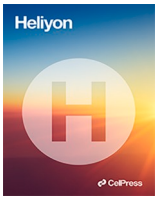
Heliyon 10 (2024) e32666

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Advanced machine learning approaches for predicting permeability in reservoir pay zones based on core analyses

Amad Hussen, Tanveer Alam Munshi, Labiba Nusrat Jahan, Mahamudul Hashan *

Department of Petroleum and Mining Engineering, Shahjalal University of Science and Technology, Sylhet 3114, Bangladesh

ARTICLE INFO

Keywords:
Reservoir characterization
Machine learning
Tree-based algorithm
Core data
Permeability modeling
Porosity
Oil saturation
Core features ranking

ABSTRACT

Permeability is the most important petrophysical characteristic for determining how fluids pass through reservoir rocks. This study aims to develop and assess intelligent computer-based models for predicting permeability. The research focuses on three novel models—Decision Tree, Bagging Tree, and Extra Trees—while also investigating previously applied techniques such as random forest, support vector regressor (SVR), and multiple variable regression (MVR). The primary dataset consists of 197 data points from a heterogeneous petroleum reservoir in the Jeanne d'Arc Basin, including laboratory-derived permeability (K), oil saturation (S_O), water saturation (S_W), grain density (ρ_{gr}), porosity (φ), and depth. The most effective machine learning models are identified by a thorough analysis that makes use of a variety of statistical metrics, such as the coefficient of the determinant (R^2), mean squared error (MSE), mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), maximum error (maxE), and minimum error (minE). Additionally, core features are ranked based on their importance in permeability modeling. This study deviates from conventional approaches by proposing an efficient means of forecasting permeability, reducing reliance on labor-intensive and time-consuming laboratory work. The findings reveal that MVR is unsuitable for permeability prediction, with all developed models outperforming it. Extra Trees emerges as the most accurate model, with an R^2 of 0.976, while random forest and bagging tree exhibit slightly lower R^2 values of 0.961 and 0.964, respectively. The ranking of these algorithms based on performance criteria is as follows: extra trees, bagging tree, random forest, SVR, decision tree, and MVR. The study also presents a detailed analysis of the impact of input parameters, highlighting porosity (φ) and water saturation (S_W) as the most influential, while grain density (ρ_{gr}), oil saturation (S_O), and depth are considered less important. This study contributes to the petroleum industry's knowledge by showcasing the inadequacy of MVR and highlighting the superior performance of machine learning models, particularly Extra Trees. The proposed models employed in this study can help engineers and researchers determine reservoir permeability quickly and accurately by using a few core attributes, reducing the dependency on resource-intensive and time-consuming laboratory work.

* Corresponding author.
E-mail address: mahmud-pme@sust.edu (M. Hashan).

https://doi.org/10.1016/j.heliyon.2024.e32666
Received 16 January 2024; Received in revised form 27 May 2024; Accepted 6 June 2024
Available online 11 June 2024

1. Introduction

The process of characterizing a reservoir involves combining a vast amount of seismic data, well logs, and geological samples. The best and most important tools for figuring out reservoir characteristics are those that use petrophysical log analysis. Since permeability is a crucial petrophysical parameter that captures the inherent dynamic flow characteristic of rocks, it is an essential component of reservoir characterization [1]. Precise permeability forecasting encompasses crucial information regarding fluid saturation distribution, prospective recoverable oil and gas yield from the reservoir, projected future exploration, appropriate production equipment selection, efficient reservoir management, and efficacious water injection plans [2]. The laboratory analysis on core plugs, which consumes a significant amount of time and materials, provides the most accurate permeability data [3]. Nevertheless, due to the

Table 1
Data-driven approaches exist in literature to attain log generated permeability.

SN	Reference	Input Log Variables	Output Variables	Sample Size	Statistical Parameter	Algorithms	Test Scores
2010	Al-Anazi & Gates [9]	GR, DT, RHOB, LLD, φ_N	K, φ	701	R MSE	BPNN GRNN SVM	R^2: 0.885 R^2: 0.930 R^2: 0.896
2011	Olatunji et al. [10]	MSFL, DT, NPHI, PHIT, RHOB, S_w	K	155	R^2 RMSE Ea	ANN SVM Type-1 Fuzzy Type-2 Fuzzy	R^2: 0.820 R^2: 0.874 R^2: 0.914 R^2: 0.921
2012	Gholami et al. [11]	DT, GR, NPHI, ROHB, PEF, MSFL, LLS, LLD,	K	175	R RMSE	SVM GRNN	R^2: 0.96 R^2: 0.94
2014	Olatunji et al. [12]	MSFL, DT, NPHI, PHIT, RHOB, S_w	K	1854	R^2, RMSE AAPRE(Ea)	SBLLM type-2 FLS	R^2: 0.869 R^2: 0.938
2014	Ahmadi et al. [13]	DT, RHOB, NPHI, PHIT	K, φ	1000	MSE R-square	LSSVM FIS GA-FIS	R^2: 0.994 R^2: 0.837 R^2: 0.962
2014	Baziar et al. [14]	GR, DT, LLD, RHOB, φ_N	K	161	R MSE AAPE	MLP-NN CANFIS SVM	R^2: 0.652 R^2: 0.789 R^2: 0.652
2014	Kaydani et al. [15]	Depth, SGR, PHIT, RHOB, CT, S_w, NPHI	K	980	MAE RMSE R^2 R	MGGp ANN-LM ANFIS GP	R2: 0.947 R^2: 0.890 R^2: 0.935 R^2: 0.812
2015	Santisukkasaem et al. [16]	Fluid residence with PRB, Pressure drop, dynamic viscosity of fluid, porosity, volumetric flow rate, particle size, length of reactor,	K	NA	AARE SSE TS R, E	ANN LR Non-LR	– – –
2017	Rafik & Kamel [17]	GR, RLLD, DT, NPHI, RHOB, SW	K, φ	927	FZI HFU	ACE GAM NNET	– – –
2017	Al-Mudhafar [18]	Shale Volume, neutron porosity, water saturation	Lithofacies K	669	RMSPE R-Square	PNN GBM	Adj *R*-sq: 0.9551 Adj *R*-sq: 0.9953
2019	Erofeev et al. [19]	Salts concentration, Formation top depth, formation bottom depth, porosity before desalination, Absolute permeability before desalination, Sample depth, Sample density, Average grain size, Color, Depth horizon	K, φ	102	R^2 MAE MSE,	LR DT RF GB XGBoost SVM NN	R^2: 0.852 R^2: 0.677 R^2: 0.775 R^2: 0.809 R^2: 0.856 R^2: 0.850 –
2020	Urang et al. [20]	RHOB, Water Saturation	K	1199	RMSE R-squared Adjust R Squared SE	ANN Nonlinear Regression (Curve fittings)	R^2: 0.9758 R^2: 0.9753
2020	Wood [21]	GR, LLD, DT, RHOB, DT, φ_N	K, φ, S_w	1000	R AAPE MSE, RMSE	Transparent open box optimized data-matching algorithm	R^2: 0.999
2021	Tian et al. [3]	Porous Media	K	1000	R	GA-ANN	R^2: 0.995
2021	Aljuboori et al. [22]	NPHI, DT, RHOB, LLD, LLS, GR, and SP, FZI	K	256	R^2	MVR Neural network	R^2: 0.46 R^2: 0.807
2021	Farouk et al. [23]	Resistivity, GR, RHOB, NPHI, Porosity	K	95	R^2, MSE	PSO-NN LS-SVM	R^2: 0.764 R^2: 0.862
2022	Hashan et al. [24]	Lithology log, porosity log, NMR log, Resistivity log	K	439	R, MSE RMSE, AAPE	MVR GPR BT SVM CNN ANN-LM ANN- SCG ANN-BR	R^2: 0.610 R^2: 0.950 R^2: 0.780 R^2: 0.940 R^2: 0.250 R^2: 0.970 R^2: 0.690 R^2: 0.980
2022	Subasi et al. [2]	Gamma Ray, depth, neutron, Electrical Resistivity, Density	K	1140	R, RAE MAE, RMSE RRSE	ANN K-NN SVM Random Forest SGB	R^2: 0.665 R^2: 0.717 R^2: 0.691 R^2: 0.784 R^2: 0.795
2022	Miah & Abir [25]	Rt, GR, RHOB, NPHI, DT	K	265	AAPE, RMSE, CC	LSSVM-CSA	–

additional expenses and effort, not all wells are cored. As a result, directly measured data are only available for a small number of wells or well bore sections. Another helpful technique to ascertain the appropriate permeability is to conduct flow experiments with representative core samples [4]. However, flow studies are costly, laborious, and complex. In addition, such studies don't yield data at every well's location. This emphasizes how important it is to precisely predict permeability using other indirect approaches in order to achieve it throughout the entire well.

The exploitation of large data sets derived from well logs to characterize heterogeneous reservoirs is difficult because of the reservoir's nonlinearity, heterogeneity, and uncertainty. The development of three-dimensional correlations for rock-fluid characteristics is a significant challenge. A powerful tool is needed to overcome these obstacles [5]. In recent times, research published across several journals has frequently employed machine learning, data-driven deep learning, and statistical techniques to handle problems related to regression and classification in the oil and gas industry [6–8]. For the purpose of obtaining rock-fluid properties from well logs, a number of computer-based intelligent techniques have been developed, including Artificial Neural Networks (ANN), Genetic Algorithms (GA), Convolutional Neural Networks (CNN), Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), Stochastic Gradient Boosting (SGB), Fuzzy-Logic (FL), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM) (Table 1).

Al-Mudhafar [26] incorporated the Bayesian Model Averaging, which reduces uncertainty in permeability modeling. Findings from Akande et al. [27] revealed that the PSO-SVR model outperforms ordinary SVR and RAND-SVR models in permeability modeling. Anifowose et al. [28] compared common and sophisticated machine learning techniques using seismic and wireline data in a carbonate reservoir. His finding reveals that a depth-matched dataset improves permeability prediction with a higher correlation coefficient. In order to forecast φ and K, Erofeev et al. [19] combined a wide range of inputs, including salt concentration, formation top depth, porosity before desalination, sample depth, formation bottom depth, absolute permeability before desalination, sample density, average grain size, color, and depth horizon. The authors used techniques such as GB, DT, ANN, XGBoost, RF, SVM, and linear regression. Al-Mudhafar [29] employed LASSO and BMA techniques to model permeability in uncored sections of a well within a sandstone reservoir. LASSO slightly outperformed BMA in core permeability prediction. Urang et al. [20] conducted experiments in the Niger Delta region to predict K using ANN and standardized nonlinear regression (curve fits), integrating RHOB and water saturation as input features. Kamali et al. [30] developed a group method of data handling (GMDH) algorithm with superior accuracy for precise permeability prediction in carbonate gas condensate reservoirs compared to established empirical correlations. These studies, in conjunction with Tables 1 and 2, suggest that intelligent-based models are favored for addressing computational and data-related problems.

However, most of the studies were performed using well log data, as tabulated in Table 1. Only five studies (included in Table 2) dealt with permeability prediction utilizing input features acquired from laboratory-generated core data. Al Khalifah et al. [31], Mahdaviara et al. [1], Topór [32], Mohammadian et al. [34], Mahdaviara et al. [33], and Kamali et al. [30] applied different computer-based intelligent techniques to derive accurate permeability from core data. As input features, the following were used: depth, grain density, pore throat radius, flow zone indicator, pore-specific surface area, porosity, irreducible water saturation, and formation resistivity factor, all of which were determined in the lab. However, no one ranked core features based on their significance in permeability modeling. The impact of water and oil saturations on a core's permeability has not yet been investigated. Moreover, comprehensive fine-tuning is not provided in these studies. The field of core data-based permeability prediction has not yet established decision tree, extra trees, or bagging tree formulation. To improve the performance of any classifier or regressor, decision tree, extra

Table 2
Data-driven methods available in the literature to attain core-data generated permeability.

Year	References	Input Log Variables	Output Variable	Input Variables Ranking	Sample Size	Statistical Parameter	Algorithms	Testing Score
2020	Al Khalifah et al. [31]	Porosity, Formation resistivity factor, pore throat diameter	K	No	130	R^2 MSE	GA ANN	R^2: 0.858 $R^{2:}$ 0.886
2021	Mahdaviara et al. [1]	Pore-specific surface area, porosity, and irreducible water saturation	K	No	66	MSE RMSE Adjusted R^2	GPR	Adjusted R^2: 0.9864 MSE: 7456
2021	Topór [32]	Depth, porosity, Grain Density	K	No	1002	MAE RMSE R^2	RF MLR	$R^{2:}$ 0.834 $R^{2:}$ 0.800
2022	Mahdaviara et al. [33]	Pore-specific surface area, porosity, and irreducible water saturation	K	No	66	MSE RMSE R^2	LSSVM-CSA MLP-LMA MLP-BR CFNN-LMA CFNN-BR GRNN	R-Squared: 0.904 R-Squared: 0.999 R-Squared: 0.996 R-Squared: 0.998 R-Squared: 0.998 R-Squared: 0.997
2022	Mohammadian et al. [34]	Porosity, connate water saturation, pore throat radius, FZI	K	No	128	R^2 MAE	XGBoost	– R^2: 0.970

trees, and bagging tree are helpful techniques. By using them, a predictor can be strengthened and balanced. When it comes to generalizing any predictor in the testing dataset, these techniques perform remarkably well. They can also assist with overfitting issues. Due to their increased computing efficiency in large-scale datasets and new optimization methodologies, these methods gained appeal. So, permeability prediction from core analysis needs a thorough investigation and optimization using decision tree, extra trees, and bagging tree. It is also necessary to evaluate the influence of core's water and oil saturations on the permeability.

In this study, machine learning models such as Multiple Variable Regression (MVR), decision tree, extra trees, bagging tree, random forest, and SVR are adopted to measure permeability. The considered input parameters are depth, porosity, oil saturation, water saturation, and grain density. These parameters are also ranked to ascertain their significance in respect to permeability. The approach proposed in this study has the potential to help engineers rapidly and precisely determine permeability using only a few core features so that laborious, costly, time-consuming, and monotonous laboratory procedures can be reduced. Although machine learning approaches offer several advantages, they require substantial amounts of high-quality data, which can be a challenge for researchers due to the difficulty of obtaining accurate core data in a laboratory environment.

The goal of the current study is to fill in the research gap by systematically achieving multiple novelties, which may be categorized as follows.

- In the realm of reservoir permeability modeling, intelligent computer-based models, such as decision tree, extra trees, and bagging tree are formulated for the first time. Previously used MVR, SVR, and random forest techniques are also tested to get a thorough conclusion.
- Investigation into how a core's permeability is affected by its water and oil saturations is performed.
- Laboratory-derived core features are ranked based on their significance in permeability modeling.

The remainder of the text is divided into the following sections: the fundamentals of several smart computer-based approaches employed in this work are critically discussed in Section 2. In Section 3, data collection and preparation are covered, along with a synopsis of the instruments and processes employed here. Results and discussions based on important data are presented in Section 4. Section 5 offers a synopsis of the findings and recommendations.

2. Theory

2.1. Multiple variable regression

The multiple variable regression (MVR) method is an advanced version of regression analysis that incorporates numerous predictor variables to evaluate the extent of linear correlation between the independent and dependent variables. The prediction model can be expressed using Equation (1), where X_1, X_2, ..., X_p stand for independent predictor variables, Y refers to the criterion variable, β_1, β_2, ..., β_p are regression coefficients, and e denotes residual error [35]. The least-square solution of the coefficient (β_i), for which $''e''$ becomes zero, is given by Equation (2).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + e \tag{1}$$

$$\hat{\beta} = \left(x^T x\right)^{-1} X^T Y \tag{2}$$

2.2. Decision tree

The decision tree method has undergone extensive research and application across various domains, establishing itself as a dependable and effective tool. It works by dividing the input data into feature values and building a tree iteratively, where each leaf node represents a class label or regression value, and each core node represents a feature [36]. Splitting keeps going until the dataset is all cleaned up. Variance is used by entropy analysis and Gini analysis to assess impurity. The binary splitting approach cannot be used to find the smallest sum of squares. Predictive modeling divides the predictor space into a set of high-dimensional boxes of different sizes. The average of the values in a given box is used to determine the prediction value for an observation that falls inside of it [37]. However, decision tree may not adjust well to new data since they are sensitive to the data that they are trained on. Pruning and ensemble approaches are two of the many solutions that have been suggested for these problems [38].

2.3. Bagging tree

Bagging, also known as bootstrap aggregating, is a popular ensemble strategy that seeks to enhance the stability and accuracy of prediction models. This method is popular in machine learning and has demonstrated encouraging outcomes in improving the predictive power of models [36]. Specific data points might not be correctly predicted by a regression tree working alone, but they can be correctly predicted by an ensemble of regression trees. Using various subsets of the training data, numerous decision tree models are created by applying the Bagging technique. The final prediction is then generated by combining these models via voting or averaging. Three essential steps make up bagging tree modeling: building $B \in N$ bootstrap samples; training the model with the *b-th* bootstrap sample to produce $f\hat{f}^{*b}(x)$; and utilizing Equation (3) to estimate the average of the estimators.

Suppose we have a model that fits our training data set $Z = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, obtaining the prediction $\hat{f}(x)$ at input X. By using the aggregation or bagging method, we can reduce the variance of our prediction by averaging its variance over multiple bootstrap samples. For each bootstrap sample Z^{*b}, where $b = 1, 2, \ldots, B$, we are able to fit our model by providing a prediction $\hat{f}^{*b}(x)$.

$$\widehat{f}_{bag}(x)\frac{1}{B}\sum_{b=1}^{B}\widehat{f}^{*b}(x) \tag{3}$$

The aggregation of predictions from multiple regression trees results in the formation of a bagged tree, which displays significantly reduced bias and variance compared to an individual decision tree. As bagging lowers bias and variance, the predictive performance is enhanced [39].

2.4. Random forest

The random forest approach is an example of an ensemble methodology that combines multiple decision trees to enhance the precision of the model. It usually uses a greater number of trees. The purpose of using a large number of trees is to improve prediction stability and reduce the variance of the ensemble. This approach can handle high dimensional data with several characteristics and is not prone to overfitting. Bootstrap resampling is employed to mitigate the issue of overfitting [19]. Initially, random forest selects m inputs at random from a set of p inputs. The best variable and split point are then chosen from a set of m inputs. It then splits the given node into two distinct daughter nodes. The threshold value is selected completely arbitrarily, which lowers the computational cost of the model and provides an edge over other bagging ensemble models. It can also rank the input attributes. This method might not work when starting with smaller datasets since the bootstrap samples might not accurately represent the entire dataset [40].

2.5. Extremely randomize tree (extra trees)

The extra trees method, an extension of the random forest methodology, uses a whole feature set for every tree and selects splitting points at random instead of conducting a comprehensive search for the optimal split. The model's performance is affected differently by three parameters: K, n_{min}, and M. More specifically, K controls the strength of the attribute selection procedure, n_{min} controls the intensity of the output noise averaging, and the parameter M determines the extent of variance reduction achieved by the ensemble model aggregation. The parameters can be automatically (e.g., via cross-validation) or manually modified to the particularities of the situation. To reduce variance more effectively than weaker randomization methods, extra trees uses ensemble averaging paired with a random selection of the cut-point and attribute, which justifies this approach from a bias-variance perspective. Due to their increased randomness in feature selection and splitting, they tend to have higher model variance but lower model bias. Randomization effectively manages noisy data and mitigates the risk of overfitting. This method has the potential to attain favorable outcomes with fewer trees owing to their intrinsic randomness. Despite their significant tendency to overfit noisy data, extra trees might be more adept at handling complex patterns in the data. The main advantage of this method is its computational efficiency, which goes hand in hand with its precision [41]. Extra trees distinguish themselves from other tree-based methods by their approach to node separation. While other tree-based models use random cut-points, extra trees utilize the complete learning sample. By aggregating the outcomes of many decision trees and employing either an arithmetic mean (for regression tasks) or a majority vote (for classification tasks), a conclusive prediction can be obtained [42].

2.6. Support vector regressor (SVR)

A lot of research has been done on the support vector regressor (SVR) in many different applications, such as regression, anomaly detection, and classification. It functions according to three basic principles: minimizing the likelihood of misclassifying test data, accounting for unobserved data in the model, and lessening the number of samples drawn from an unspecified probability distribution [9]. Boundaries are used by the SVR for both regression and classification. It splits or regresses data points using borders that are similar to streets. A hyperplane is a line that runs in the middle of the street at a constant distance from each side. Boundary lines are the sides of the street. A street's width is dependent on its margin. To handle nonlinear problems, the SVR uses several economical and computationally efficient kernel transformations. It is possible for SVR's kernel functions to extract both linear and nonlinear relationships from data. Radial basis function (RBF), sigmoid, linear, polynomial, and custom kernels are examples of kernel functions. The unique characteristics of the data and the specific task at hand must be taken into consideration while choosing a kernel. A kernel transformation is used to convert the data collection into a higher dimensional space in order to accomplish linear separation. Unlike classical regression methods that focus on enhancing accuracy and precision by means of the "best fit" of the data, SVR employs a fixed error threshold (ϵ). The solution provided in Equation (4) is used in SVR training.

$$\begin{aligned} &\textit{minimize}\ \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{n}\left(\xi_i^* + \xi_i\right) \\ &\textit{subject to}\ y_i < w, x_i \succ b \leq \in +\xi_i^*; < w, x_i > +b - y_i \leq \in +\xi_i \end{aligned} \tag{4}$$

Where w stands for the weight vector that is learned during training, C stands for the regularization parameter, x_i represents the *i-th*

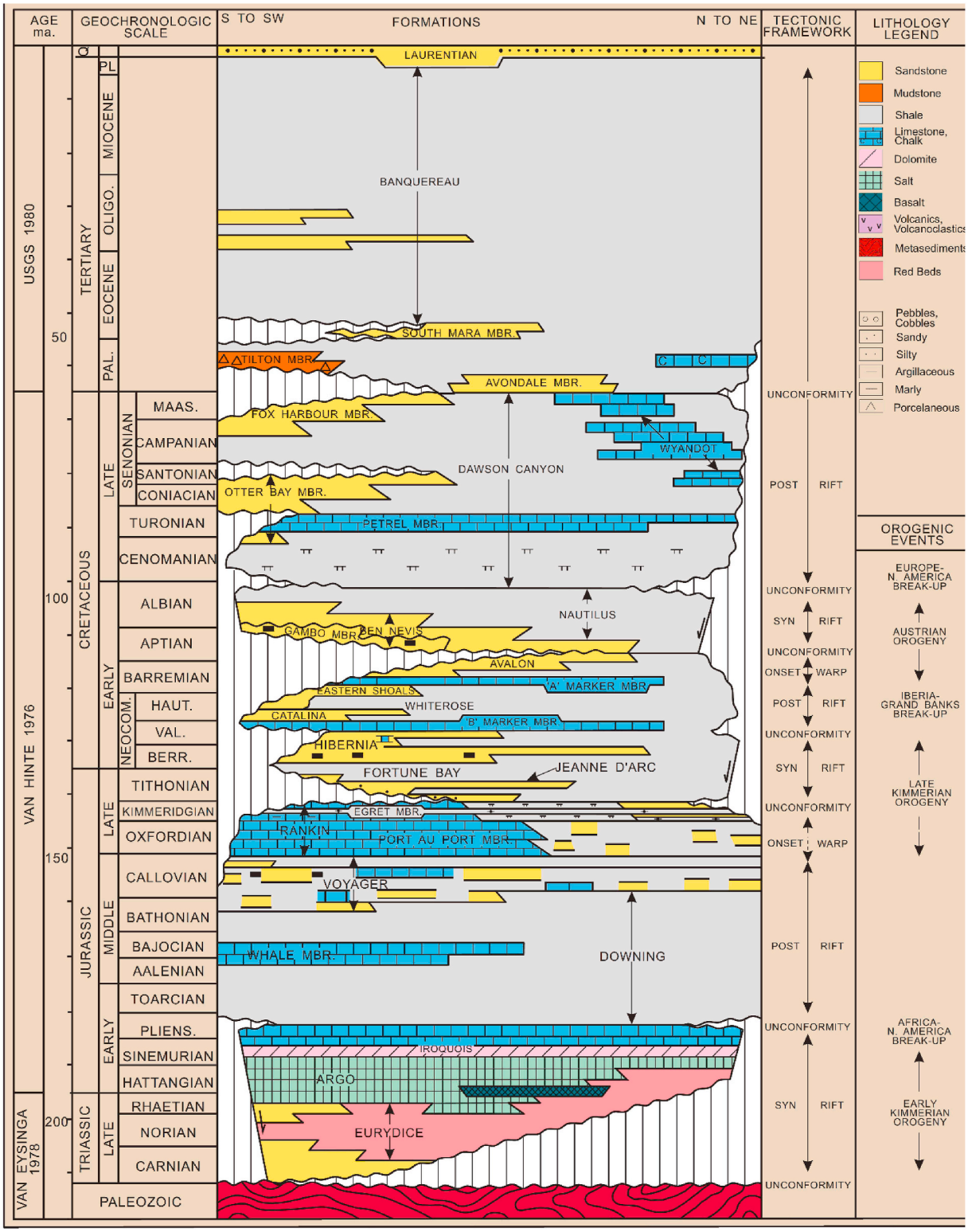


Fig. 1. Litho-Stratigraphic description of Jeanne D'arc Basin (modified from Sinclair & Canada [43]).

training event, y_i indicates the training label, and ξ_i denotes the distance between the decision boundary and the predicted values outside the boundary.

3. Materials and methods

3.1. Experimental data

A total of 197 core plugs with a diameter of 38.1 mm are collected from a well in the Jeanne d'Arc sedimentary basin and analyzed in a laboratory to gather petrophysical characteristics. The well is located in a clastic reservoir dominated by quartz, consisting mainly of fine-grained sandstone with a small amount of shale. This reservoir was formed in a marine shoreface depositional environment. The litho-stratigraphic information is presented in Fig. 1. During the Late Triassic to Middle Jurassic period, extension processes shaped the fill of the Jeanne d'Arc basin, characterized by continental red beds, evaporates, and carbonates. These formations are overlain by marine mudstones, shallow marine sandstones, and the diverse lithology of the Rankin Formation. This geological unit encompasses a basin with diverse features and contains the prolific Egret Member, which was formed in a low energy-restricted marine environment. The second mega sequence, delineated by a sequence boundary, includes river-to-marine deposits from the Kimmeridgian and Tithonian periods in the Jeanne d'Arc Formation. These deposits are subsequently overlain by the Fortune Bay and Hibernia Formations.

The Jeanne d'Arc Formation consists of coarse-grained conglomeratic fluvial deposits. These deposits are organized into eight phases of sedimentation. The subsequent layer, known as the Fortune Bay Formation, is composed of marine shales and siltstones that were deposited in offshore environments. These formations exhibit dynamic stratigraphy, which unveils tectonic events and regional regressions. In addition, the Eastern Shoals Formation is composed of shallow to marginal-marine calcareous sandstone and oolitic limestone. It is followed by the Avalon Formation, which is characterized by marine and marginal-marine deposits that coarsen upward, completing the basin's sedimentary record. In summary, the sedimentary record in the Jeanne d'Arc basin reflects dynamic tectonic events, leading to distinct mega-sequences with varied depositional environments and lithologies [43].

In the present study, laboratory-derived porosity (φ), grain density (ρ_{gr}), oil saturation (S_O), and water saturation (S_W) are used as input features, whereas permeability (K) is selected as output level. Depth has also been added as an input feature because the physical properties of the rocks in some reservoirs are strongly influenced by depth [2].

The statistical information on the investigated data applied in this study is displayed in Table 3. The observed dataset exhibits significant variations in magnitudes across individual samples, which can be attributed to the complex structure of the reservoir. The degree to which the input variables and target value correlate with each other is tabulated in Table 4. Table 4 illustrates that the input variables have a moderate to significant influence on permeability modeling, which justifies the use of these features as inputs.

A random process has divided the original data into training (80 %) and testing (20 %) data. After the model is trained, it is tested using unseen data to evaluate the developed model's success. The Python programming environment is used for all the programming tasks associated with this study. The computation is facilitated by customized computer code.

3.2. Cross validation

Cross-validation is a statistical technique utilized in statistical modeling to ensure the precision of prediction and modeling procedures. It can be implemented in different ways: K-fold cross-validation, Leave-one-out cross-validation, Random Subsampling, Stratified K-fold cross-validation, etc.

In K-Fold Cross-Validation, the dataset is divided into approximately equal-sized folds or subsets. The model is trained K times, each time with K-1 folds as training data and the remaining folds as validation data. As a result of the various dataset experiments, all data points are ultimately utilized for both training and testing, which is an advantage of K-Fold Cross-validation. Fig. 2 represents the typical five-fold cross-validation diagram [44].

Leave-one-out cross-validation (LOOCV) uses K folds equal to the number of samples. Each iteration validates one data point and trains the model on the remaining N-1 data points. These stages are repeated N times, where N is the dataset's sample count. Random subsampling is done by dividing the original dataset into two parts: train and test. The common ratios include 70–30, 80–20, or

Table 3
Statistical summary of core data.

Parameters	Depth	φ	ρ_{gr}	S_w	S_o	K
count	197	197	197	197	197	197
mean	2394.992	0.153695	2668.376	0.303915	0.199554	60.33086
max	2450.1	0.255	2880	0.73	0.389	724
min	2344	0.017	2640	0.025	0.041	0.01
25 %	2376.7	0.123	2650	0.209667	0.157	0.54
50 %	2392.7	0.157	2660	0.287	0.199667	6.34
75 %	2413	0.199	2680	0.432	0.241	54.4
STD	28.8948	0.05541	27.20	0.136808	0.062382	130.355
Skewness	0.17070	−0.55055	3.62	0.5619	−0.037836	3.2148
Kurtosis	−0.71110	−0.05125	21.70	−0.062423	0.058507	10.6483

Table 4
Pearson correlation matrix for input-output features.

	Depth	φ	ρ_{gr}	S_w	S_o	K
Depth	1	-	–	–	–	–
φ	0.265404	1	–	–	–	–
ρ_{gr}	−0.353625	−0.446268	1	–	–	–
S_w	−0.505329	−0.310288	0.222499	1	–	–
S_o	−0.129731	0.083848	0.098782	0.094245	1	–
K	0.386863	0.576862	−0.22873	−0.385444	0.256786	1

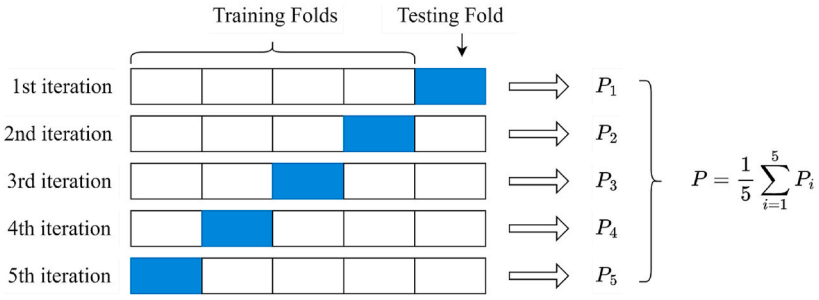


Fig. 2. A schematic diagram of five-fold cross-validation.

depending on the size of the dataset. Random subsampling ensures external prediction and provides confidence in diverse datasets with the same parameters [45].

3.3. Data transformation

Regression and machine learning approaches require statistically well-dispersed predictor and criterion variables with low measurement and instrument errors. Some machine learning models, such as SVR and MVR, can function efficiently with standardized data. Hence, in these instances, the data have been standardized by utilizing the mean as well as the standard deviation. Building the best models with the least amount of error, the highest level of accuracy, and the greatest degree of generalizability requires standardized data [46,47]. A standard normal distribution is achieved through Z-score standardization. It scales the data to one standard deviation and zero mean by subtracting the mean and dividing it by the standard deviation. This compares variables with different units or scales. Outliers and linear relationships can be identified using Z-scores, which represent how many standard deviations a data point is from the mean. Statistics and machine learning use z-score standardization to ensure consistency and comparability. Equation (5) is used in the data preprocessing stage to standardize all laboratory-derived petrophysical properties. Because of data standardization, all inputs have zero means and one standard deviation [47–50].

$$x' = \frac{x - \bar{x}}{\sigma} \tag{5}$$

Table 5
Model evaluation metrics.

Statistical Parameter	Mathematical Equation	Nomenclature
Coefficient of determinant (R^2)	$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(Y_i - \overline{y})^2}$	$\widehat{Y_i}$ stands for the predicted output of i-th sample, Y_i refers to the corresponding actual value of the i-th sample, $\widehat{Y}$ represents the estimated output for the target, and Y indicates the corresponding actual output. N is the number of observations in the sample, and p is the number of predictors (independent variables) in the model.
Adjusted R^2	Adjusted $R^2 = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$	
Mean Absolute Error (MAE)	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert Y_i - \widehat{Y_i}\rvert$	
Mean Squared Error (MSE)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2$	
RMSE	$\sqrt{MSE}$	
Mean Absolute Percentage Error (MAPE)	$MAPE = \frac{1}{n}\sum\frac{\lvert Y_i - \widehat{Y_i}\rvert}{Y_i}$	
Maximum Error	$maxE = max(\lvert Y_i - \widehat{Y}_i\rvert)$	
Minimum Error	$minE = min(\lvert Y_i - Y\rvert)$	

where, x' refers to the standardized data, $\bar{x}$ stands for the mean value of dataset, x denotes each value in the dataset, and σ represents the standard deviation of the data.

3.4. Comparison of statistical parameters

The objective of this study is to evaluate and confirm the effectiveness of newly developed bagging tree, decision tree, and extra trees models, as well as to effectively examine numerous previously used models such as MVR, SVR, and random forest. The statistical performance metrics used to evaluate the performance of the methods studied here include the coefficient of determination (R^2), mean-squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), maximum-error (maxE), and minimum-error (minE). Table 5 provides a mathematical representation of these popular and often used statistical indicators.

3.5. Permeability prediction using MVR and tree-based approaches

Based on laboratory-derived core data, permeability is predicted in this study using statistical models like MVR and tree-based ensemble techniques, such as decision tree, bagging tree, extra trees, and random forest. Core features such as water saturation (S_W), grain density (ρ_{gr}), oil saturation (S_O), depth, and porosity (φ) are used as inputs in regression and machine learning techniques, while the reservoir's permeability (K) is used as an output.

The GridsearchCV is used to identify the ideal set of parameters from a range of hyperparameter combinations during the training phase. For every tree-based model, the optimum hyperparameter values are listed in Table 6. Fig. 3 illustrates a typical tree-based modeling procedure. At first, the acquisition of essential data and the subsequent preprocessing operations establish the basis for further computing. Subsequently, meticulous selection of inputs and outputs, combined with dataset partitioning, guarantees the integrity of the model's training procedures. The iterative process of selecting models, modifying parameters, and performing cross-validation enables the development of a prediction framework, which in turn allows for the evaluation of statistical metrics to assess performance. Ultimately, the model is exported to generate predictions for novel instances, enabling its use in real-world scenarios.

3.6. Permeability prediction using support vector regressor (SVR)

This study develops SVR models using a variety of kernel transformations, including linear, Gaussian, polynomial, and radial basis functions. The optimization of the model is conducted through different runs in both the training and validation stages. In the validation stage, an accurate SVR model is built using a K-value of ten cross-validations and appropriate C, gamma, and kernel values. First, it is important to choose the right kernel to determine the optimal values for these parameters. In SVR, "kernel values" refer to the results of a specific kernel function applied to pairs of data points. These values represent the relationships or similarities between data points in a higher-dimensional space, revealing complex patterns within the data. Common kernels, such as linear, polynomial, and Gaussian, transform input data to help identify underlying structures. The SVR algorithm uses these kernel values to build a regression model that best fits the data while taking into account the accuracy-smoothness trade-off.

The values for *gamma* and *C* require an adjustment after the selected *kernel*. The present work uses a grid search technique to find the best values for the *C*, *gamma*, and *kernel* parameters. The ideal values of the hyperparameters for *C*, *gamma*, and *kernel* are displayed in Table 6. A step-by-step schematic illustrating the SVR modeling is presented in Fig. 4. The approach used in the SVR flowchart is

Table 6
The optimum hyperparameter values for SVR and tree-based models.

Algorithm	Hyperparameter	Optimized Value
Random Forest	n_estimators	54
	Criterion	absolute_error
	max_features	4
	max_depth	5
	max_leaf_nodes	2
	min_weight_fraction_leaf	0
	min_samples_split	2
Decision Tree	max_depth	5
Bagging	n_estimators	100
	max_samples	0.8
Extra Trees	n_estimators	100
	Criterion	absolute_error
	max_features	None
	max_depth	5
	min_samples_leaf	1
	min_samples_split	2
SVR	C	900
	Kernel	rbf
	Gamma	auto

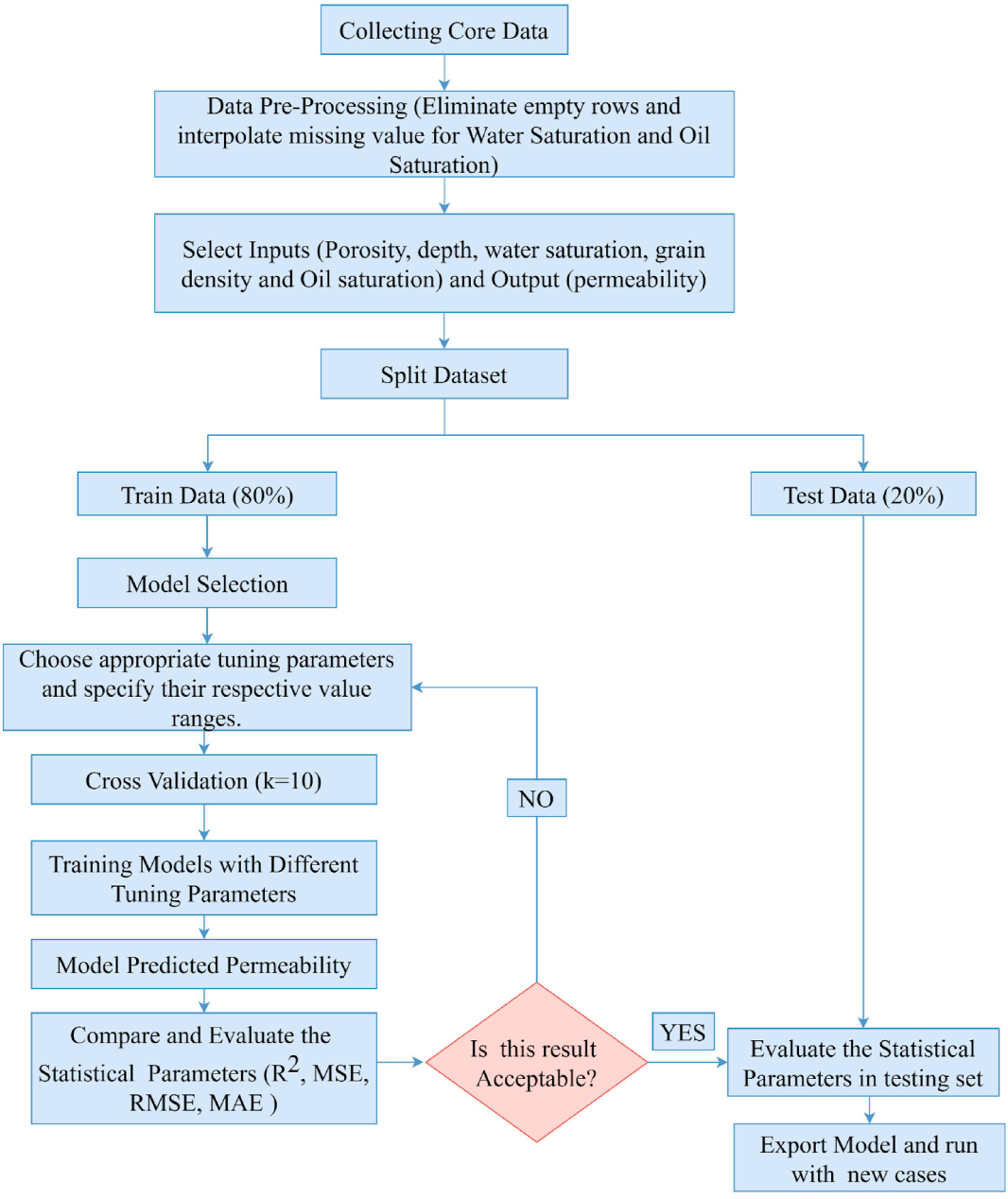


Fig. 3. Key steps to determine permeability from core data using tree-based modeling.

comparable to that of tree-based modeling. An important difference between SVR and tree-based modeling is that, in the case of SVR, the dataset is standardized. While both SVR and tree-based modeling involve gridsearch pipelines, the tuning parameters vary among models.

3.7. Ranking of core features

Laboratory-derived core features are ranked based on their significance in permeability modeling. Ranking core features helps determine which input variables or features most affect permeability prediction. Numerous input features add complexity to a machine learning model. Sometimes, some of the input features don't add any significance to the model. Moreover, it increases training time. For this reason, input feature numbers are reduced using different dimensionality reduction techniques such as principal component analysis or linear discriminant analysis. Feature ranking helps reduce input feature numbers by carefully discarding low-significant features. It plays a critical part in reducing dimensionality in the model.

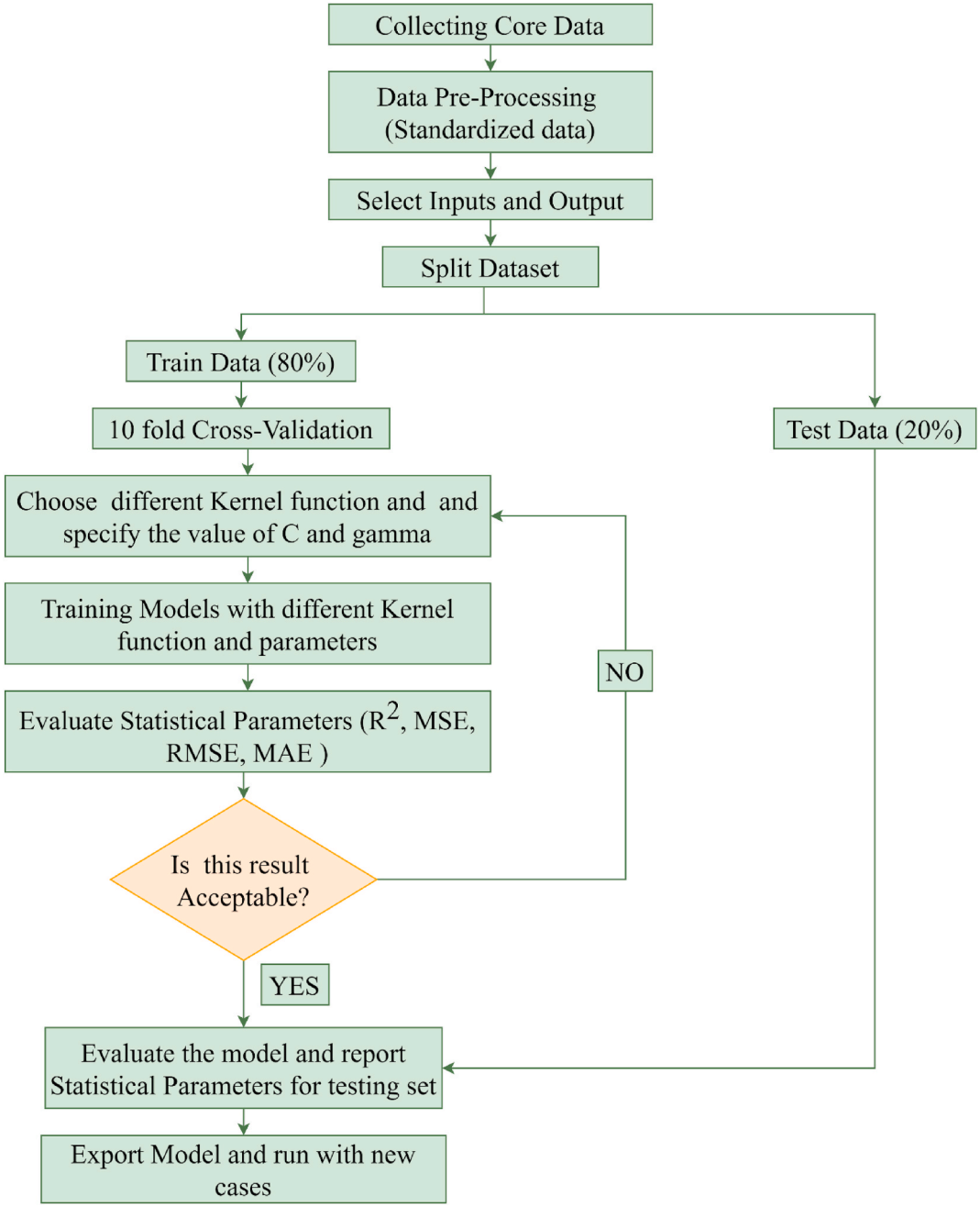


Fig. 4. Key steps to determine permeability from core data using SVR modeling.

This study employs two distinct ways of ranking: algorithmic and manual. Both strategies make use of the extra trees predictive model. One input parameter at a time is eliminated to build the model using the manual approach. To accommodate the differences in the distribution of data, the model is run sixteen times using distinct random states. The next step is to rank using the R^2 values, assigning a rank of 1 to the lowest value, a rank of 2 to the second lowest value, and so on. MSE assigns a quantitative value of 1 to the value with the largest magnitude, a value of 2 to the value with the second highest magnitude, and subsequent values are assigned in ascending order. MAE, MAPE, and maxE go through the same procedure. All the ranks are added up to determine the cumulative rank, with a lower value having a more significant influence. This procedure is repeated to rank all the input variables.

The algorithmic method is an impurity-based strategy in which bootstrapped samples and randomized feature selection are used by extra trees to produce numerous decision trees. The impurity measure, also known as entropy or Gini impurity, assesses each feature's impact on model accuracy. The feature importance is then calculated by adding up the impurity reduction that each feature in the

ensemble of trees achieves. Finding the difference between the impurity of the parent node and the weighted sum of the impurities in the child nodes is the way of computing the impurity decrease of a feature.

4. Results and discussion

4.1. Results

4.1.1. Permeability prediction using MVR model

Pearson correlation evaluates the degree and direction of a linear relationship between two continuous variables. Table 4 depicts the matrix argument values for each pair of variables' correlation, which vary from −1 to 1. A positive correlation coefficient represents a positive association between two variables. If the correlation coefficient is negative, the variables are negatively correlated.

Porosity (φ), depth, and oil saturation (S_O) are positively correlated with permeability while water saturation (S_W) and grain density (ρ) have negative correlation coefficient with permeability (Table 4). These findings are supported by the scatter plots of predictor variables versus the criterion variable (Fig. 5). As a result, it can be said that the following factors are important input variables for permeability prediction: depth, φ, ρ_{gr}, S_W, and S_O.

Equation (6), which is the best-correlated MVR equation, is obtained by the stepwise procedure. Table 7 provides a summary of the MVR model's predicted performance parameters for equation (6). The coefficient of the determinant (R^2) for the MVR model is 0.63, suggesting that the predictive ability of the MVR is not significant. The R^2 value for the MVR model is significantly lower than that of all the other models examined. Additionally, the remaining statistical parameters (error metrics) for MVR are exceptionally high, indicating that MVR is not a reliable model for predicting permeability based on core data. In some cases, the model predicted a negative value of permeability, which is unacceptable in reservoir characterization.

$$K = 54.65 - 20.70 \times S_w + 56.35 \times \varphi + 24.17 \times depth + 26.86 \times S_o + 4.70\rho_{gr} \quad (6)$$

4.1.2. Permeability prediction using decision tree (DT), bagging tree (BT), random forest (RF), extra trees (ET), and support vector regressor (SVR)

The R^2 values for the testing datasets of SVR and all tree-based models (DT, BT, RF, and ET) are significantly closer to 0.95 (Table 7 and Fig. 6). These models also have very low R^2 variances. While MAPE and minE exhibit substantial fluctuations, other performance metrics like MSE, RMSE, MAE, and maxE display a moderate degree of variance, indicating a hierarchy of these models. The RMSE, MAE, and maxE results are higher for the DT and RF. Nevertheless, BT and SVR have lower RMSE and MAE values than ET, DT, and RF

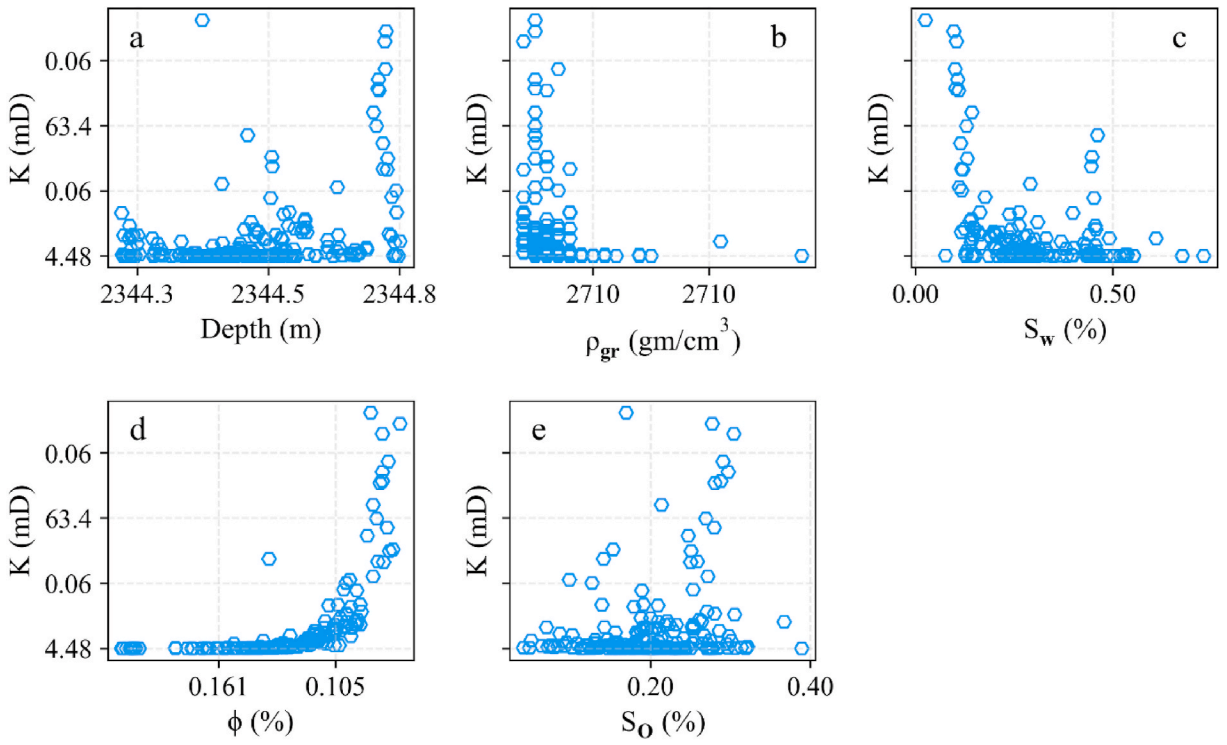


Fig. 5. Laboratory-derived core properties plotted against corresponding permeabilities. (a) K vs Depth. (b) K vs ρ_{gr}. (c) K vs S_w. (d) K vs φ. (e) K vs S_o.

Table 7
Statistical index of prediction efficiency for MVR, Tree-based models, and SVR.

Models	R^2	Adjusted R^2	MAE	MSE	RMSE	MAPE	maxE	minE
MVR	0.613	0.549	59.017	8421.679	91.77	341.81	578.247	1.132
DT	0.942	0.941	24.773	1721.222	41.488	0.602	120	0
BT	0.964	0.96	10.647	407.953	20.198	20.937	81.792	0.006
RF	0.961	0.957	16.988	1026.627	32.041	3.981	125.136	0.027
ET	0.976	0.974	9.186	253.527	15.923	2.625	50.506	0.028
SVR	0.951	0.942	15.724	625.647	25.013	17.215	85.734	0.037

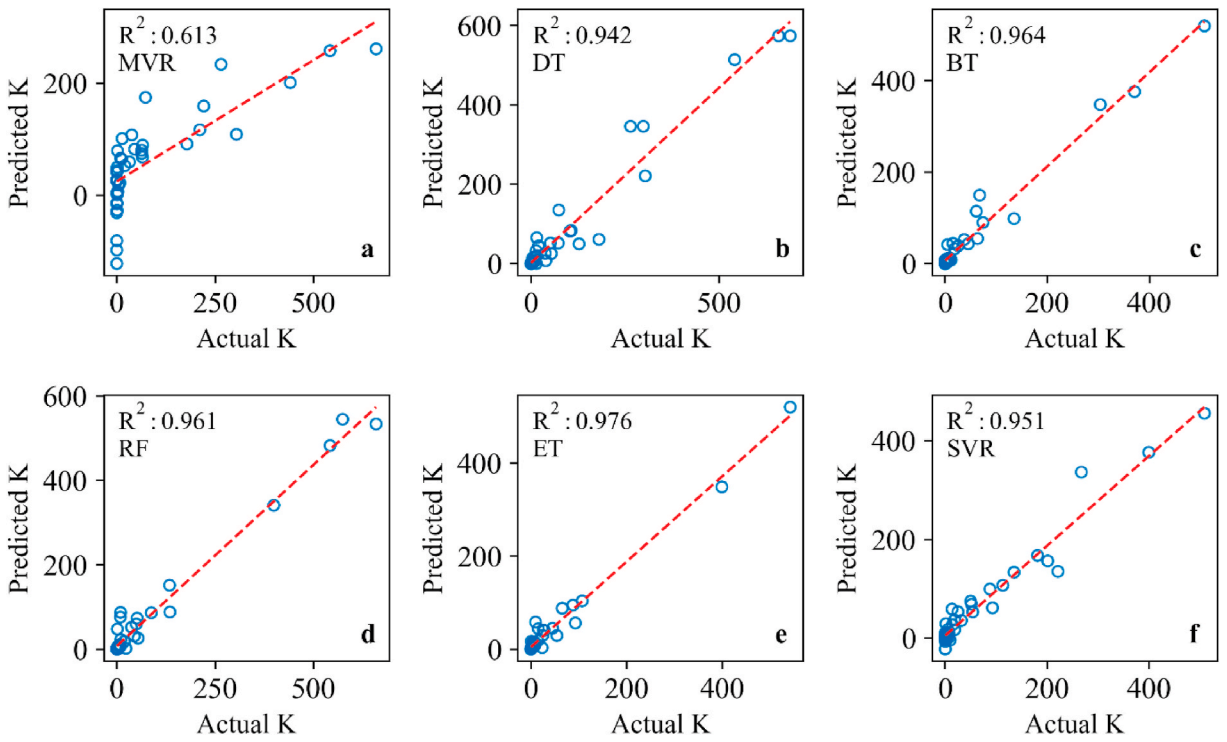


Fig. 6. Graphical presentation of R^2 for testing data (a: MVR, b: DT, c: BT, d: RF, e: ET, f: SVR).

but have higher MAPE values. ET, BT, and RF outperform other models on all statistical measures. The order of accuracy (higher to lower) of the SVR and tree-based models is described as follows based on the MSE, MAE, MAPE, maxE, and minE scores: ET, BT, RF, DT, and SVR.

4.1.3. Investigation of K-fold cross validation

The performance of the Extra Trees model was assessed using the K-fold cross-validation technique, employing various numbers of folds (K) including 3-fold, 5-fold, 7-fold, 9-fold, and 10-fold. The R^2 testing value indicates that the majority of values are approximately 0.97. The most effective configuration is the ten-fold cross-validation, with a remarkable R^2 of 0.976, whereas 9-fold witnessed the lowest value among all folds (Table 8).

The maximum and minimum absolute errors verify the model's consistency across multiple folds. The findings reveal that the model has good predictive power and ability to be generalized across diverse subsets of data. The order of accuracy (higher to lower) of the different k-fold based on the R^2, adjusted R^2, MSE, MAE, MAPE, maxE, and minE scores: 10-fold, 5-fold, 3-fold, 7-fold, and 9-fold.

4.1.4. Input features ranking

The results of the manual ranking method are listed in Table 9, while Fig. 7 displays the results of the algorithmic ranking. For permeability prediction, manual and algorithmic ranking suggest that the most important features are φ and S_W, whereas the least important variables are ρ_{gr}, S_O, and depth. Feature ranking can help to determine the permeability quickly and accurately using a few core features, reducing the need for labor-intensive, costly, and time-consuming laboratory work.

Table 8
Performance metrics of Extra Trees with different K-fold cross validation.

K	R^2	Adjusted R^2	MAE	MSE	MAPE	RMSE	maxE	minE
03	0.974	0.973	13.270	553.674	3.789	23.530	94.159	9.7E-17
05	0.975	0.974	12.541	530.556	4.327	23.034	83.000	9.7E-17
07	0.970	0.969	13.591	629.199	4.774	25.084	87.288	9.7E-17
09	0.968	0.966	13.455	685.000	5.055	26.172	113.818	9.7E-17
10	0.976	0.976	9.186	253.527	2.625	15.923	50.507	2.8E-02

Table 9
Core features ranking using manual approach.

Excluded parameter	Performance metrics							Ranking					Total Rank
	R^2	MAE	MSE	MAPE	EV	maxE	MinE	R^2	MAE	MSE	MAPE	MaxE	
φ	0.58	39.53	6989.93	89.67	0.59	374.02	0.13	1	1	1	1	1	4
S_w	0.79	22.81	3624.42	4.86	0.80	296.23	0.01	2	2	2	4	2	10
ρ_{gr}	0.82	22.00	3264.10	6.59	0.82	271.35	0.01	3	5	3	3	3	14
S_o	0.82	22.05	3143.60	7.70	0.83	256.61	0.00	4	4	4	2	4	14
Depth	0.85	22.54	2614.84	3.66	0.85	229.33	0.00	5	3	5	5	5	18

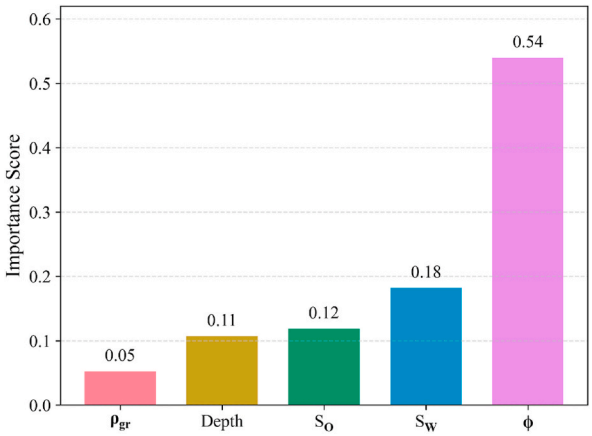


Fig. 7. Core features ranking using algorithmic approach.

4.2. Discussions

4.2.1. Models' proficiency and shortcomings

In order to enhance the estimation accuracy, the current work formulates machine learning models—bagging tree, decision tree, and extra trees—for the first time in the permeability modeling of petroleum reservoirs based on core analysis. Another scientific novelty of this work is the examination of the effects of water and oil saturations on a core's permeability. This research also ranks core features according to their importance in permeability modeling.

Each model's prediction performance is assessed using the statistical indices that are compiled in Table 5. A list of these indices for the models under study is tabulated in Tables 7 and 8. A comprehensive comparison of the models is illustrated in Fig. 8. It has been noted that MVR performs incompetently. While the R^2 displays similar trends, the RMSE, MAE, and maxE exhibit considerable variations in DT, BT, RF, ET, and SVR. The MAPE is one of the most important indicators in statistical analysis. The higher value of MAPE experienced by BT and SVR indicates that these are less accurate models. The RF and ET demonstrate a moderate degree of minE, however the SVR has a noticeably higher minE. The DT and BT approaches result in a minimal minE.

A total of 197 data points plotted in terms of depth in Fig. 9 comprehensively demonstrates the comparison of the actual and predicted permeability values for studied statistical and machine learning models. All models except MVR correctly predicted the permeability value.

The prediction performance of MVR is relatively poor, and MVR also predicted negative permeability value in several instances. Permeability is almost precisely predicted by DT, BT, and RF. The overall accuracy ratings of tree-based modeling approaches indicate that these models could be viable candidates for precisely characterizing a reservoir.

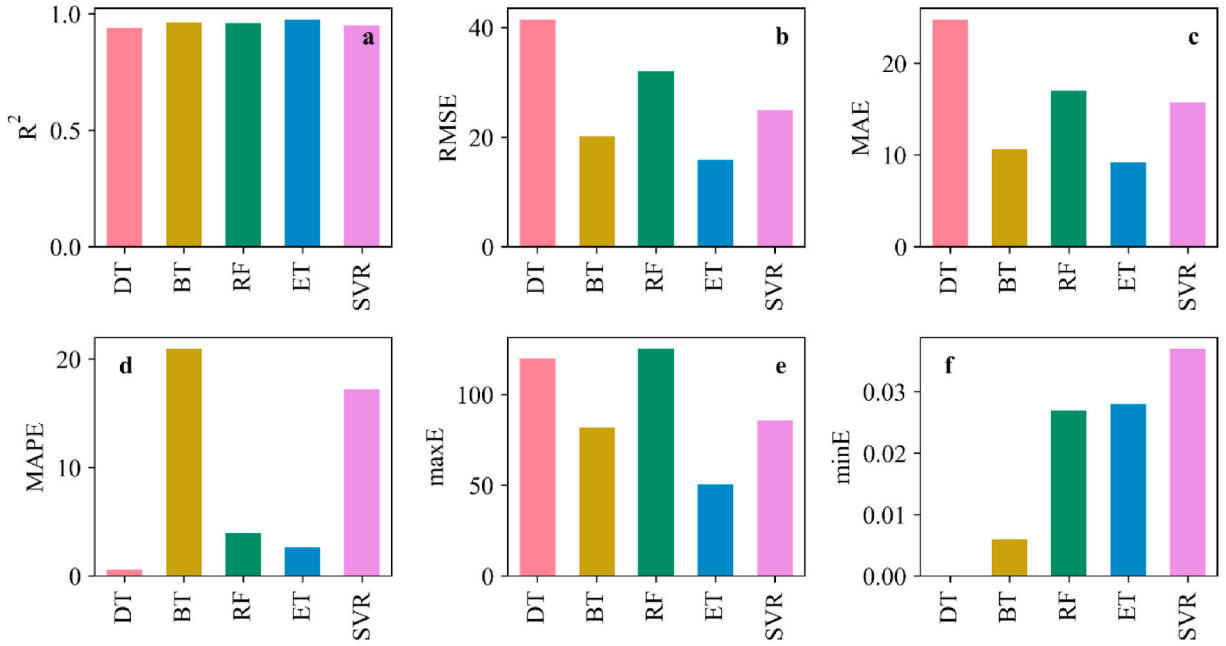


Fig. 8. Statistical indices of Decision Tree, Bagging Tree, Random Forest, Extra Trees, and SVR (a: R^2, b: RMSE, c: MAE, d: MAPE, e: maxE, f: minE).

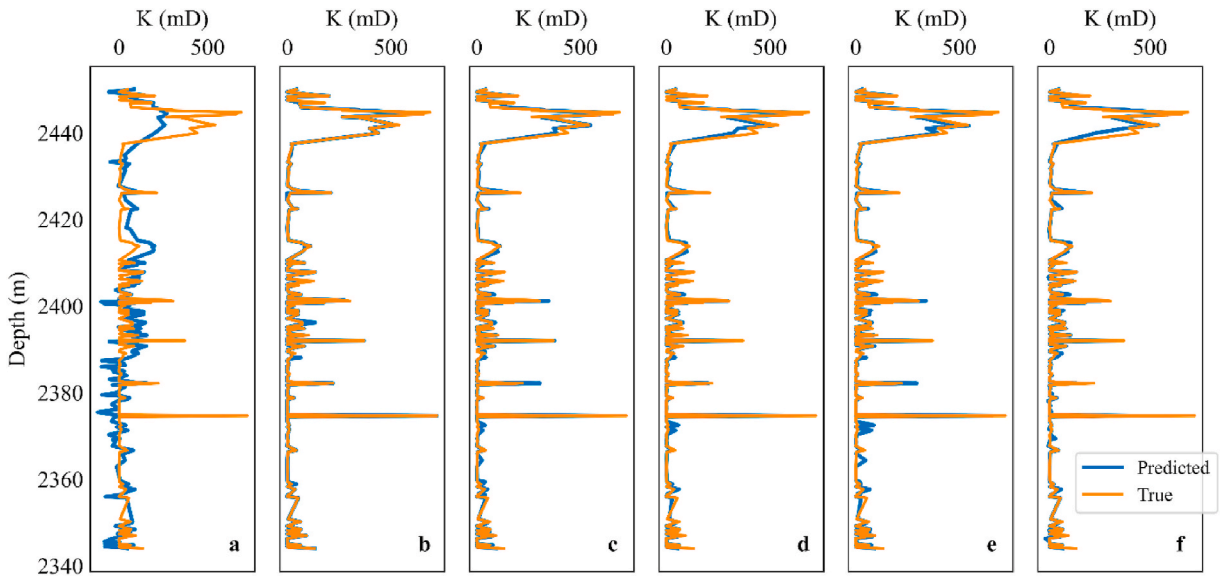


Fig. 9. Comparison of the actual and predicted permeability of MVR, DT, BT, ET, RF, and SVR (a: MVR, b: DT, c: BT, d: ET, e: RF, f: SVR).

4.2.2. Feature importance and cross-validation

In order to forecast permeability, engineers and operators can save time and effort by using the variable ranking to help them choose a few features instead of doing a whole core analysis. The input parameters ranking indicates that the variables ρ_{gr}, S_O, and depth are the least relevant for permeability prediction, whereas φ and S_W are the most significant inputs. Cross-validation investigation suggests that the value of K in k-fold cross-validation is chosen based on different information such as the size of the dataset, computing efficiency of k-fold cross-validation, etc.

4.2.3. Significance of this study

The developed DT, BT, and RF can help researchers effectively characterize heterogeneous petroleum reservoirs. Furthermore, with a few laboratory-derived core features, the proposed method can help researchers determine the permeability quickly and accurately, reducing the amount of work in the lab and the overall cost of the experiment. This work will probably pave the way for the use of

computer-based machine-learning techniques in petroleum engineering research. Rather than relying solely on tried-and-true empirical algorithms and laboratory-intensive works, novel machine learning models will be applied in evaluating a reservoir's rock-fluid properties and quality, doing an economic analysis, and deciding on drilling operations, field development, and reservoir management.

5. Conclusion

Six machine learning algorithms, including MVR, decision tree, bagging tree, random forest, extra trees, and SVR, are used in this work to predict permeability from core data. Among these, in core data-based permeability modeling, decision tree, bagging tree, and extra trees are developed for the first time. Examining how oil and water saturations affect a core's permeability and prioritizing core features in permeability modeling are two more scientific novelties. The following are the investigation's main findings.

1. The MVR cannot be used to forecast core data-based permeability since it is too unreliable. Every model-extra trees, bagging tree, random forest, SVR, and decision tree-developed in this study outperformed the MVR.
2. The random forest, bagging tree, and extra trees methods accurately predict permeability. With an R^2 of 0.976, extra trees exhibits the highest prediction accuracy, whereas random forest and bagging tree demonstrate slightly lower R^2 values of 0.961 and 0.964, respectively. Based on performance criteria, the investigated approaches are ranked as follows: extra trees, bagging tree, random forest, SVR, decision tree, and MVR.
3. The performance of the 10-fold cross-validation is superior to that of the 3-fold, 5-fold, 7-fold, and 9-fold cross-validations when extra trees are used. This reveals that employing a higher number of folds for validation might be advantageous in this scenario.
4. Based on the ranking of the input parameters, it is evident that the most significant inputs for permeability prediction are φ and S_w. On the other hand, the variables ρ_{gr}, S_O, and depth have the least significant importance.

The results indicate that the optimized random forest, bagging tree, and extra trees algorithms are good choices for permeability prediction. Regretfully, there are no set standards for choosing the right model. To find the best machine-learning model for this task, more investigation is required. Potential future strategies could involve adding advanced deep learning models and integrating core data from different locations.

Data availability statement

Data will be made available on request.

CRediT authorship contribution statement

Amad Hussen: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Tanveer Alam Munshi:** Writing – review & editing, Writing – original draft, Formal analysis, Data curation, Conceptualization. **Labiba Nusrat Jahan:** Writing – review & editing, Funding acquisition. **Mahamudul Hashan:** Writing – review & editing, Writing – original draft, Project administration, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors thank the Shahjalal University of Science and Technology (SUST) Research Center and the Ministry of Science and Technology, Bangladesh (R&D project) for providing financial support to accomplish this research.

Abbreviations, variables, parameters, symbols and Greek letters

$\widehat{Y}_i$	the predicted value of i-th sample
Y_i	corresponding true value of the i-th sample
$\overline{Y}$	Mean value of Y_i
ξ_i	Distance between the boundary line and predicted values beyond the street
ξ_i^*	Slack variable
ANFIS	Adaptive neuro fuzzy inference system
ANN	Artificial neural network
CNN	Convolutional neural network
BT	Bagging Tree

DT	Decision Tree
DT	transit time
ET	Extra Trees
FDT	Fuzzy decision tree
FIS	Fuzzy interface system
GA	Genetic algorithm
MAE	Mean absolute error
MAPE	Mean absolute percentage Error
FZI	Flow Zone Indicator
GR	Gamma ray
GRNN	General regression Neural Network
K	Permeability
LLD	Latero-resistivity log deep
LLS	Latero-resistivity log shallow
maxE	Maximum Error
minE	Minimum Error
MSFL	micro-spherical focused resistivity log
MVR	Multiple Variable Regression
N	number of variables
NPHI	neutron porosity
R^2	Coefficient of determinant
Relu	Rectified linear unit
RF	Random Forest
RHOB	Bulk density
RMSE	Root Mean Squared Error
MSE	Mean squared error
MVR	Multiple Variable Regression
SVM	Support Vector machine
SVR	Support Vector Regressor
LM	Levenberg–Marquardt
S_o	Oil Saturation
SP	Spontaneous Potential
STD	Standard Deviation
S_w	Water Saturation
S_{wc}	Connate Water Saturation
ρ_{gr}	Grain density
σ	Standard Deviation
φ	Porosity
φ_N	Neutron Porosity
Ω	Regularization function
η	Learning rate
e	Residual error

References

[1] M. Mahdaviara, A. Rostami, F. Keivanimehr, K. Shahbazi, Accurate determination of permeability in carbonate reservoirs using Gaussian Process Regression, J. Petrol. Sci. Eng. 196 (2021) 107807.

[2] A. Subasi, M.F. El-Amin, T. Darwich, M. Dossary, Permeability prediction of petroleum reservoirs using stochastic gradient boosting regression, J. Ambient Intell. Hum. Comput. 13 (7) (2022) 3555–3564.

[3] J. Tian, C. Qi, Y. Sun, Z.M. Yaseen, B.T. Pham, Permeability prediction of porous media using a combination of computational fluid dynamics and hybrid machine learning methods, Eng. Comput. 37 (4) (2021) 3455–3471.

[4] R. Sander, Z. Pan, L.D. Connell, Laboratory measurement of low permeability unconventional gas reservoir rocks: a review of experimental methods, J. Nat. Gas Sci. Eng. 37 (2017) 248–279.

[5] H. Kaydani, A. Mohebbi, A. Baghaie, Permeability prediction based on reservoir zonation by a hybrid neural genetic algorithm in one of the Iranian heterogeneous oil reservoirs, J. Petrol. Sci. Eng. 78 (2) (2011) 497–504.

[6] W.J. Al-Mudhafar, Integrating lithofacies and well logging data into smooth generalized additive model for improved permeability estimation: zubair formation, South Rumaila oil field, Mar. Geophys. Res. 40 (3) (2019) 315–332.

[7] H. Gamal, S. Elkatatny, Prediction model based on an artificial neural network for rock porosity, Arabian J. Sci. Eng. 47 (9) (2022) 11211–11221.

[8] R. Bhattacherjee, K. Botchway, J.C. Pashin, G. Chakraborty, P. Bikkina, Machine learning-based prediction of CO2 fugacity coefficients: application to estimation of CO2 solubility in aqueous brines as a function of pressure, temperature, and salinity, Int. J. Greenh. Gas Control 128 (2023) 103971.

[9] A.F. Al-Anazi, I.D. Gates, Support vector regression for porosity prediction in a heterogeneous reservoir: a comparative study, Comput. Geosci. 36 (2010) 1494–1503.

[10] S.O. Olatunji, A. Selamat, A. Abdulraheem, Modeling the permeability of carbonate reservoir using type-2 fuzzy logic systems, Comput. Ind. 62 (2) (2011) 147–163.
[11] R. Gholami, A.R. Shahraki, M. Jamali Paghaleh, Prediction of hydrocarbon reservoirs permeability using support vector machine, Math. Probl Eng. 2012 (2012) 1–18.
[12] S.O. Olatunji, A. Selamat, A.A. Abdul Raheem, Improved sensitivity based linear learning method for permeability prediction of carbonate reservoir using interval type-2 fuzzy logic system, Appl. Soft Comput. 14 (2014) 144–155.
[13] M.-A. Ahmadi, M.R. Ahmadi, S.M. Hosseini, M. Ebadi, Connectionist model predicts the porosity and permeability of petroleum reservoirs by means of petrophysical logs: application of artificial intelligence, J. Petrol. Sci. Eng. 123 (2014) 183–200.
[14] S. Baziar, M. Tadayoni, M. Nabi-Bidhendi, M. Khalili, Prediction of permeability in a tight gas reservoir by using three soft computing approaches: a comparative study, J. Nat. Gas Sci. Eng. 21 (2014) 718–724.
[15] H. Kaydani, A. Mohebbi, M. Eftekhari, Permeability estimation in heterogeneous oil reservoirs by multi-gene genetic programming algorithm, J. Petrol. Sci. Eng. 123 (2014) 201–206.
[16] U. Santisukkasaem, F. Olawuyi, P. Oye, D.B. Das, Artificial neural network (ANN) for evaluating permeability decline in permeable reactive barrier (PRB), Environ. Processes 2 (2) (2015) 291–307.
[17] B. Rafik, B. Kamel, Prediction of permeability and porosity from well log data using the nonparametric regression with multivariate analysis and neural network, Hassi R'Mel Field, Algeria, Egyptian Journal of Petroleum 26 (3) (2017) 763–778.
[18] W.J. Al-Mudhafar, Integrating well log interpretations for lithofacies classification and permeability modeling through advanced machine learning algorithms, J. Pet. Explor. Prod. Technol. 7 (4) (2017) 1023–1033.
[19] A. Erofeev, D. Orlov, A. Ryzhov, D. Koroteev, Prediction of porosity and permeability alteration based on machine learning algorithms, Transport Porous Media 128 (2) (2019) 677–700.
[20] J.G. Urang, E.D. Ebong, A.E. Akpan, E.I. Akaerue, A new approach for porosity and permeability prediction from well logs using artificial neural network and curve fitting techniques: a case study of Niger Delta, Nigeria, J. Appl. Geophys. 183 (2020) 104207.
[21] D.A. Wood, Predicting porosity, permeability and water saturation applying an optimized nearest-neighbour, machine-learning and data-mining network of well-log data, J. Petrol. Sci. Eng. 184 (2020) 106587.
[22] F.A. Aljuboori, J.H. Lee, K.A. Elraies, K.D. Stephen, Using statistical approaches in permeability prediction in highly heterogeneous carbonate reservoirs, Carbonates Evaporites 36 (3) (2021) 49.
[23] S. Farouk, S. Sen, S.S. Ganguli, H. Abuseda, A. Debnath, Petrophysical assessment and permeability modeling utilizing core data and machine learning approaches – a study from the Badr El Din-1 field, Egypt, Mar. Petrol. Geol. 133 (2021) 105265.
[24] M. Hashan, T.A. Munshi, A. Zaman, L.N. Jahan, Empirical, statistical, and connectionist methods coupled with log variables ranking for the prediction of pore network permeability in a heterogeneous oil reservoir, Geomech. Geophy. Geo-Energy and Geo-Res. 8 (4) (2022) 117.
[25] M.I. Miah, M.A.N. Abir, Hybrid connectionist models to investigate the effects on petrophysical variables for permeability prediction, in: P. Vasant, I. Zelinka, G.-W. Weber (Eds.), Intelligent Computing & Optimization, 371, Springer International Publishing, 2022, pp. 647–656.
[26] W. Al-Mudhafar, Integrating bayesian model averaging for uncertainty reduction in permeability modeling. In *Offshore Technology Conference*, OTC, 2015, May. OTC-25646).
[27] Kabiru O. Akande, Taoreed O. Owolabi, Sunday O. Olatunji, A. AbdulRaheem, A hybrid particle swarm optimization and support vector regression model for modelling permeability prediction of hydrocarbon reservoir, J. Petrol. Sci. Eng. 150 (2017) 43–53.
[28] F. Anifowose, A. Abdulraheem, A. Al-Shuhail, A parametric study of machine learning techniques in petroleum reservoir permeability prediction by integrating seismic attributes and wireline data, J. Petrol. Sci. Eng. 176 (2019) 762–774.
[29] W.J. Al-Mudhafar, Bayesian and LASSO regressions for comparative permeability modeling of sandstone reservoirs, Nat. Resour. Res. 28 (1) (2019) 47–62.
[30] M. Zanganeh Kamali, S. Davoodi, H. Ghorbani, D.A. Wood, N. Mohamadian, S. Lajmorak, V.S. Rukavishnikov, F. Taherizade, S.S. Band, Permeability prediction of heterogeneous carbonate gas condensate reservoirs applying group method of data handling, Mar. Petrol. Geol. 139 (2022) 105597.
[31] H. Al Khalifah, P.W.J. Glover, P. Lorinczi, Permeability prediction and diagenesis in tight carbonates using machine learning techniques, Mar. Petrol. Geol. 112 (2020) 104096.
[32] T. Topór, Application of machine learning algorithms to predict permeability in tight sandstone formations, Nafta Gaz. 77 (5) (2021) 283–292.
[33] M. Mahdaviara, A. Larestani, M. Nait Amar, A. Hemmati-Sarapardeh, On the evaluation of permeability of heterogeneous carbonate reservoirs using rigorous data-driven techniques, J. Petrol. Sci. Eng. 208 (2022) 109685.
[34] E. Mohammadian, M. Kheirollahi, B. Liu, M. Ostadhassan, M. Sabet, A case study of petrophysical rock typing and permeability prediction using machine learning in a heterogenous carbonate reservoir in Iran, Sci. Rep. 12 (1) (2022) 4505.
[35] B. Balan, S. Mohaghegh, S. Ameri, State-of-the-art in permeability determination from well log data: Part 1- A comparative study, model development, Paper presented at the SPE Eastern Regional Meeting, Morgantown, West Virginia (September 1995), https://doi.org/10.2118/30978-MS.
[36] G. James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor, Tree-based methods, in: An Introduction to Statistical Learning: with Applications in Python, Springer International Publishing, Cham, 2023, pp. 303–335.
[37] L. Rokach, O. Maimon, Decision Trees. *Data Mining and Knowledge Discovery Handbook*, 2005, pp. 165–192.
[38] M. Bramer, Avoiding Overfitting of Decision Trees. *Principles Of Data Mining*, 2007, pp. 121–136.
[39] L.N. Jahan, T.A. Munshi, S.S. Sutradhor, M. Hashan, A comparative study of empirical, statistical, and soft computing methods coupled with feature ranking for the prediction of water saturation in a heterogeneous oil reservoir, Acta Geophys. 69 (5) (2021) 1697–1715.
[40] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[41] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (1) (2006) 3–42.
[42] E.E. Okoro, T. Obomanu, S.E. Sanni, D.I. Olatunji, P. Igbinedion, Application of artificial intelligence in predicting the dynamics of bottom hole pressure for under-balanced drilling: extra tree compared with feed forward neural network model, Petroleum 8 (2) (2022) 227–236.
[43] I.K. Sinclair, Petroleum resources of the Jeanne d'Arc basin and environs. *Grand Banks, Newfoundland* (No. 8), Geological Survey of Canada, 1992.
[44] M. Rahimi, M.A. Riahi, Reservoir facies classification based on random forest and geostatistics methods in an offshore oilfield, J. Appl. Geophys. 201 (2022) 104640.
[45] W.J. Al-Mudhafar, Incorporation of bootstrapping and cross-validation for efficient multivariate facies and petrophysical modeling, in: SPE Rocky Mountain Petroleum Technology Conference/Low-Permeability Reservoirs Symposium, 2016, May, pp. SPE–180277. SPE.
[46] V. Patel, A.J. Flisher, S. Hetrick, P. McGorry, Mental health of young people: a global public-health challenge, Lancet 369 (9569) (2007) 1302–1313.
[47] S. Mesroghli, E. Jorjani, Estimation of gross calorific value based on coal analysis using regression and artificial neural networks, Int. J. Coal Geol. 79 (2009).
[48] B. Demuth Howard, M.H. Beale, Neural Network Toolbox: for Use with MATLAB®, MathWorks, 2000.
[49] B. Wang, X. Wang, Z. Chen, A hybrid framework for reservoir characterization using fuzzy ranking and an artificial neural network, Comput. Geosci. 57 (2013) 1–10.
[50] R. Ashena, G. Thonhauser, Application of artificial neural networks in geoscience and petroleum industry, Artificial intelligent approaches in petroleum geosciences (2015) 127–166.