Journal of Petroleum Exploration and Production Technology
https://doi.org/10.1007/s13202-024-01800-z

ORIGINAL PAPER - EXPLORATION GEOPHYSICS

Multiple linear regression and gene expression programming to predict fracture density from conventional well logs of basement metamorphic rocks

Muhammad Luqman Hasan[1,2] · Tivadar M. Tóth[1]

Received: 28 May 2023 / Accepted: 26 March 2024

Abstract
Fracture identification and evaluation requires data from various resources, such as image logs, core samples, seismic data, and conventional well logs for a meaningful interpretation. However, several wells have some missing data; for instance, expensive cost run for image logs, cost concern for core samples, and occasionally unsuccessful core retrieving process. Thus, a majority of the current research is focused on predicting fracture based on conventional well log data. Interpreting fractures information is very important especially to develop reservoir model and to plan for drilling and field development. This study employed statistical methods such as multiple linear regression (MLR), principal component analysis (PCA), and gene expression programming (GEP) to predict fracture density from conventional well log data. This study explored three wells from a basement metamorphic rock with ten conventional logs of gamma rays, thorium, potassium, uranium, deep resistivity, flushed zone resistivity, bulk density, neutron porosity, sonic porosity, and photoelectric effect. Four different methods were used to predict the fracture density, and the results show that predicting fracture density is possible using MLR, PCA, and GEP. However, GEP predicted the best fracture density with $R^2 > 0.86$ for all investigated wells, although it had limited use in predicting fracture density. All methods used highlighted that flushed zone resistivity and uranium content are the two most significant well log parameters to predict fracture density. GEP was efficient for use in metamorphic rocks as it works well for conventional well log data as the data is nonlinear, and GEP uses nonlinear algorithms.

Keywords Fractured reservoir · Fracture density · Multiple linear regression · Well-logging · Gene expression programming · Metamorphic basement

Abbreviations

AG	Amphibolite and amphibole biotite gneiss
ETs	Expression trees
GA	Genetic algorithm
GEP	Gene expression programming
GP	Gene programming
MLR	Multiple linear regression
OG	Orthogneiss
PC	Principal component
PCA	Principal component analysis
SG	Sillimanite and garnet-bearing biotite gneiss

List of symbols

D	Bulk density, g/cm^3
FD	Fracture density, m^{-1}
GR	Gamma ray, API
K	Potassium, ppm
N	Neutron porosity, v/v
P10	Number of fractures per meter, m^{-1}
PE	Photoelectric effect, b/e
RD	Deep resistivity, Ω·m
RXO	Flushed zone resistivity, Ω·m
S	Sonic porosity, us/f
TH	Thorium, ppm
U	Uranium, ppm

✉ Muhammad Luqman Hasan
luqman@geo.u-szeged.hu

Tivadar M. Tóth
mtoth@geo.u-szeged.hu

[1] Department of Mineralogy, Geochemistry and Petrology, University of Szeged, Egyetem Utca 2, Szeged 6722, Hungary

[2] Department of Petroleum Engineering, Universiti Teknologi PETRONAS, 32610 Bandar Seri Iskandar, Perak Darul Ridzuan, Malaysia

Published online: 20 April 2024

Introduction

Fracture interpretation and evaluation are very important in analyzing reservoir characteristics. Fractures improve the fluid flow inside the rocks (Li et al. 2021; Rajabi et al. 2021; Gao et al. 2023), and in some cases, can serve as hydrocarbon sinks (Vass et al. 2018; Gamal et al. 2022). Fracture detection requires data from core samples and image logs for accurate results (Shalaby and Islam 2017; Delavar 2022). However, imaging logging could be costly; hence, this is not a feasible option (Yang et al. 2017; Hussein 2022), and many wells, especially old wells, have no image logs (Delavar 2022). Coring can also be problematic to a certain extent as core retrieval processes are not always successful (Zazoun 2013), especially when coring in a highly consolidated formation (Abdideh 2016). The core analysis process can be costly and time consuming as well (Yang et al. 2017; Hussein 2022).

Conventional well logs can be used as an indirect method to evaluate fracture (Delavar 2022). Some of the previous studies outline the guidelines to indicate fracture indirectly from the conventional well logs (Serra 1986; Verga et al. 2000; Martinez et al. 2002; Ellis and Singer 2007). In recent years, several studies have been performed to explore the potential of conventional well log data to evaluate internal fractures when image logs and core samples are missing, especially for old wells. To date, many of these improvements are reliable enough to achieve this objective, and most of them leverage statistical methods and machine-learning applications (Tokhmechi et al. 2009a, 2009b; Tokhmchi et al. 2010; Ja'fari et al. 2012; Aghli et al. 2017, 2016, 2020; Pei and Zhang 2022; Qiu et al. 2022).

Several studies have been performed on fractures in carbonate rocks (Shalaby and Islam 2017; Gamal et al. 2022; Hussein 2022). Owing to the advantage and widely available conventional well log data, Gamal et al. (2022) proposed an integrated workflow to characterize and quantitatively analyze the fracture of three carbonate rock wells. Although a single definite well log cannot precisely confirm the presence of fracture, an integration of all available well logs can be an advantage for indirect measurement. However, the study also concluded that without spectral gamma ray information, a gamma ray log is not a conclusive tool for fracture detection. Hussein (2022) analyzed the conventional well log data (caliper, gamma ray, neutron, density, and sonic log) of two carbonate wells. The study utilized the gamma ray log by calculating the shale volume and compared the calculated values to the shale volume obtained by neutron-density logs; the positive difference indicated the fracture zones.

A few studies reported the use of only conventional well logs to evaluate fractures in reservoirs; and most of the papers agreed that porosity logs (density, neutron and sonic) also play an important role in analyzing the fracture characteristics (Lyu et al. 2016), with each log having its own function in the fractured zones. Further, the secondary porosity index can be determined using these porosity logs. Secondary porosity is one of the fracture indicators, although other reasons could also contribute to secondary porosity (Shalaby and Islam 2017). Aghli et al. (2020) also studied the potential of using sonic and resistivity conventional logs to determine fracture parameters. Lyu et al. (2016) reported that for fracture intensities > 1 m^{-1}, all three porosity logs together with the caliper and resistivity logs display fracture responses to a certain extent.

The studies show that conventional well logs can be used to predict fracture by analyzing the properties of each data. Further, some other studies reported the integration of these conventional well logs with other methods. Luo and Tang (2013) applied Monte Carlo simulation to four well log parameters for the identification of fractures and reduction of uncertainty problems. Abdideh (2016) employed multiple regression analysis by analysing four log parameters which are caliper, sonic, density, and photoelectric factor to predict fracture density (FD). A recent study by Tóth et al. (2023) employed multiple regression analysis to explore the relationship between FD and geophysical log data. The study explored the potential of claystone as a potential nuclear waste repository and concluded that multiple regression analysis is a good tool to be used to predict FD for older wells without image logs. Two of the most influencing geophysical parameters are resistivity and density logs. Aghli et al. (2016) used a differentiation method to analyze the fracture responses by conventional logs; however, adequate understanding of the study area structure and stratigraphy is important when applying this method. In addition, wavelet transform is one of the methods employed to improve fracture detection. Yang et al. (2017) studied the relevance of using wavelet transform to metamorphic rocks well log data and is one the few studies that report the use of wavelet transform for fractured crystalline rocks, specifically metamorphic rocks. The majority of the published literature has focused on the fracture in sedimentary rocks. In this study by Yang et al. (2017), wavelet transform was applied to an integrated curve called as the fractured integrated index. The study used density, caliper, deep resistivity (RD), and acoustic logs as the main well log data for fracture detection; the application of wavelet transform was successful for this type of reservoir as the method can detect the same fracture numbers as the image logs.

Machine learning and deep learning have evolved rapidly each day and provide many solutions to different problems. It is indeed of the powerful prediction methods to help solve reservoir and petroleum problems especially using the conventional well logs (Zhang et al. 2023a, b). In terms of fracture studies, one of the widely studied fracture parameters is FD, which many previous studies have attempted to predict using different machine-learning algorithms. Li et al. (2021) proposed a novel methodology to estimate FD and orientation from azimuthal elastic impedance difference using singular value decomposition. Li et al. (2018) predicted FD using acoustic logging signals as inputs and implemented a machine-learning technique, which is a genetic algorithm-support vector machine. It was able to classify fractures into low, medium, and high fracture densities. The model was proven accurate using the image logs. Pei and Zhang (2022) used a multi-layer perceptron (MLP) machine-learning algorithm to predict fracture parameters including FD; the algorithm that had been tested for carbonate reservoirs showed 82% accuracy in prediction. Rajabi et al. (2021) presented four hybrid models to predict FD using 12 input parameters; MLP with a combination of particle swarm optimizer (PSO) was suitable in predicting FD of a carbonate oil field. The same methodology was adopted by Gao et al. (2023) by using the two similar machine-learning-based predictions (MLP-PSO and MLP-Genetic Algorithm (GA)) as Rajabi et al. (2021). The two new algorithms were also used—the least squares support vector machine combined with PSO and with GA. Although Rajabi et al. (2021) reported that MLP-PSO was a better methodology for predicting the FD based on their dataset, Gao et al. (2023) concluded that least squares support vector machine-PSO improved the overall prediction as it worked for unstructured and non-dependent data, had lower adjustment parameters, and fast convergence speed. Delavar (2022) agreed with Gao et al. (2023) since the support vector machine in combination with other methods, i.e., radial basis function and gray wolf optimizer, was superior in accuracy for overall fractures detection. The focus of this study was not on FD.

Realizing the simplicity in the use of GA, Ferreira (2001) combined GA with gene programming and developed a new method known as gene expression programming (GEP). This newly developed method has been widely applied in several fields to solve different issues (Algaifi et al. 2021; Chu et al. 2021; Afrasiabian and Eftekhari 2022; Hassan et al. 2022; Ari and Alagoz 2023). Since the inception of this GEP method, it has been studied rigorously in the literature. Undeniably, the applications of GEP are mostly concentrated in civil engineering studies. For example, rock strength tests such as uniaxial compression strength test study was done by Jahed Armaghani et al. (2018) and İnce et al. (2019), Jalal and Iqbal (2023) compared the unconfined compression strength prediction between GEP and multigene expression programming, and Zhang and Zhang (2024) predicted coefficient of permeability of soils using GEP and compared the performance with some other predictive models. Nonetheless, there are a few studies of GEP in the geoscience and petroleum engineering field, however, it is still in its infancy stage. A few studies that utilized GEP is the study by Zhang et al. (2023a) which explored the potential of using GEP to the hydraulic fracturing effects specifically fracture complexity index after fracturing process. The applied GEP method was able to be used for this application. Esmaeilpour et al. (2024) aimed to develop a more precise model to calculate equations of state that deal with two-phase geofluids properties. The current simulation models for this calculation tend to give unimportant parameter in their solving model. The implementation of GEP solved this issue and Esmaeilpour et al. (2024) proposed GenEOS for accurate and efficient computation of two-phase fluid properties mixtures. Other than that, various topics were also studied such as Shahabi-Ghahfarokhy et al. (2022) predicted the density of pure hydrocarbons and their mixtures, Rostami et al. (2017) predicted the CO_2 solubility in crude oil, Lv et al. (2023) also studied on the fluid solubility but together with group method of data handling and GEP, the solubility of CO_2-N_2 gas mixtures in aqueous solutions was predicted. In short, GEP has been used widely in many applications and field, however, to authors' best knowledge, GEP has not been studied to predict FD based on conventional well-log parameters as the inputs and especially predicting FD in a metamorphic reservoir.

The aim of this study was to predict the FD in a fractured metamorphic hydrocarbon reservoir in the event that image logs or core samples are not available. The prediction was based on the 10 conventional well logs used for this study: gamma ray (GR), spectral gamma ray (thorium (TH), potassium (K), uranium (U)), RD, flushed zone resistivity (RXO), bulk density (D), neutron porosity (N), sonic porosity (S), and the photoelectric effect (PE). Multiple linear regression (MLR), principal component analysis (PCA), and GEP were used to determine the best method for predicting the FD.

Geological setting

The study area was the Mezősas field, located in the northern rim of the Békés Basin (Fig. 1), which is the largest and deepest sub-basin in the Pannonian Basin in Southeast Hungary (Tóth et al. 2000). This fractured buried-hill hydrocarbon reservoir has been active for decades.

The area has undergone several subsequent geological events that have resulted in a very complex mosaic of basement structures (Tari et al. 1992, 1999). The entire basement reservoir comprises Variscan metamorphic rocks evidenced by previous petrological studies (Tóth et al. 2000;

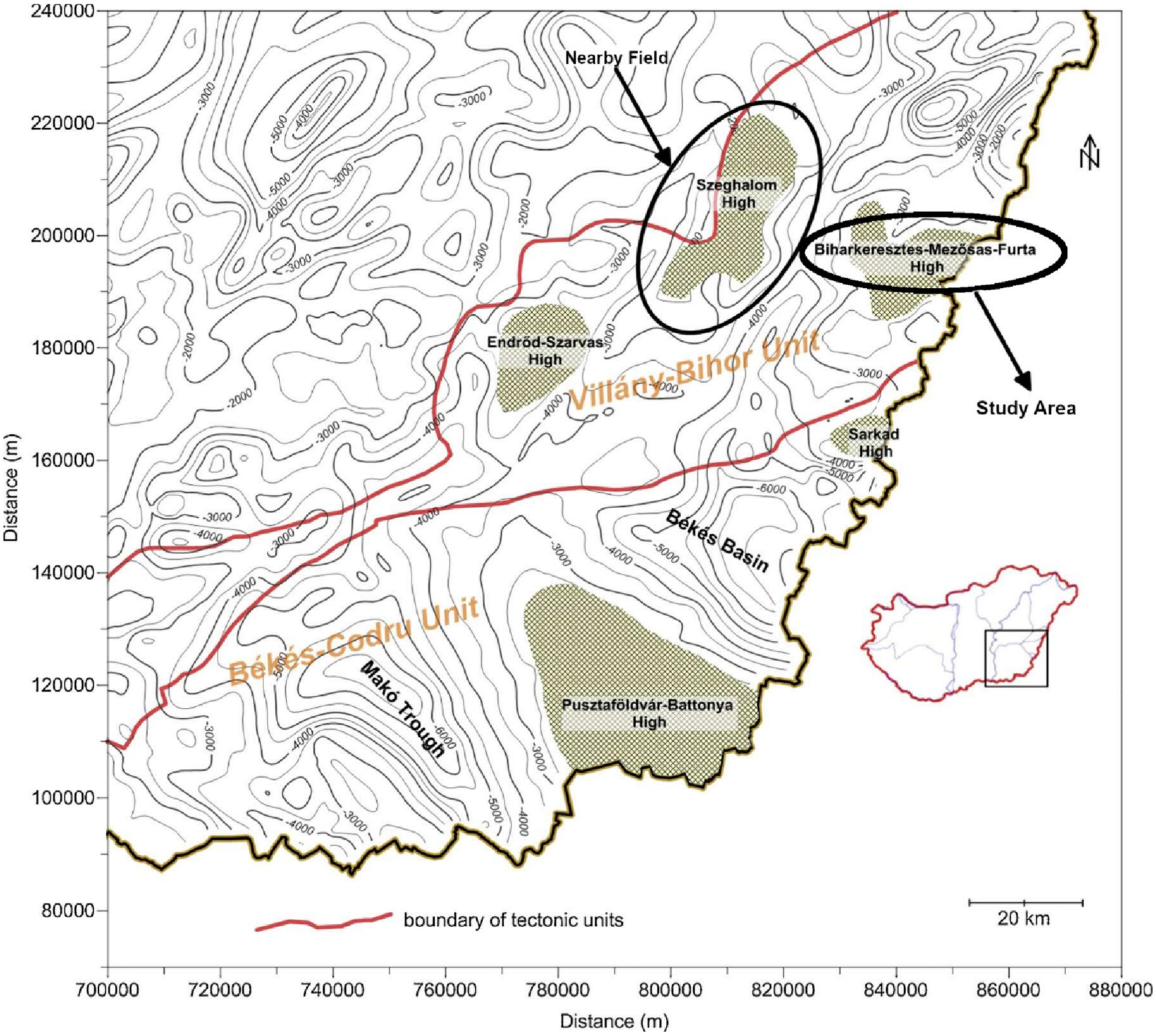

Fig. 1 The map showing the study area and the surrounding field (modified from Vass et al. 2018)

Tóth and Schubert 2018). The Eoalpine compressional tectonic evolution during the Cretaceous formed the complex nappe systems throughout the metamorphic basement of the Pannonian Basin. As a result, the basement at present is made of blocks of intermediate and high-grade metamorphic rocks with significantly different metamorphic evolutions (Tóth et al. 2021) separated by post-metamorphic structural elements (Molnár et al. 2015; Hasan et al. 2023).

Due to the formation of the Pannonian Basin during the Neogene, normal fault systems became active, and small pull-apart basins were formed, making the previous basement structure even more complicated (Albu and Pápa 1992; Tari et al. 1999). Due to these motions, a few domes or crystalline highs became exhumed in the Miocene, such as the Szeghalom Dome (SzD), which is located near Mezősas and had been extensively studied (Tóth et al. 2000; Juhász et al. 2002; Molnár et al. 2015; Vass et al. 2018). Based on the petrological and structural similarities, SzD can be a reference field for the study area (Tóth et al. 2000, 2021; Tóth and Zachar 2006). In this study, three major lithological units have been proposed based on detailed petrology and well log analysis. The lowermost rock body of the basement reservoir is dominated by orthogneiss (OG), which is covered by sillimanite and garnet-bearing biotite gneiss realm (SG), while at the topmost layer, amphibolite and amphibole biotite gneiss (AG) have been reported (Molnár et al. 2015). The OG is derived from the medium-grade metamorphism of an igneous intrusion (Tóth and Schubert 2018), while the SG has a sedimentary protolith (paragneiss). The same sequence of rock units was previously determined for the Mezősas field by petrological (Tóth and Zachar 2006) and well log (Hasan et al. 2023) analyses. Hasan et al. (2023) also proved a two-part internal structure for the topmost amphibolite unit as shown in Fig. 2.

Concerning the reservoir geological aspects, much evidence shows how the internal structure of the basement

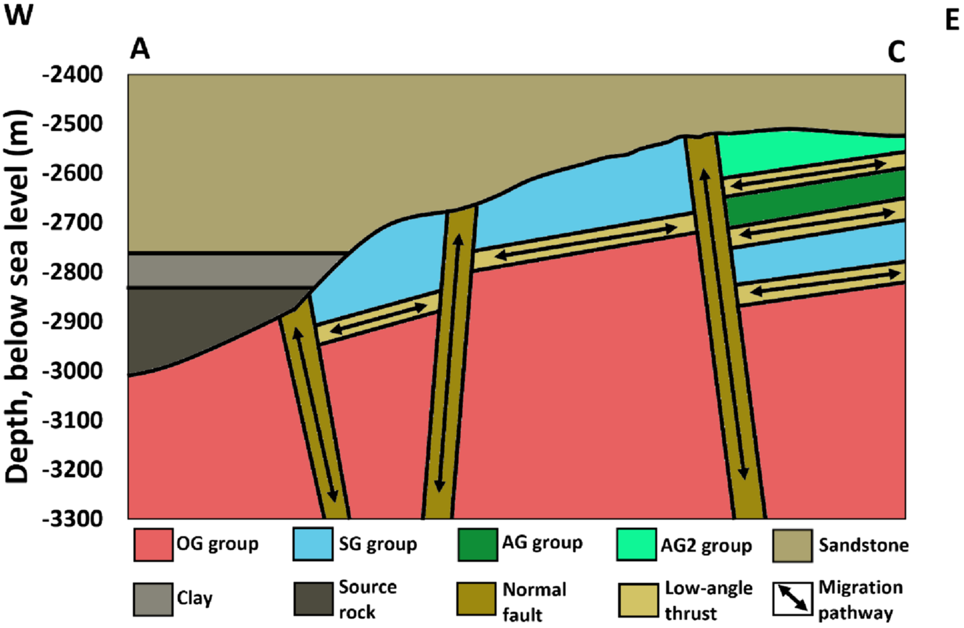

Fig. 2 The proposed model of OG-SG-AG sequence (Hasan et al. 2023). OG: orthogneiss; SG: sillimanite and garnet-bearing biotite gneiss; AG: amphibolite; AG2: amphibole-biotite gneiss

provides the migration pathways and storage capacity for hydrocarbons. The study by Molnár et al. (2015) proved that low-angle thrust faults that separate the lithological realms are responsible for hydrocarbon migration from adjacent over-pressured deep sub-basins. By studying hydrocarbon fluid inclusions and organic geochemical fingerprints, the eminent role of the thrust sheets was demonstrated during paleomigration activity through the basement block (Juhász et al. 2002; Schubert et al. 2007; Tóth et al. 2020). The integration of structural evolution and the fracture network geometry suggested that the topmost amphibolite bodies (AG block) have the highest storage capacity in the entire metamorphic basement reservoir due to the mutual interconnectivity of the microfracture system (Tóth et al. 2020).

Methodology

Data availability

Three wells were selected for this study—Wells 21, 22, and 25. From the well log analysis, the wells penetrated the basement of the reservoir and the different types of lithology. For this study, only the top basement section was considered for the analysis as it contained amphibolite. Therefore, a single lithology was considered for fracture evaluation for the purpose of standardization. This is mainly because the amphibolite lithology had more fractures as compared to SG and OG; and acted as a conduit for the flow of the fluid (Molnár et al. 2015). All three wells had ten measured log parameters—GR, spectral gamma ray (TH, K, and U), RD, RXO, D, N, S, and PE. All three wells had image logs for the FD determination. The descriptive statistics are shown in Table 1.

Flow chart

The borehole televiewer image log was analyzed, the interpretation of fractures was hand-picked, dip angle and fracture direction measurements were conducted. Fractures on image logs normally showing sinusoid features with more dip than structural dip as shown in Fig. 3. A few types of fractures were determined which were bedding, fault or microfault, fracture, induced fracture, and partial fracture. These interpretations were carefully marked on the image logs and FD results were extracted. This is the only parameter from borehole televiewer that is needed for this study. An example of image log including the interpretation of Well 21 is shown in Fig. 3. This figure also includes other conventional logs used in further analysis such as GR, resistivity, and porosity logs. After the image logs interpretation was done, at each depth, the number of fractures shown are calculated. The FD used in this study was known as P10, which was calculated by counting the number of fractures per meter along the wellbore. The FD data were standardized with the other 10 conventional well log parameters. Since the P10 FD data was used, the well logs were also standardized accordingly by calculating the well logs value per meter along the well depth. Data was then checked for missing values and outliers. All conditions including the normal distribution of the dependent variable

Table 1 Descriptive statistics of Wells 21, 22, and 25 including the combination *n* of all three wells

Well (Depth)		GR (API)	K (%)	Th (ppm)	U (ppm)	RD (ohm.m)	RXO (ohm.m)	D (g/cc)	N (v/v)	S (us/f)	PE
21	Min	25.06	1.33	0.78	0	9.36	1.58	2.52	4.24	50.85	3.97
(2584 m	Max	134.42	5.54	14.63	2.03	609.84	1126.55	3.02	23.05	93.05	12.16
–	Mean	61.14	2.58	4.11	0.39	66.51	59.61	2.77	13.12	61.6	6.74
2694 m)	SD	22.8	0.78	2.68	0.31	62.21	101.65	0.08	3.84	6.4	1.08
22	Min	20.51	0.98	0.85	0	16.25	12.97	2.49	3.16	47.28	2.96
(2618 m	Max	143.65	6.09	19.91	4.21	581.27	416.63	3.01	25.3	82.31	16.19
–	Mean	60.82	2.8	5.43	0.54	103.33	89.69	2.76	13.65	58.64	6.26
2738 m)	SD	32.04	0.97	4.35	0.59	94.74	84.45	0.09	4.47	6.65	1.97
25	Min	23.8	1.1	0.88	0.35	15.19	0.1	2.46	2.02	47.39	2.64
(2532 m	Max	130.47	4.81	16.94	8.21	327.26	5717.52	3.1	25.1	76.42	12.27
–	Mean	77.32	2.77	5.07	3.13	94.29	311.47	2.68	13.6	60.9	5.61
2654 m)	SD	18.57	0.68	2.31	1.85	62.15	2109.49	0.11	4.32	5.6	1.31
All wells (21, 22, 25)	Min	20.51	0.98	0.78	0.00	9.36	0.10	2.46	2.02	47.28	2.64
	Max	143.65	6.09	19.91	8.21	609.84	5717.52	3.10	25.30	93.05	16.19
	Mean	66.69	2.72	4.88	1.40	88.37	157.46	2.73	8.76	60.38	6.19
	SD	26.20	0.82	3.27	1.72	76.09	1253.45	0.10	7.16	6.34	1.57

Min minimum value, *Max* maximum value, *SD* standard deviation, *GR* Gamma ray, *K* potassium, *TH* thorium, *U* uranium, *RD* deep resistivity, *RXO* flushed zone resistivity, *D* bulk density, *N* neutron porosity, *S* sonic porosity, *PE* photoelectric effect

and correlations of independent variables for PCA and MLR were checked.

Since three wells were considered for this study, the first method was to use a single well as a training set and extrapolate the equation generated using MLR for the other two wells. This is a simple and straightforward first method. However, the results from this first method were not good. Hence, a second method was executed. For the second method, to improve the results of the first method, PCA was applied to the original dataset. The MLR was then applied to the new variables derived from the PCA. Since all three wells provided different PCA results, the entire dataset from all three wells was combined to generate PCA results, and regression analysis was applied to the new set of variables.

After the PCA results were generated by the statistical tool, which is the IBM SPSS Statistics 24, the data had to be divided into two which were the training and testing sets. The training data set was the one that will be used to generate MLR equation, and the equation generated will be applied to the testing set. The data can be divided into two ways which are using the random data division or separating the data manually into two. Since the data from all three wells have different range and standard deviation, the random selection method to divide the data into training and testing sets was not used in this case to avoid skewed results and biasness because some data from one well might be selected more than the other wells. Hence, one of the practical ways to divide the data was by separating the data into upper and lower sections of the well in which 70% of the data from the upper section of each well was selected as a training set for MLR (Jahed Armaghani et al. 2018; Afrasiabian and Eftekhari 2022). The equation generated was applied to the data for the lower section of the well.

Upon executing the second method, it was observed that PCA did not significantly improve the results. However, the method to separate the data into upper and lower sections of the wells was suitable to be used. Thus, for the third method, the original data set (without involving PCA results) was reused, but instead of using a single well, the training set was a combination of all upper sections of three wells combined. For the last method, the nonlinear regression method using GEP was implemented. This method also used a combination of data from all three wells. To simplify, this study attempted to determine the best possible way to incorporate MLR, PCA, and nonlinear techniques for predicting FD. The workflow is provided below and also summarized in Fig. 4. The workflow also shows the design of each method listed below.

1. Using one well as a training set method
2. Using PCA and regression methods
3. Using combination data of all wells as a training set method
4. Using GEP method

A preliminary study was conducted to understand the relationship between fracture density and all ten well log parameters that were used in this study. It is essential that this relationship to be established first to see if there is any strong correlation between one well log parameter to the

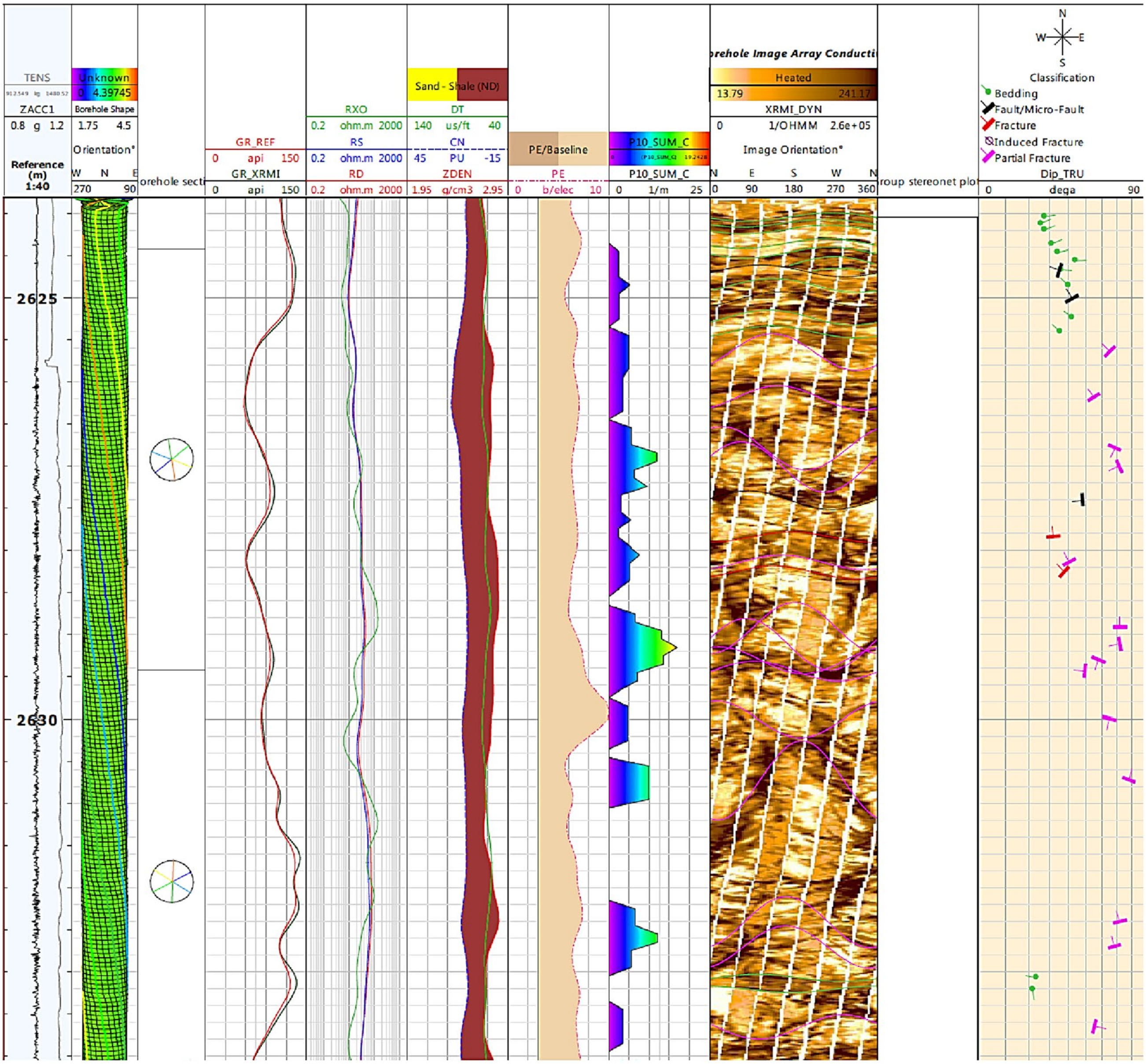

Fig. 3 The example of image log and conventional logs used in the analysis

fracture density. If more than half of the parameters show strong correlation with FD, then it would be best to select these parameters as input parameters and exclude the other remaining parameters for the next step. The results of these correlations are shown in Fig. 5. Only one well, which is Well 21 results are shown in this figure for simplicity. It is well noted that all three wells exhibit about similar results. From this figure, it shows that none of the well log parameters is highly correlated with FD. The highest correlation based on R^2 is shown by neutron porosity with 0.76. This study on correlation between input and output parameters is important so that significant parameters can be selected for the next step. However, based on the study on this correlation, it shows that all parameters are selected for the next step since only two parameters show R^2 value above 0.5 and the other parameters show R^2 value below 0.2. Hence, eliminating most of the parameters would not be a good idea. Therefore, for the next step, all parameters will be included.

PCA

PCA has applications in several fields related to petroleum engineering and geosciences studies (Tóth 2012; Konaté et al. 2017; Geng et al. 2021; Ren et al. 2023). PCA is a dimensionality reduction method that transforms the original

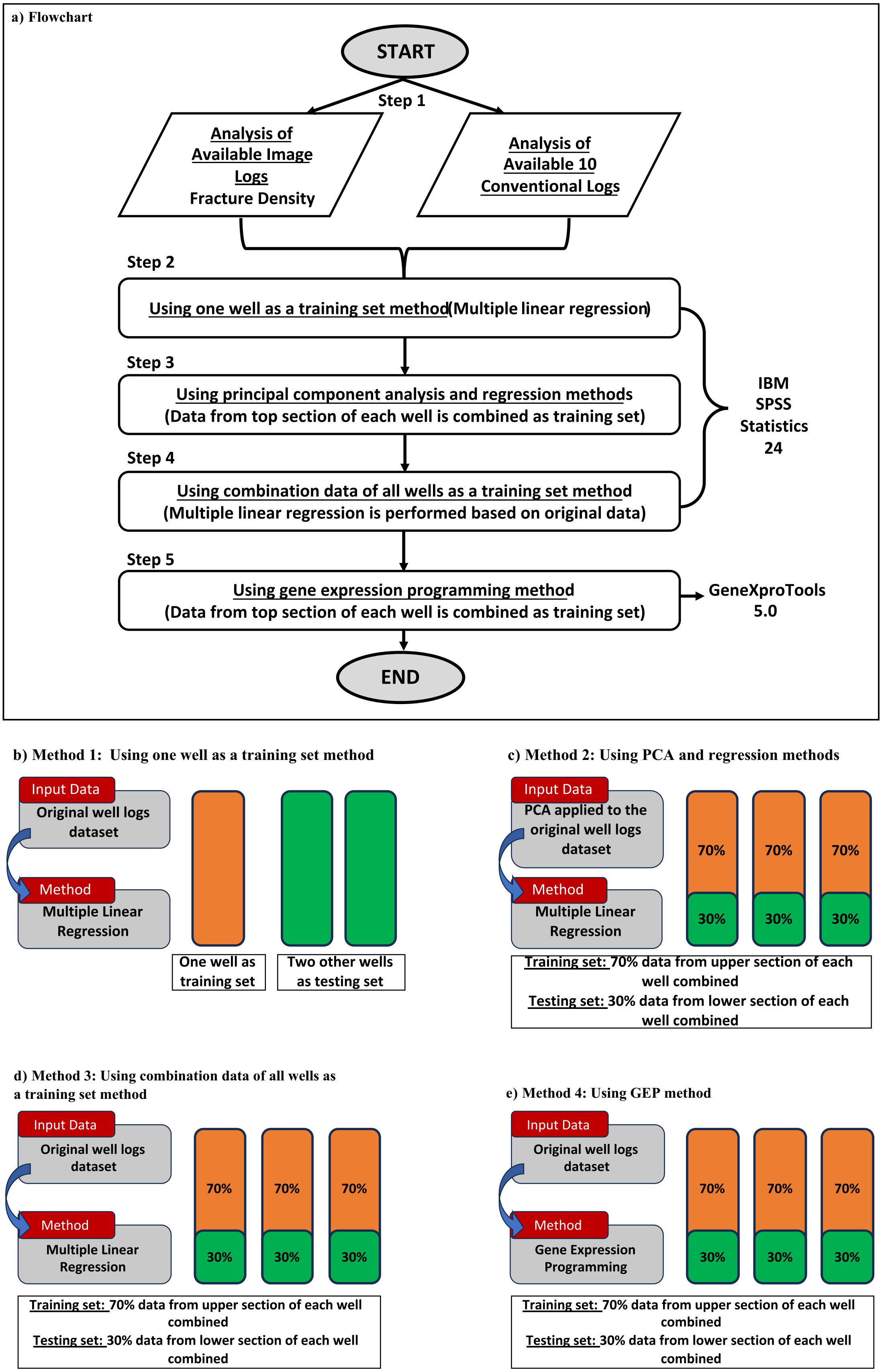
a) Flowchart
START
Step 1
Analysis of Available Image Logs
Fracture Density
Analysis of Available 10 Conventional Logs
Step 2
Using one well as a training set method(Multiple linear regression)
Step 3
Using principal component analysis and regression methods
(Data from top section of each well is combined as training set)
Step 4
Using combination data of all wells as a training set method
(Multiple linear regression is performed based on original data)
IBM SPSS Statistics 24
Step 5
Using gene expression programming method
(Data from top section of each well is combined as training set)
GeneXproTools 5.0
END
b) Method 1: Using one well as a training set method
Input Data
Original well logs dataset
Method
Multiple Linear Regression
One well as training set
Two other wells as testing set
c) Method 2: Using PCA and regression methods
Input Data
PCA applied to the original well logs dataset
Method
Multiple Linear Regression
70%
70%
70%
30%
30%
30%
Training set: 70% data from upper section of each well combined
Testing set: 30% data from lower section of each well combined
d) Method 3: Using combination data of all wells as a training set method
Input Data
Original well logs dataset
Method
Multiple Linear Regression
70%
70%
70%
30%
30%
30%
Training set: 70% data from upper section of each well combined
Testing set: 30% data from lower section of each well combined
e) Method 4: Using GEP method
Input Data
Original well logs dataset
Method
Gene Expression Programming
70%
70%
70%
30%
30%
30%
Training set: 70% data from upper section of each well combined
Testing set: 30% data from lower section of each well combined

Fig. 4 The flowchart used for this study. **a** The overall flowchart, **b** Method 1 workflow, **c** Method 2 workflow, **d** Method 3 workflow, **e** Method 4 workflow

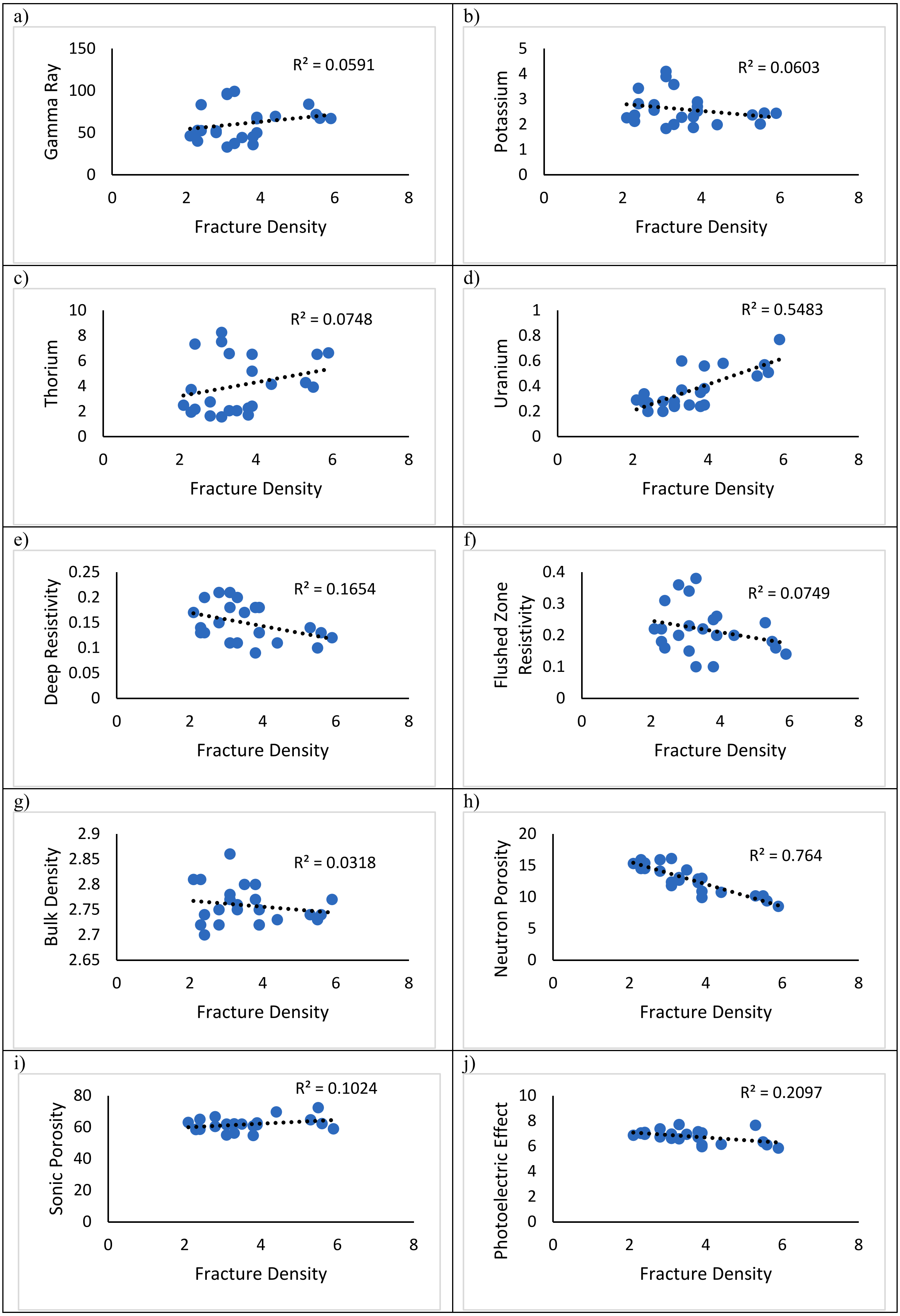

Fig. 5 The correlation between each well log parameter with fracture density. Only uranium content and neutron porosity correlated quite well with fracture density

variables into a new dataset or uncorrelated variables (Tiryaki 2008). The new variables known as principal components (PCs) are transformed using orthogonal transformation and are the linear combination of all the original variables (Li et al. 2018). The original variables might or might not be related to one another. The advantage of PCA is that it reduces a large number of variables into a smaller number while retaining information (Habibi et al. 2014). A deep understanding of the data is required to interpret the results so that the interpreted variables are meaningful and useful for further analysis. Since PCs are not intercorrelated, a PC can be described in length without having to refer to the other PC (Konaté et al. 2015). The determination of the number of PCs to be included in the analysis is normally determined by the eigenvalue. The number of PCs is determined by the Kaiser criterion, wherein the PCs are selected if the eigenvalue > 1; this is also known as the eigenvalue one criterion (Kaiser 1960). In this study, the PCA was performed on the original 10 log parameters. The three wells used in this study yielded three different PCA results; thus, data from all three wells were combined for further PCA.

MLR

MLR has been widely used in geological and petroleum engineering studies (Habibi et al. 2014; Alizadeh et al. 2022; Khosravi et al. 2022; Cai et al. 2023; Yuan et al. 2023; Tóth et al. 2023). It is a statistical technique used to produce a model based on one dependent variable and several independent variables (Cai et al. 2023). In the linear regression model, the assumption is that the response variable (or equation produced) is a linear function of the model parameters, and the residuals are normally distributed (Enayatollahi et al. 2014). The general expression of MLR is as follows:

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon \tag{1}$$

where y is the dependent variable, β_1, β_2, …, β_p are the regression coefficients, x_1, x_2, …, x_p are the independent variables, and ϵ is the regression constant.

In this study, backward selection was employed for MLR as several well log parameters were used in this study, and it was better to explore and test all parameters and select the best parameters. The software used for the study was IBM SPSS 24. This method considers all available variables or parameters in the beginning and removes the least significant variables until there is no parameter left or when the stopping condition is met (Mantel 1970; Heinze et al. 2018; Dunkler et al. 2014).

The variables are removed normally as their presence contributes to lowered R^2 value and the higher p-value of the model, or when variables are eliminated, it can cause a reduced residual sum of squares (Harrell 2001). The stopping condition is normally determined by the p-value, and the model or regression analysis stops when all remaining variables have a p-value smaller than the pre-set value of 0.05 (Heinze et al. 2018; Afrasiabian and Eftekhari 2022).

GEP

Ferreira (2001) introduced a GEP that combined the advantages of GA and gene programming (GP) and eliminated its limitations. GA was developed by Holland John (1975) based on the theory of evolution. It is well-known as the optimization algorithm that mimics natural selection processes that involve population modification procedures (Gao et al. 2023). GAs have been widely used in fracture detection studies (Rajabi et al. 2021; Gao et al. 2023) as it helps in feature selection and can be integrated with other methods for hybrid algorithm development. GA elements are linear strings of fixed length (chromosomes), while GP elements are nonlinear entities of different sizes and shapes (Ferreira 2001; Sharifi and Moghbeli 2020). GP, also traditionally known as tree-based GP, is the implementation of GA that operates as a powerful regression procedure for nonlinear, nondifferentiable, discrete, and continuous problems (Sharifi and Moghbeli 2020).

GEP combines the GA and GP; the chromosomes in GEP are encoded as linear strings of fixed length and are expressed as nonlinear entities of different sizes and shapes. GEP uses populations of individuals and introduces genetic variation using one or more genetic operators (Aydogan et al. 2023). It utilizes chromosomes that can express and create strings in the form of functions and mathematical relations (Afrasiabian and Eftekhari 2022). Chromosomes in GEP can contain one or more genes. GEP offers a complete genotype/phenotype system. The phenotype in GEP is the expression trees, and genes in GEP are expressed as genotypes. The main determinants in GEP are chromosomes and expression trees (Aydogan et al. 2023) as can be observed in Fig. 6. It shows a simple chromosome structure with four heads (function) and four tails (terminal). The function codes in chromosomes include mathematical expressions, such as plus, minus, times, multiply, and square root. An example from Fig. 6b shows that Q is actually a square root function. The terminal codes are variables (examples: x, y, z) and constants (examples: 1.2 and 3.11).

The GEP process is initiated by a random initial population generation of a specific size. Each of these initial populations or chromosomes is then assessed against fitness function over various groups of fitness situations. The chromosomes are then selected according to their fitness value in which the most fitted chromosomes have more chances

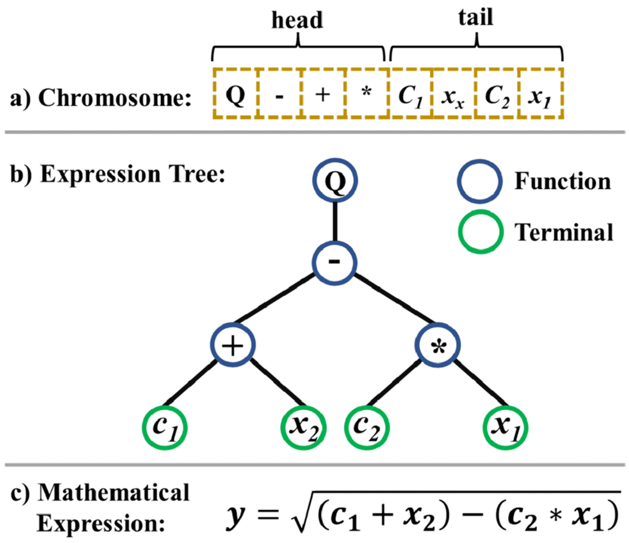

Fig. 6 Example of a classic chromosome structure **a** Chromosome with a gene structure having a head and a tail. **b** Expression tree structure with function and terminal. **c** Mathematical expression example based on this expression tree (modified from Ferreira 2001; Ari and Alagoz 2023)

to be selected to proceed to the next generation. After the selections, these are reproduced with some modifications performed by different genetic operators, such as mutation, insertion sequence transposition, inversion, gene recombination, and gene transposition. GEP uses the simplest criteria and further permits the development of complex and nonlinear programs due to multigenic behavior because of the genetic process at the chromosome level (Javed et al. 2020). The flowchart of GEP is shown in Fig. 7. In this study, GeneXproTools 5.0 was used to run the GEP for predicting the FD.

GEP is a robust method that can handle different types of data. Obviously, the input parameters should be consistent, any missing data is handled correctly, extreme outlier values should also be removed or treated appropriately. The significance of input parameters should also be checked if the input parameters have any meaningful interpretation to the target output. In this study, all 10 well log parameters are studied and determined to be significant to predict FD. The *GR* and spectral *GR* such as *K*, TH and *U* are very useful to determine the lithology of the rock. However, the use of these GR logs especially spectral GR can be useful give an indication of fracture existence specially uranium because uranium is soluble to both water and hydrocarbons that filled up the formation fractures. The resistivity logs are proven to be useful to indicate fractures indirectly as well because the RXO can read the mud filtrate invasion of the flushed zone if there are large fracture openings and these RXO values can be compared with RD to ensure that there is mud filtrate filling up the fractures. In terms of porosity logs, if there is mud invasion into the fractures, the overlay of the D-N logs can indicate this scenario by showing the sharp drop in D and sharp increase in N values. S porosity log will show a cycle skipping if there is a fracture existence. Lastly, if barite-loaded muds enter the fractures, this scenario will be captured on PE log by showing high reading of values. These are all reported in Aghli et al. (2016), Aghli et al. (2017), Shalaby and Islam (2017), Aghli et al. (2020), Hussein (2022), Gamal et al. (2022). Therefore, all these

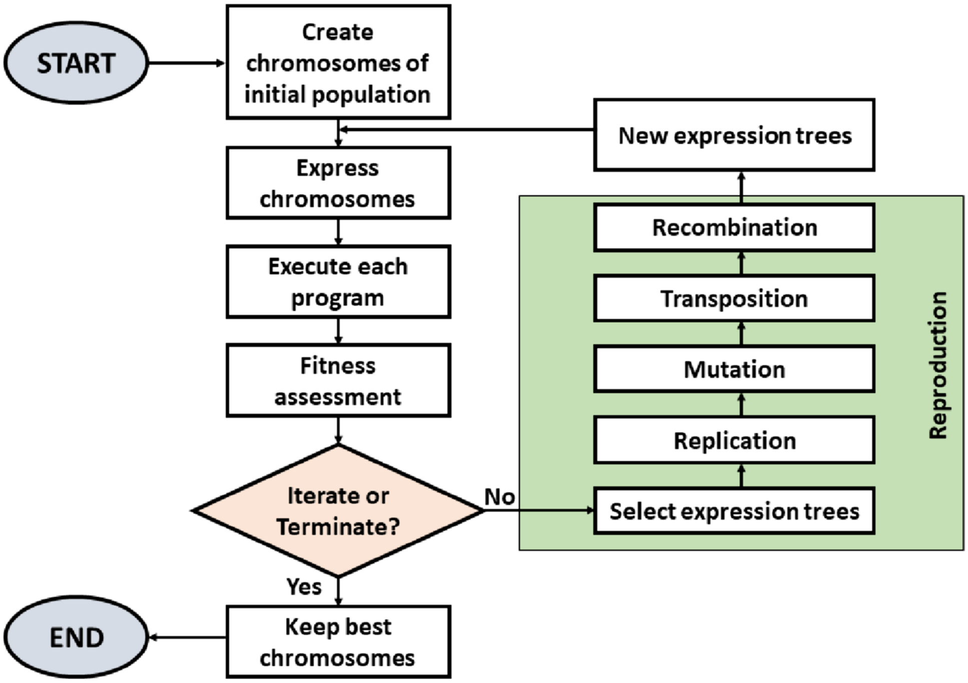

Fig. 7 The common workflow of gene expression programming (modified from Ferreira 2001; Afrasiabian and Eftekhari 2022; Ari and Alagoz 2023)

logs used in this study are important and significant to fracture determination.

Results

Using one well as a training set method

In this study, three wells with image logs were analyzed and used as the training set for regression analysis. Only the amphibolite section of the Mezősas field was analyzed in this study to maintain the consistency of the data as described in the previous section.

The purpose of designing the workflow was to select the best well based on the data and regression analysis and use that well as the training well for the other wells with the same lithology. Therefore, the results of the first section show the selection of the best well. Basically, all three wells were processed with regression analysis, and the generated equation was applied to the other two wells.

As can be observed from Fig. 8, the measured FD was obtained from the image log. Equation Well 21 is the regression analysis result for Well 21. The equation generated is as follows:

$$\mathrm{FD}(21) = 25.481 + 3.966 \times U - 4.005 \times RXO - 8.145 \times D \tag{2}$$

where FD is fracture density, 21 refers to Well 21 data, *U* refers to uranium content log value, *RXO* is the flushed zone resistivity, and *D* is the bulk density log value. This equation was then applied to Wells 22 and 25. The equations generated for Wells 22 and 25 were applied to Well 21. The equation from Well 22 is as follows:

$$\mathrm{FD}(22) = 2.871 - 1.245 \times K + 0.522 \times \mathrm{TH} - 3.022 \times U - 29.728 \times RXO + 2.258 \times D \tag{3}$$

where *K* is the potassium log value, TH refers to thorium content, *U* is the uranium content, *RXO* refers to the flushed zone resistivity, and *D* is the bulk density log value.

The equation from Well 25 is as follows:

$$\mathrm{FD}(25) = 3.147 - 0.257 \times \mathrm{TH} + 27.137 \times RD - 21.55 \times RXO \tag{4}$$

where TH refers to the thorium content, *RD* is the deep resistivity value, and *RXO* is the flushed zone resistivity.

From Fig. 8a, equations from Wells 22 and 25 did not satisfy the data from Well 21 as it shows that both equations did not correspond to the measured FD of Well 21. This observation was supported by the R^2 value. The R^2 value of measured versus predicted FD value from different wells is shown in Table 2.

From the table, the R^2 value was 0.721 for the measured versus predicted FD value of Well 21 using the equation from Well 21, which is acceptable but not high. However, when predictions of Well 21 FD were made using equations from Wells 22 and 25, the R^2 values were low, as seen in Table 2. The R^2 values of Well 21 for predicted FD were 0.052 and 0.026 when using equations from Well 22 and 25, respectively. Considering the results for Well 22 as observed in Fig. 8b, the equation from Well 21 was not suitable, especially for the top section of the well as

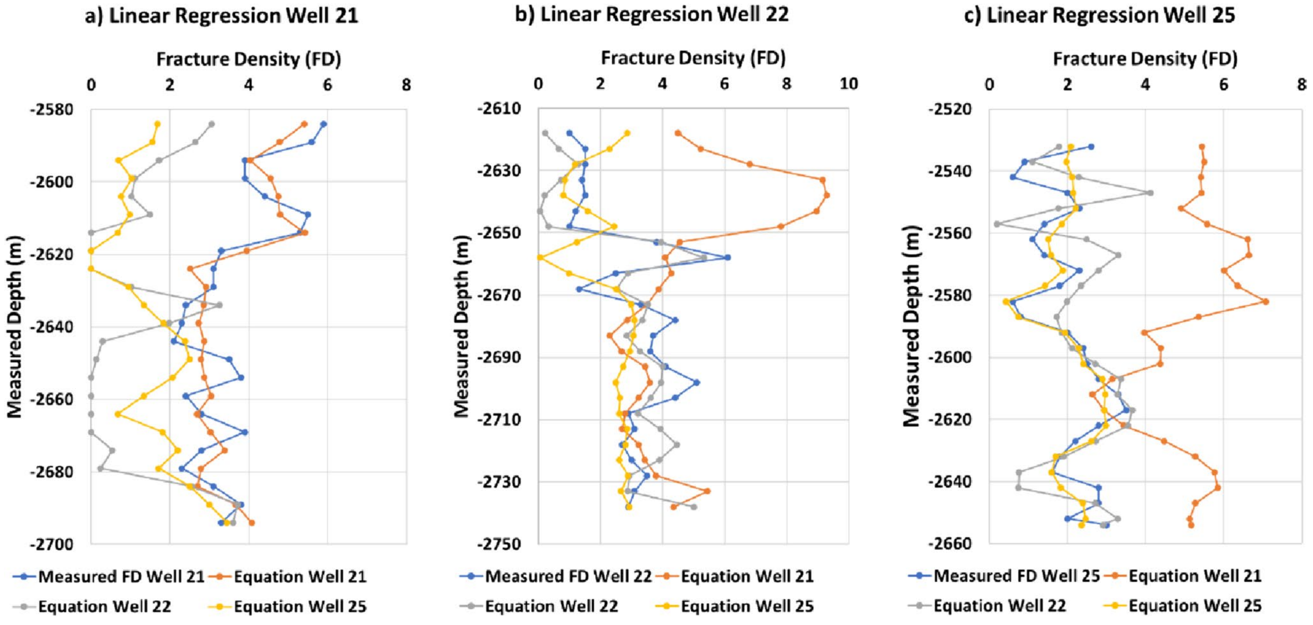

Fig. 8 The results for the first method, which uses one well as the training set and applies the derived equation to the other two wells

Table 2 The R^2 values based on results in Fig. 8

R^2 value	Equation Well 21	Equation well 22	Equation Well 25
Measured FD Well 21	0.721	0.053	0.026
Measured FD Well 22	0.039	0.666	0.006
Measured FD Well 25	0.461	0.172	0.598

The R^2 values were generated from the plot of measured FD versus predicted FD from the multiple linear regression equations

FD fracture density

it predicted a considerably high FD and a low R^2 value of 0.039, as shown in Table 2. From Fig. 8b, the equation from Well 25 provided an optimum prediction for Well 22, and the predicted value followed the same trend as the data from Well 22. However, the R^2 value is 0.006, which is very low, as shown in Table 2.

The equation from Well 21 was also not useful in predicting the FD for Well 25 as the predicted values were far off from the measured FD of Well 25. The results also did not follow the trend of the measured FD, and the R^2 value was 0.4607. Equation from Well 22, however, showed a slightly better prediction trend when applied to the dataset of Well 25 from a depth of 2580 to 2640 m. However, the R^2 value using the equation from Well 21 was better as compared to that when using the equation from Well 22 as the R^2 value for predicting the FD value for Well 25 using the equation from Well 22 was 0.172, which is low.

As observed in Table 2, the R^2 value was always the highest when the equation for a particular well was applied to its own data. For example, for Well 22, the R^2 value of measured versus predicted value was 0.666 and 0.598 for Well 25. As observed from the first method, Well 21 performed slightly better than the other two wells. To summarize, the first method to select only one well as the training well failed because no single equation from any well can be applied for the dataset from the other wells, warranting the need for other methods.

Using PCA and regression methods

PCA was performed to analyze the contribution of each log parameter toward FD and to simplify the variables by reducing the multicollinearity effects. It was conducted by first determining the number of relevant PCs with eigenvalue > 1. This is portrayed in the results and scree plot shown in Table 3 and Fig. 9, respectively. The scree plot also shows the eigenvalue versus the PCs and the cumulative variance against the PCs. From the figure, three PCs satisfy the eigenvalue criterion. The cumulative variance percentage shows that for the first three PCs, the cumulative value is 80.83% which means that 80.83% of the first three PCs contain the variation of the original log parameters or variables. The original 10 well log parameters were analyzed and Varimax rotation was applied, resulting that from 10 original parameters, the parameters were reduced to 8 due to some parameters were loaded onto several different PCs which were GR and Th, hence both of these parameters were removed to get better results. This has also been listed in Table 3.

PCA reduced the 10 log parameters into three uncorrelated PCs with 80.83% of the dataset information. The advantage of this method is that PCA removes the collinearity between the original variables. From Table 3, the first PC

Table 3 The results of PCA show the initial generated PC with the eigenvalue, variance, and cumulative variance

Principal component (PC)	Eigenvalue	Variance (%)	Cumulative variance (%)
1	3.683	46.033	46.033
2	1.654	20.676	66.710
3	1.129	14.116	80.826
4	0.595	7.441	88.267
5	0.388	4.856	93.122
6	0.279	3.491	96.614
7	0.213	2.666	99.280
8	0.058	0.720	100.000

The first three PCs have an eigenvalue > 1 and a cumulative variance of 80.826%

PCA principal component analysis, *PC* principal components

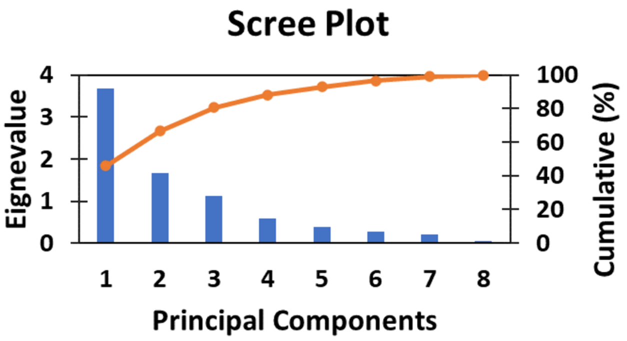

Fig. 9 The scree plot from the PCA results shows three PCs with eigenvalue of more than 1

has an eigenvalue of 3.683, the second PC of 1.654, and the third PC of 1.129.

The three PCs were independent of each other. The Varimax rotation was applied to understand if the variables were important or unimportant in rotated space as shown in Table 4. In the rotated loading, the first PC shows five variables highly correlated to PC1, two variables corresponding to PC2, and only one PC correlated to PC3. The other two variables were excluded from the analysis as their correlations with other variables, GR and K, were very high. The variables having PC loading > 0.5 belonged to the particular PC. For GR and K, both variables had PC loading > 0.5 in two different PCs, and thus, both variables were excluded to improve the results.

The first PC had five variables with PC loading > 0.5—RXO, RD, N, photoelectric effect, and S. The second PC had two other variables—U and D. The last PC had only one variable, K. Based on these results, the three new variables or PCs were newly categorized as fluid effects, fracture effects, and metamorphic lithology. These three new parameters were then verified against the FD and processed for regression analysis. As described in Sect. "Flow Chart", data from all three wells were combined for PCA as individual wells yielded different PCA results. Due to this, it was no longer possible to use the first method for regression analysis as described before. The approach here was to divide each well into upper and lower sections. A total of 70% of the data belonging to the upper section of the wells was grouped as a training set and processed for regression analysis. The generated equation was then applied to the lower section of the well with 30% of the data. The results are shown in Eq. 5:

$$\mathrm{FD(PCA)} = 2.870 - 0.143(\mathrm{PC1}) + 0.510(\mathrm{PC2}) + 0.180(\mathrm{PC3}) \tag{5}$$

where FD (PCA) is the FD function generated based on PCA results; PCA is the principal component analysis, and PC is the principal components.

The equation was applied to the lower section of each well as shown in Fig. 10. The predicted FD results are shown in green lines in Fig. 10.

The plots of predicted versus measured FD values are shown in Fig. 11, and the R^2 values were calculated to verify the validity of the second method. As shown in Figs. 10 and 11, the R^2 value for Wells 21, 22, and 25 were −0.492, 0.420, and 0.595, which were low. The predicted FD values also did not show the same trend as the measured values as shown in Fig. 10. However, the results were improved as compared to the first method. For Well 21 (Fig. 11a), the slope was negative, and the R^2 was −0.492. A negative slope indicated that the predicted values were lower than the measured values.

Using combination data of all wells as training set method

Since the second method showed improved results, the approach of separating the data to upper and lower sections was retained. The third approach was to combine some data from the upper section of each well and use them altogether for the regression analysis. However, instead of using the new variables from PCA results, this third method used the original 10 log parameters as the PCA results did not significantly improve the results. The combined data were also checked for normal distribution. This method was easier to use since the data would not be chosen at random but selected at a known depth. After the equation was generated, it was applied to the lower section of each well. The same data separation method was applied 70% of the data from the upper section was used as the training set, and the remaining was used as the testing set. The results for this section are shown in Fig. 12. The equation generated from this method is as follows:

$$\mathrm{FD(3)} = -36.265 - 0.351 \times N + 13.323 \times D + 0.122 \times S \tag{6}$$

where N is the neutron porosity value, D is the bulk density value, and S is the sonic porosity value. This new equation is labeled as FD (3) to indicate the third method used in this study.

From Fig. 12, the orange line was the selected data for the training set. After the equation from this training set had been generated, it was applied to the lower section of each well. The results are shown in the green line in Fig. 12. As can be observed, these results are better than the previous

Table 4 The results of PCA after applying Varimax rotation to show the three most significant components with the variables grouped together

Well log	Principal component (PC)		
	1	2	3
RXO	0.975	0.144	0.185
RD	0.959	0.024	0.165
N	0.721	−0.023	−0.421
PEF	0.653	−0.227	−0.377
S	0.578	−0.471	0.154
U	0.042	−0.871	−0.204
D	0.117	0.862	−0.285
K	0.117	−0.094	0.899

The original variables belong to a certain PC for PC loading value > 0.5

PCA principal component analysis, *PC* principal components

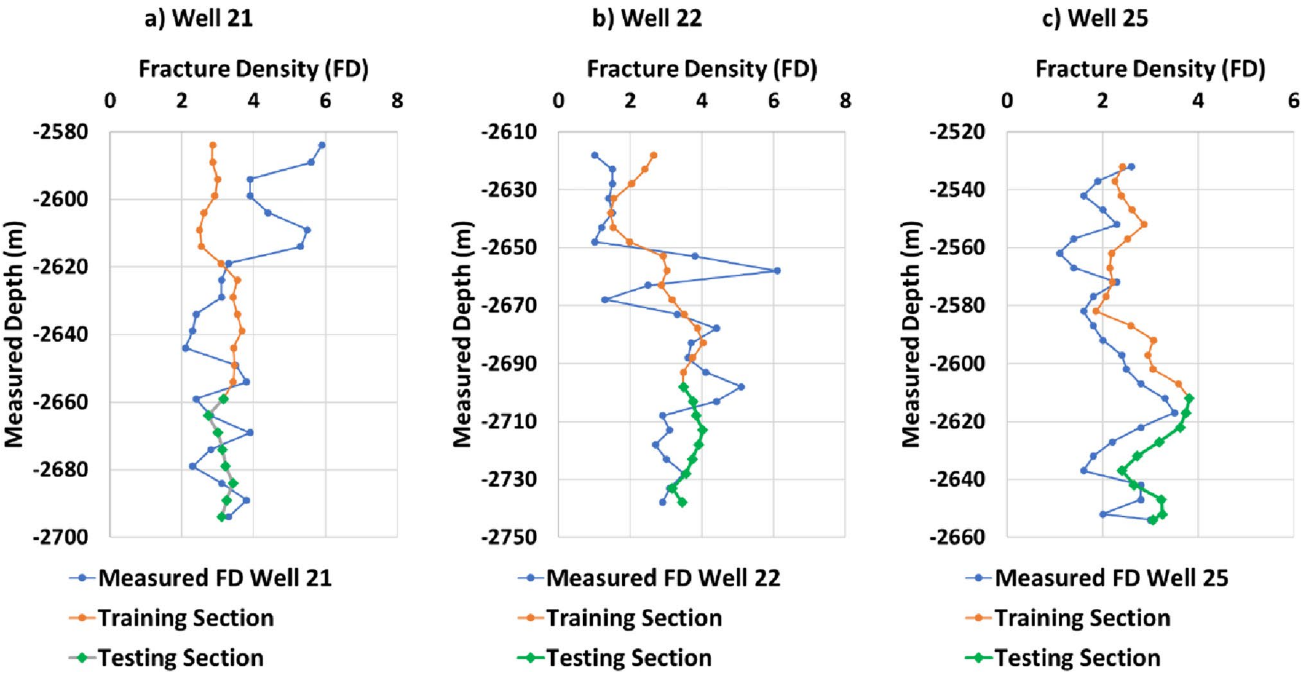

Fig. 10 The results of fracture density prediction based on the new variables generated from the principal component analysis. The orange line was the data used for the training section. The data from the upper section of all three wells were combined together, and Eq. 5 generated was applied to the lower section of the well as shown in green lines

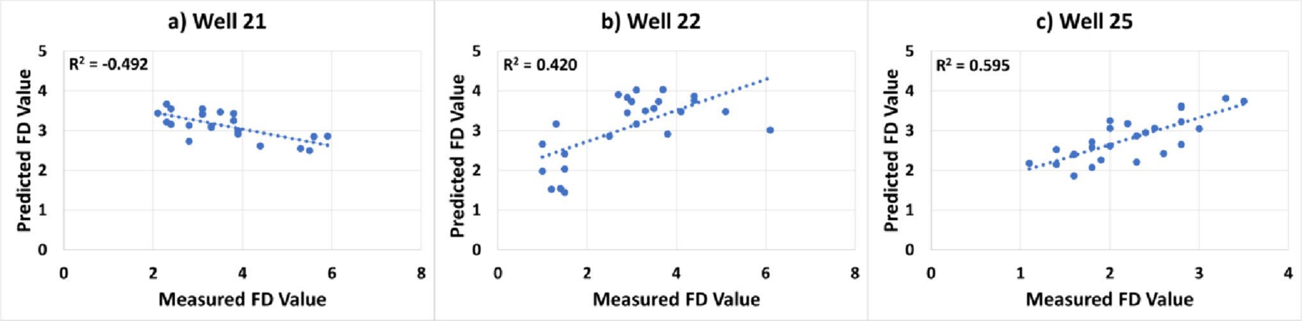

Fig. 11 The plot of the predicted FD value generated from Eq. 5 versus the measured FD value. These plots are from the results shown in Fig. 8. *FD* fracture density

methods. The predicted values show a better trend. The R^2 values for each well are also improved as can be seen in Fig. 13. For example, from Fig. 13a, the R^2 value when the predicted versus measured FD value was plotted for Well 21 is 0.748. Similar graphs were plotted for Wells 22 and 25 showing R^2 values of 0.724 and 0.745, respectively.

It can be summarized that if using the MLR method, all data from all wells can be combined for regression analysis for a better prediction of FD values.

Using GEP method

The approach of separating the upper and lower sections of the well was maintained for the nonlinear regression analysis in the gene expression tool. The training set was the same as described earlier. GEP results were produced in terms of expression trees (ETs) as seen in Fig. 14. From this figure, there are four sub-ETs that made up Eq. 7. Equation 7 is a combination of all sub-ETs. Each of the expression trees is rewritten in a mathematical equation form as seen from Eqs. 8–11. Equation 8 is the mathematical equation for sub-ET 1, Eq. 9 for sub-ET 2 and such on.

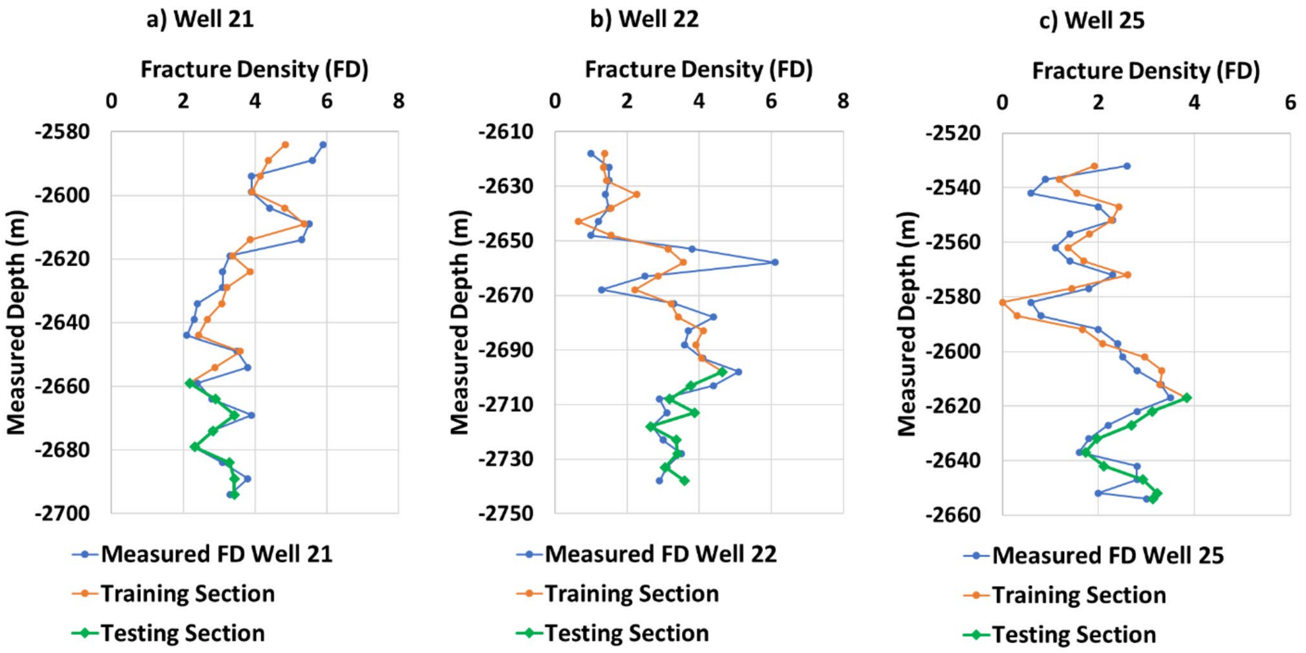

Fig. 12 The results of multiple linear regression based on the method of separating the well data into upper and lower sections. The original 10 well logs data were used in this method. Equation 6 was applied to the lower section of each well as shown in green lines

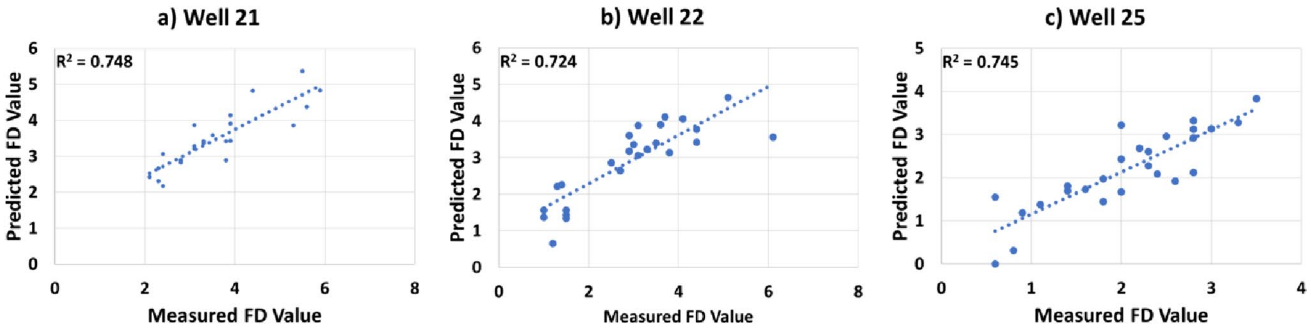

Fig. 13 The R^2 values form the plot of predicted versus measured FD values based on the third method and results in Fig. 10. *FD* fracture density

$$\mathrm{FD(GEP)} = y_1 + y_2 + y_3 + y_4 \tag{7}$$

$$y_1 = \mathrm{Tanh}\left[(N - ((8.78 + U) \times A\tan(K))) - \left(U^2 + \ln(RXO)\right)\right] \tag{8}$$

$$y_2 = (-2.88 + D) \times \left(\frac{1}{6 \cdot 89 - 4.08} + \left(\frac{0.86}{RXO} + (1 - RD)\right)\right) \tag{9}$$

$$\begin{aligned} y_3 = \min\Bigg(\left(\frac{RXO}{RD} \times (1 - RD)\right) \\ +(\min(RXO, RD) \times (-7.36)), \left(\frac{-5.98 + \mathrm{PE}}{2}\right) \times 1.70\Bigg) \end{aligned} \tag{10}$$

$$y_4 = \frac{21.86}{\min\left(\frac{\max(6.16,\mathrm{Th}) - 6.52D}{2}, \frac{\left(-\frac{\mathrm{Th}}{1.27} - 3.60\right)}{2} + N\right)} \tag{11}$$

This GEP Eq. 7 was the one used to generate results for FD prediction and the results can be observed in Figs. 15 and 16. Figure 15 shows the plot of measured versus predicted FD values. The training set is plotted in orange line and the testing set in green. For all wells, the training set showed similarity to the measured FD values. When the nonlinear equation was applied to the testing data set, the predicted FD values were also similar to the measured FD values.

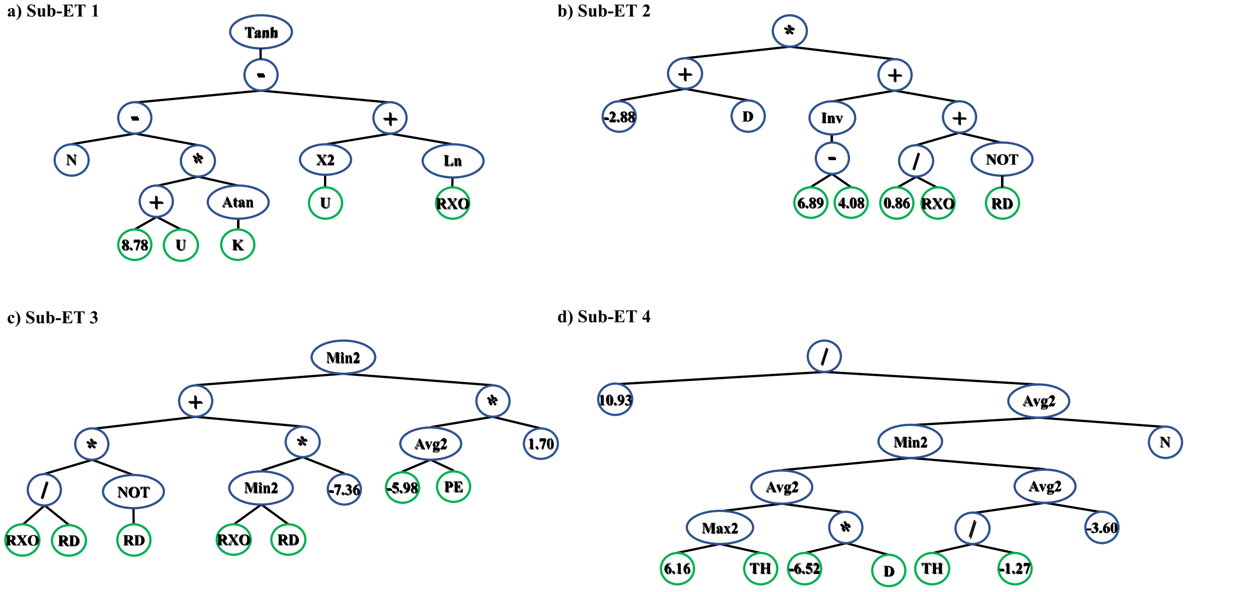

Fig. 14 Expression Tree of Gene Expression Programming results with all sub-expression trees

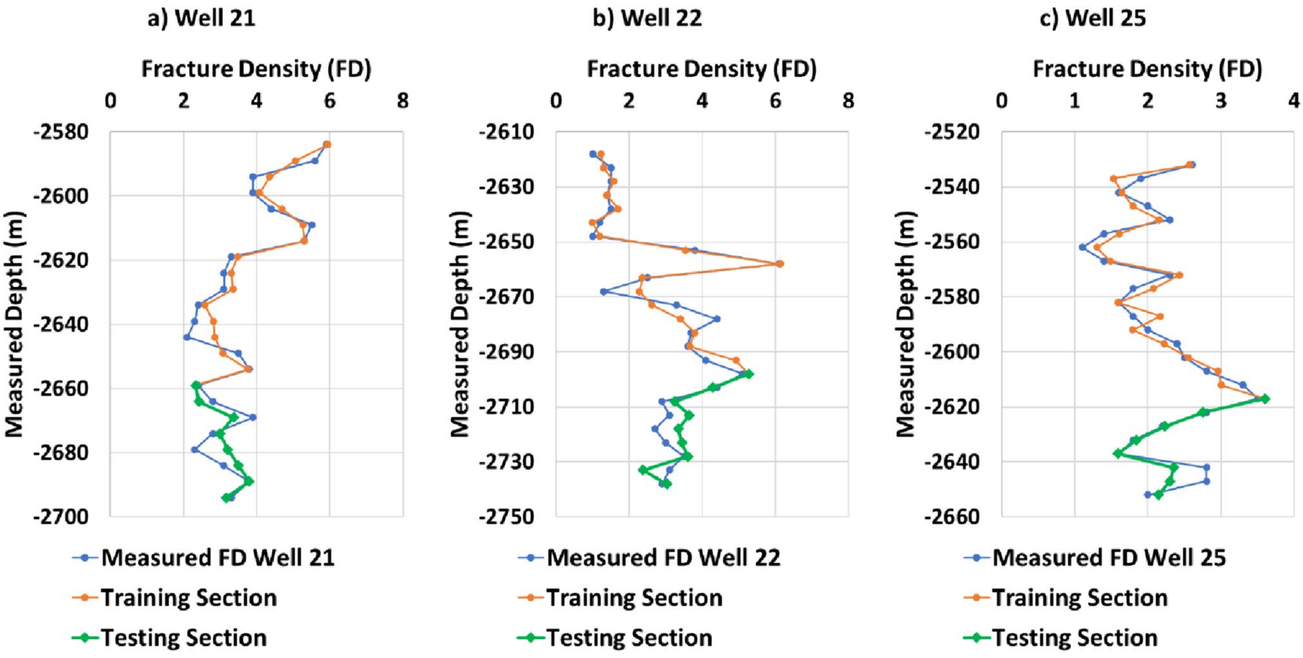

Fig. 15 The results are based on gene expression programming method. The prediction of the fracture density values was better and improved significantly as compared to the earlier methods. The R^2 values also improved as shown in Fig. 16; the R^2 values for Wells 21, 22, and 25 for predicted versus plotted FD values were 0.891, 0.893, and 0.869

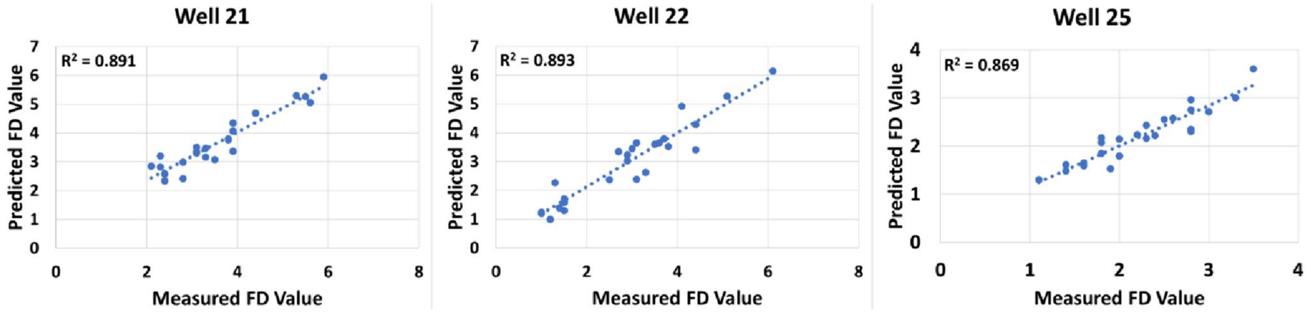

Fig. 16 The R^2 results are based on the results shown in Fig. 12. The GEP model prediction improved the R^2 values

Discussion

FD responses on conventional well logs

In this study, three wells with FD data were analyzed using different methods to be used as a model for predicting the FD of wells in case of non-availability of image logs or core samples. Many predictive models in the literature have discussed different complex algorithms to be used for FD prediction (Gao et al. 2023; Rajabi et al. 2021). Here, we present a direct and more economical way of using regression analysis and GEP to determine FD.

Well log responses are complicated in metamorphic rocks due to their complex system, and predicting fractures using the conventional logs would be helpful. A slight change in the log responses will yield different results; therefore, careful interpretation of well log data is crucial. The first method indicated that it was not suitable to use one well as a training set for other wells even for the same lithology. This could be since the first method was mainly based on the linear function of regression analysis, while the well log data are most nonlinear (Delavar 2022). The regression analysis of the first method could only predict the FD of its own data, i.e., the same data from the same well. Also, the data ranges in wells are quite different. For example, in Well 25, the standard deviation values for U and RXO were dissimilar and had a larger range as compared to that for the data from Wells 21 and 22.

However, all three equations (i.e., Eqs. 1–3) had RXO as one of the variables with an inverse relationship with FD. This is true, especially when a formation is drilled with water-based mud, the flushed zone near the wellbore will typically be filled with mud filtrate. This will cause the values of RXO to be lower than that for formation resistivity (or RD). This scenario was particularly accurate in the case of Well 25, as one of the variables for Eq. 4 was RD. The RD was positive in this equation, and RXO was negative, which clearly shows that in the case of fracture, on the log signals, there will be a crossover showing that the resistivity near the wellbore (flushed zone) will be lower, and the formation resistivity will be higher for the fracture zones. In this case, all three wells showed consistent results after MLR was conducted on each individual well. The role of RXO was also significant in the equation produced by GEP, as shown in Eq. 7. During the implementation of GEP to the dataset, all variables were included in the simulation, and GEP was selected as the best and the most significant variable in predicting FD. RXO also played a significant role in using the nonlinear approach.

Among the three porosity logs—D, N, and S—D was the best quantitative indicator of fracture as it showed reduced values for open fractures. This is because density and porosity have an inverse relationship, and during fractures, there will be an increase in porosity values, hence reducing the density of the rock. This particular case can be observed in Eq. 2 of Well 21, where D was negative, which proves the decrease in density values for fractures.

The well log interpretation and reservoir characterization are nonlinear approaches in reservoir engineering (Han and Bian 2018); therefore fracture analysis is better predicted with the nonlinear method to improve the accuracy classification (Delavar 2022).

FD influences PCA

Although the results from PCA were not used in the third and fourth methods in this study, they contribute to better insights for understanding the role of conventional well logs in the case of fracture evaluation.

From PCA results, the approach was to reduce the 10 variables into smaller variables while retaining information. The results show that three new variables could be categorized as fluid effects, fracture effects, and metamorphic lithology. The fluid effects variable contained the original RXO, RD, CN, PE, and AC. The fracture effect variables contained D and U variables, and the last new variable of metamorphic lithology contained the K variable. Conventional logs behave as a function of several factors such as lithology, porosity type, fluid flow, heterogeneity, and fractures (Aghli et al. 2020). From PCA results, the newly generated variables were in line with these cases. For the first variable, i.e., the fluid effects, the five original variables were correctly grouped. In case of fractures, the flushed zones would be filled with mud filtrate, which could affect the RXO. In addition, the fractures in the metamorphic rocks would also store fluids, such as hydrocarbons or water, and eventually influence the RD. Neutron porosity indicated the hydrogen content in the fractures filled with fluids in a similar manner to sonic porosity. Since the open fractures were the pathway for the fluids, the fluids existing in the fractures contributed to low sonic wave velocity. As a result, the presence of fractures will increase the sonic travel time.

The next variable from PCA was fracture effects, and this variable included two original parameters, i.e., U and D. Fractures are mostly influenced by the deformation (Mancktelow 2009; Li et al. 2022) and mineralization processes (Peacock and Mann 2005). The U is normally dissolved in fluids, and zones with high U concentrations can be associated with fracture zones (Hussein 2022). Therefore, spectral GR is very useful to analyze the presence of fractures. The high U log readings can indicate fluid migration along the fracture openings (Zazoun 2013). In addition, D is also a good indicator for fracture evaluation. D log has an influence on the FD (Aghli et al. 2020). The presence of fractures can heavily reduce the formation or the rock compaction; consequently, it will reduce the D values shown on the well log. The relation of bulk density and uranium content was opposite for fractures; the presence of fractures reduced the D value but increased the U content values. From Eq. 5, it was observed that this new variable—fracture effects had a positive influence on the FD function, which showed a strong relationship of FD with U and D.

Conclusions

This study aimed to predict fracture density (FD) in case where image logs or core samples are not available. This study explored a few methods in order to ensure that the simplest and most economical way could be performed to predict the FD. Conventional well logs are mostly available for all wells; therefore, the study leverage on the availability of these conventional well logs. Based on the methods used, conventional well logs could be very useful to predict FD. Some of the important conclusions of this study are listed below:

1. This work is one of the earlier studies that explore gene expression programming (GEP) to predict fracture density for metamorphic basement rock, specifically amphibolite and amphibolite biotite gneiss type of lithology. Based on the findings, the method works well to predict the FD with R2 values of at least above 0.86.
2. Different ways of separating input data were used in this study for multiple linear regression (MLR) and gene expression programming. This study proposed to separate data into 70% upper section of the well for training dataset and 30% of lower section of the well for testing dataset. The method proposed is reliable and able to work in both MLR and GEP.
3. This study showcased a few ways to generate a good predictive model for FD, essentially, this study compares two different methods which are linear and nonlinear approaches. Linear approach was done using MLR and nonlinear was done using GEP and the results showed that nonlinear approach using GEP gives better predictive model since most of the well log input parameters are nonlinear and nonlinear approaches are better to tackle nonlinear problems.
4. Different log responses were studied to investigate its relationship to FD and the findings of this study are consistent with previous published literature. For instance, flushed zone resistivity played an important role in terms of FD; for high FD, the flushed zone resistivity readings tended to be lower. As predicted, gamma ray alone did not contribute substantially to the FD prediction. However, as observed in the literature, spectral gamma rays could be very useful since the individual component of the gamma ray could be analyzed appropriately. Uranium content tended to contribute the highest in predicting FD. Porosity logs also influenced the prediction of FD, wherein bulk density showed the most effects on the FD.
5. The four proposed methods have their own advantages and disadvantages. The first method seemed to work well for its own data but failed to predict the FD of nearby wells. The second method using principal component analysis (PCA) was good in terms of analyzing the conventional well logs and how they contributed to the FD prediction; however, the newly categorized variables were not able to provide desirable results, although they improved the results from the first method. The integration of PCA into the regression analysis was considered favorable but a fine tuning to the data was necessary to be able to use the variables from the PCA.

For recommendation, future studies could explore the same method proposed in this study for various different metamorphic lithologies. Since this study only uses the amphibolite and amphibolite biotite gneiss, different studies exploring various lithologies and comparing the differences would be an interesting topic. In addition, various other methods could be employed and compared with the methods used in this study as well. For instance, multiple nonlinear regression analysis from other programs could be used to further improve the results. Also, some dynamic data could be added as input as well, for instance, water saturation, permeability, well test, and mud loss data during drilling to further improve the prediction.

Acknowledgements MOL (Hungarian Oil and Gas Plc.) is thanked for raw well-log data.

Author contributions MLH: conceptualization, methodology, validation, formal analysis, investigation, writing—original draft, writing—review and editing, visualization. TMT: conceptualization, methodology, resources, writing—review and editing, supervision.

Funding Open access funding provided by University of Szeged. The research was financially supported by the National Research, Development and Innovation Office (Grant No. K-138919) and the Open Access Fund by the University of Szeged (6312).

Data availability Some or all data used during the study are proprietary or confidential in nature and may only be provided with restrictions.

Declarations

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Ethical approval The authors confirm that this manuscript has not been previously published and is not currently under consideration by any other journal. Each named author has substantially contributed to conducting the underlying research and drafting this manuscript. Additionally, all of the authors have approved the contents of this paper. They have also agreed to the submission policies of Journal of Petroleum Exploration and Production Technology.

References

Abdideh M (2016) Estimation of the fracture density in reservoir rock using regression analysis of the petrophysical data. Geod Cartogr 42(3):85–91. https://doi.org/10.3846/20296991.2016.1226384

Afrasiabian B, Eftekhari M (2022) Prediction of mode I fracture toughness of rock using linear multiple regression and gene expression programming. J Rock Mech Geotech Eng 14(5):1421–1432. https://doi.org/10.1016/j.jrmge.2022.03.008

Aghli G, Soleimani B, Moussavi-Harami R, Mohammadian R (2016) Fractured zones detection using conventional petrophysical logs by differentiation method and its correlation with image logs. J Petrol Sci Eng 142:152–162. https://doi.org/10.1016/j.petrol.2016.02.002

Aghli G, Soleimani B, Tabatabai SS, Zahmatkesh I (2017) Calculation of fracture parameters and their effect on porosity and permeability using image logs and petrophysical data in carbonate Asmari reservoir, SW Iran. Arab J Geosci 10:1–14. https://doi.org/10.1007/s12517-017-3047-4

Aghli G, Moussavi-Harami R, Tokhmechi B (2020) Integration of sonic and resistivity conventional logs for identification of fracture parameters in the carbonate reservoirs (A case study, Carbonate Asmari Formation, Zagros Basin, SW Iran). J Petrol Sci Eng 186:106728. https://doi.org/10.1016/j.petrol.2019.106728

Albu I, Papa A (1992) Application of high-resolution seismics in studying reservoir characteristics of hydrocarbon deposits in Hungary. Geophysics 57(8):1068–1088. https://doi.org/10.1190/1.1443319

Algaifi HA, Alqarni AS, Alyousef R, Bakar SA, Ibrahim MW, Shahidan S, Ibrahim M, Salami BA (2021) Mathematical prediction of the compressive strength of bacterial concrete using gene expression programming. Ain Shams Eng J 12(4):3629–3639. https://doi.org/10.1016/j.asej.2021.04.008

Alizadeh S, Ta S, Ray AK, Lakshminarayanan S (2022) Determination of density and viscosity of crude oil samples from FTIR data using multivariate regression variable selection and classification. IFAC-Papers Online 55(7):845–850. https://doi.org/10.1016/j.ifacol.2022.07.550

Ari D, Alagoz BB (2023) A differential evolutionary chromosomal gene expression programming technique for electronic nose applications. Appl Soft Comput 136:110093. https://doi.org/10.1016/j.asoc.2023.110093

Aydogan MS, Alacali S, Arslan G (2023) Prediction of moment redistribution capacity in reinforced concrete beams using gene expression programming. Structures 47:2209–2219. https://doi.org/10.1016/j.istruc.2022.12.054

Cai M, Wang Y, Zhao W, Shi X, Li T (2023) Study on local brittleness of rock based on multiple linear regression method: case study of shahejie formation. Geofluids. https://doi.org/10.1155/2023/6189068

Chu HH, Khan MA, Javed M, Zafar A, Khan MI, Alabduljabbar H, Qayyum S (2021) Sustainable use of fly-ash: use of gene-expression programming (GEP) and multi-expression programming (MEP) for forecasting the compressive strength geopolymer concrete. Ain Shams Eng J 12(4):3603–3617. https://doi.org/10.1016/j.asej.2021.03.018

Delavar MR (2022) Hybrid machine learning approaches for classification and detection of fractures in carbonate reservoir. J Petrol Sci Eng 208:109327. https://doi.org/10.1016/j.petrol.2021.109327

Dunkler D, Plischke M, Leffondré K, Heinze G (2014) Augmented backward elimination: a pragmatic and purposeful way to develop statistical models. PLoS ONE 9(11):e113677. https://doi.org/10.1371/journal.pone.0113677

Ellis DV, Singer JM (2007) Well logging for earth scientists, vol 692. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-4602-5

Enayatollahi I, Aghajani Bazzazi A, Asadi A (2014) Comparison between neural networks and multiple regression analysis to predict rock fragmentation in open-pit mines. Rock Mech Rock Eng 47:799–807. https://doi.org/10.1007/s00603-013-0415-6

Esmaeilpour M, Nitschke F, Kohl T (2024) GenEOS: an accurate equation of state for the fast calculation of two-phase geofluids properties based on gene expression programming. Comput Phys Commun 297:109068. https://doi.org/10.1016/j.cpc.2023.109068

Ferreira C (2001) Gene expression programming a new adaptive algorithm for solving problems. arXiv preprint cs. https://doi.org/10.48550/arXiv.cs/0102027

Gamal M, El-Araby AA, El-Barkooky AN, Hassan A (2022) Detection and characterization of fractures in the Eocene Thebes formation using conventional well logs in October field, Gulf of Suez, Egypt. Egypt J Pet 31(3):1–9. https://doi.org/10.1016/j.ejpe.2022.06.001

Gao G, Hazbeh O, Davoodi S, Tabasi S, Rajabi M, Ghorbani H, Radwan AE, Csaba M, Mosavi AH (2023) Prediction of fracture density in a gas reservoir using robust computational approaches. Front Earth Sci 10:1023578. https://doi.org/10.3389/feart.2022.1023578

Geng X, Qi M, Liu J, He C, Li Y (2021) Application of principal component analysis on water flooding effect evaluation in natural edge-bottom water reservoir. J Pet Explor Prod 11:439–449. https://doi.org/10.1007/s13202-020-01055-4

Habibi MJ, Mokhtari AR, Baghbanan A, Namdari S (2014) Prediction of permeability in dual fracture media by multivariate regression analysis. J Petrol Sci Eng 120:194–201. https://doi.org/10.1016/j.petrol.2014.06.016

Han B, Bian X (2018) A hybrid PSO-SVM-based model for determination of oil recovery factor in the low-permeability reservoir. Petroleum 4(1):43–49. https://doi.org/10.1016/j.petlm.2017.06.001

Harrell FE (2001) Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis, vol 608. Springer, New York. https://doi.org/10.1007/978-1-4757-3462-1

Hasan ML, Tóth TM (2023) Localization of potential migration pathways inside a fractured metamorphic hydrocarbon reservoir using well log evaluation (Mezősas field, Pannonian Basin). Geoenergy Sci Eng 225:211710. https://doi.org/10.1016/j.geoen.2023.211710

Hassan WH, Hussein HH, Alshammari MH, Jalal HK, Rasheed SE (2022) Evaluation of gene expression programming and artificial neural networks in PyTorch for the prediction of local scour depth around a bridge pier. Results Eng 13:100353. https://doi.org/10.1016/j.rineng.2022.100353

Heinze G, Wallisch C, Dunkler D (2018) Variable selection–a review and recommendations for the practicing statistician. Biom J 60(3):431–449. https://doi.org/10.1002/bimj.201700067

Holland John H (1975) Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor

Hussein HS (2022) Carbonate fractures from conventional well log data, Kometan Formation, Northern Iraq case study. J Appl Geophys 206:104810. https://doi.org/10.1016/j.jappgeo.2022.104810

İnce İ, Bozdağ A, Fener M, Kahraman S (2019) Estimation of uniaxial compressive strength of pyroclastic rocks (Cappadocia, Turkey) by gene expression programming. Arab J Geosci 12:1–13. https://doi.org/10.1007/s12517-019-4953-4

Ja'fari A, Kadkhodaie-Ilkhchi A, Sharghi Y, Ghanavati K (2012) Fracture density estimation from petrophysical log data using the adaptive neuro-fuzzy inference system. J Geophys Eng 9(1):105–114. https://doi.org/10.1088/1742-2132/9/1/013

Jahed Armaghani D, Safari V, Fahimifar A, Mohd Amin MF, Monjezi M, Mohammadi MA (2018) Uniaxial compressive strength prediction through a new technique based on gene expression programming. Neural Comput Appl 30:3523–3532. https://doi.org/10.1007/s00521-017-2939-2

Jalal FE, Iqbal M (2023) Unconfined compression strength modelling of expansive soils for sustainable construction: GEP vs MEP. Environ Earth Sci 82(14):364. https://doi.org/10.1007/s12665-023-11049-0

Javed MF, Amin MN, Shah MI, Khan K, Iftikhar B, Farooq F, Aslam F, Alyousef R, Alabduljabbar H (2020) Applications of gene expression programming and regression techniques for estimating compressive strength of bagasse ash based concrete. Crystals 10(9):737. https://doi.org/10.3390/cryst10090737

Juhász A, Tóth TM, Ramseyer K, Matter A (2002) Connected fluid evolution in the fractured crystalline basement and overlying sediments, Pannonian Basin, SE Hungary. Chem Geol 182:91–120. https://doi.org/10.1016/S0009-2541(01)00269-8

Kaiser HF (1960) The application of electronic computers to factor analysis. Educ Psychol Measur 20(1):141–151. https://doi.org/10.1177/001316446002000116

Khosravi M, Tabasi S, Eldien HH, Motahari MR, Alizadeh SM (2022) Evaluation and prediction of the rock static and dynamic parameters. J Appl Geophys 199:104581. https://doi.org/10.1016/j.jappgeo.2022.104581

Konaté AA, Pan H, Ma H, Cao X, Ziggah YY, Oloo M, Khan N (2015) Application of dimensionality reduction technique to improve geophysical log data classification performance in crystalline rocks. J Petrol Sci Eng 133:633–645. https://doi.org/10.1016/j.petrol.2015.06.035

Konaté AA, Ma H, Pan H, Qin Z, Ahmed HA (2017) Lithology and mineralogy recognition from geochemical logging tool data using multivariate statistical analysis. Appl Radiat Isot 128:55–67. https://doi.org/10.1016/j.apradiso.2017.06.041

Li T, Wang R, Wang Z, Zhao M, Li L (2018) Prediction of fracture density using genetic algorithm support vector machine based on acoustic logging data. Geophysics 83(2):D49–D60. https://doi.org/10.1190/geo2017-0229.1

Li L, Zhang GZ, Liu JZ, Han L, Zhang JJ (2021) Estimation of fracture density and orientation from azimuthal elastic impedance difference through singular value decomposition. Pet Sci 18(6):1675–1688. https://doi.org/10.1016/j.petsci.2021.09.037

Li Z, Wang L, Li W (2022) Mechanical behavior and fracture characteristics of rock with prefabricated crack under different triaxial stress conditions. Minerals 12(6):673. https://doi.org/10.3390/min12060673

Luo HY, Tang YM (2013) Application of Monte Carlo to improve the accuracy of identifying fracture by conventional logs. In: Applied mechanics and materials, vol 295. Trans Tech Publications Ltd, pp 3237–3242. https://doi.org/10.4028/www.scientific.net/AMM.295-298.3237

Lv Q, Zhou T, Zheng R, Nakhaei-Kohani R, Riazi M, Hemmati-Sarapardeh A, Wang W (2023) Application of group method of data handling and gene expression programming for predicting solubility of CO_2-N_2 gas mixture in brine. Fuel 332:126025. https://doi.org/10.1016/j.fuel.2022.126025

Lyu W, Zeng L, Liu Z, Liu G, Zu K (2016) Fracture responses of conventional logs in tight-oil sandstones: a case study of the Upper Triassic Yanchang Formation in southwest Ordos Basin. China AAPG Bull 100(9):1399–1417. https://doi.org/10.1306/04041615129

Mancktelow NS (2009) Fracture and flow in natural rock deformation. Trabajos Geología 29:29–35

Mantel N (1970) Why stepdown procedures in variable selection. Technometrics 12(3):621–625. https://doi.org/10.1080/00401706.1970.10488701

Martinez LP, Hughes RG, Wiggins ML (2002) Identification and characterization of naturally fractured reservoirs using conventional well logs. The University of Oklahoma, pp. 1–23

Molnár L, Tóth TM, Schubert F (2015) Structural controls on petroleum migration and entrapment within the faulted basement blocks of Szeghalom Dome (Pannonian Basin, SE Hungary). Geologia Croatica 68(3):247–259. https://doi.org/10.4154/GC.2015.19

Peacock DCP, Mann A (2005) Evaluation of the controls on fracturing in reservoir rocks. J Pet Geol 28(4):385–396. https://doi.org/10.1111/j.1747-5457.2005.tb00089.x

Pei J, Zhang Y (2022) Prediction of reservoir fracture parameters based on the multi-layer perceptron machine-learning method: a case study of ordovician and cambrian carbonate rocks in Nanpu Sag, Bohai Bay Basin. China Processes 10(11):2445. https://doi.org/10.3390/pr10112445

Qiu X, Tan C, Lu Y, Yin S (2022) Evaluation of fractures using conventional and FMI logs, and 3D seismic interpretation in continental tight sandstone reservoir. Open Geosci 14(1):530–543. https://doi.org/10.1515/geo-2022-0372

Rajabi M, Beheshtian S, Davoodi S, Ghorbani H, Mohamadian N, Radwan AE, Alvar MA (2021) Novel hybrid machine learning optimizer algorithms to prediction of fracture density by petrophysical data. J Pet Explor Prod Technol 11:4375–4397. https://doi.org/10.1007/s13202-021-01321-z

Ren Q, Zhang H, Zhang D, Zhao X (2023) Lithology identification using principal component analysis and particle swarm optimization fuzzy decision tree. J Petrol Sci Eng 220:111233. https://doi.org/10.1016/j.petrol.2022.111233

Rostami A, Arabloo M, Kamari A, Mohammadi AH (2017) Modeling of CO_2 solubility in crude oil during carbon dioxide enhanced oil recovery using gene expression programming. Fuel 210:768–782. https://doi.org/10.1016/j.fuel.2017.08.110

Schubert F, Diamond LW, Tóth TM (2007) Fluid-inclusion evidence of petroleum migration through a buried metamorphic dome in the Pannonian Basin, Hungary. Chem Geol 244(3–4):357–381. https://doi.org/10.1016/j.chemgeo.2007.05.019

Serra O (1986) Advanced interpretation of wireline logs: Schlumberger well services, Houston, Document No. M-090028, pp 2–3

Shahabi-Ghahfarokhy A, Nakhaei-Kohani R, Amar MN, Hemmati-Sarapardeh A (2022) Modelling density of pure and binary mixtures of normal alkanes: comparison of hybrid soft computing techniques, gene expression programming, and equations of state. J Petrol Sci Eng 208:109737. https://doi.org/10.1016/j.petrol.2021.109737

Shalaby MR, Islam MA (2017) Fracture detection using conventional well logging in carbonate Matulla Formation, Geisum oil field, southern Gulf of Suez, Egypt. J Pet Explor Produ Technol 7:977–989. https://doi.org/10.1007/s13202-017-0343-1

Sharifi Y, Moghbeli A (2020) New predictive models via gene expression programming and multiple nonlinear regression for SFRC beams. J Market Res 9(6):14294–14306. https://doi.org/10.1016/j.jmrt.2020.10.026

Tari G, Horváth F, Rumpler J (1992) Styles of extension in the Pannonian Basin. Tectonophysics 208(1–3):203–219. https://doi.org/10.1016/0040-1951(92)90345-7

Tari G, Dövényi P, Dunkl I, Horváth F, Lenkey L, Stefanescu M, Szafian P, Tóth T (1999) Lithospheric structure of the Pannonian basin derived from seismic, gravity and geothermal data. Geol Soc London Spec Publ 156:215–250. https://doi.org/10.1144/GSL.SP.1999.156.01.12

Tiryaki B (2008) Predicting intact rock strength for mechanical excavation using multivariate statistics, artificial neural networks, and regression trees. Eng Geol 99(1–2):51–60. https://doi.org/10.1016/j.enggeo.2008.02.003

Tokhmchi B, Memarian H, Rezaee MR (2010) Estimation of the fracture density in fractured zones using petrophysical logs. J Petrol Sci Eng 72(1–2):206–213. https://doi.org/10.1016/j.petrol.2010.03.018

Tokhmechi B, Memarian H, Noubari HA, Moshiri B (2009a) A novel approach proposed for fractured zone detection using petrophysical logs. J Geophys Eng 6(4):365–373. https://doi.org/10.1088/1742-2132/6/4/004

Tokhmechi B, Memarian H, Rasouli V, Noubari HA, Moshiri B (2009b) Fracture detection from water saturation log data using a Fourier–wavelet approach. J Petrol Sci Eng 69(1–2):129–138. https://doi.org/10.1016/j.petrol.2009.08.005

Tóth TM (2012) Geochemistry of Variscan amphibolites from the metamorphic basement of the Körös Complex (Tisza Block, Hungary). Carpathian J Earth Environ Sci 7(3):5–18

Tóth TM, Schubert F (2018) Evolution of the arc-derived orthogneiss recorded in exotic xenoliths of the Koros Complex (Tisza Megaunit, SE Hungary). J Geosci 63:21–46. https://doi.org/10.3190/jgeosci.253

Tóth TM, Zachar J (2006) Petrology and deformation history of the metamorphic basement in the Mezősas-Furta crystalline high (SE Hungary). Acta Geol Hung 49(2):165–188. https://doi.org/10.1556/ageol.49.2006.2.4

Tóth TM, Schubert F, Zachar J (2000) Neogene exhumation of the variscan szeghalom dome, pannonian basin, E. Hungary. Geol J 35(3–4):265–284. https://doi.org/10.1002/gj.861

Tóth TM, Molnár L, Körmös S, Czirbus N, Schubert F (2020) Localisation of ancient migration pathways inside a fractured metamorphic hydrocarbon reservoir in south-east hungary. Appl Sci 10(20):7321. https://doi.org/10.3390/app10207321

Tóth TM, Fiser-Nagy Á, Kondor H, Molnár L, Schubert F, Vargáné Tóth I, Zachar J (2021) The metamorphic basement of the great hungarian plain: from zwischengebirge towards a variegated mosaic. Földtani Közlöny 151(1):3–26. https://doi.org/10.23928/foldt.kozl.2021.151.1.3

Tóth E, Hrabovszki E, Tóth TM (2023) Using geophysical log data to predict the fracture density in a claystone host rock for storing high-level nuclear waste. Acta Geod Geophys. https://doi.org/10.1007/s40328-023-00407-w

Vass I, Tóth TM, Szanyi J, Kovács B (2018) Hybrid numerical modelling of fluid and heat transport between the overpressured and gravitational flow systems of the Pannonian Basin. Geothermics 72:268–276. https://doi.org/10.1016/j.geothermics.2017.11.013

Verga FM, Carugo C, Chelini V, Maglione R, Bacco GD (2000) Detection and characterization of fractures in naturally fractured reservoirs. In: SPE Annual Technical Conference and Exhibition. OnePetro. https://doi.org/10.2118/63266-MS

Yang H, Pan H, Wu A, Luo M, Konaté AA, Meng Q (2017) Application of well logs integration and wavelet transform to improve fracture zones detection in metamorphic rocks. J Petrol Sci Eng 157:716–723. https://doi.org/10.1016/j.petrol.2017.07.057

Yuan Z, Chen L, Liu G, Shao W, Zhang Y, Yang W (2023) Physics-based Bayesian linear regression model for predicting length of mixed oil. Geoenergy Sci Eng 223:211466. https://doi.org/10.1016/j.geoen.2023.211466

Zazoun RS (2013) Fracture density estimation from core and conventional well logs data using artificial neural networks: the Cambro-Ordovician reservoir of Mesdar oil field, Algeria. J Afr Earth Sci 83:55–73. https://doi.org/10.1016/j.jafrearsci.2013.03.003

Zhang R, Zhang S (2024) Coefficient of permeability prediction of soils using gene expression programming. Eng Appl Artif Intell 128:107504. https://doi.org/10.1016/j.engappai.2023.107504

Zhang L, Wang Z, Xu R, Cheng H, Ren L, Lin R (2023a) Modeling and analysis of hydraulic fracture complexity index in sandy conglomerate reservoirs based on genetic expression programming—A case study in Xinjiang Oilfield. Front Earth Sci 10:1051184. https://doi.org/10.3389/feart.2022.1051184

Zhang Y, Zhang X, Sun Y, Gong A, Li M (2023b) An adaptive ensemble learning by opposite Multi-Verse Optimizer and Its application on fluid identification for unconventional oil reservoirs. Front Earth Sci 11:1116664. https://doi.org/10.3389/feart.2023.1116664

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.