




Article

A Comparison of Machine Learning Algorithms in Predicting Lithofacies: Case Studies from Norway and Kazakhstan

Timur Merembayev ¹, Darkhan Kurmangaliyev ², Bakhbergen Bekbauov ² and Yerlan Amanbek ^{1,*}

¹ Department of Mathematics, Nazarbayev University, Nur-Sultan 010000, Kazakhstan; timur.merembayev@gmail.com

² KazMunayGas Engineering LLP; Nur-Sultan 010000, Kazakhstan; D.Kurmangaliyev@kmg.kz (D.K.); B.Bekbauov@kmg.kz (B.B.)

* Correspondence: yerlan.amanbek@nu.edu.kz

Abstract: Defining distinctive areas of the physical properties of rocks plays an important role in reservoir evaluation and hydrocarbon production as core data are challenging to obtain from all wells. In this work, we study the evaluation of lithofacies values using the machine learning algorithms in the determination of classification from various well log data of Kazakhstan and Norway. We also use the wavelet-transformed data in machine learning algorithms to identify geological properties from the well log data. Numerical results are presented for the multiple oil and gas reservoir data which contain more than 90 released wells from Norway and 10 wells from the Kazakhstan field. We have compared the machine learning algorithms including KNN, Decision Tree, Random Forest, XGBoost, and LightGBM. The evaluation of the model score is conducted by using metrics such as accuracy, Hamming loss, and penalty matrix. In addition, the influence of the dataset features on the prediction is investigated using the machine learning algorithms. The result of research shows that the Random Forest model has the best score among considered algorithms. In addition, the results are consistent with outcome of the SHapley Additive exPlanations (SHAP) framework.

Keywords: machine learning; well log data; lithology classification



Citation: Merembayev, T.; Kurmangaliyev, D.; Bekbauov, B.; Amanbek, Y. A Comparison of Machine Learning Algorithms in Predicting Lithofacies: Case Studies from Norway and Kazakhstan. *Energies* **2021**, *14*, 1896. <https://doi.org/10.3390/en14071896>

Academic Editor: Abbas Mardani

Received: 13 February 2021

Accepted: 22 March 2021

Published: 29 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

It is important to understand the geological structure of formations based on the provided data in many applications. The key features of the complex subsurface can be defined by geophysicists. The geophysicists' experience on the finding of lithotypes can help to improve the accuracy of the labels in the well logs. Such experience requires many hours of work and additional data from different sources such as seismic survey, cores, etc. One of the possible solutions is to use the machine learning algorithms to accelerate the accurate prediction process in a systematic way. To achieve appropriate accuracy of results, the data-driven algorithms require a large amount of data which should be used in a balanced way in the training procedure. Traditionally, the most common features of a region are identified by geophysicists and then uncommon features are estimated by additional well log data using the knowledge of relationships among lithotypes such as PS, RHOB, and NPHI.

To determine lithotypes, geophysicists perform work in stages: first, Shale and sandstone are determined, often gamma logs are used, sometimes for control of PS, RHOB, and NPHI. After these rocks, the isolation of uncommon lithotypes is made only by their characteristic features. The more features (curves), the better the accuracy of determining the lithotypes. An inexperienced geologist without knowledge of the geology of the field may not accurately determine similar lithotypes; therefore, the use of trained models will solve the problem of the lack of knowledge among geophysicists about the field.

Machine learning can be an effective tool to enrich geoscience workflows. Geostatistical approaches were proposed in many studies [1–4] to reduce the uncertainty of the subsurface property of using the large datasets.

There are several works regarding application of data analysis methods for mining areas [5,6]. The importance of lithofacies detection for uranium mining is discussed and investigated in [7,8] using machine learning algorithms to solve multilabel lithofacies classification. The in situ leaching of uranium requires a better understanding of the permeable and impermeable rock types.

The authors of [9] have made comparisons of machine learning algorithms using scikit-learning framework (MLPClassifier, the DecisionTreeClassifier, the RandomForestClassifier, and the SVC) for data from offshore wells. Algorithms have been applied to three standard data templates and a practical data template in a lithology classification problem for wells from International Ocean Discovery Program (IODP) Expeditions. We used a dataset from the lithology subdivision in GP (group GP), G1 (group 1), G2 (group 2), and G3 (group 3). The comparison analysis showed that the multilayer perceptron MLP method had better results in the lithology classification for the practical template: lithology of the G2 group.

In [10], the authors proposed using embedded feature selection (EFS) and LightGBM to predict the permeability of a reservoir. Result of EFS was generated based on five features: DEPTH, AC, DEN, FMIT, and GR out of 22 features and was equal 0.9457 (R2). Furthermore, the authors made a comparison of several methods of selection: the mutual information regression (MIR) in FFS and the recursive feature elimination (RFE) in WFS. Commonly used feature selection methods include filter feature selection (FFS) and wrapper feature selection (WFS). The same comparison was done for LightGBM: Random Forest and XGBoost. The best result was from EFS+LightGBM: R2 of 0.9712, RMSE of 0.5959.

The authors of [11,12] presented the application of oil production exploration and development data to generate high-performance predictive models and optimal classifications of geology, reservoirs, and fluid characteristics. The deep learning algorithms have the perspective to solve problems in geoscience in piratically lithology classification as well [13–15].

In [16], the authors investigated data preprocessing methods for well logs such as a dimensional reduction and wavelet analysis in order to improve the accuracy of the group method of data handling (GMDH) for lithological classification. Wavelet analysis was used for the decomposition of the log signals for the algorithm (GMDH). The authors of [17] proposed using the continuous wavelet transform of the well log data to detect geological boundaries. One of the applications of the wavelet coefficient is to measure the edge of the boundary strength. The boundary strength is a measure of the geological thickness of units. In the method, instead of solving multivariate classification, additional features were generated to detect the boundaries of the formations. The multi-element geochemical data taken from 259 drill holes were studied and its efficiency was shown for the data with a maximum depth of 600 m.

In this paper, we investigate the prediction of lithofacies using machine learning algorithms for the geological data of Kazakhstan and Norway. We consider machine learning methods such as KNN, Decision Tree, Random Forest, XGBoost, and LightGBM with and without wavelet transformed data. Gamma-ray (GR), medium deep reading resistivity measurement (RMED), compressional waves sonic log (DTC), neutron porosity log (NPHI), bulk density log (RHOB), etc. are considered as the input data of the machine learning models. In addition, the results of the supervised learning are provided in the SHapley Additive exPlanations (SHAP) visualization framework by indicating significant well logs. Our research question is the following: how can some supervised machine learning algorithms accurately predict lithofacies based on the geophysical well log data from Norway and Kazakhstan fields?

The rest of the paper is organized as follows. In the next section, we describe the wavelet transformation, data analysis, and machine learning algorithms. Numerical results of algorithms are presented in Section 3. Section 4 concludes the paper.

2. Methodology

We first describe the wavelet transformation and then the flowchart of workflow for machine learning algorithms. Next, the data analysis and data preparation are presented. We briefly describe the considered machine learning algorithms for supervised multi-labels classification.

2.1. Wavelet Transformation

We use the Gaussian wavelet transformation for the edge detection in the geology formation. The second-order derivative of the Gaussian function is also known as the Mexican hat wavelet. Inflection points of the Mexican hat wavelet represent edges of objects in the signal. Application of wavelet transformation to the given signal generates new artificial data which can be useful for further analysis.

The physical meaning of the wavelet transform is to calculate the joint energy spectrum of signals in the frequency-time domain and identify both the frequency and time information of the distinct modes [18].

Wavelet transformation decomposes a geophysical log into a combination of signals at different frequencies. It allows determining what frequency bands of log is noise and what frequency band is actual data. It provides a one-to-one mapping of the original log, so we can go back and forward between the original and transformed data.

The integral wavelet transform of a function $f(x)$ with respect to a mother wavelet is given by

$$W_{\psi}(s, \tau) = \int_{-\infty}^{+\infty} f(x) \psi_{s, \tau}(x) dx \quad (1)$$

where

$$\psi_{s, \tau}(x) = \frac{1}{\sqrt{s}} \psi\left(\frac{x - \tau}{s}\right) \quad (2)$$

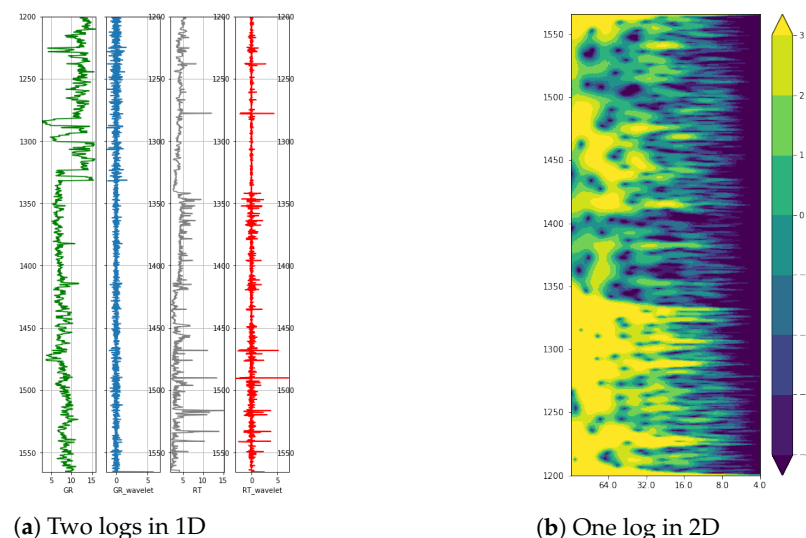
where $s > 0$, τ are the scale factor and shift, respectively.

For creation wavelet transformation, we used the Ricker wavelet, also known as the “Mexican hat wavelet”:

$$\psi(x) = \frac{2}{\sqrt{3\sigma}\pi^{1/4}} \left(1 - \left(\frac{x}{\sigma}\right)^2\right) \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad (3)$$

To illustrate the above explanation, we conduct wavelet transformation of the geophysical logs from Kazakhstan, see Figure 1.

To better display the result of the wavelet transformation of logs we use a log scale in Figure 1b. Figure 1a shows its application to the wavelet transform for two logs.



(a) Two logs in 1D

(b) One log in 2D

Figure 1. Result of applying continuous wavelet transformation.

We have followed the general workflow of a machine learning classifier which is illustrated in Figure 2. Our process of the classifier model consists of the following steps:

1. Data preprocessing.
2. Application of the wavelet transformation to generate new features.
3. Finding hyperparameters and construction of machine learning algorithm as a classifier of lithofacies.
4. Training of the model on the well log data with the labeled lithology by geophysicist or geologist.
5. Evaluation of the trained model of classifier according to specified score based on the test dataset.

The initial stage is started with a generation of new features from the current well logs. Next, training of the model for the new dataset, which includes wavelet-transformed well logs, is performed. The trained model is evaluated by estimation of the accuracy on the test dataset.

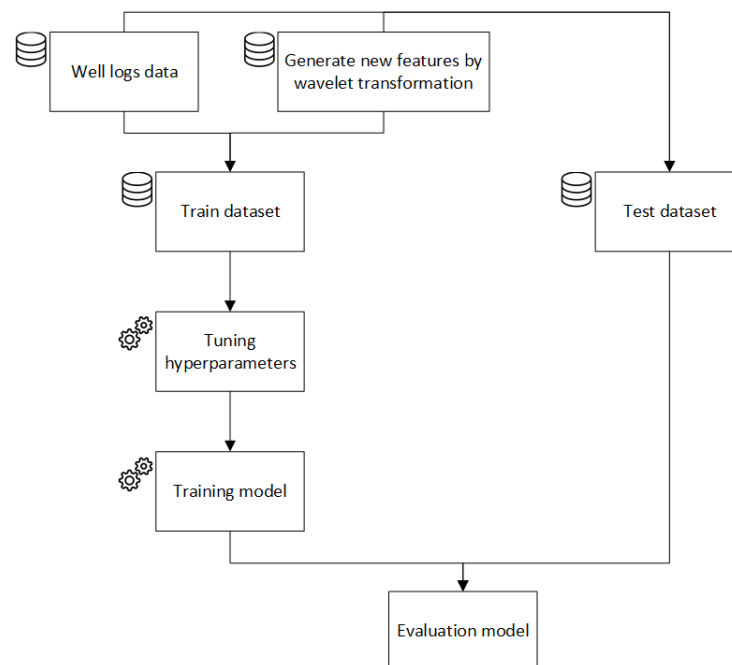


Figure 2. Flowchart of workflow for machine learning algorithms.

2.2. Data Analysis

We consider the well log data from an offshore field in the North Sea, near Norway. The study area contains 98 wells with a maximum depth of 5000 m. Dataset consists of interpreted lithofacies and well logs, 22 wireline log curves including gamma ray (GR), medium deep reading resistivity measurement (RMED), compressional waves sonic log (DTC), neutron porosity log (NPHI), and bulk density log (RHOB) and others. Digital measurements were recorded at 0.1 m intervals, see Table 1 and 2 for abbreviations and descriptions of the dataset.

Table 1. The well log abbreviations.

Log Name	Log Description
LITHOFACIES_LITHOLOGY	Interpreted Lithofacies
RDEP	Deep Reading Resistivity measurement
RSHA	Shallow Reading Resistivity measurement
RMED	Medium Deep Reading Resistivity measurement
RXO	Flushed Zone Resistivity measurement
RMIC	Micro Resistivity measurement
SP	Self Potential Log
DTS	Shear wave sonic log (us/ft)
DTC	Compressional waves sonic log (us(ft))
NPHI	Neutron Porosity log
PEF	Photo Electric Factor log
GR	Gamma Ray Log
RHOB	Bulk Density Log
DRHO	Density Correction log
CALI	Caliper log
BS	Borehole size
DCAL	Differential Caliper log
ROPA	Average Rate of Penetration
SGR	Spectra Gamma Ray log
MUDWEIGHT	Weight of Drilling Mud
ROP	Rate of Penetration
DEPTH_MD	Measured Depth
x_loc	X location of sample
y_loc	Y location of sample
z_loc	Z(TVDSS) location of sample

The interpreted lithofacies contains 12 classes. Lithofacies type corresponds to codes (number) which are used in machine learning training and prediction: 0: Sandstone, 1: Sandstone/Shale, 2: Shale, 3: Marl, 4: Dolomite, 5: Limestone, 6: Chalk, 7: Halite, 8: Anhydrite, 9: Tuff, 10: Coal, 11: Basement.

For data exploration we use a library Cegal <https://github.com/cegaldev/cegaltools>, accessed on 22 March 2021, which is the geoscience tool for loading, plotting, and evaluating well log data using python script. It is also an interactive tool to visualize data details and dependence. Figure 3 shows one well with its logs.

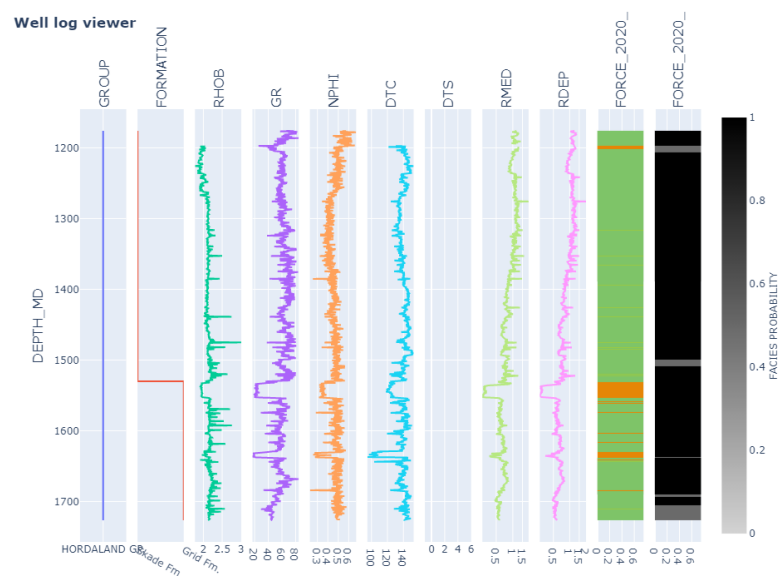


Figure 3. Visualization of well logs.

Distribution of lithology types in log scale are presented in Figure 4. We have a similar distribution of lithology classes for training and test datasets.

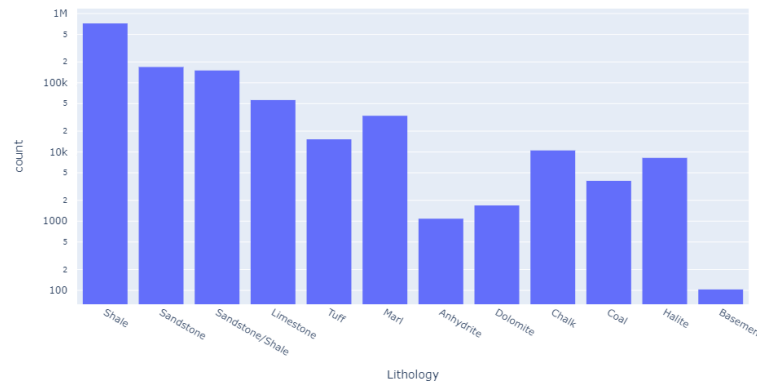


Figure 4. Histogram of lithology facies in log scale.

Table 2. The descriptions and abbreviations for full dataset.

Statistic Parameter	DEPTH_MD	CALI	RSHA	RMED	RDEP	RHOB	GR	NPHI	PEF	DTC	SP	BS
mean	2184.1	12.2	5.8	4.8	10.6	2.0	70.9	0.2	3.6	105.5	44.3	7.0
standard deviation	997.2	5.0	74.1	53.8	113.4	0.8	34.2	0.2	8.9	40.8	70.9	6.4
min	136.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	−999.0	0.0
25%	1418.6	8.9	0.0	0.9	0.9	2.0	47.6	0.0	0.0	83.7	0.0	0.0
50%	2076.6	12.4	0.6	1.4	1.4	2.2	68.4	0.2	2.9	105.3	40.4	8.5
75%	2864.4	15.7	1.5	2.6	2.5	2.5	89.0	0.4	4.6	139.3	70.4	12.3
max	5436.6	28.3	2193.9	1988.6	1999.9	3.5	1077.0	1.0	383.1	320.5	526.5	26.0

2.3. Data Preparation

The dataset contains some missing data. Key reasons for missing data are technical problems during acquisition data, cost optimization during geophysical logging, human factor, and others. We utilize the Missingo library [19] to detect the data gap from provided dataset. It helps to define logs with their location. In Figure 5, one well is presented and well logs contain missing data such as missing for full depth of well or with some gaps.

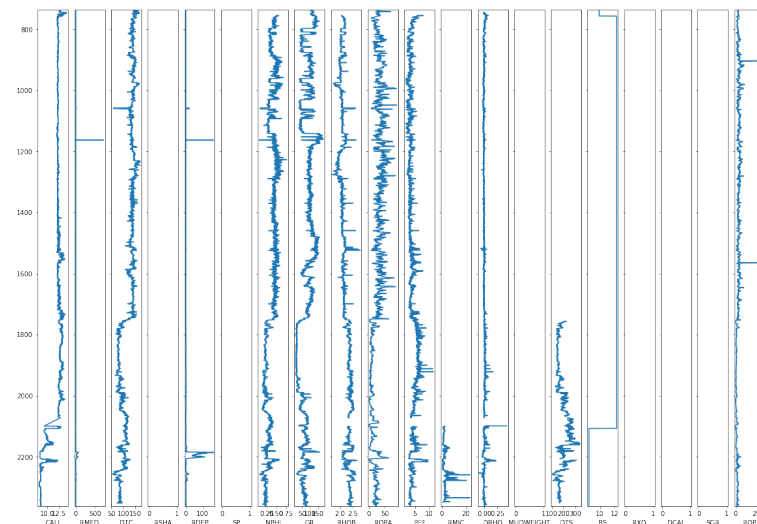


Figure 5. Real value of logs with missing data.

After careful study and statistical analysis of logs for missing data, we decided to concentrate the following logs, which have a smaller percentage of missing data: DEPTH_MD, CALI, RSHA, RMED, RDEP, RHOB, GR, NPHI, PEF, DTC, SP, and BS.

2.4. Algorithms

There are various machine learning algorithms, and each algorithm has its own advantages and disadvantages for solving geoscience problems. In this paper, we made a comparison analysis of five algorithms: K-nearest neighbors (kNN) [20], the Decision Tree [21], the Random Forest Classifier (RFC) [22], and the extreme gradient boosting (XGBoost) [23], LightGBM [24]. They are also explored with and without the generation additional features obtained from the wavelet transformation. In this research, we used “scikit-learn” [25], the developed python framework for utilization of kNN, Decision Tree, and Random Forest classifier. XGBoost and LightGBM have their own python framework.

K-Nearest Neighbors (kNN) is a machine learning method that has been used for data mining [20]. Each point (data point) has location in a multidimensional space, where the space consists of axis or features of current datasets. The trained model defines an optimal count of neighbors for the trained dataset and when we have a new (test) data point the model finds the K nearest neighbors for the test dataset. KNN has the advantage of being nonparametric. The method is sensitive to scale, so standardizing data is mandatory to eliminate differences in scale. It can be an issue when the dataset is very large, the application of special methods can solve the issue to decrease the space.

Decision tree methods are data mining methods, and they have been successfully used for classification problems. Decision trees were developed by Morgan and Sonquist in 1963, and they applied the algorithm for determinants of social conditions [21]. One advantage of the decision trees is that they are computationally fast and can handle high-dimensional data. On the other hand, a single decision tree can overfit on the data and the algorithm is greedy; therefore, it keeps growing deeper in the tree.

The random forest was introduced by Breiman as a learning tree classifier of an ensemble [22]. The key idea of the algorithm is to take the values of a random vector from an aggregated bootstrap sample (train dataset) and then to train many decision trees. However, the trained tree can have a lot of trees, thus it requires more computational resources.

The main advantage of the XGBoost is parallelization. XGBoost is a scalable version of the gradient boosting machine algorithm and showed efficiency in several machine learning applications. In [23], the XGBoost is an ensemble of classification and regression trees and works for data with nonlinear features. The key idea is to use weak trees and enhancement of trees accuracy for each iteration, taking account the error in prediction from the previous result of a weak tree, the next tree classifier is trained to take into account the error of the already trained ensemble.

LightGBM is a relatively new framework and has a wide application in machine learning/data science applications. The main issue of gradient boosting algorithms is that the algorithm processes all data to gain the result of possible separation points, which impacts performance. This method has been modified to improve the optimal search technique [24].

Based on the train dataset we calculated the main hyperparameters for Random Forest, see Table 3. The main hyperparameters for XGBoost and LightGBM are presented in Tables 4 and 5, respectively.

Table 3. Main hyperparameters for Random Forest Classification.

Hyperparameter	Symbol	Parameter Value
The number of trees in the forest	n_estimators	200
The maximum depth of the trees	max_depth	70
The minimum number of samples required to be at a leaf node	min_samples_leaf	1
The minimum number of samples required to split an internal node	min_samples_split	2
The number of features for the best split	max_features	10

Table 4. Main hyperparameters for XGBoost Classification.

Hyperparameter	Symbol	Parameter Value
Number of boosted trees to fit	n_estimator	526
Minimum sum of instance weight	min_child_weight	11
Maximum depth of a tree	max_depth	12
Minimum loss reduction required to make a further partition on a leaf node of the tree	gamma	8
L1 regularization term on weights	lambda	1.36
L2 regularization term on weights	alpha	0.23
Boosting learning rate	learning_rate	0.73

Table 5. Main hyperparameters for LightGBM Classification.

Hyperparameter	Symbol	Parameter Value
Number of boosted trees to fit	n_estimator	216
Minimum sum of instance weight	min_child_weight	4.12
Maximum depth of a tree	max_depth	11
Minimum loss reduction	min_split_gain	0.08
L1 regularization term on weights	lambda_l1	2.69
L2 regularization term on weights	lambda_l2	4.27
Boosting learning rate	learning_rate	0.05

The prediction performance of the algorithms is evaluated by tree statistical quality indicators: Jaccard metrics (accuracy), Hamming loss, and Penalty metrics. The reader is referred to Table 5.

The Jaccard metric is computed as

$$L_{Hamming}(y, \hat{y}) = \frac{1}{n_{labels}} \sum_{j=0}^{n_{labels}-1} 1(\hat{y}_j \neq y_j). \quad (4)$$

The Hamming loss is defined as

$$J(y_i, \hat{y}_i) = \frac{|y_i \cap \hat{y}_i|}{|y_i \cup \hat{y}_i|}. \quad (5)$$

To estimate the accuracy of models, a penalty matrix is used and is derived from the averaged input of a representative sample. This allows for petrophysical unreasonable predictions to be scored by a degree of “wrongness”.

The scoring matrix is defined as follows:

$$S = -\frac{1}{N} \sum_{i=0}^N A_{\hat{y}_i y_i} \quad (6)$$

where N is the number of samples, \hat{y}_i is the true lithology label, and y_i is the predicted lithology label.

3. Results

Computations are performed on a desktop machine (3.2 GHz Intel Core i7 8700 processor) with 32 GB RAM. Tuning hyperparameters and cross-validations operations are a time-consuming, therefore they are computed in parallel mode using eight cores.

3.1. Lithofacies Prediction for the Norway Data

The comparison of the selected algorithms has been performed on 12 features and additional seven features generated by wavelet transformation, for a total of 19 features. Table 6 shows scores of models on the test dataset by the Jaccard metrics (accuracy), Hamming loss, and Penalty metrics. We observe that the RFC has the highest score on the test set with metrics Accuracy, Penalty matrix, and Hamming loss of 0.948, -0.1289 , and 0.0473, respectively. Thus, RFC was selected to provide a detailed analysis of lithofacies classification. The classification report for the RFC model (12 features) can be found in

Table 7. By evaluating the precision information from Table 7, we noticed that the lowest value were computed for Dolomite (4) and Coal (10). A reason for such values could be the lack of representation of these lithofacies classes in the dataset.

Table 6. Comparison of three scores models on the test dataset.

Models	Original Dataset (12)			Original and Generated Features (19)		
	Accuracy	Penalty matrix	Hamming loss	Accuracy	Penalty matrix	Hamming loss
kNN	0.926	−0.1796	0.0672	0.801	−0.5237	0.1969
Random Forest	0.948	−0.1289	0.0473	0.938	−0.1697	0.0624
Decision Tree	0.820	−0.4810	0.1832	0.8167	−0.4810	0.1826
XGBoost	0.855	−0.3812	0.1418	0.8621	−0.3681	0.1631
LightGBM	0.897	−0.2600	0.0984	0.9013	−0.2599	0.1378

Table 7. Classification report of RFC.

Lithofacies Class	Precision	Recall	f1-Score	Support
0	0.94	0.95	0.94	33,697
1	0.89	0.92	0.9	29,227
2	0.98	0.96	0.97	147,278
3	0.9	0.94	0.92	6447
4	0.46	0.87	0.61	185
5	0.81	0.94	0.87	9746
6	0.97	0.97	0.97	2085
7	1	0.99	0.99	1684
8	0.93	0.94	0.93	198
9	0.94	0.97	0.96	2954
10	0.73	0.9	0.8	586
11	1	1	1	16
accuracy			0.95	234,103
macro avg	0.88	0.95	0.91	234,103
weighted avg	0.96	0.95	0.95	234,103

To understand the good accuracy of the RFC model for lithology classification, we use the SHAP package to verify the results, which are consistent with another study [26]. SHAP is a good tool for explanation of the different models and it provides an important value for each features. SHAP builds an explanatory model for a single row–prediction pair to explain a result of prediction. The SHAP values are calculated by averaging the values over all possible features.

SHAP does not enable us to determine the probabilities of predicted classes in the multi-label classification. The explanation models (tree and kernel) cannot output probabilities due to the constraint associated with nonlinear transformations, but it provides the raw margin values of the objective function which fit the model.

Figure 6 shows the global importance for 12 classes which was calculated as the average of absolute SHAP values. SHAP ranks the input features by the mean SHAP value, the amount of the value provides the importance of the feature in prediction of certain class (higher means more influential). The GR feature influences on the model prediction in all lithology classes, other features have less influence if compared with GR feature.

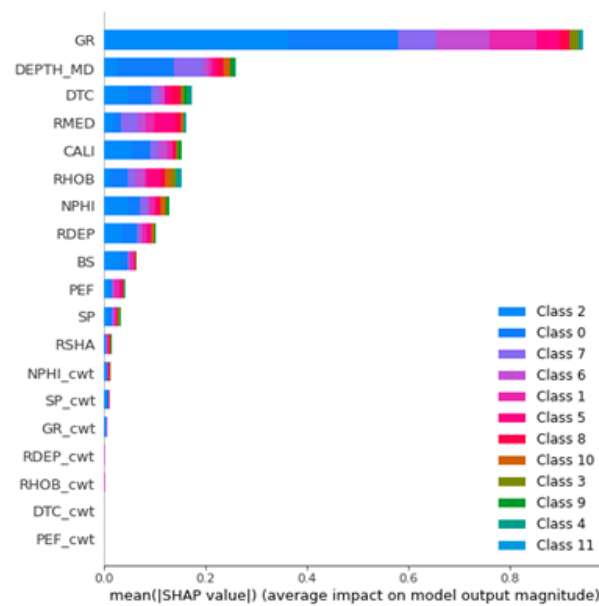


Figure 6. Importance factor of each input variable for the features.

Figure 7 gives additional explanation of the model, the influence of input features on the model prediction of lithology classes, and their distributions. SHAP calculates a Shapley value for input features and instances, and plots it on the figure. The y -axis is the input features with an order of importance for the model prediction from top to bottom. Each dot on plots is colored by the value of the selected variable, from low (blue) to high (red). SHAP chooses the selected variable for each feature based on its correlation values. Figure 7a–l illustrates the influence of features on each lithology classes. We note that the GR feature has high SHAP values and it impacts the model prediction of the following lithology classes: Sandstone, Limestone, Chalk, Halite, Anhydrite, Coal. However, for Sandstone/Shale, Shale, Marl, and Basement classes; the GR feature tends to have negative SHAP values for their lithology classes. We can see the influence of the GR, DTC, and RHOB variables on almost all lithology classes. On the other hand, some lithology classes such as Tuff and Coal have different important features in the model prediction.

Due to different nature of the Coal properties from the other classes it was found that RHOB and NPHI were significant features in the prediction of Coal. Moreover, RHOB, DTC, RMED, and GR were dominant features in the forecast of Dolomite and Limestone lithology.

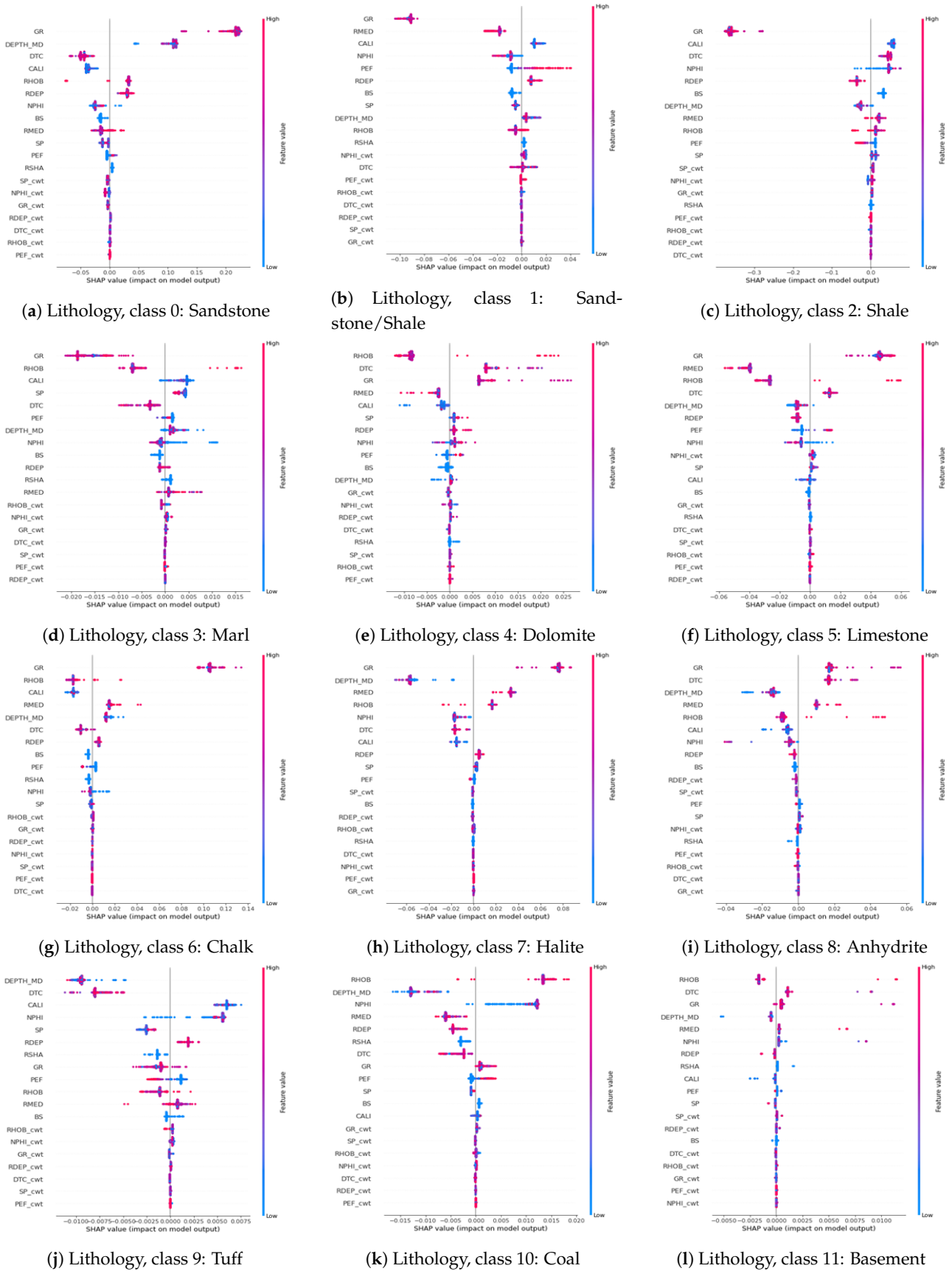


Figure 7. Summary plots for influence models prediction by classes for the Norway data.

3.2. Lithofacies Prediction for the Kazakhstan Data

We carried out numerical experiments in the above-mentioned way for wells in the Kazakhstan oil and gas field; the study area contains 10 wells with a maximum depth of 1700 m. The lithology for the field primarily consists of clay, coal, limestone, dolomite, and sand. The data contain well logs such as thermal neutron porosity, caliper, gamma-ray, temperature, resistivity, sonic, and others. The information from well logs was recorded at every foot of the formation where it is logged across.

Data was split into train and test datasets split to 75% and 25%, respectively. In Figure 8, the distribution of lithologic types for train and test dataset in log scale is presented, and distributions have a similar shape. The total dataset is 59,423 rows and 23 features, the train dataset contains rows 47,538 and 23 features, and the test dataset contains 11,885 rows and 23 features.

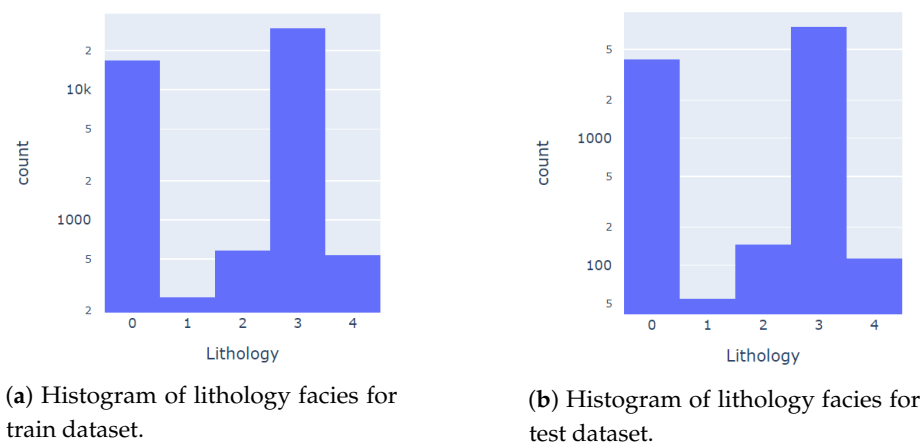


Figure 8. Summary plots for various failure modes of columns.

Based on the result for the Norway dataset, we used the Random Forest Classifier for data from the Kazakhstan field which showed a better result on three metrics. In Table 8, there are three scores that summarize the performance of the Random Forest Classifier on the test datasets for the different lithofacies types. The Random Forest Classifier shows a precisely result as well.

Table 8. Comparison three scores of Random Forest Classifier for the test dataset.

Models	Original Dataset (12)			Original and Generated Features (19)		
	Accuracy	Penalty matrix	Hamming loss	Accuracy	Penalty matrix	Hamming loss
Random Forest	0.977	−0.061	0.0227	0.975	−0.068	0.0253

Class 2 (dolomite) was not precisely predicted, see Table 9. The reason for such values can be an imbalanced dataset.

Table 9. Classification report of RFC for Kazakhstan field test data.

Lithofacies Class	Precision	Recall	f1-Score	Support
0	0.97	0.99	0.98	4045
1	0.78	0.89	0.83	47
2	0.38	0.88	0.53	64
3	0.99	0.97	0.98	7620
4	0.96	0.99	0.97	109
accuracy			0.98	11,885
macro avg	0.82	0.94	0.86	11,885
weighted avg	0.98	0.98	0.98	11,885

Figure 9 shows the global importance for five classes. The PHIE (prediction of effective porosity) and PHIT (prediction of total porosity) features influence the model prediction in all Clay (3) and Sand (0) classes.

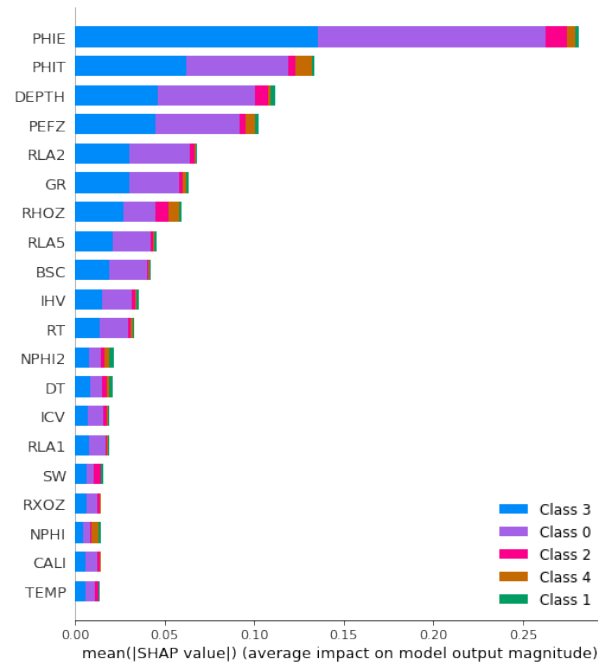


Figure 9. Importance factor of Random forest model for each input variable for the features.

Figure 10 shows the influence of input features on the model prediction of lithology classes. The SHAP values of the Sand class are higher for PHIE and PHIT features. The colors of PHIE and PHIT values indicate some threshold that can split the positive and negative influence of PHIE and PHIT features on the model prediction, see Figure 10a. The SHAP values of the Limestone class are higher for PHIE and PHIT features, see Figure 10b. The model found dependence by depth for the Limestone, also likely the class is located on defined depth for this field. The SHAP values of Dolomite, Clay, and Coal classes are higher for PHIE, PHIT, PEFZ, and RHOZ features, see Figure 10c,d. The High values of PHIE and RHOZ features are a positive influence on the model prediction of Dolomite. Lower values of PHIE and PHIT features are a positive influence on model prediction of Clay. Lower values of PHIE, PEFZ, and RHOZ features are a positive influence on model prediction of Coal.

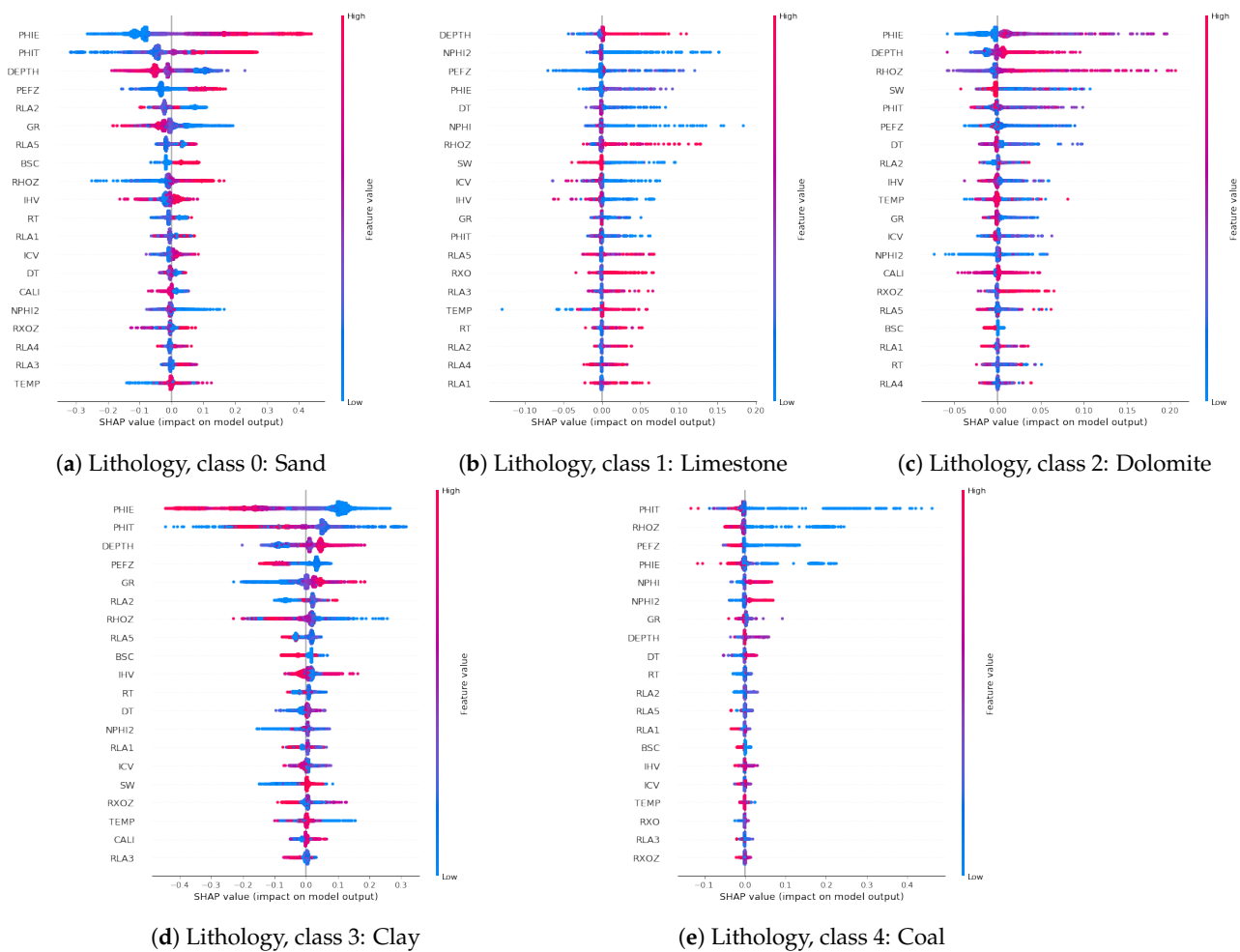


Figure 10. Summary plots for influence models prediction by classes for Kazakhstan dataset.

4. Conclusions

This paper analyzes the supervised learning algorithms for the well log data from Norway and Kazakhstan with or without the additional wavelet-transformed features. Our focus was on the data of offshore and onshore reservoirs. The findings suggest that our fitted Random Forest model shows the best results among the considered algorithms. The cross-validation methodology was applied in the machine learning models. Machine learning algorithms, in particular Random Forest method, can be integrated to specific geophysical software to proceed with a lithology classification automatically based on well logs without using information about sludge or core samples, and others. This process can improve efficiency of finding solution for some geophysical interpretation problems.

The nature of the decision tree methods (kNN, Random Forest, Decision Tree, etc.) is verified as set of good methods for the well log data, as it enables solving the nonlinear problem of the lithological classifications. The random forest model has an accuracy of 0.948, penalty matrix score of -0.1289 , hamming loss score of 0.0473 for 12 features and an accuracy of 0.938, penalty matrix of -0.1697 , and hamming loss of 0.0624 for 19 features including features which were generated from wavelet transformation of the data. Scores of algorithms that used the data and wavelet-transformed data are similar to scores of algorithms that trained only on the data without wavelet transformation. However, we believe that such additional features could help for different problems (regression) in geoscience such as identification of permeability or porosity.

We used the SHAP framework to explore the impact of features on the targeted classification and to detect the complex relationships between features. The result of the SHAP in our dataset showed that the significant features on a prediction of some lithology

classes were GR, DTC, and RHOB. However, some classes such as Tuff and Coal can be detected by other features (NPHI and RDEP).

In our future research, we intend to concentrate on deep learning algorithm such as 1D-CNN, LSTM, and RNN for prediction of multi-label lithofacies classification, porosity, and permeability using the well log data.

Author Contributions: Conceptualization, T.M. and Y.A.; methodology, T.M. and Y.A.; software, T.M.; validation, D.K. and T.M.; formal analysis, T.M. and D.K.; resources, D.K. and B.B.; writing—original draft preparation, T.M., D.K. and Y.A.; writing—review and editing, T.M., B.B. and Y.A.; visualization, T.M.; supervision, Y.A.; funding acquisition, Y.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Nazarbayev University, Grant No. 110119FD4502, the SPG fund and the Ministry of Education and Science of the Republic of Kazakhstan, Grant No. AP08052762.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing is not applicable.

Acknowledgments: T.M. and Y.A. wish to acknowledge the research grant, No AP08052762, from the Ministry of Education and the Nazarbayev University Faculty Development Competitive Research Grant (NUFDRCG), Grant No 110119FD4502.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ohl, D.; Raef, A. Rock formation characterization for carbon dioxide geosequestration: 3D seismic amplitude and coherency anomalies, and seismic petrophysical facies classification, Wellington and Anson-Bates Fields, Kansas, USA. *J. Appl. Geophys.* **2014**, *103*, 221–231. [[CrossRef](#)]
2. Wang, X.; Yang, S.; Zhao, Y.; Wang, Y. Improved pore structure prediction based on MICP with a data mining and machine learning system approach in Mesozoic strata of Gaoqing field, Jiyang depression. *J. Pet. Sci. Eng.* **2018**, *171*, 362–393. [[CrossRef](#)]
3. Amanbek, Y.; Merembayev, T.; Srinivasan, S. Framework of Fracture Network Modeling using Conditioned Data with Sequential Gaussian Simulation. *arXiv* **2020**, arXiv:2003.01327.
4. Sun, Z.; Jiang, B.; Li, X.; Li, J.; Xiao, K. A Data-Driven Approach for Lithology Identification Based on Parameter-Optimized Ensemble Learning. *Energies* **2020**, *13*, 3903. [[CrossRef](#)]
5. Ai, X.; Wang, H.; Sun, B. Automatic Identification of Sedimentary Facies Based on a Support Vector Machine in the Arysium Graben, Kazakhstan. *Appl. Sci.* **2019**, *9*, 4489. [[CrossRef](#)]
6. Osintseva, N.; Danko, D.; Priezzhev, I.; Iskaziyeu, K.; Ryzhkov, V. Combination of classic geological/geophysical data analysis and machine learning: Brownfield sweet spots case study of the middle Jurassic Formation in Western Kazakhstan. In *SEG Technical Program Expanded Abstracts 2020*; Society of Exploration Geophysicists: Tulsa, OK, USA, 2020; pp. 2176–2180.
7. Merembayev, T.; Yunussov, R.; Yedilkhan, A. Machine learning algorithms for classification geology data from well logging. In Proceedings of the 2018 14th International Conference on Electronics Computer and Computation (ICECCO), Kaskelen, Kazakhstan, 29 November–1 December 2018; pp. 206–212.
8. Merembayev, T.; Yunussov, R.; Yedilkhan, A. Machine learning algorithms for stratigraphy classification on uranium deposits. *Proc. Comput. Sci.* **2019**, *150*, 46–52. [[CrossRef](#)]
9. Bressan, T.S.; de Souza, M.K.; Girelli, T.J.; Junior, F.C. Evaluation of machine learning methods for lithology classification using geophysical data. *Comput. Geosci.* **2020**, *139*, 104475. [[CrossRef](#)]
10. Zhou, K.; Hu, Y.; Pan, H.; Kong, L.; Liu, J.; Huang, Z.; Chen, T. Fast prediction of reservoir permeability based on embedded feature selection and LightGBM using direct logging data. *Measur. Sci. Technol.* **2020**, *31*, 045101. [[CrossRef](#)]
11. Tan, F.; Luo, G.; Wang, D.; Chen, Y. Evaluation of complex petroleum reservoirs based on data mining methods. *Comput. Geosci.* **2017**, *21*, 151–165. [[CrossRef](#)]
12. Kanaev, I.S. Automated Missed Pay Zones Detection Method Based on BV10 Member Data of Samotlorskoe Field. In *SPE Russian Petroleum Technology Conference*; Society of Petroleum Engineers: Houston, TX, USA, 2020.
13. Al-Mudhafar, W.J. Integrating well log interpretations for lithofacies classification and permeability modeling through advanced machine learning algorithms. *J. Pet. Explor. Prod. Technol.* **2017**, *7*, 1023–1033. [[CrossRef](#)]
14. Kim, S.; Kim, K.H.; Min, B.; Lim, J.; Lee, K. Generation of synthetic density log data using deep learning algorithm at the Golden field in Alberta, Canada. *Geofluids* **2020**, *26*. [[CrossRef](#)]

15. Zhang, D.; Yuntian, C.; Jin, M. Synthetic well logs generation via Recurrent Neural Networks. *Pet. Explor. Dev.* **2018**, *45*, 629–639. [[CrossRef](#)]
16. Shen, C.; Asante-Okyere, S.; Yevenyo Ziggah, Y.; Wang, L.; Zhu, X. Group method of data handling (GMDH) lithology identification based on wavelet analysis and dimensionality reduction as well log data pre-processing techniques. *Energies* **2019**, *12*, 1509. [[CrossRef](#)]
17. Hill, E.J.; Pearce, M.A.; Stromberg, J.M. Improving automated geological logging of drill holes by incorporating multiscale spatial methods. *Math. Geosci.* **2020**, *53*, 1–33. [[CrossRef](#)]
18. Pathak, R.S. *The Wavelet Transform*; Springer Science & Business Media: Berlin, Germany, 2009; Volume 4; p. 178.
19. Bilogur, A. Missingno: A missing data visualization suite. *J. Open Source Softw.* **2018**, *3*, 547. [[CrossRef](#)]
20. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [[CrossRef](#)]
21. Rokach, L.; Maimon, O. Decision trees. In *Data Mining and Knowledge Discovery Handbook*; Springer: Berlin, Germany, 2005; pp. 165–192.
22. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
23. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y. Xgboost: Extreme gradient boosting. In *Microsoft. R Package Version 0.4-2*; R Package Vignette: Madison, WI, USA, 2015; pp. 1–4.
24. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Proces. Syst.* **2017**, *30*, 3146–3154.
25. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
26. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Proces. Syst.* **2017**, *30*, 4765–4774.