

Comparison Of Facies Estimation Using Support Vector Machine (SVM) And K-Nearest Neighbor (KNN) Algorithm Based on Well Log Data

Urip Nurwijayanto Prabowo^{1*}, Akmal Ferdiyan¹, Sukmaji Anom Raharjo¹, Sehah¹,
Arya Dwi Candra²

¹ Department of Physics, FMIPA, Universitas Jenderal Soedirman, Jl. dr. Soeparno 61, Purwokerto 53123, Indonesia

² Oil and Gas Production Engineering Department, Politeknik Energi dan Mineral Akamigas, 58315, Indonesia

*Corresponding author: urip.np@unsoed.ac.id

Received : October 22, 2022

Received in revised from August 05, 2023

Accepted : August 20, 2023

Online : August 30, 2023

Abstract – Facies classification is the process of identifying rock lithology based on indirect measurements such as well log measurements. Usually, the facies are classified manually by experienced geologists, so it takes a long time and is less efficient. In this paper, two machine learning (Support vector machine and K-Nearest Neighbor) were adopted to increase the effectiveness and shorten the time process of facies classification in Z Field, Indonesia. The machine learning algorithm was carried out in 4 steps, i.e. data selection, training phase, verification, and validation stage. The machine learning input data are density log, gamma ray log, resistivity log, SP log; and the output facies target are Sandstone, Siltstone, Claystone, and Limestone. The data is divided into train data for the training process and test data to validate the machine learning output. In Support vector machine results, the training accuracy is 70.1% and the testing accuracy is 47.4%, while in K-Nearest Neighbor results, the training accuracy is 70.1% and the testing accuracy is 63.3%. This result showed K-Nearest Neighbor has better accuracy than the support vector machine in facies classification in the Z field.

Keywords: log data, SVM, KNN, facies

Introduction

Facies classification is an identification of rock lithology based on indirect measurements such as well log measurements. The results of facies classification can be used to correlate important reservoir characteristic and facilitate building field scale reservoir models (Al-Anazi & Gates, 2010). Generally, facies classification process is carried out manually by experienced geologists, so it takes a relatively long time and is inefficient (Babikir et al., 2023; Sudarmaji et al., 2021). To overcome the shortcomings, many machine learning applications have been carried out on well-log data to perform facies classification automatically (Al-Mudhafar, 2017; Al-Mudhafar et al., 2022; M. Hall, 2016; Imamverdiyev & Sukhostat, 2019; Sudarmaji et al., 2021). Machine Learning (ML) derives a complex relationship between variables by correlating, clustering, or partitioning the well-log data (Wood, 2020). It can increase the effectiveness and efficiency in finding solutions to geophysical interpretation problems on complex data (Dell'Aversana, 2019; Merembayev et al., 2021) and reservoir quality assessments with lower cost and time (Dixit et al., 2020).

Machine learning is divided into unsupervised and supervised learning algorithms. Unsupervised learning is an algorithm that discovers output without any prior knowledge or given output data and supervised learning tries to discover the output by mapping the relationship of a given input to given output using testing and implementing steps (Singh et al., 2020). The researchers compared several machine learning algorithms in facies classification problem using well-logging data and the results conclude that the supervised learning algorithm has a better facies estimation accuracy than the unsupervised learning algorithm (Hardanto & Wulandhari, 2021; Mohamed et al., 2019). Many supervised algorithms have been adopted for facies classification such as support vector machine (B. Hall, 2016), K-nearest neighbor (Pratama et al., 2020), decision tree (Rahimi & Riahi, 2022),

probabilistic neural networks (Al-Mudhafar, 2017), random forest (Ippolito et al., 2021) and boosting algorithm (Al-Mudhafar et al., 2022).

In this study, we implemented a supervised learning algorithm i.e. support vector machine (SVM) and K-Nearest Neighbor (KNN) to automatically classified the facies of field Z, Indonesia. We examined the two supervised algorithms to achieve the best accuracy in facies classification.

Materials and Methods

This study uses log data from field Z, Indonesia that was acquired in 2010. The data contains 3 wells which have 4 parameters of well logging data (input) resulting from wireline logging measurements and 1 facies label (output). The facies data used in this study are based on mud log data. Well data from wireline logging measurements used in this study are density (FDC), gamma ray (GRST), resistivity (RT), and SP (SP). The facies labels used in the well data are (1) Sandstone, (2) Siltstone, (3) Claystone, and (4) Limestone.

This study was carried out in 4 steps. First, in the data selection stage, two wells are selected as training data and one well as test data. The amount of training data is three times of the test data. The training data from the two wells were then separated by 80% for the training phase and 20% for the verification phase. The parameter data used in this study can be seen in Figure 1. The training and test data were normalized so that the input features have a uniform scale to help in faster convergence of the algorithm

Second, in training phase, the machine learning algorithm is trained to recognize and form a pattern of relationships between input and output. In this study, the input used is 80% of the training data from wireline logging measurements and the output used is facies data. Third, the verification steps to measure the accuracy of the machine learning training results. At this stage, 20% of the training data that were not used in the training phase were used. Machine learning algorithms make estimations on this data and the results are matched with existing facies data. The level of estimation accuracy is determined based on the confusion matrix. Fourth, the validation steps to validate machine learning capabilities in making estimates using different data sets. If the previous stage used data from the same well, then the validation stage used data from different wells, namely test data. The results of the facies estimations will be compared with the existing facies data to determine the level of accuracy of the estimation.

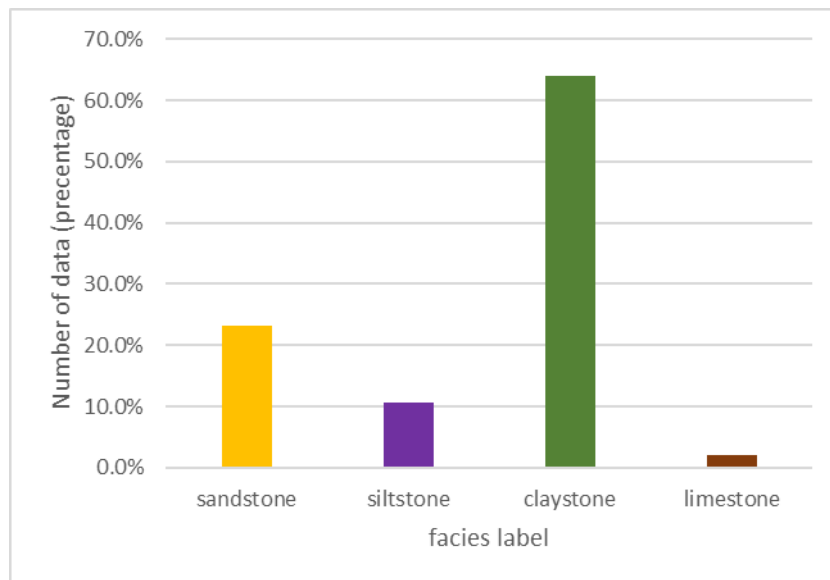


Figure 1. Distribution of facies on training data

This study uses SVM and KNN algorithms. SVM algorithms based on statistical learning techniques proposed by (Vapink, 1995) and used to perform classification and regression (Drams, 2019). The SVM algorithm divides data attribute space using hyperplane (decision-making surface) (Dell'Aversana, 2019). SVM calculate the optimal hyperplane distance between each data and hyperplane to clustering the data. The optimal

hyperplane is the optimal/closest distance between the data points and the hyperplane (decision-making surface) that called the support vector (Hardanto & Wulandhari, 2021).

K-Nearest Neighbors (KNN) is a supervised-machine learning method that maps each point (data point) at a location in multidimensional space. (Merembayev et al., 2021). This algorithm aims to classify new objects into certain groups based on the majority value of the category of the number of (k) nearest neighbors so determining the value of k is very important. (Pratama et al., 2020). This algorithm is sensitive to scale, so data standardization is needed to eliminate scale differences (Merembayev et al., 2021). The advantage of the KNN algorithm is that it can respond quickly to changes in the input data. While the drawback of KNN is that it often has difficulty drawing accurate conclusions and is not strong enough when dealing with data that is full of noise (Pratama et al., 2020).

In this study, the accuracy of the algorithm is based on the Confusion Matrix. Confusion Matrix is a table consisting of the number of rows of testing data estimated to be true and false by the model (Sudarmaji et al., 2021). The calculation of accuracy using the confusion matrix table is shown in Equation 1.

$$accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (1)$$

where TP is True Positive which is positive data that is estimated to be correct. TN is True Negative which is negative data that is estimated to be correct. FP is False Positive, which is positive data that is estimated to be wrong. FN is False Negative, it is negative data which is estimated wrong.

Results

SVM

The estimation accuracy of the SVM algorithm is influenced by the C parameter (regularization factor) and the gamma parameter. The C parameter controls the trade-off between the complexity of the model and its ability to reproduce outliers in the data. This imposes a limit on the number of support vectors which then allows for variations in hyperplane roughness (Halotel et al., 2020). A large C value will try to correctly classify more examples of the training data, but will probably follow the training data pattern too much and fail to generalize when classifying new/test data. If C is too small, then the model will not be good at fitting outliers and will have a large error in the training data (Mohamed et al., 2019). The Gamma parameter (γ) describes the size of the radial basis function or how far two vectors in the feature space need to be considered close (Mohamed et al., 2019). Poor parameter selection can lead to over or underfitting (Halotel et al., 2020).

In this study, the selection of parameters C and gamma was based on a cross-validation graph (C and gamma values) using training and test data as shown in Figure 2. Optimal parameters are achieved when the cross-validation error of parameters C and gamma reaches a minimum (Halotel et al., 2020). Based on Figure 2, in this research data, the optimal value is obtained at C = 10 and gamma = 1.

The results of SVM algorithm accuracy on the training data are determined based on the confusion matrix in Table 1. The training results show that SVM can correctly estimate 23 Sandstone data, 145 Claystone data, and 1 Limestone data and cannot estimate Siltstone data correctly from the training dataset. The SVM is incorrect estimate 39 Sandstone data as Claystone, 23 Siltstone data as Claystone, 5 Claystone data as Sandstone, and 4 Limestone data as Claystone.

The accuracy of the SVM algorithm on the test data is determined based on the confusion matrix in Table 2. The test data in Table 2 was not included in the training process to evaluate the estimations accuracy of trained machine learning on data with different characteristics from the training data. The trained SVM can estimate 26 Sandstone data, 164 Claystone data and cannot estimate the Siltstone and Limestone data from test data. The trained SVM is incorrect estimate 43 Sandstone data as Claystone, 14 Siltstone data as Claystone, 150 Claystone data as Sandstone, and 4 Limestone data as Claystone. The accuracy of training and test data was calculated using Equation 1 based on the correct estimations. The comparison between the estimation results of the facies and the actual facies on the test data can be seen in Figure 3.

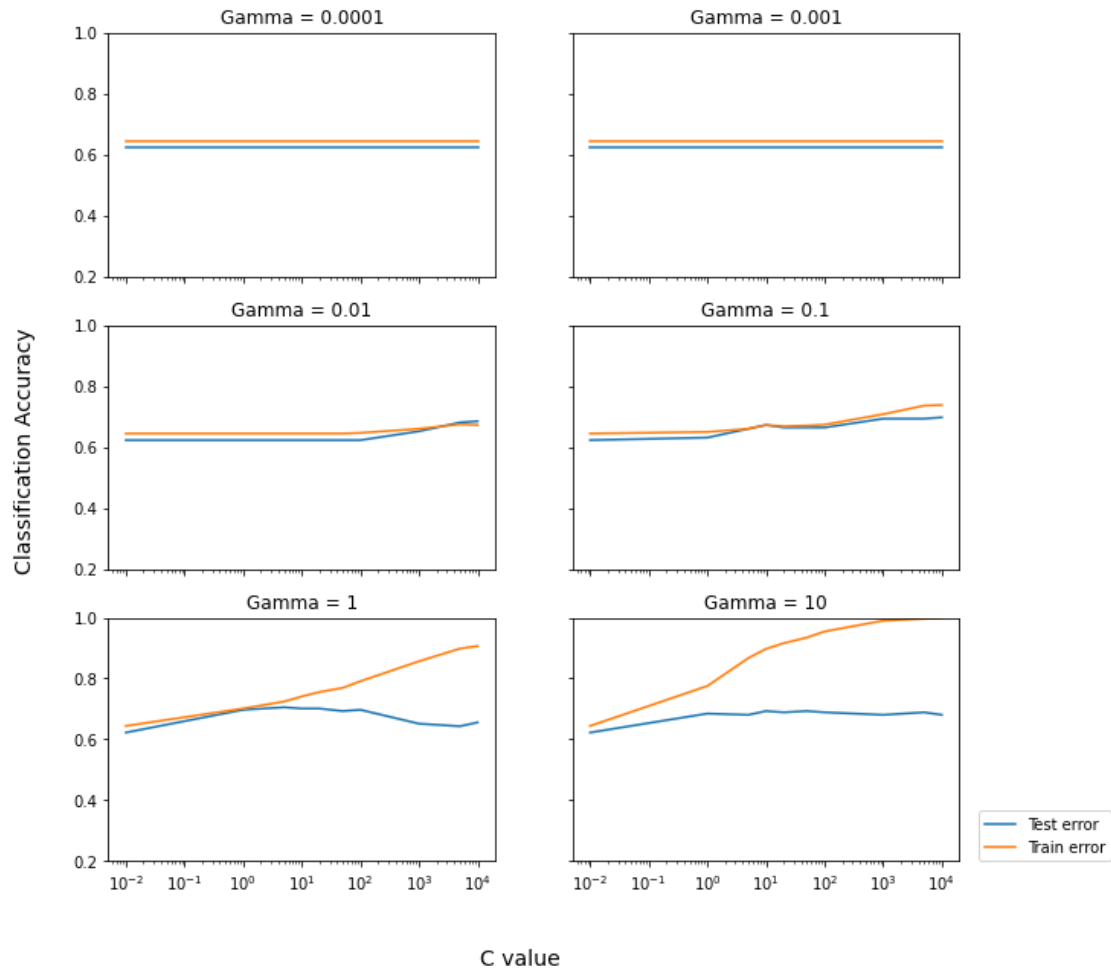


Figure 2. Cross-validation of parameter values of C and gamma.

Table 1. Confusion matrix of training data

Prediction True	Sandstone	Siltstone	Claystone	Limestone	Total
Sandstone	23	0	39	1	63
Siltstone	0	0	23	0	23
Claystone	5	0	145	0	150
Limestone	0	0	4	1	5

Table 2. Confusion matrix of test data

Prediction True	Sandstone	Siltstone	Claystone	Limestone	Total
Sandstone	26	0	43	0	69
Siltstone	0	0	14	0	14
Claystone	150	0	164	0	314
Limestone	0	0	4	0	4

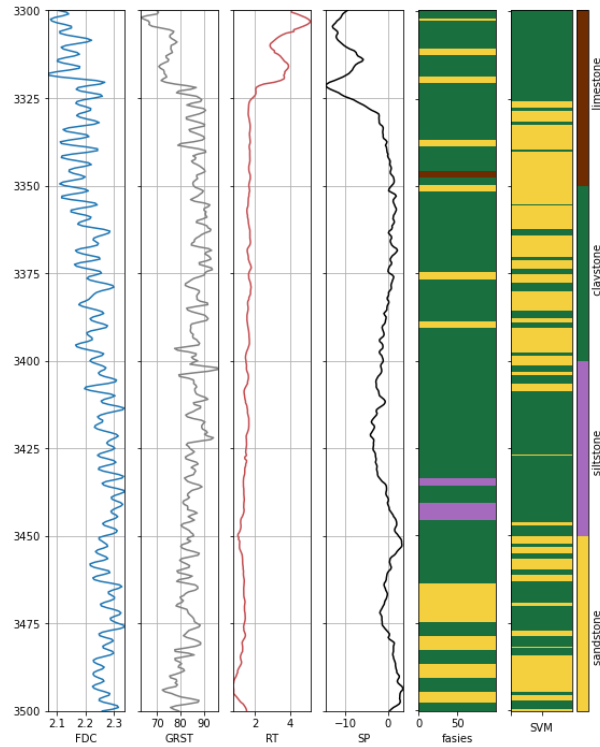


Figure 3. Comparison between the estimation results of the facies and the actual facies on the test data

KNN

The estimation accuracy in the KNN algorithm is influenced by the use of the parameter value k (number of nearest neighbors). In this study, we used the variation of the value of k with an interval of 10 and the results of the accuracy of the facies estimation based on the confusion matrix can be seen in Figure 4. The estimation accuracy in training data tends to decrease with the increase of k value and the estimation accuracy in test data is increased by the increase of k value. Figure 5 shows the comparison between the actual facies on the test data and the estimation results of KNN in all k intervals.

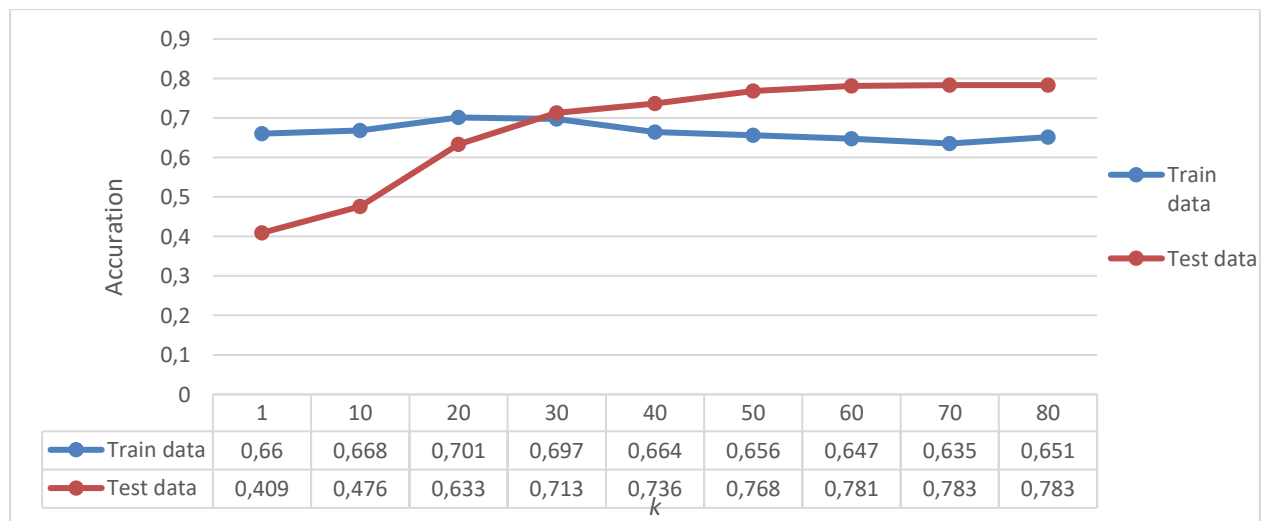


Figure 4. Accuracy of facies estimation on the KNN algorithm

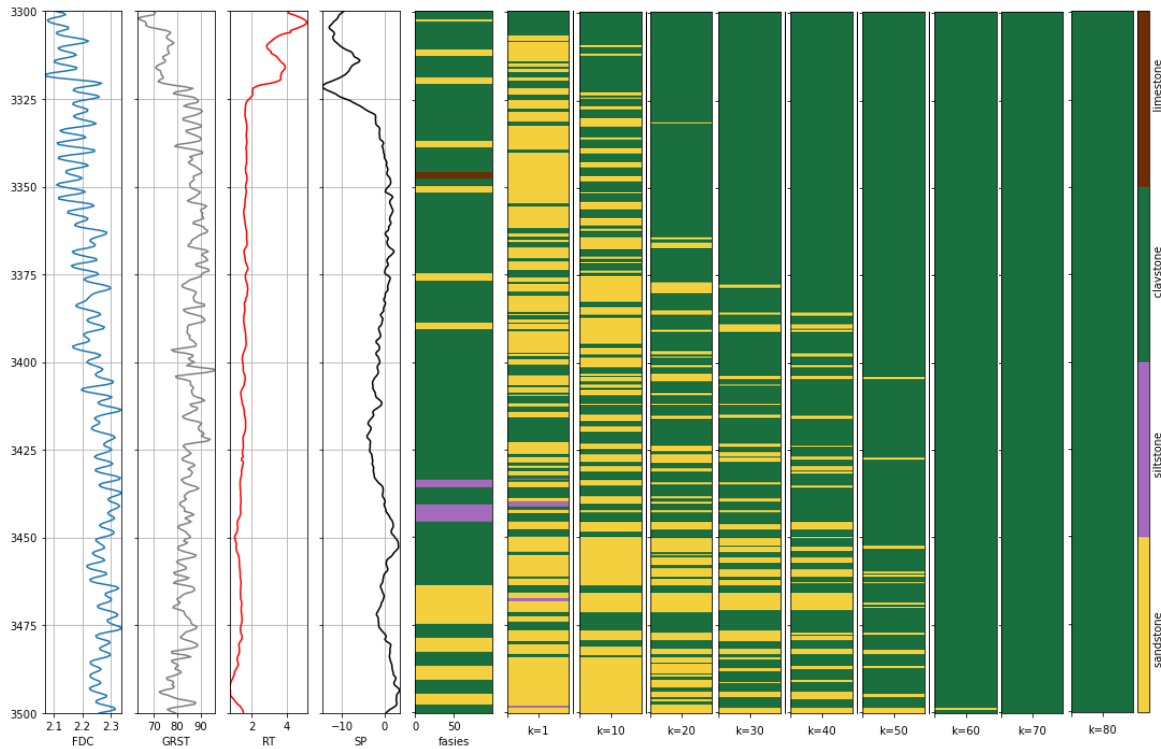


Figure 5. Comparison of well log data, facies, and facies prediction results for each k value

Discussion

SVM

Facies that have been estimated correctly are shown by the diagonal matrix in Table 1 and the accuracy of the training data facies estimation is 0.701. Table 1 also shows that the algorithm has not been able to estimate the Siltstone facies and can only estimate 1 Limestone facies correctly. This is probably due to the small amount of training data for the Siltstone and Limestone facies (Figure 1), which makes the algorithm unable to recognize the pattern from the output of the facies. The SVM is supervised learning which conducted data classification based on specified labels, so it failed to predict classification without or with less supervision from label data (Hardanto & Wulandhari, 2021). The SVM is only good at dealing with datasets that have a large number of features with respect to the number of samples (Kotsiantis, 2007). The training data should be an even distribution of the number of facies labels.

The results of the accuracy of the SVM algorithm facies estimation on the test data (validation stage) are determined based on the confusion matrix in Table 2 and the facies estimation accuracy on the test data is 0.474. The estimation accuracy of the test data is lower on the training data because, in the training phase, the algorithm adjusts the input and output patterns, while in the test data (validation stage), SVM applies the pattern that has been compiled to the new data/test data. Figure 3 and Table 2 also show that the SVM algorithm has not been able to estimate the Siltstone and Limestone facies. This result is following the accuracy of the training data which cannot estimate the facies.

KNN

Figure 4 shows that the accuracy of the facies estimation on the test data gets higher with the greater value of k used. Meanwhile, the accuracy of the estimation of test data from k = 1 to k = 20 increased but began to decrease from the value of k = 30 to the value of k = 80. From the value of k=30, the accuracy of the test data is higher than the training data, which indicates that the algorithm failed to learn the input-output pattern in the training data. This may be because the amount of training data is too small and the facies distribution is uneven

so that with a large number of k values, it will assume that almost all facies labels are the same neighbors and different facies labels with a small number are considered outliers which are not taken into account. Machine learning algorithms are highly dependent on the amount of data and the data gap in the input dataset can impact the prediction accuracy (Mandal & Rezaee, 2019). Prediction accuracy of machine learning can be improved by adding more input features and increasing the number of training datasets.

In Figure 5, the value of $k=30$ to $k=80$ shows that the algorithm failed to recognize the input-output pattern in the training data so it was dominant in estimating the facies as Claystone. Claystone has the most facies distribution in the training data (Figure 1), so a large k value causes the algorithm to guess all the closest data as Claystone facies.

In the results of the KNN algorithm, the best facies estimation accuracy is obtained at a value of $k = 20$ because the accuracy of the training data and test data is not much different and the accuracy of the training data is higher. Figure 5 shows that the KNN algorithm is also not able to estimate the Siltstone and Limestone facies as in the SVM algorithm. The facies estimation accuracy value in the KNN algorithm is higher than in the SVM algorithm. So, the KNN can handle the data distribution of the Z field better than SVM.

Conclusion

The support vector machine (SVM) and K-Nearest Neighbor (KNN) has adopted to automatically classified the facies of field Z, Indonesia. At the training stage, machine learning parameters must be determined correctly because they greatly affect the results of prediction accuracy such as C and γ parameters in SVM and k values in KNN. The results of facies estimation using the KNN algorithm have better accuracy than SVM. However, the two algorithms have not been able to determine the Siltstone and Limestone facies due to the small number of these facies in the training data so the algorithm has not recognized the facies input-output pattern yet. An even distribution of facies labels on the training data is needed so that the algorithm can recognize patterns and can estimate all types of facies correctly.

Acknowledgment

The author would like to thank LPPM Unsoed for its support in this research. Thanks to Brandon Hall for providing a tutorial on facies estimation using Python on GitHub.

References

- Al-Anazi, A., & Gates, I. D. (2010). On the capability of support vector machines to classify lithology from well logs. *Natural Resources Research*, 19(2), 125–139. <https://doi.org/10.1007/s11053-010-9118-9>
- Al-Mudhafar, W. J. (2017). Integrating well log interpretations for lithofacies classification and permeability modeling through advanced machine learning algorithms. *Journal of Petroleum Exploration and Production Technology*, 7(4), 1023–1033. <https://doi.org/10.1007/s13202-017-0360-0>
- Al-Mudhafar, W. J., Abbas, M. A., & Wood, D. A. (2022). Performance evaluation of boosting machine learning algorithms for lithofacies classification in heterogeneous carbonate reservoirs. *Marine and Petroleum Geology*, 145. <https://doi.org/10.1016/j.marpetgeo.2022.105886>
- Babikir, I., Elsaadany, M., Sajid, M., & Laudon, C. (2023). On the training sample size and classification performance: An experimental evaluation in seismic facies classification. *Geoenergy Science and Engineering*, 226. <https://doi.org/10.1016/j.geoen.2023.211809>
- Dell'Aversana, P. (2019). Comparison of different machine learning algorithms for lithofacies classification from well logs. *Bollettino Di Geofisica Teorica Ed Applicata*, 60(1), 69–80. <https://doi.org/10.4430/bgta0256>
- Dixit, N., Mccolgan, P., & Kusler, K. (2020). Machine Learning-Based Probabilistic Lithofacies Prediction from Conventional Well Logs : A Case from the Umiat Oil Field of Alaska. *Energies*, 13(4862), 1–15.
- Dramschi, J. S. (2019). Machine Learning in 4D Seismic Data Analysis Deep Neural Networks in Geophysics [Dissertation]. Technical University Denmark.
- Hall, B. (2016). Facies classification using machine learning. *Leading Edge*, 35(10), 906–909. <https://doi.org/10.1190/tle35100906.1>
- Hall, M. (2016). A user guide to the geophysical tutorials. *The Leading Edge*, February 2016, 190–191.

- Halotel, J., Demyanov, V., & Gardiner, A. (2020). Value of Geologically Derived Features in Machine Learning Facies Classification. *Mathematical Geosciences*, 52(1), 5–29. <https://doi.org/10.1007/s11004-019-09838-0>
- Hardanto, L. T., & Wulandhari, L. A. (2021). Lithofacies Classification Using Supervised and Semi-Supervised Machine Learning Approach. *International Journal on Advanced Science, Engineering and Information Technology*, 11(2), 542–548. <https://doi.org/10.18517/ijaseit.11.2.11764>
- Imamverdiyev, Y., & Sukhostat, L. (2019). Lithological facies classification using deep convolutional neural network. *Journal of Petroleum Science and Engineering*, 174, 216–228. <https://doi.org/10.1016/j.petrol.2018.11.023>
- Ippolito, M., Ferguson, J., & Jenson, F. (2021). Improving facies prediction by combining supervised and unsupervised learning methods. *Journal of Petroleum Science and Engineering*, 200. <https://doi.org/10.1016/j.petrol.2020.108300>
- Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. In *Informatica* (Vol. 31).
- Mandal, P. P., & Rezaee, R. (2019). Facies classification with different machine learning algorithm – An efficient artificial intelligence technique for improved classification. *ASEG Extended Abstract*, 2019(1), 1–6. <https://doi.org/10.1080/22020586.2019.12072918>
- Merembayev, T., Kurmangaliyev, D., Bekbauov, B., & Amanbek, Y. (2021). A Comparison of Machine Learning Algorithms in Predicting Lithofacies: Case Studies from Norway and Kazakhstan. *Energies*, 14(1896), 1–16.
- Mohamed, I. M., Mohamed, S., Mazher, I., & Chester, P. (2019). Formation lithology classification: Insights into machine learning methods. *Proceedings - SPE Annual Technical Conference and Exhibition*, 2019-September. <https://doi.org/10.2118/196096-ms>
- Pratama, H., Syahputra, L. A., Albany, M. F., Abdullah, A., Kusuma, E., Astuti, W., Mujihardi, B., & Suhardja, S. K. (2020). Early Results of Comparison between K-Nearest Neighbor and Artificial Neural Network Method for Facies Estimation. *Geofisika*, 18(01), 7–13.
- Rahimi, M., & Riahi, M. A. (2022). Reservoir facies classification based on random forest and geostatistics methods in an offshore oilfield. *Journal of Applied Geophysics*, 201. <https://doi.org/10.1016/j.jappgeo.2022.104640>
- Singh, H., Seol, Y., & Myshakin, E. M. (2020). Automated Well-Log Processing and Lithology Classification by Identifying Optimal Features through Unsupervised and Supervised Machine-Learning Algorithms. *SPE Journal*, 25(5), 2778–2800. <https://doi.org/10.2118/202477-PA>
- Sudarmaji, Winata, E., Pratama, P., Hidayat, W., & Prakoso, S. (2021). The Implementation of Machine Learning in Lithofacies Classification using Multi-well Logs Data. *Aceh International Journal of Science and Technology*, 10(April), 9–17. <https://doi.org/10.13170/aijst.10.1.18749>
- Vapink, V. (1995). *The nature of statistical learning theory*. Springer.
- Wood, D. A. (2020). Predicting porosity, permeability and water saturation applying an optimized nearest-neighbour, machine-learning and data-mining network of well-log data. *Journal of Petroleum Science and Engineering*, 184. <https://doi.org/10.1016/j.petrol.2019.106587>