Vietnam Academy of Science and Technology

Vietnam Journal of Marine Science and Technology

journal homepage: vjs.ac.vn/index.php/jmst

# Applying machine learning for hydraulic flow unit classification and permeability prediction: case study from carbonate reservoir in the Southern part, Song Hong basin

**Dung Nguyen Trung[1,*], Man Ha Quang[1], Hoa Truong Khac[1], Huong Phan Thien[2], Hoang Cu Minh[1], Hong Nguyen Viet[3]**

[1]*Petrovietnam Exploration Production Corporation, Hanoi, Vietnam*
[2]*Faculty of Oil and Gas, Hanoi University of Mining and Geology, Hanoi, Vietnam*
[3]*Schlumberger Vietnam, Hochiminh city, Vietnam*

Received: 29 March 2024; Accepted: 18 May 2024

## ABSTRACT

Machine learning (ML) is an artificial intelligence (AI) that enables computer systems to classify, cluster, identify, and analyze vast and complex sets of data while eliminating the need for explicit instructions and programming. For decades, machine learning has become helpful for complex reservoir characterization such as carbonate reservoirs. Permeability prediction from well logs is a significant challenge, especially when the core data is rarely available due to its high cost. In this study, we aimed to bridge this gap by demonstrating the practical application of integrating Hydraulic Flow Units (HFU) and machine learning methods. Our goal was to provide a reliable estimation of permeability using core and wireline logging data in the complex Middle Miocene carbonate reservoir of the CX gas field in the southern part of the Song Hong basin. In the first step, due to the reservoir's heterogeneity, the core plug dataset was classified into 5 HFUs based on the flow zone indicators (FZI) concept from the modified Kozeny-Carman equation using unsupervised machine learning - K-means method. The porosity - permeability for each HFU was defined after HFU clustering. In the second step, we designed three different workflows to predict permeability and HFU using supervised machine learning from a combination of core and log data. These workflows were rigorously test and compared with the core data. The most accurate result was chosen as the base, providing a high confidence level in our predictions' reliability.

**Keywords:** Hydraulic Flow Unit, machine learning, carbonate reservoir, clustering, permeability prediction, Song Hong basin.

[*]Corresponding author at: Petrovietnam Exploration Production Corporation, 26th Floor, Charm Vit Tower, 117 Tran Duy Hung St., Cau Giay Dist.,Hanoi, Vietnam. *E-mail addresses:* dungnt-tktdtn@pvep.com.vn

https://doi.org/10.15625/1859-3097/18654
ISSN 1859-3097; e-ISSN 2815-5904/

## INTRODUCTION

Carbonate reservoirs challenges for engineers and geologists to characterize because of their complexity due to depositional and diagenetic processes. The extreme petrophysical heterogeneity found in carbonate reservoirs is demonstrated by the wide variability observed, especially in porosity-permeability crossplots of core data analysis [1–3].

Permeability is an essential parameter for reservoir description. Generally, estimating permeability from well logs is the lowest-cost method, but predicting permeability in heterogeneous carbonates from well log data represents challenging and complex problems. Applying well-known and described mathematical models for such reservoirs' porosity-permeability relationship is complex.

Characterization of carbonate reservoirs into hydraulic flow units (HFU) is a practical way of reservoir zonation [4–8, 9]. The presence of distinct units with petrophysical characteristics such as porosity, permeability, water saturation, pore throat radius, storage, and flow capacities helps researchers to establish strong reservoir characterization. Moreover, the permeability can be calculated from the porosity-permeability relationship built for each HFU.

A hydraulic flow unit is the representative volume of total reservoir rock within which geological properties that control fluid flow are internally consistent and predictably different from properties of other rocks [4]. A flow unit is a reservoir zone that is continuous laterally and vertically and has similar flow and bedding characteristics.

However, how does one correctly cluster the HFU, and what is the optimal number of HFUs that need to be classified? How can permeability be predicted from well-log data, and how accurate is the predicted permeability? These issues will be resolved by using modern machine-learning techniques.

The application of machine learning (ML) has become popular in recent decades. ML is an artificial intelligence (AI) that enables computer systems to classify, cluster, identify, and analyze vast and complex sets of data while eliminating the need for explicit instructions and programming.

Machine learning is a practical empirical approach for regression and/or classification (supervised or unsupervised) of nonlinear systems. Such systems can be massively multivariate, involving a few or literally thousands of variables [10].

Unsupervised machine learning methods will be used to group the HFU, and the results will be compared to choose the best one. The elbow method will define the optimal number of HFUs that must be classified for the study area.

Three workflows will be used to predict permeability directly from wireline logging data or FZI/HFU data. Then, permeability can be estimated using the Amaefule equation. Series of supervised machine-learning methods will be used for these purposes.

## GEOLOGICAL SETTING

The CX gas field is in the southern part of the Song Hong basin on Triton horst (Fig. 1). The Song Hong basin extends to more than 220,000 $km^2$ from northern to central Vietnam, mainly over fractured and weathered Mesozoic basement. The greater East Vietnam Sea area has a relatively simple tectonic history. In the Eocene to Early Oligocene, extension (rifting) was initiated, followed by movements along the Red River/East Vietnam Boundary Fault Zone approximately 45 million years ago. The Red River Fault is a major left-lateral strike-slip fault system caused by the collision of the India and Asia tectonic plates. Motion on the Red River Fault System continued into the Early Miocene, and marine conditions, triggering carbonate platform growth, were established East of the fault zone on the top of structural highs. Carbonate growth was widespread on structural highs offshore Central Vietnam throughout the Early and Middle Miocene. Regional uplift of the mainland resulted in an increased influx of siliciclastics from the West, leading to stressed carbonate growth and finally to the drowning of the Tri Ton host carbonate platform during the Late Miocene [11].

The stratigraphic succession can be summarized as in Fig. 2.

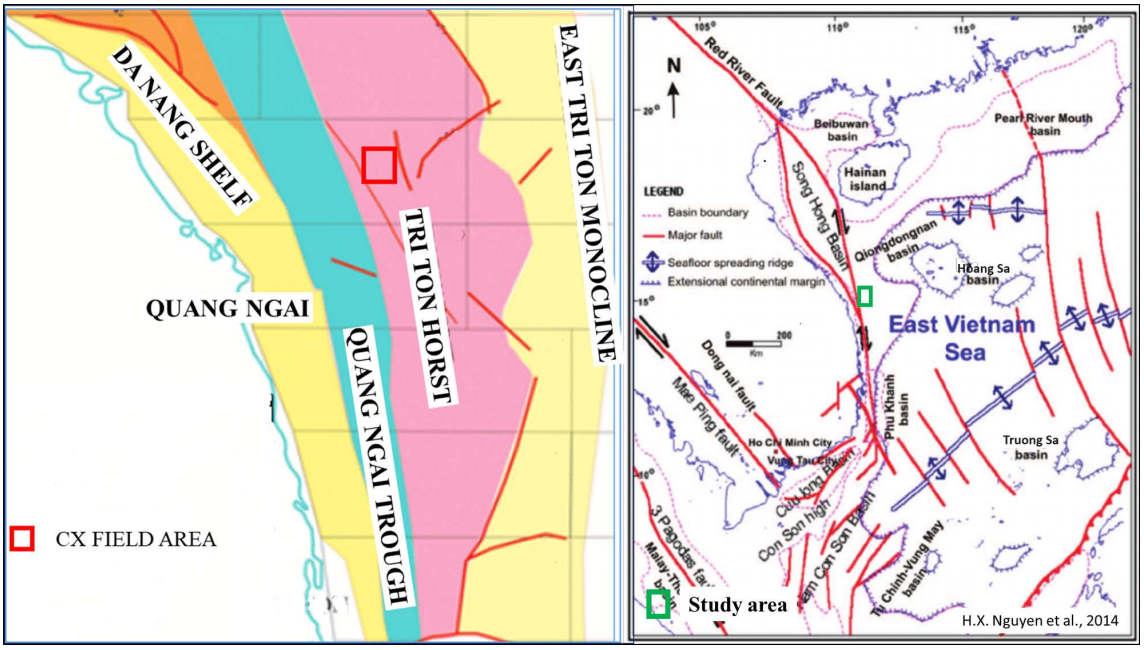


*Figure 1.* Study area location

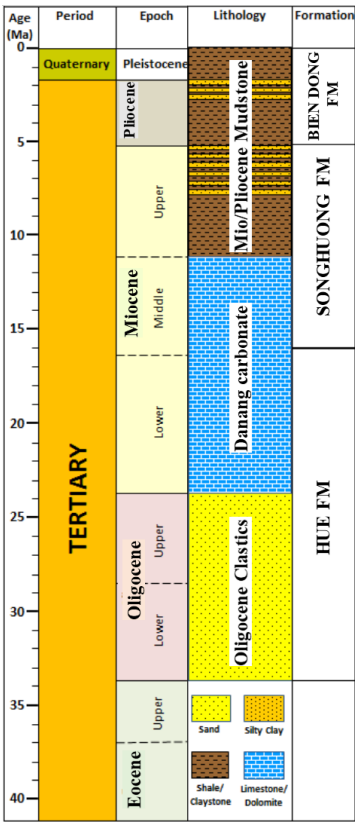


*Figure 2.* Triton horst area - Generalized stratigraphic column [12]

Middle Miocene age carbonate platforms of the Da Nang Formation nucleated on older, remnant syn-rift highs.

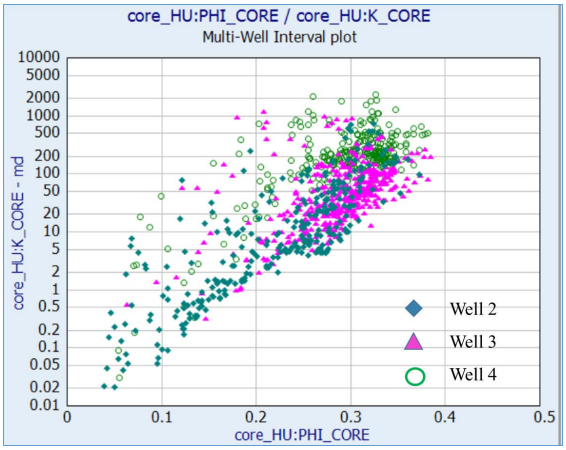


*Figure 3.* PHI_core and K_core of CX field crossplot

Reservoir rocks are the Middle Miocene age carbonates that developed on an isolated platform (length approximately 100 km, and width approximately 15 km) on top of Triton Horst structural high. The reservoir is heterogeneous with different types of pores:

interparticle, vuggy and fractures, with wide range of porosity from several percent to over 35% and high variation of permeability, from less than 1 mD to over 2,000 mD (Fig. 3).

## METHODOLOGY AND DATASET

From the original Kozeny-Carman equation, Amaefule et al., (1993) introduced two auxiliary factors: PHIZ, the normalized porosity (1), and RQI, the reservoir quality index (2). This results in a new formula that defines the Flow Zone Indicator (FZI) (3) regarding porosity - permeability relationships and accurately approximates the reservoir quality for given sedimentary facies.

The basis of HFU classification is to identify data classes that fall into that plot as a log-log graph of RQI and PHIZ. Permeability is calculated from the HFU of a sample by substituting for the mean value of FZI in the equation (4) below:

$$PHIZ = \left( \frac{\phi_e}{1-\phi_e} \right) \quad (1)$$

$$RQI = 0.0314 \sqrt{\frac{k}{\phi_e}} \quad (2)$$

$$FZI = RQI/PHIZ \quad (3)$$

$$K = 1{,}014.24 (FZI)^2 \frac{\phi_e^3}{(1-\phi_e)^2} \quad (4)$$

The study was done in 2 steps: the first was to classify the HFU using core data, and the second was to predict the permeability K - HFU for an uncored interval using the integration of well log data and core data.

The workflow applied to the study is shown in the Figure 4.

The FZI method has been widely used for HFU classification for over two decades. Although traditional methods using a probabilistic plot or histogram of FZI are still being used for HFU classification, it shows the limitation that one cannot see clearly on the chart the classified HFU as the groups of points are not clearly separated, the histogram does not reveal the distribution of the groups (Figure 5).

So, for the *first step of the study*, unsupervised machine learning with widely used methods such as K-means, Ward's hierarchical, Self-Organizing Maps and Fuzzy-C Mean has been used [12–15]. The optimal number of HFU will be chosen based on the Elbow method.

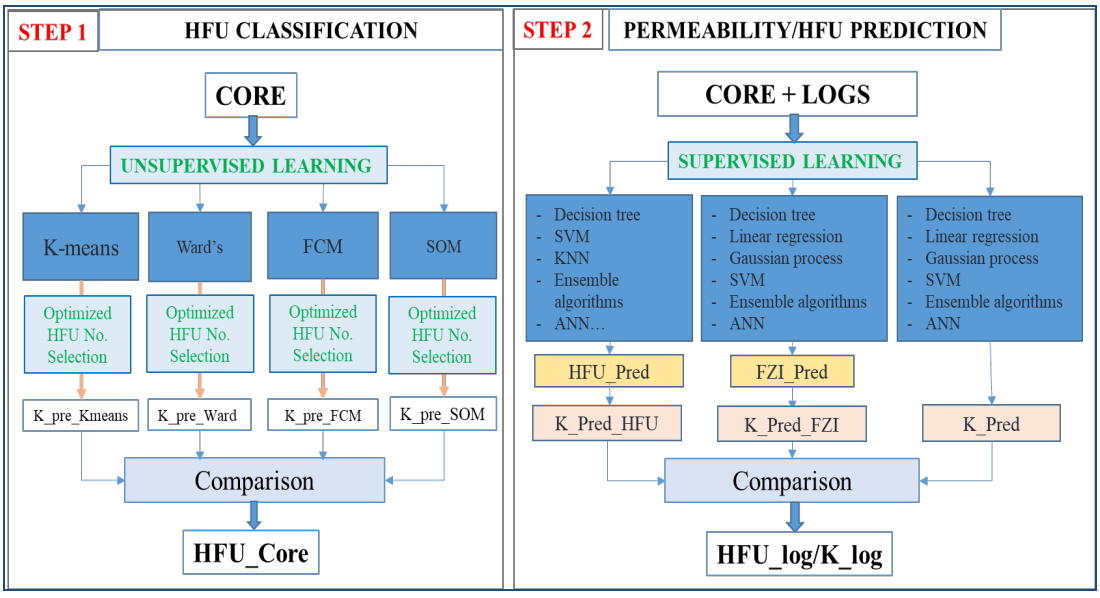


*Figure 4.* Workflows for the study

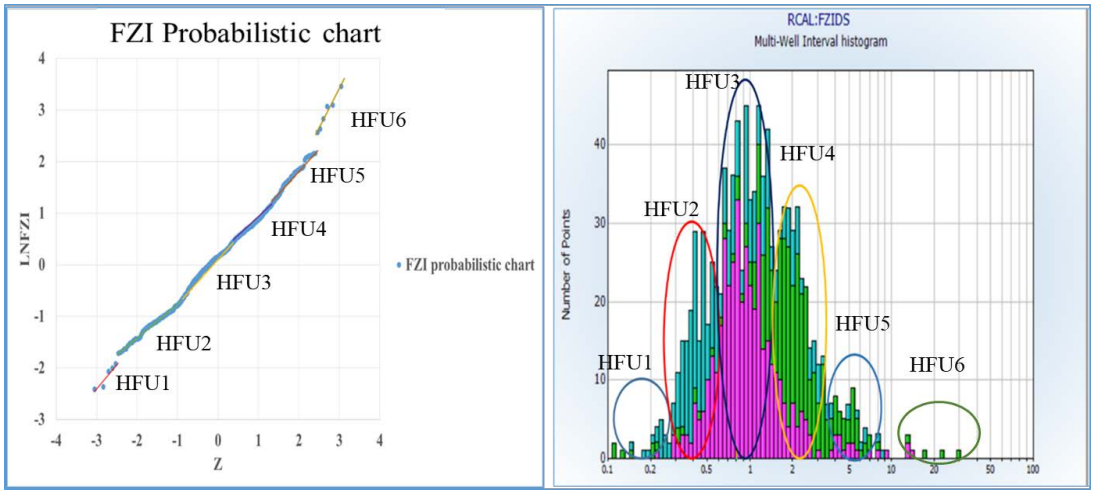


*Figure 5.* Probabilistic chart and histogram for HFU clustering (the clusters cannot be seen clearly on both figures)

## K-means clustering

K-means clustering is a typical unsupervised learning algorithm that tries to group a data set into K clusters so that items in the same cluster are close together while items in different clusters are dispersed. Hence, the K-means clustering process shrinks the distance between points in the same cluster and expands the distance between points when the clusters differ. The advantage of this clustering technique is speed. However, the problem with K-means is that the number of clusters must be known in advance, which is often a non-trivial task. Moreover, this method performs well with spherical clusters and when each cluster has equal numbers for observations. The algorithm works ineffectively with clusters of unusual size. The K-means method uses the following steps:

Step 1: Choose the number of clusters K;

Step 2: Select k random points from the data as center values;

Step 3: Assign all the points to the closest cluster centers;

Step 4: Recompute the new center values;

Step 5: Repeat steps 3 and 4 until termination conditions are reached.

## Ward's hierarchical algorithm

Hierarchical clustering is a popular clustering technique. Each data point forms a discrete cluster in Ward's agglomerative hierarchical technique. Iteratively, similar discrete clusters merge into superclusters. The data point similarity is measured by summing the square of the distances between them. The hierarchical clustering output is a dendrogram that shows the cluster hierarchy.

This approach works well if the data is spherical, multivariate, and normally distributed. Additionally, the clustering of this method is good only if there is an equal number of data points in each population. Ward's Hierarchical clustering uses the following steps:

Step 1: Calculate the proximity of individual points and consider all the data points as individual clusters;

Step 2: Merge similar clusters to form a single cluster;

Step 3: Recalculate the proximity of new clusters;

Step 4: Repeat steps 2 and 3 until termination conditions are reached.

## Self-organizing map

The Self-organizing map (SOM) algorithm implemented in this study is an adaptive learning process. Neurons learn to represent discrete input data. The neuron that best approximates an input vector becomes the winning neuron. Other neurons learn to represent similar inputs. Neurons are then

placed at the nodes in a lattice to convert multi-dimensional data into a 1D and 2D discrete map. The SOM clustering method uses the following steps:

Step 1: Initialize random weight vector;

Step 2: Choose a random input vector from the training data;

Step 3: Check all neurons to define the winning one that is the Best Matching Unit (BMU);

Step 4: Update the neuron winner by calculating the neighborhood of the BMU, noting that the number of neighbors decreases over time;

Step 5: Repeat steps 2–4 until termination conditions are reached.

### Fuzzy C-means clustering

Fuzzy clustering is a powerful unsupervised learning method for analyzing data and constructing models. It is more natural than arbitrary hard clustering such as K-means. This method is derived from fuzzy logic, suitable for solving ambiguous problems. The data points can belong to multiple clusters with different membership degrees. Fuzzy C-means uses fuzzy partitioning, meaning that a data point can belong to any or all groups. The degree of membership is graded between 0 and 1. This method also needs prior information on number clusters.

The fuzzy c-means clustering uses the following steps:

Step 1: Set the number of clusters k;

Step 2: Randomly initialize k center values;

Step 3: Calculate the membership degree of each data point;

Step 4: Calculate new center values;

Step 5: Repeat steps 3 and 4 until termination conditions are reached.

The elbow method is a heuristic method used to determine the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use (Fig. 6). The elbow method is based on the idea that as the number of clusters increases, the variation within each cluster decreases. However, some point, the improvement becomes negligible. This point is called the elbow, indicating the optimal number of data clusters.

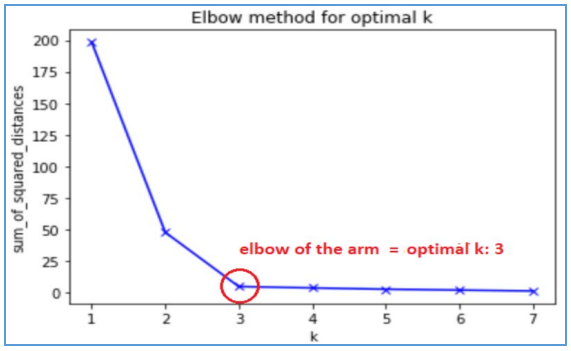


*Figure 6.* Example of Elbow method to define optimal cluster number

*In the second step of the study*, the supervised machine learning methods were used for K and HFU prediction.

Supervised machine learning uses a known dataset (called the training dataset) to train an algorithm with a known set of input data (called features or predictors) and known responses to make predictions. The training dataset includes labeled input data that pair with desired outputs or response values. From it, the supervised learning algorithm seeks to create a model by discovering relationships between the predictors and response. Then, it makes predictions of the response values for a new dataset.

In Figure 4, supervised machine learning with classification algorithms were used for the HFU prediction. The classification algorithms are used for categorical response values, where the data can be separated into specific classes. In this study, HFUs were classified in the first step, and almost all classification algorithms, including Logistic regression, Support vector machines (SVM), Neural networks, Naïve Bayes classifier, Decision trees, Discriminant analysis, Nearest neighbors (kNN), Ensemble classification were applied to predict HFU. After that, permeability K_HFU_Pred can be estimated using the K-PHI relationship for each HFU from the 1$^{st}$ step.

As FZI and K are continuous-response values, regression algorithms were used to predict them. Common regression algorithms, including Linear regression, Nonlinear regression, Generalized

linear models, Decision trees, Neural networks, Gaussian Process Regression, Support vector machine Regression, Ensemble Regression were applied for FZI and K prediction. Once FZI is predicted, HFU number can be defined and K_FZI_pred can be estimated using the K-PHI relationship of each HFU defined in the 1st step of the study.

The base method will be chosen as the most appropriate, with a high correlation coefficient and low Root mean squared error (RMSE).

The predicted permeability of K_HFU_pred, K_FZI_pred and K_pred will be compared with core data, and the method with the most accurate predicted permeability will be used.

### Dataset

Well 2, Well 3, Well 4 were cored and logged through the gas column in the carbonate reservoir. The study dataset contained over 1,000 core plugs of RCAL and conventional logging data (Gamma-ray, Deep Resistivity, Shallow Resistivity, Micro Resistivity, Density, Neutron, Compressional and Shear Sonic logs) over the Middle Miocene carbonate reservoir of 3 wells: 2, 3 and 4. Laboratory core measurements included porosity (PHI_Core) and absolute permeability (K_Core), taken from various depths of carbonate section in the above wells. The log data quality is good and can be used for quantitative analysis.

The open hole logs in Well-1 were also acquired, but the log data quality is poor, so the quantitative analysis could not be done.

## RESULTS AND DISCUSSION

### Unsupervised machine learning for HFU clustering

The results of the unsupervised machine learning methods are summarized in Table 1 and Figs. 7–10. The elbow point clearly shows the optimal number of HFUs for each method used for clustering. As seen in the Figures, the optimum number of K-means and Ward's method is 5 (five); the optimum number of SOM and FCM methods is 4.

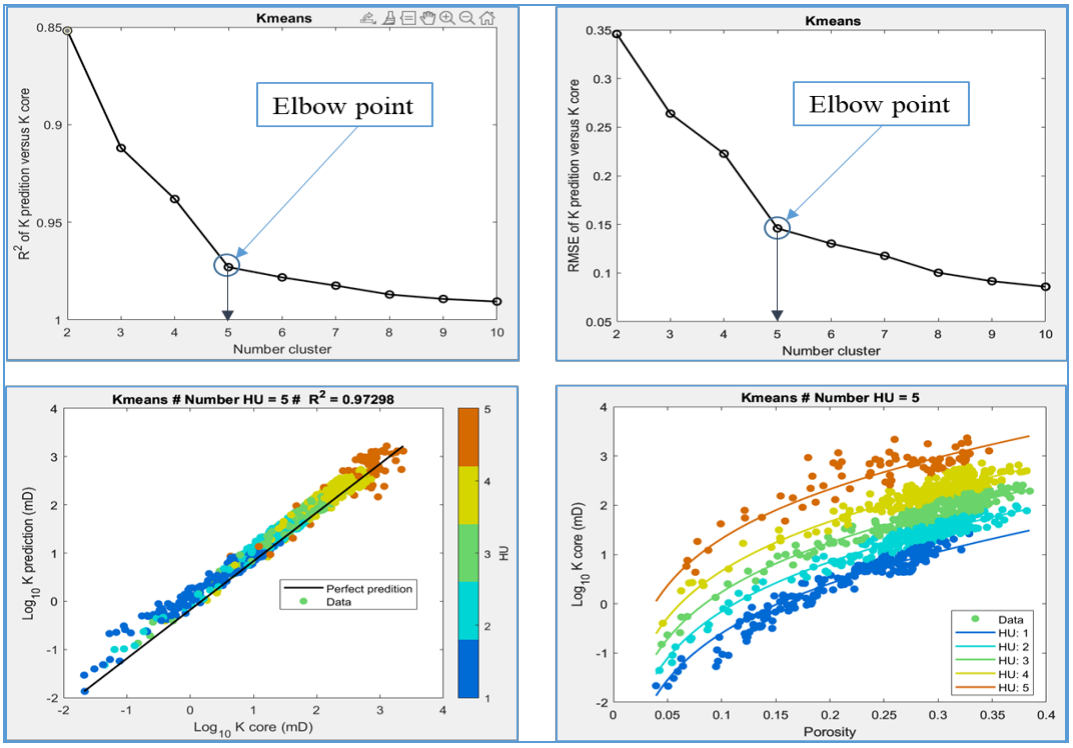


*Figure 7.* K-means clustering result with 5 HFU

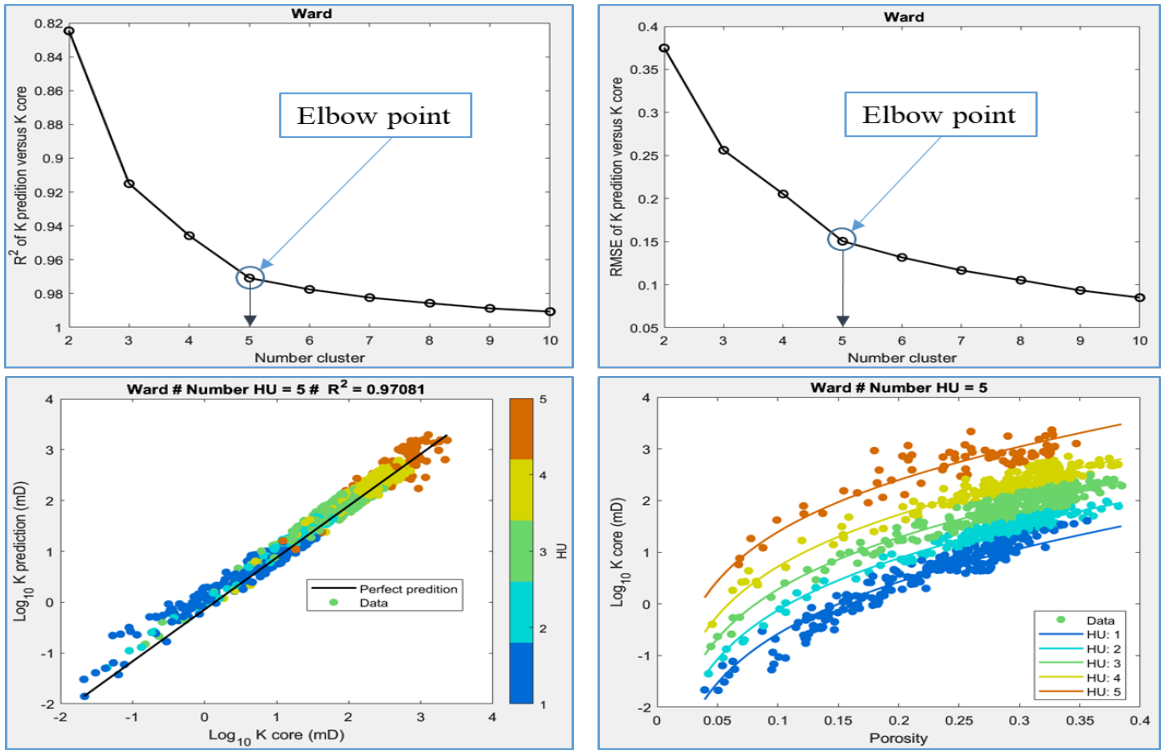


*Figure 8.* Ward's Hierarchical clustering result with 5 HFU

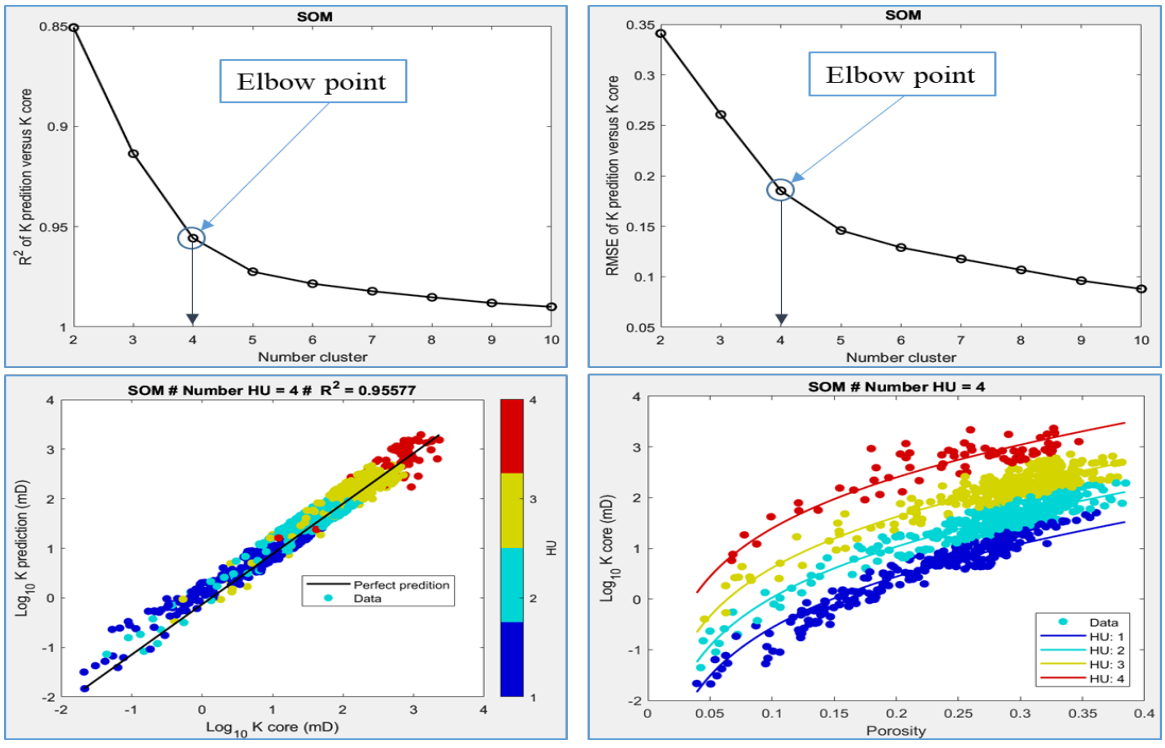


*Figure 9.* SOM clustering result with 4HFU

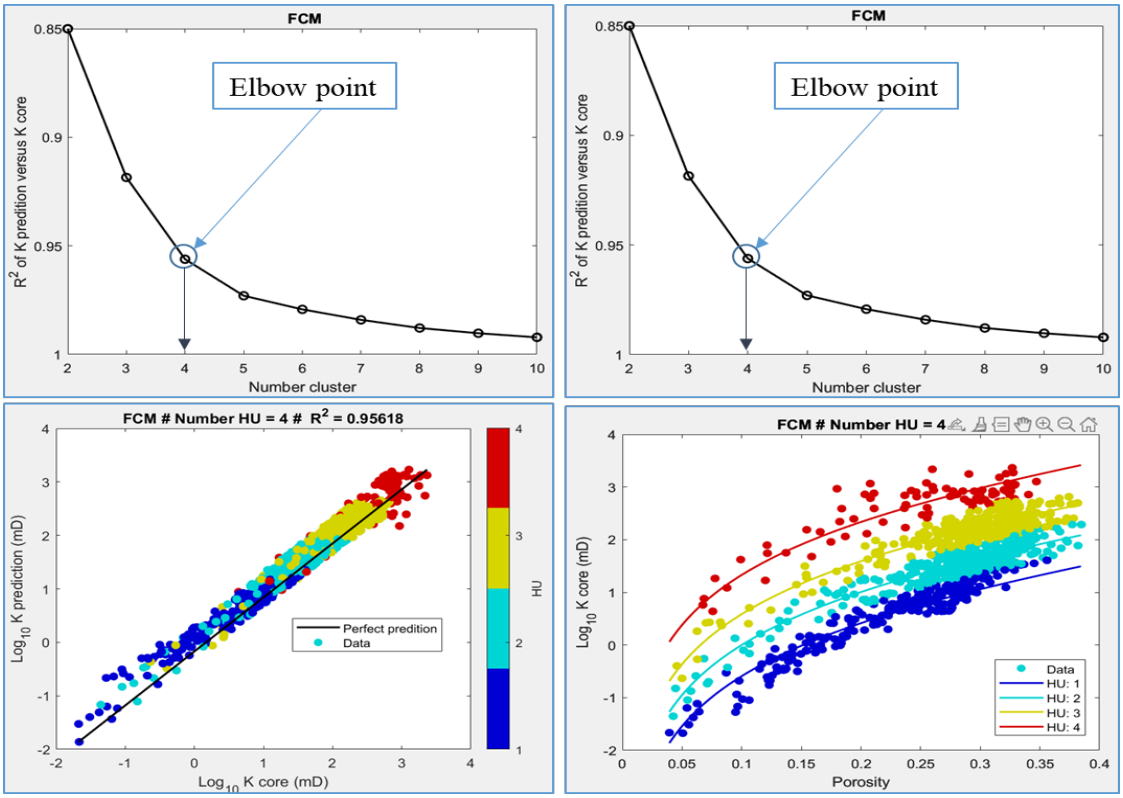


*Figure 10.* FCM clustering result with 4 HFU

*Table 1.* Comparison of R2 and RMSE of clustering methods (Calculated permeability using clustering result K_prediction vs K_core)

| | K-means (5 HFU) | FCM (4 HFU) | Ward (5 HFU) | SOM (4 HFU) |
|---|---|---|---|---|
| R2 | 0.9730 | 0.9562 | 0.9708 | 0.9558 |
| RMSE | 0.1459 | 0.1840 | 0.1505 | 0.1851 |

Table 1 shows the correlation coefficient (R2) and root mean square error (RMSE) of K_core and the calculated permeability K_prediction for each method. The table shows that R2 and RMSE are the same for all applied methods. All methods give the calculated permeability K_prediction with a high correlation coefficient R2 and a low RMSE value, with the K-means method with 5 HFUs giving the highest R2 and the lowest RMSE. Thus, the clustering results based on the K-means method will be used for the next steps.

Table 2 and Figure 11 show the final simple statistics of FZI and the distribution of K_core, PHI_core, and FZI_core from the K-mean method results for 5 HFU groups.

*Table 2.* Statistic of FZI and K-PHI relationship for classified HFU

| | HFU 1 | HFU 2 | HFU 3 | HFU 4 | HFU 5 |
|---|---|---|---|---|---|
| FZI range | 0.2243–0.5795 | 0.5795–0.9374 | 0.9374–1.5 | 1.50–2.811 | 2.811–10.2688 |
| FZI_mean | 0.4433 | 0.7566 | 1.1915 | 1.9679 | 4.4991 |
| K-PHI relationship | K = 204.375*PHI^3/ (1 – PHI)^2 | K = 595.341*PHI^3/ (1 – PHI)^2 | K = 1476.46*PHI^3/ (1 – PHI)^2 | K = 4001.38*PHI^3/ (1 – PHI)^2 | K = 21051.6*PHI^3/ (1 – PHI)^2 |

Figure 11 shows the distribution of K_core, PHI_core and FZI_core for each classified HFU.

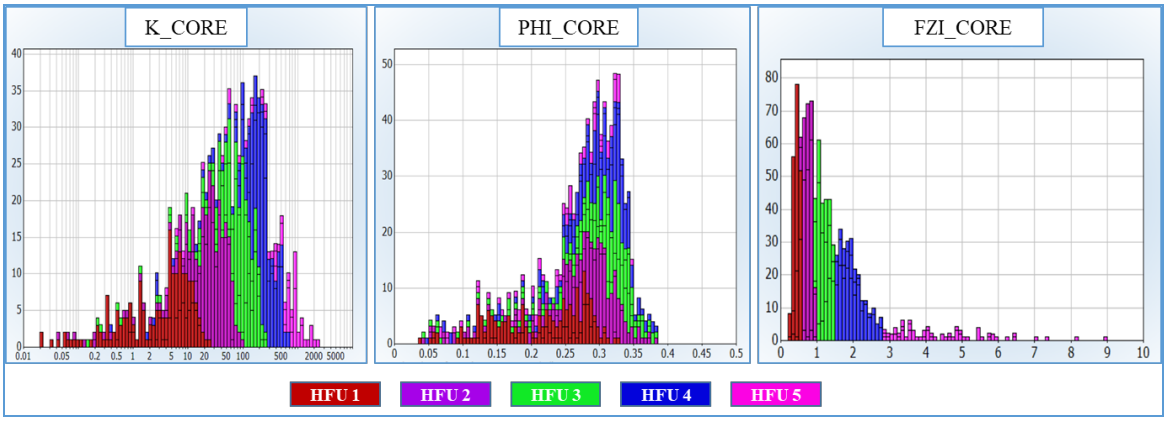


*Figure 11.* Histogram distribution of K_core, PHI_core and FZI_core for each classified HFU

## Supervised machine learning for K-HFU prediction

### *Data preparation*

Core and wireline logging data were checked for depth match and outlier removal. The correlation analysis was done to check the strength of the core and log curves (GR, RD, RS, MSFL, RHOB, NPHI, DTC, DTS) correlations.

Table 3 shows the degree of correlation between well logs and the K_core, FZI_core and HFU_core. All the above logging data can be used for supervised machine learning.

*Table 3.* Covariance of Core data and wireline logging data

| | PHI_CORE | K_CORE | FZI_CORE | HFU_CORE | GR | RD | RS | MSFL | RHOB | NPHI | DTC | DTS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PHI_CORE | 1 | | | | | | | | | | | |
| PHI_CORE | 0.758 | 1 | | | | | | | | | | |
| PHI_CORE | 0.024 | 0.614 | 1 | | | | | | | | | |
| PHI_CORE | 0.269 | 0.789 | 0.796 | 1 | | | | | | | | |
| PHI_CORE | -0.204 | -0.515 | -0.478 | -0.616 | 1 | | | | | | | |
| PHI_CORE | 0.291 | 0.461 | 0.305 | 0.498 | -0.497 | 1 | | | | | | |
| PHI_CORE | 0.194 | 0.387 | 0.295 | 0.469 | -0.431 | 0.936 | 1 | | | | | |
| PHI_CORE | -0.422 | 0.029 | 0.481 | 0.453 | -0.455 | 0.426 | 0.494 | 1 | | | | |
| PHI_CORE | -0.845 | -0.762 | -0.192 | -0.367 | 0.339 | -0.496 | -0.384 | 0.210 | 1 | | | |
| PHI_CORE | 0.668 | 0.480 | -0.084 | 0.122 | -0.275 | 0.055 | 0.004 | -0.399 | -0.567 | 1 | | |
| PHI_CORE | 0.835 | 0.711 | 0.126 | 0.303 | -0.314 | 0.451 | 0.368 | -0.259 | -0.928 | 0.663 | 1 | |
| PHI_CORE | 0.627 | 0.604 | 0.222 | 0.359 | -0.412 | 0.487 | 0.429 | -0.042 | -0.715 | 0.439 | 0.769 | 1 |

### *Training*

For all 3 suggested workflows, the responses will be FZI (calculated using equation (3)), K (from core analysis) and HFU (classified in first step), respectively, and the predictors will be GR, RD, RS, MSFL, RHOB, NPHI, DTC and DTS.

The core-log data was divided into training data (60%), validating data (25%) and testing data (15%) to ensure that the training results were good, not underfitting or overfitting.

*Table 4.* Validation and testing result for the machine learning models applied for HFU prediction

| No. | Model Type | | HFU PREDICTION | | | |
|---|---|---|---|---|---|---|
| | | | Accuracy (%) | | Total Cost | |
| | | | Validation | Test | Validation | Test |
| 1 | Tree | Fine | 70.7 | 79.1 | 219 | 52 |
| 2 | | Medium | 63.1 | 64.7 | 276 | 88 |
| 3 | | Coarse | 54.9 | 51.4 | 337 | 121 |
| 4 | Discriminant | Linear | 55.9 | 55.8 | 330 | 110 |
| 5 | | Quadratic | 58.7 | 60.6 | 309 | 98 |
| 6 | Logistic | Efficient Logistic Regression | 40.8 | 45.0 | 443 | 137 |
| 7 | Naive Bayes | Gaussian | 52.1 | 57.0 | 358 | 107 |
| 8 | | Kernel | 58.2 | 63.9 | 313 | 90 |
| 9 | SVM | Linear | 57.4 | 58.6 | 319 | 103 |
| 10 | | Quadratic | 69.8 | 77.5 | 226 | 56 |
| 11 | | Cubic | 78.9 | 80.3 | 158 | 49 |
| 12 | | Fine Gaussian | 75.8 | 82.3 | 181 | 44 |
| 13 | | Medium Gaussian | 64.4 | 66.3 | 266 | 84 |
| 14 | | Coarse Gaussian | 51.1 | 52.2 | 366 | 119 |
| 15 | KNN | Fine | 80.5 | 85.9 | 146 | 35 |
| 16 | | Medium | 62.7 | 67.5 | 279 | 81 |
| 17 | | Coarse | 49.2 | 51.8 | 380 | 120 |
| 18 | | Cosine | 65.4 | 67.9 | 259 | 80 |
| 19 | | Cubic | 63.0 | 69.9 | 277 | 75 |
| 20 | | Weighted | 80.7 | 85.9 | 144 | 35 |
| 21 | Ensemble | Boosted tree | 69.1 | 73.9 | 231 | 65 |
| 22 | | Bagged tree | 81.7 | 84.3 | 137 | 39 |
| 23 | | Subspace discriminant | 55.2 | 55.0 | 335 | 112 |
| 24 | | Subspace KNN | 78.7 | 85.1 | 159 | 37 |
| 25 | | RUSBoosted tree | 67.6 | 70.3 | 242 | 74 |
| 26 | | Optimizable | 82.5 | 85.9 | 131 | 35 |
| 27 | Neural Network | Narrow | 69.0 | 70.7 | 232 | 73 |
| 28 | | Medium | 77.0 | 81.5 | 172 | 46 |
| 29 | | Wide | 79.4 | 85.9 | 154 | 35 |
| 30 | | Bilayered | 72.2 | 77.5 | 208 | 56 |
| 31 | | Trilayered | 78.1 | 79.5 | 164 | 51 |

Overfitting is a machine learning behavior that occurs when the model is so closely aligned to the training data that it does not know how to respond to new data. It can happen when the R2 of training data is much higher than the R2 of testing data.

Underfitting is the opposite of overfitting; the model does not align well with the training data or generalize well to new data. It can happen when R2 of training data is much lower than R2 of testing data.

Tables 4, 5 show the validation and testing results for three workflows.

Tables 4, 5, we can see that the Exponent Gaussian processing regression model gave the best results, with the highest R2 and

lowest RMSE for validation and testing data for the FZI and K predictions. For HFU prediction, the optimizable ensemble method gave the best result. Thus, the models were chosen for application to the entire Middle Miocene carbonate interval in three wells.

*Table 5.* Validation and testing result for the machine learning models applied for FZI and K prediction

| No. | Model Type | | FZI prediction | | | | K prediction | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | RMSE | | RSquared | | RMSE | | RSquared | |
| | | | Validation | Test | Validation | Test | Validation | Test | Validation | Test |
| 1 | Linear regression | Linear | 0.895 | 0.732 | 0.502 | 0.484 | 0.516 | 0.463 | 0.679 | 0.717 |
| 2 | | Interactions | 0.933 | 0.678 | 0.459 | 0.557 | 0.504 | 0.418 | 0.693 | 0.77 |
| 3 | | Robust | 1.120 | 0.843 | 0.220 | 0.315 | 0.518 | 0.459 | 0.677 | 0.723 |
| 4 | | Stepwise | 0.905 | 0.687 | 0.491 | 0.545 | 0.512 | 0.463 | 0.684 | 0.718 |
| 5 | Decision tree | Fine | 0.616 | 0.549 | 0.764 | 0.709 | 0.481 | 0.332 | 0.722 | 0.855 |
| 6 | | Medium | 0.736 | 0.615 | 0.663 | 0.636 | 0.481 | 0.355 | 0.721 | 0.834 |
| 7 | | Coarse | 0.816 | 0.672 | 0.587 | 0.564 | 0.543 | 0.437 | 0.644 | 0.749 |
| 8 | SVM | Linear | 1.002 | 0.755 | 0.376 | 0.450 | 0.518 | 0.461 | 0.676 | 0.72 |
| 9 | | Quadratic | 0.872 | 0.609 | 0.527 | 0.642 | 0.489 | 0.394 | 0.712 | 0.795 |
| 10 | | Cubic | 0.904 | 0.490 | 0.492 | 0.769 | 0.614 | 0.32 | 0.546 | 0.865 |
| 11 | | Fine | 0.802 | 0.453 | 0.600 | 0.803 | 0.432 | 0.282 | 0.775 | 0.896 |
| 12 | | Medium | 0.871 | 0.567 | 0.528 | 0.690 | 0.457 | 0.35 | 0.748 | 0.839 |
| 13 | | Coarse | 1.026 | 0.761 | 0.345 | 0.442 | 0.523 | 0.454 | 0.67 | 0.729 |
| 14 | Ensemble | Boosted | 0.583 | 0.507 | 0.789 | 0.752 | 0.428 | 0.343 | 0.779 | 0.845 |
| 15 | | Bagged | 0.605 | 0.494 | 0.773 | 0.764 | 0.416 | 0.287 | 0.792 | 0.892 |
| 16 | Gaussian process regression | Square exponential | 0.539 | 0.487 | 0.819 | 0.772 | 0.416 | 0.286 | 0.791 | 0.892 |
| 17 | | Matern 5/2 | 0.494 | 0.479 | 0.848 | 0.779 | 0.361 | 0.267 | 0.843 | 0.906 |
| 18 | | Exponential | 0.459 | 0.395 | 0.869 | 0.850 | 0.320 | 0.235 | 0.870 | 0.927 |
| 19 | | Rational quadratic | 0.463 | 0.423 | 0.867 | 0.828 | 0.323 | 0.235 | 0.860 | 0.927 |
| 20 | Neural network | Narrow | 0.694 | 0.593 | 0.701 | 0.661 | 0.454 | 0.352 | 0.752 | 0.837 |
| 21 | | Medium | 0.669 | 0.555 | 0.722 | 0.703 | 0.452 | 0.337 | 0.754 | 0.851 |
| 22 | | Wide | 0.632 | 0.458 | 0.752 | 0.798 | 0.5 | 0.341 | 0.699 | 0.846 |
| 23 | | Bilayered | 0.882 | 0.593 | 0.517 | 0.661 | 0.449 | 0.38 | 0.757 | 0.81 |
| 24 | | Trilayered | 0.654 | 0.834 | 0.734 | 0.329 | 0.481 | 0.315 | 0.721 | 0.869 |

### *Applying the prediction models for all study interval*

The trained models were applied to three wells' Middle Miocene carbonate interval. The K_HFU_Pred, K_FZI_Pred estimated using the K-PHI relationship from 1st step of the study (Table 2) and predicted K_pred were checked against the K_core to make sure the prediction result was good and to choose the best workflow for permeability prediction.

The charts in Fig. 12 show the correlation between predicted permeability from each workflow with K_core.

The predicted permeability and HFU using each workflow for wells 2, 3 and 4 in the study area are shown in Figs. 13–15.

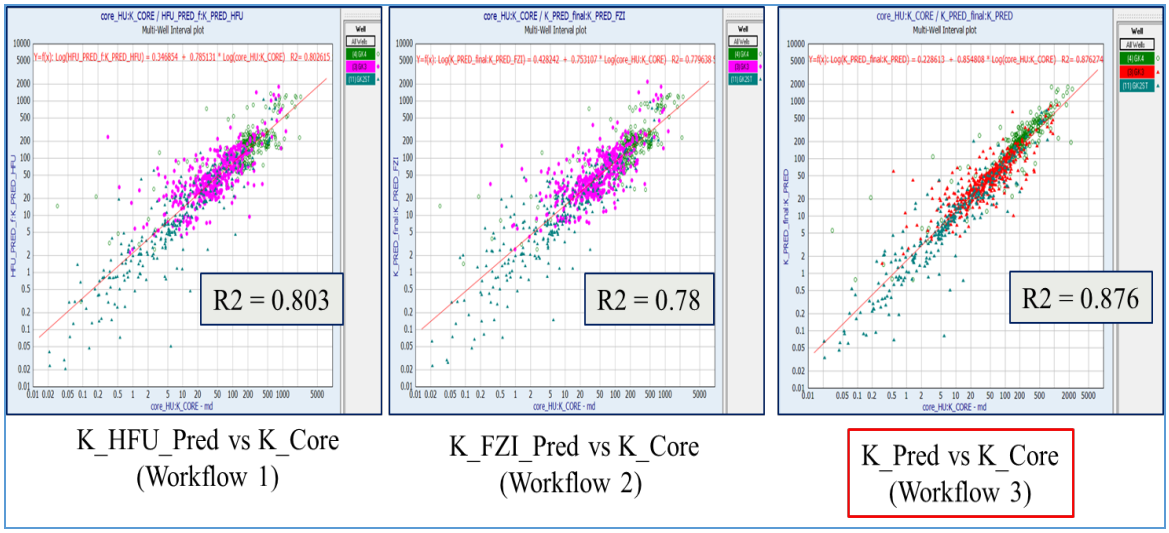


*Figure 12.* Correlation between estimated permeability from each workflow with K_core

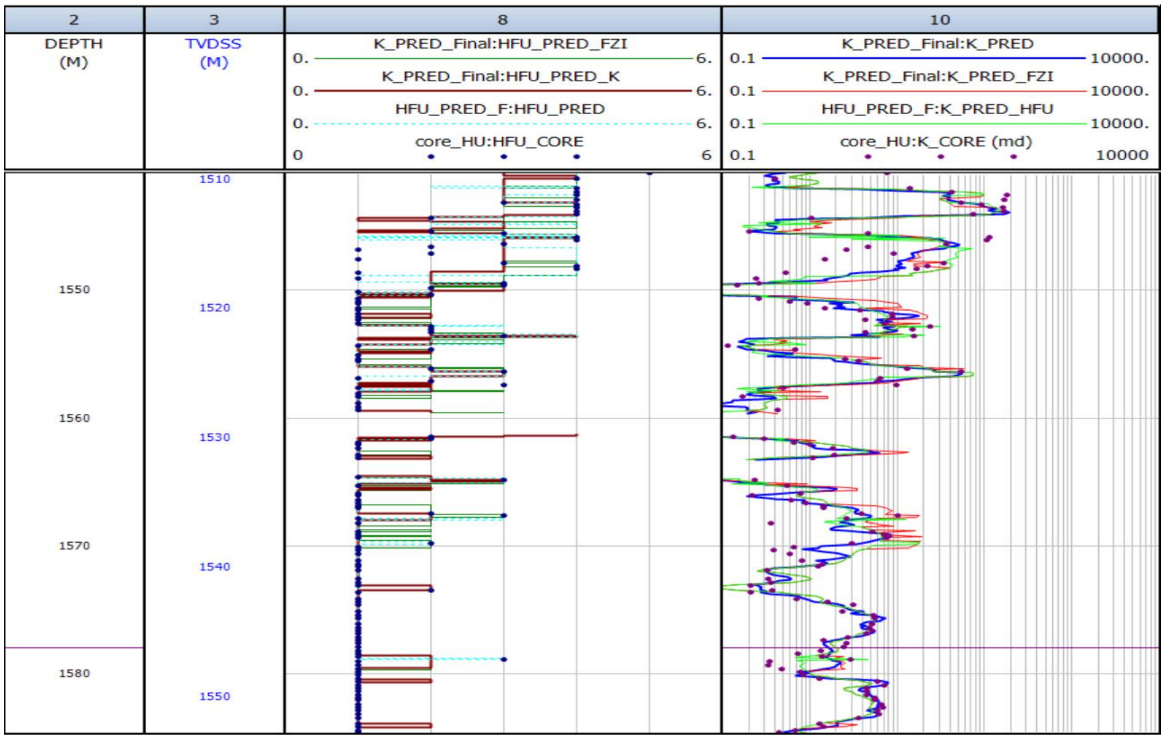


*Figure 13.* The predicted HFU and permeability vs HFU_core and K_core in Well 2

The Figures above show that all workflows' predicted K and HFU are matched with core data over the interested intervals.

Figures 12–15 show that although the correlation coefficient R2 between the predicted permeability from 3 workflows and K_core is relatively high, the directly predicted permeability using machine learning from workflow 3 gives the best correlation coefficient. Therefore, this permeability prediction result will be used as the base case for the study.

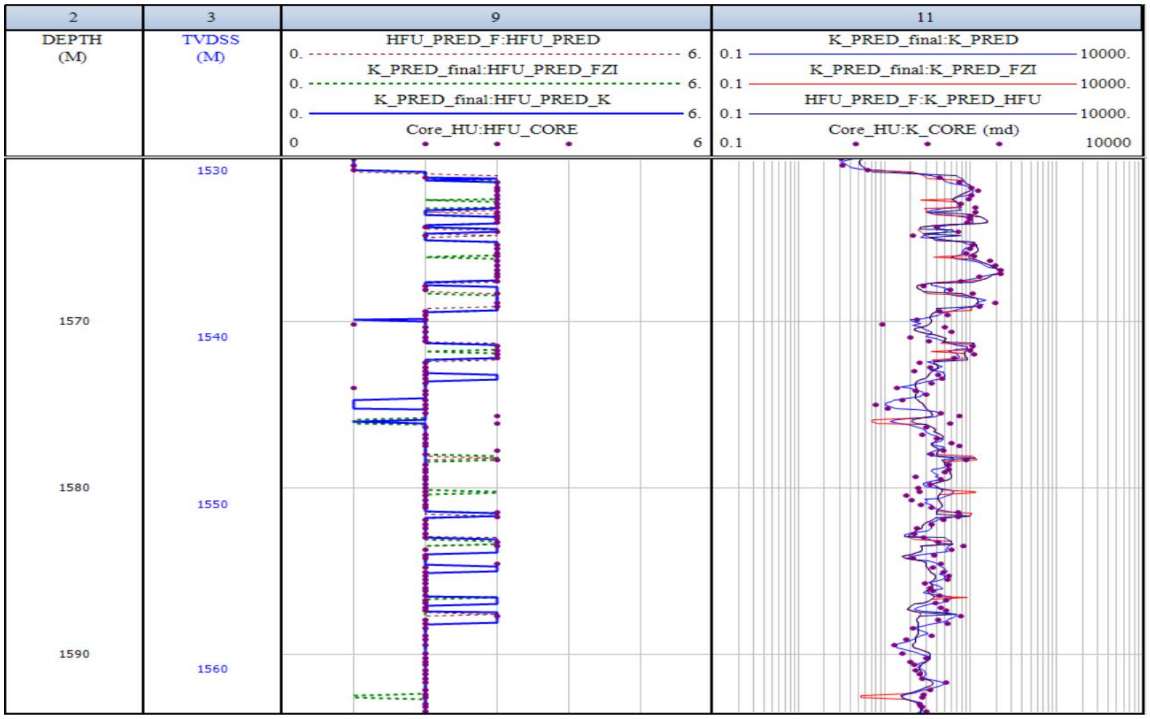


*Figure 14.* The predicted HFU and permeability in Well 3

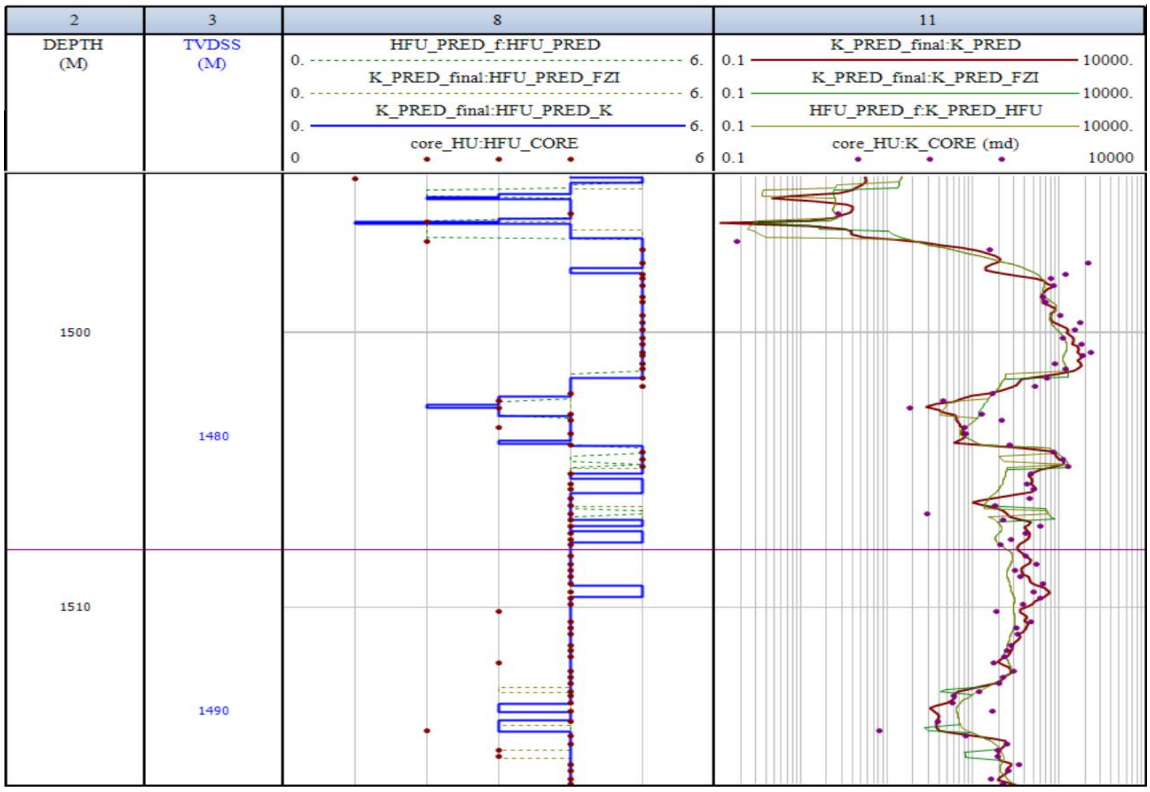


*Figure 15.* The predicted HFU and permeability in Well 4

## CONCLUSION

Applying machine learning in reservoir characterization, especially for heterogeneous carbonate reservoirs, is very meaningful as the results help improve hydrocarbon reserve estimation and reservoir performance prediction.

For the study area:

Unsupervised machine learning methods are helpful for HFU-clustering, and the K-means method gave the best results for the study dataset with **five** classified HFUs.

All workflows built for permeability prediction using machine learning methods for the study area gave good results with a high R2 compared with K_core.

The result of the study can be used to improve the field's 3D geological and 3D dynamic models.

**Contribution of authors:** Dung Nguyen Trung: outline, develop and finalize the manuscript, HFU classification; Man Ha Quang, Hoa Truong Khac, and Hong Nguyen Viet: permeability prediction; Huong Phan Thien and Hoang Cu Minh: review and conclusion. All authors have read and agreed to the published version of the manuscript.

**Abbreviations:** GR: gamma ray, API; R2: correlation coefficient; RMSE: root mean squared error; DTC: compressional sonic log, µs/ft; DTS: shear sonic log, µs/ft; HFU: hyraulic flow unit; FZI: flow zone indicator; FCM: fuzzy C mean; RQI: rock quality index; SOM: self-organised map; RD: deep resistivity lateral log, ohm.m; RS: shallow resistivity lateral log, ohm.m; MSFL: micro spherical focus log, ohm.m; RHOB: density log, g/cm$^3$; NPHI: neutron porosity, v/v; PHI: porosity, v/v; K: permeability, mD; K_FZI_Pred: estimated permeability from predicted FZI, mD; K_HFU_Pred: estimated permeability from predicted HFU, mD; K_Pred: predicted permeability using machine learning method, mD; SVM: support vector machine; KNN: K-nearest neighbor; K_Core: core permeability, mD; PHI_Core: core porosity, v/v; FZI_Core: core FZI; HFU_Core: core HFU.

*Acknowledgments:* The authors gratefully acknowledge the support given by Hanoi University of mining and geology (HUMG), Petrovietnam Exploration and Production Corporation (PVEP).

## REFERENCES

[1] Mazzullo, S. J., Rieke, H. H., and Chilingarian, G. V. (Eds.), 1996. Carbonate reservoir characterization: A geologic-engineering analysis, part II. *Elsevier*. pp. 201–205, 254–258.

[2] Lucia, F. J., Kerans, C., and Jennings Jr, J. W., 2003. Carbonate reservoir characterization. *Journal of Petroleum Technology*, *55*(06), 70–72. https://doi.org/10.2118/82071-JPT

[3] Lucia, F. J., 2007. Limestone Reservoirs. *Carbonate Reservoir Characterization: An Integrated Approach*, 181–215. doi: 10.1007/978-3-540-72742-2_6

[4] Ebanks Jr, W. J., 1987. Flow unit concept-integrated approach to reservoir description for engineering projects. *AAPG (Am. Assoc. Pet. Geol.) Bull.;(United States)*, *71*(CONF-870606-).

[5] Amaefule, J. O., Altunbay, M., Tiab, D., Kersey, D. G., and Keelan, D. K., 1993. Enhanced reservoir description: using core and log data to identify hydraulic (flow) units and predict permeability in uncored intervals/wells. In *SPE Annual Technical Conference and Exhibition?*, pp. SPE–26436. https://doi.org/10.2118/26436-MS

[6] Abbaszadeh, M., Fujii, H., and Fujimoto, F., 1996. Permeability prediction by hydraulic flow units—theory and applications. *SPE Formation Evaluation*, *11*(04), 263–271. https://doi.org/10.2118/30158-PA

[7] Mohaghegh, S., Balan, B., and Ameri, S., 1997. Permeability determination from well log data. *SPE formation evaluation*, *12*(03), 170–174. doi: 10.2118/30978-PA

[8] Babadagli, T., and Al-Salmi, S., 2002. Improvement of permeability prediction for carbonate reservoirs using well log data. In *SPE Asia Pacific Oil and Gas Conference and Exhibition*, pp. SPE–77889.

[9] Man, H. Q., Hien, D. H., Thong, K. D., Dung, B. V., Hoa, N. M., Hoa, T. K., Kieu, N. V., and Ngoc, P. Q., 2021. Hydraulic flow unit classification and prediction using machine learning techniques: A case study from the Nam Con Son basin, offshore Vietnam. *Energies*, *14*(22), 7714. https://doi.org/10.3390/en14227714

[10] Cuong, T. X., Anh, V. T., and Mai, L. C., 2015. Tectonic development of the Tri Ton horst and adjacent areas, offshore South Vietnam. In *AAPG Datapages/Search and Discovery Article #90236 © 2015 Asia Pacific Region, Geoscience Technology Workshop, Tectonic Evolution and Sedimentation of South China Sea Region, Kota Kinabalu, Sabah, Malaysia, May 26–27, 2015*.

[11] Lary, D. J., Alavi, A. H., Gandomi, A. H., and Walker, A. L., 2016. Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, *7*(1), 3–10. https://doi.org/10.1016/j.gsf.2015.07.003

[12] Ciaburro, G., 2017. MATLAB for machine learning. *Packt Publishing Ltd*. pp. 1–359.

[13] Rashid, M., Luo, M., Ashraf, U., Hussain, W., Ali, N., Rahman, N., Hussain, S., Martyushev, D. A., Thanh, H. V., and Anees, A., 2022. Reservoir quality prediction of gas-bearing carbonate sediments in the Qadirpur field: Insights from advanced machine learning approaches of SOM and cluster analysis. *Minerals*, *13*(1), 29. https://doi.org/10.3390/min13010029

[14] Ha, M. Q., Nguyen, H. M., Bui, D. V., Nguyen, H. V., Truong, H. K., and Pham, N. Q., 2023. Improving carbonate reservoir characterization by applying rock typing methods: a case study from the Nam Con Son Basin, offshore Vietnam. *Journal of Mining and Earth Sciences 64*(1), 38–49. doi: 10.46326/JMES.2023.64(1).05

[15] Strohmenger, C. J., Meyer, L., Walley, D. S., Yusoff, M. M., Lyons, D. Y., Sutton, J., Rivers, J. M., von Schnurbein, B., and Phong, N. X., 2018. Reservoir characterisation of the Middle Miocene Ca Voi Xanh isolated carbonate platform. *Petrovietnam Journal*, *6*, 10–24. https://doi.org/10.25073/petrovietnam%20journal.v6i0.186