Capillarity

Vol. 5, No. 5, p. 91-104, 2022

Original article

Prediction of permeability of tight sandstones from mercury injection capillary pressure tests assisted by a machine-learning approach

Jassem Abbasi[1]*, Jiuyu Zhao[2], Sameer Ahmed[1], Liang Jiao[3], Pål Østebø Andersen[1,4], Jianchao Cai[2]

[1]*Department of Energy Resources, University of Stavanger, Stavanger 4036, Norway*

[2]*State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum, Beijing 102249, P. R. China*

[3]*School of Geophysics and Geomatics, China University of Geosciences, Wuhan 430074, P. R. China*

[4]*The National IOR Centre of Norway, University of Stavanger, Stavanger 4036, Norway*

Keywords:
Support vector machine
MICP test
capillary pressure
permeability

Cited as:
Abbasi, J., Zhao, J., Ahmed, S., Jiao, L., Andersen, P., Cai, J. Prediction of permeability of tight sandstones from mercury injection capillary pressure tests assisted by a machine-learning approach. Capillarity, 2022, 5(5): 91-104.
https://doi.org/10.46690/capi.2022.05.02

Abstract:
Mercury injection capillary pressure analysis is a methodology for determining different petrophysical properties, including bulk density, porosity, and pore throat distribution. In this work, distinct parameters derived from mercury injection capillary pressure tests was considered for the prediction of permeability by coupling machine learning and theoretical approaches in a dataset composed of 246 tight sandstone samples. After quality checking the dataset, the feature selection was carried out by correlation analysis of different theoretical permeability models and statistical parameters with the measured permeability. Finally, porosity, median capillary pressure, Winland model, and mean pore-throat radius (corresponding to the saturation range 0.4-0.8) were chosen as the input features of the machine learning model. As the machine learning approach, a support vector machine (SVM) model with a radial basis function kernel was proposed. Furthermore, the model and its metaparameters were trained with a particle swarm optimization (PSO) algorithm to avoid over-fitting or under-fitting. In contradiction to the theoretical models, the implemented SVM-PSO model could acceptably predict the experimentally measured permeability values with an R^2 rate of over 0.88 for training and testing datasets. The introduced approach could reduce the mean relative errors from about 10 to values less than 0.45. The improvements were more significant for low permeability samples. This successful implementation shows the potential of coupled usage of theoretical and machine learning methodologies for improved prediction of permeability of tight sandstone rocks.

1. Introduction

Multiphase flow in porous media happens in many artificial and natural processes, such as nonaqueous phase liquids transport, CO_2 storage, enhanced oil recovery (Ahmad et al., 2016; Blunt, 2017). The relative distribution and movement of wetting and non-wetting phases in porous media are highly influenced by the capillary behavior of phases and controlled by two critical parameters. In classic definitions, these are the capillary pressure (P_c) and the relative permeability of the phases, both considered to be functions of the saturation of the wetting phase (Lin et al., 2018). Also, the efficiency of operations in oil recovery or geological storage is highly dependent on the permeability of the rock, and the correct prediction of its distribution in areal and lateral directions affects decisions and technical solutions. The rock permeability is directly dependent on the geometrical attributes of rocks, such as porosity, pore-size distribution, and pore network coordination number (Menke et al., 2021).

Yandy Scientific Press

*Corresponding author.
E-mail address: jassem.abbasi@uis.no (J. Abbasi); 2020310039@student.cup.edu.cn (J. Zhao); sameerahmed442@gmail.com (S. Ahmed); jiaoliang@cug.edu.cn (L. Jiao); pal.andersen@uis.no (P. Andersen); caijc@cup.edu.cn (J. Cai).

Received September 5, 2022; revised September 27, 2022; accepted October 10, 2022; available online October 15, 2022.

Capillary pressure is defined as the pressure difference across the interface between two immiscible phases. The wetting condition of the porous media, interfacial properties of the phases, and the pore geometry of the rock are the determining parameters in the capillary behavior of porous media. The porous plate method, centrifuge method, and mercury injection capillary pressure (MICP) method are the three most routine approaches for capillary pressure function measurement in geological rocks (Abbasi and Andersen, 2022). In MICP tests, the mercury as a non-wetting phase is injected into the core samples up to high pressures (to 4,000 bars). The porosity can be calculated from the total volume of mercury injected at the maximum pressure (McPhee et al., 2015). Also, the saturation at each pressure stage is determined as the volume fraction of mercury that has entered the rock. The P_c at each stage is equal to the mercury injection pressure. This test also gives insightful information related to the pore structure of the rock (Jiao et al., 2020). However, the procedure leads to permanent loss of the core samples due to retaining mercury in the pores after withdrawal. The measured P_c at each saturation stage is related to effective pore throat size by the Young-Laplace equation:

$$P_c = \frac{2\sigma\cos\theta}{r} \quad (1)$$

where σ is mercury/air interfacial tension, θ is the contact angle, and r is the pore throat radius.

The permeability of cores is routinely measured in single-phase liquid or gas injection tests. Also, due to the high measurement time of tight samples, the pressure decay method may be used (Jones, 1997). However, for decades, researchers have tried to extract permeability from other measurements, such as P_c tests, especially MICP tests. Purcell (1949) provided an analytical relation for the prediction of permeability from porosity and MICP curve properties, such as the fraction of volume occupied by mercury and capillary pressure, but assumed the porous medium was a bundle of tubes and had to apply a correction factor. Swanson (1981) defined the Swanson parameter, which is the maximum value of the curve of mercury saturation divided by P_c (S_{Hg}/P_c) plotted versus mercury saturation (S_{Hg}), for the prediction of permeability from Pc data for clean sandstones and carbonates, separately. However, Xiao et al. (2014) concluded that the Swanson equation is not successful in tight sandstones due to the ambiguous Swanson parameter values. Xiao et al. (2017) showed that in homogeneous sandstone rocks, the average pore throat radius can be calculated to find accurate permeability predictions. By analysis of tight gas sands, Rezaee et al. (2012) found that the dominant pore throat radius is in the mercury saturation of 10% and that this point correlated with permeability.

In recent years, the application of machine learning (ML) in geosciences has been growing rapidly due to the large volume of available data that needs processing and analysis (Karpatne et al., 2019). Several applications of ML have been found in the rock-fluid properties (Hébert et al., 2020), and upscaling of porosity-permeability calculations from pore-scale images (Menke et al., 2021). Due to the importance of permeability, many of the studies are focused on the prediction of permeability from different data sources, including special core analysis tests (Erofeev et al., 2019), wireline logs (Zhang et al., 2021), and nuclear magnetic resonance tests (Zhang et al., 2017). Feng et al. (2020) applied a support vector machine (SVM) algorithm to 22 sets of MICP data and found that it is superior to currently available permeability models. However, it was found that previous studies rarely considered tight sandstone rocks. Tight sandstones are known as highly heterogeneous with a mixture of pore types, including primary intergranular pores, secondary dissolution pores, and fractures (Zhao et al., 2022). This complex structure of tight sandstones brings uncertainties in the predictions of permeabilities using regular models. However, ML models have shown potential for the prediction of complex phenomena in various applications. In this work, it is tried to analyze the validity of previously developed MICP permeability models for tight sandstone rocks using a dataset much larger than the previous works, and then develop an ML model for a more accurate prediction of permeability. It is focused on synergizing the capabilities of the ML model and the previously developed theoretical models. Accordingly, at first, the validity of current permeability models (i.e., Swanson, Purcell, Parachor, Fractal, and Winland models) is investigated over a large set of MICP tests related to tight sandstones, which are gathered from all over the world. Then, the pore-scale characteristics of the Pc curves that are most correlated to the rock permeability is determined. Then, by mixing the most correlated theoretical and statistical features, it is focused on the deployment of ML-aided models for the estimation of tight sandstone permeabilities from MICP tests.

In the following, the theory behind the deployed models is introduced, and the details of the developed ML model are provided. Then, the results obtained in the process of feature engineering and the model training and testing are provided. The work ends with a conclusion.

2. Methodology

The methodology used in this work is based on an ML algorithm coupled with an optimization tool. The main workflow is shown in Fig. 1. The work started with data-gathering from industrial and literature data. Afterward, data cleaning and anomaly analysis were carried out. Then, by analytical and statistical analysis of possible relevant ML input features, the most suitable parameters for the prediction of rock permeabilities are investigated, and features with the best results are selected. Finally, the ML model is trained, and the results are validated against the testing dataset. More discussions about the methodology are provided in the coming sections.

2.1 Data gathering

In this work, a dataset with an overall of 248 samples related to tight sandstone rocks was gathered. Each sample included the measurements of porosity, permeability, and MICP curves. Some of the data was gathered from the Ordos Basin, located at the junction of the eastern tectonic domain and the western tectonic domain, which is expected to be

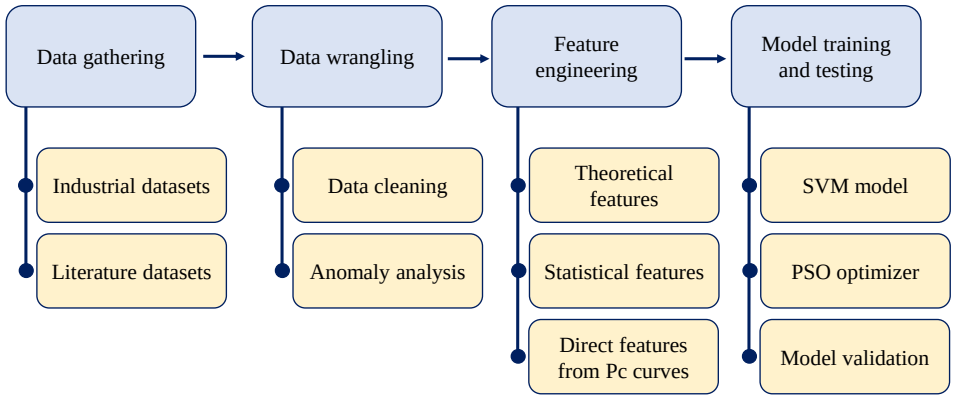

Fig. 1. The flowchart introducing the workflow of the development of the ML model for the permeability prediction of tight sandstones.

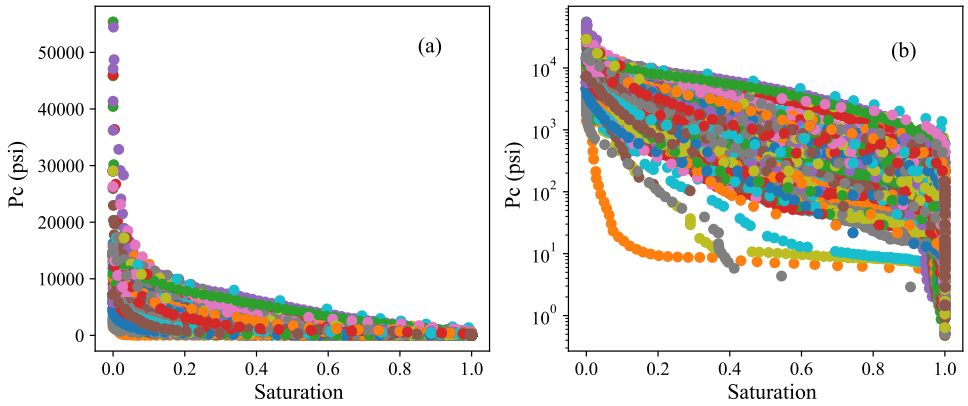

Fig. 2. The mercury injection Pc curves for all the MICP samples in (a) cartesian scale and (b) Semi-log scale.

the second-largest petroliferous basin in China (Wang and Wang, 2013). The study area is in the Yishaan Slope of the Ordos Basin, which is the main area of tight (sandstone) oil exploration and development in China. These tight cores were taken from different layers of more than 20 wells in the Yanchang Formation. A total of 172 cores were tested by Poremaster PM-33-13 for MICP data, and CMS-300 was used to obtain the porosity and permeability of these cores (Changtao et al., 2018; Fan et al., 2019). Also, 76 MICP samples were gathered from different sources in the literature (Rezaee et al., 2012; Eslami et al., 2013; Xiao et al., 2014, 2017; Tran et al., 2018; Wang et al., 2018, 2019; Arabjamaloei et al., 2019; Liu et al., 2020). Fig. 2 shows the MICP curves for all cases in both cartesian and logarithmic scales, where there are large variations in the range of Pc values. Also, the statistical distributions of porosity and permeability values of the tests are provided in Fig. 3. The porosities are in the range of 0.03-0.26, mainly in the range of 0.06-0.15. Also, permeabilities are in the range of 0.0005 to 4 mD, and mainly in the range of 0.008 to 0.1 mD, which is expectable for tight sandstones.

2.2 Permeability correlations

Many researchers have developed theoretical models for the prediction of rock permeability by interpretation of MICP tests. These models may be partially based on theoretical models or empirical datasets. In this section, the most prominent correlations utilized in our work are introduced.

2.2.1 Swanson permeability

To provide a method for the prediction of rock brine permeability from capillary measurements, Swanson (1981) provided a correlation by introducing the Swanson parameter, which equals the maximum point of the curve when (S_{Hg}/P_c), is plotted versus S_{Hg}. This point is closely related to the condition in which the non-wetting phase partially fills the effective pore volumes and has a determining role in controlling fluid flow in the rock pore system. For sandstone rocks, rock permeability K (m^2) is calculated by:

$$K = 0.015 \left(\frac{S_{Hg}}{P_c} \right)^{2.109}_{\max} \tag{2}$$

2.2.2 Purcell equation

Purcell developed a semi-analytical equation to determine the relationship between the permeability of a porous medium and its porosity and P_c curve using a bundle-of-tubes assumption combined with Poiseuille's equation (Purcell, 1949):

$$K = (6.47e^{-4})\phi \int_{S_{Hg}=0}^{S_{Hg}=100} \frac{\mathrm{d}S_{Hg}}{P_c^2} \tag{3}$$

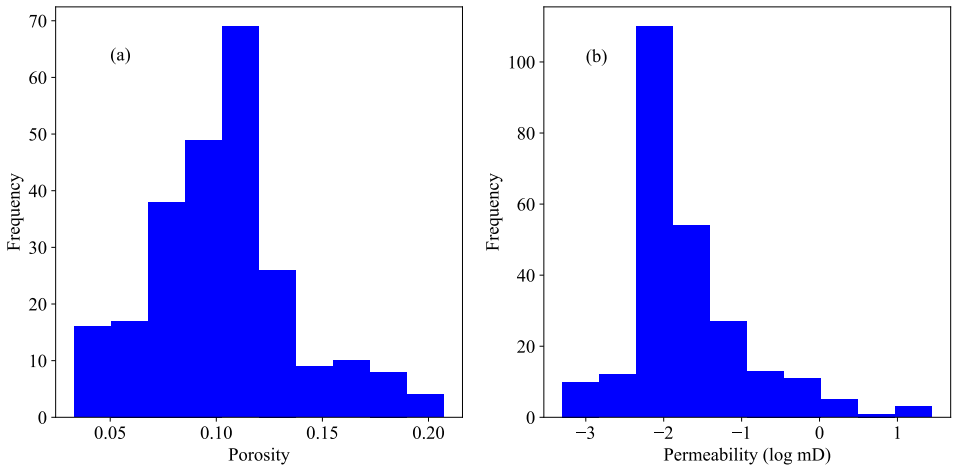

Fig. 3. Histogram distribution of (a) porosity and (b) permeability values. Porosity shows a normal distribution, while permeability shows a log-normal distribution.

where ϕ is the porosity.

The equation was verified over 26 MICP tests from sandstone rocks and showed that the permeability can be well predicted using the porosity and P_c curves.

2.2.3 Parachor equation

By extending the Swanson model, Guo et al. (2004) introduced a correlation that best fitted their MICP data by deriving the capillary pressure Parachor parameter, $(S_{Hg}/P_c^2)_{\max}$, that is defined as the maximum value of the S_{Hg}/P_c^2 curve plotted versus mercury saturation. The Parachor permeability model is defined as:

$$K = (5.29e^{-7})\left(\frac{S_{Hg}}{P_c^2}\right)_{\max} \tag{4}$$

2.2.4 Winland equation

Winland presented an empirical correlation that relates the average pore radius and porosity to the rock air permeability (Kolodzie, 1980). After regression with different parameters, he found that the radius corresponding to the mercury saturation of 35% had the best correlation with the permeability. However, the optimum corresponding saturation may be different in various cases. In this work, the reference correlation is used for the analysis of the permeability measurements:

$$\log r_{35} = 7.82 + 0.588 \log K_{air} - 0.864 \log \phi \tag{5}$$

where K_{air} is the air permeability. In this equation, the pore radius corresponding to the saturation of $S_{Hg} = 0.35$ is calculated by:

$$r_{35} = \frac{2\sigma \cos\theta}{P_c(S_{Hg})} \tag{6}$$

2.2.5 Fractal analysis of pore structure

The fractal theory is widely used for the analysis of pore structures in sandstone rocks and was first proposed by Burn and Mandelbrot (1984). They found that the size distribution of the pores in sponges follows the power law. Actually, the cumulative distribution of pores with a size greater than or equal to λ has been confirmed to follow:

$$N(L \geq \lambda) = \left(\frac{\lambda_{\max}}{\lambda}\right)^{D_f} \tag{7}$$

where D_f is defined as the pore-size fractal dimension and $\lambda_{\max}$ is the maximum pore size. The fractal dimension represents the fractal specifications of pores and especially, it gives insightful information about the complexity of the pore network of the rocks and can be calculated using different approaches, such as scanning electron microscopy, thin section analysis, X-ray computed tomography scans and MICP tests. There are several methods for the calculation of the fractal dimension of rocks from MICP data, including two-dimensional (2D) capillary tube models, three-dimensional (3D) sphere models, thermodynamic models, and 3D capillary tube models (Ge et al., 2016). Wang et al. (2018) showed that the 3D capillary tube model is the most appropriate model for the prediction of rock properties and pore structures from MICP tests. In one of these models, Li (2010) proposed a relation between mercury saturation and the measured P_c:

$$S_{Hg} \propto P_c^{-(2-D_f)} \tag{8}$$

Considering this relationship, the fractal dimension of the 3D capillary tube model is calculated by plotting the log-log scale of S_{Hg} and P_c curve. Assuming the slope of the $\log S_{Hg} - \log P_c$ curve is m, the fractal dimension is defined as (Li, 2010):

$$D_f = m + 2 \tag{9}$$

Afterward, the specific surface area of the rock is calculated by:

$$S_p = \frac{3}{r_g}\frac{1-\phi}{\phi} \tag{10}$$

where r_g is the average grain radius. The permeability (m^2) is then calculated by:

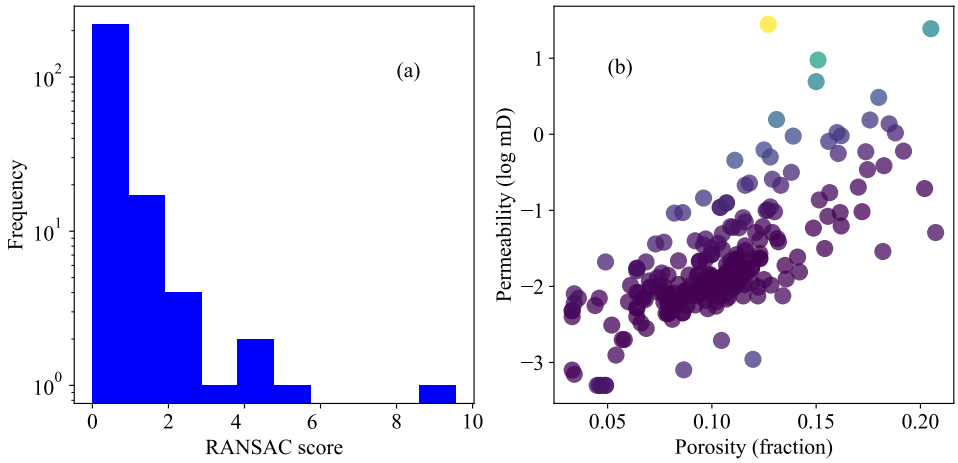

Fig. 4. The results of anomaly analysis applied to the dataset. (a) The histogram of RANSAC outlier analysis scores; higher magnitudes correspond to the outlier, (b) anomaly analysis where the colors show the outlier score of each point. Points with the lowest outlier score have darker colors, while the points with high RANSAC scores are colored in yellow.

$$K = 1.6\left(\frac{1-\phi}{S_p}\right)\left(\frac{0.952\phi^2}{1-\phi}\right)^{2/(D_f-1)} \tag{11}$$

2.3 Anomaly analysis

Outliers (anomalies) are defined as a group of the original samples in the dataset that show abnormal behavior in comparison to the majority of the population. These abnormalities can also be due to erroneous measurements/calculations or even the presence of rare cases in the population. There are several methods for the detection of outliers in the literature. In this work, a random sample consensus (RANSAC) method is used. This method follows an automatic non-deterministic algorithm and fits a model on random subsets of inliers from the complete data set (Choi et al., 2009). The advantage is that it can find the outliers in multidimensional data sets with a high level of accuracy even when a substantial fraction of anomalies is present in the dataset. The results of anomaly detection of the cores are shown in Fig. 4. In this process, the porosity-permeability data is applied for anomaly detection. Fig. 4(a) shows the histogram distribution of anomaly scores found by the RANSAC algorithm. The points with higher scores are more likely to be considered anomalies. In Fig. 4(b), the color of points is set based on the anomaly score provided by the RANSAC algorithm. It shows that there may be an abnormality in Porosity-Permeability scatter trend in cases with medium porosity but very high permeability. These cases can be due to the presence of non-reported fissures/fractures in the cores, or also can be due to measurement errors. However, it cannot be confidently considered a measurement error.

Since this work is intended to maintain the generality of the model, the suspected anomaly points are not removed from the training model. This will help to ensure that the developed model will be adequately general to be used in future predictions.

2.4 Kendal's Tau correlation coefficient

There are different models to quantify the dependency of two quantities, like Pearson, Spearman, and Kendal's Tau correlations. Different works showed the relative advantage of Kendal's Tau model in normally distributed datasets, and also its less sensitivity to discrepancies in the data (Croux and Dehon, 2010; Puth et al., 2015). In this work, Kendal's Tau correlation coefficient is applied, which is defined as the similarity of the ordering of the data when separately ranked based on each parameter (Kendall, 1976). This parameter is explicitly defined as below for any pair of (x_i, x_j) and (y_i, y_j):

$$\tau = \frac{2}{n(n-1)}\sum_{i<j}\mathrm{sgn}(x_i - x_j)\mathrm{sgn}(y_i - y_j) \tag{12}$$

where *sgn* is the sign function. Two parameters are completely uncorrelated if τ converges to zero and are directly or inversely correlated if τ equals 1 or -1, respectively.

2.5 Input features

The main part of the study is designing proper input features for the training of the SVM-PSO model. The most routine feature for the prediction of permeability is rock porosity. Fig. 4(b) shows the correlation between the porosity and permeability values. There is a linearly increasing trend in the permeability (log) versus porosity scatter plot. Kendal's tau coefficient is 0.54, showing that the porosity generally is a determining factor for the permeability prediction. So, the rock porosity is used as a relevant input feature.

Also, to utilize the previously developed theoretical models, different permeability models developed for permeability predictions from MICP tests was analyzed. These permeability models include Swanson, Purcell, Winland, Parachor, and Fractal models (the theories behind these equations are provided in the previous sections). The most relevant and correlated predictions will be used as input features of the ML model. This strategy helps in combining physics-based theories with data-based models to make use of their ad-

vantages. Also, generally, one of the drawbacks of the ML approaches is that they do not realize the physics, and in extreme cases, they can provide non-physical predictions, especially for out-of-training range predictions (Karniadakis et al., 2021). Hopefully, this coupling helps in reducing the extrapolation weaknesses of ML models.

Moreover, plus the previous features, two statistically calculated parameters are extracted from the P_c curves:

- **Average P_c** ($\overline{P}_c$): The arithmetic average value of P_c for each point in capillary pressure curves as:

$$\overline{P}_c = \frac{1}{n}\sum_{i=1}^{n}(P_c)_i \tag{13}$$

- **Median P_c** (MedPc): The P_c value lies in the middle of the P_c curve vector. To increase the accuracy of the median calculations, using a linear interpolation approach, the P_c points were recalculated from the experimental data with the same S_{Hg} interval of 0.02.

Also, in another part of the feature engineering process, it has been decided to directly import some parts of the P_c curve as the pore-throat radius. In the calculations, P_c points related to the different saturation ranges are extracted from the P_c curves. The extracted P_c values are then converted to the pore-throat radius using Eq. 6 and then their arithmetic average values are used as features. More information about the validity of these values and their correlation with rock permeability is provided in the next sections.

2.6 Support vector machine

The SVM algorithm was first suggested by Cortes and Vapnik (1995) and is a subset of supervised learning methods, and primarily introduced for classification (pattern recognition) purposes in projects like handwriting and face recognition. However, it was also successful in regression problems. In its simplest form, SVM uses a linear fitting hyperplane to regress on the dataset with a minimal error margin. Given the training dataset $(x_i, y_i)_{i=1}^{n}$ (n is the number of training samples), where x_i is the matrix of input variables (with dimension 1 by I, where I is the number of inlet variables), and y_i is the output variable, which in this work is the logarithm of permeability ($\log K_i$), $i = 1, \ldots, N$, the hyperplane equation in two-dimensional (x, y) space is defined as a subspace of dimension $n-1$:

$$f(x) = \vec{w}\cdot\vec{x} + b \tag{14}$$

where scalar b is defined as the offset of the regression line, and the vector w (1 by I) is called the weight vector and defines a direction perpendicular to the hyperplane. The prediction function above requires a small w. The regression parameters of the hyperplane function are calculated by minimizing the objective function:

$$\frac{1}{2}\omega^T\cdot\omega \tag{15}$$

Subject to the constraints:

Table 1. The meta parameters related to the PSO algorithm and SVM model.

Metaparameters	Value
Number of dimensions	2
Number of particles	10
C1	0.7
C2	0.7
Iterations	600
Kernel	RBF
Tolerance	1E-3
Objective function	$1-R^2$

$$\begin{cases}\vec{w}\cdot\vec{x}+b-\vec{y}\le\varepsilon\\ \vec{y}-\vec{w}\cdot\vec{x}-b\le\varepsilon\end{cases} \tag{16}$$

where ε is the distance within which no penalty is associated with the training loss function with points predicted within a distance epsilon from the actual value (see Table 1 for the parameters used in this work). In this equation, the error values less than ε should be ignored. To increase the generalization capability of the model, the relaxation variable (ζ) is introduced to include the errors associated with the points where the target error values exceed ε (see Fig. 5):

$$\frac{1}{2}\omega^T\cdot\omega + C\sum_{i=1}^{n}\zeta_i \tag{17}$$

Subject to the constraints:

$$\begin{cases}\vec{w}\cdot\vec{x}+b-\vec{y}\le\varepsilon+\zeta\\ \vec{y}-\vec{w}\cdot\vec{x}-b\le\varepsilon+\zeta\end{cases} \tag{18}$$

The constant $C > 0$ determines the trade-off between the flatness of the model (model complexity) and the amount to which prediction errors larger than ε are tolerated.

However, in complex classification or regression problems, SVM maps nonlinear regression problems from low-dimensional feature spaces into linear regression problems with higher-dimensional feature spaces. The mapping of parameters in space is achieved by using nonlinear transforming functions, which are called kernel functions. There are a variety of kernel functions, including radial basis function (RBF), polynomial functions, and gaussian functions. The RBF kernel function (1 by I) in SVM regression problems is defined as:

$$k(x_i, x_j) = \exp\left[-\gamma(x_i - x_j)^2\right] \tag{19}$$

The γ represents the distribution width in the kernel function that tunes the prediction accuracy. Also, x and x' are two different observations in the dataset. To solve this nonlinear problem, the main methodology is to construct a Lagrange function from the objective function, by introducing a set of variables (c), and the corresponding constraints that will be called the primal objective function. By applying the suitable kernel function in the SVM model, it is tried to maximize the

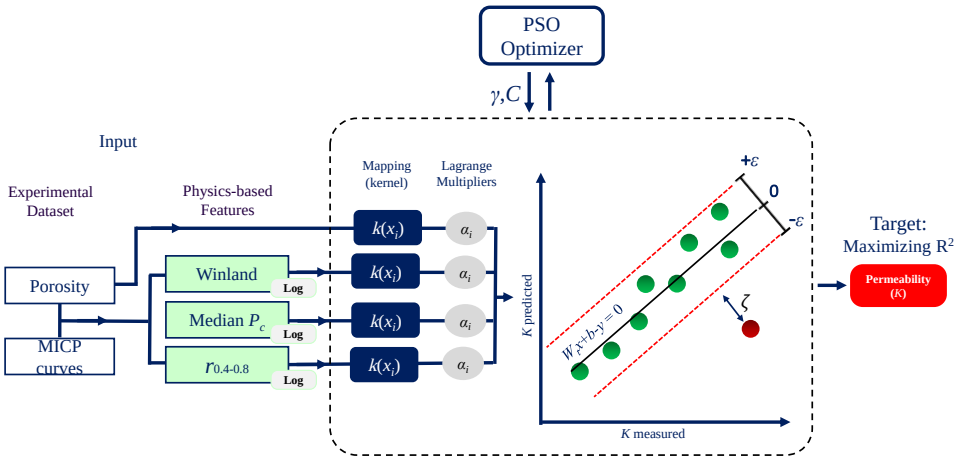

Fig. 5. The flowchart showing the main specifications of the implemented approach, including the feature engineering and the SVM model, coupled with the PSO algorithm. Before passing the data into the optimizer, the abstracted properties of the MICP curves are extracted using the physics-based approaches.

below function (Smola and Schölkopf, 2004):

$$-\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)-\varepsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)+\sum_{i,j=1}^{n}y_i(\alpha_i-\alpha_i^*) \tag{20}$$

Subject to the constraint:

$$\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0 \text{ and } \alpha_i,\alpha_i^*\in[0,C] \tag{21}$$

In this operation, the Lagrangian multipliers (α_i and α_i^*) are optimized for each sample to minimize the error between the measured and predicted permeabilities. Likewise, the expansion of the original hyperplane equation may be written as:

$$f(x)=\sum_{i=1}^{n}\alpha_i-\alpha_i^*k(x_i,x)+b \tag{22}$$

After the calculation of Lagrange multipliers α_i and α_i^*, by considering that $K(x_i,x_j)=\Phi^T(x_i)\cdot\Phi(x_i)$, one can find an optimal weight vector of the hyperplane as:

$$\omega=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)\Phi(x_i) \tag{23}$$

where the SVM tends to be overfittin g in large C values, while in small C cases, the SVM model leans towards underfitting. The support vector regression operation equation above has three metaparameters (C, γ, and ε). Since it is accepted that the accuracy of an SVM model relies on a correct setting of these metaparameters, these values need to be optimized in the training stage.The values of these metaparameters control the learning speed and generalization ability of the model.

2.7 Particle swarm optimization

The tuning parameters of different ML algorithms, which may have large impacts on the efficiency of the training process, need to be optimized during the training process. In this work, particle swarm optimization (PSO) is applied as the optimization algorithm. This nature-inspired algorithm is one of many evolutionary optimization methods that can minimize the objective function by iterative improvement of the solutions. The method works by considering a population (called swarms) of the possible solutions (particles). The population of swarms moves through the search space until the optimum solution is found. The advantage of the PSO algorithm is its tolerance in non-homogeneous conditions. This algorithm was originally introduced by Kennedy and Eberhart (1995), but a large number of variants were introduced after that. The position of a particle from x_κ^i (it can be C or γ in the SVM model) will be evolved to $x_{\kappa+1}^i$ as:

$$x_{\kappa+1}^i=x_\kappa^i+v_{\kappa+1}^i \tag{24}$$

The subscript κ indicates the increment of time. p_κ^i is the optimum position of the swarm i at time κ so far, while p_κ^g represents the global optimum position for all swarms at time κ. r_1 and r_2 are random values between 0 and 1. Also, c_1 and c_2 are the cognitive and social scaling parameters, respectively, which are selected such that $c_1=c_2=2$ to give a mean equal to 1 when they are multiplied by r_1 and r_2. More information on the theoretical aspects related to this algorithm is provided in Kameyama (2009).

2.8 Statistical assessment

To analyze the fitting quality of the models and predictions, different statistical parameters were used. These are introduced in the following. The coefficient of determination or R^2 score is defined as the value of changes in the dependent variable that is predictable from the independent variable:

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-x_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2} \tag{25}$$

In this equation, y_i is the measured variable, x_i is the predicted variable, and $\bar{y}_i$ is the overall mean of the measured vector. The mean absolute error (MAE), as is clear from its name, is the average value of all absolute errors:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{26}$$

Also, to have a good understanding of the error in comparison to the true value, the relative error is defined as:

$$\text{Relative Error} = \frac{|x_i - y_i|}{y_i} \tag{27}$$

where the average of all points is called mean relative error (MRE).

3. Model implementation

In this work, the support vector machine regressor algorithm is applied for the prediction of rock permeability from MICP test results. Due to the complexity of the investigated phenomena, it is decided to use the RBF kernel because of its scaling capabilities. As shown, the SVM model coupled with the RBF kernel has 3 metaparameters that control the accuracy and generality of the model. Since both SVM and RBF models have metaparameters that need to be optimized, here the PSO algorithm is coupled with the SVM model. Fig. 5 shows the simplified flowchart of the used model. The PSO optimizer is coupled with the SVM model to find the optimum value meta parameters and the coefficient of determination (R^2 score) is selected as the objective function to be maximized. The regularization coefficient (C) and RBF width parameter (ω) are imported as the tuning variables to fit the model. The epsilon distance is selected to be a low value of 0.05 to improve the accuracy of the model. So, the number of dimensions in the PSO model is 2 meaning that particles are crawling in a 2D space where the axes are C and ω. Also, 10 particles (the swarm) are applied in the PSO optimizer to locate optimized values of the metaparameters. More information about the implemented model is provided in Table 1. The iteration number of 600 is selected for the swarm to search for the metaparameters.

After randomly dividing the dataset, 80% of the samples were used for the training of the model, and the others were selected as the testing dataset (no validation dataset is needed in SVM models). It should be mentioned that the SVM models are not scale-invariant, meaning that the respective scale of variables influences the accuracy of the model. Because of that, it is important to reduce the scale of all inputs to the same scale of magnitude. So, except for the porosity, for the other features, the logarithmic values have been used.

4. Results

After introducing the main specifications of the workflow, in this section, the results of feature design and selection, and ML model training and testing are provided.

4.1 Feature engineering

In this section, it is tried to find the most correlated input features for the ML model. At first, the capability of empirical or semi-analytical correlations in the prediction of rock permeability is investigated. Fig. 6 compares the predictions with the actual values of permeability, and Fig. 7 compares the MAE and RMSE for the predictions, as well as Kendal's correlation coefficients. Fig. 6(a) shows the porosity-permeability correlation for the dataset. Also, the calculated permeability values for 5 different models, including Swanson, Purcell, Winland, Parachor, and Fractal models, are compared with the actual values in Figs. 6(b)-6(f). The mathematical details of the correlations are provided in the previous sections. As it is shown in Fig. 6, the predictions of these models almost follow the increasing trend of permeability. However, there are some outliers in the calculated values that show that purely relying on the specific points of the capillary curve may lead to significant errors, or in the other words, any measurement errors in these curves may lead to large prediction errors.

Fig. 7 compares the prediction errors (MAE and RMSE) for different permeability models, and also their correlation coefficients with the true rock permeability. It is shown that the predictions related to the Winland model, Swanson model, and linear regression of the Porosity-Permeability scatter plot had the minimum errors. The fractal model resulted in the highest error values. Fig. 6 shows that the fractal model could predict permeability for a large fraction of the dataset while a small portion of the data had large errors. Purcell and Parachor models almost showed similar results and prediction errors compared to each other. On the other hand, the calculated Kendall's tau in Table 2 shows that the Winland r_{35} and Porosity-Permeability relationship have the highest correlation with the permeability values ($\tau > 0.5$). Other correlations have τ values of less than 0.4. In ML models, the importance of the correlation coefficient is undeniable. So, in this work, considering the RMSE and correlation coefficients, the rock porosity and the Winland model were selected as the most suitable theory-based features for the ML-based prediction of rock permeability.

To find more insightful features, it is tried to extract more parameters from the P_c curves. To do so, the correlation of a series of statistical parameters with the rock permeability was examined. Firstly, two statistical parameters of $\overline{P}_c$ and $MedP_c$ in the capillary pressure curve are chosen as the representatives of the rock pore characteristics. Fig. 8 shows the correlation of these two statistical parameters with rock permeability. From this figure and Table 3, it is clear that the median value of capillary pressure is more correlated to the permeability, with Kendal's coefficient of -0.599, which is significant in comparison to the previous correlations. From this difference between mean and median P_c features, it can be concluded that the permeability of rocks in our dataset mostly follows the characteristics of medium size pores, and both very large and very small pores do not significantly influence the permeability.

Fig. 9 shows the scatter plot related to the average equivalent pore-throat radius ($\overline{r}$) values for different saturation ranges of capillary pressure curves. It is clear from the figure that overall, the permeability is correlated with the $\overline{r}$ for all of the S_{Hg} ranges. However, it was tried to find the ranges with the highest correlation values. As Table 3 shows Kendal's correlation coefficient values for these ranges, the most optimum sampling range is 0.4-0.8 ($\overline{r}_{0.4-0.8}$), where the calculated pore radiuses have the most correlations with

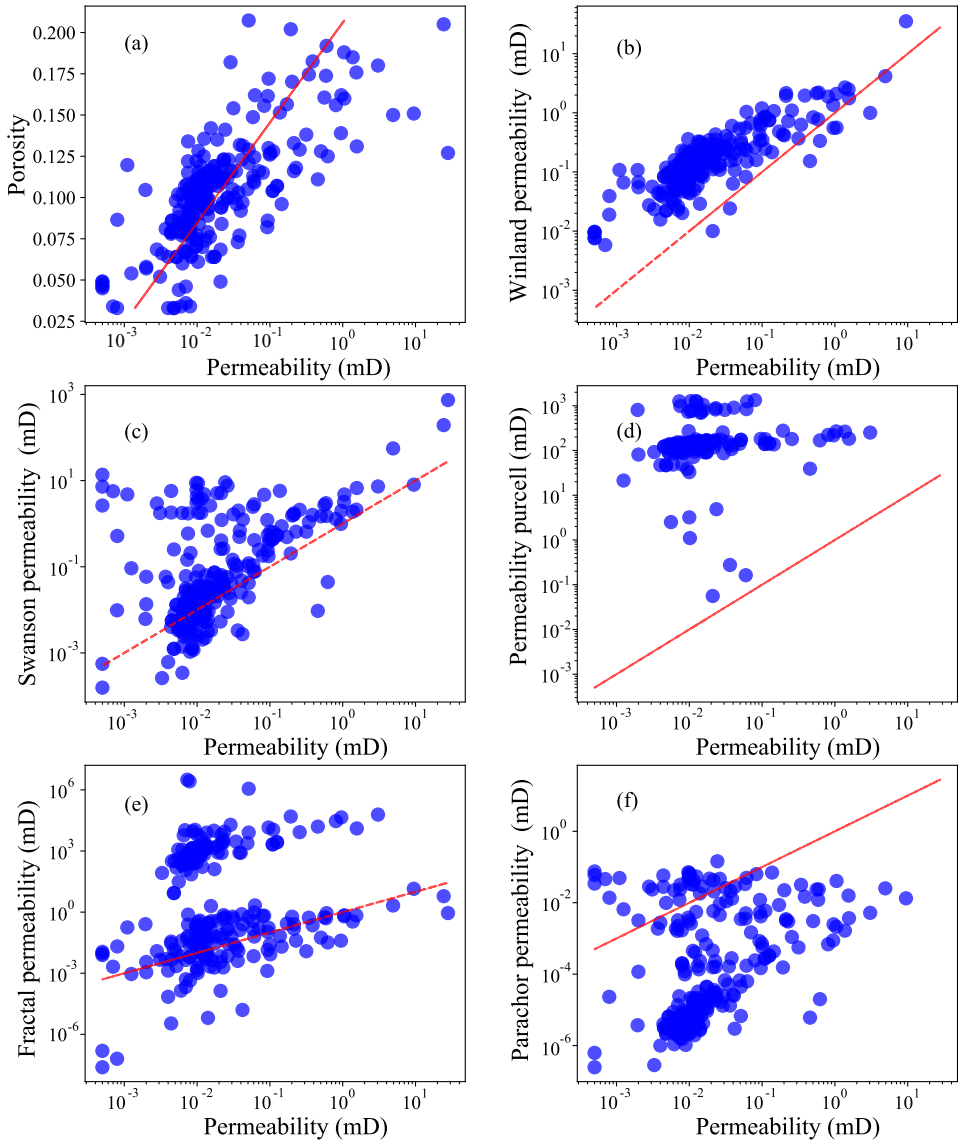

Fig. 6. The calculated values of permeability for different correlations. (a) Porosity-Permeability plot, (b) winland method, (c) swanson method, (d) purcell method, (e) fractal model, (f) parachor model. The red lines show the true permeability values.

Table 2. The statistical evaluation of the empirical permeability models and their correlation with the measured permeabilities.

Models	MAE (Log mD)	RMSE (Log mD)	Kendal's correlation coefficient
Porosity-permeability	0.34	0.38	0.545
Winland	0.92	0.49	0.645
Swanson	0.77	0.67	0.387
Purcell	2.48	3.42	0.283
Parachor	1.81	2.37	0.294
Fractal	2.40	5.16	0.151

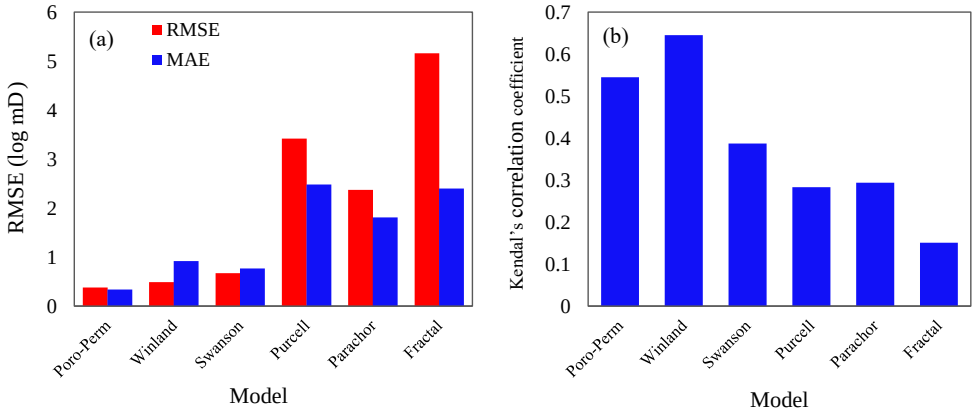

Fig. 7. Comparison of the results of the different permeability predictions of different correlations and comparing them with the linear porosity/permeability regression error (Phi-Perm), (a) MAE and RMSE and (b) Kendal's correlation coefficient.

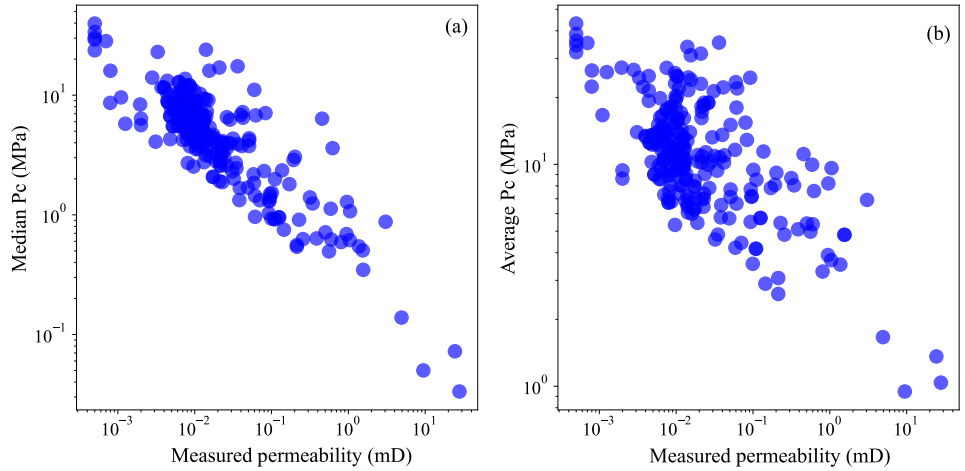

Fig. 8. Correlation of the capillary pressure curve derived features with the measured permeability. More details are provided in Table 3. (a) Average capillary pressure and (b) Median capillary pressure.

the rock permeability. This is in line with the results of previous studies like Kolodzie (1980). Actually, the pores filled in this saturation range have the highest population and highest contribution to rock permeability. Of course, $\bar{r}_{0.6-0.8}$ had higher correlation coefficient than $\bar{r}_{0.4-0.8}$, but since their differences were insignificant, it was decided to choose the $\bar{r}_{0.4-0.8}$ to keep the generality of the model and reduce the chance of being affected by measurement errors. On the other hand, the saturation ranges of 0-0.2 have the worst results, which is understandable considering that the low saturation values are related to the tightest pores. These pores cannot be sufficiently effective in the fluid transmissibility of rock.

Finally, the optimum combination of features is selected based on Kendal's correlation coefficient and also the RMSE of predictions. Since the training of the ML model is based on the relevance of the features, the features with a correlation coefficient higher than 0.5 were prioritized (Fig. 10). So, the following are selected as the inputs of the SVM-PSO model:

- Porosity (unit: fraction)
- Winland permeability model (unit: log mD)
- The median value of capillary pressure, MedPc (unit: log bar)
- The average pore-throat radius is calculated from P_c curves at the wetting phase saturation range of 0.4 to 0.8, $\bar{r}_{0.4-0.8}$ (unit: log μm)

The parentheses show the units of measurement used. Except for the porosity, other features have been transferred to the log scale before usage.

4.2 Model training and validation

After adding the training dataset (80% of the dataset) to the implemented SVM-PSO model, the model is trained with the RBF kernel, a Nelder-Mead minimizer (Avriel, 2003), and 600 epochs (see the training loss during the training in Fig. 11(a)). To avoid overfitting the model, the fitting of predicted values to the measured permeabilities is cross validated after finishing the whole optimization stage. It is found that the fitting scores of test data and training data tend to have an almost similar value.

The final cross-plot of permeability for the training data is shown in Fig. 11(b). For the training data, the R^2 score was 0.89, which is significant in comparison to the theory-based

Table 3. The Kendal's correlation coefficient of statistical and pore-throat radius features extracted from the capillary pressure curves with the measured permeability values.

Feature type	Feature	Kendal's correlation coefficient
Statistical	Average P_c	-0.380
	Median P_c	-0.599
Mean pore-throat radius	$\bar{r}_{0-0.2}$	0.273
	$\bar{r}_{0.2-0.4}$	0.450
	$\bar{r}_{0.4-0.6}$	0.563
	$\bar{r}_{0.6-0.8}$	0.587
	$\bar{r}_{0.8-1.0}$	0.476
	$\bar{r}_{0.4-0.8}$	0.580

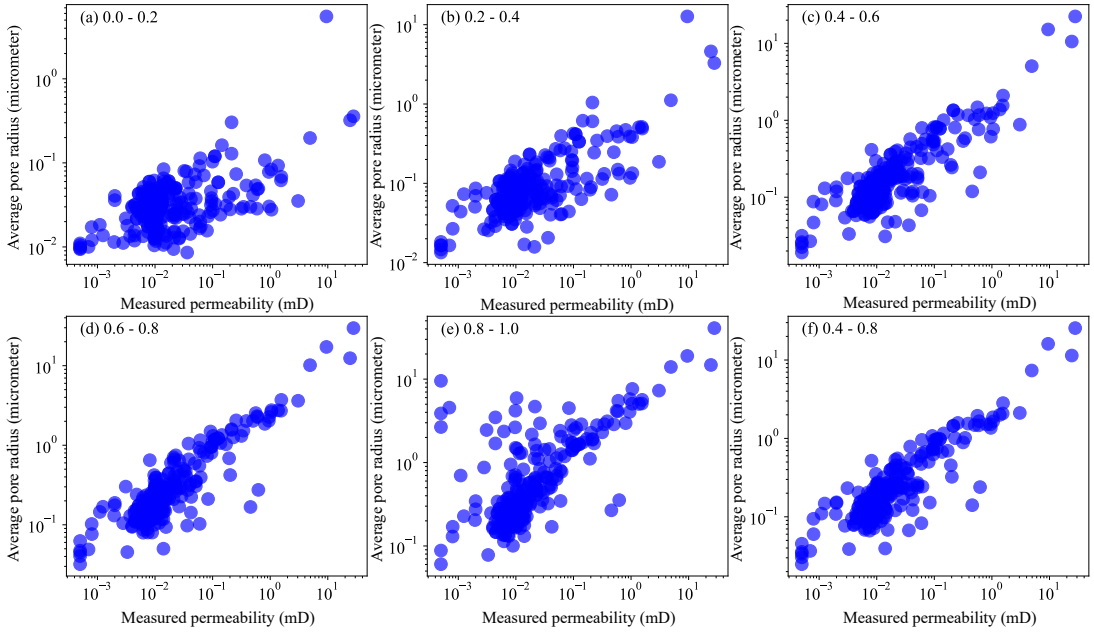

Fig. 9. Correlation of the pore radiuses (capillary pressure curves) versus the measured permeability for different saturation ranges.

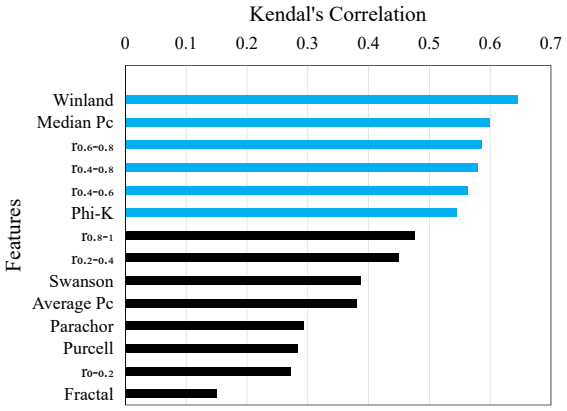

Fig. 10. The tornado chart related to the value of Kendal's correlation coefficient (absolute value) ordered from high to low. The features with a τ score higher than 0.5 are selected for the model development process.

models. Fig. 11(c) shows the predicted values of permeability in the testing stage. Compared with the training data, the permeability predictions show an acceptable accuracy. As Fig. 11(d) shows, the error of predictions was significantly lower than correlations like the Winland equation. The detailed statistics on the comparison of the SVM-PSO model are shown in Table 4, which shows that the correlation coefficient score has been improved up to around 0.74. In the input features, the maximum correlation coefficient was not higher than 0.645. Also, for the testing data, the MRE value for the Winland equation was around 10 (meaning that the error of predictions was at the level of 10 times the true values, on average), while this value was reduced to the level of 0.47 in the SVM-PSO model. It shows a significant improvement in the level of permeability predictions for the SVM-PSO model. Fig. 12 compares the permeabilities predicted by Winland and SVM-PSO models for the testing data. From the Winland

Table 4. The R^2 accuracy, RMSE, MRE, and Kendal's correlation coefficient related to the predictions of the SVM-PSO model for both training and testing data..

Algorithm	R^2	RMSE (log mD)	MRE (fraction)	Correlation coefficient
SVM-PSO training	0.89	0.25	0.35	0.72
SVM-PSO testing	0.87	0.27	0.47	0.74
SVM-PSO total	0.88	0.26	0.40	0.72

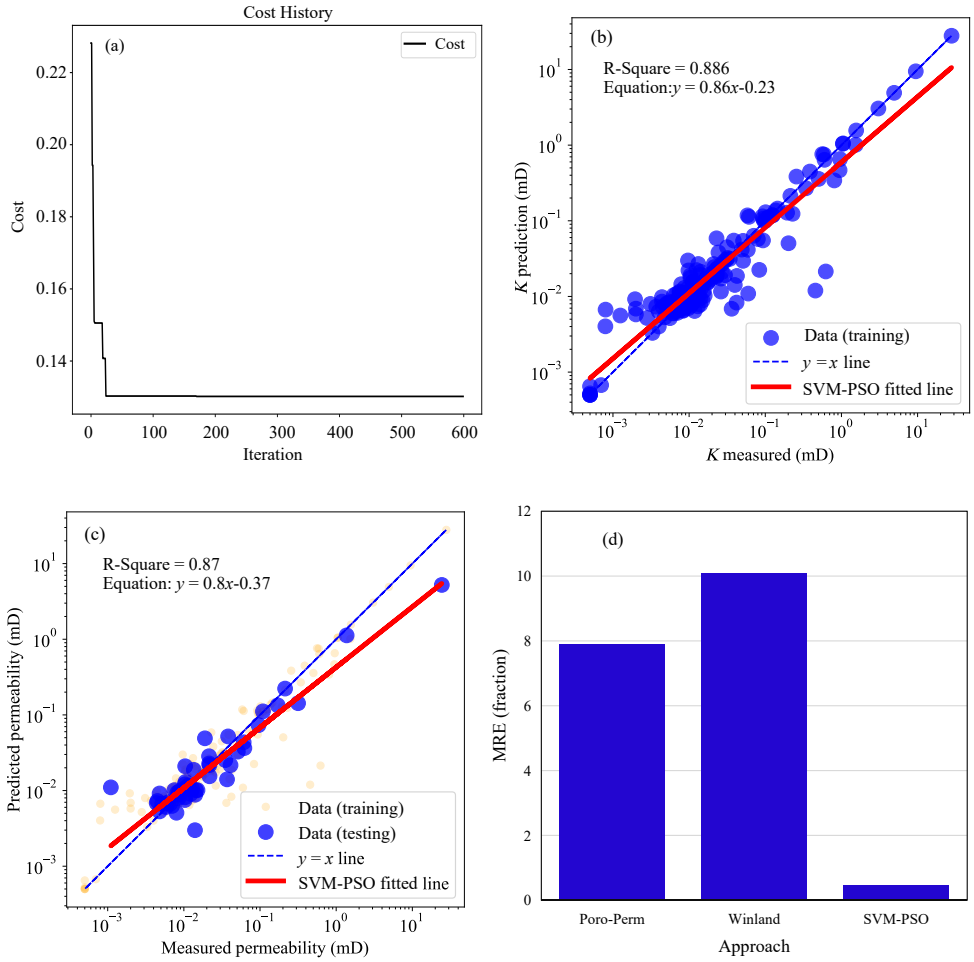

Fig. 11. The results of training and testing of the Support Vector Machine model using the PSO optimization algorithm. (a) comparing the predicted values and actual measured values. The red line shows the fitted curves on the predicted values and the blue line shows the $y = x$ trend, (b) the convergence trend of the PSO model for iterations, (c) the SVM-PSO model predictions for the test data. Orange points in the background (low transparency) show the training results and (d) the MRE related to the SVM-PSO algorithm, compared to the Winland correlation predictions, for the testing data.

predictions, it can be concluded that there are significant errors in low permeability cases, which is an indication of the necessity of the development of a new model. The SVM-PSO model could reduce the level of errors by 1 to 2 orders of magnitude to the mean MRE of 0.40, which is significant for a permeability estimator of tight sandstones.

The obtained results show that using ML models in combination with theoretical and statistical calculations is helpful in improving permeability predictions. It is observed that training the model using a larger number of datasets can improve the generality of the SVM-PSO models. In that case, using Convolutional Deep Learning approaches can be helpful for a more accurate interpretation of MICP curves. Also, using smooth capillary pressure curves can decrease prediction errors.

5. Conclusion

In this work, a large dataset of MICP tests related to tight sandstones from all over the world is gathered and an SVM-PSO ML model is used for the prediction of permeability from the MICP data.

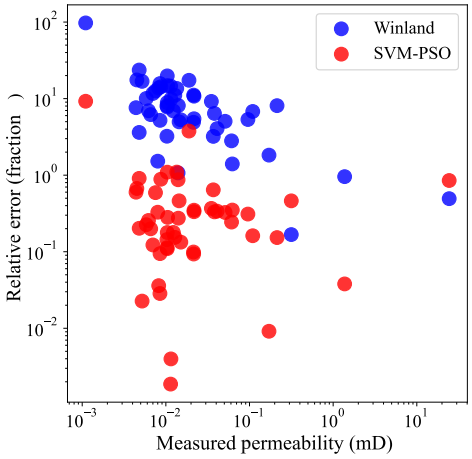

Fig. 12. Comparison of the relative error in the testing data for Winland and SVM-PSO models.

- The permeability predictions of the analytical models (Parachor, Purcell, Winland, Fractal, and Swanson models) were not accurate enough, especially for the rocks with lower permeabilities. The most accurate model was the Winland r_{35} equation with a correlation coefficient of 0.645 and an MRE of 10.
- To improve the predictions, a study was carried out to find the most relevant features to be imported as the input variables to the ML model. Porosity data and Winland permeability predictions are chosen as the input features of the ML model due to their good correlations with the true permeabilities.
- Furthermore, the median P_c and the $\bar{r}_{0.4-0.8}$ (mean pore radius for the S_{Hg} range of 0.4-0.8), which best correlated with the rock permeability measurements, were chosen as the input features. It is concluded that the permeability of the rocks was mostly dependent on the medium size pore-throats, not the largest or smallest ones.
- The SVM-PSO ML model fitted the actual permeability values with an R^2 of 0.89 and an MRE of 0.37, significantly more accurate than the theoretically based permeability model predictions. The model could significantly improve the predictions for the low-permeability rocks, where the theoretical correlations performed weakly.
- The results showed that merging machine-learning algorithms with current empirical or physics-based models can significantly improve permeability modeling practices. Using this approach, both physics-driven and data-driven approaches can synergize in the development of more reliable approaches.

Acknowledgement

A version of this paper is previously presented at SPWLA 2022 conference, in Stavanger, Norway. The authors appreciate SPWLA for permission to publish the paper in a journal. Also, Andersen acknowledges the Research Council of Norway and the industry partners, ConocoPhillips Skandinavia AS, Aker BP ASA, Vår Energi AS, Equinor ASA, Neptune Energy Norge AS, Lundin Norway AS, Halliburton AS, Schlumberger Norge AS, and Wintershall DEA, of The National IOR Centre of Norway for support.

Conflict of interest

The authors declare no competing interest.

References

Abbasi, J., Andersen, P. Ø. Theoretical comparison of two setups for capillary pressure measurement by centrifuge. Heliyon, 2022, 8(9): e10656.

Ahmad, N., Wörman, A., Sanchez-Vila, X., et al. Injection of CO_2-saturated brine in geological reservoir, A way to enhanced storage safety. International Journal of Greenhouse Gas Control, 2016, 54: 129-144.

Arabjamaloei, R., Daniels, D., Ebeltoft, E., et al. Validation of permeability and relative permeability data using mercury injection capillary pressure data. E3S Web of Conferences, 2019, 89: 01001.

Avriel, M. Nonlinear Programming: Analysis and Methods. Massachusetts, USA, Courier Corporation, 2003.

Blunt, M. J. Multiphase Flow in Permeable Media. London, UK, Cambridge University Press, 2017.

Burn, R. P., Mandelbrot B. B. The Fractal Geometry of Nature. New York, USA, W. H. Freeman, 1984.

Changtao, Y., Shuyuan, L., Hailong, W., et al. Pore structure characteristics and methane adsorption and desorption properties of marine shale in Sichuan Province, China. RSC Advances, 2018, 8: 6436-6443.

Choi, S., Kim, T., Yu, W. Performance evaluation of RANSAC family. Paper Presented at British Machine Vision Conference, London, UK, 7-10 September, 2009.

Cortes, C., Vapnik, V. Support-vector networks. Machine Learning, 1995, 20: 273-297.

Croux, C., Dehon, C. Influence functions of the spearman and kendall correlation measures. Statistical Methods & Applications, 2010, 19: 497-515.

Erofeev, A., Orlov, D., Ryzhov, A., et al. Prediction of porosity and permeability alteration based on machine learning algorithms. Transport in Porous Media, 2019, 128: 677-700.

Eslami, M., Kadkhodaie-Ilkhchi, A., Sharghi, Y., et al. Construction of synthetic capillary pressure curves from the joint use of NMR log data and conventional well logs. Journal of Petroleum Science and Engineering, 2013, 111: 50-58.

Fan, C., Zhong, C., Zhang, Y., et al. Geological factors controlling the accumulation and high yield of marine-Facies shale gas: Case study of the wufeng-longmaxi formation in the dingshan area of southeast sichuan, China. Acta Geologica Sinica, 2019, 93: 536-560.

Feng, F., Wang, P., Wei, Z., et al. A new method for predicting the permeability of sandstone in deep reservoirs. Geofluids, 2020, 2020: 8844464.

Ge, X., Fan, Y., Deng, S., et al. An improvement of the fractal

theory and its application in pore structure evaluation and permeability estimation. Journal of Geophysical Research: Solid Earth, 2016, 121(9): 6333-6345.

Guo, B., Ghalambor, A., Duan, S. Correlation between sandstone permeability and capillary pressure curves. Journal of Petroleum Science and Engineering, 2004, 43(3-4): 239-246.

Hébert, V., Porcher, T., Planes, V., et al. Digital core repository coupled with machine learning as a tool to classify and assess petrophysical rock properties. E3S Web of Conferences, 2020, 146: 01003.

Jiao, L., Andersen, P. Ø., Zhou, J., et al. Applications of mercury intrusion capillary pressure for pore structures: A review. Capillarity, 2020, 3(4): 62-74.

Jones, S. C. A technique for faster pulse-decay permeability measurements in tight rocks. SPE Formation Evaluation, 1997, 12(1): 19-25.

Kameyama, K. Particle swarm optimization-a survey. IEICE Transactions on Information and Systems, 2009, 92(7): 1354-1361.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., et al. Physics-informed machine learning. Nature Reviews Physics, 2021, 3(6): 422-440.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., et al. Machine Learning for the Geosciences: Challenges and opportunities. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(8): 1544-1554.

Kendall, M. G. Rank Correlation Methods 4th edn. London, UK, High Wycombe, 1976.

Kennedy, J., Eberhart, R. Particle swarm optimization. Paper Presented at Proceedings of ICNN'95-International Conference on Neural Networks, Perth, WA, 27 November-1 December, 1995.

Kolodzie, S. Analysis of pore throat size and use of the waxman-Smits equation to determine ooip in spindle field, colorado. Paper SPE-9382-MS Presented at the SPE Annual Technical Conference and Exhibition, Dallas, Texas, 21-24 September, 1980.

Li, K. Analytical derivation of brooks-corey type capillary pressure models using fractal geometry and evaluation of rock heterogeneity. Journal of Petroleum Science and Engineering, 2010, 73(1-2): 20-26.

Lin, Q., Bijeljic, B., Pini, R., et al. Imaging and measurement of pore-scale interfacial curvature to determine capillary pressure simultaneously with relative permeability. Water Resources Research, 2018, 54(9): 7046-7060.

Liu, M., Xie, R., Li, C., et al. Determining the segmentation point for calculating the fractal dimension from mercury injection capillary pressure curves in tight sandstone. Journal of Geophysics and Engineering, 2018, 15(4): 1350-1362.

Liu, Y., Xian, C., Li, Z., et al. A new classification system of lithic-rich tight sandstone and its application to diagnosis high-quality reservoirs. Advances in Geo-Energy Research, 2020, 4(3): 286-295.

McPhee, C., Reed, J., Zubizarreta, I. Core Analysis: A Best Practice Guide. Amsterdam, Netherlands, Elsevier, 2015.

Menke, H. P., Maes, J., Geiger, S. Upscaling the porosity-permeability relationship of a microporous carbonate for darcy-scale flow with machine learning. Scientific Reports, 2021, 11(1): 2625.

Purcell, W. R. Capillary pressures-their measurement using mercury and the calculation of permeability therefrom. Journal of Petroleum Technology, 1949, 1: 39-48.

Puth, M. T., Neuhäuser, M., Ruxton, G. D. Effective use of spearman's and kendall's correlation coefficients for association between two measured traits. Animal Behaviour, 2015, 102: 77-84.

Rezaee, R., Saeedi, A., Clennell, B. Tight gas sands permeability estimation from mercury injection capillary pressure and nuclear magnetic resonance data. Journal of Petroleum Science and Engineering, 2012, 88: 92-99.

Smola, A. J., Schölkopf, B. A tutorial on support vector regression. Statistical Computations, 2004, 14(3): 199-222.

Swanson, B. F. Simple correlation between permeabilities and mercury capillary pressures. Journal of Petroleum Technology, 1981, 33(12): 2498-2504.

Tran, H., Sakhaee-Pour, A., Bryant, S. L. A simple relation for estimating shale permeability. Transport in Porous Media, 2018, 124(3): 883-901.

Wang, F., Jiao, L., Liu, Z., et al. Fractal analysis of pore structures in low permeability sandstones using mercury intrusion porosimetry. Journal of Porous Media, 2018, 21(11): 1097-1119.

Wang, F., Yang, K., You, J., et al. Analysis of pore size distribution and fractal dimension in tight sandstone with mercury intrusion porosimetry. Results in Physics, 2019, 13: 102283.

Wang, J., Wang, J. Low-amplitude structures and oil-gas enrichment on the yishaan slope, ordos basin. Petroleum Exploration and Development, 2013, 40(1): 52-60.

Xiao, L., Liu, D., Wang, H., et al. The applicability analysis of models for permeability prediction using mercury injection capillary pressure (MICP) data. Journal of Petroleum Science and Engineering, 2017, 156: 589-593.

Xiao, L., Liu, X., Zou, C., et al. Comparative study of models for predicting permeability from nuclear magnetic resonance (NMR) logs in two Chinese tight sandstone reservoirs. Acta Geophysica, 2014, 62(1): 116-141.

Zhang, C., Cheng, Y., Zhang, C. An improved method for predicting permeability by combining electrical measurements and mercury injection capillary pressure data. Journal of Geophysics and Engineering, 2017, 14(1): 132-142.

Zhang, G., Wang, Z., Mohaghegh, S., et al. Pattern visualization and understanding of machine learning models for permeability prediction in tight sandstone reservoirs. Journal of Petroleum Science and Engineering, 2021, 200: 108142.

Zhao, G., Li, X., Liu, M., et al. Reservoir characteristics of tight sandstone and sweet spot prediction of dibei gas field in eastern kuqa depression, northwest China. Energies, 2022, 15(9): 3135.