Oil Family Typing Using a Hybrid Model of Self-Organizing Map and Artificial Neural Network

Majid Safaei-Farouji and Amir Mousavi

December 22, 2021

Oil Family Typing Using a Hybrid Model of Self-Organizing Map and Artificial Neural Network

Majid Safaei-Farouji [1], Amir Mousavi [2]

[1] School of Geology, College of Science, University of Tehran, Tehran, Iran; majid.safaei@ut.ac.ir

[2] Oxford Brookes University, Oxford, UK

Abstract

Identifying the number of oil families in petroleum basins provides practical and valuable information in petroleum geochemistry studies from exploration to development. Oil family grouping helps us track migration pathways, identify the number of active source rock(s), and examine the reservoir continuity. To date, almost in all oil family typing studies, common statistical methods such as principal component analysis (PCA) and hierarchical clustering analysis (HCA) have been used. However, there is no publication regarding using artificial neural networks (ANNs) for examining the oil families in petroleum basins. Hence, oil family typing requires novel, not overused and common techniques. This paper is the first report of oil family typing using ANNs as robust computational methods. To this end, a self-organization map (SOM) neural network associated with three clustering validity indices were employed on oil samples belonging to the Iranian part of the Persian Gulf' oilfields. For the SOM network, at first, ten default clusters were selected. Afterwards, three effective clustering validity coefficients, namely Calinski-Harabasz (CH), Silhouette indexes (SI) and Davies-Bouldin (DB), were operated to find the optimum number of clusters. Accordingly, among ten default clusters, the maximum CH (62) and SI (0.58) were acquired for four clusters. Likewise, the lowest DB (0.8) was obtained for four clusters. Thus, all three validation coefficients introduced four clusters as the optimum number of clusters or oil families. The number of oil families identified in the present report is consistent with those previously reported by other researchers in the same study area. However, the techniques used in the present paper, which have not been implemented so far, can be introduced as more straightforward for clustering purposes in the oil family typing than those of common and overused methods of PCA and HCA.

Keywords: Oil family typing, self-organization map, clustering validity indexes

1. Introduction

Identifying the relationship between oil samples and grouping them, known as oil family classification, as a part of petroleum system studies, plays a paramount role in various aspects of the oil industry, including exploration, development, etc. The primary outcomes of oil family typing are detecting migration pathways and evaluating the continuity between different oil reservoirs[1].

It is for a long time that geochemists use the statistical techniques PCA and HCA to group oil families in petroleum basins[2,3] . However, it is an undeniable fact that artificial intelligence (AI) and machine learning (ML) systems are developing on a regular basis and provide various applications for scientists[4–8], and petroleum geochemists are no exception. AI and ML techniques in petroleum-related studies have been widely used in recent years. Amar et al[9] used Ml approaches to model oil-brine interfacial tension at high pressure and high salinity conditions. Mazloom et al[10] used AI algorithms to estimate asphalten adsorbtion in nonocomposites. Rostami et al[11] utilized ANNs for predicting the natural gas viscosity. Mokarizadeh et al[5] implemented ANNs and ML algorithms to determine the solubility of SO_2 in ionic liquids. Hemmati-Sarapardeh et al[12] conducted the modeling natural gas compressibility using a kind of ANN. Amooie et al[13] took advantage of ML methods for geological carbon storage studies. Menad et al[4] estimated the solubility of CO_2 in brine via advaned ML techniques. Razghandi et al[14] predicted under-saturated crude oil viscosity by ML algorithms. Bolandi et al[15] evaluated source rock characteristics by ML methods. Bolandi et al[16] studeied the organic facies of source rocks by combining ML and ANNs. Tabatabaei et al[17] utilized ML algorithm for estimation of total organic carbon (TOC) from well log data. Naghizadeh et al[18] estimated viscosity of CO_2–N_2 gaseous mixtures by smart ML models.

Kadkhodaie-Ilkhchi et al[19] integrated endividual smart ML models with a committee machine intelligent system to approximate TOC from petrophisical well logs. Ghiasi-Freez et al[20] used committee machines to predict permeability from petrographic image analysis. Tohidi-Hosseini et al[6] predicted solution gas-oil Ratio via a robust ML system. Esfahani et al[21] implemented ML paradigms for determination of natural gas density. Hajirezaie et al[22] employed a powerful ML algorithm to estimate under-saturated reservoir oil viscosity. Karkevandi-Talkhooncheh et al[23] used the adaptive neuro fuzzy interface system optimized with evolutionary algorithms for modeling CO_2-crude oil minimum miscibility pressure. Barati-Harooni et al[24] employed different ML and AI frameworks to predict minimum miscibility pressure (MMP) in enhanced oil recovery (EOR) process by N_2 flooding. Amiri-Ramsheh et al[25] conducted an study about modeling of wax disappearence temperature (WDT) using different AI and ML methods. Mohammadi et al[26] employed a powerful ML technique to model hydrogen solubility in hydrocarbons. Moosanezhad-Kermani et al[27] employed a kind of ANN for modelling of carbon dioxide solubility in ionic liquids. Rezaei et al[28] implemented a radial basis function neural network with evolutionary algorithms for modelling of gas viscosity at high pressure and high temperature conditions. Khamehchi et al[29] utilized divers ML and AI systems to model viscosity of light and intermediate dead oil systems. In addition to the mentioned studies, recently researchers used AI and ML for organic geochemistry purposes. For example, Safaei-Farouji and Kadkhodaie[30] used intelligent AI and ML methods for estimation of kerogen type from petrophisical well logs. Collectively, even though AI and ML methods have been used in various petroleum-related firlds, oil family typing using an artificial neural network is missing. ANNs have various applications that one of which is clustering[31–33]. Therefore, oil family grouping as a kind of clustering problem can be solved via ANNs.

The SOM function as an artificial neural network proposed by [34] maps multidimensional data to a two-dimension space. This space is created with the help of a competitive and unsupervised learning process. SOM neural network preserves the topological properties of the input space by utilizing a neighborhood function. Actually, the resulting map illustrates the relationship between input patterns. [35,36].

The primary use of SOM is clustering and other types of unsupervised classifications [35,36]. So far, for oil family grouping, limited common statistical methods, such as PCA and HCA, have been used, but using artificial neural networks is entirely missing. Rabbani et al[2] geochemically analyzed thirty-three oil samples from several oil fields in the Persian Gulf's Iranian sector. They defined four main oil families through statistical methods of PCA and HCA. Mashhadi and Rabbani[37] also geochemically investigated twenty oil samples from oil fields in the Iranian part of the Persian Gulf. They identified two distinct genetic oil families using PCA analysis. In another study, Hosseini et al[3] based on the study of fourteen oil samples from the eastern Iranian sector of the Persian Gulf and implementing HCA, identified two different oil families.

Petroleum geochemistry studies of the examined area have been conducted by previous researches [2,3,37]; correspondingly, in the present paper, we focus on using a SOM neural network as a novel paradigm to determine oil families in the region. Indeed, the present study enables us to relate our outcomes to previously published works in the study area while using more database and introducing a new method for oil family typing.

In the following introduction, the method used and recent works are generally explained. The second part of the paper is devoted to the data preparation and methodology. Then, the obtained results are discussed in the third section. Ultimately, the final part of the study provides a summary of the findings.

2. Materials and Methods

Collectively, 60 oil samples were collected from the literature [2,3,37]. These samples belong to different oilfields in the Iranian part of the Persian Gulf. This Gulf and its coastal regions are home to about two-thirds of the world's proven oil reserves (715 billion barrels) [38]. The examined oilfields include Dorood, Kharg, Aboozar, Foroozan, Salman, Resalat, Reshadat, Balal, Bahregansar, Souroush, Nowrouz, Sirri A, Sirri C, Sirri D, and Sirri E. The location map of the studied oil field is given in figure 1. Also, the detailed geochemical and biomarker analysis of the studied crude oil samples can be found in Hosseiny et al[3], Mashhadi and Rabbani[37], and Rabbani et al[2]. Table. 1 summarizes the 16 geochemical and biomarker parameters used as inputs for the SOM network.

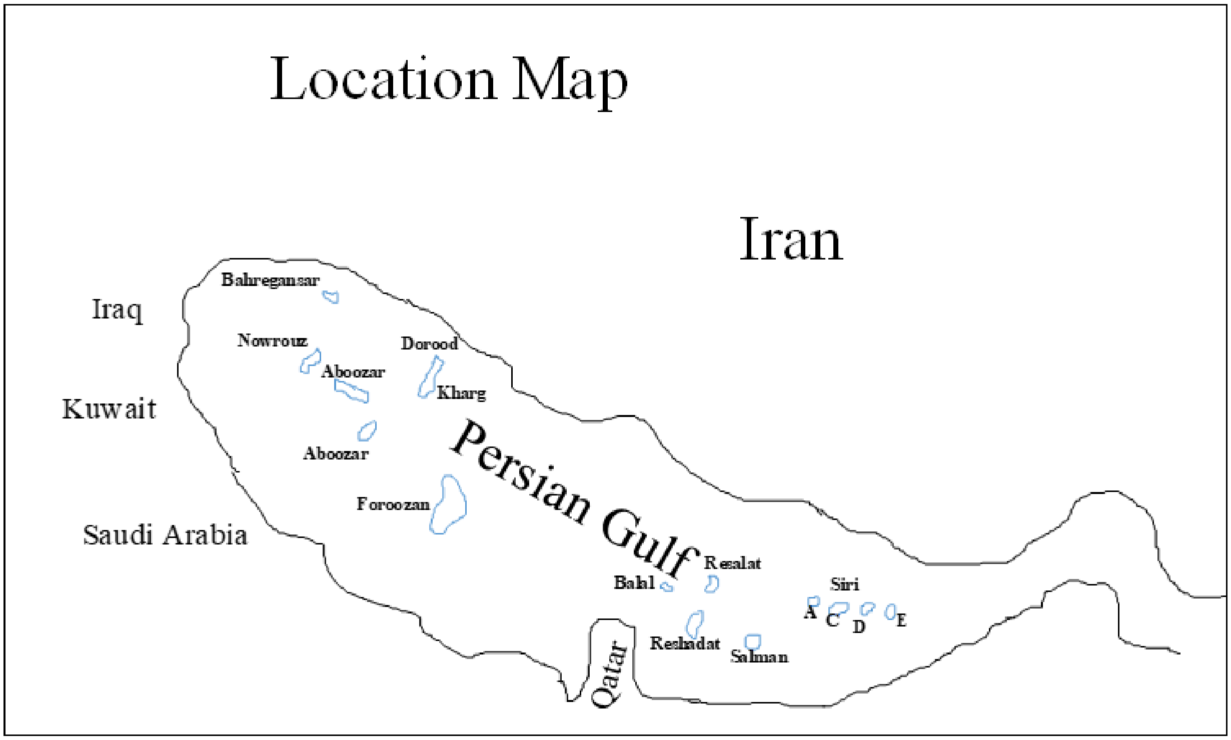


Figure 1. The geographical map of the studied oil fields.

2.1. Principal component analysis

The first stage in this study was using PCA to decrease data dimensions. Since sixteen different geochemical and biomarker parameters were implemented as inputs, it was mandatory to diminish dimensions to illustrate data and provide graph results[39,40]. Accordingly, the data dimensions or components were decreased from sixteen to three using PCA.

2.2.Creating the self-organizing map (SOM) network

Artificial neural networks mimic the learning process in the human brain. A key component in processing a neural network is the neurons that receive the inputs and generate the outputs using nonlinear operations. The SOM artificial neural network can learn complex and high-dimension data and extract a visible cluster set [34]. The process of SOM network training consists of two repetitive phases. The first phase selects the best mapping unit (neural network neurons) to adapt to input data. The second phase is to update the mapping to provide the best representation and display input data [41].

The process of selecting the best unit to conform to the input data (best adaptive unit or BMO) is based on the minimum distance (usually the Euclidean distance). Then in the update phase, each BMU and its neighboring units (within a given radius) move closer to the input data and fully comply with it. This neighborhood radius decreases with each phase selected and updated, eventually leading to a final (two-dimensional) mapping [42].

The SOM network is composed of an input layer of nodes, and an output layer of neurons, in which the grouping of the inputs is formed [43]. The output layer is called the competitive layer because the competitive role of the network during the training process takes place at this layer. A competitive layer is a two-dimensional plane structured with m neurons while accommodating an input of n neurons. Each input layer neuron with different weight values is connected to the competing layer neurons, and also, a series of minor connections are made between the competing layer neurons [44]. The number of neurons may vary from a few tens to a few thousand. Each neuron is assigned a dimensional vector d with weight m, of which d is the same dimension as the input vectors. Neurons are connected to their neighboring neurons by a neighborhood relationship that affects the topology or structure of the map. Common topologies are square, hexagonal, triangular or irregular grids [45].

As depicted in Figure 2, the SOM neural network consists of a set of M=m×m processing neurons. Suppose these M neurons are organized on a grid in a plane. In that case, the obtained network is two-dimensional because this network projects multi-dimensional input vectors onto a two-dimensional surface; for a given network, the input vector x is composed of a fixed dimension n. In the array, the n components of the input vector x (i.e., x_1, x_2, . . ., x_n) are connected to each neuron. For a connection from the ith component of the input vector to the jth neuron, a synaptic weight wij is assigned. Thus, an n-dimensional vector wj of synaptic weights is related to each neuron j[46].

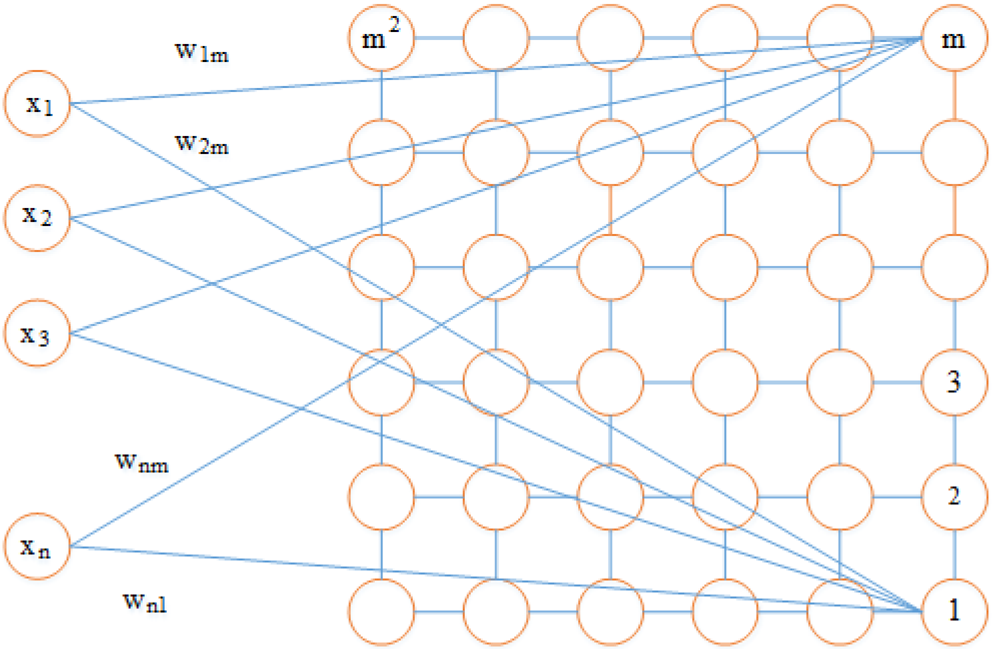


Figure 2. The main structure of a SOM neural network.

In brief, the process of the SOM network is as following[46]:

1. Calculate the distance between the pattern (X) and all neural neurons[46]

$$d_{ij} = \| x_k - w_{ij} \| \qquad (1)$$

2. Select the nearest neuron as the winning neuron[46]

$$w_{ij:}\ d_{ij} = \min(d_{mn}) \qquad (2)$$

3. Update each neuron according to the neighbourhood function[46].

$$w_i = w_{ij} + \alpha_h\, (w_{winner}, w_{ij})\ \|x_k - w_{ij}\| \qquad (3)$$

The value of coefficient a reduces the effect of different weights[46].

This process is repeated until a specific stopping criterion is reached. Often the criterion for stopping is a certain number of repetitions. To stabilize the convergence and stability of the map, the learning rate and neighbourhood radius are reduced in each iteration. Therefore, convergence will tend to zero. The measuring distance between the vectors is the Euclidean distance [46].

2.3.Clustering Validity Indices

The clustering validity indexes commonly are used associated with a clustering algorithm. According to the selected index, to determine the exact number of clusters, either minimum or maximum index value aids to figure out the optimum number of clusters (k) [47].

Generally, validity indexes can be grouped into internal and external. Internal indexes employ the information related to the data itself, whilst external indices, such as labels, are implemented by external information. Internal measures can improve clustering algorithms. By contrast, external measures can be used merely for validation. Internal indices are generally employed to determine k value [48–51].

In this paper, for the SOM neural network, three efficient internal coefficients, including DB, CH, and SH, were implemented to determine the optimum number of clusters for oil samples. Initially, a number of 10 classes were selected for the SOM network. The model was developed based on these clusters; then, the optimum number of classes as the optimum number of oil families was recognized using the coefficients.

2.3.1. David-Bouldin (DB) Index:

This index aims to minimize the average distance between each cluster and the most similar one. The minimum value for the DB index indicates the optimum number of clusters or oil families[52].

This index is described as[52]:

$$DB = \frac{1}{k}\sum_{i=1}^{k} max_{j\neq i}\{\, D_{i,j}\} \qquad (4)$$

In which $D_{i,j}$ shows the within-to-between cluster distance ratio for the *i*th and *j*th clusters. $D_{i,j}$ can be defined as[52]:

$$D_{i,j} = \frac{\bar{d}_i+\bar{d}_j}{d_{i,j}} \qquad (5)$$

Where di represents the mean distance between each point in the ith cluster and the cluster's centroid, $d_{i,j}$ denotes the Euclidean distance between the centroids of the *i*th and *j*th clusters. The optimum clustering solution possesses the lowest DB index value [52].

2.3.2. Calinski–Harabasz (CH) Index

CH index [53] demonstrates the quality of clustering solution based on the average sum of squares between and within a cluster. It can be measured as[47]:

$$CH = \frac{SSB}{SSW} \times \frac{(n-k)}{(k-1)} \qquad (6)$$

In which SSB shows the average between-cluster sum of squares. SSW indicates the average within-cluster sum of squares, *k* represents the number of clusters, and *n* denotes the number of observations. The average *SSB* is calculated as bellows[47]:

$$SSB = \sum_{i=1}^{k} n_i \; \|m_i - \mu\|^2 \qquad (7)$$

Where m_i is the centroid of cluster I, μ shows the mean of all data points, and $\|m_i - \mu\|$ typifies the Euclidean distance between the centroid of the cluster and the mean of all data points. The formulation of mean SSW is computed as bellows[47]:

$$SSw = \sum_{i=1}^{n} \sum_{x \epsilon p_i} \|x - m_i\|^2 \qquad (8)$$

In which k indicates the number of clusters, x is a sample, p_i demonstrates the *ith* cluster, m_i shows the centroid of the cluster p_i, and $\|x - m_i\|$ is Euclidean distance between sample and centroid of the cluster[47].

A higher CH quantity epitomizes a better data clustering outcome or the optimum number of questionable clusters. Therefore, high SSB and low SSW numbers give a well-separated cluster [47].

2.3.3. Silhouette Index (SH)

SH index [54] demonstrates how close every data point is to other data points within a cluster and how well clusters are detached from each other. Simply put, it operates based on the distance between each point between and within clusters. The highest silhouette quantity indicates the optimum number of clusters (k) [55].

$$sp(i) = \frac{b(i) - a(i)}{Max\{a(i), b(i)\}} \qquad (9)$$

In which $sp(i)$ is named silhouette width of point. *a* (*i*) shows the mean distance between the *ith* point and all the points in the clusters *Pi*, (*i* = 1, 2, . . ., *n*). *b* (*i*) displays the most minor of these distances. Hence, it can be observed that the silhouette value will be between 1 and -1. For every clustering, the average index of all *sp* (*i*) is employed [47]. The detailed feature of the SOM network used for clustering in the present study is given in Table. 2.

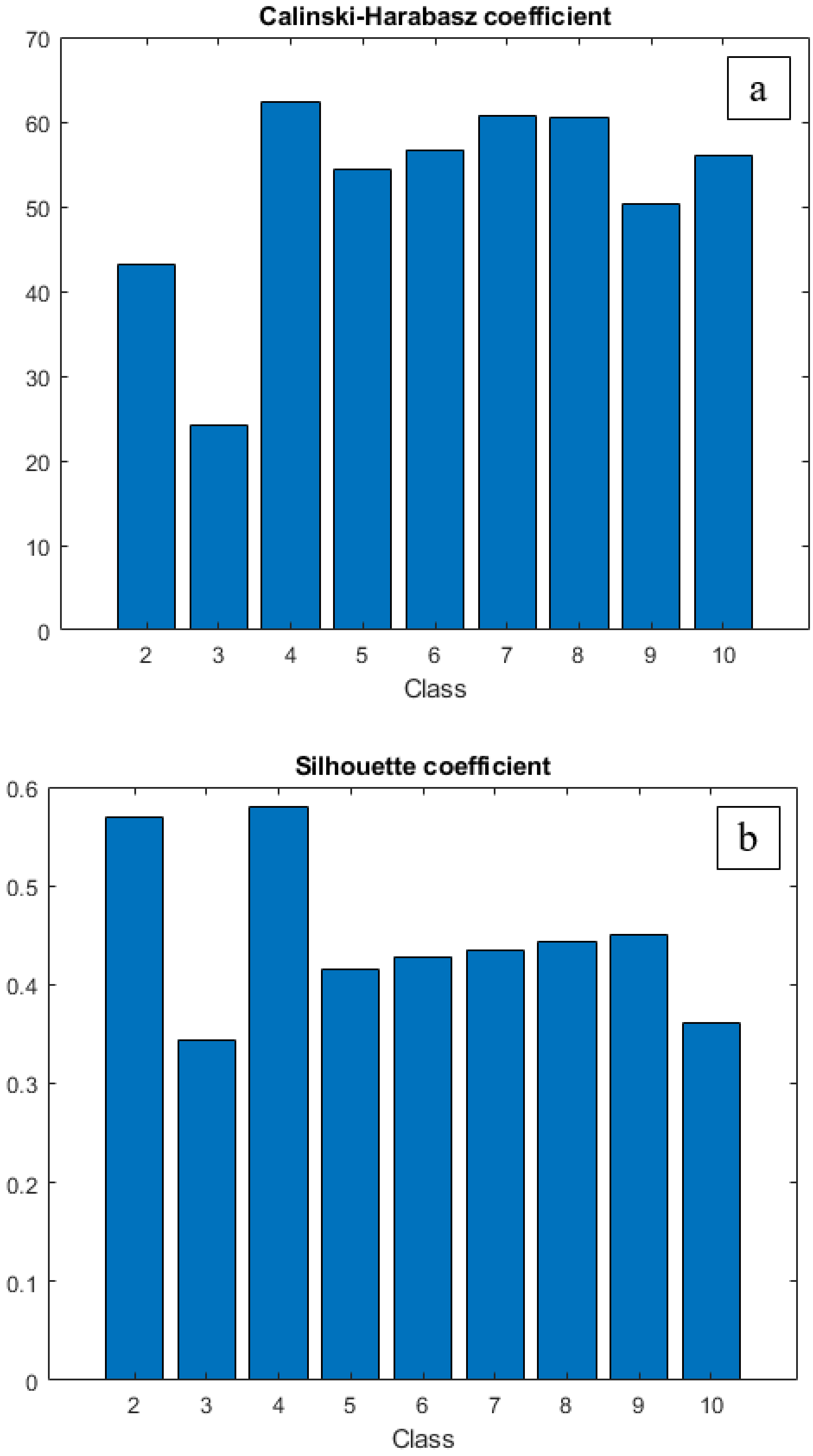
Table.2. The features selected for the SOM network.

Tolology	Distance	CoverSteps	InitNeighbor
Hextop	Linkdist	100	1

3. Results

Ten clusters as the default numbers have been defined for the SOM network as the definite number of clusters or oil families is unknown. The samples were distributed in these clusters. Nevertheless, the principal objective of this study is to find the optimum number of clusters and hence oil families among these defined clusters. Therefore, validity indices were employed.

Regarding clustering validity coefficients, the maximum values of CH (62) and SI (58) parameters were determined for four clusters (Figures 3a & b). Additionally, the minimum DB coefficient (0.8) was achieved for four clusters (Figure 3c). This means that all three used clustering validity indices showed four clusters as the optimum number of clusters. Figure 4 in a 3-D shape typifies four clusters identified by SOM neural network. Therefore, it can be concluded that four oil families exist in the Iranian part of the Persian Gulf. In other words, at least four different source rocks have generated the reservoir oils.


Calinski-Harabasz coefficient
a
70
60
50
40
30
20
10
0
2
3
4
5
6
7
8
9
10
Class
Silhouette coefficient
b
0.6
0.5
0.4
0.3
0.2
0.1
0
2
3
4
5
6
7
8
9
10
Class

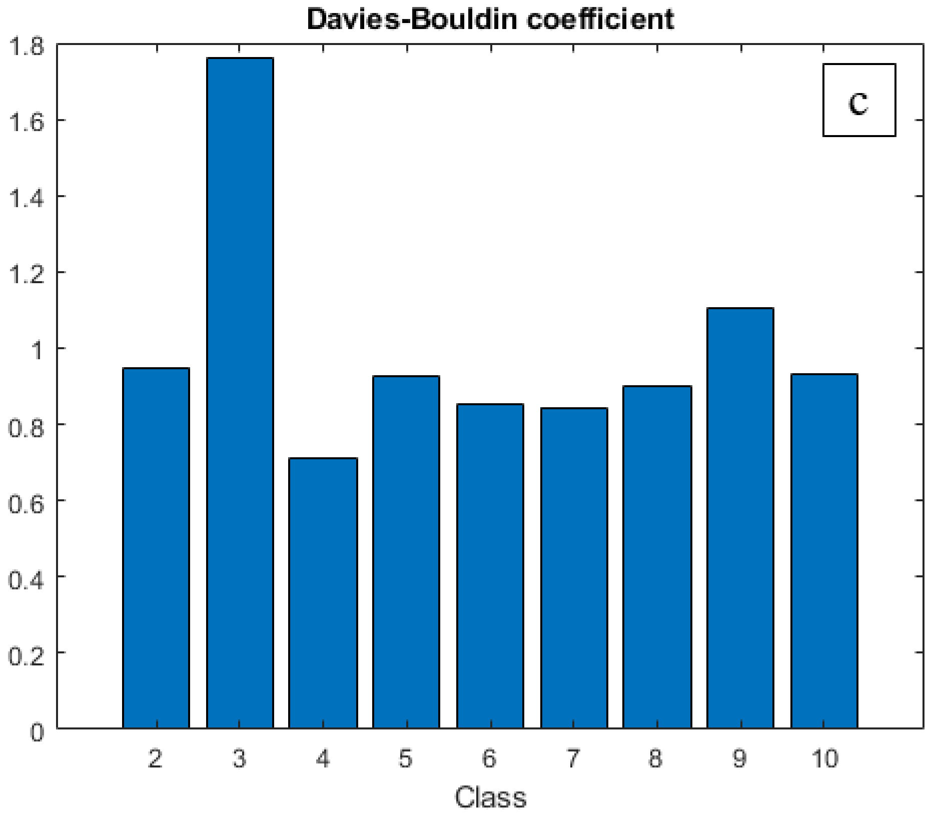


Figure3. The outcomes obtained by CH (a), H (b), and DB coefficients (c) demonstrating the optimum number of clusters.

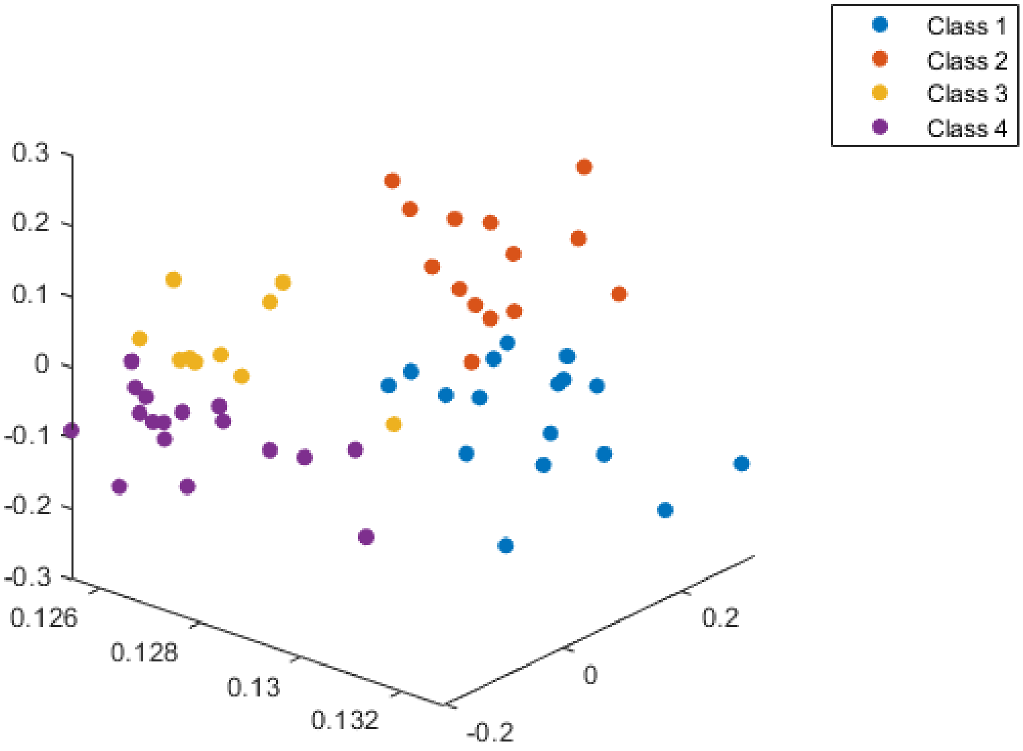


Figure 4. The schematic of the SOM clustering results illustrating four oil families.

Based on the SOM network's obtained result, cluster I consist of crude oil samples from Foroozan, Aboozar, Balal, Resalat, Reshadat, Salman, Bahregansar, and Souroush oilfields. Cluster II is composed of crude oils from Foroozan, Kharg, Dorood, Balal, Salman, and Nowrouz oilfields. Cluster III contains oil samples from Resalat, Reshadat, Hendijan, Nousrat, Siri A, Siri C, Siri D, and Siri E oilfields. Finally, crude oil samples from Kharg, Dorood, Aboozar, Reshadat, Bahregansar, Nousrat, Sirri A, Sirri C, Sirri D, and Sirri E were grouped into cluster IV.

Overall, the SOM artificial neural network employed in the present paper grouped crude oil samples into four clusters and demonstrated four oil families in the studied area. Hanifa-Tuwaiq, Garau, Diyab member of Surmeh Formation, Kazhdumi, Sarvak, Khatiya, and Ahmadi member of Sarvak Formation are regarded as the possible source rocks in the region [2]. The identified number of oil families are consistent with those suggested by Rabbani et al[2]. Nonetheless, only thirty-three samples were analysed in the mentioned research. However, sixty crude oil samples were analysed to identify oil families in the present paper to reach more reliable results.

4. Conclusions

Lack of novelty in previous studies was the main reason for which we decided to find a new method for identifying oil families, a vital study, in petroleum basins. Thus, an SOM neural network was selected for this purpose. In creating the SOM network, ten clusters were initially defined in the network. Then, three effective clustering validity coefficients were implemented to identify the optimum number of clusters based on geochemical and biomarker characteristics of oil samples used as inputs for the network. The maximum CH and SI coefficients were acquired for four clusters. Similarly, the lowest DB coefficient was obtained for four clusters among ten defined clusters. Accordingly, all three validation indices introduced four clusters as the optimum number of clusters, hence the number of oil families. Finally, it should be noted that, while some statistical methods such as PCA or HCA can be employed for oil family typing, these approaches have become over-used, and petroleum geochemistry studies and specifically oil family grouping demands novel paradigms. Accordingly, this paper introduced the SOM artificial neural network as a quick and easy-to-use method, which could be great asses for geochemists in petroleum geochemistry studies for classification purposes.

Abbreviation table

Abbreviation	Full Name
PCA	Principal Component Analysis
HCA	Hierarchical Clustering Analysis

ANNs	Artificial Neural Networks
CH	Calinski-Harabasz
DB	Davies-Bouldin
SI	Silhouette indexes
AI	Artificial Intelligence
ML	Machine Learning

Declaration of interests

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

Appendix

Table 1: biomarker parameters used as inputs for the SOM network.

Oilfield	% C_{27} steran	% C_{28} steran	% C_{29} steran	Steranes/ Terpanes	$C_{19}t/C_{23}t$	$C_{22}t/C_{21}t$	$C_{24}t/C_{23}t$	$C_{26}t/C_{25}t$	$C_{24}Tet/C_{23}t$	$C_{28}BNH/C_{30}H$	$C_{29}H/C_{30}H$	$C_{30}DiaH/C30H$	$Gam/C_{31}HR$	$C_{35}H/C_{34}H$	$d^{13}CSAT$ (‰)	$d^{13}CARO($‰)
Foroozan	39.01	22.15	38.84	0.15	0.22	1.02	0.26	0.43	1.44	0.02	1.34	0.01	0	1.17	-27.1	-27.2
Foroozan	41.84	21.32	36.83	0.17	0.25	0.94	0.29	0.48	1.37	0.02	1.36	0.03	0	1.15	-27.2	-27.2
Foroozan	36.19	22.35	41.46	0.14	0.25	0.92	0.28	0.43	1.3	0.02	1.42	0.01	0	1.02	-27.3	-27.1
Foroozan	42.75	16.33	40.92	0.13	0.24	1.04	0.3	0.46	1.31	0.03	1.52	0.01	0	0.8	-27	-26.7
Foroozan	40.7	18.42	40.88	0.16	0.16	1.1	0.39	0.16	1.29	0.02	1.36	0.01	0	0.67	-27.4	-27.2
Kharg	40.01	22.59	37.4	0.19	0.21	0.77	0.29	0.43	1.42	0.02	1.4	0.01	0.13	1.12	-27.3	-27
Kharg	41.39	20.23	38.38	0.22	0.23	0.83	0.32	0.4	1.52	0.02	1.48	0.01	0	0.91	-27.3	-26.9
Kharg	33.24	29.06	37.7	0.34	0.19	0.74	0.39	0.4	1.26	0.03	1.21	0.02	0.24	0.83	-27.3	-27.1
Dorood	38.1	21.26	40.64	0.16	0.21	0.92	0.3	0.39	1.49	0.02	1.46	0.01	0	0.9	-27.4	-27.2
Dorood	42.6	22.63	34.77	0.16	0.17	1.09	0.25	0.38	1.51	0.01	1.53	0.01	0.04	1.01	-27.4	-27.2
Dorood	31	29.78	40.3	0.35	0.14	0.56	0.56	0.42	1.08	0.04	1	0.03	0.24	0.83	-27.3	-27
Aboozar	33.17	25.18	41.65	0.24	0.18	0.69	0.41	0.41	1.49	0.02	1.14	0.01	0	0.83	-27.7	-27.4
Aboozar	34.33	28.7	36.97	0.3	0.13	0.35	0.56	0.48	1.09	0.04	0.96	0.09	0	0.69	-28.5	-27.1
Balal	43.2	22	34.81	0.3	0.83	0.42	0.53	0.5	2.22	0.07	1	0.23	0.25	0.65	-27.3	-26.6
Balal	34.51	20.45	45.04	0.16	0.3	0.75	0.37	0.48	1.65	0.02	1.03	0.02	0.1	1.04	-27.1	-26.8
Balal	32.3	21.1	46.7	0.15	0.25	0.65	0.43	0.47	1.85	0.02	0.45	0.05	0.14	1.01	-27.5	-27.2
Balal	38.3	22	39.7	0.3	0.86	0.34	0.61	0.63	2.55	0.09	0.98	0.12	0.53	0.64	-27.39	-27.08
Balal	36.3	23.2	40.5	0.3	0.78	0.39	0.61	0.53	2.51	0.06	1.07	0.15	0.19	0.63	-27.66	-27.04
Resalat	36.21	21.08	42.71	0.2	0.39	0.72	0.38	0.4	1.75	0.02	1.03	0.05	0	0.93	-27.2	-26.5
Resalat	35.36	19.59	45.05	0.19	0.58	0.65	0.46	0.37	1.89	0.04	1.21	0.04	0	0.97	-27.1	-26.6

Resalat	39.55	26.21	34.25	0.23	0.04	0.97	0.29	0.39	0.34	0.04	0.98	0.01	0.14	1.03	-27	-26.4
Resalat	37.9	29.1	33.1	0.21	0.04	1.08	0.28	0.4	0.36	0.03	1.11	0	0.32	1.35	27.2	26.5
Resalat	31.9	22.2	45.9	0.18	0.47	0.8	0.42	0.44	2.07	0.01	0.92	0.03	0.29	0.97	27.2	-26.6
Resalat	33	21.3	45.7	0.11	0.37	1	0.32	0.69	2.01	0.04	1.2	0.09	0.22	0.97	-26.3	-26
Resalat	34.38	22.85	42.76	0.14	0.38	0.95	0.29	0.65	1.89	0.04	1.21	0.04	0	0.97	-26.3	-26
Reshadat	35.36	19.59	45.05	0.19	0.58	0.65	0.46	0.37	2.08	0.03	0.9	0.06	0	0.85	-27.1	-26.6
Reshadat	35.08	29.91	35.01	0.25	0.1	0.58	0.43	0.43	0.58	0.03	0.93	0.02	0.14	0.86	-27.3	-26.2
Reshadat	36.2	28.9	34.8	0.21	0.08	0.57	0.28	0.43	0.59	0.03	0.95	0	0.32	0.91	-27.2	-26.5
Reshadat	33	21.9	45.1	0.2	0.53	0.56	0.45	0.36	2.3	0.02	0.98	0.04	0.33	1.08	-27.6	-26.9
Salman	34.24	23.34	42.41	0.25	0.55	0.62	0.46	0.45	2.03	0.03	1.03	0.05	0	0.97	-27.2	-26.3
Salman	31.13	21.75	47.12	0.24	0.56	0.57	0.47	0.48	2.06	0.03	1.06	0.06	0.12	0.81	-27.2	-26.4
Salman	30.43	16.41	53.16	0.17	0.35	0.79	0.66	0.59	1.74	0.02	0.97	0.04	0	0.62	-27.1	-26.7
Salman	34.8	20.4	44.8	0.17	0.33	0.8	0.38	0.46	1.89	0.01	1.03	0.03	0.12	0.99	-27.3	-26.8
Salman	35.1	25	39.9	0.23	0.35	0.9	0.37	0.42	1.43	0.01	1	0.03	0.15	1	-27	-26.7
Salman	37.4	21.4	41.2	0.21	0.45	0.82	0.44	0.5	2.21	0.01	1.02	0.05	0.28	0.97	-27.3	-26.7
Salman	34.5	19.7	45.9	0.16	0.34	0.64	0.4	0.41	2	0.02	1.06	0.03	0.09	1.02	-27.4	-26.7
Salman	40	20	40	0.26	0.55	0.72	0.49	0.51	2.17	0.01	0.92	0.06	0.13	0.94	-27.4	-26.8
Bahregansar	28.83	24.75	46.42	0.33	0.24	0.57	0.46	0.67	1.25	0	1	0.03	0.13	1.2	-27.34	-27.08
Bahregansar	34.47	30.73	34.8	0.36	0.09	0.41	0.65	0.81	0.53	0	0.68	0.05	0.23	0.97	-28.21	-27.06
Nowrouz	40	20	41	0.14	0.17	1.1	0.26	0.77	1.66	0	1.18	0.01	0.15	1.22	-27.87	-27.54
Souroush	30.46	20.98	48.56	0.22	0.13	0.73	0.36	0.8	1.29	0	1.07	0.03	0.16	1.15	-28.11	-27.6
Hendijan	35	27	38	0.57	0.1	0.39	0.74	0.74	0.53	0	0.68	0.05	0.23	0.97	-28.33	-27.08
Sirri D	34.56	31.52	33.92	0.24	0.07	0.58	0.4	0.43	0.56	0.04	0.86	0.02	0.09	0.97	-27.1	-26.2
Sirri D	32	31.9	36.1	0.3	0.14	0.54	0.47	0.24	0.53	0.07	0.85	0.02	0.15	0.92	-27.1	-26.3
Sirri D	37	30	33	0.27	0.07	0.67	0.41	0.54	0.51	0.04	0.76	0.02	0.22	0.95	-27.3	-26.3
Nousrat	34.54	30.11	35.34	0.24	0.09	0.57	0.44	0.42	0.55	0.04	0.94	0.02	0.08	0.86	-27.1	-26.5
Nousrat	37	29.9	33.1	0.21	0.07	0.67	0.37	0.44	0.52	0.03	0.91	0.02	0.26	1.27	-27.3	-26.1
Nousrat	40	29	31	0.22	0.07	0.87	0.36	0.58	0.37	0.04	0.96	0.02	0.24	0.96	-26.9	-26.3
Nousrat	34.54	30.11	35.34	0.24	0.09	0.57	0.44	0.42	0.55	0.04	0.94	0.02	0.08	0.86	-27.1	-26.5
Sirri E	36.15	31.9	31.96	0.22	0.1	0.49	0.43	0.44	0.6	0.03	0.89	0.01	0.06	1.1	-27.2	-26.2
Sirri E	38.1	31.1	30.8	0.21	0.1	0.48	0.45	0.39	0.67	0.04	0.85	0.02	0.25	1.28	-27.4	-26.1
Sirri E	37.2	31.6	31.3	0.23	0.1	0.49	0.46	0.38	0.67	0.05	0.79	0.02	0.28	1.27	-27.3	-26
Sirri E	32.5	34	33.6	0.27	0.14	0.42	0.5	0.23	0.63	0.02	0.78	0.02	0.25	1.18	-27.1	-26.4
Sirri E	34.5	34	31.5	0.25	0.14	0.42	0.5	0.35	0.59	0.02	0.85	0.02	0.25	1.14	-26.5	-26.6
Sirri A	33.78	31.62	34.6	0.26	0.13	0.47	0.47	0.4	0.66	0.05	0.78	0.02	0.12	0.12	-27.1	-26
Sirri A	35.6	31.6	32.8	0.26	0.11	0.47	0.51	0.46	0.78	0.05	0.77	0.02	0.26	1.16	-27.1	-26.2
Sirri A	36.9	30.1	33	0.24	0.12	0.4	0.5	0.48	0.75	0.05	0.83	0.02	0.3	1.18	-27.2	-26.1
Sirri A	35	32	33	0.26	0.11	0.48	0.48	0.53	0.55	0.05	0.84	0.05	0.3	1.17	-27	-26.3
Sirri C	34.39	31.51	34.1	0.25	0.09	0.56	0.41	0.39	0.57	0.04	0.84	0.02	0.11	1.38	-27.1	-25.9
Sirri C	35	31	34	0.25	0.09	0.51	0.46	0.43	0.66	0.04	0.86	0.02	0.27	1.17	-27.3	-26.2
Sirri C	39	27	34	0.26	0.06	0.58	0.43	0.6	0.48	0.04	0.73	0.02	0.25	1.29	-27.4	-26.2
Sirri C	37	29	34	0.25	0.08	0.72	0.4	0.57	0.45	0.03	0.77	0.02	0.2	0.9	-27	-26.4

C_{19}t/C_{23}t: C_{19} tricyclic terpanes/C_{23} tricyclic terpanes, **C_{22}t/C_{21}t**: C_{22} tricyclic terpanes /C_{21} tricyclic terpanes, **C_{24}t/C_{23}t**: C_{24} tricyclic terpanes /C_{23} tricyclic terpanes, **C_{26}t/C_{25}t**: C_{26} tricyclic terpanes /C_{25} tricyclic terpanes, **C_{24}Tet/C_{23}t**: C_{24} tetracyclic terpanes /C_{23} tricyclic terpanes, **C_{28}BNH/C_{30}H**: C_{28}Bisnorhopane/C_{30}Hopane, **C_{29}H/C_{30}H**: C_{29}Hopane/C_{30}Hopane, **C_{30}DiaH/C30H**: C_{30}DiaHopane/C30Hopane, **Gam/C_{31}HR**: Gamacerane/C_{31}HopaneRatio, **C_{35}H/C_{34}H**: C_{35}Hopane/C_{34}Hopane, **d^{13}CSAT (‰)**:d^{13}CSaturate (‰), **d^{13}CARO(‰)**: d^{13}CAromaric(‰).

References

(1) Peters, K. E.; Walters, C. C.; and Moldowan, J. . *The Biomarker Guide: Biomarkers and Isotopes in Petroleum Systems and Earth Historye*; 2005.

(2) Rabbani, A. R.; Kotarba, M. J.; Baniasad, A. R.; Hosseiny, E.; Wieclaw, D. Geochemical Characteristics and Genetic Types of the Crude Oils from the Iranian Sector of the Persian Gulf. *Org. Geochem.* **2014**, *70*, 29–43.

(3) Hosseiny, E.; Rabbani, A. R.; Moallemi, S. A. Oil Families and Migration Paths by Biological Markers in the Eastern Iranian Sector of Persian Gulf. *J. Pet. Sci. Eng.* **2017**, *150*, 54–68.

(4) Menad, N. .; Hemmati-Sarapardeh, A.; Varamesh, A.; Shamshirband, S. Predicting Solubility of CO2 in Brine by Advanced Machine Learning Systems: Application to Carbon Capture and Sequestration. *J. CO2 Util.* **2019**, *33*, 83–95.

(5) Mokarizadeh, H.; Atashrouz, S.; Mirshekar, H.; Hemmati-Sarapardeh, A.; Pour, A. . Comparison of LSSVM Model Results with Artificial Neural Network Model for Determination of the Solubility of SO2 in Ionic Liquids. *J. Mol. Liq.* **2020**, *304*, 112771.

(6) Tohidi-Hosseini, S. .; Hajirezaie, S.; Hashemi-Doulatabadi, M.; Hemmati-Sarapardeh, A.; Mohammadi, A. . Toward Prediction of Petroleum Reservoir Fluids Properties: A Rigorous Model for Estimation of Solution Gas-Oil Ratio. *J. Nat. Gas Sci. Eng.* **2016**, *29*, 506–516.

(7) Shateri, M.; Sobhanigavgani, Z.; Alinasab, A.; Varamesh, A.; Hemmati-Sarapardeh, A.; Mosavi, A. Comparative Analysis of Machine Learning Models for Nanofluids Viscosity Assessment. *Nanomaterials* **2020**, *10(9)*, 1767.

(8) Mosavi, A.; Ozturk, P.; Chau, K. . Flood Prediction Using Machine Learning Models: Literature Review. *Water* **2018**, *10(11)*, 1536.

(9) Amar, M. .; Shateri, M.; Hemmati-Sarapardeh, A.; Alamatsaz, A. Modeling Oil-Brine Interfacial Tension at High Pressure and High Salinity Conditions. *J. Pet. Sci. Eng.* **2019**, *183*, 106413.

(10) Mazloom, M. .; Rezaei, F.; Hemmati-Sarapardeh, A.; Husein, M. .; Zendehboudi, S.; Bemani, A. Artificial Intelligence Based Methods for Asphaltenes Adsorption by Nanocomposites: Application of Group Method of Data Handling, Least Squares Support Vector Machine, and Artificial Neural Networks. *Nanomaterials* **2020**, *10(5)*, 890.

(11) Rostami, A.; Hemmati-Sarapardeh, A.; Shamshirband, S. Rigorous Prognostication of Natural Gas Viscosity: Smart Modeling and Comparative Study. *Fuel* **2018**, *222*, 766–778.

(12) Hemmati-Sarapardeh, A.; Hajirezaie, S.; Soltanian, M. .; Mosavi, A.; Nabipour, N.; Shamshirband, S.; Chau, K. . Modeling Natural Gas Compressibility Factor Using a Hybrid Group Method of Data Handling. *Eng. Appl. Comput. Fluid Mech.* **2020**, *14(1)*, 27–37.

(13) Amooie, M. .; Hemmati-Sarapardeh, A.; Karan, K.; Husein, M. .; Soltanian, M. .; Dabir, B. Data-Driven Modeling of Interfacial Tension in Impure CO2-Brine Systems with Implications for Geological Carbon Storage. *nternational J. Greenh. Gas Control* **2019**, *90*, 102811.

(14) Razghandi, M.; Hemmati-Sarapardeh, A.; Rashidi, F.; Dabir, B.; Shamshirband, S. Smart

Models for Predicting Under-Saturated Crude Oil Viscosity: A Comparative Study. *Energy Sources, Part A Recover. Util. Environ. Eff.* **2019**, *41(19)*, 2326–2333.

(15) Bolandi, V.; Kadkhodaie-Ilkhchi, A.; Alizadeh, B.; Tahmorasi, J.; Farzi, R. Source Rock Characterization of the Albian Kazhdumi Formation by Integrating Well Logs and Geochemical Data in the Azadegan Oilfield, Abadan Plain, SW Iran. *J. Pet. Sci. Eng.* **2015**, *133*, 167–176.

(16) Bolandi, V.; Kadkhodaie, A.; Farzi, R. Analyzing Organic Richness of Source Rocks from Well Log Data by Using SVM and ANN Classifiers: A Case Study from the Kazhdumi Formation, the Persian Gulf Basin, Offshore Iran. *J. Pet. Sci. Eng.* **2017**, *151*, 224–234.

(17) Tabatabaei, S. M. E.; Kadkhodaie-Ilkhchi, A.; Hosseini, Z.; Moghaddam, A. A. A Hybrid Stochastic-Gradient Optimization to Estimating Total Organic Carbon from Petrophysical Data: A Case Study from the Ahwaz Oilfield, SW Iran. *Comput. Geosci.* **2015**, *127*, 35–43.

(18) Naghizadeh, A.; Larestani, A.; Nait Amar, M.; Hemmati-Sarapardeh, A. Predicting Viscosity of CO2–N2 Gaseous Mixtures Using Advanced Intelligent Schemes. *J. Pet. Sci. Eng.* **2022**, *208*, 109359. https://doi.org/10.1016/j.petrol.2021.109359.

(19) Kadkhodaie-Ilkhchi, A.; Rahimpour-Bonab, H.; Rezaee, M. A Committee Machine with Intelligent Systems for Estimation of Total Organic Carbon Content from Petrophysical Data: An Example from Kangan and Dalan Reservoirs in South Pars Gas Field, Iran. *Comput. Geosci.* **2009**, *35(3)*, 459–474.

(20) Ghiasi-Freez, J.; Kadkhodaie-Ilkhchi, A.; Ziaii, M. The Application of Committee Machine with Intelligent Systems to the Prediction of Permeability from Petrographic Image Analysis and Well Logs Data: A Case Study from the South Pars Gas Field, South Iran. *Pet. Sci. Technol.* **2012**, *30(20)*, 2122–2136.

(21) Esfahani, S.; Baselizadeh, S.; Hemmati-Sarapardeh, A. On Determination of Natural Gas Density: Least Square Support Vector Machine Modeling Approach. *J. Nat. Gas Sci. Eng.* **2015**, *22*, 348–358.

(22) Hajirezaie, S.; Pajouhandeh, A.; Hemmati-Sarapardeh, A.; Pournik, M.; Dabir, B. Development of a Robust Model for Prediction of Under-Saturated Reservoir Oil Viscosity. *J. Mol. Liq.* **2017**, *229*, 89–97.

(23) Karkevandi-Talkhooncheh, A.; Hajirezaie, S.; Hemmati-Sarapardeh, A.; Husein, M. .; Karan, K.; Sharifi, M. Application of Adaptive Neuro Fuzzy Interface System Optimized with Evolutionary Algorithms for Modeling CO2-Crude Oil Minimum Miscibility Pressure. *Fuel* **2017**, *205*, 34–45.

(24) Barati-Harooni, A.; Najafi-Marghmaleki, A.; Hoseinpour, S. .; Tatar, A.; Karkevandi-Talkhooncheh, A.; Hemmati-Sarapardeh, A.; Mohammadi, A. . Estimation of Minimum Miscibility Pressure (MMP) in Enhanced Oil Recovery (EOR) Process by N2 Flooding Using Different Computational Schemes. *Fuel* **2019**, *235*, 1455–1474.

(25) Amiri-Ramsheh, B.; Safaei-Farouji, M.; Larestani, A.; Zabihi, R.; Hemmati-Sarapardeh, A. Modeling of Wax Disappearance Temperature (WDT) Using Soft Computing

Approaches: Tree-Based Models and Hybrid Models. *J. Pet. Sci. Eng.* **2021**, *208*, 109774. https://doi.org/10.1016/j.petrol.2021.109774.

(26) Mohammadi, M. .; Hadavimoghaddam, F.; Pourmahdi, M.; Atashrouz, S.; Munir, M. .; Hemmati-Sarapardeh, A.; Mosavi, A. .; Mohaddespour, A. Modeling Hydrogen Solubility in Hydrocarbons Using Extreme Gradient Boosting and Equations of State. *Sci. Rep.* **2021**, *11(1)*, 1–20.

(27) Moosanezhad-Kermani, H.; Rezaei, F.; Hemmati-Sarapardeh, A.; Band, S. .; Mosavi, A. Modeling of Carbon Dioxide Solubility in Ionic Liquids Based on Group Method of Data Handling. *Eng. Appl. Comput. Fluid Mech.* **2021**, *15(1)*, 23–42.

(28) Rezaei, F.; Jafari, S.; Hemmati-Sarapardeh, A.; Mohammadi, A. . Modeling of Gas Viscosity at High Pressure-High Temperature Conditions: Integrating Radial Basis Function Neural Network with Evolutionary Algorithms. *J. Pet. Sci. Eng.* **2022**, *208*, 109328.

(29) Khamehchi, E.; Mahdiani, M. .; Amooie, M. .; Hemmati-Sarapardeh, A. Modeling Viscosity of Light and Intermediate Dead Oil Systems Using Advanced Computational Frameworks and Artificial Neural Networks. *J. Pet. Sci. Eng.* **2020**, *193*, 107388.

(30) Safaei-Farouji, M.; Kadkhodaie, A. Application of Ensemble Machine Learning Methods for Kerogen Type Estimation from Petrophysical Well Logs. *J. Pet. Sci. Eng.* **2021**, *208*, 109455.

(31) Balakrishnan, P. .; Cooper, M. .; Jacob, V. .; Lewis, P. . A Study of the Classification Capabilities of Neural Networks Using Unsupervised Learning: A Comparison WithK-Means Clustering. *Psychometrika* **1994**, *59(4)*, 509–525.

(32) Erilli, N. .; Yolcu, U.; Eğrioğlu, E.; Aladağ, Ç. .; Öner, Y. Determining the Most Proper Number of Cluster in Fuzzy Clustering by Using Artificial Neural Networks. *Expert Syst. Appl.* **2011**, *38(3)*, 2248–2252.

(33) Du, K. . Clustering: A Neural Network Approach. *Neural networks* **2010**, *23(1)*, 89–107.

(34) Kohonen, T. *Self-Organization and Associative Memory*; 1989.

(35) Cabanes, G.; Bennani, Y. Learning the Number of Clusters in Self Organizing Map. In *INTECH Open Access Publisher*; 2010; pp 14–28.

(36) Ghaseminezhad, M. H.; Karami, A. A Novel Self-Organizing Map (SOM) Neural Network for Discrete Groups of Data Clustering. *Appl. Soft Comput.* **2011**, *11(4)*, 3771–3778.

(37) Mashhadi, Z. S.; Rabbani, A. R. Organic Geochemistry of Crude Oils and Cretaceous Source Rocks in the Iranian Sector of the Persian Gulf: An Oil–Oil and Oil–Source Rock Correlation Study. *Int. J. Coal Geol.* **2015**, *146*, 118–144.

(38) Rabbani, A. . Petroleum Geochemistry, Offshore SE Iran. *Geochemistry Int.* **2007**, *45(11)*, 1164–1172.

(39) Abdi, H.; Williams, L. . Principal Component Analysis. *Wiley Interdiscip. Rev. Comput.*

Stat. **2010**, *2(4)*, 433–459.

(40) Ringnér, M. What Is Principal Component Analysis? *Nat. Biotechnol.* **2008**, *26(3)*, 303–304.

(41) Kohonen, T. The Self-Organizing Map. *Neurocomputing* **1998**, *21(1)*, 1–6.

(42) Clark, S.; Sarlin, P.; Sharma, A.; Sisson, S. A. Increasing Dependence on Foreign Water Resources? An Assessment of Trends in Global Virtual Water Flows Using a Self-Organizing Time Map. *Ecol. Inform.* **2015**, *26(2)*, 192–202.

(43) Dayhoff, J. . Neural Network Architectures: An Introduction. *Van Nostrand Reinhold Co* **1990**.

(44) Wang, Z.; Bian, S.; Liu, Y.; Liu, Z. The Load Characteristics Classification and Synthesis of Substations in Large Area Power Grid. *Int. J. Electr. Power Energy Syst.* **2013**, *48*, 71–82.

(45) Vesanto, J. SOM-Based Data Visualization Methods. *Intell. data Anal.* **1999**, *3(2)*, 111–126.

(46) Kohonen, T.; Kaski, S. Self-Organized Formation of Various Invariantfeaturefiters in the Adaptive-Subspace SOM”. *Neural Comput.* **1997**, *9*, 1321–1344.

(47) Ünlü, R.; Xanthopoulos, P. Estimating the Number of Clusters in a Dataset via Consensus Clustering. *Expert Syst. Appl.* **2019**, *125*, 33–39.

(48) Chiang, M. M. .; Mirkin, B. Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads. *J. Classif.* **2010**, *27(1)*, 3–40.

(49) Liang, J.; Zhao, X.; Li, D.; Cao, F.; ADang, C. Determining the Number of Clusters Using Information Entropy for Mixed Data. *Pattern Recognit.* **2012**, *45(6)*, 2251–2265.

(50) Rendón, E.; Abundez, I.; Arizmendi, A.; Quiroz, E. . Internal versus External Cluster Validation Indexes. *Int. J. Comput. Commun.* **2011**, *5(1)*, 27–34.

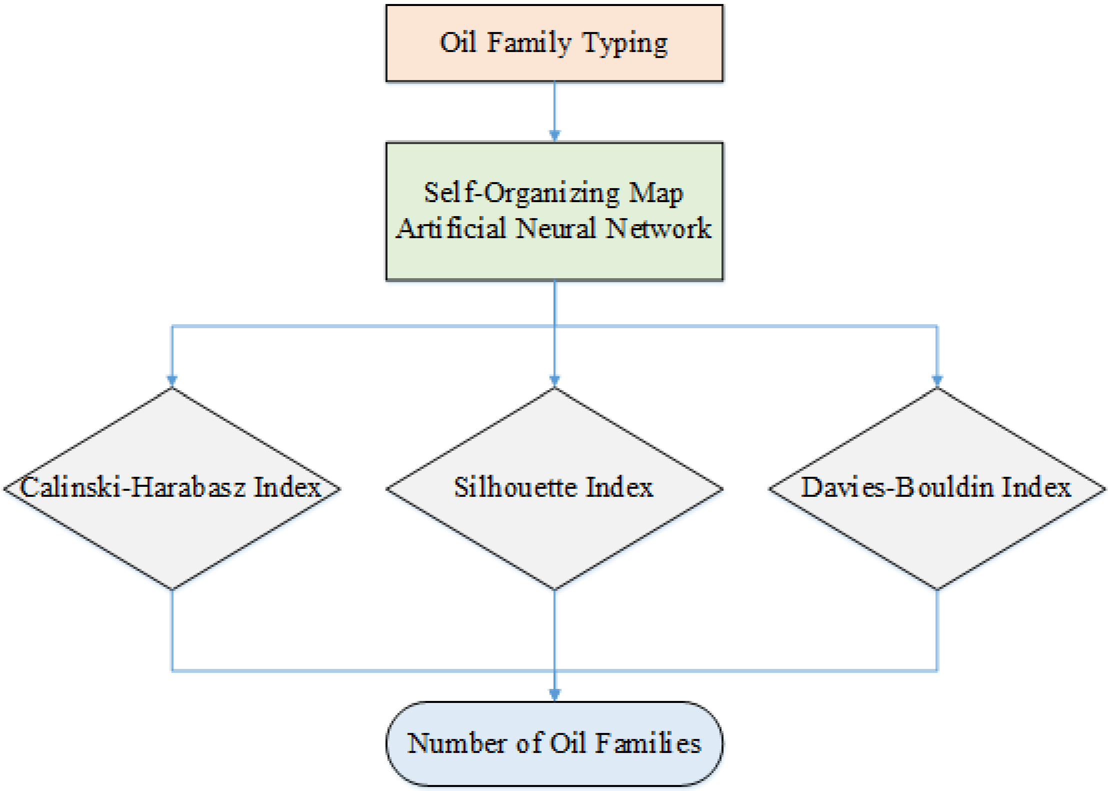
(51) Wang, K.; Wang, B.; Peng, L. CVAP: Validation for Cluster Analyses. *Data Sci. J.* **2009**, *8*, 88–93.

(52) Davies, D. .; Bouldin, D. . A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *2*, 224–227.

(53) Caliński, T.; Harabasz, J. A Dendrite Method for Cluster Analysis. *Commun. Stat. Methods* **1974**, *3(1)*, 1–27.

(54) Rousseeuw, P. . Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65.

(55) Liu, Y.; Li, Z.; Xiong, H.; Gao, X.; Wu, J. Understanding of Internal Clustering Validation Measures. In *2010 IEEE international conference on data mining*; IEEE, 2010; pp 911–916.


Oil Family Typing
Self-Organizing Map
Artificial Neural Network
Calinski-Harabasz Index
Silhouette Index
Davies-Bouldin Index
Number of Oil Families

TOC Graphic