



Contrasting machine learning regression algorithms used for the estimation of permeability from well log data

Naman Khilrani¹ · Piyush Prajapati¹ · Atul Kumar Patidar¹

Received: 12 May 2021 / Accepted: 4 September 2021 / Published online: 26 September 2021
© Saudi Society for Geosciences 2021

Abstract

The oil and gas industry has slowly shifted its focus to a more data science-driven interpretation approach from the last decade. The petrophysical data analysis using advanced statistical and machine learning methods has been widely accepted due to reducing uncertainties and predicting more accurate data trends than conventional methods. The same approach is reflected in permeability estimation, where regression models are used during the well log data interpretation. This becomes necessary because amassing permeability data by other means is economically unfavorable and time-consuming. The exploration and production giants like Equinor employ variations on elementary methods like simple linear regression (SLR) to establish regression models with accuracies up to 0.98 R^2 scores. However, in recent years, various advanced machine learning algorithms have been developed that could be utilized in oil and gas data analysis and modeling with greater accuracy, coupled with thorough data cleaning and outlier removal. The current study demonstrates the application of modern machine learning algorithms to analyze drilling data from two wells of Equinor's Volve field located at the Norwegian Continental Shelf (NCS) and compare the outcomes with conventional analysis methods. The core data analysis of wells F-15/9-19A and F-15 B&BT2 is used, and a relationship is established between permeability and data obtained from wireline logging (prominently used variables) where a relationship can be observed consistently are porosity (PHIF) and shale volume (VSH). The goodness of fit of the correlations is thus obtained by calculating the " R^2 score," which gives an estimate about the accuracy of the regression models. Current study aims to compare the efficiency of four major regression algorithms, SLR, Lasso regression (LR), multiple linear regression (MLR), and support vector regression (SVR), in the estimation of Klinkenberg core corrected permeability (KLOGH) using porosity and shale volume. After performing a thorough cleaning and outlier filtering from the dataset, it was found that over 1000 iterations, SVR peaked when it came to R^2 score value (SVR, 0.88), while MLR performed the best on average (MLR, 0.77).

Keywords Machine learning · Klinkenberg core corrected permeability · Regression · Core data · R^2 score

Introduction

Digital modernization of the oil and gas industries, mainly in wireline logging techniques, has led to the accessibility of more reliable and accurate subsurface information rapidly and cost-effectively. Creating vital information by systematic investigation of significant datasets can assist in better decision-making. It also helps to improve operation efficiency,

reliability, and productivity of any project as cost reduction and quality assurance are the primary focus for any industry. The recent developments and digital boom have significantly improved data interpretation and visualization capabilities through advanced statistical, artificial intelligence (AI), and machine learning (ML) methods, which help us to find out the discrete or concealed information that has led to the more straightforward observation of trends hidden inside these massive datasets (Guan et al. 2019; Hong Li et al. 2020; Doveton and Prenskey 1992). These modern analytical techniques are rapid, cost-effective, and capable of retrieving more accurate information from enormous datasets available with any organization. The recent advancements in AI/ML techniques help data scientists and managers to drive meaningful insights from raw data and are exceptionally beneficial for the oil and gas

Responsible Editor: Santanu Banerjee

✉ Atul Kumar Patidar
atulpatidar@gmail.com; apatidar@ddn.upes.ac.in

¹ Department of Petroleum Engineering and Earth Sciences, University of Petroleum and Energy Studies, Dehradun, India

industry (Al-Bulushi et al. 2012; Alkinani et al. 2019; Zanjani et al. 2020). Petrophysical analysis using well log interpretation and its correlation with several other techniques like geological field data, core analysis, and production data can assist in reservoir rock characterization. The petrophysical properties such as porosity and permeability are a complex blend of chemical and mechanical origin sedimentary processes (lithification, diagenesis); therefore, they demand precise analysis of geological and geophysical characters of rocks (Otoo and Hodgetts 2020; Skalinski and Kenter 2015; Zhong et al. 2020; Yang et al. 2020). The estimation of shale volume in the reservoir reveals the amount of clay in the reservoir formation. The increasing shale amount indicates the reduction of porosity as well as the permeability of the reservoirs. The porosity in any reservoir rock is estimated as a fraction of voids over the total rock volume, while permeability is defined as a measure of gas/fluid flow through a porous medium (Singh 2019). Permeability is a function of porosity, shale volume, water saturation, and other reservoir properties (Adeniran et al. 2019). But the accurate prediction of this internal engineering property of the reservoir rock is difficult and requires variable sets of data to analyze the effects of reservoir facies heterogeneity in subsurface geological conditions. Generally, Darcy's law is used to measure the permeability of a drilled core (Wadsworth et al. 2020). This method gives a fairly accurate idea about the permeability of a core, but it is economically infeasible to extract a core from every drilled well of the field.

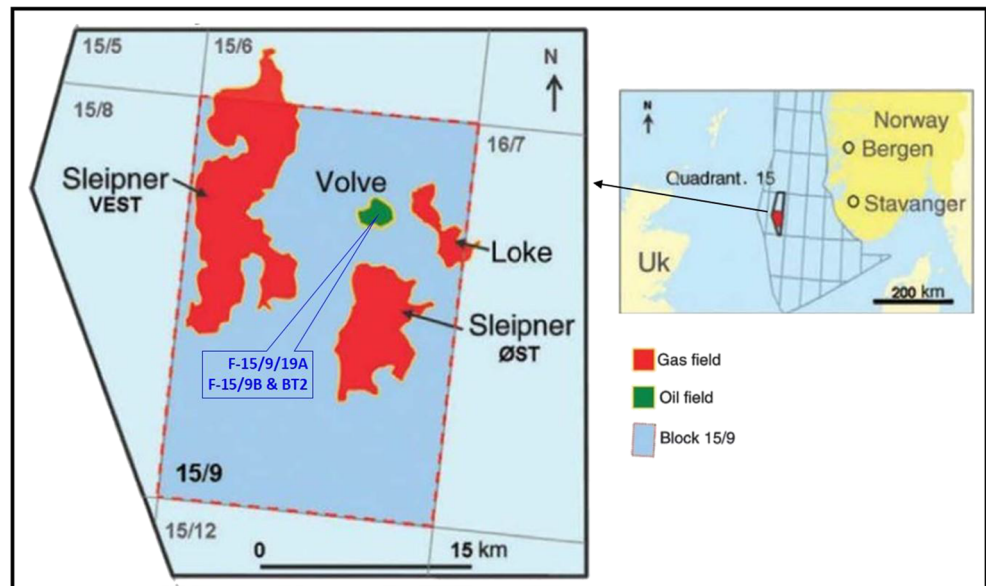
Previously published studies demonstrate the calculation of water saturation from well log interpretation for predicting permeability of the formation using conventional and the MLR technique only (Alger et al. 1963; Das and Chatterjee 2018; Morteza et al. 2014; Wendt et al. 1986). A range of statistical and computer-based algorithms like least squares support vector machine (LSSVM), imperialist competitive algorithm (ICA), and artificial neural network (ANN) were used in the investigation and prediction of permeability using well log data but mostly associated with errors during sensitivity analysis and low R^2 score (Mohaghegh et al. 1994). Additionally, most of the previous studies either concentrate on a single variable (water saturation) for making correlations or employ ambiguous and unclean datasets (Ahmadi and Chen 2019; Wood 2020). The approach can be improved by employing simple yet powerful algorithms on complete, thorough, and clean datasets that have been treated for outlier removal and return relevant results between dependent and independent variable correlations. The reliable and straightforward relationships between porosity and permeability exist, but their application is limited to homogenous reservoir rocks due to consistent petrophysical properties. Therefore, the complex reservoir conditions are always associated with uncertain predictions which lead to poor subsurface correlation (Tixier 1949; Timur 1968; Coats and Dumanoir 1974; Donovan 1984). As a result, empirical and regression modeling

techniques have become standard practices in the oil and gas industry for permeability estimation using well logs, especially for the wells having no reservoir cores. It also reduces human errors and the extended time required for laboratory measurements. The modern machine learning algorithms are cost-effective, have better predictive capabilities, and provide any required range of probabilities for data analytics.

The present study focuses on the precise computation of permeability using different regression modeling techniques, especially for scenarios where core data is not available. This study demonstrated the application of four machine learning techniques SLR, LR, MLR, and SVR to analyze and forecast the permeability of the Volve oil field petrophysical dataset having shale volume and porosity as independent parameters. The open-sourced datasets of Equinor's Volve field (<https://data.equinor.com>) an oil field on the Norwegian Continental Shelf have been used in the present study. The field is located 200 km west of Stavanger at the southern end of the Norwegian sector (Fig.1). The operator Equinor and the Volve license partners, ExxonMobil and Bayern gas, have made the repository of all subsurface and operating data from this oil field available for research and analysis purposes. Petrophysical and coring data used in the current study is taken from two wells, i.e., F-15/9/19A and F-15/9B&BT2, due to the availability of selected parameters like initial Klinkenberg corrected core permeability, porosity, and shale volume. Python programs were developed to apply SLR, LR, MLR, and SVR methods and analyze core data of the abovementioned wells. Finally, the predicted outputs are compared with available test datasets to calculate the accuracy of designed regression models. These rapid and cost-effective soft-computing regression techniques show outstanding results of permeability forecasting with an accuracy of over 0.99 R^2 scores (in the case of well F-15B&BT2), using Gaussian process regression including variables porosity and shale volume. A comparative analysis is also made between the four regression algorithms as mentioned above to estimate the best fit method for assessing large petrophysical datasets. The results indicated that SVR had the best peak performance compared to MLR, SLR, and LR, in decreasing bias. However, this does not mean that SVR is inherently the superior method compared to the other three in every situation.

The subsequent sections of this paper are showing the background of the study area and brief details of the adopted algorithms in the "Background of the study area and adopted algorithms" section. The details of the adopted methodologies and criteria of the machine learning model section with a process flow chart are given in the "Methodologies adopted" section. The next important section, "Result and discussion," deals with interpretation, description, and comparison of achieved outcomes using various algorithms. Finally, the "Conclusion" section is kept at the end.

Fig. 1 Location map of the Volve field located at the southern end of the Norwegian Continental Shelf (Ravasi et al. 2015). Wells F-15/9/19A and F-15/9B & BT2 used in the present research work are situated in the Volve oil field shown by green color in the above map under block 15/9



Background of the study area and adopted algorithms

The block 15/9 of the Volve field has proven commercial quantities of hydrocarbon discoveries. It lies 200 km west of Stavanger at the southern end of the Norwegian sector with an average water depth of 80 m. The field is situated 5 km north of the Sleipner Vest field. The Jurassic age Hugin formation acts as the central reservoir unit of this field (Lervik 2006; Otoo and Hodgetts 2020). The thickness of the Hugin formation is estimated to range between 5 and 200m with an approximate reservoir depth of 2700–3000 m, which can vary in different regions due to post-depositional erosional processes (Folkestad and Satur 2008). The first successful discovery well was drilled in Volve 1993, and the development and operation (PDO) plan was approved in 2005. Initially, field development was planned with the jack-up processing drilling facility and the vessel “Navion Saga” which was used for storing stabilized oil. According to the published literature and the reports of the Norwegian petroleum directorate, the hydrocarbon production from this field was started in 2008, and finally in 2016, decommission decision was taken after 8.5 years of successful operation life (www.equinor.com). This was twice more than as long as initially planned. Volve produced with a peak rate of about 56,000 barrels per day and delivered a total of 63 million barrels of oil with a recovery rate of 54% of reserve estimates (Sen and Ganguli 2019). As mentioned in the Introduction section that the different machine learning techniques are used to analyze the well logs of Volve oil field, the brief details of the used algorithms in current research work are as follows.

Simple linear regression (SLR)

It estimates the relationship between one or more independent variables and a dependent variable by minimizing the sum of the squares in the difference between the observed and predicted values. In the LR, a dependent variable and one or more independent variables relationship is modeled by fitting a linear equation. Herein, a single scalar predictor variable X is predicted using a simple scalar responsive vector Y (Table 1). This fits a linear model that will best minimize the residual sum of squares between the observed responses in the dataset and the responses predicted by the linear approximation (Uyanık and Güler 2013).

Lasso regression (LR)

The LR or least absolute shrinkage is a regression analysis method that employs both regularization and variable selection to attain maximum possible prediction accuracy and interpretability from the original data. It selects a reduced set of covariates for use in a model for higher accuracy (Liu et al. 2020).

Multi-linear regression (MLR)

It is an extended context of LR. It estimates the linear relationship between multiple explanatory (independent) variables and a single response variable. The only difference between SLR and MLR is the number of independent variables. The basis of MLR is highly dependent on the assumption of a linear relationship between both the dependent and independent variables (Pereira 2004; Uyanık and Güler 2013). No major correlation between the independent variables is

Table 1 Regression algorithms with equations and formulas

S.N	Regression	Formulas	References
1	Linear regression	Minimum $\sum_{i=1}^n (y_i - w_i x_i)^2$	Where y_i is the target, w_i is the coefficient, and x_i is the predictor(feature) Also, formula and calculations for linear regression uses $y_i = \beta_0 + \beta_1 x_{i1}$
2	Lasso regression	$\sum_{i=1}^n (y_i - y)^2 + \lambda \sum_{j=1}^p \beta_j $	Where β_j is the slope of the line and λ is the tuning parameter
3	Multi-linear regression	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \epsilon$	Where y_i is the target or predicted variable, β_1 and β_2 are regression coefficients, β_p represents the slope coefficient, and ϵ is the model's error term
4	Support vector regression	Minimum $\frac{1}{2} \ w\ ^2$	Constraint $y_i - w_i x_i < \epsilon$

assumed. Essentially, it is an extension of ordinary least squares that uses more than one independent variable.

Support vector regression (SVR)

The objective was to find a function $f(x)$ which has at most ϵ deviation from the obtained targets for all the training data and at the same time is as flat as possible using SVR. It gives the flexibility to define how much error is acceptable in our model and will find an appropriate line (or hyperplane in a higher dimension) to fit the data (Jap et al. 2015). In contrast to ordinary least squares (OLS), the objective of SVR is to minimize the coefficient. Three different kernels are used to represent the trend in SVR, i.e., RBF, linear, and polynomial.

Methodologies adopted

The Volve field data is evaluated upon in several aspects to make it suitable for the comparative analysis. A flow chart is shown in Fig. 2 illustrating the methodology adopted in the current research work. The stage-wise flow of data collection from open-sourced Equinor data repository and its cleaning process of null values and outliers, the preparation of MLR models, and evaluated R^2 scores comparison of prepared models for comparison purpose are highlighted.

The null values and the data outliers were removed using Scikit-Learn (or Sklearn) (Arnold et al. 2011). Data outliers are unknown data spikes, which are bizarrely different from other elements of the dataset. Z-score function from Scikit-Learn was used to remove data outliers, while null values were eliminated using .dropna() function as shown in Appendices 1, 2, and 3.

The comparative analysis between the SLR, LR, MLR, and SVR methods is used to predict the best technique for getting results with higher accuracy. The goal was to establish the data-oriented accurate estimation of petrophysical parameters

from the raw datasets. Essentially, this boils down to establishing a statistical relationship between a response variable (Y) and an explanatory variable (X) (Wiener et al. 1991). This can be done by employing a regression analysis to model the distribution of variables concerning one another and derive a relationship. The type of model to be put to use depends on the distribution of Y for X on a plane. Continuous and normal distribution warrant the use of LR; a binary distribution, logistic regression; Poisson or multinomial distribution; log-linear analysis; and so on. With the help of modeling, we try to estimate the predictor variable(s) effect on the magnitude of a response variable.

The methods show a significant dependence on the porosity and shale volume of the system. Theoretically, permeability increases with an increase in porosity, and decreases due to a higher amount of shale volume, because the increase in shale volume accounts for the blockage of the path to the hydrocarbon flow, subsequently reducing the permeability of the system (Yao and Holditch 1993). Permeability is defined as the measure of the ease with which a fluid flows through a porous medium (Fossen and Bale 2007; Jia et al. 2019). It is a critical aspect to be accounted for in any reservoir analysis. Permeability data can be obtained in laboratories (core analysis), in reservoirs (pressure transient tests), and through well logs (Yao and Holditch 1993). However, the conventional methods of predicting permeability are time-consuming (add non-productive time) therefore considering economically unfavorable. Therefore, rapid and economically viable methods of AI/ML have been used to account for the results of the prediction of permeability using available datasets. This regression approach in the oil and gas sector, for permeability estimation, can be used to build high-accuracy fluid flow models. Ultimately, the results of predicted reservoir properties through conventional and ML approaches can be applied in the planning of the newly proposed well to enhance the geological chance of success (GCF)(Pereira 2004; Wendt et al. 1986).

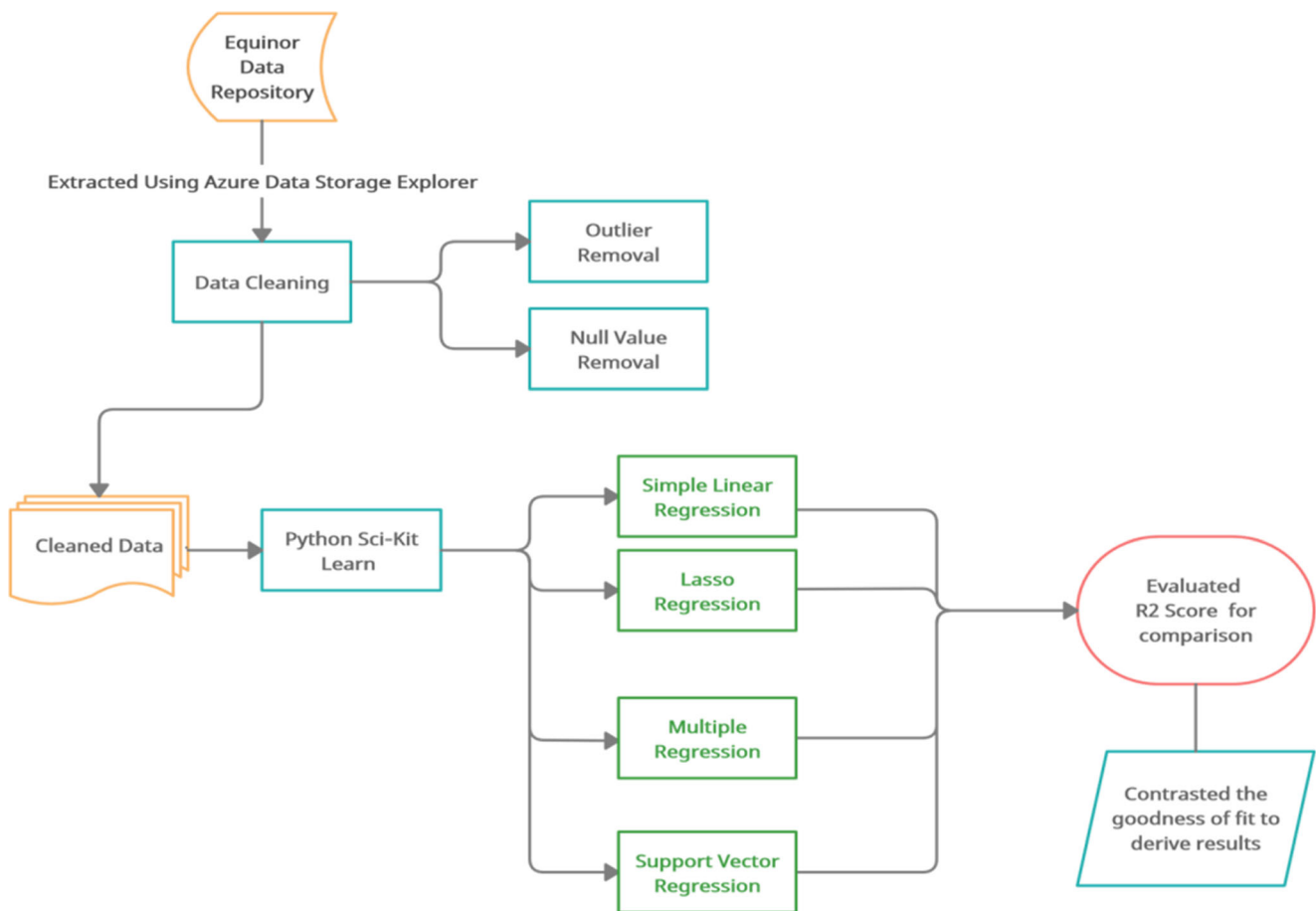


Fig. 2 Flow chart illustrating various stages of data collection, conditioning, and working methodology

ML model selection

Every model has its advantage and disadvantages; therefore, it is crucial to discern the type of regression algorithm to be used by subsequently plotting the data on a scatter plot. For simple linear data, LR algorithm can be used, but if the data is non-linear, data transformation can increase the model's accuracy. In rare cases, if the data transformation fails for the non-linear selection model, then more complex models can be utilized. The workflow identifies the type of distribution from the scatter plot and recognizes if it resembles a known mathematical function. If the data resembles a linear function, the model utilized is linear, an exponential model for exponential curves, etc. The forward selection consists of fitting the data to a rudimentary mathematical model, evaluating the data fitness (commonly referred to as goodness of a fit), and eventually moving on to more complicated models to obtain comparatively better correlations. However, the backward selection also aims toward getting a model with a desired goodness of fit. Still, the difference between the two models lies in the fact that the backward selection begins with starting the most complicated model to fit the data and then simplifying it down as per need. This study employed a forward selection model. We

started with relatively simple, two-dimensional models, including those formed by SLR and LR. We then moved on to more complex, three-dimensional models employing two and, in some cases, even three independent models variables. The problems in the oil and gas industries can be categorized as statistical problems or regression problems. The advanced mathematical approach empowers computers to have decision-making capability using AI/ML-based mathematical and statistical ways of approaching solutions (Y Liu and Chen 1999; Yang Liu et al. 2019; Zanjani et al. 2020). Understanding the permeability of a system is a critical aspect that needs to be considered for hydrocarbon exploitation and production. This involves the regressive analysis study of data from a drilled core, which is then extrapolated for the whole formation to develop the conceptual model of permeability (Letham and Bustin 2016; J. Li and Sultan 2017).

In the current research work above discussed, four methods (SLR, LR, MLR, SVR) are used to estimate Klinkenberg core corrected permeability using porosity and shale volume. Data from wells F-15/9-B&BT2 and F-15/9-19A were used to develop four regression models, and their R^2 scores were computed to compare the goodness of fit. In addition to this, a correlation plot for the two wells was formed using available petrophysical data

(Figs. 3 & 4). This correlation plot lends valuable insight into (i) how the different petrophysical datasets are related to one another, on a scale of -1 to 1 , in which -1 depicts a robust negative correlation between the two parameters and 1 depicts a strong positive correlation, and (ii) the probability distribution of values for all individual variables (seen on the diagonal elements of the CORRPlot matrix). Figure 3 contains a correlation matrix of the data from well F-15/9-19A, between different variables logKL, PHIF, density-porosity (PORD), SW, and VSH. The matrix aims to show a correlation between these parameters, while the elements on the diagonal include individual distributions of the data. The lines shown in red represent an elementary trend between the two variables of a plot. In contrast, the number on the top left represents Pearson's correlation coefficient for the two datasets. A positive value trending toward 1.0 represents a robust and direct relation, while a value trending towards -1.0 represents an inverse relation.

It can be interpreted from the figures that logKL shows a relatively strong dependence on porosity, while inverse relationships of varying strength also exist against water saturation and shale volume. Data from Fig. 4, however, indicates the absence of a relationship against porosity but shows a strong negative correlation against shale volume (Tembely et al. 2021). On a

similar line of analysis, the correlation matrix of the second well (F-15/9-B&BT2) is prepared and shown in Fig. 4. In the paper by Wendt et al. (1986), the authors depict the relation for the prediction of permeability to be dependent on porosity-related variables, shale volume, and water saturation in decreasing order of relevance to the trained dataset. It is also discerned through the correlation matrix shown in Figs. 3 and 4, permeability data shows a positive correlation against petrophysical variables such as porosity and water saturation. In contrast, a negative correlation against shale volume is evident. The criteria for selecting the dataset is the involvement of variables that show a clear correlation with permeability (Figs. 3 & 4). Cleaning of the dataset for the removal of null values and the vast amounts of data outliers is recommended before making any interpretation.

The box plot shown in Fig. 5 resembles the dataset and its point distribution for different depth intervals through well F-15/9-19A. For a single interval, the original box represents the distribution of values, the and horizontal line dividing the boxes in two represents the median of that subset of data. At the same time, the top and bottom whiskers conveniently depict the upper and lower limits of data. Despite the spaced distribution through depths 3870–3880 m, the plot shows consistency in value distribution concerning depth. For optimal

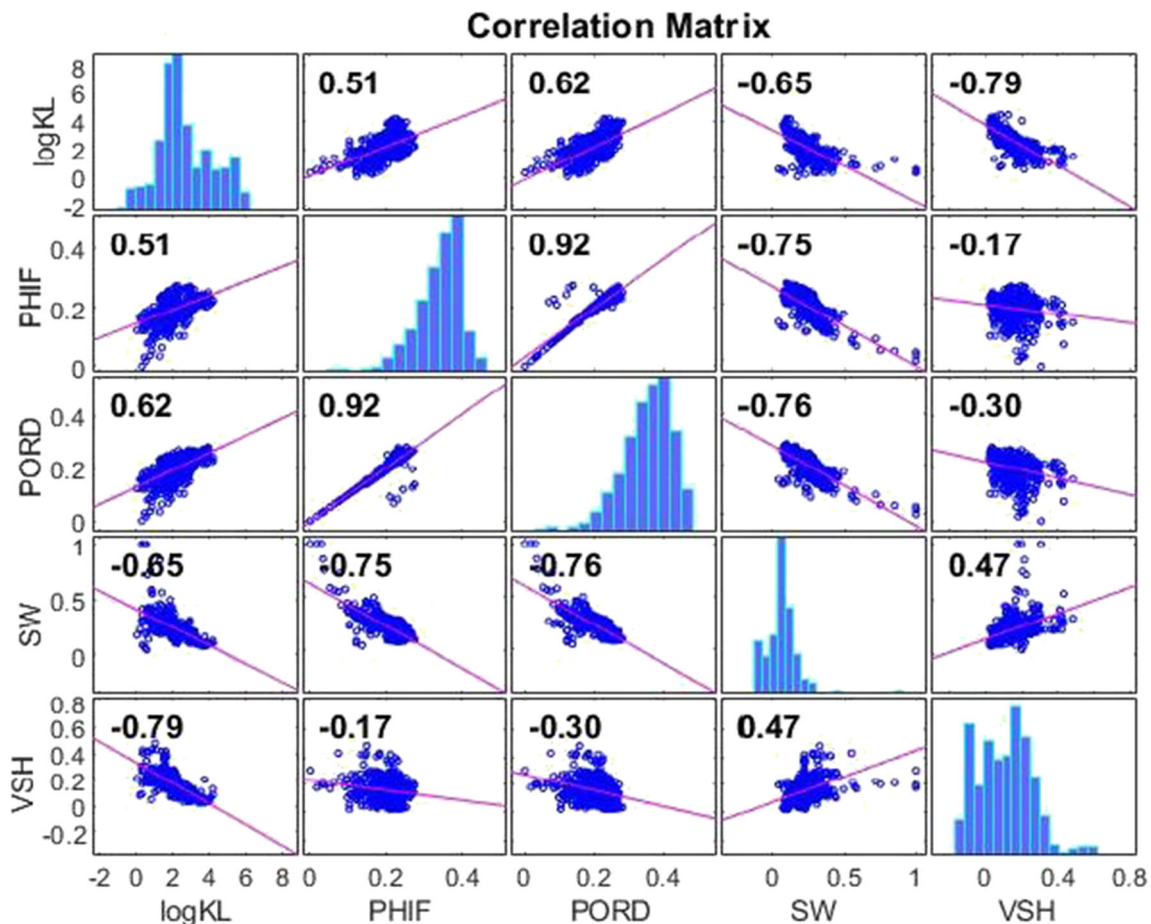


Fig. 3 Correlation matrix of various parameters of well F-15/9-19A

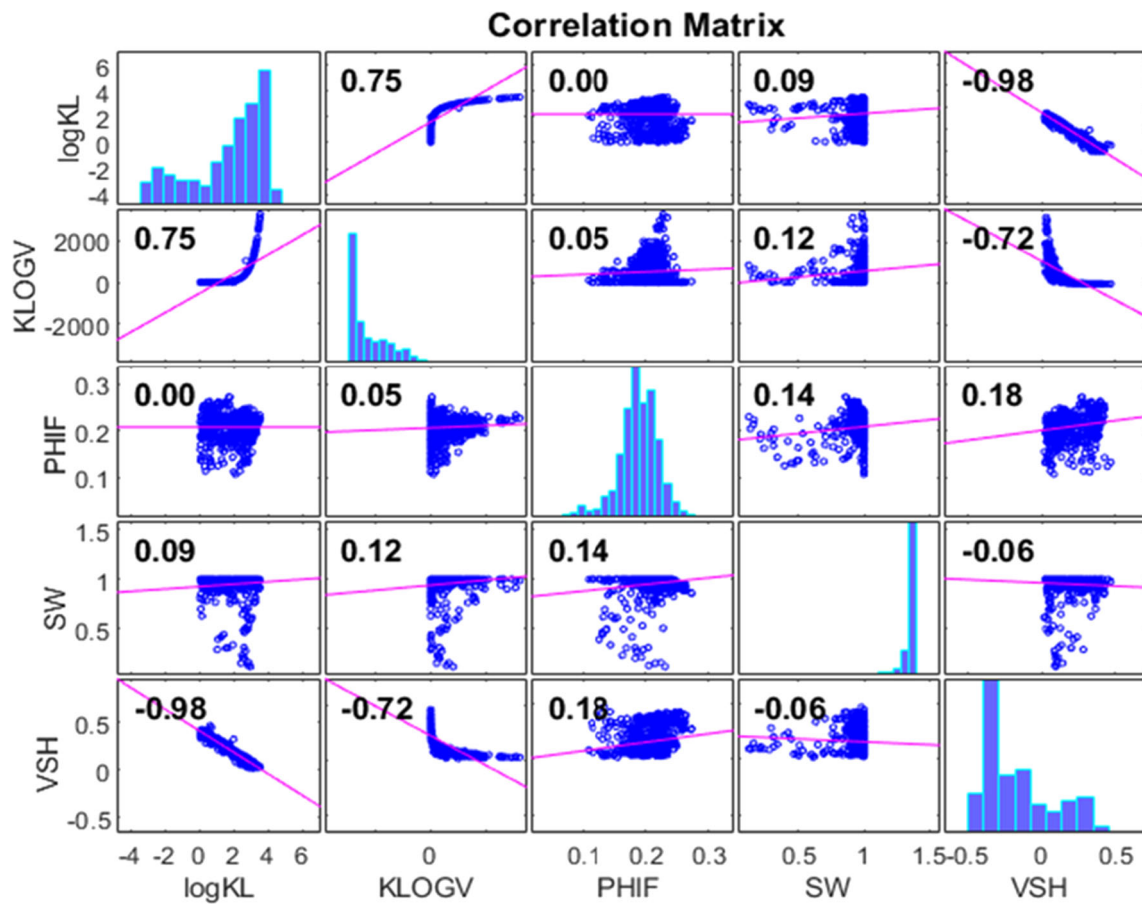


Fig. 4 Correlation matrix of various parameters of well F-15/9 B&BT2

results, it is necessary to have a dataset tuned with data that has a narrow point distribution, with values for a single interval converging toward a singular value. While such datasets

might not necessarily result from core permeability data reading, they can always be used as model datasets as a criterion to determine the dataset.

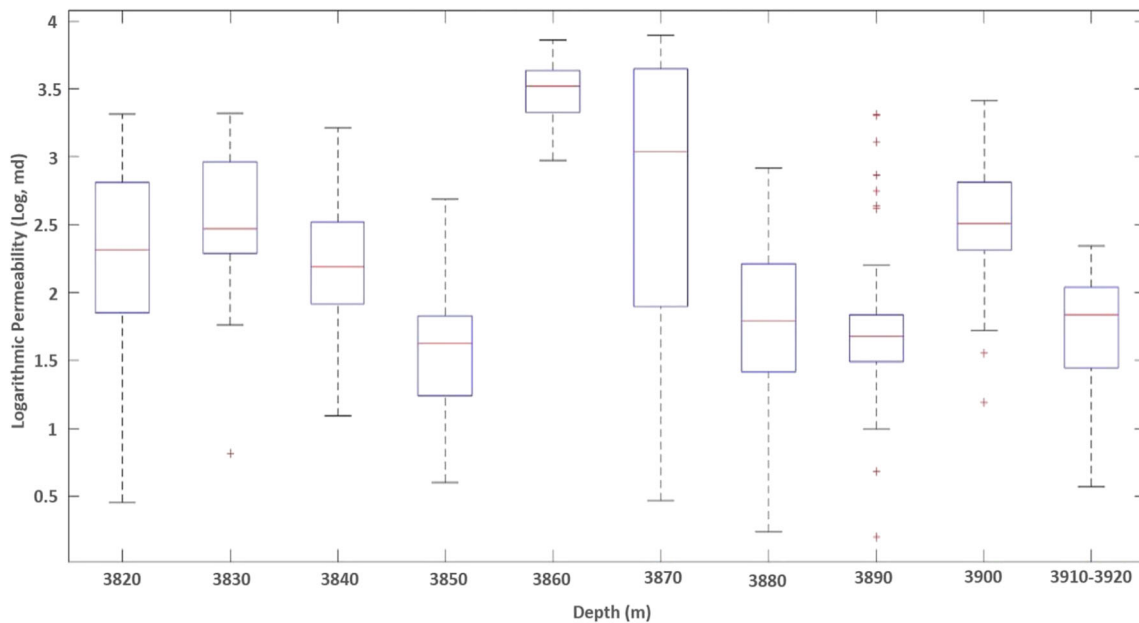


Fig. 5 The box plot describing the distribution of the logarithmic permeability [Log(md)] points against depth for the well F-15/9-19A. The central line between the boxes represents the median of the data bin, and the whiskers representing the upper and lower limits

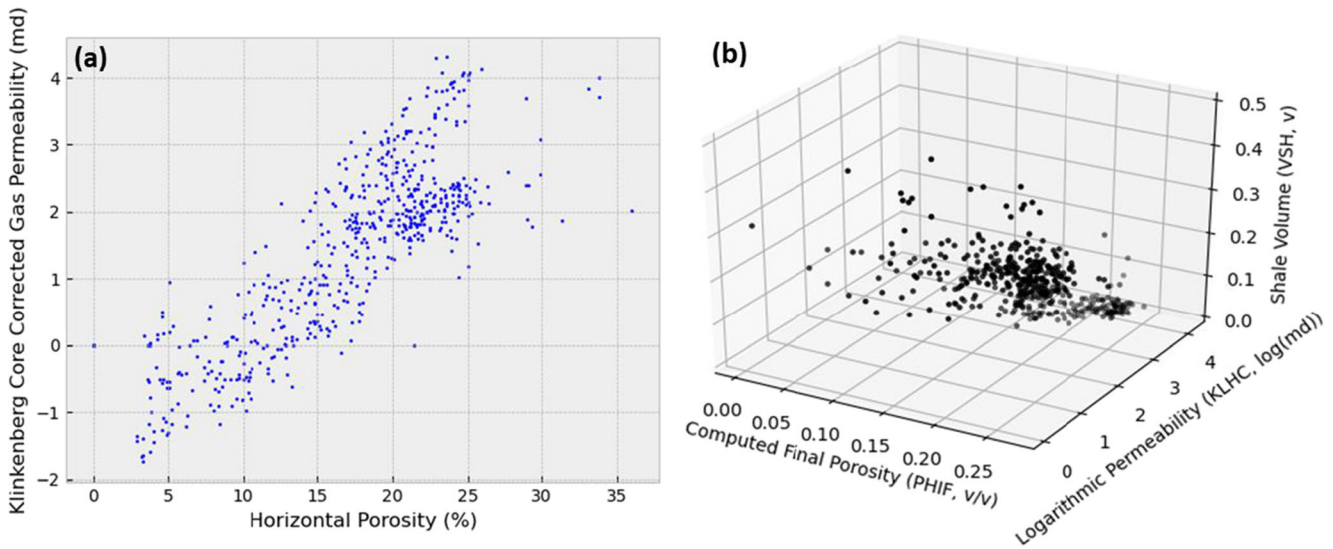


Fig. 6 Scatter plots. **a** Two-dimensional plot between the logarithm of Klinkenberg core corrected permeability and horizontal porosity. **b** Three-dimensional plot between the logarithm of Klinkenberg core corrected permeability and horizontal porosity-shale volume

Result and discussions

A linear trend was observed between the horizontal porosity and Klinkenberg core corrected permeability for SLR and LR, whereas a planar trend has been established for SVR and MLR. The output received from all four methods is compared based on their R^2 values. This statistical measure represents the percentage of variance for the dependent variable, which is explained by an independent variable in a regression model. The independent variables in our studies are horizontal porosity and shale volume, while the response variable is Klinkenberg core corrected permeability.

In the scatter plot for linear data establishing a correlation between horizontal porosity and Klinkenberg core corrected permeability, the permeability (in md) is represented as a logarithmic function due to the non-linearity of the data (Fig. 6). A point of interest is the high negative correlation of shale volume to permeability, which signifies a specific decrease in permeability seen with an increase in shale volume (Fig. 6). The low porosity could also be due to shale volume, which increases clay content in pores and negatively impacts porosity and permeability. It is also important to note that the majority of plotted points lie on a straight line, which is due to an increase in porosity with an increase in permeability (Singh 2019). The regression models created during this project were

programmed using Python and its modules (Sci-kit Learn for regression, Matplotlib for plotting, and Pandas-Numpy-dlisis for data handling). A summary of Equinor’s published mathematical relationships for selective formations using MLR published by Equinor is shown in Table 2. Additionally, two-dimensional relations have been portrayed in Fig. 7. The original points have been outlined with a scatter plot, while the formulated correlation has been displayed with the help of a regression trend line (displayed in red). Similarly, three-dimensional relations have been displayed in Fig. 8.

Figure 9 depicts a simple overlapping plot of permeability illustrated by core analysis (with applied overburden correction and correction for Klinkenberg effect), plotted on top of data predicted from two-variable [KLOGH vs PHIF and VSH] correlation used to predict permeability of the entire Volve field (Table 2) against increasing depth. It can be observed in Fig. 9a that sections A, C, D, and E show varying levels of deviation from the actual permeability data, as a consequence of the limitations of using a simple three-dimensional model for prediction, while sections B and F depict a high degree of similarity between the core data and predicted data. Figure 9b depicts a three-dimensional view of points predicted by the abovementioned correlation against actual data from core analysis. In unconventional reservoirs, the permeability and gas accumulation ability of shale gas

Table 2 Equinor’s published relationships for selective formations using multivariate regression

S. no	Mathematical relationships	Formation
1	$KLOGH = 10^{(-0.7 + 17.3*PHIF - 5*VSH)}$	Sleipner øst, Hugin formation
2	$KLOGH = 10^{(2 + 8*PHIF - 9*VSH)}$	Volve, Hugin formation
3	$KLOGH = 10^{(-3 + 32*PHIF - 2*VSH)}$	Volve, Sleipner formation
4	$KLOGH = 10^{(-1.85 + 17.4*PHIF - 3*VSH)}$	Sleipner øst, Skagerrak formation

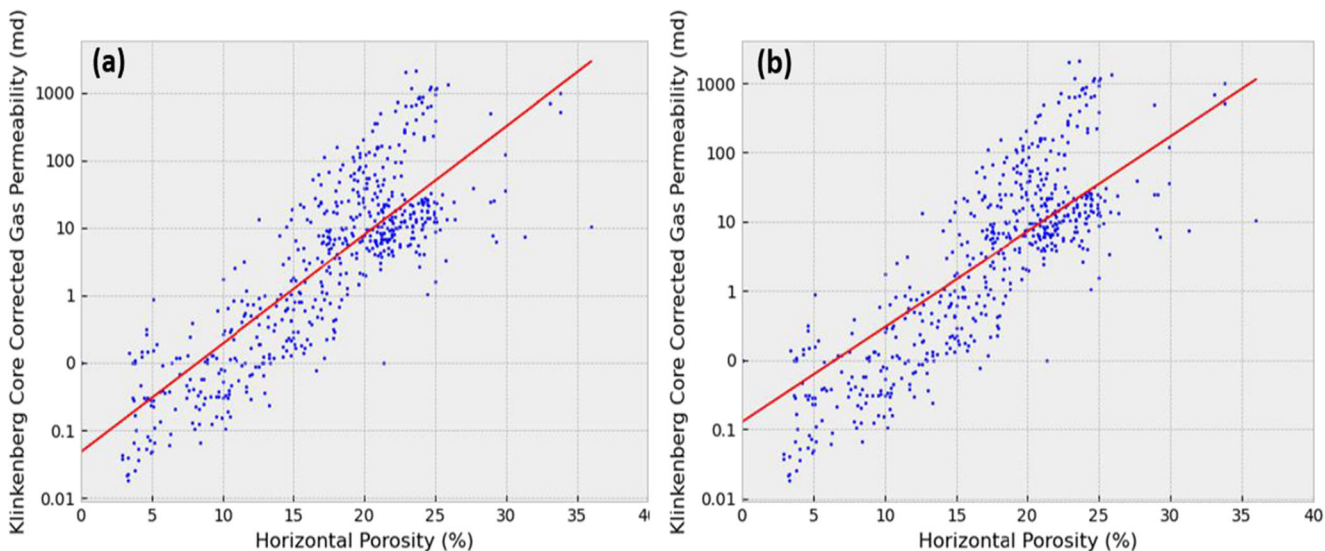


Fig. 7 Two-dimensional plot with trend line shown in red color. **a** Plotting with the SLR for horizontal porosity and logarithm of Klinkenberg core corrected permeability with the equation

$\log(k) = -1.395 + [0.1671] * HPOR$. **b** Plotting with the LR for horizontal porosity and logarithm of Klinkenberg core corrected permeability with the equation $\log(k) = -0.6574 + [0.1230] * HPOR$

reservoirs depend upon capillary pressure difference between sweet spots and surrounding rocks (Zheng et al. 2020), so in the presence of appropriate depth against log curves, accurate machine learning models can be created to predict permeability for similar shale formations throughout the field (Wen et al. 2020).

Results of the comparison between the regression algorithms as mentioned above are tabulated in Fig. 10. However, it was also observed that the goodness of fit could be further boosted by the inclusion of a third independent parameter—water saturation. In an independently formed correlation between permeability, porosity, and shale volume using MLR on a minimal dataset of <700

data points, a relatively strong correlation was established between the response and the predictor variable of the R^2 score, amounting to 0.79. However, a look at the correlation matrix directed use toward the presence of a relationship between the response variable and water saturation as shown in Figs. 3 and 4. Including water saturation data into said minimal dataset from well F-15/9-19A boosted the R^2 score from 0.79 to 0.81, an effect that cannot be neglected considering the limitations provided by the data. The authors firmly believe that the inclusion of water saturation into regression analysis done to predict permeability can lead to a high accuracy bump in the predictions made using only two independent variables.

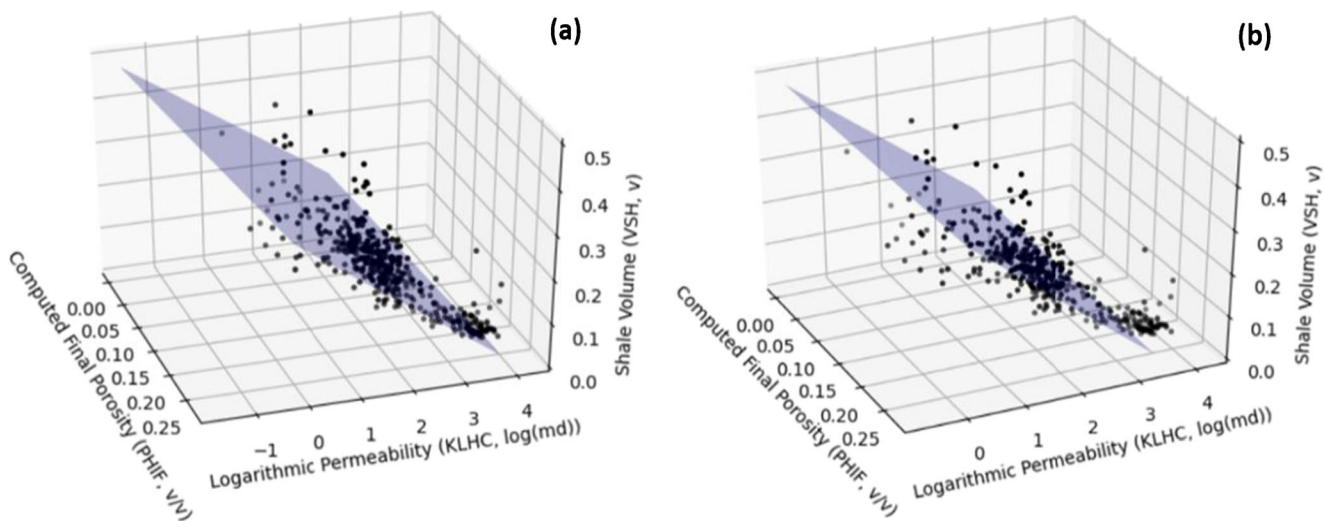


Fig. 8 Three-dimensional plot with a trend plane shown as a translucent blue plane. **a** Plotting with the MLR for horizontal porosity-shale volume and logarithm of Klinkenberg core corrected permeability with the equation $\log(k) = [1.6728] + [8.3827] * HPOR - [7.1119] * VSH$. **b** Plotting with

the SVR for horizontal porosity-shale volume and logarithm of Klinkenberg core corrected permeability with the equation $\log(k) = [2.4710] + [3.6827] * HPOR - [6.3984] * VSH$

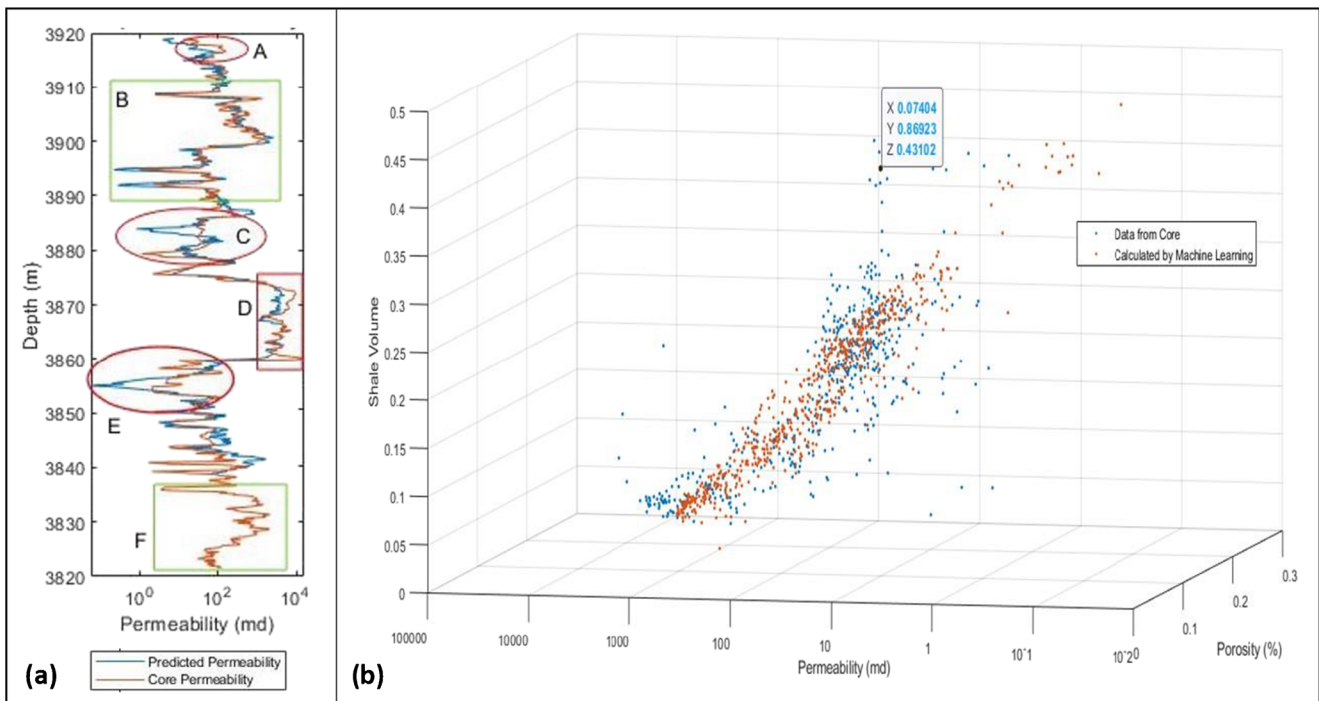


Fig. 9 a Depth vs permeability curve showcasing the variation between predicted permeability and core permeability for well F-15/9-19A. A–F is described in the text part. b A three-dimensional scatter plot showcasing

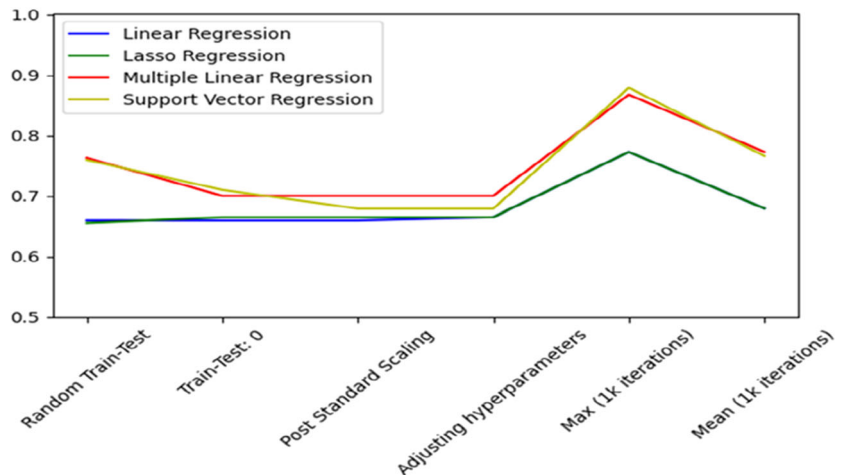
variation between points predicted via MLR vs points from core data analysis between permeability, shale volume, and porosity

The results effectively support the claims made by Wendt et al. (1986) that introducing relevant independent variables in the regression vastly increases the goodness of fit. This is made evident by a gap of effectively 0.10 that both MLR and SVR hold over Lasso and SLR in the “mean after 1000

Iterations” category. Considering LR, a considerable increase from the previous score of 0.659 to 0.665 was noticed once the value of the hyperparameter λ was decreased from 1.0 to 0.0001. According to Valenzuela et al. (2017), SVR results improve with data scaling; however, standard scaling this

Fig. 10 Comparison of the outputs of various regression models at different stages

Category	LR	SLR*	MLR	SVR
Random Train-Test Sizes	0.659948099	0.655682866	0.762850366	0.759359542
At reproducible Train-Test (0)	0.659747862	0.665220759	0.700224978	0.710060923
After Standard Scaling Data	0.659747862	0.665220759	0.700224978	0.679001737
After Adjusting hyperparameters (alpha = 0.0001)	0.665221424	0.665220759	0.700224978	0.679001737
Max after 1k Iterations	0.772576743	0.772577846	0.867431486	0.878731553
Mean After 1k iterations	0.679489599	0.679489405	0.772839444	0.766217991



dataset with Sklearn's standard scalar function resulted in a slight decrease in the model's accuracy. However, it is also important to note that including larger sets of data and relevant variables will undoubtedly improve upon the model's goodness of fit, and, at a certain point, it will lead to the condition of overfitting—resulting in a model with shallow bias, but a model that predicts values in an incredibly narrow margin, leading to high variance (or high margin of error in predicted data). This is why this study focuses on a maximum of two reservoir parameters.

Conclusions

All the models prepared during current research work show great potential for forecasting permeability using well log and core data, but SVR has shown better results among the chosen four algorithms; SLR, LR, SVR, and MLR. The performance of machine learning methods is greatly dependent on the quality of the input dataset; therefore, the performance of an algorithm must be examined for each specific dataset and problem. The following results can be concluded from the output of the current research work.

1. The inclusion of water saturation into conventional permeability prediction models can substantially boost prediction accuracy. The use of relevant petrological independent variables in the regression algorithm can significantly enhance the correctness of models and is recommended for industrial applications.
2. Despite explicitly being conceived for geophysical studies, LR fails to attain accuracies even remotely close to multivariable techniques like multiple and SVR.
3. A correlation plot depicting Pearson's correlation coefficient between well log data variables including horizontal permeability, porosity, and water saturation was created, which can be used to describe the relationships between these variables and be used to derive an idea about their distribution.
4. Considering all the models included in the scope of this paper, it can be said that MLR remains the best contender for a general use-case in the industry; however, it is observed that under specific train-test size datasets, the accuracy of SVR can be boosted over that of MLR, implying the use of SVR in fringe cases that require high goodness of fit.

Appendix 1. Python code implementing NULL value removal (NULL values represented as -999.25 in the original files) and outlier removal using Scikit Learn's Z-score functionality

```
In [1]: import dlistio
import matplotlib.pyplot as plt
import matplotlib as mpl
import numpy as np
from scipy.stats import zscore

In [2]: def clean(frame):
a = frame
a = a[a != -999.25]
zarr = zscore(a)
new = np.empty_like(a)
for index, item in enumerate(zarr):
    if abs(item) < 2.0:
        new[index] = a[index]
    else:
        new[index] = a[index-1]
print(len(new))
return new
```

Appendix 2. Python code implementing multi-linear regression

```
def mulreg(X, y, a ):
    from sklearn.model_selection import train_test_split
    X_train, X_test, y_train, y_test = train_test_split(X, y, random_state = a)

    # Fitting Multiple Linear Regression to the Training set
    from sklearn.linear_model import LinearRegression
    regressor = LinearRegression()
    regressor.fit(X_train, y_train)

    # Predicting the Test set results
    y_pred = regressor.predict(X_test)

    from sklearn.metrics import r2_score
    rscore=r2_score(y_test,y_pred)

    coefficient = regressor.coef_
    intercept = regressor.intercept_

    x_surf, y_surf = np.meshgrid(np.linspace(X.PHIF.min(), X.PHIF.max(), 100),np.linspace(X.VSH.min(), X.VSH.max(), 100))
    onlyX = pd.DataFrame({'PHIF': x_surf.ravel(), 'VSH': y_surf.ravel()})
    fittedY = regressor.predict(onlyX)
    ## convert the predicted result in an array
    fittedY=np.array(fittedY)

    return rscore, coefficient, intercept, x_surf, y_surf, fittedY
```

Appendix 3. Python code implementing support vector regression

```
def svr(X, y, a ):
    from sklearn.model_selection import train_test_split
    X_train, X_test, y_train, y_test = train_test_split(X, y, random_state = a)

    from sklearn.svm import SVR
    regressor = SVR(kernel = 'linear')
    regressor.fit(X_train, y_train)

    y_pred = regressor.predict(X_test)

    from sklearn.metrics import r2_score
    rscore = r2_score(y_test, y_pred)
    coefficient = regressor.coef_
    intercept = regressor.intercept_
    x_surf, y_surf = np.meshgrid(np.linspace(X.PHIF.min(), X.PHIF.max(), 100),np.linspace(X.VSH.min(), X.VSH.max(), 100))
    onlyX = pd.DataFrame({'PHIF': x_surf.ravel(), 'VSH': y_surf.ravel()})
    fittedY = regressor.predict(onlyX)
    ## convert the predicted result in an array
    fittedY=np.array(fittedY)

    return rscore, coefficient, intercept, x_surf, y_surf, fittedY
```

Acknowledgements The authors gratefully acknowledge the support received from the University of Petroleum and Energy Studies, Dehradun. All the authors are thankful to Equinor for the availability of well data on an open-source platform (<https://data.equinor.com>) for academic research purposes.

Author contribution Piyush Prajapati leads to collect the data, and Naman Khilrani leads to prepare the codes for data analysis. Atul Kumar Patidar conceptualizes, draws methodology, and supervised the work, and finally, all the authors synchronized the results, made interpretations, and wrote this paper.

Data availability The well log data used in the current research work is taken from the open-sourced Volve data village dataset of Equinor (2018), hosted at <https://data.equinor.com/>.

Declarations

Conflict of interest The authors declare no competing interests.

References

Adeniran AA, Adebayo AR, Salami HO, Yahaya MO, Abdulraheem A (2019) A competitive ensemble model for permeability prediction in heterogeneous oil and gas reservoirs. *Applied Computing and Geosciences* 1:100004. <https://doi.org/10.1016/j.acags.2019.100004>

- Ahmadi MA, Chen Z (2019) Comparison of machine learning methods for estimating permeability and porosity of oil reservoirs via petrophysical logs. *Petroleum* 5(3):271–284. <https://doi.org/10.1016/j.petlm.2018.06.002>
- Al-Bulushi NI, King PR, Blunt MJ, Kraaijveld M (2012) Artificial neural networks workflow and its application in the petroleum industry. *Neural Computing and Applications* 21(3):409–421. <https://doi.org/10.1007/s00521-010-0501-6>
- Alger RP, Raymer LL, Hoyle WR, Tixier MP (1963) Formation density Log applications in liquid-filled holes. *Journal of Petroleum Technology* 15(03):321–333. <https://doi.org/10.2118/435-PA>
- Alkinani HH, Al-Hameedi ATT, Dunn-Norman S, Flori RE, Alsaba MT, Amer AS (2019) Applications of artificial neural networks in the petroleum industry: a review. *SPE Middle East Oil and Gas Show and Conference (MEOS)*, Manama, Bahrain, (SPE-195072-MS). <https://doi.org/10.2118/195072-MS>
- Arnold K, Gosling J, Holmes D, Flanagan D, Odersky M, Spoon L, Venners B, Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A et al (2011) Scikit-learn: machine learning in {P}ython. *Journal of Machine Learning Research* 12(85):2825–2830. <http://jmlr.org/papers/v12/pedregosa1a.html>
- Coats GR, Dumanoir JL (1974) A new approach to improved Log-derived permeability. *The Log Analyst* 15(1):17–29
- Das B, Chatterjee R (2018) Well log data analysis for lithology and fluid identification in Krishna-Godavari basin, India. *Arabian Journal of Geosciences* 11:231. <https://doi.org/10.1007/s12517-018-3587-2>
- Donovan DT (1984) Geological survey. *Nature* 312:192. <https://doi.org/10.1038/312192a0>
- Doveton JH and Prenskey SE (1992) Geological applications of wireline logs - a synopsis of developments and trends. *The Log Analyst* 33: 286–303
- Folkestad A, Satur N (2008) Regressive and transgressive cycles in a rift-basin: depositional model and sedimentary partitioning of the Middle Jurassic Hugin Formation, Southern Viking Graben, North Sea. *Sedimentary Geology* 207(1–4):1–21. <https://doi.org/10.1016/j.sedgeo.2008.03.006>
- Fossen H, Bale A (2007) Deformation bands and their influence on fluid flow. *American Association of Petroleum Geologists Bulletin* 91(12):1685–1700. <https://doi.org/10.1306/07300706146>
- Guan Q, Zhang F, Zhang E (2019) Application prospect of knowledge graph technology in knowledge management of oil and gas exploration and development. 2019 2nd International Conference on Artificial Intelligence and Big Data. ICAIBD 2019:161–166. <https://doi.org/10.1109/ICAIBD.2019.8837003>
- Jap D, Stöttinger M, Bhasin S (2015) Support vector regression: exploiting machine learning techniques for leakage modeling. Fourth Workshop on Hardware and Architectural Support for Security and Privacy (HASP 15). Association for Computing Machinery, New York, NY, USA, Article 2:1–8. <https://doi.org/10.1145/2768566.2768568>
- Jia R, Liu B, Fu X, Gong L, Liu Z (2019) Transformation mechanism of a fault and its associated microstructures in low-porosity rocks: a case study of the Tanan depression in the Hailar-Tamtsag basin. *Journal of Marine Science and Engineering* 7(9):286. <https://doi.org/10.3390/jmse7090286>
- Lervik KS (2006) Triassic lithostratigraphy of the Northern North Sea basin. *Norsk Geologisk Tidsskrift* 86(2):93–115
- Letham EA, Bustin RM (2016) Klinkenberg gas slippage measurements as a means for shale pore structure characterization. *Geofluids* 16(2): 264–278. <https://doi.org/10.1111/gfl.12147>
- Li J, Sultan AS (2017) Klinkenberg slippage effect in the permeability computations of shale gas by the pore-scale simulations. *Journal of Natural Gas Science and Engineering* 48:197–202. <https://doi.org/10.1016/j.jngse.2016.07.041>
- Li H, Yu H, Cao N, Tian H, Cheng S (2020) Applications of artificial intelligence in oil and gas development. *Archives of Computational Methods in Engineering* 28:937–949. <https://doi.org/10.1007/s11831-020-09402-8>
- Liu Y, Chen G (1999) Optimal parameters design of oilfield surface pipeline systems using fuzzy models. *Information Sciences* 120(1): 13–21. [https://doi.org/10.1016/S0020-0255\(99\)00059-6](https://doi.org/10.1016/S0020-0255(99)00059-6)
- Liu Y, Chen S, Guan B, Xu P (2019) Layout optimization of large-scale oil-gas gathering system based on combined optimization strategy. *Neurocomputing* 332:159–183. <https://doi.org/10.1016/j.neucom.2018.12.021>
- Liu Y, Wei Y, Liu Y, Li W (2020) Forecasting oil price by hierarchical shrinkage in dynamic parameter models. *Discrete Dynamics in Nature and Society* 2020:29–33. <https://doi.org/10.1155/2020/6640180>
- Mohaghegh S, Arefi R, Ameri S, Hefner MH (1994) A methodological approach for reservoir heterogeneity characterization using Artificial Neural Networks. *SPE Annual Technical Conference and Exhibition, New Orleans, Louisiana, (SPE-28394-MS)*. <https://doi.org/10.2118/28394-MS>
- Morteza, A., Alireza, S., Amir, H., Amirshahriar, R., Mehdi, H., (2014) Application of progressive quasistatic (PQS) algorithm in prediction of water saturation in tight gas sandstones - a case study. Paper presented at the 20th Formation Evaluation Symposium of Japan, Chiba, Japan. Paper No. SPWLA-JFES-2014-BB
- Otoo D, Hodgetts D (2020) Porosity and permeability prediction through forward stratigraphic simulations using GPM and petrel: application in shallow marine depositional settings. *Geoscientific model development discussions* 14:2075–2095. <https://doi.org/10.5194/gmd-2020-37>
- Pereira, J. L. L. (2004) Permeability prediction from well log data using multiple permeability prediction from well log data using multiple regression analysis regression analysis. Graduate Theses, Dissertations, and Problem Reports. <https://researchrepository.wvu.edu/etd/1507>
- Ravasi M, Vasconcelos I, Curtis A, Kritski A (2015) Vector-acoustic reverse time migration of Volve ocean-bottom cable data set without up/down decomposed wavefields. *Geophysics* 80(4): S137–S150. <https://doi.org/10.1190/geo2014-0554.1>
- Sen S, Ganguli SS (2019) Estimation of pore pressure and fracture gradient in Volve Field, Norwegian North Sea. *SPE Oil and Gas India Conference and Exhibition, Mumbai, India, (SPE-194578-MS)*. <https://doi.org/10.2118/194578-ms>
- Singh NP (2019) Permeability prediction from wireline logging and core data: a case study from Assam-Arakan basin. *Journal of Petroleum Exploration and Production Technology* 9(1):297–305. <https://doi.org/10.1007/s13202-018-0459-y>
- Skalinski M, Kenter JAM (2015) Carbonate petrophysical rock typing: integrating geological attributes and petrophysical properties while linking with dynamic behaviour. *Geological Society Special Publication* 406(1):229–259. <https://doi.org/10.1144/SP406.6>
- Tembely M, AlSumaiti AM, Alameri WS (2021) Machine and deep learning for estimating the permeability of complex carbonate rock from X-ray micro-computed tomography. *Energy Reports* 7:1460–1472. <https://doi.org/10.1016/j.egy.2021.02.065>
- Timur A (1968) An investigation of permeability, porosity and residual water saturation relationships for sandstone reservoirs. *The Log Analyst* 9:3–5
- Tixier MP (1949) Evaluation of permeability from log resistivity gradients. *Oil and Gas Journal* 48:113–122
- Uyanık GK, Güler N (2013) A study on multiple linear regression analysis. *Procedia - Social and Behavioral Sciences* 106:234–240. <https://doi.org/10.1016/j.sbspro.2013.12.027>
- Valenzuela, O, Zhang, M, Selpi, S (2017) Combining support vector regression with scaling methods for highway tollgates travel time and volume predictions. *Proceedings of International Work-*

- Conference on Time Series Analysis (ITISE 2017) 1:411–421. <https://research.chalmers.se/en/publication/251312>
- Wadsworth FB, Vossen CEJ, Schmid D, Colombier M, Heap MJ, Scheu B, Dingwell DB (2020) Determination of permeability using a classic Darcy water column. *American Journal of Physics* 88(1):20–24. <https://doi.org/10.1119/10.0000296>
- Wen Z, Tao Z, Chengzao J, Xiangfang L, Keliu W, Minxia H (2020) Numerical simulation on natural gas migration and accumulation in sweet spots of tight reservoir. *Journal of Natural Gas Science and Engineering* 81:103454. <https://doi.org/10.1016/j.jngse.2020.103454>
- Wendt WA, Sakurai S, Nelson PH (1986) Permeability prediction from well logs using multiple regression. *Reservoir characterization* 181–221. <https://doi.org/10.1016/b978-0-12-434065-7.50012-5>
- Wiener JM, Rogers JA, Rogers JR, Moll RE (1991) Predicting carbonate permeabilities from wireline logs using a back-propagation neural network. *SEG Annual Meeting* 1991:285–288. <https://doi.org/10.1190/1.1888943>
- Wood DA (2020) Predicting porosity, permeability and water saturation applying an optimized nearest-neighbour, machine-learning and data-mining network of well-log data. *Journal of Petroleum Science and Engineering* 184:106587. <https://doi.org/10.1016/j.petrol.2019.106587>
- Yang E, Fang Y, Liu Y, Li Z, Wu J (2020) Research and application of microfoam selective water plugging agent in shallow low-temperature reservoirs. *Journal of Petroleum Science and Engineering* 193:107354. <https://doi.org/10.1016/j.petrol.2020.107354>
- Yao CY, Holditch SA (1993) Estimating permeability profiles using core and log data. *SPE Eastern Regional Meeting*, Pittsburgh, Pennsylvania, (SPE-26921-MS). <https://doi.org/10.2118/26921-ms>
- Zanjani MS, Salam MA, Kandara O (2020) Data-driven hydrocarbon production forecasting using machine learning techniques. *International Journal of Computer Science and Information Security* 18(6):65–72
- Zheng S, Xiangfang L, Wenyuan L, Tao Z, Minxia H, Hadi N (2020) Molecular dynamics of methane flow behavior through realistic organic nanopores under geologic shale condition: pore size and kerogen types. *Chemical Engineering Journal* 398:124341. <https://doi.org/10.1016/j.ccej.2020.124341>
- Zhong H, Yang T, Yin H, Lu J, Zhang K, Fu C (2020) Role of alkali type in chemical loss and ASP-flooding enhanced oil recovery in sandstone formations. *SPE Reservoir Evaluation and Engineering* 23(2): 431–445. <https://doi.org/10.2118/191545-PA>