NAFTA-GAZ

Nafta-Gaz 2021, no. 5, pp. 283–292, DOI: 10.18668/NG.2021.05.01

Application of machine learning algorithms to predict permeability in tight sandstone formations

Zastosowanie metod uczenia maszynowego do przewidywania przepuszczalności w formacjach zwięzłych piaskowców typu *tight gas*

Tomasz Topór

Oil and Gas Institute – National Research Institute

Abstract: The application of machine learning algorithms in petroleum geology has opened a new chapter in oil and gas exploration. Machine learning algorithms have been successfully used to predict crucial petrophysical properties when characterizing reservoirs. This study utilizes the concept of machine learning to predict permeability under confining stress conditions for samples from tight sandstone formations. The models were constructed using two machine learning algorithms of varying complexity (multiple linear regression [MLR] and random forests [RF]) and trained on a dataset that combined basic well information, basic petrophysical data, and rock type from a visual inspection of the core material. The RF algorithm underwent feature engineering to increase the number of predictors in the models. In order to check the training models' robustness, 10-fold cross-validation was performed. The MLR and RF applications demonstrated that both algorithms can accurately predict permeability under constant confining pressure (R^2 0.800 vs. 0.834). The RF accuracy was about 3% better than that of the MLR and about 6% better than the linear reference regression (LR) that utilized only porosity. Porosity was the most influential feature of the models' performance. In the case of RF, the depth was also significant in the permeability predictions, which could be evidence of hidden interactions between the variables of porosity and depth. The local interpretation revealed the common features among outliers. Both the training and testing sets had moderate-low porosity (3–10%) and a lack of fractures. In the test set, calcite or quartz cementation also led to poor permeability predictions. The workflow that utilizes the *tidymodels* concept will be further applied in more complex examples to predict spatial petrophysical features from seismic attributes using various machine learning algorithms.

Key words: machine learning, random forest, permeability prediction.

STRESZCZENIE: Zastosowanie algorytmów uczenia maszynowego w geologii naftowej otworzyło nowy rozdział w poszukiwaniu złóż ropy i gazu. Algorytmy uczenia maszynowego zostały z powodzeniem wykorzystane do przewidywania kluczowych właściwości petrofizycznych charakteryzujących złoże. W pracy zastosowano metody uczenia maszynowego do przewidywania przepuszczalności w warunkach ustalonego ciśnienia złożowego dla formacji zwięzłych piaskowców typu *tight gas*. Modele zostały skonstruowane przy użyciu algorytmów o różnym stopniu komplikacji (wielowymiarowa regresja liniowa – MLR i lasy losowe – RF), a następnie poddano je procesowi uczenia na danych zawierających podstawowe informacje o otworze, podstawowe parametry petrofizyczne oraz typ skał pochodzący z makroskopowego i mikroskopowego opisu próbek rdzeni. Typ skał został rozkodowany i poddany procesowi inżynierii cech, aby wydobyć dodatkowe zmienne do modelu. Proces uczenia na zbiorze treningowym został przeprowadzony z wykorzystaniem 10-krotnej kroswalidacji. Uzyskane wyniki pokazują, że oba algorytmy mogą przewidywać przepuszczalność z dużą dokładnością (R^2 = 0,800 dla MLR vs R^2 = 0,834 dla RF). Dokładność modelu RF jest około 3% lepsza niż MLR i około 6% lepsza w porównaniu do modelu referencyjnego (model regresji liniowej z jedną zmienną – porowatością). W przypadku obu modeli porowatość była najistotniejszym parametrem przy przewidywaniu przepuszczalności. Dodatkowo w modelu wykorzystującym lasy losowe istotną cechą okazała się głębokość próbki, co może świadczyć o dodatkowych interakcjach pomiędzy zmiennymi. Cechą wspólną próbek w zbiorze treningowym i testowym, dla których modele zadziałały ze słabą skutecznością, były porowatość od 3% do 10% i brak spękań. Dodatkowo w zbiorze testowym niska dokładność przewidywań przepuszczalności była związana z obecnością cementacji kalcytem i kwarcem. Workflow wykorzystujący stan wiedzy dotyczącej modelowania, którego trzon stanowi pakiet *tidymodels*, będzie dalej stosowany do prognozowania przestrzennych właściwości petrofizycznych na podstawie atrybutów sejsmicznych.

Słowa kluczowe: uczenie maszynowe, lasy losowe, predykcja przepuszczalności.

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

Article contributed to the Editor: 11.01.2021. Approved for publication: 22.04.2021

Introduction

Permeability is one of the most difficult petrophysical properties to determine and predict. It is also a key element in the characterization of both conventional and unconventional tight reservoirs. Sandstone gas reservoirs are usually defined as tight when their permeability is below 0.1 mD. However, for many tight sandstone formations, the average permeability is often lower than 0.01 mD (Ma et al., 2015). Usually, permeability is measured directly on core samples as a part of routine core analysis (RCA), taking into account its stress-dependency (McPhee et al., 2015; Such et al., 2015). Many attempts have been made to establish a relationship between porosity (or pore structure attributes) and permeability for different sedimentary basins, including tight sand reservoirs (e.g., Pape et al., 1999; Comisky et al., 2007; Such et al., 2007). However, due to the complexity of permeability's function, this parameter cannot be fully explained using superficial relationships.

The recent application of machine learning algorithms in petroleum geology demonstrated a new approach to solving various issues with characterizing reservoirs (Caté et al., 2017; Karpatne et al., 2019). Machine learning algorithms have been successfully used to predict reservoir porosity and permeability (Rafik and Kamel, 2017; Wu et al., 2018; Ahmadi and Chen, 2019; Erofeev et al., 2019; Male and Duncan, 2020), water saturation (Ao et al., 2019; Wood, 2020), capillary pressure (Jamshidian et al., 2018), pore pressure, geomechanical properties (Naeini et al., 2019), and mineral compositions (Rubo et al., 2019). Besides the regression problems, machine learning has been used for facies and fracture classification (Bhattacharya and Mishra, 2018). Unsupervised learning is commonly applied in rock typing (Ma et al., 2015; Meshalkin et al., 2018; Lis-Śledziona, 2019; Topór, 2020).

Machine learning utilizes many algorithms of varying degrees of complexity. The most commonly used are linear regression, regularized regression, k-nearest neighbors, decision trees, random forests, gradient boosting, and neural networks (Wendt et al., 1986; Baziar et al., 2018; Ao et al., 2019; Boehmke and Greenwel, 2020). Most of these algorithms can be used for problems of both regression and classification. They also perform differently – more sophisticated nonlinear algorithms are always better than simple linear regression. However, the improvement of accuracy comes at the expense of interpretability, which is important for model implementation. Recent advances in machine learning provide approaches to interpreting so-called "black box" models using the global and local explanation strategies (Molnar, 2019; Boehmke and Greenwel, 2020). Global interpretability helps researchers understand the model's predictions by looking at the importance of its features and how influential they are on the model's predictions and performance. As opposed to this holistic view, a local interpretation focuses on a particular observation (or a group of observations) and features that influence this observations' model prediction. Both approaches are an important part of interpretable machine learning (Molnar, 2019).

Among the machine learning methods mentioned above, the random forests (RF) algorithm is rapidly gaining attention among geoscience researchers and the petroleum community (Bhattacharya and Mishra, 2018; Brantson et al., 2018; Ao et al., 2019; Aulia et al., 2019; Bhattacharya et al., 2019; Rubo et al., 2019; Attanasi et al., 2020). This is mostly due to its very good out-of-the-box performance and ability to handle data (both nominal and continuous) that are not structurally designed (James et al., 2013). The RF algorithm is a modification of bagged trees with many de-correlated trees (Boehmke and Greenwel, 2020). The decision tree's growing process is performed using the randomization of predictors at each tree split (James et al., 2013; Boehmke and Greenwel, 2020). This operation reduced variance and improved prediction performance (James et al., 2013). It also distinguishes RF from bagging, where all predictors are used at each split. The subset of variables for each split (m_{try}) is one of the RF hyperparameters that can be tuned to improve the model's performance. Detailed information about RF algorithms with a particular emphasis on their mathematical principles can be found in Louppe (2014).

In this study, multiple linear regression (MLR) and random forests (RF) were trained to predict permeability under confining stress conditions for samples from tight sandstone formations, using basic well information (depth, basin, and formation names), petrophysical data (porosity and grain density), and rock type from a visual inspection of the core material. All data comes from US Department of Energy data in the public domain (Byrnes et al., 2009). The dataset provides a perfect opportunity to test the performance of various machine learning models on petrophysical data. It is also a valuable dataset with which to engineer features and interpret models of a global and local scale. Additionally, the paper presents a *tidymodel* workflow for modeling and machine learning using the programming language R.

This study is the second in a series about applying machine learning in the evaluation of unconventional tight reservoirs. The results from the first study with elements of unsupervised learning can be found in Topór (2020).

Methods

The workflow applied in this study utilizes the *tidymodels* packages and the latest state-of-the-art for machine learning modeling with R (R Core Team, 2018; Boehmke and Greenwel,

2020; Kuhn and Silge, 2020). The workflow consists of several steps that are accomplished by exploratory data analysis. It starts with data sampling and cross-validation (*rsample* packages), data preprocessing (*recipes*), model setup (*parsnip*), and model tuning (*tune*), and ends with model assessment (*yardstick*). All steps are joined together with the *workflows* package.

Exploratory data analysis and data preprocessing

The modeling process was performed on US Department of Energy data in the public domain (Byrnes et al., 2009). The data contain tabulated results from the petrophysical analysis performed on archive siliciclastic core material from five Rocky Mountain basins. The dataset provides a perfect opportunity to combine basic well information with petrophysical data and the petrographic description in order to predict permeability under confining stress conditions using different machine learning algorithms. Helium porosity at ambient conditions (porosity) and gas permeability collected at 4000 psi (k_conf) were measured on the same sample. In addition to the porosity and permeability, the analyzed dataset includes information about several predictive features, such as grain density (gd), basin and formation names (basin, formation), depth, and an encoded rock type (rt). The last one comes from a visual rock inspection from microscopic and lens observations on core material and plugs. Each depth interval was sampled from one to three times, and collected plugs were marked with the letters A, B, and C. In this study, only A plugs were used. To add a degree of difficulty, the routine permeability was also removed from the dataset. Detailed information about the dataset and the results of the project can be found in Byrnes et al. (2009).

The dataset consists of 1,096 observations and seven variables, four of which are of numeric type (k_conf, porosity, gd, and depth); the other three are categorical (basin, formation, and rt). Missing data comprise about 4.5% of the original dataset (Fig. 1). Most of the missing information (3.01%) comes from the outcome variable (k_conf); the others are porosity (0.91%) and grain density (0.64%). Data imputation was skipped from the modeling process (only 17 observations could be imputed for grain density and porosity), and the missing data were removed.

Increasing the number of variables can potentially improve modeling performance (especially in RF). Keeping that in mind, the rock type variable (rt) was decrypted to create another four variables. The variable was encoded so that:

- the first digit refers to basic lithology;
- the second digit represents grain size, sorting, and texture;
- the third digit represents the degree of consolidation, cementation, and the occurrence of fractures;
- the fourth digit refers to primary sedimentary structures; and
- the fifth digit represents dominant cementation or pore-filling material.

The resulting variables are lithology, fractures, sedimentary structures, and the pore-filling/cementation. Each of these new variables has several categories. For those variables that have less numerous categories, the level of the category was reduced (lumped) to not exceed five (e.g., pore-filling [sulfide pore filling, siderite, phosphate, anhydrite/gypsum, dolomite, calcite, quartz, authigenic clay, carbonaceous debris, non-pore-filling/detrital clay] was reduced to the four levels with the most observations and the rest went to the "other" group). Observations of the fractures variable composed of the categories "fractured" and "unfractured" were removed from the modeling process due to the limited number of observations in the first category (only 18). Also, 21 observations were removed from the lithology variable (one belonging to "shale" and 20 to "NA"). The basin and formation names were kept from the original database. The only modification was removing the "Sand Wash" Basin observations due to the limited number of observations. After the target and feature engineering, the

Fig. 1. Distribution of missing data in the analyzed dataset, using the visdat package
Rys. 1. Rozkład brakujących danych (pakiet visdat)

final dataset consisted of 1,002 observations and ten variables, of which half were of the nominal type.

The goal of the model was to predict permeability under constant confining pressure (k_conf) for different tight sandstone formations using basic well information, petrophysical data, and a simple rock description. Permeability is one of the essential rock features that determine reservoir quality (Ma, 2015). It is also information that cannot be directly derived from the wireline logs. The permeability ranged from 0.000001 mD to 173 mD (median: 0.00145 mD), and its distribution was highly right (or positively) skewed. Before the modeling process, the outcome variable had to be log-transformed (as presented in Fig. 2).

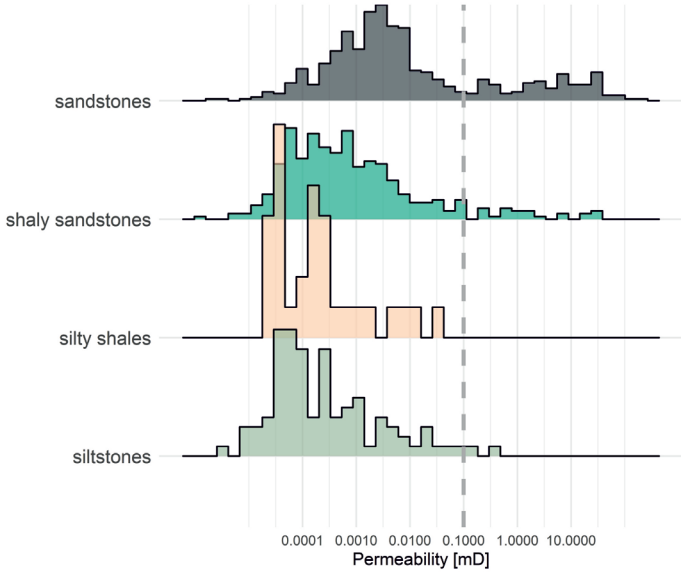

Fig. 2. Distribution of permeability across lithotypes, arranged by descending permeability median. The overwhelming majority of the samples were unconventional tight rocks (k < 0.1 mD)

Rys. 2. Rozkład przepuszczalności dla poszczególnych litotypów. Litotypy zostały uszeregowane przy użyciu mediany przepuszczalności. Zdecydowana większość próbek posiada przepuszczalność < 1 mD (skały niekonwencjonalne)

Porosity and pore structure highly influence a rock's permeability and is as important as permeability in reservoir evaluation (Clarkson et al., 2012; Ma, 2015). The dataset's porosity value ranged from 0.3% to 24.9% (median 6.5%). As expected, the highest porosity and permeability values were observed for sandstone samples and the lowest for silty shale and siltstones (Fig. 3).

The collected data come from an archived siliciclastic core from a depth of 124.1 ft (37.8 m) to 16,723.9 ft (5097.4 m). Depth can also influence the outcome variable, since a greater depth usually indicates higher effective stress and lower permeability (Jones, 1997; Shar et al., 2017). The grain density also varies across the analyzed lithotypes, ranging from 2.34 g/cm^3 to 2.84 g/cm^3 (Fig. 4). The descriptive statistic for all lithotypes is shown in Table 1.

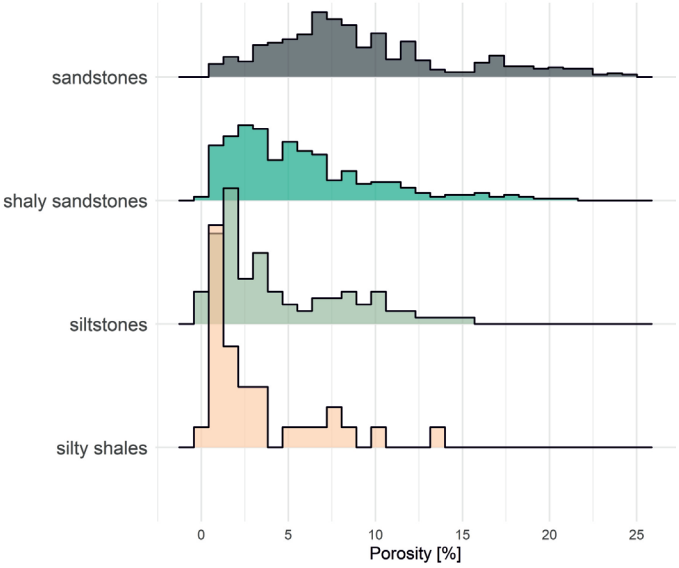

Fig. 3. Distribution of porosity across lithotypes, arranged by descending porosity median

Rys. 3. Rozkład porowatości dla poszczególnych litotypów. Litotypy zostały uszeregowane przy użyciu mediany porowatości

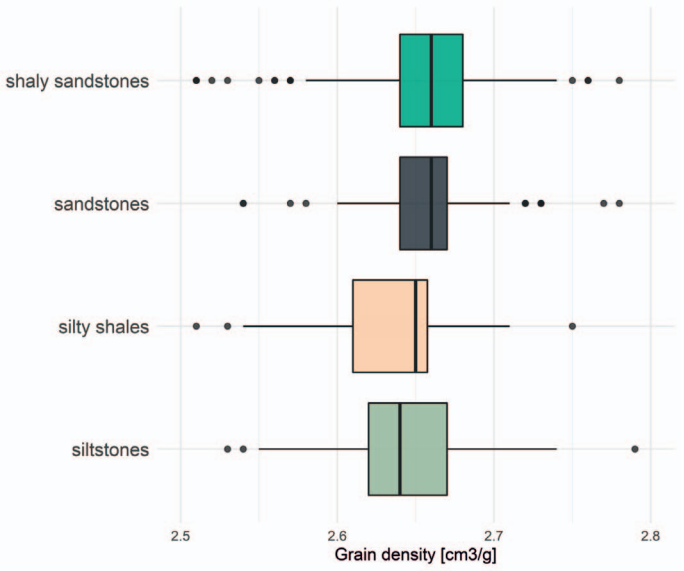

Fig. 4. Distribution of grain density across lithotypes, arranged by descending gd median (black lines in boxes)

Rys. 4. Rozkład gęstości szkieletowej dla poszczególnych litotypów. Litotypy zostały uszeregowane przy użyciu mediany gęstości szkieletowej (czarna pionowa linia)

Using the *rsample* package, the dataset was split into training and testing sets with a proportion of 0.75 (1002/752/250). The 10-fold cross-validation technique was used on the training set to create ten different resamples of analysis (training) and assessment (validation) sets (752/667/75) (Fig. 5). Such an approach produces ten different performance metrics, from which an average model performance can be calculated.

Although the data preprocessing was already partially done (data reduction, transformation, and creation of additional

Table 1. Descriptive statistics for the main petrophysical parameters
Tabela 1. Statystyki opisowe dla głównych parametrów petrofizycznych

Variables	n	mean	sd	median	min	max	range	skew	se
Sandstones									
depth [ft]	532	7844.21	3526.27	7017.00	183.20	16723.90	16540.70	0.36	152.88
k conf. [mD]	532	3.02	12.56	0.00	0.00	171.00	171.00	7.86	0.54
porosity [%]	532	9.21	5.23	7.90	0.70	24.90	24.20	0.87	0.23
grain density [g/cm^3]	532	2.66	0.03	2.66	2.54	2.78	0.24	0.24	0.00
Shaly sandstones									
Depth [ft]	342	7446.60	2926.54	7279.00	124.10	16625.10	16501.00	0.06	158.25
k conf. [mD]	342	0.46	3.26	0.00	0.00	34.00	34.00	8.35	0.18
porosity [%]	342	5.64	4.18	4.70	0.40	21.50	21.10	1.27	0.23
grain density [g/cm^3]	342	2.65	0.05	2.66	2.34	2.84	0.50	–1.52	0.00
Siltstones									
depth [ft]	97	6808.29	3760.32	6700.10	174.00	16653.80	16479.80	0.27	381.80
k conf. [mD]	97	0.01	0.05	0.00	0.00	0.44	0.44	7.87	0.00
porosity [%]	97	4.47	3.83	3.00	0.30	14.90	14.60	0.95	0.39
grain density [g/cm^3]	97	2.64	0.05	2.64	2.36	2.79	0.43	–1.47	0.01
Silty shales									
depth [ft]	31	6761.07	4566.49	6577.30	206.00	16626.00	16420.00	0.48	820.17
k conf. [mD]	31	0.00	0.01	0.00	0.00	0.03	0.03	3.56	0.00
porosity [%]	31	3.27	3.38	1.90	0.40	13.50	13.10	1.38	0.61
grain density [g/cm^3]	31	2.63	0.07	2.65	2.41	2.75	0.34	–1.00	0.01

n –sample size within this group; sd – standard deviation; se – sample standard error; k conf – gas permeability collected at 4000 psi

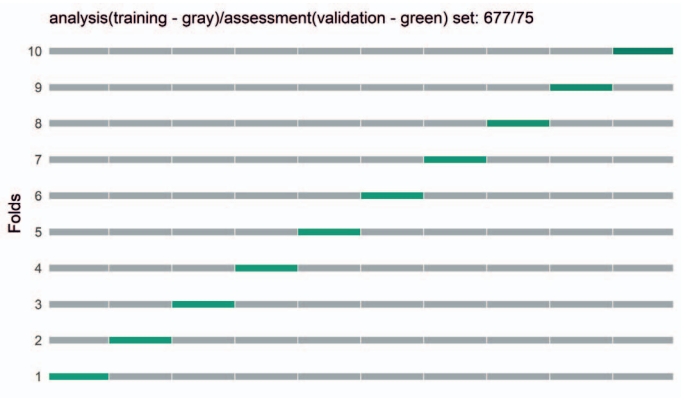

Fig. 5. Visualization of 10-fold cross-validation
Rys. 5. Wizualizacja procesu kroswalidacji

variables), the main part of preprocessing was performed with the *recipes* package. Before the modeling process, the outcome variable needs to be log-transformed and all numeric variables normalized (the center and scale should have a standard deviation of one and a mean of zero). The applied machine learning algorithms handle nominal variables, and thus preprocessing, and the creation of dummy variables was skipped. All operations were done on the training set.

Model specification

The models were trained using MLR and RF algorithms based on the same data quality assumptions. Such an approach allows comparison of the model's performance and selection of the best model. Specification of the model is required to define the mode and the engine (*parsnip* package). MLP, by default, has a "regression" mode and the "lm" engine comes from the *stat* package. In this study, RF has a "regression" mode and "randomForest" engine from the *randomForest* package. The model specification for RF also requires hyperparameters – parameters that control the learning process – to be set. Because hyperparameters cannot be learned from the data, they need to be tuned by assigning different values, training different models, and selecting based on the best results.

The main tunable arguments (hyperparameters) for the RF model are as follows:

- m_{try} – 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.

Out of these three arguments, only m_{try} and min_n were tuned. The m_{try} parameter is the split-variable randomization feature responsible for balancing low tree correlation with predictive strength. The min_n parameters control the splitting scheme and represent the number of observations needed to keep splitting nodes. A high number of trees is recommended to stabilize the error rate. The number of trees was set at 1,000 by default. A detailed description of each hyperparameter can be found in Boehmke and Greenwel (2020). The defined recipes and model specifications were combined using the *workflows* package for each model. Workflows are objects that combine all the information necessary for the modeling process (Kuhn and Silge, 2020).

Hyperparameter tuning was performed on the resamples data using a 30-grid search. This yields 30 possible combinations of the m_{try} and min_n values and 30 different models, each of which was checked against the resample data (300 models). The grid search identified the best m_{try} and min_n values for further modeling. An m_{try} of 3 and a min_n of 12 provided the highest R^2 (rsq) and the lowest standard error (se) among all possibilities.

After the tuning process, the RF model specification was updated (accounting for the best hyperparameters) together with the workflow. Both models were run on training data using defined resamples to get reliable metrics.

Model performance evaluation

The MLR and RF models' performance was evaluated using the residual mean standard error (RMSE), mean absolute error (MAE), and the rsq provided by the *yardstick* package. These indicators are also the most common metrics used to evaluate models (Kuhn and Silge, 2020; Male and Duncan, 2020). Both the RMSE and MAE have the same units as the original data, and both metrics measure the accuracy. However, because the dependent variable was log-transformed, it is hard to determine how well these models performed. When the log transformation is revoked, the RMSE and MAE take high values that do not reflect the actual accuracy. This is because of the high range of the permeability value, from 0.000001 mD to 173 mD (MAE of 1.01 means that on average, the model prediction is wrong by 2.72 mD; RMSE of 1.47 reflects standard error of prediction of ±4.09 mD, assuming a normal distribution of prediction results). Because of that, the rsq was used as the most intuitive metric.

One of the advantages of the *tidymodels* approach is that it is cohesive with the *tidyverse* framework – a collection of packages for data manipulation and visualization. It is therefore easy to combine the modeling results and visualize them in a meaningful way. Figure 6 presents a direct comparison of true and predicted permeabilities. This figure can also be used to evaluate the models' accuracy and spot any outliers: the closer observations are to the 1:1 line, the better the accuracy and the higher the rsq.

Overall, the RF performed with greater predictive accuracy than MLR (Tab. 2, Fig. 6). However, MLP seems to better predict the permeability values for those samples with an

Table 2. Main metrics from modeling results on the training dataset

Tabela 2. Metryki dla wyników modelowania na zbiorze treningowym

Model	Metric	Estimate
MLR	RMSE	1.61
RF	RMSE	1.47
MLR	MAE	1.15
RF	MAE	1.01
MLR	Rsq	0.800
RF	Rsq	0.834

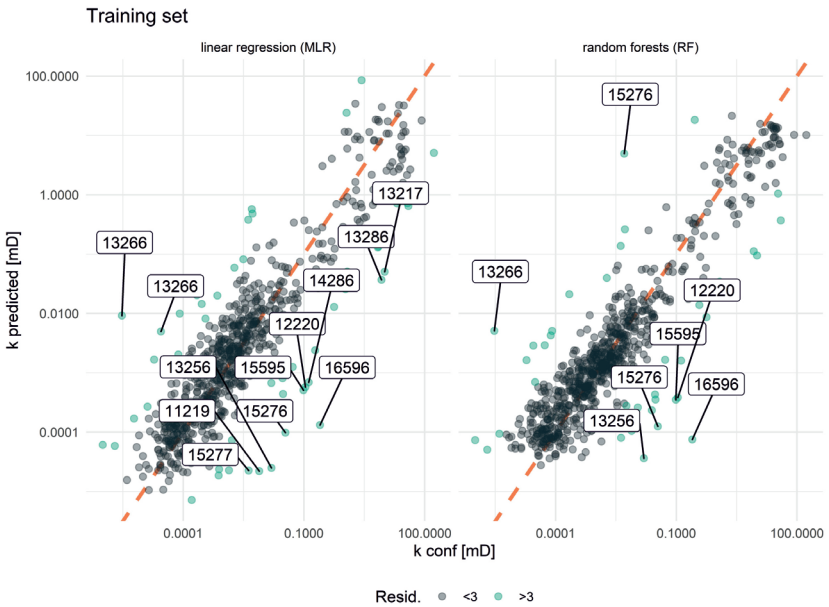

Fig. 6. Model performance on the training set. Green dots represent points for which the residual was greater than 3; labeled samples had a residual greater than 5. The dashed line represents the 1:1 line

Fig. 6. Wyniki modelowania na zbiorze treningowym. Zielone punkty reprezentują obserwacje dla których residua były większe niż 3. Etykiety reprezentują obserwacje, dla których residua były większe niż 5. Przerywana linia reprezentuje linię 1:1

extremely tight structure (k_conf < 0.0001 mD). The reason for that may lie in the linear nature of the regression problem, for which RF could not maintain linearity with the continuous transition, as MLR could. This issue was well described in the work of Ao et al. (2019), where the authors applied a linear RF algorithm to study logging regression modeling in unconventional shale reservoirs. Unfortunately, linear RF has not yet been introduced in *tidymodels*.

The resulting RMSE from MLR and RF models are comparable to the one reported by Byrnes et al. (2009). The researchers (2009) used simple linear regression (LR), MLR, and artificial neural networks (ANN) to predict permeability for tight sand formations. Byrnes et al. (2009) reported a standard error of 4.5 mD, 4.1 mD, and 3.3 mD for LR, MLR, and ANN, respectively. Their results are comparable to those obtained in this study, although models were based on slightly different data quality assumptions (a different number of observations and variables). While Byrnes et al. (2009) used calculated in-situ porosity to predict permeability, ambient porosity was used in this study.

Model interpretation

One way to interpret a model is to perform feature (variable) importance (*vip* package) as a part of the global interpretation (James et al., 2013; Boehmke and Greenwel, 2020). This operation can determine which variables are the most influential to the model. In the case of linear regression models, the feature importance is strictly based on t-statistic. Variables with the most predictive power have the highest t-statistic and the lowest p-value.

In RF, the variable importance is based on the average total reduction of loss for selected features across all trees and the permutation-based importance measure (Molnar, 2019; Boehmke and Greenwel, 2020). In this approach, the increase of prediction error (accuracy) is checked after the variable set is randomly shuffled (one variable at a time). The most important variables are those for which the largest decrease in accuracy is observed after permutation.

The variable importance analysis revealed that the most influential feature was common for both models (Fig. 7). Porosity was expected to be at the top of feature importance since this parameter controls permeability in many sedimentary basins, both conventional and unconventional reservoirs (Pape et al., 1999; Clarkson et al., 2012). For MLR, adding other variables only slightly improves the model. The situation is somewhat different for RF, for which the depth variable seems to influence the permeability predictions (Fig. 7). The RF algorithm considers the interactions between the variables and nonlinear relationships, which can improve model accuracy (finding hidden relationships in the data). The MLR, on the other hand, assumed that all relationships are linear and that there is no interaction between the variables.

Interestingly, the top three variables for RF (most influential to the model) consist of basic information that could be obtained from wireline logging tools (porosity, basic lithology, and depth). Using these three variables in permeability prediction will still provide high accuracy when RF is used. The most important information from decrypted rock type in the RF model is lithology. Sedimentary structures and the pore-filling/cementation variables have only a slight impact on the model's performance. The same can be stated for the variables of basin and formation name. This allows assuming that a simple model with only three variables could be successfully used in other tight sandstone formations of Rocky Mountain basins, which is consistent with conclusions reported by Byrnes et al. (2009).

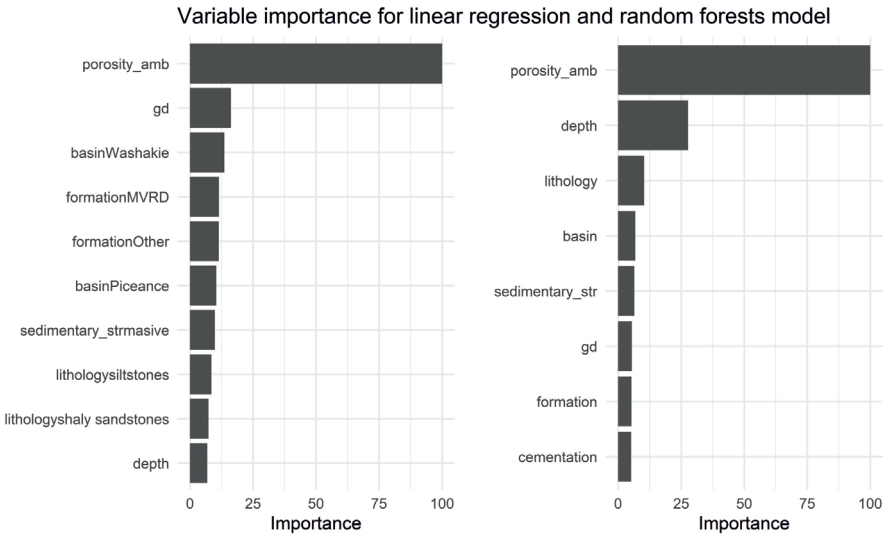

Fig. 7. Variable importance for MLR and RF. The VIP for MLR is directly related to the p-value ($p < 0.05$). The VIP for RF uses permutation-based importance measure loss

Rys. 7. Istotność zmiennych dla MLR i RF. VIP dla MLR jest bezpośrednio związana z p-watością ($p < 0.05$). VIP dla RF bazuje na funkcji straty

Neither model could correctly predict the permeability for samples with rock types 12220, 13256, 13266, 15276, 15595, or 16596 (Fig. 6). For those samples, the residual was greater than 5. Based on rock typing (Byrnes et al., 2009), these are as follows:

- 12220 – siltstones or very shaly sandstones (40–65% clay or silt), well-indurated, moderate-low porosity (3–10%), unfractured, convolute, slumped, large borrow mottled bedding, and with sulfide pore filling;
- 13256 – moderately shaly sandstones (10–40% clay or silt), well-indurated, moderate-low porosity (3–10%), unfractured, quartz pore filling with different sedimentary structures;
- 13256 – flaser bedded, discontinuous mud layers;
- 13266 – sandstone, small-scale (< 4 cm) x-laminated, ripple x-lam with a small-scale hummockey crossbed;
- 15276 – sandstone, medium-grain, well-indurated, moderate-low porosity (3–10%), unfractured, large scale (> 4 cm) trough or planar crossbed, quartz pore filling;
- 15595 – sandstone, medium-grain, indurated, mod-high porosity (> 10%), unfractured, flaser bedded, discontinuous mud layers, calcite pore filling;
- 16596 – sandstone, coarse-grain, indurated, mod-high porosity (> 10%), unfractured, flaser bedded, discontinuous mud layers, quartz pore filling.

For 12220, 15595, and 16596, there were not enough samples to train the model, which could have resulted in poor prediction. The RF prediction for samples of rock type 15276 is overestimated for one sample and underestimated for another one. Overall, there were only two out of 34 samples reported with such a high mismatch. The RF prediction for the sample of rock type 13256 is underestimated (one sample out of 17).

In addition, MLR could not predict samples with the rock types 11219, 13217, 13286, 14286, or 15277. It is difficult to identify a common feature of all the samples for which the models failed. Most of these outliers have a third digit equal to two, which is linked to a well-indurated, moderate-low-porosity (3–10%), unfractured sample. However, they have different sedimentary structures and different pore-filling material.

Model validation

The final step of modeling is to check the models' performance on the testing set (Kuhn and Silge, 2020).

The training and testing set metrics are comparable (compare Tab. 2 to Tab. 3), which means that the models do not overfit. As with the training set, several outliers (residual > 3) were not explained by the models (Fig. 8). The features shared by these outliers are number two on the third position and six and seven on the fifth position from rock type, which refers to a well-indurated, moderate-low-porosity (3–10%), unfractured sample (as in the training set), but with calcite or quartz cementation.

The rsq results obtained from the MLR and RF models were compared to those from the basic linear regression model (LR), which used only one variable to predict permeability, yet with the highest importance (k_conf ~ porosity). This comparison served as an indicator of how adding another variable improves the permeability prediction. The rsq for simple LR, performed on the testing set, shows a high initial correlation of 0.772 (Fig. 9, Tab. 3). The performance of MLR is about 3% better, and RF about 6% better, than that of LR (Tab. 3). However, MLF and RF are complex models that might be hard to implement in real life. Reducing the number of variables to three (porosity, basic lithology, and depth) would blur the line between LR and MLR (rsq of 0.77 vs. 0.78) and would lower the RF performance by about 3% (rsq: 0.80), while still leaving the RF as the best model. If more complex models are preferred or acceptable, RF seems to be a better choice for both the full set of variables and the reduced one.

Table 3. Main metrics from modeling results on the testing dataset
Tabela 3. Metryki dla wyników modelowania na zbiorze testowym

Model	Metric	Estimate
MLR	RMSE	1.59
RF	RMSE	1.48
LR (references)	RMSE	1.71
MLR	MAE	1.11
RF	MAE	0.99
LR (references)	MAE	1.25
MLR	Rsq	0.802
RF	Rsq	0.831
LR (references)	Rsq	0.772

Conclusions and recommendations for future study

This study demonstrates the application of a machine learning approach to predict permeability under confining stress conditions for tight sandstone formations using MLR and RF algorithms. The modeling results from both training and testing sets were evaluated using standard metrics and related to those presented by Byrnes et al. (2009). The RF showed slightly better performance on both the training and testing sets, with an rsq of ~0.83 compared to MLR's rsq of ~0.80. The results were also related to simple LR (reference model) with only one variable – porosity. The RF showed about 6% improvement over the LR when evaluated on the testing set.

The porosity of the MLR and RF models was the most influential variable to the models' performance, based on feature

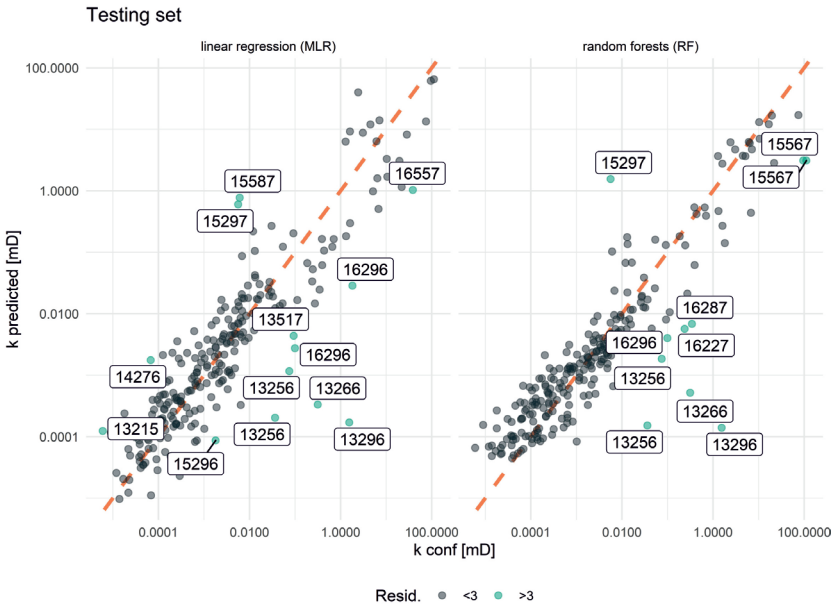

Fig. 8. Model performance on the testing set. Green dots represent points for which the residual was greater than 3; these samples are also labeled. The dashed line represents the 1:1 line

Rys. 8. Wyniki modelowania na zbiorze testowym. Zielone punkty reprezentują obserwacje dla których residua były większe niż 3. Przerywana linia reprezentuje linie 1:1

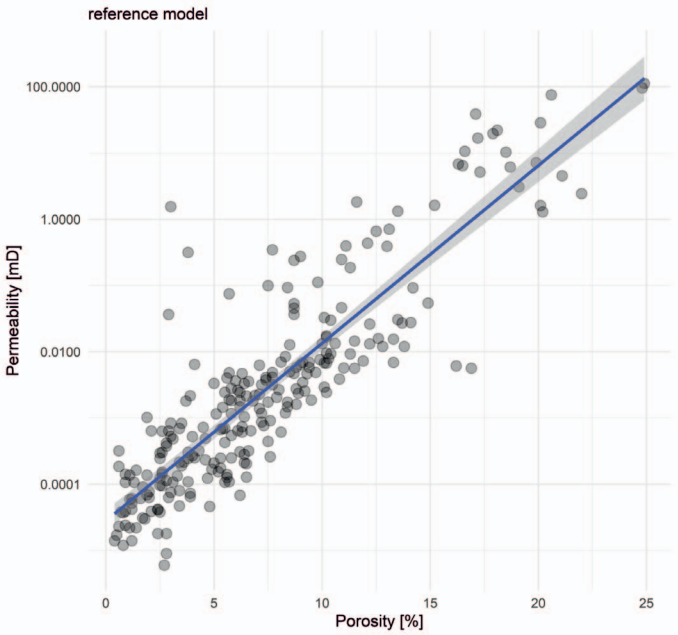

Fig. 9. Porosity-permeability cross plot with a fitted linear model (k_conf ~ porosity). This model is treated as a reference model with an rsq of 0.772

Rys. 9. Wykres porowatość-przepuszczalność z dopasowanym modelem liniowym (k_conf ~ porosity). Dopasowany model jest modelem referencyjnym z R^2 0.772

importance analysis. In RF, the depth variable had a high score that could be evidence of hidden interactions between the variables of porosity, permeability, and depth.

An attempt was made to explain the outliers – observations that were poorly predicted by the models (local interpretation). Based on the rock type characteristics provided by Byrnes et al. (2009), the common feature of outliers for both training and testing sets was moderate-low porosity (3–10%) and a lack of fractures. Also, in the testing set, calcite or quartz cementation was characteristic of outliers.

Acknowledgments I want 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 przewidywania porowatości i przepuszczalności w warunkach złożowych (in-situ)* – the work of the Oil Gas Institute – National Research Institute was commissioned by the Ministry of Science and Higher Education; order number: 0055/SG/2020, archival number: DK-4100-0043/2020.

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.

Attanasi E.D., Freeman P.A., Coburn T.C., 2020. Well predictive performance of play-wide and Subarea Random Forest models for Bakken productivity. *Journal of Petroleum Science and Engineering*, 191: 107150. DOI: 10.1016/j.petrol.2020.107150.

Aulia A., Jeong D., Saaid I.M., Kania D., Shuker M.T., El-Khatib N.A., 2019. A Random Forests-based sensitivity analysis framework for assisted history matching. *Journal of Petroleum Science and Engineering*, 181: 106237. DOI: 10.1016/j.petrol.2019.106237.

Baziar S., Shahripour H.B., Tadayoni M., Nabi-Bidhendi M., 2018. Prediction of water saturation in a tight gas sandstone reservoir by using four intelligent methods: a comparative study. *Neural Computing and Applications,* 30: 1171–1185. DOI: 10.1007/s00521-016-2729-2.

Bhattacharya S., Ghahfarokhi P.K., Carr T.R., Pantaleone S., 2019. Application of predictive data analytics to model daily hydrocarbon production using petrophysical, geomechanical, fiber-optic, completions, and surface data: A case study from the Marcellus Shale, North America. *Journal of Petroleum Science and Engineering*, 176: 702–715. DOI: 10.1016/j.petrol.2019.01.013.

Bhattacharya S., Mishra S., 2018. Applications of machine learning for facies and fracture prediction using Bayesian Network Theory and Random Forest: Case studies from the Appalachian basin, USA.

Journal of Petroleum Science and Engineering, 170: 1005–1017. DOI: 10.1016/j.petrol.2018.06.075.

Boehmke B., Greenwel B., 2020. Hands-On Machine Learning with R. *Chapman and Hall/CRC*.

Brantson E.T., Ju B., Omisore B.O., Wu D., Selase A.E., Liu N., 2018. Development of machine learning predictive models for history matching tight gas carbonate reservoir production profiles. *Journal of Geophysics and Engineering*, 15: 2235–2251. DOI: 10.1088/1742-2140/aaca44.

Byrnes A.P., Cluff R.M., Webb J.C., 2009. Analysis of Critical Permeability, Capillary Pressure, and Electrical Properties for Mesaverde Tight Gas Sandstones from Western U.S Basins. *DOE report DE-FC26-05NT42660*. http://www.kgs.ku.edu/mesaverde/ (dostęp: 31.03.2021).

Caté A., Perozzi L., Gloaguen E., Blouin M., 2017. Machine learning as a tool for geologists. *Leading Edge*, 36: 215–219. DOI: 10.1190/tle36030215.1.

Clarkson C.R., Jensen J.L., Chipperfield S., 2012. Unconventional gas reservoir evaluation: What do we have to consider? *Journal of Natural Gas Science and Engineering*, 8: 9–33. DOI: 10.1016/j.jngse.2012.01.001.

Comisky J.T., Newsham K., Rushing J.A., Blasingame T.A., 2007. A Comparative Study of Capillary-Pressure-Based Empirical Models for Estimating Absolute Permeability in Tight Gas Sands. *SPE Annual Technical Conference and Exhibition*. DOI: 10.2118/110050-MS.

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.

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

Jamshidian M., Mansouri Zadeh M., Hadian M., Moghadasi R., Mohammadzadeh O., 2018. A novel estimation method for capillary pressure curves based on routine core analysis data using artificial neural networks optimized by Cuckoo algorithm – A case study. *Fuel*, 220: 363–378. DOI: 10.1016/j.fuel.2018.01.099.

Jones S.C., 1997. A technique for faster pulse-decay permeability measurements in tight rocks. *SPE Formation Evaluation*, 12: 19–24. DOI: 10.2118/28450-pa.

Karpatne A., Ebert-Uphoff I., Ravela S., Babaie H.A., Kumar V., 2019. Machine Learning for the Geosciences: Challenges and Opportunities. *IEEE Transactions on Knowledge and Data Engineering*, 31: 1544–1554. DOI: 10.1109/TKDE.2018.2861006.

Kuhn M., Silge J., 2020. Tidy Modeling with R. https://www.tmwr.org/ (dostęp: 31.03.2021).

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

Louppe G., 2014. Understanding Random Forests: From Theory to Practice. U*niversity of Liège*. http://arxiv.org/abs/1407.7502 (dostęp: 31.03.2021).

Ma Y.Z., 2015. Unconventional resources from exploration to production, Unconventional Oil and Gas Resources Handbook: Evaluation and Development. *Elsevier Inc*. DOI: 10.1016/B978-0-12-802238-2.00001-8.

Ma Y.Z., Moore W.R., Gomez E., Clark W.J., Zhang Y., 2015. Tight Gas Sandstone Reservoirs, Part 1: Overview and Lithofacies, Unconventional Oil and Gas Resources Handbook: Evaluation and Development. *Elsevier Inc*. DOI: 10.1016/B978-0-12-802238-2.00014-6.

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.

McPhee C., Reed J., Zubizarreta I., 2015. Core analysis: A Best Practice Guide. First edit. *Elsevier, Amsterdam, Netherlands*. DOI: 10.1016/B978-0-444-63533-4.09991-1.

Meshalkin Y., Koroteev D., Popov E., Chekhonin E., Popov Y., 2018. Robotized petrophysics: Machine learning and thermal profiling for automated mapping of lithotypes in unconventionals.*Journal of Petroleum Science and Engineering*, 167: 944–948. DOI: 10.1016/j.petrol.2018.03.110.

Molnar C., 2019. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. *Leanpub book.*

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

Pape H., Clauser C., Iffland J., 1999. Permeability prediction based on fractal pore-space geometry. *Geophysics*, 64: 1447–1460. DOI: 10.1190/1.1444649.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing: Vienna*.

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.

Rubo R.A., de Carvalho Carneiro C., Michelon M.F., Gioria R. dos S., 2019. Digital petrography: Mineralogy and porosity identification using machine learning algorithms in petrographic thin section images. *Journal of Petroleum Science and Engineering*, 183: 106382. DOI: 10.1016/j.petrol.2019.106382.

Shar A.M., Mahesar A.A., Chandio A.D., Memon K.R., 2017. Impact of confining stress on permeability of tight gas sands: an experimental study. *Journal of Petroleum Exploration and Production Technology*, 7: 717–726. DOI: 10.1007/s13202-016-0296-9.

Such P., Leśniak G., Budak P., 2007. Kompleksowa metodyka badania właściwości petrofizycznych skał. *Prace Instytutu Górnictwa Naftowego i Gazownictwa*, 142: 1–69.

Such P., Dudek L., Mroczkowska-Szerszeń M., Cicha-Szot R. 2015. The influence of reservoir conditions on filtration parameters of shale rocks. *Nafta-Gaz*, 11: 827–832. DOI: 10.18668/NG2015.11.03.

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: 219–229. DOI: 10.18668/ng.2020.04.01.

Wendt W.A., Sakurai S., Nelson P.H., 1986. Permeability prediction from well logs using multiple regression., Reservoir characterization. *Academic Press, Inc*. DOI: 10.1016/b978-0-12-434065-7.50012-5.

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.

Wu J., Yin X., Xiao H., 2018. Seeing permeability from images: fast prediction with convolutional neural networks. *Science Bulletin*, 63: 1215–1222. DOI: 10.1016/j.scib.2018.08.006.

Tomasz TOPÓR, PhD
Head of Petrophysics Laboratory
Department of Geology and Geochemistry
Oil and Gas Institute – National Research Institute
Lubicz 25 A
31-503 Krakow
E-mail: *toport@inig.pl*