An advanced ensemble modeling approach for predicting carbonate reservoir porosity from seismic attributes

Tomasz Topór[1], Krzysztof Sowiżdżał[2]

[1] Oil and Gas Institute-National Research Institute, Krakow, Poland, e-mail: topor@inig.pl (corresponding author), ORCID ID: 0000-0002-5306-4636

[2] Oil and Gas Institute-National Research Institute, Krakow, Poland, e-mail: sowizdzal@inig.pl, ORCID ID: 0000-0002-6367-6273

Received: 16 January 2023; accepted: 26 June 2023; first published online: 6 September 2023

Abstract: This study uses a machine learning (ML) ensemble modeling approach to predict porosity from multiple seismic attributes in one of the most promising Main Dolomite hydrocarbon reservoirs in NW Poland. The presented workflow tests five different model types of varying complexity: K-nearest neighbors (KNN), random forests (RF), extreme gradient boosting (XGB), support vector machine (SVM), single layer neural network with multilayer perceptron (MLP). The selected models are additionally run with different configurations originating from the pre-processing stage, including Yeo–Johnson transformation (YJ) and principal component analysis (PCA). The race ANOVA method across resample data is used to tune the best hyperparameters for each model. The model candidates and the role of different pre-processors are evaluated based on standard ML metrics – coefficient of determination (R2), root mean squared error (RMSE), and mean absolute error (MAE). The model stacking is performed on five model candidates: two KNN, two XGB, and one SVM PCA with a marginal role. The results of the ensemble model showed superior accuracy over single learners, with all metrics (R2 0.890, RMSE 0.0252, MAE 0.168). It also turned out to be almost three times better than the neural net (NN) results obtained from commercial software on the same testing set (R2 0.318, RMSE 0.0628, MAE 0.0487). The spatial distribution of porosity from the ensemble model indicated areas of good reservoir properties that overlap with hydrocarbon production fields. This observation completes the evaluation of the ensemble technique results from model metrics. Overall, the proposed solution is a promising tool for better porosity prediction and understanding of heterogeneous carbonate reservoirs from multiple seismic attributes.

Keywords: machine learning, model stacking, ensemble method, carbonates, seismic attributes, porosity prediction

INTRODUCTION

The inherent feature of carbonate reservoirs is significant heterogeneity that originates from multiple stages of diagenesis. This process overprints the initial depositional environment features and shapes the pore structure characteristics, which is the most complicated in carbonate rocks. Porosity in carbonate reservoirs manifests itself as a dual or triple system with pore space, creating vugs, intercrystalline pores, and fractures of different sizes and distributions (Moore & Wade 1989, Tiab & Donaldson 2015). The complex diagenetic history of carbonate systems also dilutes the relationship between reservoir property characteristics and seismic response, making this type of reservoir extremely unpredictable in the subsurface. The statement is particularly true regarding the

https://doi.org/10.7494/geol.2023.49.3.245

Vp-porosity relationship, which is much less constrained in carbonates than in siliciclastic reservoirs (Hendry et al. 2021). The described features have long been recognized as a challenge for evaluating fundamental reservoir properties at the reservoir scale.

The recent advances in machine learning (ML) tools have shed new light on the recognition and characterization of carbonate systems (Hendry et al. 2021). One unique feature of ML algorithms is their capability to synthesize high-dimensional data and find hidden interactions between them, making them a powerful tool for studying complex and heterogeneous carbonate systems where the relationships between reservoir properties and seismic response are highly non-linear and ambiguous (Hendry et al. 2021). Multi-attribute ML processes have been successfully used for seismic facies recognition (e.g., Jesus et al. 2019, Pattnaik et al. 2020, Carvalho et al. 2022), evaluation of reservoir properties (e.g., Sinaga et al. 2019, Hou et al. 2022), detection of reservoir quality and sweet spots of carbonate reservoirs (e.g., Chen et al. 2021). Other authors have hybridized unsupervised and supervised methods for comprehensive carbonate facies classification and subsequent porosity-permeability prediction (e.g., Ferreira et al. 2021).

Another interesting strategy involves the application of ensemble models, where predictions of single learners are combined to make a final prediction (Couch & Kuhn 2022, Kuhn & Silge 2022). Although ensemble techniques are popular in single methods such as bagging, random forest, and boosting (Breiman 1996a, 2001, Freund & Schapire 1997), the technique originates in stacking many models of different types (Wolpert 1992, Breiman 1996b). Model stacking has already proven superior predictive performance in various settings and is often used as a winning solution in ML competitions (see: the winning solutions of Kaggle competitions at www.kaggle.com).

The application of model stacking (or ensemble modeling) has also received attention in reservoir characterization studies of porosity, permeability and water saturation (e.g., Adeniran et al. 2019, Bedi & Toshniwal 2019, Otchere et al. 2021a). Most of the published studies utilize multiply regression, artificial neural network (ANN), and support vector machines (SVM) as an ensemble (e.g., Chen & Lin 2006, Reza et al. 2011, Anifowose et al. 2013, 2015, Helmy et al. 2013). However, other methods such as random forest (FR) and extreme gradient boosting (XGB) also have been successfully applied for reservoir characterization (Otchere et al. 2021a, Liu et al. 2022). Both RF and XRG are known for their superior accuracy and the ability to handle unstructured data (James et al. 2013, Hall & Hall 2017), making them good solution when analyzing complex relationships between reservoir properties and seismic response (Topór & Sowiżdżał 2022).

The study demonstrates the capability of an ensemble model for the porosity prediction of carbonate reservoirs from seismic attributes. The study's novelty lies in a complex workflow involving advanced data pre-processing, enhanced models' hyperparameters tuning, model stacking with different candidate types, and comprehensive evaluation of model results using ML metrics and visual inspection with the spatial distribution of porosity.

The pre-processing data stage involves data transformation (normalization, Yeo–Johnson), data reduction (PCA), and feature engineering with the application of the unsupervised k-means method to determine seismic facies based on pre-selected seismic attributes. The study assesses the role of individual pre-processing types in porosity prediction. The tuning strategy is conducted using one of the most efficient ways of determining models' hyperparameters – the race ANOVA method. Stacking is performed on models of different levels of complexity, such as K-nearest neighbors (KNN), random forests (RF), extreme gradient boosting (XGB), support vector machine (SVM), and single layer neural network with multilayer perceptron model (MLP). These models are commonly used to build the modeling strategy and are known for their superior prediction accuracy (Kuhn & Johnson 2013, Hall & Hall 2017).

The obtained stacked model is compared to single learners as well as to neural network results (NN) obtained from commercial software.

All models are evaluated using standard ML metrics, including coefficient of determination (R2), mean squared error (RMSE), and mean absolute error (MAE). Finally, the obtained results are spatially distributed to check if the porosity patterns followed the one expected from the known research works and experience acquired during the exploration, appraisal, and field production stages.

DATASET

The analyzed dataset refers to one of the most promising Main Dolomite hydrocarbon reservoirs in NW Poland. It consists of seismic attributes calculated from a 3D seismic cube that describes the area of the carbonate platform, barriers, slope, and bottom facies. The analyzed dataset consists of more than five million (5,476,146) observations and 35 variables. In addition to the seismic attributes variable and coordinate information, the analyzed dataset included estimated porosity (Phi) as an outcome variable (2,278 observations). Porosity was derived using the standard crosslots method based on well-log profiles (neutron-acoustics, neutron-density).

EXPLANATORY DATA ANALYSIS (EDA)

The analyzed hydrocarbon system is highly diverse in terms of storage properties. Porosity varies from 0% to 31.2% (mean 11.3%, median 6.8%), and its distribution is skewed with a high population of porosity less than 5% (Fig. 1).

According to Mikołajewski & Wróbel (2005), the Main Dolomite carbonate rocks consist of both vogues and intercrystalline porosity, complicating the presence of microfractures which are common in this type of rock. The deposition environment initially controlled the pore network development, but the most important factor was the multiple stages of diagenetic changes (Kotarba & Wagner 2007). Burial diagenesis in the analyzed hydrocarbon system deteriorated and enhanced storage and filtration properties. Compaction, recrystallization, and cementation (mainly with anhydrite and dolomite) had the most adverse impact on the pore network, while the dissolution of carbonate grains with pore fluids enriched in CO_2 and fracturing (in micro and macro scales) positively impacted reservoir properties (Mikołajewski & Wróbel 2005, Kotarba & Wagner 2007).

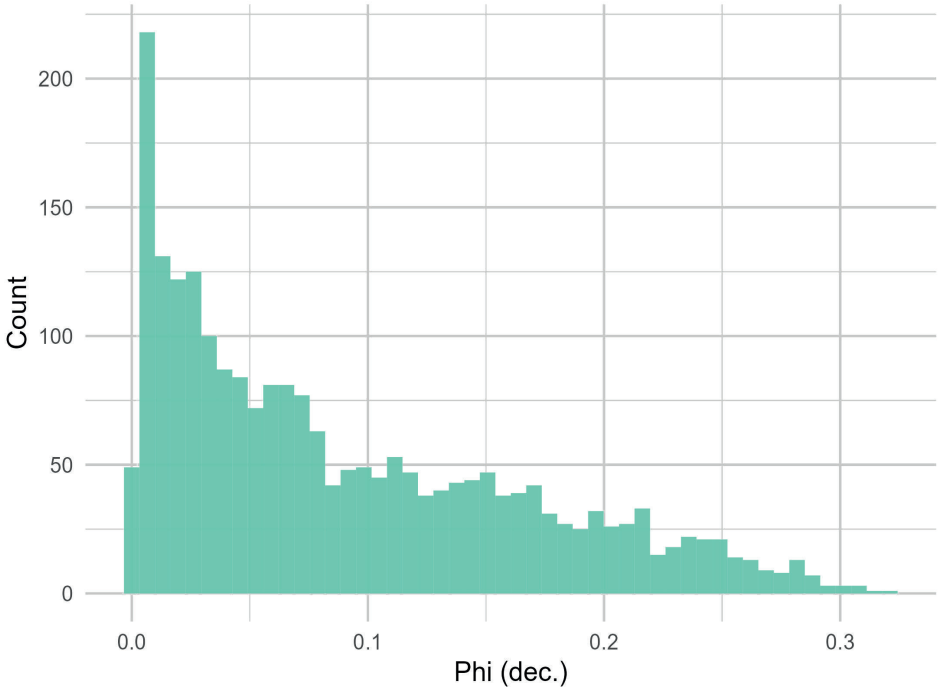

Fig. 1. *Distribution of porosity (Phi) in the analyzed dataset*

The investigation of the relationship between seismic attributes and log-based porosity data was conducted within the framework of the 3D grid into which seismic data were resampled, and well-log profiles upscaled. It required the acceptance of a compromise between the loss of well-log data accuracy resulting from the upscaling procedure (the arithmetic average method was implemented) and the apparent increase of seismic data resolution resulting from its downscaling. Nonetheless, this procedure ensured both datasets were transformed into a standard 3D grid domain, allowing further analysis using machine learning methods.

The relationship between porosity and seismic attributes in carbonate reservoirs is complex, making porosity prediction challenging. The correlation between porosity and seismic variables diverges from linear and monotonic relations, expressed in its low Spearman's rank correlation coefficient (Fig. 2). The correlation matrix presented in Figure 2 reveals another problem, which complicates prediction and may lead to uncertainty in ML modeling. This issue is a high cross-correlation (collinearity) between seismic attributes. The cross-correlated variables were grouped using hierarchical clustering and the Lance–Williams dissimilarity update formula with Ward's method. The method uses the classical sum-of-squares criterion, producing groups that minimize within-group dispersion (Ward 1963, Lance & Williams 1966, Murtagh & Legendre 2014). The number of clusters was arbitrarily set to eight (Fig. 2).

Figure 3 shows the most significant variables for porosity prediction. Although these variables are statistically significant (p-value < 0.05), their correlation coefficients are weak and below 0.5. Additionally, some of them are also highly cross-correlated (Fig. 3). Several of the listed seismic attributes are important for seismic characterization study, including the top-ranked attribute, acoustic impedance (Rel_AI), which is commonly used to target reservoir potential (Hendry et al. 2021). A similar role can be assigned to Vp and its low-velocity anomalies within carbonate strata.

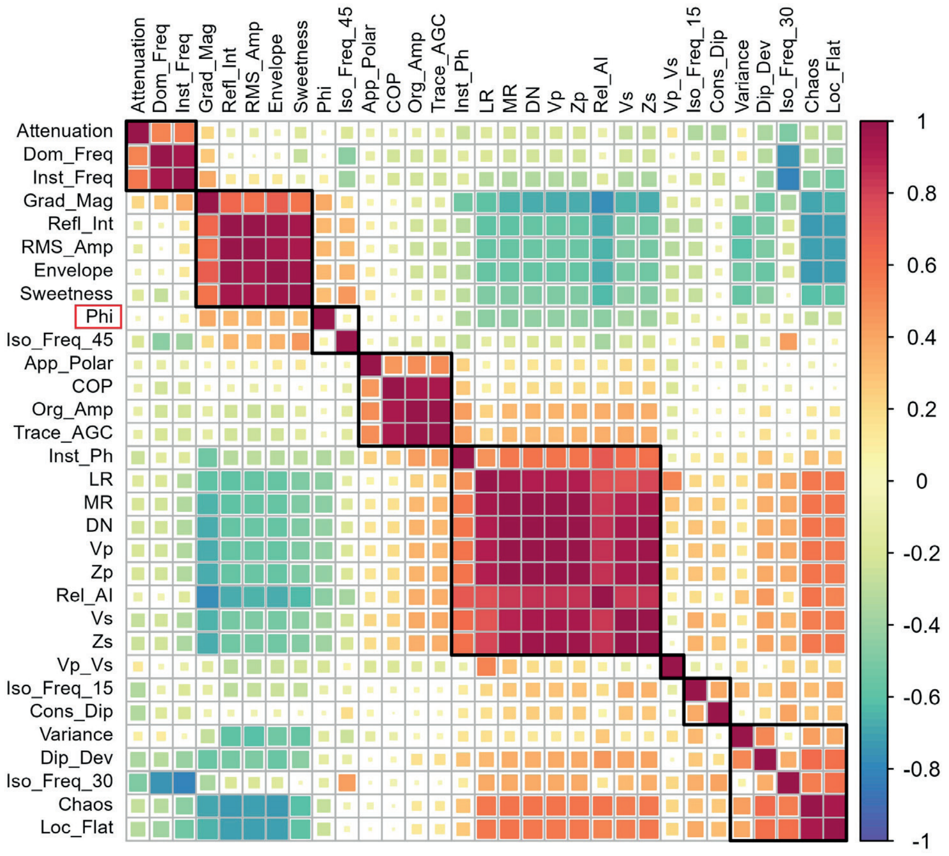

Fig. 2. *A correlation matrix with marked highly cross-correlated variables based on hierarchical clustering. An outcome variable (Phi) was highlighted. A description of the individual seismic attribute is presented in Appendix 1*

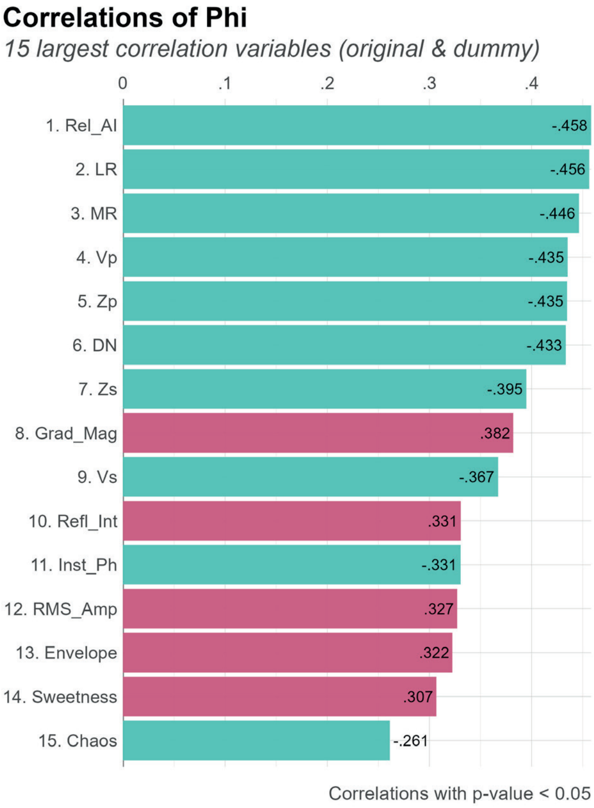

Fig. 3. *Fifteen (15) most significant correlation variables for porosity prediction with their correlation coefficients*

Generally, seismic attributes resulting from simultaneous seismic inversion workflow and those related to seismic signal amplitude reveal the highest correlation with porosity. Since they record similar information contained in seismic data, they will be further processed into predictive models before their implementation.

METHODS

The workflow applied in this study uses the tidymodels framework and the latest concepts developed by R Core Team for modeling and machine learning (Kuhn & Silge 2020, Kuhn & Silge 2022, R Core Team 2022). The workflow is coupled with a 3D visualization of the obtained results from Petrel software (Fig. 4).

Data pre-processing and feature engineering

The EDA reveals several problems with variables that need to be considered in the pre-processing stage before the modeling. The pre-processing and feature engineering was performed using a recipe package from tidymodels meta-packages (Kuhn & Silge 2022).

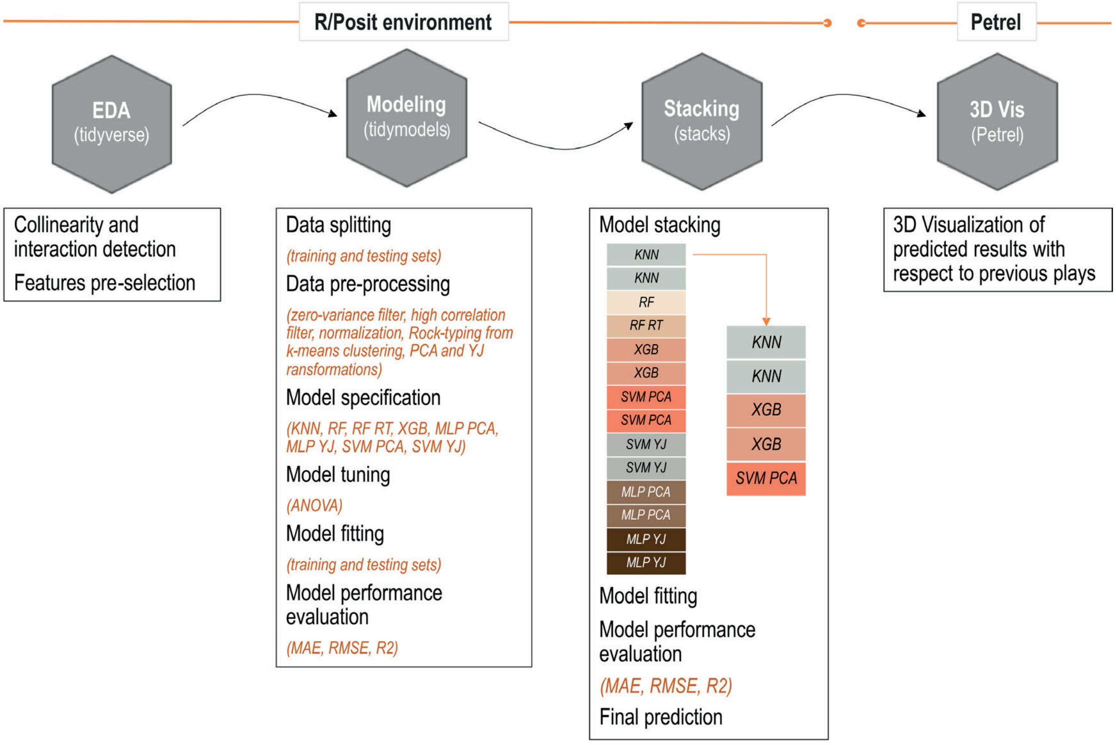
R/Posit environment
Petrel
EDA (tidyverse)
Modeling (tidymodels)
Stacking (stacks)
3D Vis (Petrel)
Collinearity and interaction detection
Features pre-selection
Data splitting
(training and testing sets)
Data pre-processing
(zero-variance filter, high correlation filter, normalization, Rock-typing from k-means clustering, PCA and YJ ransformations)
Model specification
(KNN, RF, RF RT, XGB, MLP PCA, MLP YJ, SVM PCA, SVM YJ)
Model tuning
(ANOVA)
Model fitting
(training and testing sets)
Model performance evaluation
(MAE, RMSE, R2)
Model stacking
KNN
KNN
RF
RF RT
XGB
XGB
SVM PCA
SVM PCA
SVM YJ
SVM YJ
MLP PCA
MLP PCA
MLP YJ
MLP YJ
KNN
KNN
XGB
XGB
SVM PCA
Model fitting
Model performance evaluation
(MAE, RMSE, R2)
Final prediction
3D Visualization of predicted results with respect to previous plays

Fig. 4. *Workflow of the study coupling R/Posit environment and Petrel software*

Since the study uses stacked models for porosity prediction that involves several models of different levels of complexity, the pre-processing stage was grouped for the model types.

The KNN, RF, and XGB use three-step data pre-processing that involves the zero-variance filter, high correlation filter, and data normalization (base pre-processing). The applied filters remove variables containing only a single value and those with significant absolute Spearman correlations with other variables. The normalization step scales numeric data to have a standard deviation of one and a mean of zero. The applied procedure reduced the number of variables to 21 numerical features.

The RF was additionally run with an extra class variable that indicates seismic facies. This new feature was derived from unsupervised *k*-means classification and a combination of stratigraphic and structural seismic attributes (Rel_AI, RMS_Amp, Loc_Flat, Chaos, Variance, Vp, Dip_Dev) (Randen & Sønneland 2005). A similar approach was described by Ferreira et al. (2021). The number of clusters was set to potentially represent three facies of carbonate platforms, build-up, and debris. The *k*-means clustering was performed using Euclidean distances measurement and the Hartigan–Wong algorithm (Hartigan & Wong 1979). A detail of the *k*-means method and its application for rock-typing can be found in Topór (2020).

The data pre-processing and feature engineering for the MLP and SVM models was additionally extended for the Yeo–Johnson transformation. Its main role was transforming continuous variables to be more normally distributed (Yeo & Johnson, 2000). The algorithm is similar to the Box-Cox but does not require the input variables to be strictly positive, which is important when handling seismic attributes, which are skewed and frequently with negative values. Both algorithms were also run in a variant after principal component analysis (PCA). This method transforms variables into a set of artificial components, which capture the maximum amount of information in the original variables and, in the same way, combat significant inter-variables correlations in a data set (Jolliffe 2010). The number of components was arbitrarily set to five.

The presented pre-processing stage produced eight different model configurations: KNN, RF, RF with seismic facies (defined as rock-type: RF RT), XGB, MLP with Yeo–Johnson transformation (MLP YJ) MLP with PCA (MLP PCA), SVM with Yeo–Johnson transformation (SVM YJ), and SVM with PCA (SVM PCA).

The dataset was split into a training set and a test set using a 0.8 ratio. The 10-fold cross-validation was used on the training set to obtain ten resampling sets for analysis and assessment. In addition, the outcome variable was used to conduct stratified sampling. This operation helps ensure that the resamples have equivalent proportions of porosity range as in the original data set (see Fig. 1).

Model description and specification

Stacking was performed on a suite of high-performance models such as KNN, RF, XGB, MLP, and SVM, which, self-alone or combined, are frequently used to build the modeling strategy (Kuhn & Johnson 2013). Part of them (XGB and RF) were also top models in the contest organized by the Society of Exploration Geophysicists, which involved interpreting data from well-log analysis (Hall & Hall 2017). Before stacking, the models were run individually on resamples.

The tidymodels workflow required the specification of the mode and engine of the model. The mode is common for all models and is set to regression. The regression models were trained using the KNN algorithm with “kknn” engine, RF with “ranger” engine, XGB with “xgboost” engine, MLP with “nnet” engine, and SVM with “kernlab” engine.

The KNN is a simple algorithm that relies on the *k* most similar data points and sophisticated distance metrics to generate accurate predictions. Because the model is based on *k* nearest neighbors, it is inherently local and cannot be summarized by a closed-form model (Molnar 2019, Boehmke & Greenwell 2020). It also means that the right *k* number will determine its performance. Besides the number of neighbors, the KNN has two other hyperparameters that can be tuned. These are a type of kernel function used to weight distances (weight_func) and a single number for the parameter used in calculating Minkowski distance (dist_power). Kuhn & Johnson (2013) and Boehmke & Greenwell (2020) provide details of the method with the full specification.

The RF and XGB are decision tree algorithms with the exact model representation and inference but a different training algorithm. While RF creates an extensive collection of de-correlated (independent) trees to improve predictive performance, the XGB builds an ensemble of shallow trees in sequence, where each tree learns and improves on the previous (Yoav & Schapire 1997, Boehmke & Greenwell 2020). The number of hyperparameters also distinguishes the two algorithms. For the RF, the operator needs to specify the number of predictors randomly sampled at each tree split (mtry) and the minimum number of observations in a node required for further split (min_n). The number of trees within the ensemble (trees) was left as the default (500). The XGB uses the gradient descent optimization algorithm to update model parameters. Two extra hyperparameters regulate the algorithm – the number representing the rate at which the boosting algorithm adapts from iteration to iteration (learn_rate) and the number for the reduction in the loss function required to split further (loss_reduction). Both RF and XGB were extensively described by Topór (2021) and Topór & Sowiżdżał (2022).

The SVM model uses kernel-induced feature space to find a fitting hyperplane with good generalization based on original features. Using a robust loss function (ε-insensitive loss), the algorithm tries to form a margin around the regression hyperplane of ε width that includes as many observations within the margin as possible. The observations whose residual satisfy $r(x, y) \pm \varepsilon$ form the support vectors that define the margin. The kernel function can capture complex non-linear relationships (Boehmke & Greenwell 2020). The SVM algorithm has four hyperparameters, but only two were tuned in this study – the cost of predicting a sample within or on the wrong side of the margin (cost) and the positive number for the polynomial degree (degree). The remaining hyperparameters, such as the polynomial scaling factor (scale_factor) and the ε in the SVM insensitive loss function (margin), were set by default.

The MLP is a supplement of a single-layer, feed-forward neural network (NN) and defines a multilayer perceptron model. The model has three layers – the input layer, which receives the input signal to processes; the output layer, which performs prediction; and the in-between hidden layer, which performs non-linear transformations to assign weights of the inputs provided to the network. The process is done through the activation function that is automatically set depending on the type of the outcome variable. The number of units (units) is a crucial hyperparameter that determines model performance. Besides the hidden layer, two other parameters can be tuned – the penalty, which defines the amount of regularization to simplify the model, and epochs, which establishes the length of training. Both hyperparameters help overcome overfitting and enhance generalization performance, which is an ability to appropriately adapt to new, previously unseen data. Details of the method are provided by Abirami & Chitra (2020).

Tuning strategy

The modeling process involves testing eight models before stacking. Selecting the best hyperparameters for each model using a classical approach with a grid search would be both time and resource-consuming but, more importantly, inefficient. As a result, grid searching was performed via racing with ANOVA models. The method computes RMSE for selected tuning parameters, pre-defined in model specification, across resample data. The algorithm tests the statistical significance of tuning parameter combinations and eliminates those which are not prospective using the repeated measure ANOVA model (Kuhn 2014).

The applied tuning strategy involved testing 120 model combinations (15 for each model type) to select the best settings for each model.

RESULTS AND DISCUSSION

The performance of individual regression models was assessed after fitting the final models (with tuned hyperparameters) to the training set and then evaluating them with the testing set. This operation eliminates the issue of overfitting that can occur in the training set. The assessment was performed using standard ML metrics – R2, RMSE, and MAE (Table 1). The RMSE and MAE have the same units as the outcome variable, and R2 ranges from 0 to 1. Since the distribution of the outcome variable is skewed, the MAE and R2 seem to reflect the model accuracy best. The results of R2 revealed that the models with the highest predictive power are KNN, XGB, and RF (Table 1). The MAE 0.0173–0.0202 for the best models reflects the error of porosity prediction of ~1.7–2.0% (porosity as a percentage).

Table 1
Model metrics for the best model type from the tuning results on the testing set

Model	R2	RMSE	MAE
KNN	0.875	0.0266	0.0173
XGB	0.868	0.0273	0.0191
RF RT	0.857	0.0285	0.0202
RF	0.856	0.0286	0.0202
SVM YJ MLP PCA	0.722 0.612	0.0402 0.0471	0.0276 0.0362
SVM PCA	0.415	0.0583	0.0414
MLP YJ	0.343	0.0610	0.0464

The top rank of ensemble models (XGB and RF) is not surprising. Both models are known for their computational efficiency and superior accuracy (Chen & Guestrin 2016, Hall & Hall 2017). XGB and RF are designed to deal with collinearity, handle data that are not structurally designed, and consider the hidden relationships between the variables (James et al. 2013, Kuhn & Johnson 2013). These features could play a key role when modeling porosity from seismic attributes for which colinearity is a significant issue and non-linear relationships are common. The results showed that the XGB and RF could predict with 87% and 86% accuracy, respectively. The accuracy of RF with seismic facies variable (RF RT) was on the same level as base RF. These results neglected the role of *k*-means classification in porosity prediction for the studied carbonate platform, though the method is commonly used to evaluate seismic carbonate facies (Ferreira et al. 2021, Carvalho et al. 2022). Ferreira et al. (2021) demonstrated the application of an unsupervised classification that differentiates between carbonate platform, build-ups, and the debris seismic facies in Bare carbonate formation from Brazil. The obtained results were used for porosity and permeability prediction with the MLP method. The authors, however, did not report the porosity and permeability modeling results without the seismic facies, making the statement that the high heterogeneity of the formation restricts the direct and marked correlation between the mapped seismic facies and petrophysical properties. The case could be similar for the studied formation.

What is surprising, though, is the high rank of the KNN model, which is not very popular in predicting reservoir properties such as porosity. Raheem & Shuker (2021) have used KNN and recurrent neural networks (RNN) with one of the long- and short-term memory (LSTM) algorithms to predict porosity from seismic attributes. The authors, however, showed superior accuracy of LSTM over KNN. Wardhana & Pratama (2021) used KNN to predict porosity from well-log data. They showed that KNN performed with better accuracy when compared to artificial neural networks (ANN) and support vector regression (SVR). Other application of KNN in formation evaluation focuses on classification problems and rock-typing (e.g., Al-Amri et al. 2017, Hou et al. 2022). The testing set results show that KNN is highly efficient and can predict porosity with the same accuracy as RF and XGB (Table 1). The high score of the KNN model on the testing set also excludes the possible overfitting issue.

Although SVM and MLP have a long history in various aspects of reservoir characterization studies (Dramsch 2020, Otchere et al. 2021b), their performance is not satisfactory for the analyzed dataset (Table 1). The accuracy of these models varies with respect to the applied pre-processing. The PCA is frequently used to decrease the redundancy of the seismic attributes before the modeling, which is a common issue in seismic characterization studies (Chopra et al. 2018, Jesus et al. 2019, Carvalho et al. 2022). The results showed that PCA had a limited effect on improving the MLP model but significantly deteriorated the performance of SVM (Table 1). The Yeo–Johnson transformation positively affected the predictive power of the SVM model, placing it in the fifth position in the overall ranking with the accuracy of 72% (Table 1).

The model stacking was performed with the eight model definitions and 14 most promising candidate members, where two come from the KNN model, two from the RF model (one for each model configuration), two arise from the XGB model, four from the SVM models (two for each model configuration), and four from the MLP model definition (two for each model configuration). To predict porosity from each candidate member, the models were blended and fitted using an elastic net model (Hastie et al. 2015, Couch & Kuhn 2022). The penalty and model mixture for the elastic net model was evaluated based on the lowest RMSE (default settings) and set for 0.0001 and 1, respectively. The penalty parameter helps to overcome overfitting by forcing the regression estimator to shrink its coefficients toward 0.

Of 14 possible candidate members, the ensemble retained five with non-zero stacking coefficients. These models belong to two KNN, two XGB, and one SVM with PCA transformation (Fig. 5).

The obtained coefficients create weightings for each of the member models in a final model prediction:

$$\begin{aligned}Phi_{ensemble\ pred.} = &-0.00046 + 0.33487 \times \text{KNN}(\text{No } 4) + \\ &+ 0.13263 \times \text{KNN}(\text{No } 9) + \\ &+ 0.22525 \times \text{XGB}(\text{No } 14) + \\ &+ 0.31699 \times \text{XGB}(\text{No } 2) + \\ &+ 0.00004 \times \text{SVM PCA}(\text{No } 12)\end{aligned}$$

The presence of SVM PCA is surprising, but its role in porosity prediction is still marginal, with weights of 0.00004. The final ensemble model assessment was evaluated on the testing set. The stack model shows an enhanced accuracy over individual models in all model metrics (Table 2). The performance of the stacked model is ~1% better than the best from the individual models (KNN and XGB). Although this may not be impressive, this one extra percent in the Kaggle competitions may tilt the balance and secure victory. In a seismic reservoir characterization study, models with the best predictive accuracy may potentially reduce the uncertainty that is inherently connected with predictions over millions of observations between the wells and for which there is no labeling data.

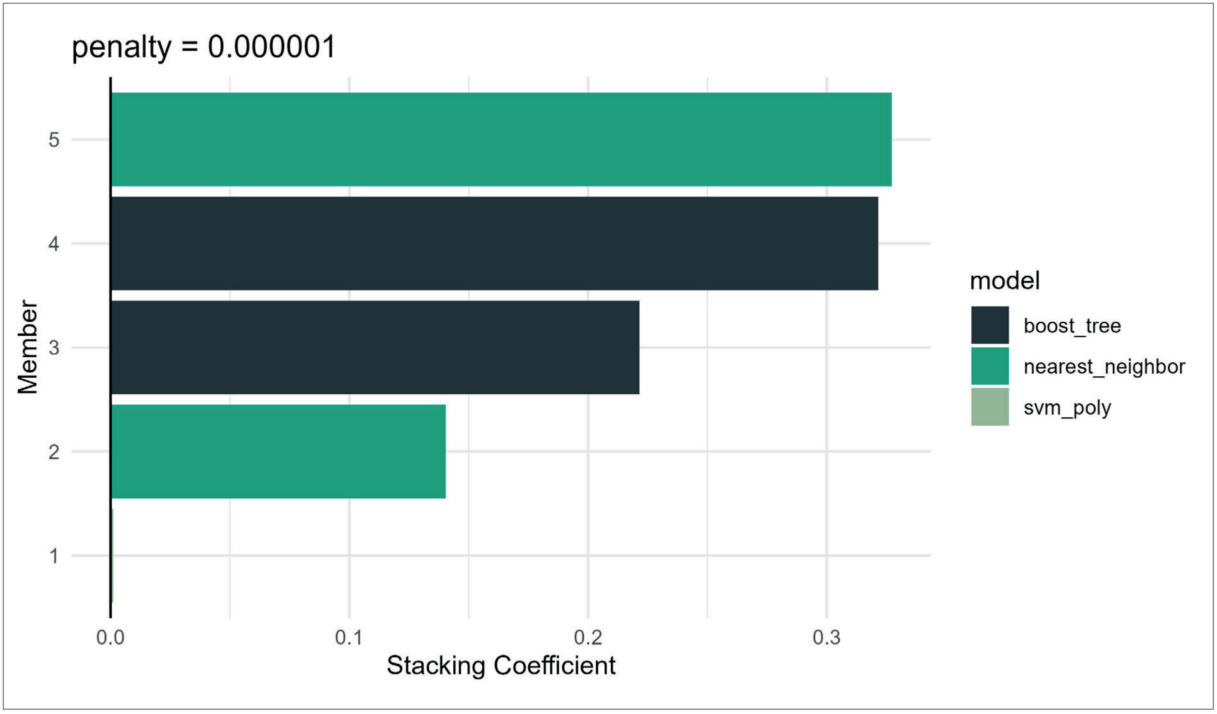

Fig. 5. *Model stacking coefficients for selected candidates*

Table 2
Model metrics for the stack model, selected member candidates, and the NN model from commercial software

Model	R2	RMSE	MAE
.pred (ensemble model)	0.890	0.0252	0.0168
KNN (No 9)	0.884	0.0259	0.0171
KNN (No 4)	0.880	0.0263	0.0170
XGB (No 2)	0.876	0.0268	0.0187
XGB (No 14)	0.875	0.0271	0.0174
SVM PCA (No 12)	0.436	0.0579	0.0418
NN (from Petrel)	0.318	0.0628	0.0487

The most important result, however, is a comparison between the ensemble model and the one obtained from commercial software using the neural net – NN (Fig. 6, Table 2). The prediction accuracy of NN (with no data pre-processing) is extremely low at the level of this obtained from MLP with Yeo–Johnson transformation (MLP YJ). Thus the direct application of neural nets results, in this case, is not recommended. To date, this drawback has been overcome by treating NN results as a secondary variable in the co-kriging form of deterministic or stochastic algorithms in which the relationship between primary (modeled) and secondary data is expressed with the Pearson correlation coefficient (in this case, equal to 0.69).

The spatial distribution of porosity from the model stacking is the last verification step. The obtained results should be consistent (or preferably outperform) with the previous research works in the studied area. This proved its validity in the processes of field appraisal and dynamic model calibration and simulation confirmed in the filed production output (Jędrzejowska-Tyczkowska 2003, Malaga et al. 2006, Papiernik et al. 2009). In most cases for the studied area, zones with high porosity values represent exploration fields with high hydrocarbon production. Figure 7A shows that high-porosity zones from the ensemble model are connected to production wells. A cross-section through the example areas of a carbonate platform, barriers, and slope structures revealed another distinguishing feature of the studied area – that zones with the best reservoir properties are located on the western slope of the carbonate platform (Fig. 7B). These sediments are linked to the debris facies (Mikołajewski & Wróbel 2005, Jaworowski & Mikołajewski 2007). Although the derived seismic facies did not capture this feature, the results from the ensemble model clearly showed enhanced porosities in these areas.

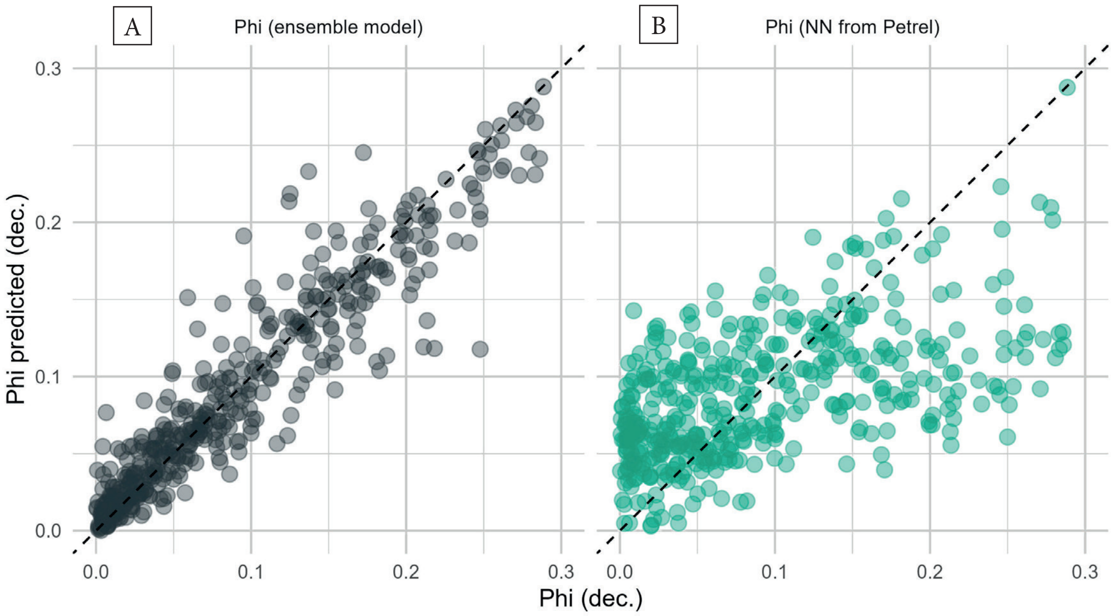

Fig. 6. *Comparison between Phi modeling results of the ensemble model (A) and NN (Petrel) (B)*

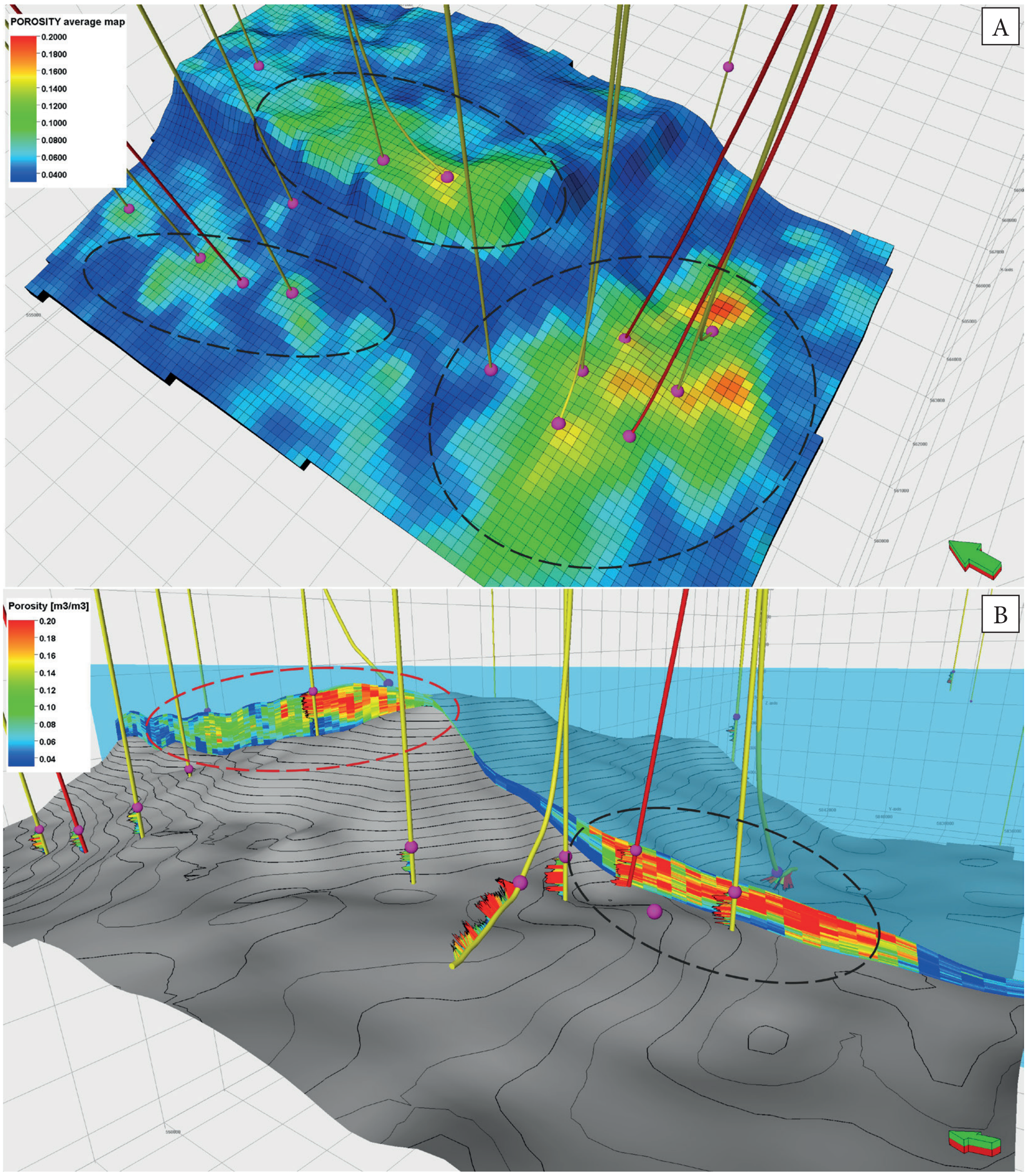

Fig. 7. *Graphic presentation of the results of stacking model implementation into the 3D geological model: A) average map of porosity distribution predicted from ensemble model; an outlines of confirmed oil and gas accumulations (expressed by increased porosity values – marked with ellipsoidal shapes) are apparent, which are additionally validated by production yield; B) cross-section through the carbonate platform with porosity distribution from ensemble model revealing high convergence with porosity values interpreted along boreholes; the barrier between oil accumulation (marked with a black ellipse) and gas accumulation (marked with a red ellipse) resulting from reduced porosity values is clearly visible*

CONCLUSIONS

The study presents an advanced ensemble modeling approach for predicting carbonate reservoir porosity from seismic attributes. The approach combined complex data pre-processing (data filtration, transformation, feature engineering) and tuning strategy with the race ANOVA method to create eight efficient models: KNN, RF, RF RT (RT with a class variable representing seismic facies), XGB, and MLP and SVM with two configurations each – with 5 PCA components, and with variable transform using Yeo–Johnson method. The model candidates were evaluated based on standard ML metrics – R2, RMSE, and MAE, and the role of each model configuration was assessed.

The results of the testing set revealed that the models with the highest predictive power are KNN, XGB, and RF, for which the MAE varies from 0.0173 to 0.0202. These results reflect the error of porosity prediction on the level of ~1.7–2.0%. The accuracy of the RF RT model did not improve significantly, with an extra class feature neglecting the role of seismic facies variable from *k*-means for porosity prediction in the studied carbonate reservoir. The PCA and Yeo–Johnson transformation significantly impacted MLP and SVM models, making the SVM YJ the best of the tested MLP and SVM models.

The model stacking was performed with the eight model definitions and 14 most prospective candidate members, from which five models were selected: two KNN, two XGB, and one SVM PCA. The latter one had a minor impact on a final ensemble model accuracy. Although not significant, the performance of the ensemble model showed superior accuracy over the best single learners, with all metrics (R2 0.890, RMSE 0.0252, MAE 0.168). The predicting power of an ensemble model best reflects the direct comparison with the results obtained with NN from commercial software. This comparison showed that the ensemble model is almost three times better than NN (R2 0.318, RMSE 0.0628, MAE 0.0487).

The spatial distribution of ensemble porosity results consisted of an additional assessment of model correctness. The obtained spatial distribution follows the one observed from the previous appraisal and exploration works, where high porosity zones occur in the slopes of the carbonate platform.

Finally, the results showed that the proposed advanced ML ensemble approach is promising for the better prediction of porosity in heterogeneous carbonate reservoirs from multi-seismic attributes.

The study was financed based on the statutory work entitled "Hybrydowe techniki uczenia maszynowego (meta-learning) w prognozowaniu parametrów petrofizycznych złoża" – the work of the Oil and Gas Institute – National Research Institute commissioned by the Ministry of Education and Science; order number 0048/SG/2022, archive number DK-4100-36/22.

REFERENCES

Abirami S. & Chitra P., 2020. Energy-efficient edge based real-time healthcare support system. [in:] Pethuru R. & Preetha E. (eds.), *The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases*, Advances in Computers, 117, Elsevier, 339–368. https://doi.org/10.1016/bs.adcom.2019.09.007.

Adeniran A.A., Adebayo A.R., Salami H.O., Yahaya M.O. & Abdulraheem A., 2019. A competitive ensemble model for permeability prediction in heterogeneous oil and gas reservoirs. *Applied Computing and Geosciences*, 14, 2070. https://doi.org/10.1016/j.acags.2019.100004.

Al-Amri M., Mahmoud M., Elkatatny S., Al-Yousef H. & Al-Ghamdi T., 2017. Integrated petrophysical and reservoir characterization workflow to enhance permeability and water saturation prediction. *Journal of African Earth Sciences*, 131, 105–116. https://doi.org/10.1016/j.jafrearsci.2017.04.014.

Anifowose F., Labadin J. & Abdulraheem A., 2013. Ensemble learning model for petroleum reservoir characterization: a case of feed-forward back-propagation neural networks. [in:] Li J., Cao L., Wang C., Tan K.C., Liu B., Pei J. & Tseng V.S. (eds.), *Trends and Applications in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science,7867, Springer, Berlin, Heidelberg, 71–82. https://doi.org/10.1007/978-3-642-40319-4_7.

Anifowose F., Labadin J. & Abdulraheem A., 2015. Improving the prediction of petroleum reservoir characterization with a stacked generalization ensemble model of support vector machines. *Applied Soft Computing*, 26, 483–496. https://doi.org/10.1016/j.asoc.2014.10.017.

Bedi J. & Toshniwal D., 2019. PP-NFR: an improved hybrid learning approach for porosity prediction from seismic attributes using non-linear feature reduction. *Journal of Applied Geophysics*, 166, 22–32. https://doi.org/10.1016/j.jappgeo.2019.04.015.

Boehmke B. & Greenwell B., 2020. *Hands-On Machine Learning with R*. The R Series, Chapman and Hall/CRC Press. https://bradleyboehmke.github.io/HOML/ [access: 1.09.2023].

Breiman L., 1996a. Bagging predictors. *Machine Learning*, 24(2), 123–40. https://doi.org/10.1007/BF00058655.

Breiman L., 1996b. Stacked regressions. *Machine Learning*, 24(1), 49–64. https://doi.org/10.1007/BF00117832.

Breiman L., 2001a. Random Forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Carvalho R., Gonzalez M.G. & Lupinacci E.W., 2022. Characterizing seismic facies in a carbonate reservoir, using machine learning offshore Brazil. *World Oil*. https://www.worldoil.com/magazine/2022/june-2022/features/characterizing-seismic-facies-in-a-carbonate-reservoir-using-machine-learning-offshore-brazil/ [access: 1.09.2023].

Chen C.H. & Lin Z.S., 2006. A committee machine with empirical formulas for permeability prediction. *Computers & Geosciences*, 32(4), 485–496. https://doi.org/10.1016/j.cageo.2005.08.003.

Chen T. & Guestrin C., 2016. XGBoost: A Scalable Tree Boosting System. [in:] *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data*, ACM, New York, 785–794. https://doi.org/10.1145/2939672.2939785.

Chen Y., Zhao L., Pan J., Li C., Xu M., Li K., Zhang F. & Geng J., 2021. Deep carbonate reservoir characterisation using multi-seismic attributes via machine learning with physical constraints. *Journal of Geophysics and Engineering*, 18(5), 761–775. https://doi.org/10.1093/jge/gxab049.

Chopra S., Lubo-Robles D. & Marfurt K., 2018. Some machine learning applications in seismic interpretation. *AAPG Explorer*, Search and Discovery Article, 42270.

Couch A. & Kuhn J., 2022. stacks: Stacked Ensemble Modeling with Tidy Data Principles. *Journal of Open Source Software*, 7(75), 4471. https://doi.org/10.21105/joss.04471.

Dramsch J.S., 2020. 70 Years of Machine Learning in Geoscience in Review. [in:] Moseley B. & Krischer L. (eds.), *Machine Learning in Geosciences*, Advances in Geophysics, 61, Elsevier, 1–55. https://doi.org/10.1016/bs.agph.2020.08.002.

Ferreira D.J., Dias R.M. & Lupinacci W.M., 2021. Seismic pattern classification integrated with permeability-porosity evaluation for reservoir characterization of presalt carbonates in the Buzios Field, Brazil. *Journal of Petroleum Science and Engineering*, 201, 108441. https://doi.org/10.1016/j.petrol.2021.108441.

Freund Y. & Schapire R., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119–139. https://doi.org/10.1006/jcss.1997.1504.

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

Hartigan J.A. & Wong M.A., 1979. K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), 100–108. https://doi.org/10.2307/2346830.

Hastie T., Tibshirani R. & Wainwright M., 2015. *Statistical Learning with Sparsity*. Monographs on Statistics and Applied Probability, 143, Chapman & Hall/CRC Press.

Helmy T., Rahman S., Hossain M.I. & Abdelraheem A., 2013. Non-linear heterogeneous ensemble model for permeability prediction of oil reservoirs. *Arabian Journal for Science and Engineering*, 38, 1379–1395. https://doi.org/10.1007/s13369-013-0588-z.

Hendry J., Burgess P., Hunt D., Janson X. & Zampetti V. (eds.), 2021. *Seismic Characterization of Carbonate Platforms and Reservoirs*. Geological Society, London, Special Publications, 509. https://doi.org/10.1144/SP509-2021-51.

Hou J., Zhao L., Zeng X., Zhao W., Chen Y., Li J., Wang S. et al., 2022. Characterization and evaluation of carbonate reservoir pore structure based on machine learning. *Energies*, 15(9), 7126. https://doi.org/10.3390/ en15197126.

James G., Witten D., Hastie T. & Tibshirani R., 2013. *An Introduction to Statistical Learning with Applications in R*. Springer Texts in Statistic, 103, Springer, New York. https://doi.org/10.1007/978-1-4614-7138-7.

Jaworowski K & Wróblewski Z., 2007. Oil- and gas-bearing sediments of the Main Dolomite (Ca2) in the Międzychód region: A depositional model and the problem of the boundary between the second and third depositional sequences in the Polish Zechstein Basin. *Przegląd Geologiczny*, 55(12/1), 1017–1024.

Jędrzejowska-Tyczkowska H., 2003. *Sejsmicznie konsystentne estymatory złoża węglowodorów*. Prace Instytutu Górnictwa Naftowego i Gazownictwa, 123, IGNiG, Kraków.

Jesus C., Azul, M.O, Lupinacci W.M. & Machado L., 2019. Multiattribute framework analysis for the identifcation of carbonate mounds in the Brazilian presalt zone. *Interpretation*, 7(2), T467–T476. https://doi.org/10.1190/INT-2018-0004.1.

Jolliffe I.T. (ed.), 2010. *Principal Component Analysis*. 2nd ed. Springer Series in Statistics, Springer, New York.

Kotarba M. & Wagner R., 2007. Generation potential of the Zechstein Main Dolomite (Ca2) carbonates in the Gorzów Wielkopolski – Międzychód – Lubiatów area: Geological and geochemical approach to microbial-algal source rock. *Przegląd Geologiczny*, 55(12/1), 1025–1036.

Kuhn M., 2014. *Futility analysis in the cross-validation of machine learning models*. https://doi.org/10.48550/arXiv.1405.6974.

Kuhn M. & Johnson K., 2013. *Applied Predictive Modeling*. Springer, New York. https://doi.org/10.1007/978-1-4614-6849-3.

Kuhn M. & Silge J., 2022. *Tidy Modeling with R: A Framework for Modeling in the Tidyverse*. O'Reilly Media. https://www.tmwr.org/ [access: 1.09.2023].

Lance G.N. & Williams W.T., 1966. A general theory of classifactory sorting strategies, I. Hierarchical systems. *The Computer Journal*, 9(4), 373–380. https://doi.org/10.1093/comjnl/9.4.373.

Malaga M., Solarski T. & Wolnowski T., 2006. Modelowanie geostatystyczne dolomitu głównego w rejonie Międzychód – Sieraków. *Prace Instytutu Nafty i Gazu*, 137, 1067–1071.

Mikołajewski Z. & Wróbel M., 2005. Petrografia i diageneza utworów cechsztyńskiego dolomitu głównego (Ca2) w rejonie złoża ropy naftowej Lubiatów (Polska Zachodnia). *Przegląd Geologiczny*, 53(4), 335–336.

Molnar C., 2019. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book [access: 1.09.2023].

Moore C.H. & Wade W.J., 1989. The Classification and Nature of Carbonate Porosity. [in:] Moore C.H. (ed.), *Carbonate Diagenesis and Porosity*, Developments in Sedimentology, 46, Elsevier, 21–41. https://doi.org/10.1016/S0070-4571(08)71056-2.

Murtagh F. & Legendre P., 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification*, 31, 274–295. https://doi.org/10.1007/s00357-014-9161-z.

Otchere D.A., Gana, T.O.A., Gholami R. & Lawal M., 2021a. A novel custom ensemble learning model for an improved reservoir permeability and water saturation prediction. *Journal of Natural Gas Science and Engineering*, 91, 103962. https://doi.org/10.1016/j.jngse.2021.103962.

Otchere D., Arbi A., Ganat T.O., Gholami R. & Ridha S., 2021b. 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. https://doi.org/10.1016/j.petrol.2020.108182.

Papiernik B., Machowski G., Słupczyński K. & Semyrka R., 2009. Geologiczny model rejonu akumulacji ropnogazowej Lubiatów-Międzychód Grotów. *Geologia: kwartalnik Akademii Górniczo-Hutniczej im. Stanisława Staszica w Krakowie*, 35(2/1), 175–182.

Pattnaik S., Chen S., Helba A. & Ma S., 2020. *Automated carbonate rock facies identification with deep learning*. Paper presented at the SPE Annual Technical Conference and Exhibition, 26–29 October 2020, Virtual, SPE-201673. https://doi.org/10.2118/201673-MS.

R Core Team, 2022. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Raheem M.W. & Skuker A.K., 2021. Prediction by reservoir porosity using micro-seismic attribute analysis by machine learning algorithms in an Iraqi Oil Field. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(14), 3324–3332.

Randen T. & Sønneland L., 2005. Atlas of 3D Seismic Attributes. [in:] Iske A. & Randen T. (eds.), *Mathematical Methods and Modelling in Hydrocarbon Exploration and Production*, Mathematics in Industry, 7, Springer, Berlin, Heidelberg, 23–46. https://doi.org/10.1007/3-540-26493-0_2

Reza S., Ali K., Behrouz R., Mohammad Y. & Saeed K., 2011. A committee machine approach for predicting permeability from well log data: A case study from a heterogeneous carbonate reservoir, Balal oil field, Persian Gulf. *Journal of Geopercia*, 1(2), 1–10. https://doi.org/10.22059/jgeope.2011.23279.

Sinaga T., Mohammad R. & Haidar, M., 2019. Porosity prediction using neural network based on seismic inversion and seismic attributes. *E3S Web of Conference*, 125, 15006. https://doi.org/10.1051/e3sconf/201912515006.

Tiab D. & Donaldson E.C., 2015. *Petrophysics Theory and Practice of Measuring Reservoir Rock and Fluid Transport Properties*. 4rd ed. Gulf Professional Publishing, Elsevier.

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. https://doi.org/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. https://doi.org/10.18668/NG.2021.05.01.

Topór T. & Sowiżdżał K., 2022. Application of machine learning tools for seismic reservoir characterization study of porosity and saturation type. *Nafta-Gaz*, 78(3), 165–175.

Ward J.H., Jr., 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845.

Wardhana S.G. & Pratama H., 2021. Leveraging Artificial Neural Network, Support Vector Regression, and K-Nearest Neighbors for Porosity Prediction in X Well: An Integrated Petrophysics and Machine Learning. [in:] *Proceedings of Joint Convention Bandung (JCB), November 23rd – 25th 2021*, Ikatan Ahli Teknik Perminyakan Indonesia, Jakarta, 744–747.

Wolpert D., 1992. Stacked generalization. *Neural Networks*, 5(2), 241–59. https://doi.org/10.1016/S0893-6080(05)80023-1.

Yeo I.-K. & Johnson R.A., 2000. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4), 954–959. http://www.jstor.org/stable/2673623.

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(1), 119–139. https://doi.org/10.1006/jcss.1997.1504.

APPENDIX 1

Description of the individual seismic attributes used in the study

Abbreviations name of attribute	Name of attribute	Description
Grad_Mag	Gradient magnitude	The magnitude of the instantaneous gradient computed in three-dimensions of the sample neighborhood
Refl_Int	Reflection intensity	Reflection intensity is the average amplitude over a specified window (default 9 samples) multiplied by the sample interval
RMS_Amp	RMS amplitude	RMS Amplitude computes root mean squares on instantaneous trace samples over a specified window
Envelope	Envelope	The total instantaneous energy of the analytic signal (the complex trace), independent of phase
Sweetness	Sweetness	Sweetness is the implementation of two combined attributes (envelope and instantaneous frequency) and is used for the identification of features where the overall energy signatures change in the seismic data
Attenuation	t* attenuation	The differential loss of high frequencies relative to low frequencies as measured above and below the point of interest. t* attenuation is a patented seismic attribute for indicating open fractures within the seismic volume based on windowed frequency attenuation
Dom_Freq	Dominant frequency	An attribute is calculated as the hypotenuse between instantaneous frequency and instantaneous bandwidth. This attribute, in combination with Instantaneous bandwidth, serves as a supplement to the Instantaneous frequency, as the three attributes reveal the time-varying spectral properties of seismic data
Inst_Freq	Instantaneous frequency	An attribute that helps to measure the cyclicity of geological intervals and can be helpful in cross-correlation across faults. The time derivative of phase, $w = d(phase)/dt$
Inst_Ph	Instantaneous phase	The instantaneous phase is a good indicator of continuities, faults, pinch-outs, bed interfaces, sequence boundaries, and regions of on-lap patterns – the argument of the analytic signal, $phase = \text{arctg}(g/f)$. The attribute is calculated on a sample-by-sample basis without regard for the waveform
LR	Lambda*Rho	An attribute obtained from AVO inversion using moduli and density relationships to impedance. Lambda-Lamé parameter of incompressibility; Rho-density
DN	Bulk density	Rock bulk density estimated from simultaneous seismic inversion procedure
Vp_Vs	P-wave velocity	The velocity of P-wave estimated with simultaneous seismic inversion procedure
Zp	P-wave impedance	The acoustic impedance of P-wave derived from simultaneous seismic inversion
Rel_AI	Relative acoustic impedance	Relative acoustic impedance is a running sum of regularly sampled amplitude values. It is calculated by integrating the seismic trace, passing the result through a high-pass Butterworth filter to reduce potentially introduced low-frequency noise
Vs	S-wave velocity	The velocity of S-wave estimated with simultaneous seismic inversion procedure
Zs	S-wave impedance	The acoustic impedance of S-wave derived from simultaneous seismic inversion
MR	Mu*Rho	An attribute obtained from AVO inversion using moduli and density relationships to impedance. Mu-Lamé parameter of rigidity; Rho-density
App_Polar	Apparent polarity	The polarity of the instantaneous phase calculated at the local amplitude extreme
COP	Cosine of phase	The cosine of the instantaneous phase, also known as normalized amplitude; helps to enhance the definition of structural delineations. Used together with instantaneous phase for comparison
Org_Amp	Original amplitude	The real part of the analytical signal $f(t)$

Abbreviations name of attribute	Name of attribute	Description
Trace_AGC	Trace AGC (automatic gain control)	The trace AGC (iterative) volume attribute automatically scales the instantaneous amplitude samples with the local root mean square (RMS) amplitude level, computed over a user-specified vertical window, and has the option to apply multiple RMS iterations, in order to get more well-behaved (smooth) scaling function
Iso_Freq_15	Iso-frequency component (15 Hz)	The contribution of individual frequencies (here 15 Hz) to the make-up of the input seismic signal. The desired frequency is the isolated frequency component to extract from the input seismic volume
Cons_Dip	Consistent dip	Accurate volumetric dip estimation
Variance	Variance	The estimation of local variance in the signal
Dip_Dev	Dip deviation	Tracking of the rapid changes in the dip orientation field. The difference between the dip trend and the instantaneous dip
Iso_Freq_30	Iso-frequency component (30 Hz)	The contribution of individual frequencies (here 30 Hz) to the make-up of the input seismic signal. The desired frequency is the isolated frequency component to extract from the input seismic volume
Chaos	Chaos	The chaotic signal pattern contained within seismic data is a measure of the lack of organization in the dip and azimuth estimation method
Loc_Flat	Local flatness	The variance of the orientation field to identify the uniformity of the signal within the orientation estimation range
Vp_Vs	Vp/Vs	P-wave and S-wave velocity ratio
Phi	Porosity	Porosity derived from well-log interpretation
Iso_Freq_45	Iso-frequency component (45 Hz)	The contribution of individual frequencies (here 45 Hz) to the make-up of the input seismic signal. The desired frequency is the isolated frequency component to extract from the input seismic volume