of artificial intelligence aimed to enable machines to extract patterns from data based on mathematical statistics, numerical methods, optimization, probability theory, discrete analysis, geometry, etc.	

Symbol	Definition	Unit
AdaBoost	One of ensemble ML model that works by weighting the observations, putting more weight on difficult to classify instances and less on those already handled well.	
Gradient Boosting	The development of AdaBoost algorithm that identifies the shortcomings by using high weight data points. Gradient boosting performs the same by using gradients in the loss function	
Random Forest	Ensemble ML model that has algorithm to build multiple decision trees and merges them together to get a more accurate and stable prediction.	
Linear Regression	A supervised machine Learning model in which the model finds the best fit linear line between the independent and dependent variable.	

REFERENCES

Economides, M. J., Hill, A. D., Ehlig-Economides, C. & **Up, D. Z**., 2013. *Petroleum Production System.* 2nd ed. Masscachusetts: Prentice Hall.

Economides, M. J. & **Nolte, K. G**., 2010. *Reservoir Stimulation.* 3rd ed. New Jersey: Prentice Hall.

Friedman., J. H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics,* 29(5), pp. 1189 - 1232.

Holditch, S. A. & **Ma, Y. Z.**, 2016. *Unconventional Oil and Gas Resources Handbook: Evaluation and Development*:Elsevier/Gulf Professional Publishing.

Kang, S., 2021. k-Nearest Neighbor Learning with graph neural networks. *Mathematics,* 9(8), p. 830.

Makhotin, I., Koroteev,, D. & **Burnaev, E.,** 2019. Gradient boosting to boost the efficiency of hydraulic fracturing. *Journal of Petroleum Exploration and Production Technology,* pp. 1-7.

Mutalova, R. F., Mozorov, A.D., Osiptsov, A.A., Vainshtein, A.L., Burnaev, E.V., Shel, E.V., & **Paderin, G.V.,** 2019. Machine learning on field data for hydraulic fracturing design optimization. *Journal of Petroleum Science & Engineering. Special Issue: Petroleum Data Science,* pp. 1-21.

Pearson, K., 1920. Notes on the history of correlation. *Biometrika,* 13(1), pp. 25-45.

Ribarič, M. & **Šušteršič, L**., 2017. *Empirical formulas for prediction of experimental data & appendix,* Slovenia: Jožef Stefan Institute, Ljubljana.

Smola, A. & **Viswanathan, S. V. N.,** 2008. *Introduction to machine learning.* Cambridge: Cambridge University Press.

Temizel, C., Canbaz, C.H., Palabiyik, Y., Aydin, H., Tran, M., Ozyurtkan, M.H., Yurukcu, M., & **Johnson, P.,** 2021. *A thorough review of machine learning applications in oil and gas industry.* Virtual, SPE/IATMI.