https://dx.doi.org/10.1016/j.petrol.2021.109853

Comparison of Machine Learning Techniques for Predicting Porosity of Chalk

Meysam Nourani[1], Najeh Alali[2], Saeed Samadianfard[3], Shahab S. Band[4*], Kwok-wing Chau[5], C.M. Shu

[1]Reservoir Geology Department, Geological Survey of Denmark and Greenland (GEUS), Copenhagen, Denmark;

[2]College of Petroleum Engineering, Al-Ayen University, Thi-Gar, 64001, Iraq;

[3]Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Tabriz, Iran,

[4]Future Technology Research Center, National Yunlin University of Science and Technology, Douliou, Yunlin 64002, Taiwan, ROC;

[5]Department of Civil and Environmental Engineering, Hong Kong Polytechnic University, Hong Kong, People's Republic of China

(*corresponding author email: shamshirbands@yuntech.edu.tw)

Comparison of Machine Learning Techniques for Predicting Porosity of Chalk

Abstract

Precise and fast estimation of porosity is a vital element of reservoir characterization. A new technology for fast and reliable porosity prediction of chalk samples is presented by applying machine learning methods and X-ray fluorescence (XRF) elemental analysis. Input parameters of prediction models are based on rapid and accurate elemental analysis of chalk samples obtained from Hand-held X-ray fluorescence (HH-XRF) measurements. The intelligent models, including Random Forest (RF), Multilayer perceptron (MLP), Random Forest integrated by Genetic Algorithm (GA-RF) and Multilayer Perceptron integrated by Genetic Algorithm (GA-MLP), are trained and tested based on samples consisting of outcrop chalk samples from Rørdal and Stevns Klint and core samples from Ekofisk Formation in the North Sea. Results are evaluated by sustainability index (SI), determination coefficient (R^2), correlation coefficient (CC), and Willmott's Index of agreement (WI). Results indicate that the combination of GA-RF intelligent method with XRF elemental analysis successfully provides an accurate model by 0.99, 0.02, 0.995 and 0.99 respectively for CC, SI, WI and R^2, respectively.

Keywords: Porosity, Chalk, Hand-held X-ray fluorescence, Random Forest, Multilayer perceptron, Random Forest Optimized by Genetic Algorithm, Multilayer Perceptron Optimized by Genetic Algorithm

XRF	X-ray fluorescence	GA	Genetic Algorithm
HH-XRF	Hand-held X-ray fluorescence	WI	Willmott's Index of agreement
RF	Random Forest	CC	Correlation coefficient
MLP	Multilayer perceptron	SI	Sustainability index
MLP-GA	Multilayer Perceptron integrated by Genetic Algorithm	R2	Determination coefficient

NMR	Nuclear magnetic resonance	AI	Artificial intelligence
ML	Machine learning	GP	Genetic programming
ANN	Artificial neural network	FL	Fuzzy logic
CNN	Convolutional neural network	TOC	Total organic carbon
DT	Decision tree	EA	Evolutionary algorithms
MLP	Multi layered perceptron	RF	Random forest

1 Introduction

Chalk creates an economically strategic lithology, which supplies large hydrocarbon reserves across Texas and northwest Europe. Chalk consists of lithified carbonate ooze that is integrated during burial through cementation and compaction [1, 2]. More than 80% of carbonate ooze is made of cocoolite parts, nannoconids and foraminifers, with an indefinite amount of clay minerals, whose values are variable and not constant [3-5]. After deposition, much of chalk from North Sea Central Graben moved and the reworking of the chalk influences reservoirs today. The formations from which oil and gas are produced, Ekofisk, Tor and Hod Formations, display some resemblances across the basin, but also variations from field to field [6].

The porosity of rock is the ratio between the pore space volume to the bulk volume of the rock and is expressed as a percentage [7, 8]. Interconnected or effective porosity, which is defined as the volume of connected pores to total bulk rock volume, is generally interested in reservoir engineering [9]. Typical porosity values generally vary from 5 to 30%. 15% porosity is a very typical value [10]. The chalk porosity can be closely related to the initial sediment composition and diagenetic background [11]. The high porosity of chalk in North Sea mines has made it famous and popular (about 20 to 40%) [9]. The Tor Formation has better porosity conditions than other chalk Formations [12]. Due to the over-pressured nature of the basin in North Sea of Denmark, depth reduction cannot be a factor in reducing the chalk porosity [13]. But generally, the chalk porosity decreases substantially with depth. Diagenetic variations would usually reason porosity reduction from around 70-80% at the surface to about 10% at burial depths of 2 km. Several features such as the presence of hydrocarbons, halokinesis, overpressure, the burial depth and post-depositional tectonics have prevented diagenesis from applying its maximum potential in the North Sea. Therefore, chalk reservoirs have reserved their high porosities [14, 15].

Rock effective porosity can be achieved from conducting laboratory measurements through core sample analysis. A range of laboratory methods such as imbibition, mercury injection and gas expansion methods is available for determination of sample pore space volume in core analysis [16, 17]. The great majority of pore volume determinations on North Sea chalk samples during the last approximately 30 years have been measured by gas expansion method [17]. Bulk volume can be measured by submersing the sample in a mercury bath, or by using a mercury displacement pump, or by caliper techniques [16, 17]. Porosity can also be estimated using open-hole well logs such as sonic, density, neutron and nuclear magnetic resonance (NMR) logs [18, 19]. However, core analysis provides accurate and reproducible porosity data [20], which is relatively time consuming, expensive and not always accessible.

Hand-held X-ray fluorescence (HH-XRF) has proved to be a rapid, powerful, reliable and stable tool for field-based or laboratory, geochemical characterization [21-23]. Previously HH-XRF and principal component analysis (PCA) have been successfully used to consider the relationship between concentrations of elements and porosity of chalk samples. Nourani et al. (2019) reported that the chalk porosity can be effectively controlled by aluminum, Fe, K, calcium and silicon. Aluminum, calcium and silicon contents of chalk emanate from clay, calcite and silica, respectively [24, 25]. Chakraborty et. al (2017) [26] employed support vector machine based classifier using portable XRF to estimate the calcium concentration of 75 soils. Findings indicated that the carbonate formation staged on only 22.6% of the samples. Ca content of intact aggregates had a correlation by about (r=0.89) vs. ground soil samples.

With the advent of new technologies, including topics related to artificial intelligence (AI) [27], various sectors of industry, including the oil and gas industry at different levels of their performance have been greatly affected by these technologies. Machine learning (ML), as one of the popular subsets of artificial intelligence, has always been used in various topics [28]. In this way, Rostami et al (2018) employed Least-Square Support Vector Machine for providing a new platform as a correlative model for CO2 solubility. Results were evaluated by the average absolute relative deviation and coefficient of determination by comparing the predicted and target values. Accordingly, it was concluded that the proposed technique could successfully cope with the task for improving the problem statement [29]. Saghafi and Arablo (2018) proposed a novel technique using Genetic Programming (GP) platform for the estimation of the gas condensate compressibility factor in the presence of dew point pressure. The results were analyzed using sensitivity analysis based on Spearman and Pearson approaches to conclude the effect of each input parameter on the target variable [30]. Okwu and Nwachukwu (2018) developed state-of-the-art fuzzy logic applications in petroleum exploration and production operations considering non-deterministic input variables, main challenges and possible solutions using fuzzy logic analysis [31]. Sircar et al (2021) provided a comprehensive state-of-the-art in the context of evaluating ML-based techniques for data processing in different terms of upstream oil and gas industries. Besides, the study discussed the main limitations, research gaps and the future perspectives for achieving a smart development in the field [32]. As it is clear, ML-methods have a deep influence in the field of oil and gas, which is due to its high reliability and accuracy in various operations. For this reason, the application of ML-methods in specialized branches of oil and gas has also become important.

Alnahwi and G. Loucks (2019) [33] employed ML-based artificial neural networks (ANNs) analysis of X-ray fluorescence data to estimate mineralogies and also evaluate the quality of the developed models. The online Neural

Designer software was also employed to conduct the modeling process. Quantitative laboratory-measured X-ray diffraction mineralogies and total organic carbon (TOC) were employed to perform high-resolution semiquantitative modeling, and to generate mineralogic and organic matter models. Findings indicated that the proposed method was a promising method. Zhao et al. (2020) [34] employed a ML model based on random forest algorithm to develop an analytical approach for the special core analysis dataset that illustrated a key missing feature in the prediction process. The study conducted the missing feature and proposed the proper characteristics in combination with in-situ fluid saturations. Andrianov and Nick (2000) [35] employed ML-based analytical method along with the discrete fracture simulations to generate a dual porosity model. Accordingly, a pixelated representation technique was employed to characterize the fracture geometry. Then, a convolutional neural network (CNN), as ML-based model, was used to map the fracture parametrization and the upscaled parameters.

Results of the analyzes performed by these studies showed that the application of ML-based methods (from simple to complex) had become a popular method in recent years in the field of the study. All recommendations about future perspectives have one thing in common, and that is the application of new and hybrid ML methods in different types of data using different evolutionary algorithms (EA). This action leads to increasing trust in ML-based methods and finding the strengths and weaknesses of these methods in a fully practical way. It should be noted that the use of ML-based methods has no limitations in terms of dataset type and method of analysis. On the other hand, the models used should not only be more accurate and have a simpler training process, but also need to reduce the time to perform computations and analyzes as much as possible. One of the advantages of using ANN-based methods with hybrid architecture and using EA algorithms is their simplicity, less process time than deep learning-based methods and their high sustainability. Therefore, the objective of this study is to investigate the abilities of artificial intelligence techniques for rapid and accurate estimation of porosity for chalk samples. This paper deals with a comparison of different models for predicting porosity of chalk samples by coupling a ML concept and elemental analysis of chalk obtained from HH-XRF. The ML approach is calibrated and tested via outcrop chalk samples from Rørdal and Stevns Klint and core samples from Ekofisk Formation in the North Sea. In addition, different intelligent methods, namely RF, MLP, RF-GA and MLP-GA, are undertaken and their accuracies are compared.

2 Modeling techniques

ANNs are mathematical/computational models for distinguishing nonlinear relationships connecting inputs to outputs in complicated systems [36-38]. This category of ML is inspired from human brain system mostly by familiarizing the conception of biological neurons [39-41]. RF, MLP, GA-RF and GA-MLP are recognized methods and applied for modeling purposes in various complicated engineering tasks [42-45].

2.1. RF method

RF provides an exceptional mixture of model interpretability and prediction accuracy among famous ML methods. The random collective strategies employed in RF aid it to accomplish better generalizations in addition to accurate predictions [46, 47]. Many types of applications can be predicted by RF accurately. It can evaluate the sensitivity of each feature in model training process. In addition, the trained model can successfully evaluate and measure two-by-two proximity between samples [48-50]. RF is a set of tree-based estimators $h(x.\theta_k)$, k= 1…..K where θ_k refer to independent and identically distributed random vectors and x denotes the target vector of length p with associated random vector. RF-based estimation is an unweighted average over the set with the following expression $h(x) = \left(\frac{1}{\mathrm{K}}\right)\sum_{k=1}^{K} h(x,\theta_k)$ [51].

2.2. MLP method

MLP can be developed to estimate any measurable dataset and function. It proves no preliminary assumptions about the dataset. It can be developed to generalize when introduced with hidden data. MLP can estimate nonlinear functions [52-54]. It is defined by fully connected nodes in the next and previous layer, and has been considered as providing a nonlinear recognition between corresponding output and input vectors [55]. MLP can have one or more hidden layers followed by an output layer [56]. The nodes are linked by output signals and weights, which are a function of the summation of the independent variables (as input matrix) to the node implemented by an activation function, or a transfer function. It is the compliance of many nonlinear transfer functions that aids MLP to estimate non-linear functions. The result of a node is scaled by the connecting weight. Then it can be considered as a feed to the nodes in the next layer [57-59].

2.3. GA-RF method

GA-RF method principally includes two main phases: parameter tuning and RF optimization. In general, the parameter regulation mostly detects optimal values of RF's parameters, like the maximum decision tree depth, the forest scale and the number of split features. Next, RF optimization is followed by means of GA to find an optimal combination of DTs in the optimized RF with the aim of maximizing the profit score by investigating actual and potential returns and losses [60, 61].

2.4. GA-MLP method

MLP is a feed-forward, supervised NN architecture. Back propagation (BP) training algorithm can be employed for reducing the error between the target value and network output. MLP structure and learning parameters are required to decide for enhancing the testing performance. As these parameters are usually selected randomly, detecting variables that produce the highest test accuracy is a time-consuming process. In GA-MLP method, network structure and learning parameter of PB algorithm are improved to achieve an efficient and faster weight-update process by employing GA [62-64].

3 Data acquisition and preparation

The data base includes porosity and HH-XRF experiments on core samples from Ekofisk Formation in the North Sea and outcrop chalk samples from Rørdal and Stevns Klint (ST). Plug samples are gathered from the Rørdal quarry near Aalborg, Denmark. The quarry characterizes a probable exposure for Tor organization hydrocarbon reservoirs in the North Sea [1, 65, 66]. Plugs are dried in oven at 60 °C for 40 h before conducting experiments. The grain volume is computed with Boyle's Law and a double-chambered Helium porosity meter. The bulk volume is determined by Archimedes principle in the presence of a submerging plug in a mercury bath. The pore volume is computed by measuring bulk and grain volumes [66]. XRF experiments are performed using a Niton^{TM}Xl3tGoldd+ HH-XRF device. The used HH-XRF is implemented by an Ag anode that measures at 6–50 kV and up to 200 µA, and provides semi-quantitative element doping [21]. HH-XRF is measured for a total of 43 elements, whilst 5 of these elements are considered in this study. Porosity values and measured HH-XRF for 5 elements of outcrop and North Sea chalk samples are listed in Tables 1 and 2, respectively [24]. In addition, a statistical summary of porosities and XRF elemental analysis data in entire, training and test datasets is given in Table 3.

Table 1. Porosity and measured HH-XRF for 5 elements of outcrop chalk samples from Rørdal and ST [24].

No.	**Location**	**Sample ID**	**φ (%)**	**Al (%)**	**Si (%)**	**Ca (%)**	**Fe (%)**	**K (%)**
1	Rørdal	23	43.1	0.27	2.44	44.97	0.15	0.10
2	Rørdal	33	44.1	0.17	3.06	45.10	0.11	0.06
3	Rørdal	45	47.5	0.26	1.29	46.55	0.11	0.05
4	Rørdal	127	43.6	0.05	1.21	47.07	0.09	0.07
5	Rørdal	186	44.1	0.31	3.01	44.86	0.13	0.11
6	Rørdal	187	40.4	0.62	3.91	43.02	0.25	0.22
7	Rørdal	192	28.8	0.88	4.68	42.01	0.25	0.26
8	Rørdal	194	47.0	0.22	1.64	47.12	0.11	0.07
9	Rørdal	201	45.1	0.21	3.38	44.96	0.13	0.07
10	Rørdal	244	47.1	0.05	2.69	46.19	0.09	0.07
11	Rørdal	246	47.6	0.30	3.58	44.68	0.10	0.07
12	Rørdal	261	47.3	0.16	1.44	46.81	0.08	0.05
13	Rørdal	289	45.0	0.34	3.18	44.65	0.13	0.12
14	Rørdal	405	31.3	0.77	5.23	38.08	0.38	0.66
15	Rørdal	466	41.2	0.53	2.49	45.26	0.23	0.16
16	Rørdal	469	48.2	0.30	2.44	46.50	0.07	0.06
17	ST	MTB20	47.3	0.05	0.56	48.33	0.04	0.07
18	ST	MTI2	47.1	0.14	0.61	48.90	0.04	0.03
19	ST	MT7	47.0	0.19	0.95	48.09	0.07	0.07
20	ST	MTB6	47.3	0.10	0.47	48.92	0.06	0.03
21	ST	MT9	47.6	0.10	0.60	48.63	0.05	0.04

22	ST	MT81	46.9	0.10	0.41	49.27	0.03	0.03
23	ST	MT10	46.0	0.20	0.81	48.19	0.07	0.05
24	ST	MT15	48.8	0.10	0.46	48.67	0.03	0.03
25	ST	MT49	47.8	0.10	0.46	49.18	0.03	0.05
26	ST	MT52	49.2	0.10	0.43	49.17	0.03	0.05
27	ST	MT64	48.1	0.10	0.38	49.36	0.03	0.05

Table 2. Porosity and measured HH-XRF for 5 elements of chalk samples from Ekofisk Formation [24].

No.	Formation	Sample ID	φ (%)	Al (%)	Si (%)	Ca (%)	Fe (%)	K (%)
1	Ekofisk	2	35.7	0.22	2.84	45.34	0.07	0.13
2	Ekofisk	3	35.6	0.14	2.80	45.49	0.07	0.05
3	Ekofisk	4	34.8	0.22	2.85	45.75	0.07	0.09
4	Ekofisk	5	39.5	0.05	2.19	46.58	0.07	0.05
5	Ekofisk	6	39.4	0.05	2.10	46.80	0.06	0.04
6	Ekofisk	7	39.3	0.05	2.07	46.65	0.06	0.05
7	Ekofisk	8	39.2	0.05	2.17	46.90	0.07	0.09
8	Ekofisk	9	36.3	0.19	3.16	45.27	0.07	0.07
9	Ekofisk	10	37.1	0.23	3.23	45.40	0.07	0.08
10	Ekofisk	11	37.2	0.12	3.48	45.22	0.07	0.05
11	Ekofisk	12	37.1	0.05	2.32	46.73	0.13	0.07
12	Ekofisk	13	37.3	0.05	2.17	46.88	0.07	0.04
13	Ekofisk	14	38.1	0.05	2.22	46.96	0.07	0.05
14	Ekofisk	15	38.1	0.05	2.40	46.78	0.06	0.04
15	Ekofisk	16	40.0	0.05	2.27	46.86	0.10	0.05
16	Ekofisk	17	39.1	0.05	2.41	46.55	0.10	0.05

17	Ekofisk	18	38.8	0.05	2.59	46.35	0.12	0.05
18	Ekofisk	19	40.7	0.05	2.22	46.94	0.10	0.05
19	Ekofisk	20	40.3	0.05	2.39	46.66	0.11	0.04
20	Ekofisk	21	40.5	0.05	2.51	46.69	0.11	0.05
21	Ekofisk	22	37.8	0.24	4.13	44.47	0.13	0.08
22	Ekofisk	23	37.7	0.27	4.13	44.61	0.14	0.06
23	Ekofisk	24	36.1	0.11	3.79	45.30	0.12	0.04
24	Ekofisk	25	33.6	0.22	6.20	42.46	0.13	0.05
25	Ekofisk	26	35.6	0.13	6.44	42.42	0.09	0.05
26	Ekofisk	27	35.1	0.24	6.45	42.06	0.10	0.05
27	Ekofisk	28	35.1	0.05	6.51	41.99	0.11	0.05

Table 3. Statistical characteristics of the dataset

	Variable (%)	**Mean**	**Minimum**	**Maximum**	**Standard deviation**	**Coefficient of variation**	**Skewness**
Entire data	Al	0.181	0.050	0.881	0.175	0.966	2.249
	Si	2.591	0.380	6.510	1.594	0.615	0.762
	Ca	46.013	38.078	49.359	2.211	0.048	-0.994
	Fe	0.099	0.029	0.379	0.062	0.624	2.296
	K	0.079	0.029	0.665	0.091	1.155	5.203
	Porosity	41.281	28.800	49.180	5.212	0.126	-0.136
training	Al	0.178	0.050	0.881	0.168	0.944	2.523
	Si	2.422	0.380	6.510	1.547	0.639	0.672
	Ca	46.230	41.994	49.359	1.994	0.043	-0.357

	Fe	0.091	0.029	0.252	0.051	0.557	1.585
	K	0.069	0.029	0.259	0.047	0.685	2.709
	Porosity	42.116	28.800	49.180	5.385	0.128	-0.452
	Al	0.188	0.050	0.767	0.194	1.031	1.825
	Si	2.928	0.461	6.455	1.676	0.572	0.895
test	Ca	45.578	38.078	49.185	2.600	0.057	-1.401
	Fe	0.115	0.030	0.379	0.079	0.688	2.279
	K	0.099	0.042	0.665	0.144	1.454	3.627
	Porosity	39.610	31.300	47.760	4.532	0.114	0.453

As it is clear from Table 3, Ca and porosity include the highest content of the utilized data. Ca, with mean value of 46.013%, varies in the range of 38.078 to 49.359 % with a standard deviation of 2.211% and coefficient of variation 0.048%. Porosity, with a mean value of 41.281%, varies in the range of 29.8 to 49.18% with a standard deviation of 5.212% and coefficient of variation 0.126%. K has the lowest portion of the dataset with a mean value of 0.079%, which varies in the range of 0.029 to 0.665% with a standard deviation of 0.091% and coefficient of variation of 1.155%.

4. Methodology

Development of the models are performed by employing Al, Si, Ca, Fe and K as independent variables for the prediction of porosity. Two robust models RF and hybrid GA-RF are developed and compared with MLP and hybrid GA-MLP models in terms of accuracy. Figs.1 and 2 present flowcharts of GA-MLP and GA-RF methods, respectively. Besides, due to the fact that there is not any direct way for splitting the entire data to training and testing sets, different proportions were implemented in previous studies, e.g. Choubin (2020) utilized 63% of data for training, whereas Qasem et al., (2019) and Kargar et al., (2020) implemented 67% of data, Dodangeh et al., (2019), Asadi et al., (2020), Shabani et al., (2020) and Samadianfard et al., (2020) exploited 70% of the entire data for developing the models. Thus, for the model development is the current research, data was split into training (67%) and testing (33%). Then, the accuracy is evaluated by the most frequently used performance parameters, namely CC, SI, WI and R^2. These parameters compare the target and output values and generate indexes for evaluating the model performance as well as its accuracy [28]. Table 4 presents the parameters related to RF and hybrid GA-RF models, which are generated in the developing phase. Parameters A, B, C, D, E, F and G are related to Random Forest. number_of_trees, Random Forest. maximal_depth, Random Forest. confidence, Random Forest. minimal_leaf_size, Random Forest. minimal_size_for_split, Random Forest. number_of_prepruning_alternatives, and Random Forest. subset_ratio, respectively. 10^4 evaluations of the objective function are implemented for GA. The number of chromosomes is set as 10^2 and the maximum number of iterations is set as 10^3.

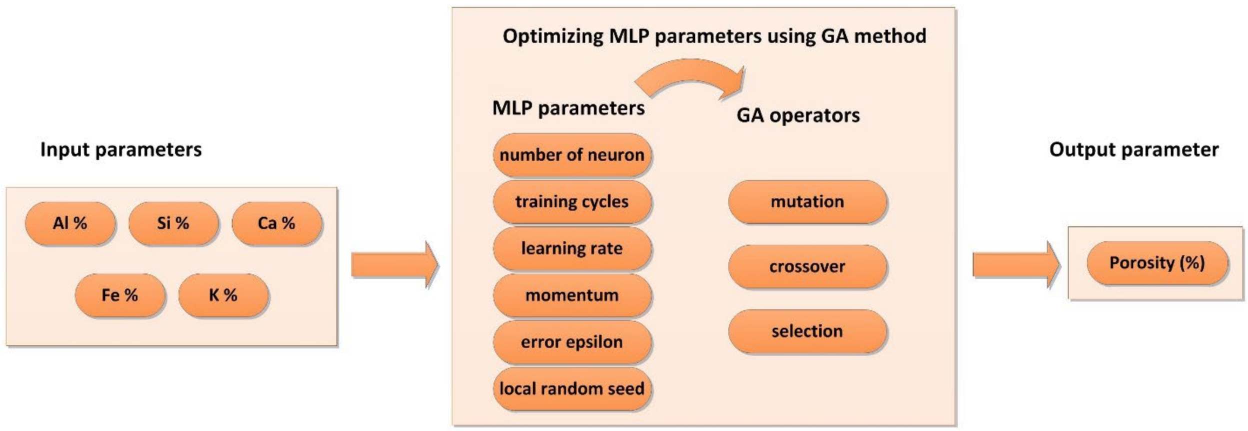

Fig.1. Flowchart of GA-MLP model.

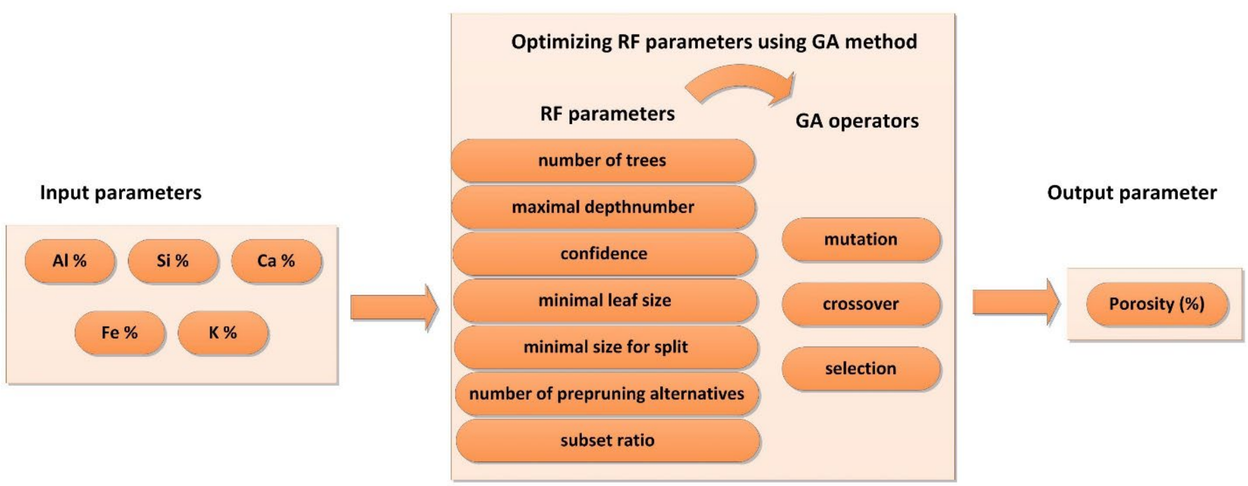

Fig.2. Flowchart of GA-RF model.

Table 4. Parameters of RF and GA-RF models.

Model	Parameter						
	A[1]	B[2]	C[3]	D[4]	E[5]	F[6]	G[7]
RF	100	10	0.100	2	4	3	0.200
GA-RF	81	36	0.219	1	80	55	0.303

[1] A: Random Forest.number_of_trees. [2] B: Random Forest.maximal_depth. [3] C: Random Forest.confidence. [4] D: Random Forest.minimal_leaf_size. [5] E: Random Forest.minimal_size_for_split. [6] F: Random Forest.number_of_prepruning_alternatives. [7] G: Random Forest.subset_ratio

Table 5 presents parameters related to MLP and MLP-GA models. Parameters A, B, C, D, E, F and G are related to Neural Net. training_cycles, Neural Net. learning_rate, Neural Net. Momentum, Neural Net. error_epsilon and Neural Net. local_random_seed

Table 5. Parameters of MLP and GA-MLP models.

Model	Parameter				
	A	B	C	D	E
MLP	200	0.0100	0.9000	0.0001	1992
GA-MLP-1	77	0.3992	0.5456	Infinity	77

[1] A: Neural Net.training_cycles. [2] B: Neural Net.learning_rate. [3] C: Neural Net.momentum. [4] D: Random Forest.minimal_leaf_size. [5] E: Neural Net.error_epsilon.

5 Results and Discussion

In order to inspect realizations of the suggested intelligent models and to make comparison between their accuracies, graphical error assessment and statistical analysis criteria are computed. Correlation coefficient (CC), Willmott's Index of agreement (WI), Scattered Index (SI), and coefficient of determination (R^2), which are normally applied in regression analysis, are considered in this research. The mathematical formulas of these statistical criteria are enclosed in Appendix A. A comprehensive assessment of the performances of the suggested models using CC, SI, WI and R^2 coefficients are collected in Table 6. Cross plots of the predicted chalk porosity by the intelligent models against the real data from the measurements are demonstrated in Fig. 3. More points close to the unit slope line shows lower deviations between the real data and model predictions in this type of plot. Fig. 3 shows that most of the data points estimated by RF and GA-RF intelligent methods are located close to the unit slope line, verifying their high degree of accuracy to predict chalk porosity.

Table 6. General results of computations for the studied models.

Model	**Statistical parameters**			
	CC	**SI**	**WI**	**R^2**
RF	0.99	0.02	0.997	0.98
MLP	0.90	0.05	0.943	0.82
GA-RF	0.99	0.02	0.995	0.99
GA-MLP	0.93	0.04	0.964	0.87

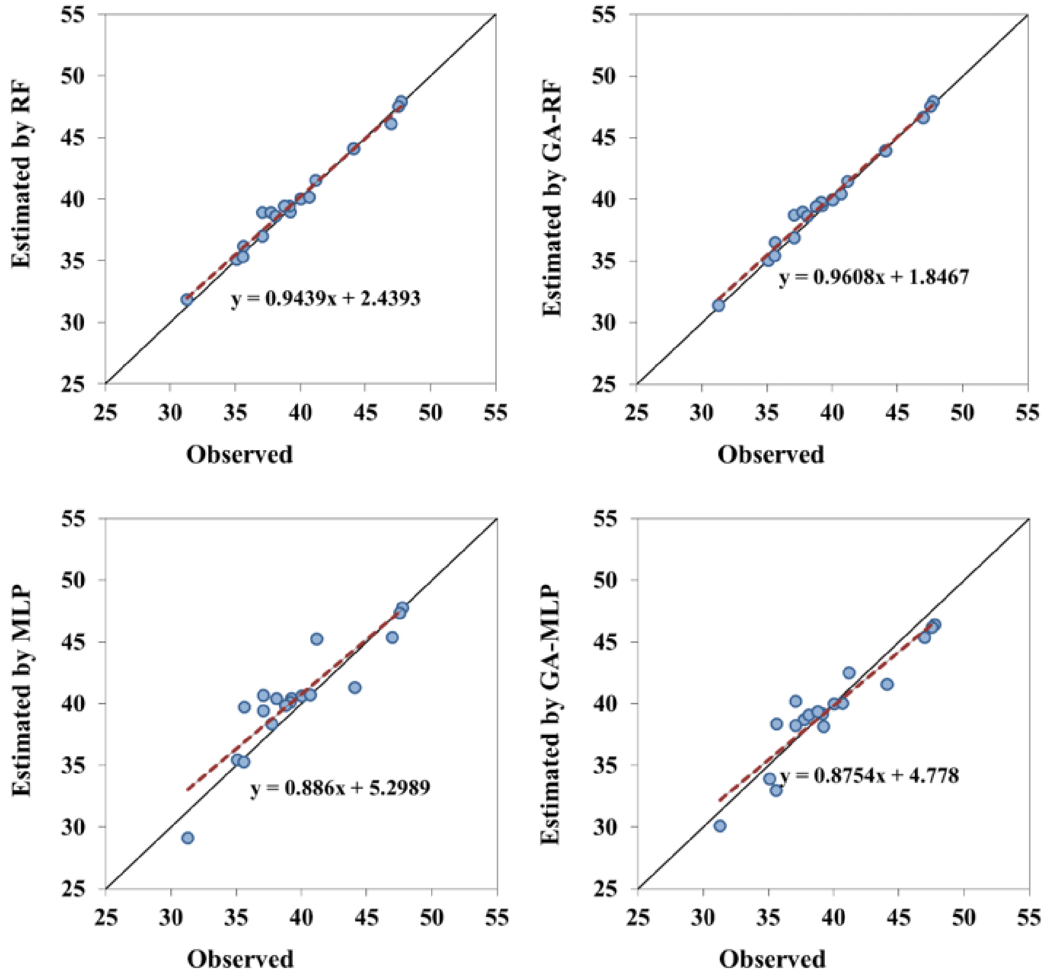

Fig.3. Scatter plots of target and estimated values.

In addition, the developed smart models is compared with Nourani et al. [24]. The previous research has established an empirical index, I index, based on multivariate descriptor relationships to link Ca, Si, Al, Fe and K elements in outcrop chalk samples to porosity as follows:

$$\mathrm{I}_I = \mathrm{Ca} + \mathrm{Si} - 20\,Al^{4.7} - 4.7Fe + 2.7K^{1.4} \quad (1)$$

where Ca, Si, Al, Fe and K are percentages of calcium, silicon, aluminum, iron and potassium elements in chalk samples, measured by HH-XRF. Moreover, Nourani et al. showed that a three-component Partial Least Square (PLS) model can predict chalk porosity with high satisfactory validation results. Fig. 4 shows bar plots of R^2 for six different models including the developed smart models in this study, and PLS and empirical models from the previous research. It should be noted that in order to ensure the fairness of the comparison, the considered R^2 value (0.95) for empirical

I index is only valid for outcrop chalk samples, whereas R^2 values for the rest of models involve both outcrop and reservoir chalk samples. Statistical criteria in Table 6 show high degree of performances for both GA-RF and RF methods. However, according to Table 6 and Fig. 4, GA-RF shows slightly higher R^2 coefficient than RF. Therefore, GA-RF is the most reliable model considering its lowest SI (0.02) and highest R^2 (0.99) values. Similarly, it can be concluded from Fig. 4 and Tables 6 that the discussed models for predicting chalk porosity follow the accuracy ranking shown below:

GA-RF > RF > I index > PLS > GA-MLP > MLP

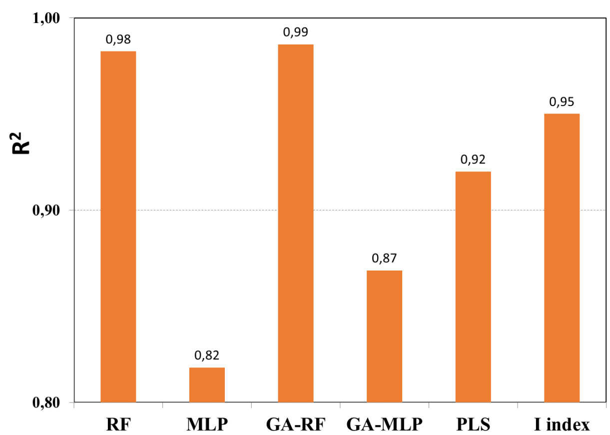

Fig. 4. Bar graphs of R^2 values.

For the purpose of getting a profound insight into the accuracy of model predictions, Figs 5 and 6 demonstrate a comparison between the experimental porosity data and the predicted chalk porosity for testing phase. A very good agreement between experimental and predicted data by GA-RF and RF methods is obtained in testing phase, as shown in Fig. 5. Besides, Fig. 6 shows fairly good predictions by GA-MLP and MLP methods.

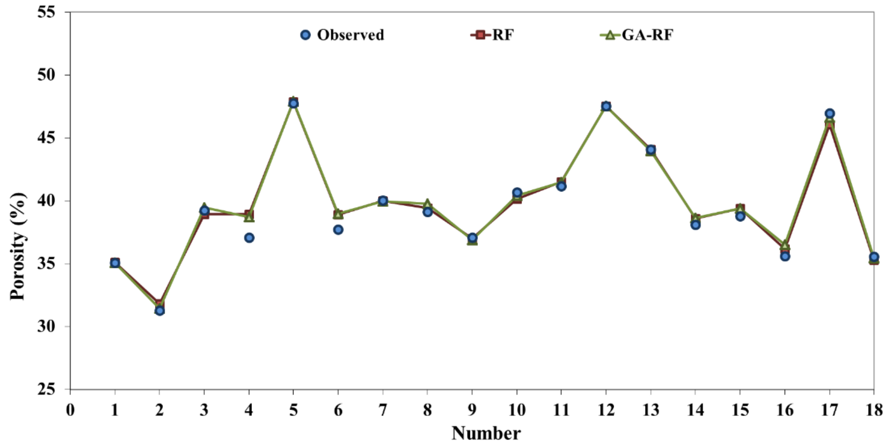

Fig. 5. Observed and estimated values of studied parameters with RF and GA-RF models.

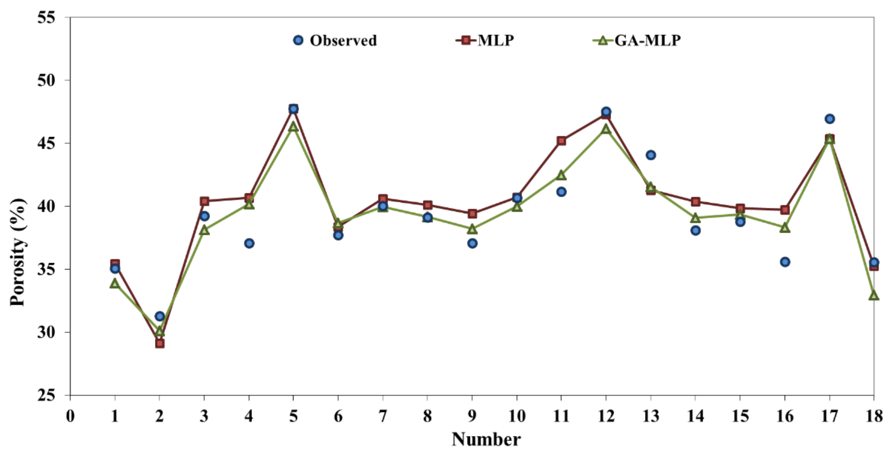

Fig. 6. Target and estimated values of studied parameters with MLP and GA-MLP models.

According to Fig. 6, it can be observed that the deviation of the developed method from observed data (as target values) is small except number 2, 4, 9, 11, 13 and 16. However, the difference between target and predicted values for Ga-MLP is lower than that of the single MLP. This can be due to the characteristics of GA-MLP, which employs GA for forming the weight and bias values of MLP and in fact plays as a training algorithm role [67] and sets the weight and bias values to reduce the training error. In fact, it considers the weight and bias values as a cost-function and optimizes the problem to reduce the cost-function. The main reason for reducing error values of GA-MLP in comparison with those of MLP model is the capability of GA in setting the weights and bias values in a proper way compared with the training algorithm of the single MLP model. RRelief-F algorithm [68] is applied to evaluate the weight of each element in porosity prediction as listed in Table 7.

Table 7. Weight of elements in porosity prediction by RReliefF algorithm.

Variable (%)	Weight (%)
Al	19
Si	33
Ca	23
Fe	17
K	8

Among five elements in Table 7, silicon plays the most substantial role in chalk porosity prediction. According to RReliefF algorithm analysis, calcium and aluminum are the second and third significant elements contributing in prediction of chalk porosity, respectively. Aluminum, calcium and silicon are the main elements present in clay, calcite and silica, respectively [24, 25]. Therefore, the quantities of these elements are proportional to the corresponding chemical compounds, such as clay, calcite and silica, which are present in chalk samples. Higher weights of Al, Ca and Si in predicting of chalk porosity are in accord with the facts that the matrix of chalk is composed mainly of calcium carbonate [69, 70], and the relatively low porosity of chalk is related to the contents of nano-quartz and clay minerals [11, 71, 72].

As it turns out, the proposed methods are able to successfully increase the accuracy of forecasting and estimation, and this leads to less network error and simulation for future applications and more accurate studies. Finding an accurate simulated model for a particular experiment reduces the cost and time of the experiment for the same experiment in the same system if the same experiment needs to be repeated. Simulation in this system also furnishes the strengths and weaknesses of the employed variables with focused and credible evidence. This is one of the reasons why researchers are always looking to produce more accurate, faster, and more reliable models in all competing scientific fields, and they are evolving over time.

6 Conclusions

Intelligent methods are suggested to provide accurate, robust and reliable models to predict the porosity of chalk samples by four ML methods, namely GA-RF, RF, GA-MLP and MLP, and using XRF elemental analysis data. Real porosity and XRF elemental analysis on outcrop chalk samples from Rørdal and Stevns Klint and core samples from Ekofisk Formation in the North Sea are used to figure out accuracy and effectiveness of the suggested predictive techniques. Results indicate that GA-RF is the most accurate model for predicting the chalk porosity in comparison with existing methods applied in this study. GA-RF demonstrates a high coefficient of determination (0.99) and very low SI value of 0.02. However, the application of ML-based methods, in addition to being successful in terms of accuracy and appropriateness of the problem, must be able to cope with challenges and disadvantages of using ML-based techniques. Challenges in applying the ML-based techniques include over-fitting and uncertainty in contact with data changes and unstructured data set. In addition, there are other challenges such as the need for ML-based techniques to have a complete data set and the need for sufficient time to complete the process. Each of these challenges needs to be addressed, which can be considered as a challenge in future studies.

Appendix A. Statistical formulas

I: Correlation coefficient (CC), expressed as:

$$CC = \frac{\left(\sum_{i=1}^{n} O_i P_i - \frac{1}{n}\sum_{i=1}^{n} O_i \sum_{i=1}^{n} P_i\right)}{\left(\sum_{i=1}^{n} {O_i}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} O_i\right)^2\right)\left(\sum_{i=1}^{n} {P_i}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} P_i\right)^2\right)}$$

II: Scattered Index (SI) follows as:

$$SI = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2}}{\bar{O}}$$

III: Willmott's Index of agreement (WI) expressed as:

$$WI = 1 - \left[\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}\left(\left|P_i - \bar{O}_i\right| + \left|O_i - \bar{O}_i\right|\right)^2}\right]$$

where O_i and P_i are the observed and predicted i^{th} value.

IV: Coefficient of determination (R^2), expressed as:

$$R^2 = \left[\frac{\left(\sum_{i=1}^{n} O_i P_i - \frac{1}{n}\sum_{i=1}^{n} O_i \sum_{i=1}^{n} P_i\right)}{\left(\sum_{i=1}^{n} {O_i}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} O_i\right)^2\right)\left(\sum_{i=1}^{n} {P_i}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} P_i\right)^2\right)}\right]^2$$

References

1. Meyer, A.G., et al., *Modifications of chalk microporosity geometry during burial–an application of mathematical morphology.* Marine and Petroleum Geology, 2019. **100**: p. 212-224.
2. Borre, M. and I.L. FABRICIUS, *Chemical and mechanical processes during burial diagenesis of chalk: an interpretation based on specific surface data of deep-sea sediments.* Sedimentology, 1998. **45**(4): p. 755-769.
3. Buls, T., et al., *Production of calcareous nannofossil ooze for sedimentological experiments.* Journal of Sedimentary Research, 2015. **85**(10): p. 1228-1237.
4. Häkansson, E., R. Bromley, and K. Perch-Nielsen, *Maastrichtian chalk of North-West Europe-pelagic shelf sediments.* Pelagic Sediments: on Land and under the Sea, Special Publication, 1974. **1**: p. 211-233.
5. Meyer, A.G., M. Nourani, and L. Stemmerik, *Description of chalk microporosity via automated mathematical morphology on scanning electron microphotographs.* Petroleum Geoscience, 2019: p. petgeo2019-018.
6. Andersen, M.A., *Petroleum research in North Sea chalk.* Rogaland Research, 1995.
7. Guo, B., K. Sun, and A. Ghalambor, *well productivity Handbook*. 2014: Elsevier.
8. Taud, H., et al., *Porosity estimation method by X-ray computed tomography.* Journal of petroleum science and engineering, 2005. **47**(3-4): p. 209-217.
9. D'Heur, M., *Porosity and hydrocarbon distribution in the North Sea chalk reservoirs.* Marine and Petroleum Geology, 1984. **1**(3): p. 211-238.
10. Wheaton, R., *Fundamentals of applied reservoir engineering: appraisal, economics and optimization*. 2016: Gulf Professional Publishing.
11. Fabricius, I.L., *Chalk: composition, diagenesis and physical properties.* Bulletin of the Geological Society of Denmark, 2007. **55**: p. 97-128.
12. Vejbæk, O., et al., *Cretaceous.* Petroleum Geological Atlas of the Southern Permian Basin Area. European Association of Geoscientists and Engineers (EAGE), Houten, The Netherlands, 2010. **195**: p. 209.
13. Fabricius, I.L., B. Røgen, and L. Gommesen, *How depositional texture and diagenesis control petrophysical and elastic properties of samples from five North Sea chalk fields.* Petroleum Geoscience, 2007. **13**(1): p. 81-95.
14. Gautier, D.L., *Kimmeridgian shales total petroleum system of the North Sea graben province*. 2005, US Geological Survey.
15. Surlyk, F., et al., *Upper cretaceous.* The Millennium Atlas: Petroleum Geology of the Central and Northern North Sea. Geological Society, London, 2003. **213**: p. 233.
16. Glover, P., *Formation evaluation MSc Course notes (Porosity)*. 2016.
17. Maas, J. and N. Springer, *JCR 7-Advanced Core Measurements "Best Practices" for Low Reservoir Quality Chalk*. 2014.
18. Xiao, L., et al., *Calculation of porosity from nuclear magnetic resonance and conventional logs in gas-bearing reservoirs.* Acta Geophysica, 2012. **60**(4): p. 1030-1042.
19. Miah, M.I., *Porosity assessment of gas reservoir using wireline log data: a case study of bokabil formation, Bangladesh.* Procedia Engineering, 2014. **90**: p. 663-668.
20. McPhee, C., J. Reed, and I. Zubizarreta, *Core analysis: a best practice guide*. 2015: Elsevier.
21. Schovsbo, N.H., et al., *Stratigraphy and geochemical composition of the Cambrian Alum Shale Formation in the Porsgrunn core, Skien-Langesund district, southern Norway.* Bulletin of the Geological Society of Denmark, 2018. **66**(1).
22. Dahl, T.W., et al., *Tracing euxinia by molybdenum concentrations in sediments using handheld X-ray fluorescence spectroscopy (HHXRF).* Chemical geology, 2013. **360**: p. 241-251.
23. Hammer, Ø. and H.H. Svensen, *Biostratigraphy and carbon and nitrogen geochemistry of the SPICE event in Cambrian low-grade metamorphic black shale, Southern Norway.* Palaeogeography, Palaeoclimatology, Palaeoecology, 2017. **468**: p. 216-227.
24. Nourani M., S.N., Meyer A. G., Sigalas L., Lorentzen H. J., Olsen D., Stemmerik L., *An index for predicting porosity in chalk by XRF*, in *DHRTC*. 2019: Copenhagen. p. 1-2.
25. Nourani, M., et al., *A predictive model for the wettability of chalk.* SN Applied Sciences, 2020. **2**(10): p. 1-12.
26. Chakraborty, S., et al., *Semiquantitative Evaluation of Secondary Carbonates via Portable X-ray Fluorescence Spectrometry.* 2017. **81**(4): p. 844-852.
27. Ardabili, S., et al. *Deep learning and machine learning in hydrological processes climate change and earth systems a systematic review*. in *International Conference on Global Research and Education*. 2019. Springer.

28. Ardabili, S., A. Mosavi, and A.R. Várkonyi-Kóczy. *Systematic review of deep learning and machine learning models in biofuels research*. in *International Conference on Global Research and Education*. 2019. Springer.
29. Rostami, A., et al., *Applying SVM framework for modeling of CO2 solubility in oil during CO2 flooding.* 2018. **214**: p. 73-87.
30. Saghafi, H., M.J.J.o.P.S. Arabloo, and Engineering, *Development of genetic programming (GP) models for gas condensate compressibility factor determination below dew point pressure.* 2018. **171**: p. 890-904.
31. Okwu, M.O., A.N.J.J.o.P.E. Nwachukwu, and P. Technology, *A review of fuzzy logic applications in petroleum exploration, production and distribution operations.* 2019. **9**(2): p. 1555-1568.
32. Sircar, A., et al., *Application of machine learning and artificial intelligence in oil and gas industry.* 2021.
33. Alnahwi, A. and R.G.J.A.B. Loucks, *Mineralogical composition and total organic carbon quantification using x-ray fluorescence data from the Upper Cretaceous Eagle Ford Group in southern Texas.* 2019. **103**(12): p. 2891-2907.
34. Zhao, B., et al., *A Hybrid Approach for the Prediction of Relative Permeability Using Machine Learning of Experimental and Numerical Proxy SCAL Data.* 2020.
35. Andrianov, N. and H.M.J.A.i.W.R. Nick, *Machine learning of dual porosity model closures from discrete fracture simulations.* 2021. **147**: p. 103810.
36. Lawal, A.I. and M.A. Idris, *An artificial neural network-based mathematical model for the prediction of blast-induced ground vibrations.* International Journal of Environmental Studies, 2020. **77**(2): p. 318-334.
37. Agatonovic-Kustrin, S. and R. Beresford, *Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research.* Journal of pharmaceutical and biomedical analysis, 2000. **22**(5): p. 717-727.
38. Nosratabadi, S., et al. *State of the art survey of deep learning and machine learning models for smart cities and urban sustainability*. in *International Conference on Global Research and Education*. 2019. Springer.
39. Fong, R.C., W.J. Scheirer, and D.D. Cox, *Using human brain activity to guide machine learning.* Scientific reports, 2018. **8**(1): p. 1-10.
40. Hassabis, D., et al., *Neuroscience-inspired artificial intelligence.* Neuron, 2017. **95**(2): p. 245-258.
41. Azam, F., *Biologically inspired modular neural networks*. 2000, Virginia Tech.
42. Babatunde, O.H., et al., *A genetic algorithm-based feature selection.* 2014.
43. PAUZI, H.M. and L. ABDULLAH, *Airborne particulate matter research: a review of forecasting methods.* Journal of Sustainability Science and Management, 2019. **14**(4): p. 189-227.
44. Kale, A. and S. Sonavane. *Optimal feature subset selection for fuzzy extreme learning machine using genetic algorithm with multilevel parameter optimization*. in *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. 2017. IEEE.
45. Das, H., et al., *A novel PSO based back propagation learning-MLP (PSO-BP-MLP) for classification*, in *Computational Intelligence in Data Mining-Volume 2*. 2015, Springer. p. 461-471.
46. Qi, Y., *Random forest for bioinformatics*, in *Ensemble machine learning*. 2012, Springer. p. 307-323.
47. Boulesteix, A.L., et al., *Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2012. **2**(6): p. 493-507.
48. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
49. Deng, H., *Guided random forest in the RRF package.* arXiv preprint arXiv:1306.0237, 2013.
50. Strobl, C. and A. Zeileis, *Danger: High power!–exploring the statistical properties of a test for random forest variable importance.* 2008.
51. Segal, M.R., *Machine learning benchmarks and random forest regression.* 2004.
52. Parlos, A.G., K.T. Chong, and A.F. Atiya, *Application of the recurrent multilayer perceptron in modeling complex process dynamics.* IEEE Transactions on Neural Networks, 1994. **5**(2): p. 255-266.
53. Pandey, P. and S. Barai, *Multilayer perceptron in damage detection of bridge structures.* Computers & structures, 1995. **54**(4): p. 597-608.
54. Meiabadi, M.S., et al., *Modeling the Producibility of 3D Printing in Polylactic Acid Using Artificial Neural Networks and Fused Filament Fabrication.* 2021. **13**(19): p. 3219.
55. Gardner, M.W. and S. Dorling, *Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences.* Atmospheric environment, 1998. **32**(14-15): p. 2627-2636.
56. Ruck, D.W., et al., *The multilayer perceptron as an approximation to a Bayes optimal discriminant function.* IEEE Transactions on Neural Networks, 1990. **1**(4): p. 296-298.
57. Gibson, G.J., S. Siu, and C. Cowen. *Multilayer perceptron structures applied to adaptive equalisers for data communications*. in *International Conference on Acoustics, Speech, and Signal Processing*. 1989. IEEE.

58. Belue, L.M. and K.W. Bauer Jr, *Determining input features for multilayer perceptrons.* Neurocomputing, 1995. **7**(2): p. 111-121.
59. Bello, M.G., *Enhanced training algorithms, and integrated training/architecture selection for multilayer perceptron networks.* IEEE Transactions on Neural networks, 1992. **3**(6): p. 864-875.
60. Ye, X., L.-a. Dong, and D. Ma, *Loan evaluation in P2P lending based on random forest optimized by genetic algorithm with profit score.* Electronic Commerce Research and Applications, 2018. **32**: p. 23-36.
61. Naghibi, S.A., K. Ahmadi, and A. Daneshi, *Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping.* Water Resources Management, 2017. **31**(9): p. 2761-2775.
62. Taşkıran, M., Z.G. Çam, and N. Kahraman. *An efficient metho to optimize multi-layer perceptron for classification of human activities*. in *2nd International Conference on Computer, Control and Communication Technologies (CCCT'15)*. 2015.
63. Juang, C.-F., *A hybrid of genetic algorithm and particle swarm optimization for recurrent network design.* IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2004. **34**(2): p. 997-1006.
64. Boeringer, D.W. and D.H. Werner, *Particle swarm optimization versus genetic algorithms for phased array synthesis.* IEEE Transactions on antennas and propagation, 2004. **52**(3): p. 771-779.
65. Surlyk, F., et al., *The cyclic Rørdal Member–a new lithostratigraphic unit of chronostratigraphic and palaeoclimatic importance in the upper Maastrichtian of Denmark.* Bulletin of the Geological Society of Denmark, 2010. **58**: p. 89-98.
66. Nourani, M., et al., *Determination of the overburden permeability of North Sea Chalk.* rock mechanics and rock engineering, 2019. **52**(6): p. 2003-2010.
67. Ecer, F., et al., *Training Multilayer Perceptron with Genetic Algorithms and Particle Swarm Optimization for Modeling Stock Price Index Prediction.* Entropy, 2020. **22**(11): p. 1239.
68. Urbanowicz, R.J., et al., *Relief-based feature selection: Introduction and review.* Journal of biomedical informatics, 2018. **85**: p. 189-203.
69. Selley, R., *SEDIMENTARY ROCKS| Mineralogy and Classification.* 2005.
70. Price, M., *Fluid flow in the Chalk of England.* Geological Society, London, Special Publications, 1987. **34**(1): p. 141-156.
71. Lindgreen, H., et al., *The tight Danian Ekofisk chalk reservoir formation in the south Arne field, North Sea: mineralogy and porosity properties.* Journal of Petroleum Geology, 2012. **35**(3): p. 291-309.
72. Lind, I. and P. Grøn, *Porosity variation in chalk.* Zbl. Geol. Paläont. Teil I, 1996. **1994**(11/12): p. 1447-1457.