Gas Processing Journal

Vol. 6, No. 2, 2018, pp. 1-14
http://gpj.ui.ac.ir
DOI: http://dx.doi.org/ 10.22108/gpj.2018.112015.1035

Transparent Machine Learning Algorithm Offers Useful Prediction Method for Natural Gas Density

David A. Wood, Abouzar Choubineh

[1] DWA Energy Limited, Lincoln, United Kingdom
[2] MSc Petroleum University of Technology, Ahwaz, Iran

Received: 2018-07-10 Revised: 2018-04-19 Accepted: 2018-10-15

Abstract: Machine-learning algorithms aid predictions for complex systems with multiple influencing variables. However, many neural-network related algorithms behave as black boxes in terms of revealing how the prediction of each data record is performed. This drawback limits their ability to provide detailed insights concerning the workings of the underlying system, or to relate predictions to specific characteristics of the underlying variables. The recently proposed transparent open box (TOB) learning network algorithm successfully addresses these issues, by revealing the exact calculation involved in the prediction of each data record. That algorithm, described in summary, can be applied in a spreadsheet or fully-coded configurations and offers significant benefits to analysis and prediction of many natural gas systems. The algorithm is applied to the prediction of natural gas density using a published dataset of 693 data records involving 14 variables (temperature and pressure plus the molecular fractions of the twelve components: methane, ethane, propane, 2-methylpropane, butane, 2-methylbutane, pentane, octane, toluene, methylcyclopentane, nitrogen and carbon dioxide). The TOB network demonstrates very high prediction accuracy (up to R2 =0.997), achieving comparable accuracy to the predictions reported (R2 =0.995) for an artificial neuralnetwork (ANN) algorithm applied to the same data set. With its high levels of transparency, the TOB learning network offers a new approach to machine learning as applied to many natural gas systems.

keywords: Predicting gasdensity; Learning networks;Multi-component natural gas; Auditable machine learning; Transparentpredictions

1. Introduction

The employment of machine learning algorithms to provide accurate predictions from complex systems governed by multiple variables with poorly defined non-linear relationships is growing. The use of system learning tools such as artificial neural networks (ANN), Adaptive Neuro-Fuzzy Inference Systems (ANFIS), support vector machines (SVM),least squares support vector machine (LSSVM), etc., are being ever more widely applied as systems learning tools (Schmidhuber, 2015).

The learning potential of ANN was recognized in the 1950's (Kleene, 1956) and has developed with a number of different algorithms now routinely exploited, such as the multilayer perceptron (MLP) (Hush & Horne, 1993; Haykin, 1995) and radial basis functions (RBF) (Broomhead & Lowe, 1988). Since its development in the 1990s ANFIS adapts ANN with a Takagi–Sugeno fuzzy inference system (Jang, 1993) and has successfully demonstrated its learning capabilities when applied to approximate and uncertain non-linear functions (Jang, Sun, & Mizutani, 1997). SVM and LSSVM algorithms, also developed in the 1990s, provide supervised learning that is successfully applied as nonlinear regression and correlation analysis (Cortes & Vapnik, 1995; Vapnik, 2000). These machine learning algorithms are now commonly applied to provide predictions to many non-linear systems, including those in the gas and oil industries. Moreover, they are also widely used ina hybrid form, coupled with various optimization algorithms such as genetic algorithms to improve their performance (Ghorbani, Ziabasharhagh, &

* Corresponding Author.
Authors' Email Address: D. A. Wood (dw@dwasolutions.com), A. Choubineh (abouzar68choubineh@gmail.com),
ISSN (Online): 2345-4172, ISSN (Print): 2322-3251

Amidpour, 2014; Ghorbani, Hamedi, Shirmohammadi, Mehrpooya, & Hamedi, 2016) and data handling capabilities (Ghorbani, Shirmohammadi, Mehrpooya, & Hamedi, 2018; Shirmohammadi, Ghorbani, Hamedi, Hamedi, & Romeo, 2015; Shirmohammadi, Soltanieh, & Romeo, 2018; Choubineh, Ghorbani, Wood, Moosavi, Khalafi, & Sadatshojaei, 2017).

However, in academia and industry, the extensive exploitation of machine learning algorithms in their many hybrid forms has polarized scientists, particularly in the oil and gas industry. Perhaps the most-contentious issue is the lack of transparency provided by neural networks regarding their inner calculations, particularly the relative weightings and adjustments made to input variables in deriving specific predictions. This often leads to them being used and viewed as blackboxes (Heinert, 2008). It requires complex and sometimes cumbersome simulations to gain insight to the ways variables are treated in their calculations. At best this turns them into "white boxes" that provide insight to the relative influences of input variables on the calculations beingmade (Elkatatny, Tariq, & Mahmoud, 2016).

This black-box condition frustrates and infuriates some scientists and industry practitioners. If it is not possible to see, quickly and in detail, how a prediction is derived from a machine learning tool, and no new fundamental insight is provided about the underlying system, then, for example, many experimental scientists see no value in such systems. Those in this camp when reviewing machine-learning studies simply dismiss them as correlation analysis of minor importance with no experimental justification. On the other hand, some oil and gas companies and their suppliers/ service companies are comfortable with a black-box approach as in some circumstances it can enhance their competitive advantage and keep their underlying data analysis confidential. Some researchers also embrace the black-box condition by their willingness to just enter the input-variable data into opaque coding (e.g., some MatLab machine-learning functions), derive accurate correlations and predictions for their objective function and make bold claims concerning the superiority of their newly developed algorithms. This leaves many practitioners blind to the inner workings of such systems. Nevertheless, the uptake of machine learning and the diversity of its applications continue to grow and have more impact on the way decisions are taken and real-time actions determined in the field.

What is urgently required are more-transparent machine learning tools that raise awareness about their underlying systems rather than obscure them. A recently-proposed algorithm, the Transparent Open-Box (TOB) network (Wood, 2018) demonstrates that it is possible to do this in such way that sufficient prediction accuracy is provided at the same time as revealing the inner network calculations involved in deriving the prediction for each data record.

In the natural gas industry density of natural gas (ρ) is dependent on several complex non-linear relationships relating to its physical and chemical characteristics. Its prediction from underlying physiochemical conditions makes it suitable for machine learning applications.

Natural gas density is an important metric contributing to the calculations of other variables relevant to numerous systems involving natural gas (e.g. pipelines, storage facilities, and underground reservoirs). It is complex because it varies significantly with respect to pressure, temperature and gas composition (AlQuraishi & Shokir, 2011). However, measuring it experimentally is time-consuming and expensive, and estimating it from PVT data involves significant assumptions about related metrics, which are difficult to define with accuracy (e.g. z-factors). Although various equations of state (EOS) are proposed for calculating ρ they have proved to be too simple and inconsistent across the full P-T and compositional ranges encountered (Elsharkawy, 2003; Farzaneh-Gord, Khamforoush, Hashemi, & Pourkhadem, 2010).

Shokir (2008) proposed a fuzzy logic method and AlQuraishi and Shokir (2011) developed the probabilistic alternating conditional expectations model to predict ρ. Since then several machine learning algorithms have been applied to accurately predictρ with various machine learning algorithms applied to various medium-sized and large databases. These include: ANN (AlQuraishi & Shokir, 2009); LSSVM (Esfahani, Baselizadeh, & Hemmati-Sarapardeh, 2015); ANN-TLBO (Choubineh, Khalafi, Kharrat, Bahreini,& Hosseini, 2017); and, ANFIS (Dehaghani & Badizad, 2017). These studies have typically achieved accuracies of predicted versus measured data with coefficients of determination of about 0.99 and very low

values of statistical error measures (e.g., root mean squared error).

Here we describe the methodology and mathematical basis of the TOB algorithm and demonstrate its application benefits using, as an example, a complex non-linear natural gas system for predicting natural gas density from mole fractions of gas composition, together with its temperature and pressure. We have selected a comprehensive published dataset (Atilhan, Aparicio, Karadas, Hall, & Alcade, 2012) of experimental measurements performed on Qatar North Field natural gas samples (693 data records), because it has been previously used for published ANN study (Choubineh, Khalafi, Kharrat, Bahreini, & Hosseini, 2017) (to predict ρ. Our objective is to show that the TOB algorithm is capable of producing comparable accuracy for predicting from this dataset as the ANN study with the additional benefit of providing transparency to each individual prediction it calculates. We make no claims that the TOB algorithm can outperform other more mathematically-complex machine-learning algorithms (ANN, ANFIS, LSSVM etc.) in terms of prediction accuracy, but rather that it can achieve acceptable levels of accuracy with the additional benefit of greater transparency.

2. The Transparent Open-box learning Network Algorithm: Methodology and Mathematical Basis

The TOB network approach was proposed and outlined by Wood (2018). There are 14 steps, divided into two stages (stage 1 and stage 2), involved in applying the TOB learning network algorithm (2018). These sequences of steps are explained here with the sequence of steps summarized in a flow diagram (Figure 1).

Step 1: Assemble data into a two-dimensional (2D) array consisting of (*N+1*) variables and *M* data records. The variables include *N* input variables plus one dependent variable (i.e., the prediction-objective dependent variable or PODV) to predict for the *M* data records.

Step 2: Sort and rank the data records into ascending or descending order of the PODV values.

Step 3: Calculate standard statistical metrics (i.e., minimum, maximum, mean, standard deviation, etc) defining the range and distribution of each of the N+*1* variables in the dataset. These statistics must include the minimum and maximum values as these are used for the normalization process described in step 4.

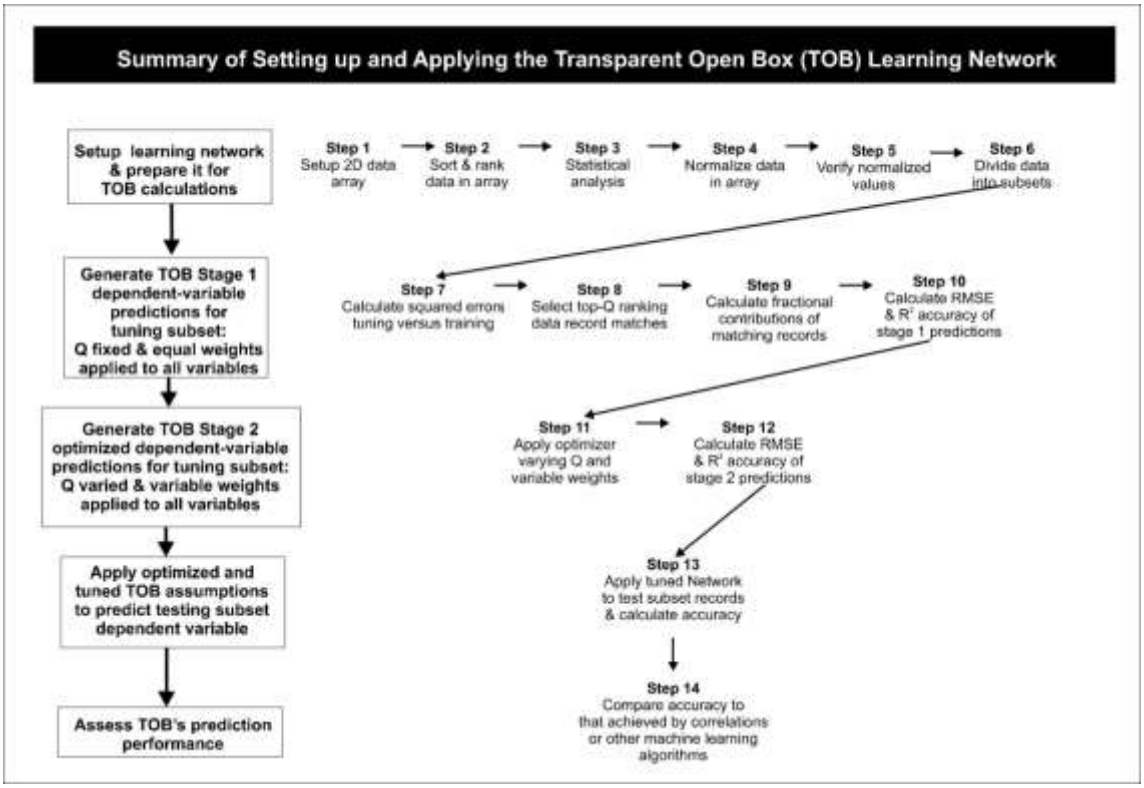


Figure 1. Flow diagram illustrating the 14 steps involved in the application of the transparent open-box (TOB) learning network algorithm (2018). The methodology and mathematical details associated with each step are described in the text.

Step 4: Normalize (*M*) data records for each (*N+1*) variable. To provide a normalized range of (-1, +1) for variable X as minimum and maximum limits use Eq. (1)

$$X_i^* = 2*[(X_i - Xmin)/(Xmax - Xmin)]-1 \quad (1)$$

Where:

X_i is the actual value of variable *X* for the i^{th} data record

Xmin is the minimum value of variable *X* that exists in the entire data set

Xmax is the maximum value of variable *X* that exists in the entire data set

X_i^* is the normalized value of variable *X* for the i^{th} data record

Step 5. Calculate the standard statistical metrics for the normalized dataset. This is not an essential step, but it provides a useful check to confirm that the normalized variables do, indeed, all fall within the range -1 to +1 as intended.

Step 6. Divide the normalized-2D array into training, tuning and testing subsets. The testing subset is kept apart and is not involved in model training or tuning. Sensitivity analysis can be conducted to establish the best division of data records (i.e., percentages of the entire dataset) to allocate to each subset. Typically, more than about 70% of the data records are allocated to the training subset. Consequently, up to about 15%, depending on the size of the data set, are then allocated to each of the other two subsets (i.e., tuning and testing). Such divisions of the data records produce meaningful levels of prediction accuracy. For a specific dataset, exact percentages allocated to each subset can be refined by running sensitivity cases.

Step 7. Calculate the squared error for each of the *N+ 1 variable* (i.e., the variable squared error, VSE) between each of the tuning records (*J* in total) and all the data records allocated to the training subset (*K* in total) as expressed in Eq. (2):

$$VSE(X)_{jk} = \left[X_k(tr) - X_j(tu)\right]^2 \quad (2)$$

Where:

$X_k(tr)$ is the value of variable *X* for the k^{th} data record in the training subset

$X_j(tu)$ is the value of variable *X* for the j^{th} data record in the tuning subset

$VSE(X)_{jk}$ is the squared error of variable *X* for the j^{th} data record in the tuning subset versus the k^{th} data record of the training subset.

Then sum the *VSE* values for each of the *N+1* variables between each data record to calculate $\sum VSE$. In this step and for TOB stage 1 up to step 10, equal weighting factors (W_n) are applied to each variable as expressed in Eq. (3):

$$\sum VSE_{jk} = \sum_{n=1}^{n=N+1} VSE(Xn)_{jk} * (Wn) \quad (3)$$

Where:

$VSE(Xn)_{jk}$ is the squared error for variable *Xn* for the j^{th} data record in the tuning subset versus the k^{th} data record of the training subset.

$\sum VSE_{jk}$ is the sum of the squared errors for all *N+1* variables for the j^{th} data record in the tuning subset versus the k^{th} data record of the training subset.

Wn is a weighting factor applied to the squared error of variable n. Each of the *N+1* weighting factors is free to be allocated an independent value between 0 and 1. However, in stage one of the algorithm (i.e. for step 7 to step 10, Figure 1) the weighting factors for all *N+1* variables are set to the same non-zero value (e.g. all *Wn* values are set to a single-constant value, e.g., they could all be set to equal 1, or all be set to 0.5, or another constant number between 0 and 1) so that no bias is introduced among the variable contributions to the match established in stage one of the algorithm.

Use $\sum VSE_{jk}$ values as the basis for ranking the matches in the training subset in ascending order of (∑VSE) for each tuning subset record.

Step 8. Select the top-*Q*-ranking data records (i.e., those with the lowest in ($\sum VSE$)) in the training subset for each tuning subset data record. Rank these high-matching records in order; so that the training subset data record with lowest ($\sum VSE$) versus data record *j* in the tuning subset is ranked as the #1 match for tuning subset record *j*. The integer value of Q is typically set to 10 for TOB stage 1 (i.e. up to step 10) and is refined later in the optimization stage two (see Step 11). These top-10-matching data records from the training subset for each tuning subset is recorded and then made available for the detailed calculation of the PODV prediction for each tuning subset data record.

Step 9. The top-ten-ranking records in the training subset for the j^{th} data record in the tuning subset each contribute a fraction to the predicted value of the dependent variable for that j^{th} data record. That fractional contribution is calculated by Eq. (4) to Eq. (6) and is

proportionalto their relative $\sum VSE$ scores for the j^{th} data record.

$$f_q = \sum VSE_{jq} / [\sum_{r=1}^{r=Q} \sum VSE_{jr}] \quad (4)$$

Where:

q and r are each one of the Q-top-ranking records from the training subset with the closest matches to the j^{th} record in the tuning subset.

f_q is the fractional contribution of one of the top-Q-ranking records for the j^{th} record in the tuning subset calculated such that Eq.(5) applies.

$$\sum_{q=1}^{q=Q} f_q = 1 \quad (5)$$

In order to ensure that the matching record with the lowest $\sum VSE_{jk}$ value contributes most to the dependent variable prediction for the j^{th} data record, Eq. (6) is then calculated involving all of the top-Q-ranking records.

$$(X_{N+1})_j^{predicted} = \sum_{q=1}^{q=Q} [(X_{N+1})_q * (1 - f_q)] \quad (6)$$

Where:

$(X_{N+1})_q$ is the dependent variable for the q^{th} data record in the training subset, which is one of the Q-top-ranking data records in the training subset for the j^{th} data record in the tuning subset.

$(X_{N+1})_j^{predicted}$ is the initially predicted value for the dependent variable for the j^{th} data record in the tuning subset (with equal weighting, as described in Step 7, applied to all the variables).

Applying Eq. (6) ensures that the rank#1 in the training subset top-matching records contributes most to the predicted values. On the other hand, the rank #Q match in the training subset contributes least to the dependent-variable prediction for the j^{th} data record in the tuning subset.

Step 10. Compute the coefficient of determination (R^2), mean square error (MSE) and root mean square error (RMSE) for the predicted versus actual or measured values of the PODV for all J data records in the tuning subset using Eq. (7) and Eq. (8).

$$R^2 = 1 - \frac{\sum_{j=1}^{j=J} (X_j^{actual} - X_j^{predicted})^2}{\sum_{j=1}^{j=J} (X_{ave}^{actual} - X_j^{predicted})^2} \quad (7)$$

$$MSE = \frac{1}{J} \sum_{j=1}^{j=J} (X_j^{actual} - X_j^{predicted})^2 \quad (8)$$

$$RMSE = \sqrt{MSE} \quad (9)$$

Where:

X_j is the dependent variable (designated $(X_{N+1})_j$ in Eq. (6)) for the j^{th} data record in the tuning subset

X_j^{actual} is the actual value of the dependent variable for the j^{th} data record

$X_j^{predicted}$ is the predicted value of the dependent variable for the j^{th} data record

X_{ave}^{actual} is the average actual value of the dependent variable for all J data records in the tuning subset.

This step represents the end of TOB stage one of the prediction process. Step 10 provides a provisionally-tuned TOB network that provides predictions based upon uniform weighting (as described in Step 7) applied to all the variables and by matching data records with those in the training subset.

The TOB stage 2 involves applying optimization to improve the accuracy of the predictions for the tuning set as a whole. TOB stage 2 also tests the optimized prediction metrics with the yet-to-be- used independent testing subset.

Step 11. Apply an optimizer to the provisionally-tuned TOB network to improve the accuracy of the PODV prediction for the tuning subset. The optimizer is set up with its objective function to minimize RMSE (Eq.9) for all J data records in the tuning subset by varying a set of optimization variables within specified constraints. These optimization variables are:

1. The weights (W_n) applied by the optimizer to each of the N input variables in Eq. (3) are allowed to vary independently between values 0 and 1. This contrast with Step 7 of stage one of the algorithm, where all the weights were initially set to the same constant number between 0 and 1. Also, in Step 11 the dependent variable (identified as variable N+1) is not involved in the optimization as it is considered as an unknown, so a zero weight is applied to it. Sometimes, very-low weights (e.g. 1.0 E-10 or less) may be selected as optimum weights for certain input variables. This very-low-weight value does not mean that such a variable is insignificant in the optimum solution. Their non-zero values, albeit small, will contribute to selecting the relative contributions of each of the top-matching records in the predictions made. This point is illustrated for an example data record from the dataset evaluated.

2. The integer values of Q (how many of the top-matching records to include in the

predictions) is allowed to vary in Eqs. (4), (5) and (6). Typically, 2<=Q<=10 is the range within which Q is allowed to vary and the optimizer selects the best value of Q from that range. Q values of higher than 10 could be used. However, the experience of applying the algorithm to multiple datasets suggests that all the top-ten matching records are not used in the optimum solutions found by the optimizer.

In this study, the standard "Solver" optimizer in Microsoft Excel is used to conduct the optimization process. Specifically, it is the GRG (Generalized Reduced Gradient) algorithm option within the Solver function that is used. GRG applies a robust non-linear-programming algorithm (Farzaneh-Gord, Khamforoush, Hashemi, & Pourkhadem, 2010). GRG is setup to "multistart" (i.e., run multiple cases each with a population of 150) and to converge to a solution value of 0.0001, if possible, for the RMSE objective function. GRG can be run directly from an Excel worksheet or as part of a visual basic for applications (VBA) code in Excel. It is possible to use other fully-coded optimizers to achieve this, but the advantages of doing it in Excel for mid-sized dataset is explained.

The optimization process accepts the top-Q matches in the training subset for each data record in the tuning subset established by step 8 of TOB stage 1. However, in TOB stage two it re-evaluates the $\sum VSE_{jk}$ scores using Eq. (3) by varying W_N in each iteration of the optimizer, and the $\sum VSE_{jq}$scores scores use Eq. (4) by varying Q in each iteration of the optimizer.

Step 12. Evaluate and compare the *RMSE* and R^2 values obtained by the optimum solution found by step 11. The statistical accuracy of the predictions derived from Step 11 typically demonstrates a significant improvement on the TOB-stage-1 predictions (from step 10). Also, this step runs and evaluates sensitivity cases with different fixed values of Q (2 to 10). All but one of these sensitivity cases is sub-optimal. However, comparing these sensitivity-case results helps to identify regions of the TOB network that might be prone to under fitting or over fitting.

Step 13. Apply the weights and Q values of the optimized learning network, tuned for the tuning and training subsets, to the independent testing subset. The *RMSE* and R^2 values obtained for the testing subset should be close to those for the optimized tuning subset. The detailed prediction calculations for each data record (testing and tuning subsets) are transparently recorded and can be reviewed to interrogate the reason for prediction outliers, if any occur. This is a useful prediction-auditing attribute of the TOB and helps to provide deeper insight to the underlying dataset and optimally-tuned network. It also generates more confidence in the reliable range for which meaningful predictions can be generated.

Step 14. Decide whether the level of accuracy achieved by the TOB is fit-for-purpose? If sodeploy it. If not? Interrogate the prediction performance of the TOB by reviewing in detail the prediction calculations for each of the data records of the tuning and testing data subsets to establish the PODV value ranges for which the network lacks sufficient accuracy. This information can help to focus the network on viable PODV ranges and establish value ranges for which the dataset is too sparse. This information can also be useful as a benchmark for assessing the performance of other machine-learning algorithms applied to the same data set.

In summary, TOB Stage 1 involves constructing a network of initial record matches from a large training subset to the individual records of a much smaller tuning subset of data records. That first stage yields a provisional prediction for the dependent variable which can usually be significantly improved upon by the optimization applied in TOB stage 2.

TOB stage 1involves standard matching and ranking algorithms between an unknown record and the multiple records in the larger training subset. That training subset should typically be comprised of more than about70% of all the data records available. In order to obtain reliable predictions across the entire PODV-value range covered by the dataset, the records included in the tuning subset and the testing subset should be distributed across the full range displayed by the dataset. It is also appropriate for the data records with the minimum and maximum PODV values to be placed in the training subset. These requirements mean that the division of the data records between the data subsets is not conducted randomly, as that might lead to sparse data coverage in certain PODV-value ranges in the training and tuning subsets.

The simple steps of TOB Stage 1 (steps 1 to 10) often generate predictions for the dependent variable of credible but sub-optimal accuracy. This highlights which data records in the training subset should be the focus of more detailed analysis for each data record in the

tuning subset. Stage 1 can often achieve impressive levels of accuracy from highly non-linear input data distributions. TOB stage 2: (steps 11 to 12) applies optimization to refine and tune the predictions derived from TOB stage 1. A comparison of the prediction results from stage 1 and stage 2 can typically reveal the respective contributions of each stage to the accuracy of the final predictions derived. Once the optimized-tuning process is completed, and the optimum tuned values of Q and W_N are established, those values are then applied to generate predictions of the dependent variable for the data records of the testing subset (TOB stage2: steps 13 and 14).

The TOB learning network can be applied using spreadsheets (e.g. Excel workbooks), which is a suitable approach for small to mid-sized data sets. It can also be set up in fully-coded formats or as a hybrid code plus spreadsheet configurations. The spreadsheet and hybrid alternatives have the attraction that the standard built-in spreadsheet optimizers can be exploited (e.g. the generalized reduced gradient, GRG, and evolutionary optimizers of Excel's Solver optimization function). That approach enables the final steps of the TOB Stage 1 and the Stage 2 prediction calculations to be displayed as simple and easily-audited formulas for each data record in the spreadsheet cells. For large datasets it is more efficient to code the TOB algorithm with suitable mathematical coding languages (i.e., Octave, R, Python, MatLab etc.).

To predict ρ from the compiled natural gas dataset (693 data records) evaluated in this study, a hybrid VBA-Excel spreadsheet configuration is used. The TOB subsets (training, tuning and testing) are initially displayed in Excel with some calculations conducted using spreadsheet formula (e.g. statistical metrics for all variables). Visual Basic for Applications (VBA) coding is then used to normalize, rank, and match the data records of the tuning and training subsets (TOB stage 1). The VBA code places the top-ten ranked matches for each tuning subset data record into an Excel sheet cell. This enables the final TOB Stage 2 optimization calculations to be conducted with Excel cell formula, enabling the Solver optimizer(s) (Frontline Solvers. Standard Excel Solver, 2018) to be applied. This approach enhances the transparency and insight to the dataset compared to the fully-coded method.

3. Application of the TOB Learning Network to Predict Gas Density

A TOB network is used to predict ρ from the dataset of experimental measurements performed on Qatar North Field natural gas samples (693 data records)published by Atilhan et al. (2012). The data records cover a temperature range of 250 K to 450 K and a pressure range of 15 to 65 MPa. They also include compositional data for each data record in the form of mole fractions for the 12 components: methane, ethane, propane, 2-methylpropane, butane, 2-methylbutane, pentane, octane, toluene, methylcyclopentane, nitrogen and carbon dioxide. A value range and mean for each variable in the full 693-record data set is provided in Table 1.

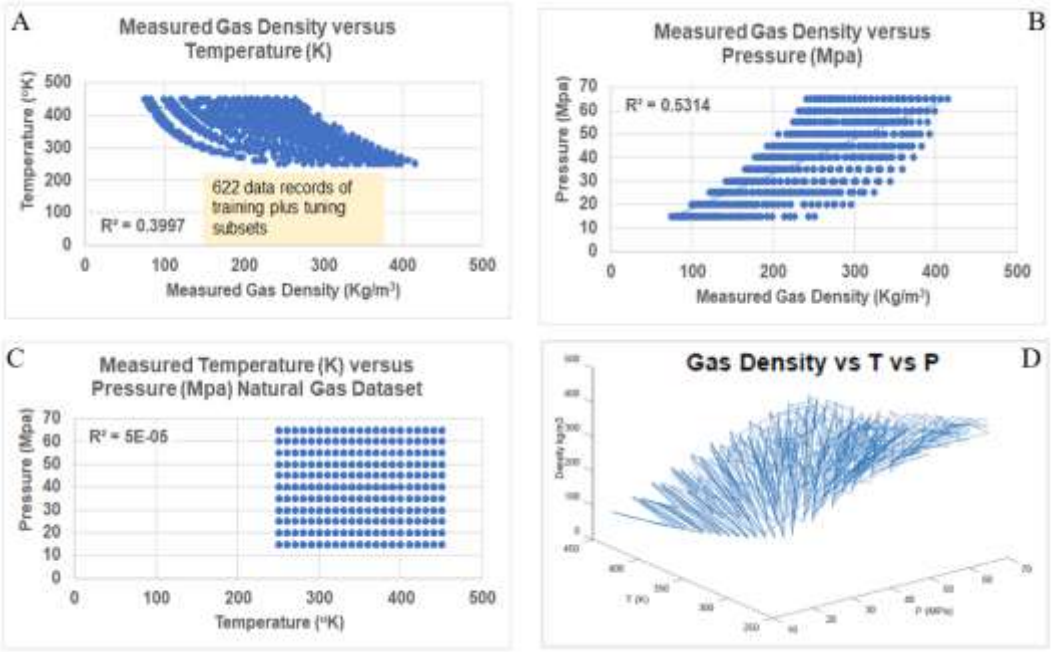
Table 1. Natural gas dataset (Atilhan, Aparicio, Karadas, Hall,& Alcade, 2012) statistical summary of data record values for fourteen input variables with gas density as the dependent variable to which the TOB learning network (Wood, 2018) is applied.

Summary of Dataset Consisting of 693 Data Records				
Input Variable Number	**Mol Fractions:**	**Min**	**Max**	**Mean**
1	**Methane**	0.8034	0.9026	0.85197
2	**Ethane**	0.05189	0.05828	0.05515
3	**Propane**	0.01878	0.02106	0.01997
4	**2- Methyl propane**	0.00384	0.00412	0.00399
5	**Butane**	0.00573	0.00641	0.006
6	**2-Methyl butane**	0.00169	0.00214	0.0019
7	**Pentane**	0.0014	0.00162	0.0015
8	**Octane**	0.00145	0.00161	0.00153
9	**Toluene**	0.0009	0.0011	0.00097
10	**Methylcyclopentane**	0.00095	0.00106	0.00101
11	**Nitrogen**	0	0.06596	0.03364
12	**Carbon Dioxide**	0	0.0438	0.02237
13	**Pressure (Mpa)**	15	65	40
14	**Temperature (K)**	250	450	350
Dependent Variable:	**Density (Kg/m³)**	75.36	415.25	244.928

The dataset is divided into training (532 data records; 77% of the complete dataset), tuning (90 data records; 13% of the complete dataset) and testing subsets (71 data records; 10% of the complete dataset) for detailed TOB network analysis.

The relationships between the key variables, P and T and ρ for the training subset are illustrated in Figures 2 A to D demonstrating the significant non-linearity and irregularity in the relationships among these variables across the entire dataset.

Table 2 and Figures 3 and 4 shows the results obtained from applying the TOB to this dataset: 1) up to step 10 (evenly-weighted-variable contributions to POV prediction) for the configured tuning set; 2) up to step 12 for the optimized tuning set; 3) up to step 12 applying the optimized TOB settings to the testing subset.



Figures 2. (A to D). Pressure, temperature and density relationships in the training set used for the TOB network application.

Table 2. Gas density prediction performance of TOB learning network applied to the 693-record data set showing solutions with a range of variable weightings applied.

Variable Description	Variable Number	Pre-optimization Equal Weightings	Best Solution Solver GRG Multi-start	Best Solution Solver Evolutionary Algorithm	Sensitivity Analysis with Q constrained to integers progressively from 10 to 2 (All cases runs with for the 90 records of the tuning subset with the Solver GRG optimizer configured in the same way)								
Q Constrained to	Integer Constraints		2 to 10	2 to 10	10	9	8	7	6	5	4	3	2
Q selected for solution	Integer #		3	3	10	9	8	7	6	5	4	3	2
Prediction Performance of Optimum and Constrained Optimum Solutions Applied to the Tuning Subset (90 records: ~ 13.0% of total dataset)													
RMSE	Kg/m³	6.8877	**5.5995**	**5.5995**	6.8638	6.7311	6.6134	6.6996	6.3217	6.3711	6.6230	5.5995	7.7695
R²	fraction	0.9939	**0.9954**	**0.9954**	0.9940	0.9941	0.9942	0.9938	0.9942	0.9942	0.9936	0.9954	0.9907
Weightings (0<=w<=1) Applied to constrained optimum solutions for the tuning subset													
Temperature	#14	0.5	**0.09888**	**0.65545**	0.78770	0.66440	0.15029	0.79172	0.44836	0.67680	0.39388	0.46541	0.59351
Pressure	#13	0.5	**0.08734**	**0.57896**	0.62375	0.55941	0.15310	0.85796	0.49286	0.69382	0.38816	0.41110	0.67403
All other variables	#1 to #12	0.5	**N/A**	**N/A**	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Ratio of T weight to P weight			**1.13211**	**1.13210**	1.26284	1.18768	0.98163	0.92280	0.90971	0.97548	1.01475	1.13211	0.88055
Prediction Performance of Optimum Solution Variable Weightings and Q Value Applied to the Testing Subset (71 records: ~ 10.2% of total dataset)													
RMSE	Kg/m³		**5.5023**										
R²	fraction		**0.9965**										

The algorithm was applied to this dataset using the combination of an Excel spreadsheet for steps 10 to 14 (enabling the use of Solver's GLG and evolutionary optimization functions) and VBA code to handle the ranking sorting normalization, record matching, and selection (steps 1 to 8).

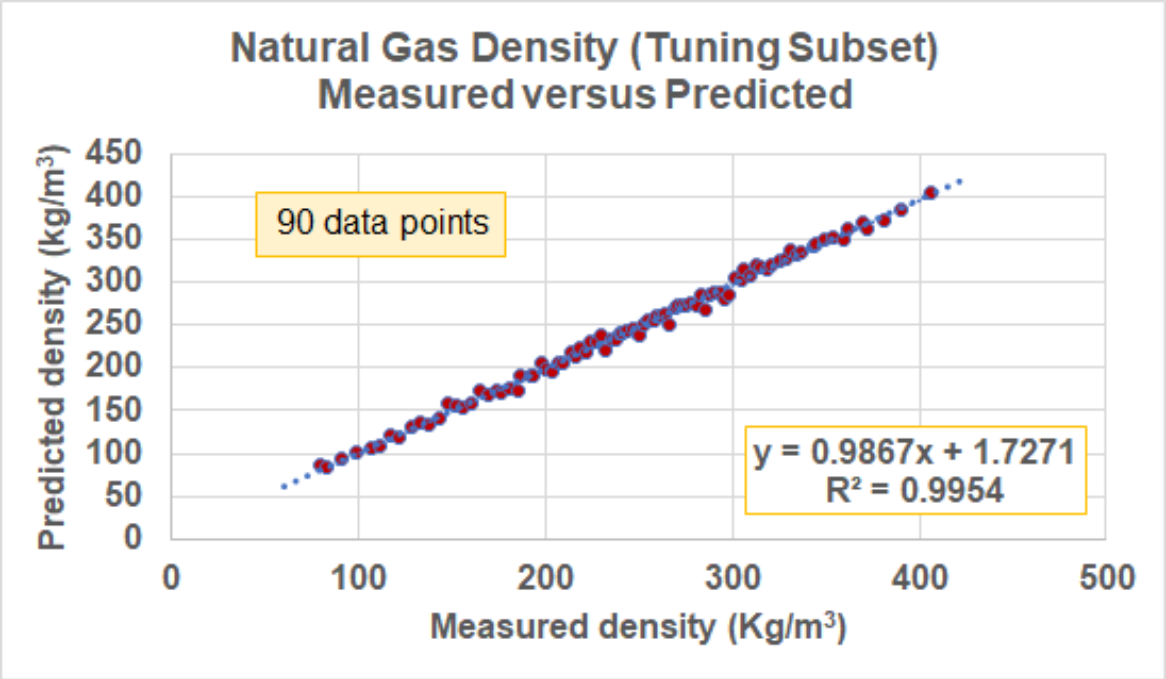


Figure 3. Predicted versus measured gas density for the tuning data set (90 records).

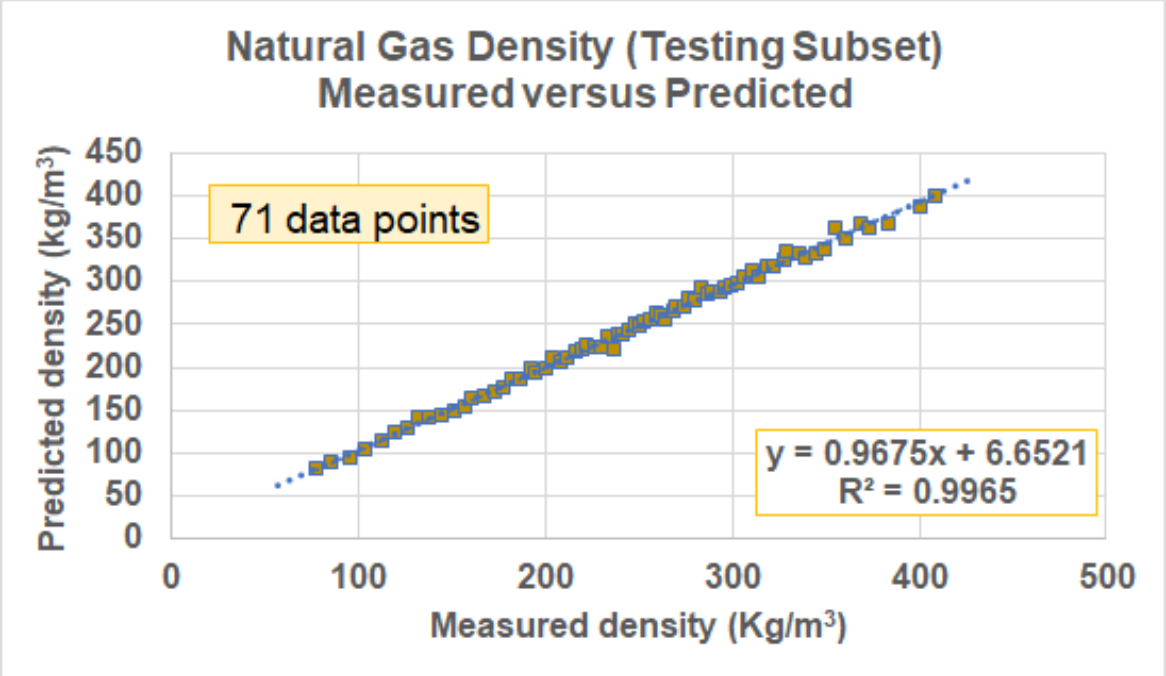


Figure 4. Predicted versus measured gas density for the testing data set (71 records).

The results reveal that the TOB algorithm can achieve very high levels of prediction accuracy (RMSE= 5.6; R^2 = 0.997 for the testing subset). This accuracy is comparable to that achieved by ANN applied to the same data set (Choubineh, Khalafi, Kharrat, Bahreini, & Hosseini, 2017) which reported RMSE = 5.28 and R^2 = 0.995.Both TOB and ANN provide superior ρ predictions than the published correlations applied to the same dataset [Azizi, Behbahani, & Isazadeh, 2010;Sanjari & Lay, 2012). RMSE and R^2 achieved by Azizi et al.'s (2010) correlation were 59.18 and 0.7, respectively. Gas density predicted using that model (Azizi, Behbahani, & Isazadeh, 2010) for the low-density range areis reasonable. However, that correlation significantly overestimates ρ in the higher-density range. Sanjari and Lay's (2012) correlation model, achieved a better gas density prediction performance (RMSE=12.6; R^2=0.97than the Azizi et al.'s (2010) correlation. Although Sanjari and Lay's (2012)model estimates gas density values lower than 340 kg/m^3 with reasonable precision, the values diverge greatly from the unit slope line (y = x) for values in the range of 340–450 kg/m^3,indicating the limitations of that model in that density range.

The proposed TOB model achieves a high level of prediction accuracy while also being able to display the exact prediction calculation for each record in each subset (i.e., identify which of the top-ranking matching records are involved and their fractional contributions to the prediction value). The algorithm achieves most of this in steps 1 to 10 (i.e., TOB stage 1) for this data set (achieving R^2 = 0.9939 for equal weighting applied to all 14 of the input variables for the tuning subset, Q=10 and without the use of an optimizer). Reviewing the prediction details of each record shows that the only variables impacting the prediction based on using the high-ranking matches are T and P. The matching and ranking of the squared errors has removed the impact of the mole fractions of the individual gas components from the final optimized (TOB stage 2) prediction calculation. The mole fractions of the gas components play an important role in selecting the top-10 ranking records in the training subset for each tuning subset data record in the credible TOB stage 1 provisional prediction. However, in TOB stage 2 the optimizer takes those top-10 matching records for each stage 1 prediction and refines them by varying the weights it applies to P and T while applying zero weight to the mole fractions of the gas components. By doing so it slightly improves on the TOB stage 1 prediction.

Table 2 highlights the results of optimization and sensitivity analysis by varying the Q value from 10to 2. The optimum Q value for this data set is 3 (yielding the minimum RMSE value). For values of Q below 3, the accuracy of the model is impaired slightly, suggestive of under fitting. The impact of varying Q on the predictions is caused by subtle changes in the weightings applied to variables T and P. As it happens for this data set the optimum weightings for these two variables in gas density prediction are close in relative magnitude (e.g. 0.5:0.5). That is why the TOB stage 1 provisional prediction managed to achieve such high prediction accuracy because it applied 0.5 weightings to all 14 input variables.

As shown in Table 2 the ratio of the weightings for T and P (w_T/w_P) varies for the optimum solutions associated with different Q values. For Q = 2, 5 to 8 that weightings ratio is less than 1. For other Q values, up to 10 it is greater than 1. The optimum value of w_T/w_P is 1.13211 for Q=3. This is very specific and useful information about how the TOB algorithm is making its optimum predictions for this data set. On the high-ranking matched records (top 3, when Q=3) it is using weights only for T and P and it is doing so in a ratio of 1.13211 to achieve the optimum prediction accuracy. Such insight to the prediction calculation cannot be readily revealed by ANN, ANFIS, SVM and LSSVM algorithms. Future work is planned to compare the performance of TOB with these other machine-learning algorithms for the prediction gas density and other complex oil and gas system.

On the positive side, the TOB methodology provides transparency to the specific calculations involved in each prediction it makes and it achieves credible levels of prediction accuracy. On the negative side, TOB cannot extrapolate its predictions beyond the minimum and maximum values of the dependent variable covered by data records in the training subset, which many other machine-learning algorithms can do. It also cannot achieve highly accurate predictions in sparsely populated regions of a training sunset. This is not necessarily a bad limitation as it inhibits the algorithm from over fitting sparse datasets; a criticism often leveled against other machine learning algorithms. We believe that these attributes make the TOB algorithm a complementary addition to the

suite of existing machine-learning algorithms and justify its use in conjunction with other machine learning algorithms to provide greater transparency to the prediction process.

4. Conclusions

The Transparent Open-Box (TOB) learning network provides a valuable tool for evaluating and deriving predictions from complex, non-linear natural-gas systems. It offers advantages and complementary capabilities to the more-traditional machine-learning networks in that:

- all its intermediate calculations and relationships are fully auditable and accessible;
- relative weightings applied to variables in optimized solutions are clearly revealed;
- it performs well with standard optimizers (e.g., Excel's Solver options) and can be also be linked easily to customized optimization algorithms;
- varying its Q-factor values readily identifies under-fitted versus over-fitted solutions

We recommend that what can be achieved in terms of prediction-performance accuracy by the transparent open-box network algorithmshould be usefulas a performance benchmark when applying less-transparent machine-learning algorithms to specific datasets.

There are many natural gas datasets to which the TOB (e.g. PVT, drilling, well-log data, reservoir and source rock analysis) learning network could be readily applied and provide more enlightening analysis and predictions than the black boxes currently applied to them.

References

AlQuraishi, A. A., & Shokir, E. M. (2009). Viscosity and density correlations for hydrocarbon gases and pure and impure gas mixtures. *Petroleum Science and Technology*, *27*(15), 1674-1689. https://doi.org/10.1080/10916460802456002

AlQuraishi, A. A., & Shokir, E. M. (2011). Artificial neural networks modeling for hydrocarbon gas viscosity and density estimation. *Journal of King Saud University-Engineering Sciences*, *23*(2), 123-129. https://doi.org/10.1016/j.jksues.2011.03.004

Atilhan, M., Aparicio, S., Karadas, F., Hall, K. R., & Alcalde, R. (2012). Isothermal PρT measurements on Qatar's North Field type synthetic natural gas mixtures using a vibrating-tube densimeter. *The Journal of Chemical Thermodynamics*, *53*, 1-8. http://dx.doi.org/10.1016/j.jct.2012.04.008

Azizi, N., Behbahani, R., & Isazadeh, M. A. (2010). An efficient correlation for calculating compressibility factor of natural gases. *Journal of Natural Gas Chemistry*, *19*(6), 642-645. https://doi.org/10.1016/S1003-9953 (09) 60081-5

Broomhead, D. S., & Lowe, D. (1988). Radial basis functions, multi-variable functional interpolation and adaptive networks (No. RSRE-MEMO-4148).Royal Signals and Radar Establishment Malvern (United Kingdom).

Choubineh, A., Ghorbani, H., Wood, D. A., Moosavi, S. R., Khalafi, E., & Sadatshojaei, E. (2017). Improved predictions of wellhead choke liquid critical-flow rates: modelling based on hybrid neural network training learning based optimization. *Fuel*, *207*, 547-560. DOI information: 10.1016/j.fuel.2017.06.131

Choubineh, A., Khalafi, E., Kharrat, R., Bahreini, A., & Hosseini, A. H. (2017).Forecasting gas density using artificial intelligence.*Petroleum Science and Technology*, *35*(9), 903-909. https://doi.org/10.1080/ 10916466.2017.1303712

Cortes, C., & Vapnik, V. (1995). Support-vector networks.*Machine learning*, *20*(3), 273-297.doi:10.1007/BF00994018.

Dehaghani, A. H. S., & Badizad, M. H. (2017).A soft computing approach for prediction of P-ρ-T behavior of natural gas using adaptive neuro-fuzzy inference system. *Petroleum*, *3*(4), 447-453. https://doi.org/10.1016/j.petlm.2016.12.004

Elkatatny, S., Tariq, Z., & Mahmoud, M. (2016).Real time prediction of drilling fluid rheological properties using artificial neural networks visible

mathematical model (white box).*Journal of Petroleum Science and Engineering*, *146*, 1202-1210.https://doi.org/10.1016/j.petrol.2016.08.021

El-M Shokir, E. M. (2008).Novel density and viscosity correlations for gases and gas mixtures containing hydrocarbon and non-hydrocarbon components. *Journal of Canadian Petroleum Technology*, *47*(10). https://doi.org/10.2118/08-10-45

Elsharkawy, A. M. (2003). Predicting volumetric and transport properties of sour gases and gas condensates using EOSs, corresponding state models, and empirical correlations. *Petroleum science and technology*, *21*(11-12), 1759-1787. https://doi.org/10.1081/LFT-120024560

Esfahani, S., Baselizadeh, S., & Hemmati-Sarapardeh, A. (2015). On determination of natural gas density: least square support vector machine modeling approach. *Journal of Natural Gas Science and Engineering*, *22*, 348-358. https://doi.org/10.1016/j.jngse.2014.12.003

Farzaneh-Gord, M., Khamforoush, A., Hashemi, S., & Pourkhadem, H. (2010).Computing thermal properties of natural gas by utilizing AGA8 equation of state.*International Journal of Chemical Engineering and Applications*, *1*(1), 20-24. ISSN: 2010-0221

Frontline Solvers.Standard Excel Solver-Limitations of Nonlinear Optimization. (2018). https://www.solver.com/standard-excel-solver-limitations-nonlinear-optimization

Ghorbani, B., Hamedi, M., Shirmohammadi, R., Mehrpooya, M., & Hamedi, M. H. (2016).A novel multi-hybrid model for estimating optimal viscosity correlations of Iranian crude oil.*Journal of Petroleum Science and Engineering*, *142*, 68-76. https://doi.org/10.1016/j.petrol.2016.01.041

Ghorbani, B., Shirmohammadi, R., Mehrpooya, M., & Hamedi, M. H. (2018). Structural, Operational and Economic Optimization of Cryogenic Natural gas plant Using NSGAII Two-Objective Genetic Algorithm. *Energy*, *159*, 410-428. DOI: 10.1016/j.energy.2018.06.078

Ghorbani, B., Ziabasharhagh, M., & Amidpour, M. (2014).A hybrid artificial neural network and genetic algorithm for predicting viscosity of Iranian crude oils.*Journal of Natural Gas Science and Engineering*, *18*, 312-323. DOI 10.1016/j.jngse.2014.03.011

Haykin, S. (1995). Neural networks: A comprehensive foundation. Pearson Prentice Hall.

Heinert, M. (2008).Artificial neural networks–how to open the black boxes.*Application of Artificial Intelligence in Engineering Geodesy (AIEG 2008), S*, 42-62. ISBN 978-3-9501492-4-1.

Hush, D. R., & Horne, B. G. (1993). Progress in supervised neural networks. *IEEE signal processing magazine*, *10*(1), 8-39. DOI: 10.1109/79.180705

Jang, J. S. (1993). ANFIS: adaptive-network-based fuzzy inference system. *IEEE transactions on systems, man, and cybernetics*, *23*(3), 665-685. doi:10.1109/21.256541

Jang, J. S. R., Sun, C. T., & Mizutani, E. (1997).Neuro-Fuzzy and Soft Computing – Prentice Hall.

Kleene, S. C. (1956). Representation of Events in Nerve Nets and Finite Automata ", Shannon and Mc Carthy (éds): Automata studies.

Sanjari, E., & Lay, E. N. (2012).An accurate empirical correlation for predicting natural gas compressibility factors.*Journal of Natural Gas Chemistry*, *21*(2), 184-188. https://doi.org/10.1016/S1003-9953(11)60352-6

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85-117.https://doi.org/10.1016/j.neunet.2014.09.003

Shirmohammadi, R., Ghorbani, B., Hamedi, M., Hamedi, M. H., & Romeo, L. M. (2015).Optimization of mixed refrigerant systems in low temperature applications by means of group method of data handling

(GMDH).*Journal of Natural Gas Science and Engineering*, *26*, 303-312. https://doi.org/10.1016/j.jngse.2015.06.028

Shirmohammadi, R., Soltanieh, M., & Romeo, L. M. (2018). Thermoeconomic analysis and optimization of post-combustion CO2 recovery unit utilizing absorption refrigeration system for a natural-gas-fired power plant. *Environmental Progress & Sustainable Energy*, *37* (3), 1075-1084. doi:10.1002/ep.12866

Vapnik, V. N. (2000).The nature of statistical learning theory. Springer-Verlag New York.

Wood, D. A. (2018). A transparent Open-Box learning network provides insight to complex systems and a performance benchmark for more-opaque machine learning algorithms. *Advances in Geo-Energy Research,2 (*2*)*, 148-162.doi: 10.26804/ager.2018.02.04.