



Article Well-Logging-Based Lithology Classification Using Machine Learning Methods for High-Quality Reservoir Identification: A Case Study of Baikouquan Formation in Mahu Area of Junggar Basin, NW China

Junlong Zhang¹, Youbin He^{1,*}, Yuan Zhang², Weifeng Li¹ and Junjie Zhang^{3,*} 💿

- ¹ School of Geosciences, Yangtze University, Wuhan 430100, China; zhang.junlong@outlook.com (J.Z.); liweifeng339@263.net (W.L.)
- ² Research Institute of Petroleum Exploration and Development, SINOPEC Jianghan Oilfield Company, Wuhan 430223, China; zhangyuan1.jhyt@sinopec.com
- ³ Global Research, RBC Capital Markets, Toronto, ON M5J 2J5, Canada
- * Correspondence: 100709@yangtzeu.edu.cn (Y.H.); junjiezhung@gmail.com (J.Z.)

Abstract: The identification of underground formation lithology is fundamental in reservoir characterization during petroleum exploration. With the increasing availability and diversity of well-logging data, automated interpretation of well-logging data is in great demand for more efficient and reliable decision making for geologists and geophysicists. This study benchmarked the performances of an array of machine learning models, from linear and nonlinear individual classifiers to ensemble methods, on the task of lithology identification. Cross-validation and Bayesian optimization were utilized to optimize the hyperparameters of different models and performances were evaluated based on the metrics of accuracy—the area under the receiver operating characteristic curve (AUC), precision, recall, and F1-score. The dataset of the study consists of well-logging data acquired from the Baikouquan formation in the Mahu Sag of the Junggar Basin, China, including 4156 labeled data points with 9 well-logging variables. Results exhibit that ensemble methods (XGBoost and RF) outperform the other two categories of machine learning methods by a material margin. Within the ensemble methods, XGBoost has the best performance, achieving an overall accuracy of 0.882 and AUC of 0.947 in classifying mudstone, sandstone, and sandy conglomerate. Among the three lithology classes, sandy conglomerate, as in the potential reservoirs in the study area, can be best distinguished with accuracy of 97%, precision of 0.888, and recall of 0.969, suggesting the XGBoost model as a strong candidate machine learning model for more efficient and accurate lithology identification and reservoir quantification for geologists.

Keywords: machine learning; supervised classification; lithology identification; well-logging; ensemble methods; gradient-boosted decision trees

1. Introduction

Lithology identification is a task of great significance in reservoir characterization for petroleum exploration and engineering [1]. It is the basis for reservoir quality assessment (e.g., porosity and permeability) and supports related geological research and drilling activities (e.g., sedimentary modeling, favorable zone prediction, and well planning) [2,3]. Well-logging has been utilized as an effective remote sensing measurement to predict underground formation lithology from a surface geophysical survey. Well-logging data contains rich geological information, which is a synthesized reflection of formation lithology and physical properties [4].

The idea of lithology identification from well-logging is to establish the relationship between petrological characteristics and logging curves. Typical lithologies are supposed to have their own specific logging responses. For example, the GR–RT (gamma ray– resistivity of true formation) crossplot is effective to recognize sandstone and mudstone in



Citation: Zhang, J.; He, Y.; Zhang, Y.; Li, W.; Zhang, J. Well-Logging-Based Lithology Classification Using Machine Learning Methods for High-Quality Reservoirs Identification: A Case Study of Baikouquan Formation in Mahu Area of Junggar Basin, NW China. *Energies* 2022, *15*, 3675. https://doi.org/ 10.3390/en15103675

Academic Editors: Mohamed Mahmoud and Zeeshan Tariq

Received: 29 March 2022 Accepted: 13 May 2022 Published: 17 May 2022

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



Copyright: © 2022 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/). conventional sand and shale reservoirs due to the fact that sandstone has relatively low GR log values and high responding RT, whereas mudstone behaves oppositely on GR and RT logs [2].

However, traditional logging interpretation depends heavily on expertise and human experience, which is labor intensive and time consuming, and often suffers from subjectiveness and inconsistency of expert experience [5]. Due to the complexity of the geological condition in unconventional reservoirs (e.g., carbonate, tight sandstone, or sandy conglomerate reservoir [6,7]) and the increasing diversity and amount of logging data, the traditional logging interpretation methods show great limitations. As a result, researchers are turning to more advanced methods for breakthroughs in lithology identification.

Machine learning techniques have been embraced by the oil and gas industry as alternative methods in addressing the complex and challenging problems it faces to enable automation, lift performance, or explore new solution paradigms [8]. With advances in algorithms, computational theories, and hardware such as graphic processing units, machine learning shows great advantages in learning complex patters and relationships from large amount of data [9]. Two primary classes of machine learning algorithms, namely, unsupervised and supervised learning methods, have been prescribed for lithology identification. Supervised learning methods use a set of training data to learn relationships between features and corresponding labels and build models that are predictive for previously unseen data. Supervised learning algorithms outperform by a substantial margin compared with nearly all unsupervised learning algorithms reported in lithology classification using well-logging data [10,11].

A wide variety of supervised learning methods have been reported in the task of lithology identification, including but not limited to Naïve Bayes [12], linear regression [13], k-Nearest Neighborhood (kNN) [14], support vector machine (SVM) [15,16], decision tree and its variants (e.g., random forests and boosting trees) [17–19], and artificial neural networks (ANN) [20,21]. However, as different experiments were carried out using different datasets with their own lithology classification schemes, it is hard to make a parallel comparison of those machine learning models. Recently, more studies have attempted to compare the performance of machine learning methods for lithology identification. Xin et al. [22] compared the performance of five machine learning methods for formation lithology identification using well-logging data from the Daniudui gas field (DGF) and the Hangjinqi gas field (HGF) and concluded that Gradient Tree Boosting classifier (GTB) and Random Forest had better accuracy than other three methods, namely, Naïve Bayes, SVM, and ANN. Dev et al. [23] tested three models from the family of gradient-boosted decision tree (GBDT) methods using data from DGF and HGF, and identified LightGBM and CatBoost as the preferred algorithms for lithology classification using well-logging data. Merembayev et al. [24] evaluated five machine learning algorithms including kNN, Decision Tree, Random Forest, XGBoost, and LightGBM on well-logging data from Norway and Kazakhastan for lithofacies classification. The results showed that Random Forest has the best score among considered algorithms.

In this study, we intend to make a more systematic and comprehensive comparison of machine learning methods for lithology identification using well-logging data. We categorize supervised machine learning methods into three groups, namely, linear individual classifiers, nonlinear individual classifiers, and ensemble methods, with increasing model complexity. We select several typical machine learning models within each group to evaluate their performance using well-logging data collected from 17 wells in our study area and try to answer three key questions:

- Do nonlinear individual classifiers always show better performance in terms of accuracy than linear individual classifiers for well-logging-based lithology classification?
- Do ensemble methods consistently outperform individual classification models and by what margin? Which (if any) is the superior ensemble method?
- How well can different lithology classes in our study be distinguished by the bestperforming models?

The rest of the paper is organized as follows: Section 2 introduces the study area and the well-logging dataset. Section 3 describes the machine learning methods included in our study for lithology identification, as well as the metrics used to evaluate their performance. Section 4 presents quantitative results of the experiments in terms of hyperparameter optimization, overall performance, and lithology classification results. Feature importance is also evaluated by the end of the section. The conclusions are summarized in Section 5.

2. Study Area and Dataset

The study area is located in the sandy conglomerate reservoir of the Baikouquan formation in the Mahu Sag of Xinjiang oilfield in the Junggar Basin (Figure 1), which is the main oil and gas exploration area in northwestern China. It was chosen for the availability of high-quality well-logging and corresponding core images.



Figure 1. Location of the study area (modified after Tao et al. [25]). (**A**) Large-scale map of Xingjiang Province, China. (**B**) Work location in northern Xinjiang Province. (**C**) Detailed map with typical coring well locations and the approximate scope of ancient fan deposits (including alluvial fans and fan deltas).

The dataset for the study consists of well-logging data with 9 properties acquired from 17 wells with close proximity to each other. Lithologic labels were interpreted from 520 m core images with 4156 data points for machine learning workflow development. The nine log properties include gamma ray (GR), self potential (SP), caliper log (CALI), shallow/medium deep/deep reading resistivity measurement (RESS/RESM/RESD), neutron porosity log (PHIN), bulk density log (RHOB), and interval transit time (DT). The description of the well-logging dataset is shown in Table 1.

The lithofacies identified from core images in Baikouquan Formation contain 3 classes: mudstone (M), sandstone (S), and sandy conglomerate (SC). The labeling scheme was designed to reduce the subjectivity that exists in core photograph interpretation and produce consistent and reliable labeling for the dataset to benchmark the performance of different machine learning algorithms. The prepared dataset consists of nine predictor variables and a lithology class as target variable.

Statistics	DEPTH	GR	SP	CALI	RESS	RESM	RESD	PHIN	RHOB	DT
mean	3820.63	55.44	-1.02	8.72	20.54	24.51	23.53	19.28	2.50	72.52
std	157.30	13.83	33.97	0.77	17.42	15.15	15.78	4.77	0.10	6.76
min	3279.63	29.29	-68.06	5.59	0.20	2.88	-1.69	9.01	1.85	58.31
25%	3806.88	45.43	-14.45	8.41	7.69	13.15	11.92	16.41	2.47	68.31
50%	3855.13	52.32	-9.53	8.50	16.48	22.14	20.92	18.21	2.52	70.52
75%	3882.38	62.63	-1.97	8.69	28.81	32.56	32.09	20.36	2.56	74.36
max	4350.25	109.375	102.362	16.82	126.388	114.335	104.926	47.4	2.681	118.61

Table 1. The description of the well-logging dataset.

3. Machine Learning Models for Lithology Classification

Machine learning has been increasingly used in data-driven discovery in geoscience to perform complex prediction tasks by learning patterns from large amounts of data, which cannot be easily done by a set of explicit rules [9]. There are four major machine learning paradigms: supervised learning, semisupervised learning, unsupervised learning, and reinforcement learning [26].

In supervised learning, the model attempts to predict a target value using a set of variables or features after learning the relationship between the predictors (the features) and the output in training. When the target variable is a categorical variable (also called a label), the problem is said to be a classification problem and the model is called a classifier.

This study explores an array of machine learning models and determines their performance in lithology classification using the well-logging data. These individual machine learning models can be broadly categorized as linear and nonlinear models, which will be detailed in Sections 3.1 and 3.2. Ensemble models are models combining individual models, which will be covered in Section 3.3.

3.1. Linear Models for Classification

Linear classification models refer to the class of classifiers that result in linear decision boundaries [27]. Linear models remain a popular choice in applications, especially when they can achieve adequate accuracy, for their straightforward implementation and better interpretability.

3.1.1. Logistic Regression

Logistic Regression (LR) is one of the most popular linear models for classification in the industry [27,28]. In the binary case, the model allows us to model the posterior probability of being 0 or 1 using a linear function of input variables or features, with a sum of one:

$$p(Y = 1|X = x) = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)}$$

$$p(Y = 0|X = x) = \frac{1}{1 + \exp(\beta_0 + \beta^T x)}$$
(1)

Applying the *logit* transformation, one obtains the log-odds ratio as

$$\log \frac{p(Y = 1 | X = x)}{p(Y = 0 | X = x)} = \beta_0 + \beta^T x$$
(2)

The input space is optimally divided by the decision boundary of hyperplane defined by $\{x|\beta_0 + \beta^T x = 0\}$ for which the log-odds ratio is zero, meaning that the posterior probability of being in one class or the other is equal.

5 of 15

3.1.2. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is another popular model that leads to linear decision boundary [28,29]. The LDA model separates two classes based on a set of observed characteristics *x* by modeling the class densities $f_1(x)$ and $f_0(x)$ of each class as multivariate normal distributions with means μ_1 and μ_{10} and the same covariance matrix Σ .

Again, we compute and investigate the log-ratio

$$\log \frac{p(Y=1|X=x)}{p(Y=0|X=x)} = \log \frac{f_1(x)}{f_0(x)} + \log \frac{\pi_1}{\pi_0} = \log \frac{\pi_1}{\pi_0} - \frac{1}{2}(\mu_1 + \mu_0)^T \Sigma^{-1}(\mu_1 + \mu_0) + x^T \Sigma^{-1}(\mu_1 + \mu_0)$$
(3)

where π_1 and π_0 are the prior probability of the two classes. The decision boundary or the hyperplane, defined by Equation (3), equals 0 and is linear in *x*. The hyperplane drawn by LDA aims to maximize the ratio of the between-group variance and the within-group variance, so the two classes can be best-separated [30].

3.2. Nonlinear Models for Classification

More advanced machine learning techniques have been developed to model complex patterns in data, which often result in nonlinear decision boundaries.

3.2.1. k-Nearest Neighbor

k-Nearest Neighbor (kNN) is a simple but effective classification method [31]. The approach consists of calculating the Euclidean distance of a new instance with each instance in the training sample that has already been labeled. Then, the class label of the new instance is assigned according to the major class of the k-nearest neighbors in the training set. kNN has the advantage of being nonparametric, but one has to carefully select k to achieve optimal classification results. The method is also sensitive to the scale of different features in multidimensional space; so, date standardization is required to eliminate the effect of scale differences in both training and test sets [32].

3.2.2. Support Vector Machine

Support Vector Machine (SVM) is one of the most widely applicable machine learning models developed by Vapnik [33]. The idea of the method is to transform the input space into a high-dimensional feature space using a nonlinear function, where two classes can be separated linearly. The goal of SVM is to find the hyperplane that maximizes the minimum distance between the hyperplane and the support vectors. Like LR and LDA, SVM was originally developed for two-class classification, then further extended to multiclass problems [34].

SVM is reported to perform well in cases where sample size is small or the number of features is more than the data points. It has good generalization in practice, and thus, a relatively low risk of overfitting. Despite its advantages, choosing the optimal kernel for SVM is a difficult task. SVM also does not directly provide probability estimates and is harder to be interpreted compared with decision-tree-based methods [35,36].

3.2.3. Decision Trees

Decision Trees (DT) are one of the most commonly used models in supervised classification and serve as the building blocks for several more-sophisticated ensemble models. DT constructs decision rules organized in a treelike structures to map input values to their target labels. In a tree structure, leaves represent labels and nonleaf nodes are features. Each branch represents a rule that leads to the final classification. The challenge lies in how to build the smallest decision trees: the best split should result in a classification with the lowest entropy or with the highest information gain. A realization of such a heuristic is C4.5 developed by Quinlan [37]. DT holds a lot of advantages, which explains its popularity in many applications, including the following: (1) easy to interpret and explain; (2) requiring relatively little effort from users for data preparation; (3) implicitly performing variable screening or feature selection. However, one key disadvantage of DT is that they tend to overfitting. Without proper pruning or limiting tree growth, they could become poor predictors [38].

3.3. Ensemble Models for Classification

To improve the performance of individual classifiers, ensemble models have been introduced. The idea of ensemble methods is to combine multiple weak learners to obtain a strong learner resulting in more accurate or robust predictions [39].

Ensemble models can be split into homogeneous and heterogeneous. Homogeneous ensemble models use only one type of classifier whereas heterogeneous ones combine different types of classifiers [40]. Two popular techniques in building homogeneous ensemble models are bagging and boosting. In bagging, k independent base classifiers are generated using bootstrapping; then, results are aggregated through majority voting. In boosting, base classifiers are built sequentially to improve the prediction of the previous outcomes [41].

In both cases, the base classifier can be any type of model, but decision tree methods are usually applied. Two such examples of bagging and boosting are Random Forest (RF) and Gradient-Boosted Decision Trees (GBDT).

3.3.1. Random Forest

Random Forest is one of the most popular bagging algorithms introduced by Breiman [42]. The algorithm starts with the generation of bootstrapped samples from the data; then, the collected decision trees are fitted to those samples. Predictions from all trees are aggregated in the inference to form the final decision via major voting in case of classification [43]. Benefiting from the randomization, RF helps in the reduction of variance, and is less likely to overfit compared with individual decision trees.

3.3.2. Extreme Gradient Boosting Trees

Extreme Gradient Boosting Trees (XGBoost) belongs to the family of gradient-boosted decision trees (GBDT). It was developed by Chen [44] and made multiple enhancements to improve the efficiency and scalability of the original GBDT methods in the implementation.

Plain gradient boosting trains each subsequent model using the residuals (the difference between the predicted and true values) or gradient, which is the reason why it is called "gradient boosting". By correcting the mistakes of the previous models, it gradually rectifies the results and improves the accuracy of predictions. XGBoost takes this one step further. It exploits the second-order derivative in the loss function formulation to accelerate the convergence of the model. XGBoost also introduces more regularization in the model formulation to control overfitting, which further improves its performance.

Built and developed for the sole purpose of model performance and computational speed, it quickly gained popularity and became the algorithm of choice for many winning solutions of machine learning competitions [45].

3.4. Experiment Setting and Parameter Tuning

To obtain steady and reliable results, we split the dataset into train, validation, and test subsets with a stratified random sampling method. This ensures that the distribution of different classes in the training and testing datasets are consistent. The testing set, consisting of 10% of the total data points in our case, is critical to evaluate the generalizability of the machine learning model resulted from training. The remaining data are further divided into train and validation subsets through cross-validation, which is also used to tune the hyperparameters for parameteric models. The higher the model complexity, the more hyperparameters there are to tune and the larger the feature space to search for the optimal hyperparameters. For hyperparameter tuning, Bayesian optimization is utilized to make

the exploration of large feature space more efficient [46]. Once the hyperparameters are determined, we train the model on the full training set and make inference on the testing set.

3.5. Model Evaluation

The performance of each model is evaluated using the following metrics: Accuracy, Recall, Precision, F1-score, and the area under the receiver operating characteristics curve (AUC).

According to the combination of actual data labels and predicted classes, the classification results can be divided into four cases: true positive (*TP*), false positive (*FP*), true negative (*TN*), and false negative (*FN*). The accuracy, defined by Equation (4), measures the percentage of correctly classified samples.

$$\operatorname{accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$
(4)

The recall is defined in Equation (5), indicating the percentage of real positive samples that are classified as positive.

$$\operatorname{recall} = \frac{TP}{TP + FN}$$
(5)

The precision is defined in Equation (6), measuring the proportion of actual positive samples within the samples that are predicted to be positive.

$$precision = \frac{TP}{TP + FP}$$
(6)

The F1 score is the harmonic mean of Recall and Precision and can be used to evaluate the model thoroughly. It is calculated as Equation (7):

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}$$
(7)

AUC is the area under the receiver operating characteristics (ROC) curve that represents the trade-off between Recall (*TPR*) and Specificity (*FPR*), given by Equations (8) and (9). As it is independent of a cutoff value, AUC is considered a better overall performance indicator than accuracy.

$$TPR = \frac{TP}{TP + FN} \tag{8}$$

$$FPR = \frac{FP}{TN + FP} \tag{9}$$

The AUC value ranges from 0.5 to 1, with 0.5 as the expected value of random prediction. A model with better overall performance has an AUC value close to 1.

4. Results and Discussion

This section comprehensively evaluates the performances of different machine learning models with increased complexity for lithology classification. For each machine learning model, we compare its prediction results based on the optimal hyperparameters tuned with respect to the metrics listed in Section 3.5. The results are presented in Sections 4.1 and 4.2. In Section 4.3, we further investigate how each lithological class can be distinguished by the best-performing machine learning methods among all models we put into the test and to what extent, and discuss the implication of how it can contribute to the identification of high-quality reservoirs in the study area. Lastly, in Section 4.4 we explore the feature importance and how the number of features affect the classification performances, studying the potential for designing more effective lithology classification systems in the future.

4.1. Hyperparameter Optimization

Table 2 presents the optimal hyperparameter settings tuned for different machine learning models in descending order of model complexity. The more complex the model, the more hyperparameters are required to tune for the model to avoid overfitting and achieve optimal performance. When the number of hyperparameters increase in the model, it becomes less feasible to use grid search to find the optimal hyperparameters as the searching space increases exponentially. Take the XGBoost model for example, consider five levels for each of the six parameters, the grid search needs to explore in total $5^6 = 15,625$ hyperparameter settings, whereas Bayesian optimization takes around 200 evaluations, equivalent to 1.3% of the workload of the grid search. A similar comparison between grid search and Bayesian optimization can be found in [18].

Table 2. Main hyperparameters tuned for parametric models.

Model	Hyperparameter	Symbol	Parameter Values
XGBoost	Boosting learning rate	learning_rate	0.08
	Subsample ratio of the training instances	subsample	0.7
	The maximum depth of a tree	max_depth	9
	The number of boosted trees	n_estimators	600
	L2 regularization term on weights	reg_alpha	0.1
	L1 regularization term on weights	reg_lambda	1.2
RF	The minimum number of samples required 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 trees in the forest	n_estimators	400
KNN	The number of neighbors to inspect	n_neighbor	3
SVM	Penalty parameter of the error term	С	1000
	Kernel coefficient for 'RBF'	gamma	0.1
DT	The minimum number of samples required at a leaf node	min_samples_leaf	1
	The minimum number of samples required to split an internal node	min_samples_split	2

4.2. Overall Performances

One of the main objectives of our study is to determine how different categories of machine learning methods perform in the task of lithology classification using well-logging data. Table 3 presents the overall performances of different machine learning models, with the better performer ranking higher in the list. It shows clearly that ensemble models perform best, followed by individual nonlinear models, while linear models rank last. The results live up to the common expectation that, in general, the classification performance improves with the increase in the model complexity.

Further, there are consistent performance results between training and testing for all different machine learning models, indicating that minimum overfitting exists in our trained models.

Values of AUC for ensemble models in both training and testing are well above 0.90, indicating very good discriminant and generalization abilities of the ensemble models. Within the ensemble models, XGBoost outperforms RF on all metrics. This is expected, as boosting methods are capable of reducing both bias and variance by increasing the expressive power of the base learner, while RF as bagging method is devised to reduce variance by subsampling the training data.

It is also noted that individual nonlinear models perform better than linear models. Linear models generally work better in situations where instances of different classes have clear boundaries and can be separated linearly, indicating that the lithology classes cannot be easily distinguished in the feature space formed by the well-logging data.

Madal	Training					Testing				
Niodel	Accuracy	AUC	Recall	Precision	F1	Accuracy	AUC	Recall	Precision	F1
XGBoost	0.852	0.920	0.735	0.847	0.844	0.882	0.947	0.769	0.880	0.876
RF	0.837	0.918	0.695	0.833	0.823	0.861	0.942	0.715	0.861	0.849
KNN	0.801	0.861	0.689	0.794	0.794	0.839	0.892	0.753	0.836	0.837
SVM	0.797	0.857	0.648	0.782	0.783	0.844	0.898	0.711	0.836	0.835
DT	0.766	0.762	0.659	0.764	0.765	0.781	0.760	0.656	0.776	0.779
LDA	0.708	0.741	0.423	0.637	0.636	$0.748 \\ 0.745$	0.793	0.451	0.692	0.685
LR	0.705	0.744	0.411	0.620	0.627		0.796	0.446	0.669	0.677

Table 3. Overall performance of different machine learning models in lithology classification, with best metrics achieved in each column are highlighted in bold.

4.3. Lithology Classification Evaluation

ROC curve and the confusion matrix are produced with optimized XGBoost and RF classifiers on test dataset, to inspect how well each lithology class can be distinguished from well-logging data in greater detail.

In Figure 2, ROC curves exhibit great generalization performances for both XGBoost and RF in classifying the three lithology classes. Among them, mudstone is the lithology class receiving the highest AUCs of 0.97 and 0.96 for XGBoost and RF, respectively. Meanwhile, sandstone gets the lowest AUCs for both classifiers, which are still above 0.92. The AUCs for sandy conglomerate are between those of mudstone and sandstone, with 0.95 and 0.94 for XGBoost and RF, respectively.



Figure 2. ROC Curve for XGBoost (**a**) and RF (**b**). M refers to mudstone, S refers to sandstone, and SC refers to sandy conglomerate.

Figure 3 highlights the confusion matrices of predictions generated by XGBoost and RF, from which we can interpret the actual classification accuracy of each lithology class. The normalized confusion matrices (Figure 3b,d) demonstrate that sandy conglomerate has the same highest classification accuracies for both XGBoost and RF at 97%. This could be attributed to the fact that sandy conglomerate accounts for 70% in the dataset and the evaluation metric for model optimization is classification accuracy.





XGBoost excels RF in the classification of mudstone and sandstone, with 5% and 11% higher classification accuracies for the two lithology classes, respectively. Classification performance of sandstone is the worst in both XGBoost and RF among the three lithology classes, which is coherent with the lowest AUCs observed in Figure 2. Mistakes are mainly concentrated in the misclassification of sandstone, as well as mudstone, into sandy conglomerate for both classifiers. This is likely due to the fact that sandy conglomerate has a mixed nature and contains samples with well logging signatures resembling that of mudstone and, especially, sandstone. It could result in overlaps between sandy conglomerate and the other two lithology classes in the feature space, thus leading to misclassification of sandstone and mudstone into sandy conglomerate.

Classification reports (Figure 4) are further derived from the confusion matrices. In our study, XGBoost model achieves both high precision of 0.888 and high recall of 0.969 for sandy conglomerate, which shows great potential to identify high-quality reservoirs in Mahu Sag [7]. In Mahu Sag, the sandy conglomerate accounts for over 90% of total oil-producing layers' thickness. It is also necessary to separate the lithofacies of tractive current sandy conglomerates (TCSC) and gravity flow sandy conglomerates (GFSC) from the lithology class of sandy conglomerate. Tractive current sandy conglomerates account for around 60% in thickness and about 90% in oil production of sandy conglomerate layers. To make the automated identification of TCSC possible, more rigorous lab work and interpretation are required to further label sandy conglomerate into sublithofacies of TCSC and GFSC, which is our next step.



Figure 4. Classification reports of each lithology class for XGBoost (a) and RF (b).

4.4. Feature Importance

Feature selection in machine learning models is a relevant consideration for many applications. Reducing irrelevant features can reduce model complexity and increase the generalization performance of the model. It also helps in designing more cost-efficient models by reducing the number of features in data collection.

In this study, we examine the model performances with reduced features using backward elimination based on importance measures. Figure 5 shows the feature importance extracted from XGBoost model of the nine well-logging variables. We then remove the feature with lowest importance one by one and retrain the model based on cross-validation. Figure 6 showcases the accuracy and AUC obtained by the XGBoost model with the decreased number of features. As can be seen from the figure, AUC decreases slowly at first but takes a sharp downturn when the feature number drop below 3. To keep the classification accuracy in testing above 0.9, we need to keep the top 6 features in the model.



Figure 5. Feature importance of XGBoost model for each input well-logging variable.



Figure 6. Accuracy and AUC plots of optimized XGBoost model with decreasing number of features.

5. Conclusions

The identification of lithology from well-logging data is an important task in reservoir characterization for petroleum exploration. Many different machine learning methods have been reported for this application. In this study, we comprehensively evaluated the performances of an array of supervised machine learning methods, from linear and nonlinear individual classifiers to ensemble methods, on lithology identification using well-logging data acquired from the sandy conglomerate reservoir of the Baikouquan formation in the Mahu Sag of the Junggar Basin, China. Cross-validation and Bayesian optimization were applied to optimize the hyperparameters of different models and their performances were evaluated on separate test dataset.

Results exhibit that ensemble methods (XGBoost and RF) perform best among the three categories of machine learning models, followed by nonlinear individual classifiers (kNN, SVM, and DT). Linear individual classifiers (LR and LDA) produce the least favorable results, indicating their disadvantages in solving the nonlinear lithology classification problem using well-logging data. Within the ensemble methods under testing, XGBoost has the best performance, with an accuracy of 0.882 and AUC of 0.947. It outperforms RF especially in the classification of sandstone with an increase in accuracy of 11%. Among the three lithology classes, sandy conglomerate—as found in the potential reservoirs in the study area—can be best distinguished with an accuracy of 97%, precision of 0.888, and recall of 0.969, suggesting the XGBoost model as a strong candidate machine learning model for more efficient and accurate lithology identification and reservoir quantification for geologists. Furthermore, we investigated the importance of well-logging variables and the impact of the number of well-logging variables as input on the classification performance. Experiments showed that at least the top three features are required for the XGBoost model to maintain comparable performance.

The study suggests ensemble methods as the more accurate and efficient machine learning models that can assist geologists in reservoir identification and lithology classification in general. The machine learning workflow established is transferable and can be applied in other geological environments. Future work will include further separation of subclasses of tractive current sandy conglomerates (TCSC) and gravity flow sandy conglomerates (GFSC) within sandy conglomerates from core images and more labeling on well-logging data. More machine learning methods within the category of boosting and beyond, such as neural networks, will be explored to find the best-performing model for distinguishing TCSC and GFSC, thus achieving better reservoir quality (e.g., permeability and porosity) assessment.

Author Contributions: Conceptualization, J.Z. (Junlong Zhang) and J.Z. (Junjie Zhang); methodology, J.Z. (Junlong Zhang) and J.Z. (Junjie Zhang); software, J.Z. (Junjie Zhang); validation, J.Z. (Junlong Zhang) and J.Z. (Junjie Zhang); formal analysis, J.Z. (Junlong Zhang) and J.Z. (Junjie Zhang); investigation, J.Z. (Junlong Zhang) and Y.Z.; project administration, W.L.; resources, W.L. and Y.H.; data curation, J.Z. (Junlong Zhang), Y.Z., and J.Z. (Junjie Zhang); writing—original draft preparation, J.Z. (Junlong Zhang) and J.Z. (Junjie Zhang); writing—review and editing, J.Z. (Junjie Zhang) and Y.H.; visualization, J.Z. (Junlong Zhang) and J.Z. (Junjie Zhang); supervision, J.Z. (Junjie Zhang) and Y.H.; funding acquisition, Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Grant NO. 41872118, Grant NO. 41472096 and Grant NO. 42002165).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing is not applicable.

Acknowledgments: We would like to thank the reviewers whose constructive comments improved the quality of this paper.

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

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area under the receiver operating characteristic curve
XGBoost	Extreme Gradient Boosting
RF	Random Forest
kNN	k-Nearest Neighborhood
SVM	Support Vector Machine
GBDT	Gradient-Boosted Decision Trees
DGF	Daniudi gas field
HGF	Hangjinqi gas field
GR	Gamma ray
SP	Self potential
CALI	Caliper log
RE	Self potential
RESS	Shallow reading resistivity measurement
RESM	Medium deep reading resistivity measurement
RESD	Deep reading resistivity measurement
PHIN	Neutron porosity log
RHOB	Bulk density log
М	Mudstone
S	Sandstone
SC	Sandy conglomerate
LR	Logistic Regression
LDA	Linear Discriminant Analysis
DT	Decision Trees
TP	True positive
FP	False positive
TN	True negative
FN	False negative

ROC	Receiver operating characteristics curve
TPR	True positive rate
FPR	False positive rate
TCSC	Tractive current sandy conglomerates
GFSC	Gravity flow sandy conglomerates

References

- 1. Buryakovsky, L.; Chilingar, G.V.; Rieke, H.H.; Shin, S. Fundamentals of the Petrophysics of Oil and Gas Reservoirs; Wiley: Hoboken, NJ, USA, 2012; p. 400.
- 2. Gu, Y.; Bao, Z.; Song, X.; Patil, S.; Ling, K. Complex lithology prediction using probabilistic neural network improved by continuous restricted Boltzmann machine and particle swarm optimization. *J. Pet. Sci. Eng.* **2019**, *179*, 966–978. [CrossRef]
- Liu, J.J.; Liu, J.C. An intelligent approach for reservoir quality evaluation in tight sandstone reservoir using gradient boosting decision tree algorithm—A case study of the Yanchang Formation, mid-eastern Ordos Basin, China. *Mar. Pet. Geol.* 2021, 126, 104939. [CrossRef]
- 4. Xie, Y.; Zhu, C.; Lu, Y.; Zhu, Z. Towards Optimization of Boosting Models for Formation Lithology Identification. *Math. Probl. Eng.* **2019**, 2019, 5309852. [CrossRef]
- Liu, H.; Wu, Y.; Cao, Y.; Lv, W.; Han, H.; Li, Z.; Chang, J. Well logging based lithology identification model establishment under data drift: A transfer learning method. *Sensors* 2020, 20, 3643. [CrossRef] [PubMed]
- 6. Zhao, Z.; He, Y.; Huang, X. Study on Fracture Characteristics and Controlling Factors of Tight Sandstone Reservoir: A Case Study on the Huagang Formation in the Xihu Depression, East China Sea Shelf Basin, China. *Lithosphere* **2021**, 2021, 1–15. [CrossRef]
- Lu, X.; Sun, D.; Xie, X.; Chen, X.; Zhang, S.; Zhang, S.; Sun, G.; Shi, J. Microfacies characteristics and reservoir potential of Triassic Baikouquan Formation, northern Mahu Sag, Junggar Basin, NW China. J. Nat. Gas Geosci. 2019, 4, 47–62. [CrossRef]
- 8. Li, W.; Hu, W.; Abubakar, A. Machine learning and data analytics for geoscience applications—Introduction. *Geophysics* 2020, *85*, WAI-WAII. [CrossRef]
- 9. Bergen, K.J.; Johnson, P.A.; De Hoop, M.V.; Beroza, G.C. Machine learning for data-driven discovery in solid Earth geoscience. *Science* 2019, 363, eaau0323. [CrossRef]
- Bhattacharya, S.; Carr, T.R.; Pal, M. Comparison of supervised and unsupervised approaches for mudstone lithofacies classification: Case studies from the Bakken and Mahantango-Marcellus Shale, USA. J. Nat. Gas Sci. Eng. 2016, 33, 1119–1133. [CrossRef]
- 11. Singh, H.; Seol, Y.; Myshakin, E.M. Automated Well-Log Processing and Lithology Classification by Identifying Optimal Features through Unsupervised and Supervised Machine-Learning Algorithms. *SPE J.* **2020**, *25*, 2778–2800. [CrossRef]
- 12. Rosid, M.S.; Haikel, S.; Haidar, M.W. Carbonate reservoir rock type classification using comparison of Naïve Bayes and Random Forest method in field "S" East Java. *AIP Conf. Proc.* **2019**, *2168*, 020019. [CrossRef]
- 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. Wang, X.; Yang, S.; Zhao, Y.; Wang, Y. Lithology identification using an optimized KNN clustering method based on entropyweighed cosine distance in Mesozoic strata of Gaoqing field, Jiyang depression. J. Pet. Sci. Eng. 2018, 166, 157–174. [CrossRef]
- 15. Al-Anazi, A.; Gates, I.D. A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs. *Eng. Geol.* **2010**, *114*, 267–277. [CrossRef]
- 16. J. Al-Mudhafar, W. Integrating kernel support vector machines for efficient rock facies classification in the main pay of Zubair formation in South Rumaila oil field, Iraq. *Model. Earth Syst. Environ.* **2017**, *3*, 12. [CrossRef]
- 17. Dev, V.A.; Eden, M.R. Formation lithology classification using scalable gradient boosted decision trees. *Comput. Chem. Eng.* **2019**, 128, 392–404. [CrossRef]
- 18. 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]
- 19. Bressan, T.S.; Kehl de Souza, M.; Girelli, T.J.; Junior, F.C. Evaluation of machine learning methods for lithology classification using geophysical data. *Comput. Geosci.* 2020, *139*, 104475. [CrossRef]
- 20. Ren, X.; Hou, J.; Song, S.; Liu, Y.; Chen, D.; Wang, X.; Dou, L. Lithology identification using well logs: A method by integrating artificial neural networks and sedimentary patterns. *J. Pet. Sci. Eng.* **2019**, *182*, 106336. [CrossRef]
- Liu, Y.; Huang, C.; Zhou, Y.; Lu, Y.; Ma, Q. The controlling factors of lacustrine shale lithofacies in the Upper Yangtze Platform (South China) using artificial neural networks. *Mar. Pet. Geol.* 2020, *118*, 104350. [CrossRef]
- Xie, Y.; Zhu, C.; Zhou, W.; Li, Z.; Liu, X.; Tu, M. Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. J. Pet. Sci. Eng. 2018, 160, 182–193. [CrossRef]
- Dev, V.A.; Eden, M.R. Gradient Boosted Decision Trees for Lithology Classification. Comput. Aided Chem. Eng. 2019, 47, 113–118. [CrossRef]
- 24. 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. [CrossRef]
- 25. Tao, J.; Zhang, C.; Qu, J.; Yu, S.; Zhu, R. A de-flat roundness method for particle shape quantitative characterization. *Arab. J. Geosci.* 2018, *15*, 414. [CrossRef]

- 26. Yu, S.; Ma, J. Deep Learning for Geophysics: Current and Future Trends. Rev. Geophys. 2021, 59, e2021RG000742. [CrossRef]
- 27. Banks, D.L.; Fienberg, S.E. Data Mining, Statistics. In *Encyclopedia of Physical Science and Technology*; Academic Press: Cambridge, MA, USA, 2003 ; pp. 247–261. [CrossRef]
- 28. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning; Springer: New York, NY, USA, 2009. doi: 10. 1007/978-0-387-84858-7. [CrossRef]
- 29. Ghojogh, B.; Ca, B.; Crowley, M.; Ca, M. Linear and quadratic discriminant analysis: Tutorial. arXiv 2019, arXiv:1906.02590.
- 30. Stanimirova, I.; Daszykowski, M.; Walczak, B. Robust Methods in Analysis of Multivariate Food Chemistry Data. *Data Handl. Sci. Technol.* **2013**, *28*, 315–340. [CrossRef]
- Mucherino, A.; Papajorgji, P.J.; Pardalos, P.M. Data Mining in Agriculture; Springer: New York, NY, USA, 2009; Volume 34. [CrossRef]
- 32. Cover, T.M.; Hart, P.E. Nearest Neighbor Pattern Classification. IEEE Trans. Inf. Theory 1967, 13, 21–27. [CrossRef]
- 33. Vapnik, V.N. The Nature of Statistical Learning Theory; Springer: New York, NY, USA, 1995. [CrossRef]
- Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) 2011, 2, 1–27. [CrossRef]
- Cristianini, N.; Shawe-Taylor, J. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods; Cambridge University Press: Cambridge, UK, 2000. [CrossRef]
- Auria, L.; Moro, R.A. Support Vector Machines (SVM) as a Technique for Solvency Analysis. SSRN Electron. J. 2008, 811, 1–16. [CrossRef]
- 37. Quinlan, J. C4.5: Programs for Machine Learning; Morgan Kaufmann Publishers: Burlington, MA, USA, 1993.
- Kotu, V.; Deshpande, B. Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner; Morgan Kaufmann Publishers: Burlington, MA, USA, 2014; pp. 1–425. [CrossRef]
- 39. Krogh, A.; Sollich, P. Statistical mechanics of ensemble learning. Phys. Rev. E 1997, 55, 811. [CrossRef]
- 40. Guidolin, M.; Pedio, M. Sharpening the Accuracy of Credit Scoring Models with Machine Learning Algorithms. In *Data Science for Economics and Finance*; Springer: New York, NY, USA, 2021; pp. 89–115. [CrossRef]
- 41. Wang, G.; Hao, J.; Ma, J.; Jiang, H. A comparative assessment of ensemble learning for credit scoring. *Expert Syst. Appl.* **2011**, 38, 223–230. [CrossRef]
- 42. Breiman, L. Random Forests. Mach. Learn. 2001, 45, 5–32. [CrossRef]
- 43. Schonlau, M.; Zou, R.Y. The random forest algorithm for statistical learning. Stata J. 2020, 20, 3–29. [CrossRef]
- 44. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
- XGBoost—Machine Learning Challenge Winning Solutions. Available online: https://github.com/dmlc/xgboost/blob/master/ demo/README.md#machine-learning-challenge-winning-solutions (accessed on 12 May 2022).
- Snoek, J.; Larochelle, H.; Adams, R.P. Practical Bayesian Optimization of Machine Learning Algorithms. *Adv. Neural Inf. Process.* Syst. 2012, 4, 2951–2959.