NAFTA-GAZ

Nafta-Gaz 2022, no. 3, pp. 165–175, DOI: 10.18668/NG.2022.03.01

Application of machine learning tools for seismic reservoir characterization study of porosity and saturation type

Zastosowanie metod uczenia maszynowego do charakterystyki porowatości i typu nasycenia przy użyciu atrybutów sejsmicznych

Tomasz Topór, Krzysztof Sowiżdżał

Oil and Gas Institute – National Research Institute

ABSTRACT: The application of machine learning (ML) tools and data-driven modeling became a standard approach for solving many problems in exploration geology and contributed to the discovery of new reservoirs. This study explores an application of machine learning ensemble methods – random forest (RF) and extreme gradient boosting (XGBoost) to derive porosity and saturation type (gas/water) in multi-horizon sandstone formations from Miocene deposits of the Carpathian Foredeep. The training of ML algorithms was divided into two stages. First, the RF algorithm was used to compute porosity based on seismic attributes and well location coordinates. The obtained results were used as an extra feature to saturation type modeling using the XGBoost algorithm. The XGBoost was run with and without well location coordinates to evaluate the influence of the spatial information for the modeling performance. The hyperparameters for each model were tuned using the Bayesian optimization algorithm. To check the training models' robustness, 10-fold cross-validation was performed. The results were evaluated using standard metrics, for regression and classification, on training and testing sets. The residual mean standard error (RMSE) for porosity prediction with RF for training and testing was close to 0.053, providing no evidence of overfitting. Feature importance analysis revealed that the most influential variables for porosity prediction were spatial coordinates and seismic attributes sweetness. The results of XGBoost modeling (variant 1) demonstrated that the algorithm could accurately predict saturation type despite the class imbalance issue. The sensitivity for XGBoost on training and testing data was high and equaled 0.862 and 0.920, respectively. The XGBoost model relied on computed porosity and spatial coordinates. The obtained sensitivity results for both training and testing sets dropped significantly by about 10% when well location coordinates were removed (variant 2). In this case, the three most influential features were computed porosity, seismic amplitude contrast, and iso-frequency component (15 Hz) attribute. The obtained results were imported to Petrel software to present the spatial distribution of porosity and saturation type. The latter parameter was given with probability distribution, which allows for identifying potential target zones enriched in gas.

Key words: machine learning, random forest, XGBoost, seismic attributes, reservoir properties prediction.

STRESZCZENIE: Metody uczenia maszynowego stanowią obecnie rutynowe narzędzie wykorzystywane przy rozwiązywaniu wielu problemów w geologii poszukiwawczej i przyczyniają się do odkrycia nowych złóż. Prezentowana praca pokazuje zastosowanie dwóch algorytmów uczenia maszynowego – lasów losowych (RF) i drzew wzmocnionych gradientowo (XGBoost) do wyznaczenia porowatości i typu nasycenia (gaz/woda) w formacjach piaskowców będących potencjalnymi horyzontami gazonośnymi w mioceńskich osadach zapadliska przedkarpackiego. Proces uczenia maszynowego został podzielony na dwa etapy. W pierwszym etapie użyto RF do obliczenia porowatości na podstawie danych pochodzących z atrybutów sejsmicznych oraz współrzędnych lokalizacji otworów. Uzyskane wyniki zostały wykorzystane jako dodatkowa cecha przy modelowaniu typu nasycenia z zastosowaniem algorytmu XGBoost. Modelowanie za pomocą XGBoost został przeprowadzone w dwóch wariantach – z wykorzystaniem lokalizacji otworów oraz bez nich w celu oceny wpływu informacji przestrzennych na wydajność modelowania. Proces strojenia hiperparametrów dla poszczególnych modeli został przeprowadzony z wykorzystaniem optymalizacji Bayesa. Wyniki procesu modelowania zostały ocenione na zbiorach treningowym i testowym przy użyciu standardowych metryk wykorzystywanych do rozwiązywania problemów regresyjnych i klasyfikacyjnych. Dodatkowo, aby wzmocnić wiarygodność modeli treningowych, przeprowadzona została 10-krotna kroswalidacja. Pierwiastek błędu średniokwadratowego (RMSE) dla wymodelowanej porowatości na zbiorach treningowym i testowym był bliski 0,053 co wskazuje na brak nadmiernego dopasowania modelu (ang. *overfitting*). Analiza istotności cech ujawniła, że zmienną najbardziej wpływającą na prognozowanie porowatości były współrzędne lokalizacji otworów oraz atrybut sejsmiczny sweetness. Wyniki modelowania XGBoost (wariant 1) wykazały, że algorytm jest w stanie dokładnie przewidywać typ nasycenia pomimo problemu z nierównowagą klas. Czułość

Corresponding author: T. Topór, e-mail: *tomasz.topor@inig.pl*

Article contributed to the Editor: 01.12.2021. Approved for publication: 22.02.2022

wykrywania potencjalnych stref gazowych w przypadku modelu XGBoost była wysoka zarówno dla zbioru treningowego, jak i testowego (0,862 i 0,920). W swoich predykcjach model opierał się głównie na wyliczonej porowatości oraz współrzędnych otworów. Czułość dla uzyskanych wyników na zbiorze treningowym i testowym spadła o około 10%, gdy usunięto współrzędne lokalizacji otworów (wariant 2 XGBoost). W tym przypadku trzema najważniejszymi cechami były obliczona porowatość oraz atrybut sejsmiczny *amplitude contrast* i atrybut *iso-frequency component* (15 Hz). Uzyskane wyniki zostały zaimportowane do programu Petrel, aby przedstawić przestrzenny rozkład porowatości i typu nasycenia. Ten ostatni parametr został przedstawiony wraz z rozkładem prawdopodobieństwa, co dało wgląd w strefy o najwyższym potencjale gazowym.

Słowa kluczowe: uczenie maszynowe, lasy losowe, drzewa wzmocnione gradientowo, atrybuty sejsmiczne, predykcja własności zbiornikowych.

Introduction

Porosity and saturation are essential parameters for characterizing reservoirs and developing oil and gas exploration strategies. Both parameters can be precisely determined from laboratory measurements or well log interpretations as continuous properties for a particular well. Although these data have a high accuracy (especially core measurements), they only identify isolated locations and leave out patches of missing data in-between wells. On the other hand, seismic attributes provide abundant lateral information at the reservoir scale but much lower resolution (Soleimani et al., 2020). The challenge for modern petroleum exploration geologists is to establish the relationship between these two datasets. However, these relationships are difficult to unveil due to the complex non-linear relationships between seismic data and reservoir parameters obtained from core measurements or well log interpretations (Feng, 2020; Liu et al., 2021).

The latest machine learning (ML) algorithms derive complex patterns that exist between variables based on implicit interactions and correlation. This unique learning capacity translates directly into the ability to make highly accurate predictions (Wood, 2020). Over the past two decades, ML algorithms and tools have made significant progress in solving the issues related to the study of reservoir characterization (Dramsch, 2020). The ML tools have been successfully applied in facies identifications and rock-typing (e.g., Lis-Śledziona, 2019; Topór, 2020; Liu et al., 2021) and in the prediction of key reservoir properties (e.g., Rafik and Kamel, 2017; Słota-Valim, 2018; Ahmadi and Chen, 2019; Ao et al., 2019; Erofeev et al., 2019; Naeini et al., 2019; Male and Duncan, 2020; Wood, 2020).

The area of seismic reservoir study mainly relies on deep learning methods. Bagheri et al. (2013) used support vector machines to analyze reservoir lithofacies of an Oil Field of Iran using 3D seismic data. The same tool was used by Na'imi et al. (2014) and Soleimani et al. (2020) to estimate porosity and water saturation. Feng (2020) used other deep learning methods and found that convolutional neural networks can achieve higher predictive accuracy in predicting reservoir porosity from seismic data than traditional artificial neural networks. Yasin et al. (2021) combined Gaussian simulation algorithms and post-stack seismic inversion using vector machine and particle swarm optimization to infer the spatial distribution of lithology and porosity from well logs and seismic data. Otchere et al. (2021) provided a comprehensive review of deep learning methods used in the petroleum industry.

Although deep learning methods prevail in reservoir characterization studies, other ML algorithms such as random forest (RF) and extreme gradient boosting (XGBoost) are also gaining attention in academia and industry. Zou et al. (2021) used the RF algorithm to predict porosity from multiple seismic attributes. Butorin (2020) used the same algorithm for probabilistic estimation of the distribution of an oil-saturated reservoir. Both ensemble methods are extremely popular machine learning algorithms for facies prediction (Bestagini et al., 2017; Saporetti et al., 2018; Kim et al., 2019). In the contest organized by The Society of Exploration Geophysicists in 2016, the XGBoost algorithm was leading in interpreting data from well-log analysis (Hall and Hall, 2017). One of the main benefits of using RF and XGBoost is their ability to handle data that are not structurally designed (James et al., 2013). Besides, they are computationally efficient, highly stable, and have superior accuracy, especially when it comes to classification problems (Chen and Guestrin, 2016; Hall and Hall, 2017).

This study used two ensemble methods – RF and XGBoost – to predict porosity and saturation type in sandstone formations from the Miocene deposits of the Carpathian Foredeep. The porosity prediction using RF was trained using seismic attributes and well location coordinates. The same variables and porosity prediction were used to classify the saturation type (water/gas) using the XGBoost algorithm. The XGBoost was also run in a second variant without well locations to evaluate the effect of spatial variables on prediction performance. The results were evaluated using training and testing sets and metrics suitable for regression (RMSE, R^2) and classification problems with class imbalance (sensitivity). In addition, the porosity computed with RF was compared with a deep learning approach from commercial software. The results were imported into Petrel software to represent the spatial distribution of porosity and saturation type.

Methods

Brief background of the RF and XGBoost algorithms

The RF and XGBoost algorithms utilize a machine learning approach called ensemble learning that combines multiple models and so-called weak learners in the prediction process. This aggregation method helps overcome the technical challenges of single predictive models, such as low accuracy, high variance, and feature noise and bias.

The RF algorithm is a modification of bagged trees with many deep, uncorrelated trees (models) that operate as an ensemble (Boehmke and Greenwell, 2020). Each model is trained parallel on a random subset of training samples and predictors (features). This feature helps to reduce variance, minimize overfitting, and improve prediction performance (James et al., 2013). It also distinguishes RF from bagging, where all predictors are used at each split. The RF has three hyperparameters that need to be set prior to the modeling process:

- mtry – the number of predictors to consider at each split;
- trees – the number of trees contained in the ensemble (forest);
- min_n – the minimum number of observations in a node for further splitting.

The author's previous paper provided detailed information about random forest hyperparameters (Topór, 2021 – based on Boehmke and Greenwell, 2020). A full description of the mathematical principles of RF algorithms is provided in Louppe (2014).

XGBoost is the most sophisticated ensemble tree method that adapts the idea of boosting weak learners using the gradient descent architecture (Boehmke and Greenwell, 2020). Gradient boosting machines build an ensemble of shallow trees in a sequence where each model learns from mistakes generated by the previous model (Yoav and Schapire, 1997). Gradient descent is an optimization algorithm that is used to update model parameters. The idea behind it is to minimize the loss function by measuring the local gradient of loss for a specified set of parameters. The parameters are sequentially adjusted in the direction of the descending gradient until they reach a minimum (zero gradients) (Boehmke and Greenwell, 2020). This algorithm can be performed on any loss function, thus handling regression and classification problems (Friedman, 2001). An essential parameter in the gradient descent architecture is a learning rate hyperparameter that controls the size of the steps. This parameter must be tuned to prevent the local minimum from being omitted or from never being reached if the learning rate is too high or too low, respectively.

XGBoost improves upon the base gradient boosting machines through system optimization and algorithmic enhancements, the most important of which appear to be additional regularization hyperparameters (Boehmke and Greenwell, 2020). Compared to RF, XGBoost has four extra hyperparameters that provide added protection against overfitting:

- tree_depth – the integer for the maximum depth of the tree;
- learn_rate – the number for the rate at which the boosting algorithm adapts from iteration-to-iteration;
- loss_reduction – the number for the reduction in the loss function required to split further;
- sample_size – the number for the proportion of data that is exposed to the fitting routine.

For detailed information about XGBoost hyperparameters, see Boehmke and Greenwell (2020 – chapter 12.5). The mathematical principles of the XGBoost algorithm can be found in (Chen and Guestrin, 2016).

ML workflow

The workflow applied in this study uses the *tidymodels* framework and the concept developed by R Core Team for modeling and machine learning (R Core Team, 2018; Kuhn and Silge, 2020). *Tidymodels* is a metapackage that uses tidyverse rules with a common philosophy of designing, grammar, and data structure.

Exploratory data analysis and data pre-processing

The analyzed dataset consists of seismic attributes calculated with a 3D seismic cube and with porosity and water saturation interpretations from well logs calibrated against core porosity measurements. The dataset refers to a field with multiple gas horizons located in the southern part of Carpathian Foredeep, filled with Middle Miocene shallow-marine sediments. In this area, the autochthonous Miocene complex is overthrusted by the Carpathian flysch nappes, which negatively impacts seismic data quality. The intervals of interest are gas-saturated sandstones reservoirs interbedded with mudstone-claystone formations.

The dataset consists of 28 variables and 787 177 observations. The outcome variables are porosity (phi) and saturation type (sat_type). Porosity was estimated based on standard methods of analyzing porosity crossplots on well-log profiles (neutron-acoustics, neutron-density). The saturation type comes from the estimated water saturation (Sw_initial) derived from the *Archie* or *Indonesia* formulas (Archie, 1942). The estimated water saturation was converted into a categorical variable of saturation type (water/gas) that represents horizons with potential high gas content (419 observations) and high water content (723 observations). The threshold used to partition Sw was set at 0.6 (Sw_initial > 0.6 indicates water horizons, Sw_initial < 0.6 shows gas horizons). This threshold still caused class imbalance issues that must be treated separately during data pre-processing.

The seismic attributes used in the modeling were generated using Petrel software (Schlumberger's trademark) and consisted of 23 parameters. The seismic attributes were derived from the acoustic impedance results of seismic inversion and attributes associated with the main features of the seismic signal, such as amplitude, frequency, phase, and polarity (Jędrzejowska-Tyczkowska, 2003). Additionally, three parameters for the spatial position of the samples (coordinates) within the grid were also used.

The completion rate for the outcome variables is extremely low (0.0015%), leaving a much smaller data set for modeling (1142 observations). A limited number of labels is common in seismic interpretation studies, where predictions are based on sparse information from well log interpretations. Since phi is of numerical type and sat_type of categorical type, the study evaluates regression and classification problems.

What is striking about this dataset is the very weak linear correlation between porosity and seismic attributes or spatial coordinates ($r < 0.3$) (Figure 1). The initial saturation used to derive the saturation type (water/gas) variable shows a link with seismic attenuation, z_coord, and a group of seismic frequency variables. There is also a negative trend between phi and Sw_initial, which is consistent with the *Archie* formula (Archie, 1950). Figure 1 also indicates a variables cross-correlation issue that needs to be solved in the data pre-processing stage.

The data was split into a training set and a test set using a ratio of 0.8. The 10-fold cross-validation technique was used on the training set to obtain ten resampling sets for analysis and assessment. This technique minimizes estimation or classification errors on unseen data. In the classification model, cross-validation was coupled with the stratification of the sat_type variable. This operation allows stratified sampling and creates more representative training and assessment sets.

Since both RF and XGBoost handle unstructured data, the data pre-processing stage was kept to a minimum. All numerical variables were normalized. A high correlation filter was applied to remove variables that had significant absolute correlations. The filter threshold was set at 0.9, and the Spearman method was used to account for non-linear relationships. In the classification models, up-sampling was applied to deal with class imbalance, as recommended by Kuhn and Silge (2020). The function replicates the rows of a dataset to equalize the occurrence of levels in a sat_type variable. All operations were done on the training set.

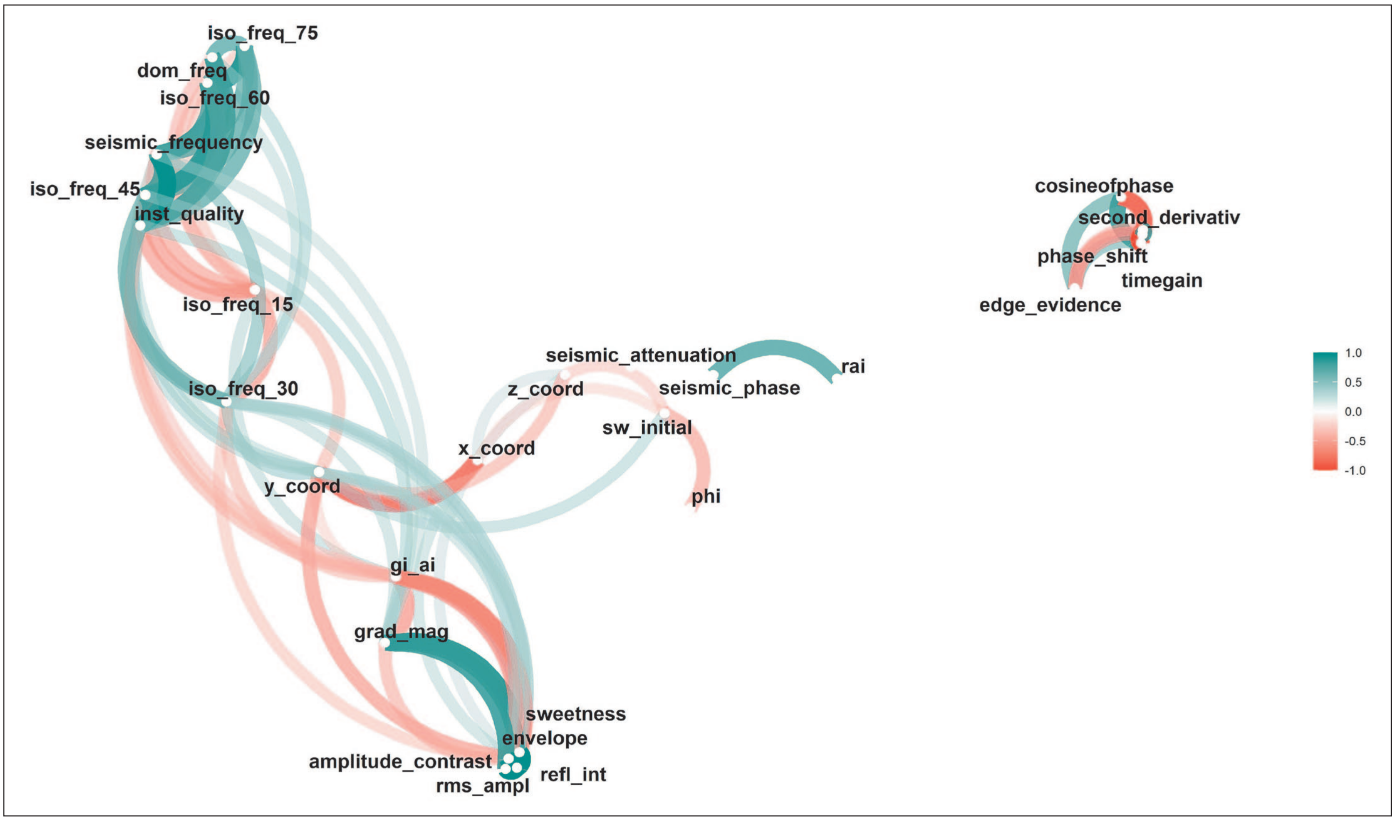

Figure 1. Correlation plot for numerical variables used in modeling. The more strongly correlated appear closer together and are connected by stronger paths. The proximity of the points is determined using multidimensional clustering. The minimum absolute value of correlations (using the Pearson method) is set at 0.3

Rysunek 1. Wykres korelacji dla zmiennych numerycznych stosowanych w modelowaniu. Zmienne, które są silniej skorelowane, pojawiają się bliżej siebie i są połączone silniejszymi ścieżkami. Bliskość punktów jest określana za pomocą grupowania wielowymiarowego. Minimalna wartość bezwzględna korelacji (metodą Pearsona) wynosi 0,3

Model specification and tuning strategy

The regression model was trained using the RF algorithm and the "randomForest" engine; the classification model using the XGBoost algorithm and the "exboost" engine. The models were trained under different data quality assumptions. The classification model used the porosity prediction from the regression modeling process as an additional variable. In addition, the classification model was run in two variants: with and without coordinates. This operation allows evaluating an input spatial variable (such as the location of wells with potentially high gas content) for performance prediction. The hyperparameters for each model were tuned using Bayesian optimization. The method uses an iterative training process to obtain combinations of tuning parameters based on previous results. The combination of hyperparameters is predicted using Gaussian process modeling and then scored based on the performance estimates of the models (Shahriari et al., 2016).

Results and Discussion

Results of hyperparameter tuning

The most efficient way to improve the prediction accuracy of machine learning models is to tune the hyperparameters of the models. Unlike RF, the predictive accuracy of XGBoost highly depends on the settings of the hyperparameters, so they require tuning before the learning process begins (Probst et al., 2019). Bayesian optimization used in this study is a promising tuning strategy method used in reservoir characterization. This method offers an improved and more accurate way of selecting a hyperparameter identification method compared to manual and grid search techniques (Otchere et al., 2021). The results of Bayesian optimization of model parameters for RF and XGBoost are presented in Table 1. The combinations of tuning parameters for the regression model were scored using the residual mean standard error (RMSE); whereas for classification models, the scoring was based on the area under the receiver operator curve (roc auc).

Performance of the models

The performance of the regression model was evaluated using RMSE and the coefficient of determination (R^2). The RMSE has the same unit as the original data, and the R^2 ranges from 0 to 1. Both metrics measure the accuracy of a model. An RMSE of ~0.053, as in the results obtained from training (0.052) and testing sets (0.053), reflects a standard error of prediction of ±5% porosity. Since the accuracy of the test set is similar to that of the training set, the RF model does not overfit.

We attempted to compare the RF modeling results with those obtained from artificial neural networks (ANN) available in Petrel software. The application of ANN proved to be ineffective, resulting in correlation coefficient values in the range of 0.2–0.3 (depending on the data pre-processing), which is distinctly lower than those produced by RF models. The low prediction accuracy of ANN may have been associated with a too high vertical resolution of the 3D grid, which was 4.5 m. Reducing the vertical resolution of the 3D grid and constructing the ANN estimation model of average porosity values for each reservoir horizon separately improved the R^2. This operation, however, changed the data quality assumption and did not allow for the comparison of the results. Moreover, it was seen as an oversimplification that neglected the variation of porosity in the vertical direction, which directly impacts the results of the modeled property. Therefore, the results obtained for RF seem promising, especially for solving estimation problems in high-resolution models or in low-thickness reservoirs.

While the results of the ANN algorithm could be used as a map of secondary average variables applied as a locally varying mean within stochastic or deterministic 3D modeling algorithms (Deutsch, 2002), the RF result can be considered as the final result of porosity prediction as well as a secondary variable for both locally varying means and co-kriging applications given the ability of the RF method to provide reasonable results within a high vertical resolution grid.

The performance of the classification models was assessed using classification accuracy and roc auc as standard metrics

Table 1. Results of Bayesian optimization of hyperparameters for RF and XGBoost

Tabela 1. Wyniki optymalizacji hiperparametrów metodą Bayesa dla modeli RF i XGBoost

Regression	Classification – variant 1	Classification – variant 2
random forest porosity (phi) prediction	XGBoost gas/water prediction (with coordinates)	XGBoost gas/water prediction (without coordinates)
mtry:23 trees: 1951 min_n: 2	mtry: 20 trees: 1001 min_n: 4 tree_depth: 12 learn_rate: 0.0431080161 loss_reduction: 0.0001291416 sample_size: 0.8987710451	mtry: 23 trees: 1558 min_n: 2 tree_depth: 15 learn_rate: 0.0035576682 loss_reduction: 0.4632878423 sample_size: 0.8631567641

Table 2. Main estimators for evaluating model performance

Tabela 2. Główne metryki dla wyników modelowania

Model		Metric	Estimate	
			training	testing
RF phi		RMSE R^2	0.052 0.402	0.053 0.391
XGBoost sat_type (gas/water)	variant 1	accuracy roc auc sensitivity	0.909 0.963 0.862	0.917 0.966 0.920
	variant 2	accuracy roc auc sensitivity	0.831 0.901 0.756	0.781 0.861 0.804

in this type of modeling (Table 2). While accuracy reflects the fraction of correctly classified observations, roc auc computes sensitivity and specificity across continuous classification thresholds (Kuhn and Silge, 2020). The higher the value of roc auc (which ranges from 0 to 1), the better the model discriminates between areas with high gas potential and high water saturation potential.

However, high accuracy may be misleading in this case since the modeling deals with class imbalance. It is much easier for the model to find the area with high water saturation that prevails in the studied dataset. Keeping in mind the goal of the study – predicting zones enriched with gas – the most reasonable parameter to evaluate the performance of classification models seems to be sensitivity (recall). This parameter reflects the fraction of positive correctly classified observations. A meaningful way to present the results of classification modeling and particularly sensitivity is through the confusion matrix (Figure 2A and 2B). In practice, the sensitivity means that on the training set, XGBoost (variant 1) was able to classify 286 as potential gas zones correctly and misclassified 46 observations by assigning them to water zones (Figure 2A). The sensitivity of the model is high and equals 0.862 (Table 2). On the testing set, these values are 80 and 7 (sensitivity equals 0.920; Figure 2B, Table 2).

Once the inputs from the spatial variables (x, y, and z coordinates) were removed, the prediction accuracy of XGBoost variant 2 dropped. The model correctly predicted 251 observations and misclassified 81 observations (Figure 3A). On the testing set, these values are 70 and 17, respectively (Figure 3B). The sensitivity for both the training and the testing set dropped significantly by about 10% compared to XGBoost variant 1 (Table 2).

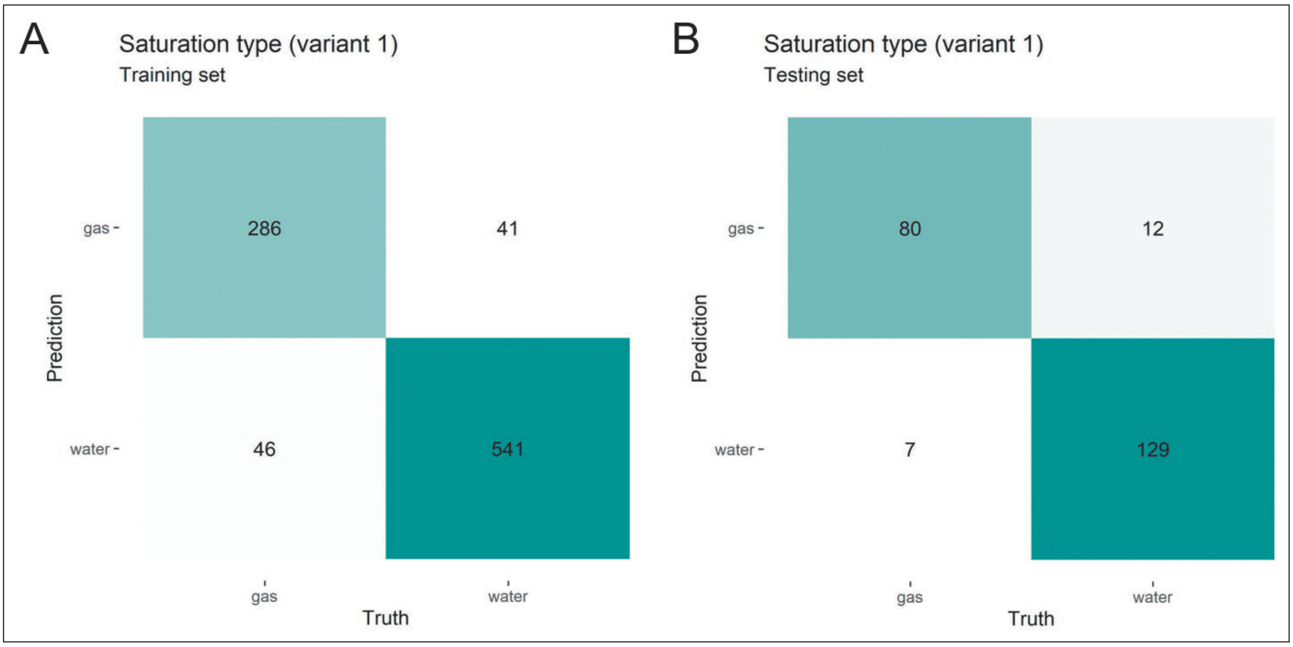

Figure 2. Confusion matrix for modeling results obtained by XGBoost variant 1 (with coordinates) for the training set (A) and the testing (B) set

Rysunek 2. Macierz błędów dla wyników modelowania uzyskanych dla wariantu 1 XGBoost (ze współrzędnymi) dla zbioru treningowego (A) i testowego (B)

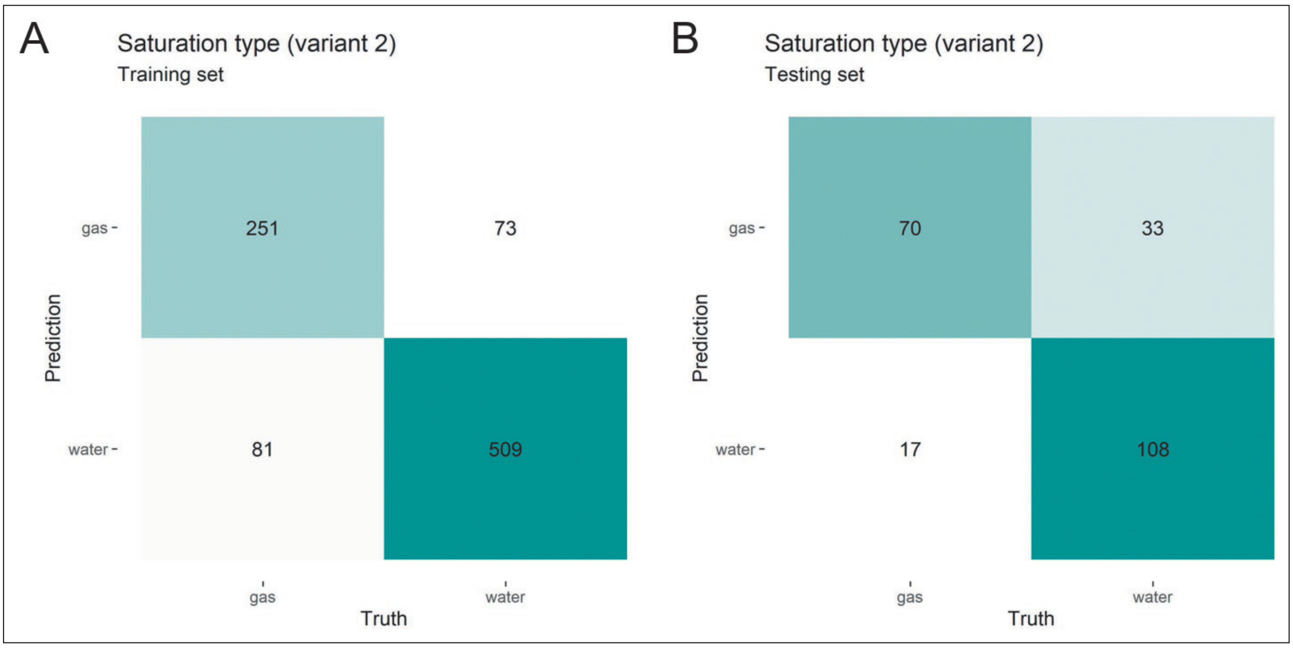

Figure 3. Confusion matrix for modeling results obtained by XGBoost variant 2 (without coordinates) for the training set (A) and the testing (B) set

Rysunek 3. Macierz błędów dla wyników modelowania uzyskanych dla wariantu 2 XGBoost (bez współrzędnych) dla zbioru treningowego (A) i testowego (B)

Interpretation of models

Evaluation of feature importance is a critical part of machine learning interpretation and explainability (Molnar, 2019). In ensemble models like RF and XGBoost, the importance of the features is calculated using impurity-based feature importance, also known as the permutation method (James et al., 2013). The idea behind this procedure is to randomly shuffle each model variable and calculate the change in model performance. The most significant features that influence model performance have higher predictive power (Boehmke and Greenwell, 2020).

According to Molnar (2019), a key step in interpreting feature importance is to apply a cross-correlation, as adding a correlated feature can decrease the importance of the associated feature. In this study, the cross-correlation issue was addressed in the pre-processing stage. Features that had absolute Spearman's rank correlation coefficient greater than 0.9 were eliminated. In the regression model, the number of features was limited to 19, while in the classification model, the number of features was reduced to 16 and 17 for variants 1 and 2, respectively.

The top 10 influential features for the regression and classification models are shown in Figure 4. For porosity prediction, spatial coordinates and sweetness are at the top (Figure 4A). The RF model strongly relies on coordinates in the prediction process and looks for a spatial connection of porosity between the studied wells.

As expected, the computed porosity plays an essential role in predicting the saturation type (gas/water) in both classification models (Figures 4B, 4C). The proximity of zones with potentially high gas/water content, as encoded in spatial coordinates, also helps the model enhance prediction performance. The role of spatial coordinates is unquestionable, as removing these variables decreases model sensitivity by ~10%. In XGBoost with no coordinates (variant 2), the model strongly relies on amplitude contrast. The model also accounts for the iso frequency component (15 Hz) and seismic attenuation attributes. The instantaneous frequency responds to wave propagation effects and depositional characteristics and can be used as a hydrocarbon or fracture zone indicator (low-frequency anomaly) and as a bed thickness indicator (Taner, 2001).

Spatial prediction of porosity and saturation type

The results of the modeling process were integrated with Petrel software to show the spatial distribution of obtained parameters: porosity and saturation type (gas/water) (Figures 5, 6A, 6B). High porosity is not always related to high gas content. This is common in Miocene reservoirs of the Carpathian Foredeep, where gas can accumulate in various lithotypes, including heterolites or even mudstones with much poorer reservoir properties (Leśniak et al., 2007; Sowiżdżał et al., 2020). On the other hand, this relationship is stronger in the gas accumulation region and disappears in the water zone, especially for variant 1, which utilizes well coordinates as a variable.

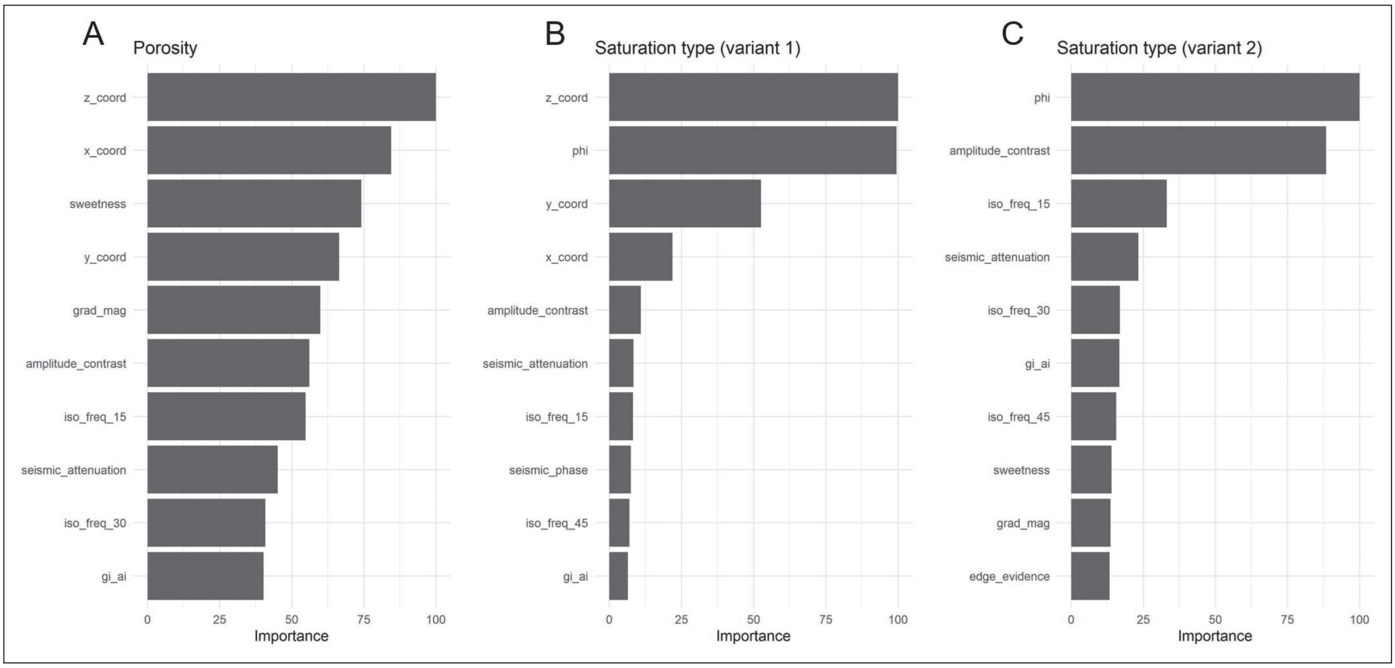

Figure 4. Variable importance for the top ten most influential features in the regression (A) and classification (B, C) models

Rysunek 4. Dziesięć najbardziej wpływowych cech dla modelu regresji (A) i modeli klasyfikacyjnych (B, C)

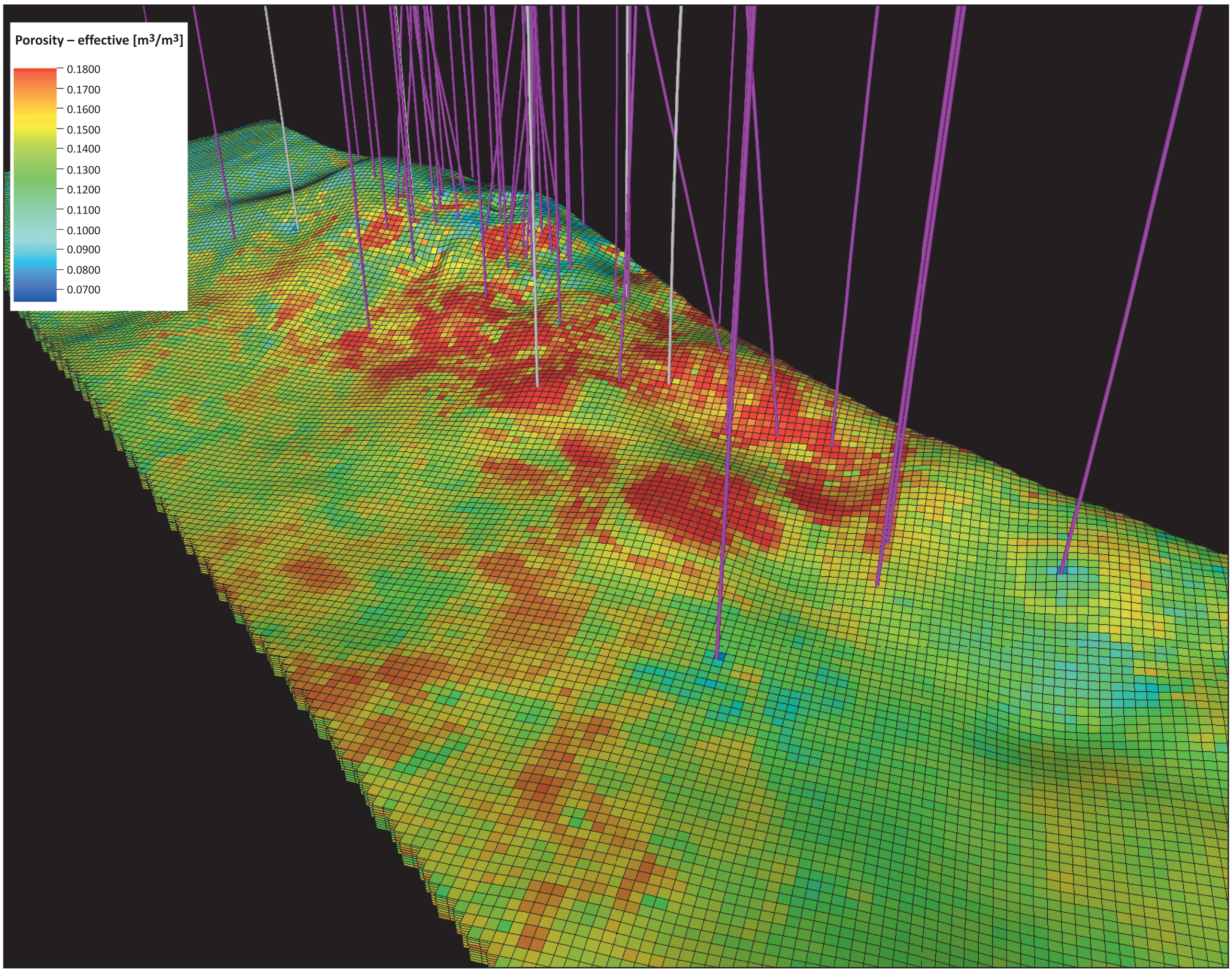

Figure 5. Distribution of porosity resulting from RF application (for selected horizon)

Rysunek 5. Rozkład porowatości na podstawie modelu RF dla wybranego horyzontu

The saturation type (gas/water) is presented in a probability distribution map in two variants for the results obtained from the XGBoost model with and without spatial coordinates (Figures 6A, 6B). This approach allows showing the potential zones enriched with gas (Sw < 60%) along with the probability of its occurrence as well as to evaluate the influence of already existing wells on the prediction of gas zones.

Despite the differences in the accuracy and sensitivity of the models across test sets, both XGBoost models indicate similar zones with potentially high gas content (Figures 6A, 6B). What is intriguing is the ability of XGBoost to reproduce the level of gas/water contact even though this variable (interpreted at well locations) was removed from the data set since it strongly suggests the extent of gas zones.

Conclusions

The study presents a data-driven approach for inferring porosity and saturation type from seismic attributes in multiple horizon gas field within sandstone formations of the Miocene strata of the Carpathian Foredeep. The modeling part of the study was performed using the tidymodels approach and state-of-the-art machine learning modeling. Porosity was inferred using the RF algorithm, and the saturation type was performed with XGBoost using two model variants: with and without well location coordinates.

The RF model had the standard prediction error of ~0.053 for both the training and testing sets with no evidence of overfitting.

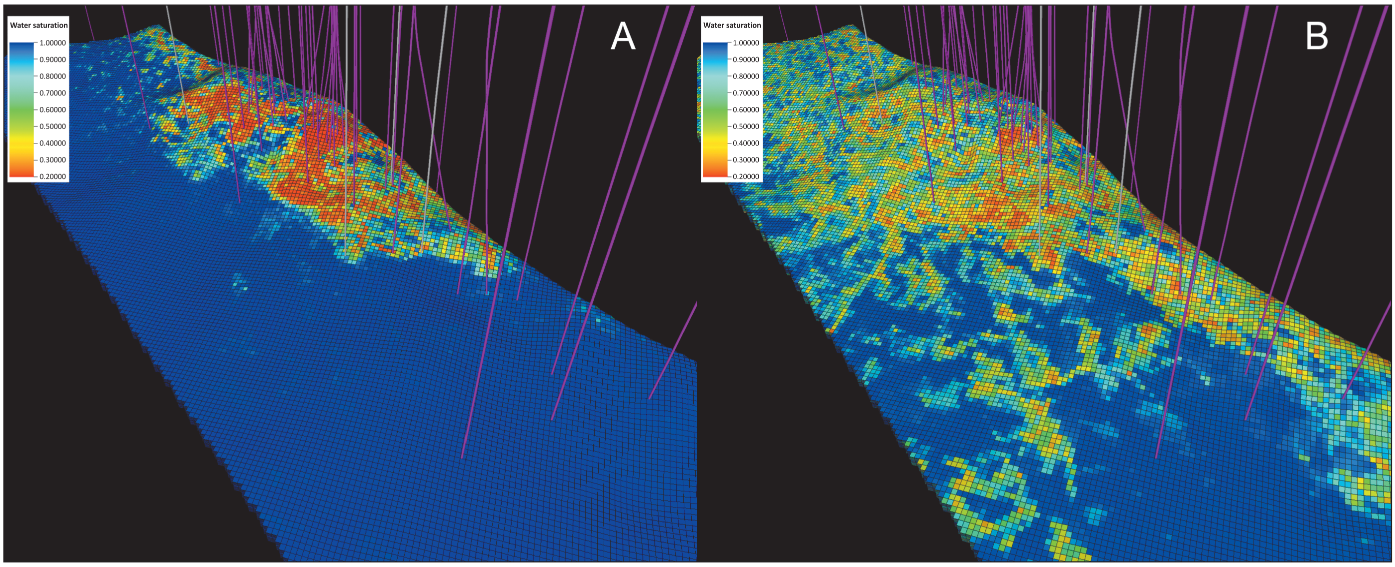

Figure 6. Distribution of gas occurrence probability for selected horizon estimated with XGBoost with coordinates (A) and without coordinates (B)

Rysunek 6. Rozkład prawdopodobieństwa wystąpienia gazu dla wybranego horyzontu oszacowany za pomocą nodelu XGBoost ze współrzędnymi (A) i bez współrzędnych (B)

The XGBoost with coordinates is more conservative in its predictions and clearly distinguishes the gas zones from the water zones with a sharp gas water contact boundary (Figure 6A). In contrast, the second variant of XGBoost indicates potential gas zones with a much wider range (albeit with lower probability) and with a more fuzzy boundary between gas and water zones (Figure 6B). Unfortunately, these predictions are not verifiable due to a lack of wells in the peripheral parts of the study area. Based on the sensitivity results from the testing set, XGBoost has the potential ability to find gas in 92% (variant 1) and 80% (variant 2) of indicated areas, depending on the data used in modeling.

Combining these results with probability distribution maps allows focusing on the most prospective locations with potential gas occurrence.

The model had much greater accuracy compared to the results obtained from ANN (Petrel software). Feature importance analysis revealed that the RF mainly relied on spatial coordinates and sweetness from seismic attributes.

Despite the class imbalance issue, the XGBoost with location coordinates demonstrated very good performance in distinguishing between potential gas and water zones. The sensitivity of the model reached 0.862 and 0.920 for the training and testing sets, respectively. The two most significant features of the model were porosity values and spatial coordinates. The capability of accurately predicting saturation type dropped by about 10% once the well location coordinates were removed from the modeling process. The second variant of the XGBoost relied on the values of porosity, amplitude contrast, and the iso-frequency component (15 Hz).

The obtained results were imported to Petrel software to show the spatial distribution of the parameters. The saturation type was presented as a probability distribution map, indicating potential target zones enriched in gas.

The presented workflow seems to be a promising tool for refining the reservoir modeling strategies, for identifying the best locations for infill drillings in appraised reservoirs or already exploited reservoirs, and finally, for decreasing uncertainty in decision-making processes.

Acknowledgments

We would like to thank Dr. Piotr Szulc for his valuable comments, which helped to improve the manuscript

This paper was written based on the statutory work entitled: *Zastosowanie technik uczenia maszynowego do prognozowania parametrów petrofizycznych skał oraz identyfikacji stref perspektywicznych* – the work of the Oil and Gas Institute – National Research Institute was commissioned by the Ministry of Education and Science; order number: 0077/SG/2021, archive number: DK-4100-0065/2021.

References

Ahmadi M.A., Chen Z., 2019. Comparison of machine learning methods for estimating permeability and porosity of oil reservoirs via petro-physical logs. *Petroleum*, 5: 271–284. DOI: 10.1016/j.petlm.2018.06.002.

Ao Y., Li H., Zhu L., Ali S., Yang Z., 2019. The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. *Journal of Petroleum Science and Engineering*, 174: 776–789. DOI: 10.1016/j.petrol.2018.11.067.

Archie G.E., 1942. Electrical Resistivity Log as an Aid in Determining Some Reservoir Characteristics. *Transactions of the AIME*, 146: 54–62. DOI: 10.2118/942054-G.

Archie G.E., 1950. Introduction to Petrophysics of Reservoir Rocks. *AAPG Bulletin*. 34: 943–961.

Bagheri M., Riahi M.A., Hashemi H., 2013. Reservoir lithofacies analysis using 3D seismic data in dissimilarity space. *Journal of Geophysics and Engineering*, 10: 1–9. DOI: 10.1088/1742-2132/10/3/035006.

Bestagini P., Lipari V., Tubaro S.T., 2017. A machine learning approach to facies classification using well logs, *in SEG Technical Program Expanded Abstracts, Society of Exploration Geophysicists*, 2137–2142. DOI: 10.1190/segam2017-17729805.1.

Boehmke B., Greenwell B., 2020. Hands-On Machine Learning with R. <https://bradleyboehmke.github.io/HOML/> (accessed: November 2021).

Butorin A.V., 2020. Interpretation of seismic inversion results using the "Random Forest". *Conference Proceedings, Data Science in Oil & Gas, Online, European Association of Geoscientists & Engineers*, 1–7. DOI: 10.3997/2214-4609.202054018.

Chen T., Guestrin C., 2016. XGBoost: A Scalable Tree Boosting System. *22nd SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA*.

Deutsch C.V., 2002. Geostatistical reservoir modeling. *Oxford University Press, New York*.

Dramsch J.S., 2020. 70 Years of Machine Learning in Geoscience in Review. *Advances in Geophysics*, 61: 1–55. DOI: 10.1016/bs.agph.2020.08.002.

Erofeev A., Orlov D., Ryzhov A., Koroteev D., 2019. Prediction of porosity and permeability alteration based on machine learning algorithms. *Transport in Porous Media*, 128: 677–700. DOI: 10.1007/s11242-019-01265-3.

Feng R., 2020. Estimation of reservoir porosity based on seismic inversion results using deep learning methods. *Journal of Natural Gas Science and Engineering*, 77: 103270. DOI: 10.1016/j.jngse.2020.103270.

Friedman J.H., 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29: 1189–1232.

Hall M., Hall B., 2017. Distributed collaborative prediction: Results of the machine learning contest. *Leading Edge*, 6: 267–269. DOI: 10.1190/tle36030267.1.

James G., Witten D., Hastie T., Tibshirani R., 2013. An Introduction to Statistical Learning with Applications in R. *Springer Texts in Statistic*. DOI: 10.1007/978-1-4614-7138-7.

Jędrzejowska-Tyczkowska H., 2003. Sejsmicznie konsystentne estymatory złoża weglowodorów. *Prace Instytutu Górnictwa Naftowego i Gazownictwa*, 123: 1–139.

Kim Y., Hardisty R., Torres E., Marfurt K.J., 2019. Seismic facies classification using random forest algorithm. *SEG International Exposition and Annual Meeting*, 2161–2165. DOI: 10.1190/segam2018-2998553.1.

Kuhn M., Silge J., 2020. Tidy Modeling with R. <https://www.tmwr.org/> (accessed: November 2021).

Leśniak G., Such P., Dziadzio P., 2007. Reservoir Properties of Miocene Sandstones in Rzeszow Area (Carpathian Foredeep, Poland). *Thrust Belts and Foreland Basins*: 397–412, DOI: 10.1007/978-3-540-69426-7_21.

Lis-Śledziona A., 2019. Petrophysical rock typing and permeability prediction in tight sandstone reservoir. *Acta Geophysica*, 67: 1895–1911. DOI: 10.1007/s11600-019-00348-5.

Liu X., Ge Q., Chen X., Li J., Chen Y., 2021. Extreme learning machine for multivariate reservoir characterization. *Journal of Petroleum Science and Engineering*, 205: 108869. DOI: 10.1016/j.petrol.2021.108869.

Louppe G., 2014. Understanding Random Forests: From Theory to Practice. *University of Liège*. <http://arxiv.org/abs/1407.7502%0A> (accessed: November 2021).

Male F., Duncan I.J., 2020. Lessons for machine learning from the analysis of porosity-permeability transforms for carbonate reservoirs. *Journal of Petroleum Science and Engineering*, 187: 106825. DOI:10.1016/j.petrol.2019.106825.

Molnar C., 2019. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. <https://christophm.github.io/interpretable-ml-book> (accessed: November 2021).

Na'imi S.R., Shadizadeh S.R., Riahi M.A., Mirzakhanian M., 2014. Estimation of reservoir porosity and water saturation based on seismic attributes using support vector regression approach. *Journal of Applied Geophysics*, 107: 93–101. DOI: 10.1016/j.jappgeo.2014.05.011.

Naeini E.Z., Green S., Rauch-Davies M., 2019. An integrated deep learning solution for petrophysics, pore pressure and geomechanics property prediction. *SPE/AAPG/SEG Unconventional Resources Technology Conference*: 1–18. DOI: 10.15530/urtec-2019-111.

Otchere D.A., Arbi Ganat T.O., Gholami R., Ridha S., 2021. Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models. *Journal of Petroleum Science and Engineering*, 200: 108182. DOI: 10.1016/j.petrol.2020.108182.

Probst P., Boulesteix A.L., Bischl B., 2019. Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, 20: 1–32.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna*. <https://www.r-project.org> (accessed: November 2021).

Rafik B., Kamel B., 2017. Prediction of permeability and porosity from well log data using the nonparametric regression with multivariate analysis and neural network, Hassi R'Mel Field, Algeria. *Egyptian Journal of Petroleum*, 26: 763–778. DOI: 10.1016/j.ejpe.2016.10.013.

Saporetti C.M., da Fonseca L.G., Pereira E., de Oliveira L.C., 2018. Machine learning approaches for petrographic classifcation of carbonate-siliciclastic rocks using well logs and textural information. *Journal of Applied Geophysics*, 155: 217–225. DOI: 10.1016/j.jappgeo.2018.06.012.

Shahriari B., Swersky K., Wang, Z., Adams R.P., De Freitas N., 2016. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104: 148–175. DOI: 10.1109/JPROC.2015.2494218.

Słota-Valim M., 2018. Określanie mechanicznych właściwości skał na podstawie właściwości fizycznych przy użyciu sztucznych sieci neuronowych. *Nafta-Gaz*, 74(5): 343–355. DOI: 10.18668/NG.2018.05.01.

Soleimani F., Hosseini E., Hajivand F., 2020. Estimation of reservoir porosity using analysis of seismic attributes in an Iranian oil field. *Journal of Petroleum Exploration and Production Technology*, 10: 1289–1316. DOI: 10.1007/s13202-020-00833-4.

Sowiżdżał K., Słoczyński T., Sowiżdżał A., Papiernik B., Machowski G., 2020. Miocene Biogas generation system in the Carpathian Foredeep (SE Poland): A basin modeling study to assess the potential of unconventional mudstone reservoirs. *Energies*, 13(7): 1838. DOI: 10.3390/en13071838.

Taner M.T., 2001. Seismic Attributes. *The Canadian Society of Exploration Geophysicists's*: 49–56.

Topór T., 2020. An integrated workflow for MICP-based rock typing: A case study of a tight-gas sandstone reservoir in the Baltic Basin (Poland). *Nafta-Gaz*, 76(4): 219–229. DOI: 10.18668/NG.2020.04.01.

Topór T., 2021. Application of machine learning algorithms to predict permeability in tight sandstone formations. *Nafta-Gaz*, 77(5): 3–12. DOI: 10.18668/NG.2021.05.01.

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: 106587. DOI: 10.1016/j.petrol.2019.106587.

Yasin Q., Sohail G.M., Khalid P., Baklouti S., Du Q., 2021. Application of machine learning tool to predict the porosity of clastic depositional system, Indus Basin, Pakistan. *Journal of Petroleum Science and Engineering*, 197: 107975. DOI: 10.1016/j.petrol.2020.107975.

Yoav F., Schapire R.E., 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55: 119–139. DOI: 10.1006/jcss.1997.1504.

Zou C., Zhao L., Xu M., Chen Y., Geng J., 2021. Porosity Prediction With Uncertainty Quantification From Multiple Seismic Attributes Using Random Forest. *Journal of Geophysical Research*. Solid Earth, 126. DOI: 10.1029/2021JB021826.

Tomasz TOPÓR, Ph.D.
Assistant professor at the Department of Geology and Geochemistry
Oil and Gas Institute – National Research Institute
Lubicz 25 A
31-503 Krakow
E-mail: *tomasz.topor@inig.pl*

Krzysztof SOWIŻDŻAŁ, Ph.D. Eng.
Deputy Director for Prospecting of Hydrocarbons Deposits
Oil and Gas Institute – National Research Institute
Lubicz 25 A
31-503 Krakow
E-mail: *krzysztof.sowizdzal@inig.pl*