Artificial intelligence approach to depositional facies characterization based on electrical log data

G E Putri¹, A Haris² and M R Septandy³

¹Reservoir Geophysics, Department of Physics, Universitas Indonesia
²Geophysics Study Program, Universitas Indonesia
³Geology Study Program, Universitas Indonesia
*E-mail: aharis@sci.ui.ac.id, m.rizqy@sci.ui.ac.id

Abstract. This research aims to determine the depositional facies from electrical log data using the gradient boosting classifier method, which comprises a powerful algorithm. The electrical logs used are gamma-ray (GR), resistivity (ILD), neutron porosity (NPHI), and density (RHOB), while the output is in the form of images. The training data consists of 4 wells in Jambi sub-Basin, South Sumatera Basin, while the testing data comprises 5 wells with gamma-ray, resistivity, NPHI, and RHOB as input. Several scenarios are used to predict the facies model, namely training and validation dataset by using and isolating facies in well combination input, and with or without feature augmentation. Furthermore, the values collected were validated using F1 score. The result showed that 85.5% and 84.7% of F1 scores were allocated to training and validation to increase accuracy in scenarios without facies isolation and with feature augmentation. Therefore, the gradient boosting classifier method is reliable enough to characterize depositional facies in the associated area of interest.

1. Introduction
It is imperative to correlate lithofacies with well log data and other geological or geophysical information described from core samples for proper connection [1]. However, such information is not always provided to the entire length of the wells, therefore, pattern recognition needs to be applied for adequate prediction. Data can be separated into several subsets, with each element consisting of similar attributes, irrespective of its type. Furthermore, it is possible to classify elements based on some common characteristics obtained from the dataset because it is the basic principle of automatic classification of a new point [2]. The numerous approaches to automatic classification of data are divided into supervised and unsupervised methods [3].

The supervised methods are based on classification labels (targets) knowledge, while the unsupervised cluster the samples into subsets through statistical and other mathematical approaches [2]. This research used the merge interpretation method with well logs and core data in Jambi Sub-Basin, South Sumatera Basin area, to classify the depositional facies.

2. Problem Formulation
Well log data contains most information capable of determining depositional facies. Therefore, through logging, it is possible to obtain a detailed description of rock formations at different depth levels by measuring a wide variety of its properties [4]. However, due to the high cost of coring, it is important to find another method capable of determining the depositional facies. One of such methods is the use of machine learning to predict facies.

The application of supervised learning algorithms in the machine enables it to recognize the complicated correlations between various data sets. It also allows for the extraction of essential
information to improve the precision of reservoir quality assessment, minimize exploration risks, and identify break-even points [5].

3. Facies Classification Using Gradient Boosting Classifier
In this research used 4 electrical logs as input, namely Gamma Ray (GR), Induction Resistivity (ILD), Neutron Porosity (NPHI), and Density (RHOB) for radioactivity formation, electrical resistivity, porosity of facies, and density of lithology in the subsurface, respectively.

After a brief determination of the lithology from the interpreted depth, the facies were classified from the basic well log form that characterizes a depositional environment such as a cylindrical, irregular, bell, funnel, symmetrical, and asymmetrical [6]. However, this research only used 3 forms in this area, namely cylindrical, symmetrical, and serrated. The cylindrical is associated with facies accumulation which is heterogenic to the shallow water environment. The symmetrical occur where there is a conformity combination of bell-funnel form and a combination of coarsening-finering upward. It forms as a result of reworked offshore build-up, from regressive to transgressive shoreface. The serrated form is associated with a storm-dominated shelf and distal deep marine slope, indicating thin sand interbedded with shale.

Based on well log form, 3 lithofacies characterizing the layer at depth d of the well is proposed as follows.
1. Distributary Channel Mudstone Sandstone (DCMdSt)
2. Shoreface Delta Mudstone Sandstone (SDMdSt)
3. Deep Marine Slope Interbedded Mudstone Sandstone (DMSMdSt)

The available attributes are concatenated from the machine learning process and associated with each depth and a feature vector as defined in equation 1.

\[ f_{d,w} = [f_{d,w}^{GR}, f_{d,w}^{Res}, f_{d,w}^{Por}, f_{d,w}^{Den}] \] (1)

In addition, each depth is associated with a class label \( c_d \), \( \epsilon \{DCMdSt, SDMdSt, DMSMdSt\} \) which indicates the facies associated with the layer.

Gradient boosting classifier works by splitting training data observations into different subsets [8]. A limited number of features are selected from each subset to train a separate decision tree, thereby solving a gradient-based minimalization problem. A new feature vector \( f \) is classified by testing each tree in the forest to provide a candidate label. After that, the results from all trees are merged into a single decision [4].

4. Research Methodology
The gradient-boosting classifier (GBC) attempts to decrease error by resampling and varying the weights for individual weak learners to increase classification accuracy [7]. This method provides a metric to assess the relative influence of each parameter included in the classifier [8]. GBC also provides the output in images, symbols, and graphics with its various characteristics.

This research uses the general supervised machine learning flow divided into training, validation, and testing phases. During training, a set of labeled observations from used wells in Jambi Sub-Basin, South Sumatera Basin, is used to determine a function that maps features to class labels. After training, the model formed must be validated with input data which is not used as a training data set with an accuracy value close to the obtained result. When the validation output is far from the training result, it is called underfitting or overfitting. In the testing phase, the classifier is used to estimate class labels from any unlabelled feature vector \( f \) obtained from new well logs [4].

The electrical logs used to characterize the depositional facies are gamma-ray (GR), resistivity (ILD), neutron porosity (NPHI), and density (RHOB). These parameters are used to train, validate and test data. This study utilized 4 wells as the training and validation data, namely E, I, K, and L, and 5 wells as the testing data, including C, G, J, M, and N. The workflow of this research is shown in Figure 1.
5. Results and Discussion
The input data used in this research are 3 facies with GR, ILD, NPHI, and RHOB logs. The training and validation data were obtained from E, I, K, and L wells, while the testing data were from C, G, J, M, and N wells. Correlations between parameters are shown in Figure 2. The result shows that correlations of the logs well illustrate the facies. This is seen through facies log plots according to the depth of each well, which is adequately described as shown in Figure 3. Therefore, to ensure the data is useful, feature distribution was carried out as shown in Figure 4.
Figure 3. Facies log plots with training and validation data, namely E, I, K, and L wells. The log of each well illustrates its facies such as distributary channel mudstone sandstone (DCMdSt) coloured in dark blue, shoreface delta mudstone sandstone (SDMdSt) coloured in light blue, and deep marine slope interbedded mudstone sandstone (DMSMdSt) coloured in purple. Lithofacies with low GR, low ILD, high NPHI and low RHOB represents low radioactivity, high resistivity, high porosity, and low density.
Figure 4. Feature Distribution is between fellow logs such as NPHI with NPHI have good facies distributions which are showed by the curves formed. Meanwhile, the facies distributions between different logs such as ILD with GR are not good enough.

Figure 5. Facies Distribution is the most common facies in each well with distributary channel mudstone sandstone (DCMdSt), followed by shoreface delta mudstone sandstone (SDMdSt), and deep marine slope interbedded mudstone sandstone (DMSMdSt).
Figure 6. Features per well for all logs with the parameters consisting of the same amount of high accuracy value.

The distributary channel mudstone sandstone facies (DCMdSt) are the most commonly found in each well, as shown in Figure 5. Meanwhile, Figure 6 shows that all input needs to be balanced to get a high accuracy value with a similar amount of data.

In this research, the authors conducted 4 scenarios to obtain the best results needed to predict lithofacies in this area. The first is training and validation dataset using isolated facies in well combination input with feature augmentation. The second is similar to the first and without feature augmentation. The third is training and validation dataset without isolated facies and with feature augmentation. Meanwhile, the last is similar to the third without feature augmentation. Furthermore, the average values of F1 and confusion matrix scores were used to strengthen the accuracy results, while the standard scalerfunction was used to normalize the functions. From the 4 scenarios model, the best average F1 score and confusion matrix score from the third scenario for both training and validation with the accuracy of the F1 score average are 0.855 (training) and 0.847 (validation), as shown in Figure 7.

Figure 7. Interaction of facies in the confusion matrix in scenarios 1 (A), 2 (B), 3 (C), and 4 (D). The best model is scenario 3 with F1-Score 0.855.
Figure 8 shows the normalized confusion matrix consisting of predicted and true labels. However, it is majorly associated with determining the value of false positives and negatives. Therefore, to acquire good accuracy, both amounts need to be smaller than true positives and true negatives. In the confusion matrix, good results are shown by the interaction between facies in producing the most significant number when compared with others. From scenario 3, the relationship between DCMdSt, SDMdSt, and DMSMdSt is obtained with values of 0.92, 0.79, and 0.68. The accuracy value of the detector is obtained by averaging diagonal values of the confusion matrix. A perfect detector tends to produce a diagonal confusion matrix which found that the best lithofacies predicted by the model is distributary channel mudstone and sandstone (DCMdSt), while the worst is deep marine slope interbedded mudstone sandstone (DMSMdSt). This is possible because the DCMdSt facies is the largest in this study while the DMSMdSt is the least. Another interesting fact is the number of facies of shoreface delta mudstone sandstone (SDMdSt) and deep marine slope interbedded mudstone sandstone (DMSMdSt) which are predicted to be distributary channel mudstone sandstone (DCMdSt). This is the basis for increasing the amount of data on the two facies to enable them to produce accurate predictions.

Figure 8. Normalized Confusion Matrix. Dark blue part shows good value, which is close to 1. This matrix shows relationships between fellow facies in scenarios 1 (A), 2 (B), 3 (C), and 4 (D).
At the end of this research, classification techniques were used to the test data collected. The selection of test data is based on wells that need to have the same input data as the training and validation such as GR, ILD, NPHI, and RHOB. The wells used as the testing data are C, G, J, M, and N. The disadvantage of using this method is that it consists of numerous facies, therefore it tends to affect the accuracy. Therefore, the alternative solution is to reduce the amount of facies to obtain accurate values.

6. Conclusion
In conclusion, the gradient boosting classifier method help in determining depositional facies in any area of interest with limited core data. The input data used to determine it are GR, ILD, NPHI, and RHOB logs, while F1 score and confusion matrix were used as the accuracy techniques. The training and validation dataset without isolated facies in scenario 3 showed an F1 score of 85.5 % on training and 84.7% on validation. This is in accordance with the result of the predicted labels to the test data, which shows the classification of depositional facies. The accuracy value is high, indicating that the gradient boosting classifier method effectively determines the process.

Acknowledgments
The authors are grateful to the Reservoir Geophysics, Department of Physics, Geophysics and Geology Study Program Universitas Indonesia for supporting this research.

References
[1] Sanchetta A C, Leite E P and Honório B C Z 2013 Facies recognition using a smoothing process through fast independent component analysis and discrete cosine transform. Computers & Geosciences 57, 175-182
[2] MacQueen J 1967 Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* Vol. 1, No. 14, pp. 281-297.

[3] Duda R O and Hart P E 1973 *Pattern classification and scene analysis* Vol. 3, pp. 731-739. New York: Wiley.

[4] Bestagini P, Lipari V and Tubaro S 2017 A machine learning approach to facies classification using well logs. In *SEG technical program expanded abstracts 2017* (pp. 2137-2142). Society of Exploration Geophysicists.

[5] Dev V A and Eden M R 2019 Formation lithology classification using scalable gradient boosted decision trees. *Computers & Chemical Engineering* **128**, 392-404.

[6] Catuneanu O, Galloway W E, Kendall C G S C, Miall A D, Posamentier H W, Strasser A and Tucker M E 2011 Sequence stratigraphy: methodology and nomenclature. *Newsletters on stratigraphy* **44**(3), 173-245.

[7] Schapire R E, Freund Y, Bartlett P and Lee W S 1998 Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics* **26**(5), 1651-1686.

[8] Friedman J H 2001 Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232.