MEDWIN PUBLISHERS
Committed to Create Value for Researchers

Petroleum & Petrochemical Engineering Journal
ISSN: 2578-4846

Enhanced Prediction of Liquid Loading in Gas Wells Using the Group Method of Data Handling: Comparative Analysis with Machine Learning Algorithms

Mgimba M*

Mbeya University of Science and Technology, Tanzania

***Corresponding author:** Melckzedeck Mgimba, Mbeya University of Science and Technology, Tanzania

Perspective Article
Volume 10 Issue 2
Received Date: May 06, 2026
Published Date: May 25, 2026
DOI: 10.23880/ppej-16000416

Abstract

Despite significant global demand, natural gas production is impeded by liquid loading, which restricts gas flow and diminishes well performance. Traditional mathematical models and limited machine learning applications have struggled to accurately predict and manage this issue, often requiring manual adjustments and offering subpar results. In response, this work develops and applies the Group Method of Data Handling (GMDH) approach to predict liquid loading and the critical gas rate more effectively. The GMDH method outperforms other algorithms like Artificial Neural Networks (ANN), Support Vector Machines (SVM), decision trees, and Naïve Bayes in predicting liquid loading, demonstrating higher accuracy (0.956), precision (0.971), and recall (0.917). While ANN and SVM show commendable performance, Naïve Bayes lags notably behind with significantly lower accuracy and recall rates. For predicting the critical gas rate, GMDH shows the lowest Root Mean Square Error (RMSE) at 43.7472 Mscf/D and the highest coefficient of determination (R2 of 0.9161), indicating a robust model compared to ANN and SVM. GMDH's efficacy is further enhanced by its ability to automatically select the best model parameters, such as neurons and layers, optimizing predictions without manual tuning. This application of GMDH not only advances the prediction of liquid loading in gas wells but also provides actionable insights for experts to anticipate and mitigate this challenge effectively, thereby enhancing gas production efficiency and well sustainability.

Keywords: Group Method of Data Handling (GMDH); Artificial Neural Network (ANN); Support Vector Machine (SVM); Decision Tree; Naïve Bayes; Liquid loading

Highlights

- Natural gas production struggles with liquid loading, which impedes gas flow and diminishes well performance, despite extensive global demand.
- Innovative GMDH methods significantly enhance liquid loading predictions, showing superior accuracy and precision over traditional models and other machine learning algorithms.
- Among tested algorithms, GMDH outperforms ANN, SVM, decision tree, and Naïve Bayes in predicting liquid loading with notable statistical metrics such as high accuracy and low RMSE for critical gas rates.
- GMDH not only yields the most effective predictions for liquid loading but also offers automatic tuning of model parameters, presenting a practical solution for gas well

CC BY

management.

Introduction

The demand for natural gas in the world is huge, based on the International Energy Agency (IEA); despite the war between Russia and Ukraine, which reduced the availability of gas, the energy demand is predicted to increase in the year between 2022 and 2026 by 1.6% [1]. Therefore, more gas is required to be produced to fill the gap available, but there are some challenges that hinder gas production, one of them being liquid loading [2]. Liquid loading happens when the velocity of a gas is less compared to critical velocity, at that velocity, the gas fails to transport liquid, hence it accumulates in some parts of the pipe. The build-up of liquid in some part of the pipe or below the gas well hinders gas flow and thus reduces gas production.

The liquid in the gas well originates from the reservoir fluid (natural gas and water vapour) due to the variation in temperature and pressure [3]. The drop-in temperature at constant pressure or increase of pressure at constant temperature can cause the water vapour to alter to liquid water, as illustrated in Figure 1 [4].

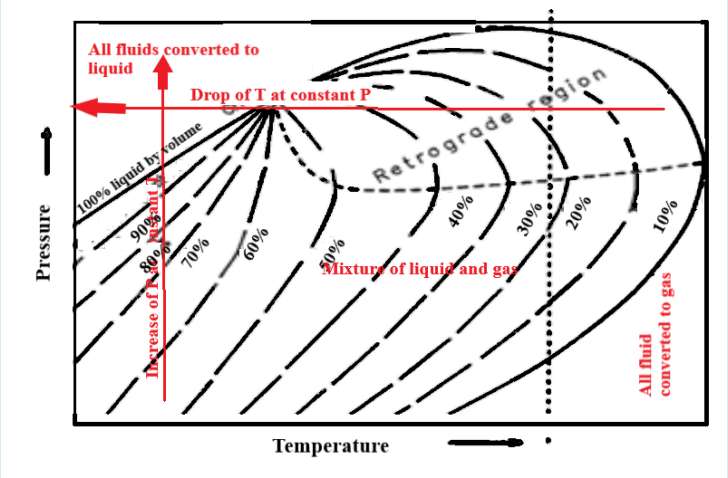


Figure 1: The Impact of Temperature and Pressure Changes on the Development of Multiphase Flow [4].

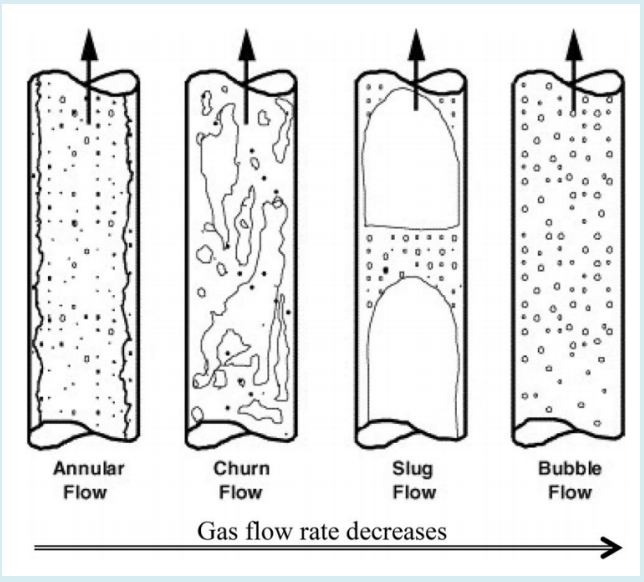


Figure 2: The Alteration of Flow Pattern when the Gas Flow Rate Decreases [5].

Based on Adaze, et al. [5], the ocuurence of liquid loading is influenced by the lower gas flowing rate. The gas flowing rate is mainly influenced by the reservoir pressure. Thus, at the beginning of the production, the reservoir pressure is high enough to lift the liquid drops to the surface. But as the production continue, the reservoir pressure is dropping causing the flowing rate to be lower enough to lift the water drop to the surface. The drop in reservoir pressure and gas flowing rate causes changes in the flow pattern from annular flow which is dominated by gas to bubble flow which is dominated by liquid, as depicted in Figure 2. Therefore, the liquid loading happens in the bubble flow pattern when the flowing rate is lower enough to lift the liquid.

When the liquid is produced for a long time or produced in large quantities, it can accumulate in some parts of the pipe [6]. The accumulation of liquid water (liquid loading) in the gas production pipe can cause friction to the flowing gas and hence limit the gas flow, reduce the efficiency of gas transportation, and affect the gas production economy [7]. The occurrence of a gas well's liquid loading causes a

Copyright© Mgimba M.

reduction in gas flow rate, and when the situation is critical, a well can be abandoned [8]. According to Alsanea, et al. [9], the gas well's liquid loading can decline or kill the gas well production. Based on He, et al. [10], the development of liquid water in the gas production pipe will cause the development of multiphase flow, which affects the pressure decrease in the production pipe and then affects the flow of gas in the pipe. Due to these facts, predicting and controlling liquid loading is essential for the longer production of the gas well and the economy of producing natural gas.

Many scholars have studied the gas well's liquid loading; for instance, Chen, et al. [11] studied the two-phase flow (liquid and gas) in gas wells. They identified that in the well with the two-phase flow, there are interaction forces between gas and liquid interphase and external forces between the production pipe wall and fluid. These kinds of forces limit the gas flow in the production pipe causing decreasing gas production. Through experiments, researchers studied the gas flow pattern in the pipe to evaluate how and when the liquid loading was formed. With the aid of those flow patterns, some scholars developed mathematical models that can be used to predict the occurrence of liquid loading [12].

One of the models developed to predict liquid loading is the Turnel droplet model, which was established by Turner, et al. [13]. This model considers the formation of liquid loading depending on the forces that are acting on the liquid drops. The liquid drop is being acted upon by two forces: drag force, which pushes the drops upward, and gravitational force (caused by the weight of the drop), which pushes the drop downward. The liquid drops will move out of the well when the drag force is higher than the gravitational force. But when the gravitational force is higher than the drag force, liquid drops will fall and fail to move out of the well [14]. Based on the Turnel droplet model, the terminal velocity, which estimates the velocity capable of carrying liquid drops out of the well, can be evaluated using Equation (1) [15]. Based on this Equation, the terminal velocity depends on the square root of the liquid density, gas density, and drop diameter. The terminal velocity indicates the velocity at which the drag force is equal to the gravitational force. When the flow rate of the well is above terminal velocity, then liquid loading will not happen; otherwise, the well is in danger of incurring liquid loading.

$$V_t = 3.617\sqrt{\frac{d(\rho_L - \rho_g)}{f C_d \rho_g}} \quad (1)$$

Whereas V_t is the terminal velocity, d is the drop diameter, ρ_L is the density of the liquid, ρ_g is the density of the gas and C_d is the drag coefficient.

The Equation (1) was improved by considering interfacial tension in evaluating critical velocity (), as shown in Equation (2). The Turnel model was applied to predict liquid loading in the field using Equation (2); the result shows that from the 90 wells that were tested, 66 wells' results were correct [16].

$$V_{ct} = 5.465\frac{\left(\sigma\left(\rho_L - \rho_g\right)\right)^{\frac{1}{4}}}{\rho_g^{\frac{1}{2}}} \quad (2)$$

Despite the good performance of the Turnel model, some shortcomings were observed. Some of the shortcomings were that the model does not work better for a well with a low gas rate and wellhead pressure of less than 500 psi [15] and neglects the occurrence of flow patterns [17]. Furthermore, the model takes into account the shape of liquid drops as a sphere, but at a high flowing rate, the shape becomes an ellipse; this affects results when predicting liquid loading.

Due to the limitation of the Turnel droplet model, Turner, et al. [13] introduced another model called the Turnel film model. The model considers the movement of liquid film as it moves in the pipe. But also, the model considers the flow patterns which are ignored in the Turnel droplet model. The idea of the film model was supported by other scholars, including Luo, et al. [13], Van't Westende [18], Veeken, et al. [19], who observed that liquid loading is evaluated by considering fluctuations in the flow patterns in the pipe rather than considering the movement of liquid drops only.

Apart from the model established by Turner, et al. [13], there are other various models that were established to predict the critical gas velocity/ flow rate for a liquid loading to happen, as illustrated in Table 1. For instance, Duns and Ros [20] introduced the dimensionless parameter (Y) to predict the onset of liquid loading. This parameter incorporates the gravitational and interfacial forces, along with the dimensions and properties of the wellbore. Liquid loading is predicted to occur when (Y) exceeds a critical value, typically between 10 and 12 for most applications. Abhulimen, et al. [21] developed a model to predict the critical velocity for liquid loading in the multiphase flow in pipelines. This model provides a simplified empirical formula to estimate the critical gas velocity required to prevent liquid accumulation based on gravitational forces and fluid properties. Furthermore, Beggs and Brill [22] introduced an empirical correlation that were used to predict liquid holdup for different flow pattern. This correlation is applicable for two phases flowing in the horizontal flow. Another method to predict the critical velocity for liquid loading was introduced by Dukler's method (Dukler-Eaton correlation). This method

Copyright© Mgimba M.

was developed in 2017 and has since been widely applied in the oil and gas industry. Dukler's method is based on extensive experimental data and provides a correlation to estimate the critical gas velocity (V_{crit4}) at which liquid loading begins to occur in gas wells. This correlation considers various factors including fluid properties, wellbore geometry, and operational [23].

Many other correlations were developed to predict critical gas velocity for the liquid loading to happen. Guo, et al. [24] modified empirical correlation that accounts for the effects of multiphase flow regime transitions (such as slug flow) on liquid loading prediction. They integrate empirical adjustments based on observed flow patterns in gas wells. Moreover, Ansari, et al. [25] incorporated machine learning techniques to develop predictive models for liquid loading onset. They used historical field data combined with machine learning algorithms to improve accuracy and account for complex interactions between fluid properties, flow dynamics, and wellbore conditions.

Methods	Mathematical Models	Citations
Duns and Ros Method of 1963	$Y = \frac{\rho_L - \rho_g}{\rho_g} * \frac{\sigma}{\rho_g D} * \frac{g}{\rho_g}$	[20]
Kabir and Hasan Method of 1990	$V_{crit1} = 3.1\left[\frac{\sigma g\left(\rho_L - \rho_g\right)}{\rho_g{}^2}\right]^{0.25}$	[26]
Berg and Brill Method of 1973	$H(S) = \frac{0.98 \wedge^{0.4846}}{N^{0.0868}}$ $H(I) = \frac{0.845 \wedge^{0.5351}}{N^{0.0173}}$ $H(D) = \frac{1.065 \wedge^{0.5824}}{N^{0.0609}}$	[22]
Dukler's Method (Dukler-Eaton Correlation) of 2017	$V_{crit4} = \left[\frac{4K}{3C_d}\right]^{\frac{1}{4}}\left[\frac{\sigma g\left(\rho_L - \rho_g\right)}{\rho_g^{\frac{1}{2}}}\right]^{\frac{1}{4}}$	[23]
Abhulimen, et al. [21]	$Q_g = 5.2759\frac{P \times A}{Z}\left[c\left[W_1\frac{(67 - 0.0031P)^{\frac{1}{4}}}{\sqrt{0.0031P}}\right] + b\left[W_2\left[\frac{(45 - 0.0031P)^{\frac{1}{4}}}{\sqrt{0.0031P}}\right]\right]\right]$	[21]
Guo and Ghalambor Method of 2005	$V_{crit6} = C_5\sqrt{\frac{\rho_L - \rho_g}{\rho_g} * \frac{\sigma}{\rho_g D}}$	[24]
Ansari et al. Method of 2018	V_{crit7} = f(variables, machine learning model)	[25]

Table 1: Correlations that were developed to predict critical gas velocity for liquid loading.

Several researchers have studied liquid loading by applying or adjusting these models. Riza, et al. [27] adjusted the droplet model by adding 20% of the Webber number. The results improved, whereas 77 out of 90 wells had a correct prediction of liquid loading. Enos [14] conducted a study by applying a 20% adjustment of the Turnel droplet model. The results were 100% efficient in predicting unloaded wells and 51% in predicting loaded wells. Nallaparaju [12] predicted the gas well's liquid loading by integrating the Turnel model, IPR and TPR curves. Through that study, Nallapuraju [12] to predict the time at which a liquid loading will happen in a gas well. According to Belfroid, et al. [28], the inclination angle highly influences the prediction of liquid loading. It was found that an inclination angle of 500 has a higher critical

Copyright© Mgimba M.

velocity (more than 40%) when compared to a vertical well. But Pagou, et al. [29] also utilised a film model to estimate the liquid loading using data available in some articles. In that work, they take into account the cause of void ratio, deviation angle, and well diameter. It was found that inclination angle and tubing diameter have a high influence on the liquid loading prediction. Therefore, the models to predict the liquid loading show good forecasts of liquid loading despite some shortcomings like underestimating the critical velocity in the well producing under loaded conditions, failure to consider multiple factors which have many effects, and the accuracy of prediction is not enough. Due to these shortcomings, some authors suggest using machine learning to predict liquid loading [25].

Liquid loading in horizontal gas wells impairs gas production and if not diagnosed in a timely manner can kill the well. Liquid loading occurs when the gas production rate declines and gas velocity drops below the critical velocity required to carry liquid to the surface. Different models used in conventional reservoirs such as droplet, film or transient multiphase flow models are also applied with modifications in unconventional gas reservoirs. However, none of these models show great success when applied to inclined and horizontal wells in shale gas reservoirs. This is due to the fact that these models are developed for vertical wells and cannot identify the right multiphase flow regime in inclined and horizontal sections of the well. It is also extremely hard to define the right liquid droplet size and shape or liquid film thickness as well as trajectory changes. Furthermore, these models cannot accurately predict the transient between annular and slug flow regimes in horizontal wells. As more wells are produced in shale gas reservoirs, a great amount of information from production control and monitoring becomes available which can be used to build a data-based smart model for real-time diagnostics of liquid loading in new wells. In this new approach, data (pressure, completion, and productions) from many wells that have experienced or have not experienced liquid loading problems in the same area will be the basis for developing the smart model.

Machine learning (ML) techniques have been applied to different engineering problems. Among the pieces that made use of machine learning techniques in forecasting pore pressure was demonstrated by Hutomo, et al. [30]. Through the integration of Eaton methods and artificial neural networks (ANNs), Hutomo, et al. [30] estimated pore pressure. In that work, the pore pressure was predicted using seismic data, which included seismic frequency, seismic amplitude, seismic amplitude, and seismic shear impedance. The compressive strength of the paste was found experimentally and then predicted by using ANN and particle swarm techniques. The accuracy of the techniques was evaluated based on the lowest mean square error (MSE) and correlation coefficient (R) [31]. Ahmed, et al. [32] used drilling parameter data and well logs to anticipate the pore pressure using an artificial neural network. He, et al. [33] used depth, gamma-ray interval transit time, and formation density to estimate the fluid pressure in the reservoir using a backpropagation (BP) neural network. The project's pore pressure projection had a 4.61% forecast inaccuracy. Keshavarzi R and Jahanbakhshi [34] used the Eaton technique and backpropagation artificial neural network (BPANN) to forecast the pore pressure gradient in the Asmari oil field in Iran. When compared to the Eaton approach, the BPANN performs better. Also, the ML techniques have been used to forecast reservoir parameters such as permeability [35], hydrocarbon content like TOC [36], sources rock potential (S1 and S2) [37] and mechanical properties of the formation like brittleness index [38].

Apart from applying ML techniques in other petroleum fields, other scholars have applied ML techniques to predict the flow of liquid in the pipe, flow pattern, and liquid hold-up [39]. For instance, Kanin, et al. [39] utilised an ML algorithm to study multiphase flow, whereby flow pattern, liquid hold-up and pressure gradient were evaluated. This study discovered that flow pattern depends on field flow rate, fluid properties, and the shape and location of the pipe. Also, Lin et al. deep learning neural networks to predict the flow patterns in the inclined pipes from velocity and angle of inclination. Furthermore, ML algorithms were introduced by Hog, et al. [40] to study liquid loading on the pipe that transports natural gas. The ML algorithms used in this work include K nearest neighbour algorithm (KNN), decision tree (DT), plain Bayesian classification (NBC), random forest (RF), ANN and support vector machine (SVM). The prediction in this work was made from the data generated from the simulation's software, OLGA. Through that software, different transportation conditions, including pressure, gas velocities, pipe size, and pipe properties, were involved in predicting the liquid loading. In this study, KNN and RF show the best performance compared to the remaining algorithms. However, this study does not involve real field data to evaluate the liquid loading. Ansari, et al. [25] applied ML methods to predict liquid loading. In that work, they employ a fuzzy pattern recognition method and unsupervised techniques to study liquid loading for the Marcellus reservoir. A total of 160 wells were employed for this task, whereas, 70% of the data were used for training, 15% for calibration, and 15% for testing. The findings of this study show great success with high reliability and accuracy. Recently, Abhulimen [21] studied liquid loading using ML techniques, including genetic algorithm, ANN, particle swarm optimisation, and Bayesian neural network. In this study, ANN provided an accuracy of 92 and 93%; particle swarm optimisation provided an accuracy of 83 and 92%; the genetic algorithm provided an accuracy of 83 and 89%, and the Bayesian neural

Copyright© Mgimba M.

network provided an accuracy of 73 and 77% for training and testing data, respectively. All ML algorithms show good performance; however, more accuracy is needed, and there is manual modification of the parameters like neurons, biases, and weight [41].

To increase the accuracy and overcome the mentioned problem, scholars introduced another ML algorithm called the group method of data handling (GMDH). The GMDH algorithm has the ability to overcome the problems of manually adjusting model parameters and improving the accuracy of prediction [42]. But also this technique succeeds in other directions of petroleum fields, including forecasting pore pressure [41], permeability [43], TOC [37,44], standpipe pressure [45] and formation temperature [46]. Thus, for the initial instance, we utilised the GMDH algorithm to forecast the liquid loading from the field data. The GMDH algorithm will be used to classify the gas flow in the production pipe and determine if there is liquid loading. Also, the GMDH will be applied to predict the critical gas rate for liquid loading to happen. The results of GMDH will be compared with other ML algorithms which have been used in previous works, including ANN, decision tree, Naïve Bayes, and SVM. Through this study, new knowledge of applying GMDH in predicting liquid loading and adapting different scenario in the multiphase flow will be added. Hence, knowledge of predicting liquid loading using ML algorithms, which are not covered well, will be expanded. The success of the technique will also help to know when liquid loading will happen and see how to control it.

The article contains fives section, starting from introduction in the first section and then methodology on the second section. Section three of these article covers the results and discussions. This section present all results which are predict by all machine learning techniques. Then, in section four some limitations of this study are mentioned and the conclussion of the article is presented in section five.

Methodology

Different ML techniques were applied to reach the goal of this research. Among the ML techniques applied were artificial neural network (ANN), group method of data handling (GMDH), Decision tree method, support vector machine (SVM), and Naïve Bayes. The techniques were used for the actual field data presented in the articles of Turner, et al. [16] and Coleman, et al. [15].

The Correlations Analysis

Before applying ML algorithms, the correlation analysis between the input and output data was studied. Therefore, the Pearson correlation was used to find the relationship between input and output data. The Pearson correlation was evaluated using Equation (4), where x_i is the value of each input data, x_{av} is the average value of input data, is the value of liquid loading and y_{av} is the average value of liquid loading data.

$$r = \frac{\sum(x_i - x_{av})(y_i - y_{av})}{\sqrt{\sum(x_i - x_{av})^2 \times \sum(y_i - y_{av})^2}} \quad (4)$$

The Pearson correlation is frequently employed because it is particularly adept at detecting relationships between linearly related variables. The three correlations exhibit a positive link when their values are closer to +1, an inverse relationship when their values are more relative to -1, and no association when their values are closer to 0. This method provides a simple means of identifying multicollinearity. The obtained attributes were utilised to generate additional predictions employing diverse machine learning models.

Artificial Neural Network (ANN)

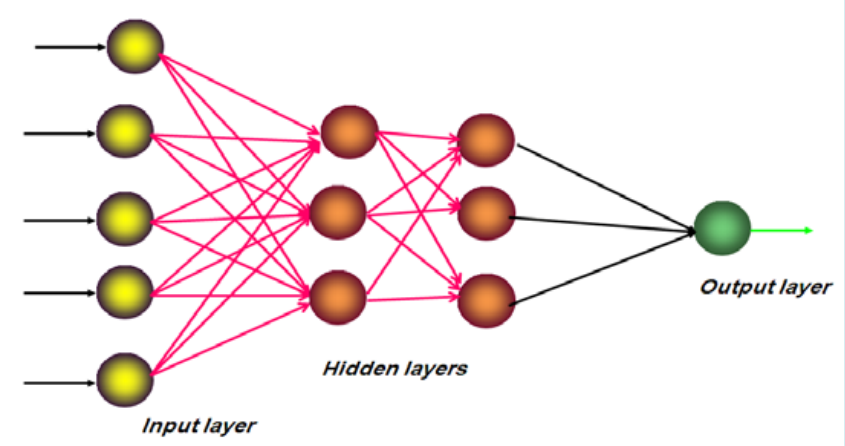
For non-linear system identification, this ANN technique is widely used [47]. The methodology has its roots in neurobiology and has been shaped by the structure of the human brain. As a result, it possesses the brain's non-linear qualities and intelligence. Equation (5), where x is the input, w is the weight given to each input, p is the feature bias, and $h_{w,p}(x)$ is the output, is used by the ANNs to predict the output. An artificial neural network (ANN) is made up of several components that are connected and arranged together using weighted networks, often known as neurons [48]. According to the ANN design, the system's neuron produces an output by adding together all its weighted inputs and multiplying it by the inputs in a non-linear activation function.

$$h_{w,p}(x) = f\left(W^T x + p\right) = \frac{1}{1 + \exp\left(-\left(w^T x + p\right)\right)} \quad (5)$$

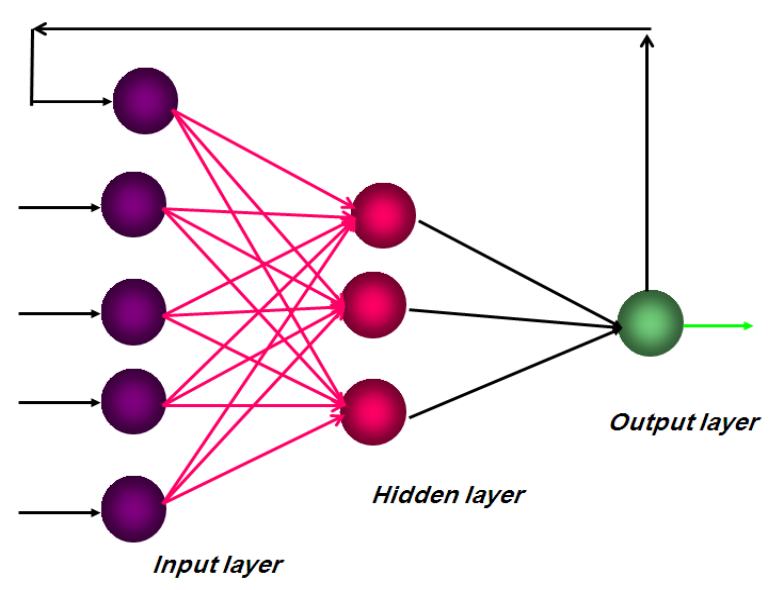
Designs for Neural Networks

There are several classifications of neural networks, each with its own design, applications, and methods. Recurrent and feedforward neural networks are the most often utilized types of neural networks [49]. The input and output layers are part of a feedforward network, and there are hidden layers in between. The input layers and the output layer are connected by these hidden layers [50]. Figure 3 shows a multilayer feedforward network architecture. Like the feedforward network, the recurrent network has a loop that transfers data from the output layer to the input layer or vice versa. This neuronal network is therefore referred regarded as a dynamic network (Figure 3, b).

Copyright© Mgimba M.



(a)



(b)

Figure 3: The Assembly of a Neural Network, (A) Is the Feedforward Network And (B) Is the Recurrent Neural Network Architecture.

Group Method of Data Handling (GMDH)

Another multi-layered, non-linear trained machine technique is called GMDH. The techniques address problems using the polynomial function, and they can handle higher order polynomials without requiring complex mathematical knowledge [47]. It is intended to provide a connection between several inputs and one output. The approach can also determine which inputs have an impact on the result. There are levels in the network, and each layer has two input neurons. The quadratic equation derived from the two inputs of the neurons is used to determine the output of each neuron. The quadratic functions are obtained by the use of linear regression analysis, and the previous layer must be trained before moving on to the next layer. The training procedure selects the best performance from each individual training neuron. The next layer appears and the training is performed after the neural with the best performance is chosen. According to Abu-Kheil [51], layers are added until the stopping requirements are met.

The Ivakhnenko polynomial Equation (6) is used to calculate each neuron's output. Firstly, the pair of variables in each layer is combined by the whole GMDH network. Next, for every group, the polynomial coefficients and

Copyright© Mgimba M.

corresponding output are assessed using the least square fitting technique. Lastly, the findings will be assessed and tested using the external accuracy requirements once the polynomial coefficients have been reviewed.

$$y = f\left(x_1, x_2\right) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 x_1^2 + w_5 x_2^2 \quad (6)$$

Equation (7) is used to analyze the model's correctness. In this equation, RE stands for the regularity criterion measure, N for the amount of samples, P for the necessary outcomes, and y for the GMDH neuron's prediction. Each neuron's output is calculated and tested using the equation. Which input is more pertinent to the network may be determined by looking at the outcomes of this equation. It also illustrates the neuron polynomial's capacity to forecast the necessary output. A smaller response value denotes a better fit, whereas a greater number denotes a worse fit.

$$RE^2 = \frac{\sum_{i=1}^{N}\left(P_i - y_i\right)^2}{\sum_{i=1}^{N} P_i^2} \quad (7)$$

The GMDH automatically selects the number of neurons and layers based on the value of , which is derived from Equation (7). As seen in Figure 4, neurons with a smaller value will be removed, while those with a bigger value will be chosen when the value of falls below a certain threshold value. Furthermore, the layer that has the lowest value is preserved. The new layer will end if the value of in the subsequent layer is greater than in the preceding layer. The ultimate outcomes of the GMDH network will be represented by the output of the neuron with the lowest in the last layer [44].

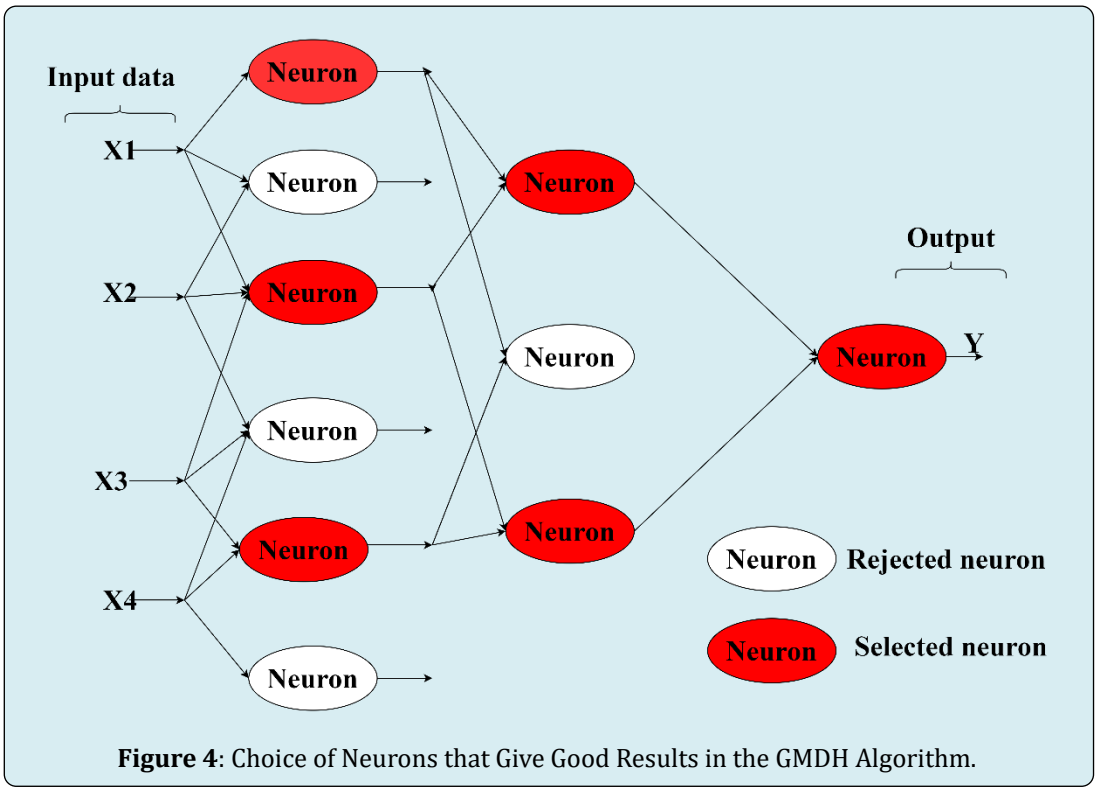


Figure 4: Choice of Neurons that Give Good Results in the GMDH Algorithm.

Applying the GMDH Algorithm

There were various phases involved in putting this idea into practice. Isolating the data into training and testing data was the initial stage. While the testing data were applied to forecast liquid loading, the training data assisted in assessing the weight of GMDH neurons. Thirty percent of the data were testing data, while the remaining seventy percent of the data were the training data. The second step was to use Equation (8), where m is the number of input variables, to identify the combination of the two inputs among all inputs. Equation (9) illustrates how each combination was then enlarged to a quadratic polynomial (Z).

$$\text{combinations} = \frac{\mathrm{m}(\mathrm{m}\text{-}1)}{2} \quad (8)$$

$$Z = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{11}x_{21} & x_{11}^2 & x_{21}^2 \\ 1 & x_{12} & x_{22} & x_{12}x_{22} & x_{12}^2 & x_{22}^2 \\ & & & & & \\ 1 & x_{1N} & x_{2N} & x_{1N}x_{2N} & x_{1N}^2 & x_{2N}^2 \end{bmatrix} \quad (9)$$

Copyright© Mgimba M.

Then, the polynomials' weights were evaluated using Equation (10).

$$w = \left(Z^T Z\right)^{-1} Z^T y \quad (10)$$

Using the chosen data from the testing group, the third stage involved calculating and testing each polynomial's output. Finally, the liquid loading forecast was estimated using Equation (11).

$$y_m = w * Z \quad (11)$$

Support Vector Machine (SVM)

The SVM is a type of ML algorithm applied to solve regression and classification problems; SVMs are mainly good at resolving binary classification difficulties, which need categorising the members of a data set into two classes. The SVM is one of the machine learning methods that works by constructing a border called a hyperplane. This hyperplane is used to separate the data into different groups or classes, as presented in Figure 5 [52]. The format of data or correlation between data will guide the application of this method, with a likely application on linear divisible or nonseparable samples. The vectors of data are evaluated, and one that shows good properties of separating data is chosen to represent the hyperplane. Other vectors that are close to the hyperplane are recognized as support vectors; the length between these vectors is referred to as the margins. The best hyperplane is the one that contains the largest margin since the largest margin indicates the presence of the smallest errors. The SVM can contain a linear or non-linear hyperplane, on which the linear hyperplane, shown in Figure 5, provides fast and good results during its application. However, some data are difficult to separate with linear hyperplanes; thus, the non-linear (2D hyperplane) can be used in those cases [53].

The mathematical hyperplane can be constructed by Equation (12) as shown in Figure 5.

$$w^T \varphi(x_i) + b = 0 \quad (12)$$

The data will be divided when they attain the condition of Equation (13) and Equation (14).

$$w^T \varphi(x_i) + b \geq +1 \quad y_i = +1 \quad (13)$$
$$w^T \varphi(x_i) + b < +1 \quad y_i = -1 \quad (14)$$

Then, Equations (11), (12), and (14) can be combined to form Equation (15). Thus, Equation (15) is used to define the hyperplane which divides the data.

$$y_i(w^T(x_i) + b \geq +1 \quad i = 1....N \quad (15)$$

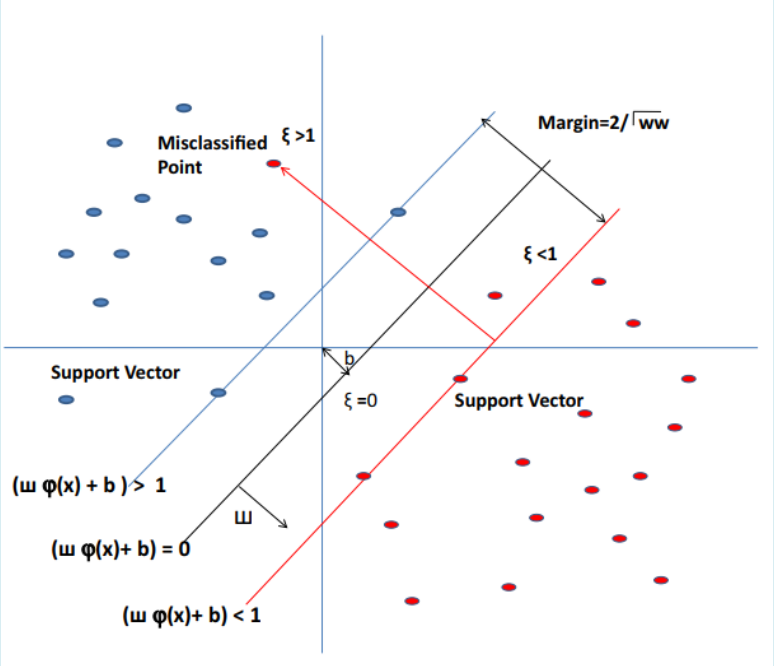


Figure 5: The Application of the SVM Method, Whereas the Hyperplane Separates Different Groups of Data, is Presented [54].

SVMs are also suitable for examining complex data that can't be divided by a simple straight line. For this case, a mathematical model called Kernels is used to generate a hyperplane, which is able to divide separate data. The most

Copyright© Mgimba M.

used kernels in SVM are linear and radial basics function (RBF) Kernel. The linear kernels are defined by Equations (16) [53].

$$k\left(x,x^{'}\right)=x.x^{T} \quad (16)$$

Where x is the input matrix, and x^T is its transpose.

The RBF kernel can be presented by Equation (17). This function is most used for non-linear problems.

$$k\left(x,x^{'}\right)=e^{-\gamma x-x_i^2}\gamma>0 \quad (1)$$

Whereas γ is the parameter for the RBF kernel, which indicates the spread of data.

Classifications by SVM techniques have been applied in different cases in the geosciences, some of them being lithology classification [55,56], modelling of 3D geological features [57] and recognition of the lithofacies of organic matter [58].

Decision Tree Techniques

The decision tree is the method commonly used to separate data into various defined groups. This method used the flow chart to divide the data into separate groups; hence, the whole structure of the division is observed as a tree, which is why the method is referred to as a decision tree [59]. An example of a decision tree can be shown in Figure 6, indicating whether people put air conditioners (RAC) in a room off or on Yu, et al. [60].

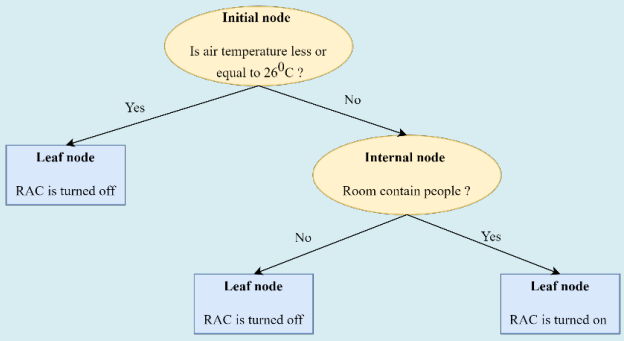


Figure 6: Decision Tree Flow Chart for People Who Turn the Air Condition (RAC) Off or Yu, et al. [60].

The decision tree method involves two processes of learning and classifying the data, as depicted in Figure 7. The gathered data are split into two groups for the learning process: the training data and the testing data. An essential component of assessing data mining models is the making of testing and training data. The majority of data sets in the database are typically chosen at random for training work, whereas the rest of the data is used for testing. Keep in mind that the testing and training sets must be disjoint yet originate from a similar [61,62]. Subsequently, the training data is fed into a decision tree-generating algorithm, which produces the decision tree. Classification and regression trees (CART) [63], C4.5 [64], and ID3 [65] are a few frequently applied decision tree generating methods. Because C4.5 is versatile and can be used with a wide range of data formats, we are using it in this project to create decision trees in conjunction with the open-source data mining program WEKA. Making forecasts against the test data is the first step in the classification process that assesses the correctness of the decision tree that was created. By contrasting the anticipated target values with the actual goal values of the testing data, one can determine how accurate a decision tree is. If the decision tree's accuracy is deemed satisfactory, it can be used to classify and predict fresh datasets; if not, the cause of the issue should be found and appropriate measures taken to address it [66].

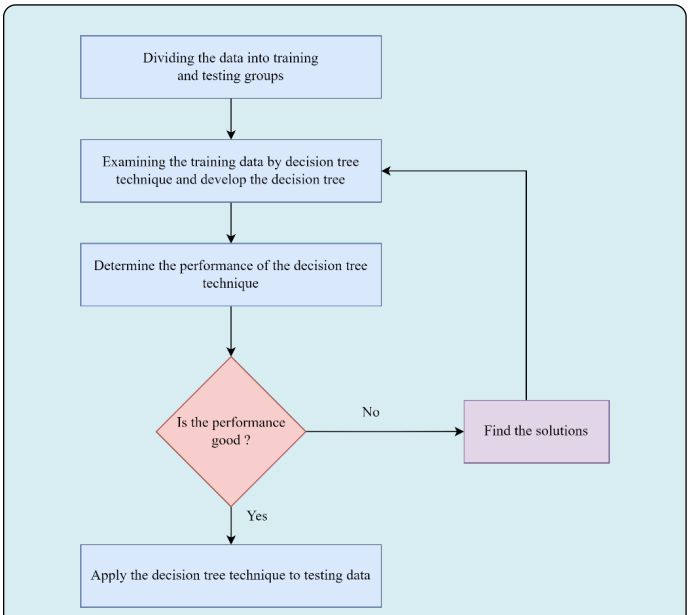


Figure 7: The Procedure for Applying the Decision Tree Method [67].

The following describes the process of creating a decision tree using the training data. The training data's entries are initially compiled into a single partition. The method selects a predictor property that "best" separates the goal class values in the partition at each iteration. The method divides the partition into child partitions so that each child partition has the similar value of the nominated attribute when a predictor attribute is picked. When any one of the following ending conditions is satisfied, the decision tree method ceases splitting the partition iteratively: first

Copyright© Mgimba M.

is when, in a partition, every record has the same target class value; second is when predictor qualities that could be utilised to separate a division further have run out. Lastly, is when there are no more records for certain predictor variable values [61].

The algorithm used to generate decision trees is avaricious. Selecting a split property that may most effectively divide the goal class values splits a partition iteratively. The split attribute selection determines the decision tree model's quality and, consequently, the future data's classification accuracy. A popular criterion measure for decision trees to assess the transparency of a partition in decision tree nodes is the information theory concept of entropy. The data subset's entropy, Di, of the ith tree node is reffered to as follows agreed a decision tree that only contains binary variables for the goal, such as LOW EUI (L) and HIGH EUI (H) (Equation (18)) [61].

$$D_i = -\left(\frac{H}{TN}\log_2\left(\frac{H}{TN}\right)+\frac{L}{TN}\log_2\left(\frac{L}{TN}\right)\right) \quad (18)$$

Whereas H stands for a number of HIGH EUI records in Di, L stands for a number of LOW EUI records in Di, and TN stand for a total number of records in Di.

The range of the entropy is 0 to 1. Observe that while Di is pure, the entropy equals 0, and when H equals L, it equals 1. A candidate splitting test will be employed at each decision tree node to assess all available characteristics and determine which attribute is best for splitting data. Furthermore, a minimal threshold entropy value (ENmin) will be established and compared with the node classification entropy (Entropy (Di)) to determine whether a node should be a leaf. If Entropy (Di) is less than ENmin, the node is a leaf and will be labelled as such. If not, a second splitting test ought to be run. In the event that more candidate splitting tests have no discernible effects on gain ratio or information gain, the test will also be terminated, and the node will be designated as stop [61,64].

Naïve Bayes (NB) Classifier

The NB classifier is one of the simple and fast methods of classifying data. This method is extensively used because it has a high theoretical foundation, high efficiency and good simplification ability [68,69]. The main assumption of the NB techniques is that the variables are conditionally independent when they are classified.

For the certain data group to be categorised, the probability of each group looking under the condition of the existence of this group is resolved, and the group to be categorised fits to the group with the maximum posterior probability. NB classifier techniques are based on Bayes theorem, on which the probability to classify the data can be evaluated by using Equation (19) [11].

$$P(C_j \setminus X) = \frac{P(C_j)P(X \setminus C_j)}{P(X)} = \frac{P(C_j)\prod_{i=1}^{n} P(X_i \setminus C_j)}{P(X)} \quad (19)$$

Whereas P( ) indicates the former probability of the group, indicates the number of features, and P () indicates the conditional probability of the feature.

If the groups have the same P(X), then criterion f(x) is evaluated using Equation (20).

$$f(x) = \text{argmax}\, P(C_j)\prod_{i=1}^{n} P(X_i \setminus C_j) \quad (20)$$

Therefore, in this study, the challenge was to determine the conditional probability of the feature $P(X_i \setminus C_j)$. Thus, the probability density was estimated by the Gaussian mixture, and then the NB classifier was built and used to predict the liquid loading.

Performance of the models on classifying the wells with and without liquid loading

The performances of all algorithms (for classification) were evaluated based on the accuracy, precision ratio and recall ratio of the results. The accuracy of the results was evaluated using Equation (21). The higher value of accuracy indicates the higher performance of the model, and the smaller value of accuracy indicates poor performance of the model. Therefore, the decision on which ML technique was best for predicting the liquid loading was evaluated based on the accuracy of the model [70].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (21)$$

Whereas TP shows true positive, TN shows a true negative, FP shows a false positive, and FN shows a false negative.

The precision ratio shows the ratio of samples that positive samples have been labelled as positive samples. The precision ratio was evaluated by using the Equation (22) [70].

$$\text{Precision} = \frac{TP+TN}{TP+FP} \quad (22)$$

Recall ratio indicates the ratio in which positive samples have been indicated as positive samples in all positive samples. Equation (23) was used to evaluate the recall ratio [70].

$$\text{Recall ratio} = \frac{TP}{TP+FN} \quad (23)$$

Copyright© Mgimba M.

Results and Discussion

Application of ML Algorithms on Classifying Wells with Liquid Loading

Before training the Coleman data, the correlation analysis was studied to determine the association between input data and liquid loading. The findings specify a linear association between input data and critical gas rate for liquid loading, as observed in Table 1. Wellhead pressure shows a strong correlation with the critical gas rate with a Person correlation of 0.769, followed by depth with a Person correlation of 0.332 and then bottom hole pressure with a Person correlation of 0.248. On the other hand, water content shows a poor correlation with the Person correlation of -0.042, followed by specific gas gravity with a Person correlation of -0.0094 and then condensate content with the Person correlation of -0.171. Since all the input data show some correlation with the critical gas rate, then all input data were used during the training and testing process. This decision is because machine learning algorithms are good at predicting even scenarios with non-linear relationships.

	Gas Specific gravity (air=1)	Depth (ft)	Tubing ID (In)	Condensate content (bbUMMscf)	Water content (bbl/MMscf)	WHFP (psia)	BHSP* (psig)	Observed (qc, Mscf/D)
Gas Specific gravity(air=1)	1.000							
Depth (ft)	0.000	1.000						
Tubing ID (In)	-0.030	0.325	1.000					
Condensate (bbUMMscf)	0.022	0.011	-0.036	1.000				
Water (bbl/MMscf)	-0.030	-0.198	-0.043	0.001	1.000			
WHFP (psia)	-0.013	0.396	0.162	-0.151	-0.138	1.000		
BHSP (psig)	-0.069	0.487	0.220	-0.171	0.076	0.470	1.000	
Observed (qc,Mscf/D)	-0.094	0.332	0.201	-0.127	-0.042	0.769	0.248	1.000

Table 1: The Correlation Between Input Data and Critical Gas Rate for Liquid Loading.

Application of ML Algorithms on Classifying Wells with Liquid Loading

The Running Time and Operating Model Parameters

The performances of ML algorithms, including GMDH, ANN, Naïve Bayes, decision tree and SVM, were determined based on running time and operating parameters, which provided the final results. These performances were evaluated using two kinds of data: data from the paper of Turner, et al. [13] and others from the paper of Coleman, et al. [15]; for simplicity, they will be called Turnel data and Coleman data. There were 107 field data reported in the article published by Turner, et al. [13] representing wells which are loaded (with liquid loading), unloaded (without liquid loading), near liquid loading, and others were questionable (they were not sure whether have liquid loading or not). Thus, the wells with liquid loading and near liquid loading were regarded as having liquid loading to tackle this situation. Meanwhile, those wells that were unloaded were regarded as not having liquid loading, and questionable wells were removed from the data. Hence, only 90 data remained indicating loaded and unloaded wells; these data were ultilised for training and testing using the five different algorithms.

In the application of ANN techniques, the data were splitted into three groups, training data and testing data, into a ratio of 70% and 30%, respectively. The results of the ANN technique in the Turnel data were attained after running the model for 22 epochs at a time of 21 seconds, with 10 hidden neurons and 1 output layer in the Turnel data, as illustrated in Table 2.

Copyright© Mgimba M.

Parameter	Initial value	Final value	Target value
Elapsed time	-	0:0:21	
Epoch	0	22	1000
Gradient	1.25	0.0825	1.0e-66
Performance	1.51	0.16	0
Validation checks	0	6	6

Table 2: The ANN model parameters which were used in predicting liquid loading.

The GMDH method predicted liquid loading in Turnel data. The initial model parameters applied by the GMDH method were 10 neurons, 15 layers, and 0.7 for classifying the testing and training data (in **the Turnel data**). However, the results were obtained after attaining 3 layers. Also, the results indicate that the application of GMDH in the Turnel data took 3 minutes, which was longer when compared to the ANN algorithm.

Other algorithms, including SVM, decision tree and Naïve Bayes, were applied to classify the Turner data. In these algorithms, all the data were treated together and trained to get the output results: wells with liquid loading or wells with no liquid loading. All these algorithms were solved with the aid of MATLAB (R2021a).

Performance of the Models

The performance of all five algorithms was evaluated by determining the accuracy, precision and recall of the data. Both GMDH and ANN's data were splitted into training data and testing data; thus, for the training data, the results of the models' performances are illustrated in Table 3. Based on these findings, the performance of GMDH was better than ANN, whereby accuracy for the GMDH algorithm was 0.984 higher than that of ANN, which was 0.935. Also, the precision and recall for the GMDH algorithm were 1.0 and 0.96, respectively, which were higher than that of ANN, which were both 0.929. These results indicate that in the GMDH algorithm, the liquid loading is well predicted when compared with ANN algorithms. Also, the wells with liquid loading are well indicated in the sample with liquid loading. Generally, the findings suggest that for both GMDH and ANN, the data were well-trained due to the fact that the accuracy, precision and recall were higher than or equal to 0.929.

Method	Accuracy	Precision	Recall
GMDH	0.984	1.000	0.960
ANN	0.935	0.929	0.929

Table 3: The performance of GMDH and ANN in the training Turner data.

As 70% of the Turner data were used to train the model, the remaining 30% were used to test the GMDH and ANN algorithms. The findings indicate that the accuracy of the GMDH algorithm was less than that of ANN, as follows: GMDH has an accuracy of 0.889, and ANN has an accuracy of 0.893. Also, the recall for the GMDH was less than that of the ANN. The recall is the only parameter higher in GMDH compared to ANN, as depicted in Table 4. But all three parameters, which indicate the performance of GMDH and ANN algorithms, are close to each other, meaning that both models have good predictions of liquid loading. Thus, the GMDH algorithm works better in the training data, but ANN performs better in the testing data.

Method	Accuracy	Precision	Recall
GMDH	0.889	0.900	0.818
ANN	0.893	0.778	0.875

Table 4: The Performance of GMDH and ANN in the Testing of Turners' Data.

The overall performance of all algorithms, including GMDH, ANN, SVM, decision tree and Naïve Bayes is illustrated in Figure 8. For the GMDH and ANN performance, the results for both training and testing data were combined. After combining the results then, they were compared with other models in which the data were not divided. The findings show that GMDH had the highest values of accuracy, precision and recall, which were 0.956, 0.971 and 0.917, respectively. The ANN follows by having higher values of accuracy, precision and recall, which were 0.922, 0.892 and 0.917, respectively. After ANN, SVM and decision tree show better results, but Naïve Bayes shows the worst results compared with other algorithms. The Naïve Bayes had the highest precision (1.0), but other parameters including accuracy and recall, were minimal when compared to different algorithms; accuracy was 0.789, and recall was 0.472. Therefore, the best presentation in predicting liquid loading in the Turner data was shown by the GMDH algorithm followed by the ANN algorithm. Also, the poor performance in predicting liquid loading was shown by the Naïve Bayes algorithm.

Copyright© Mgimba M.

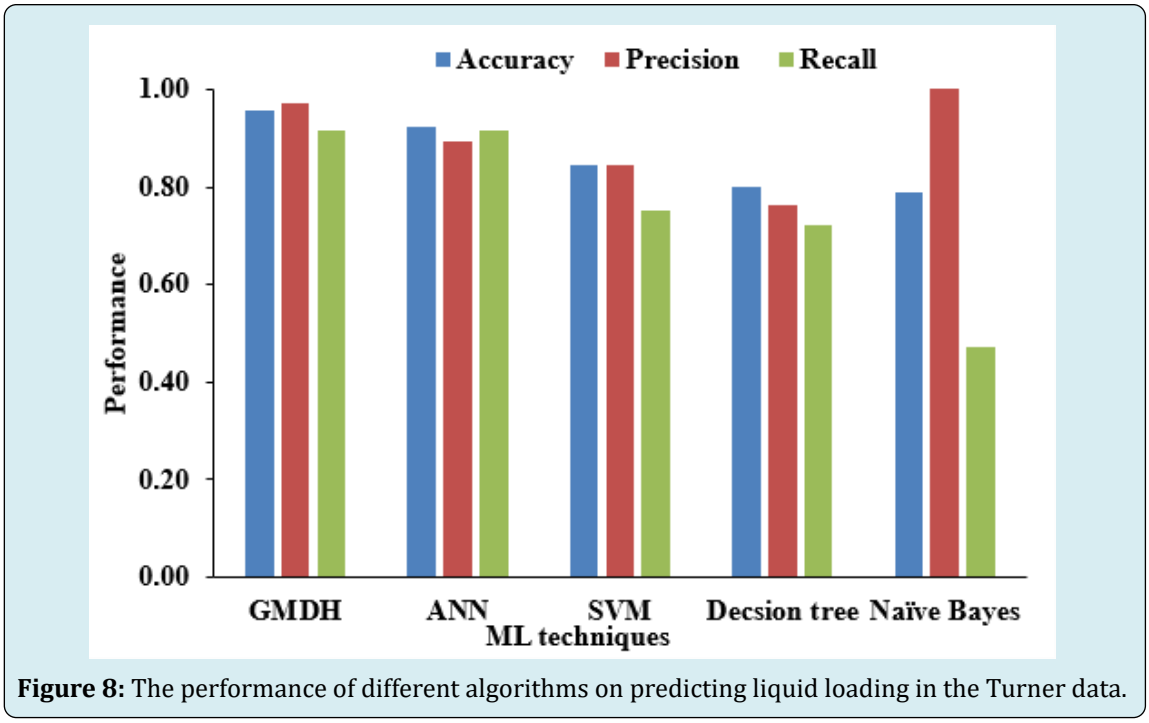


Figure 8: The performance of different algorithms on predicting liquid loading in the Turner data.

Application of ML Algorithms on Predicting Critical Gas Rate for Liquid Loading

Other data available in the article of Coleman, et al. [15] did not classify the well which is loaded and the one which is not, but on these wells, they stated the minimum gas flowing rate at which liquid loading can happen. The minimum gas flowing rate at which liquid loading can happen is referred to as the critical gas rate. The critical gas rate indicates the flowing rate at which the liquid loading can happen, such that when the gas flowing rate is below the critical gas rate, then the liquid loading can happen, and when the gas flowing rate is higher than the critical gas rate, then the loading cannot happen. Therefore, predicting critical gas rate is very significant for predicting liquid loading in the gas wells.

In this study, different algorithms were employed to forecast the critical gas rate in the gas well. These algorithms utilised the regression technique, which differs from the classification used in section 3.1. The algorithms used in the prediction of critical rate for gas loading include SVM, ANN and GMDH.

Forecast of Critical Gas Rate for Liquid Loading by Utilising SVM Algorithms

The SVM algorithm was applied to forecast the critical gas rate for liquid loading to happen in the gas well. This algorithm was used in the field data presented on the article by Coleman, et al. [13]. In these data, the critical gas rates for liquid loading were predicted from various input data, including specific gravity of gas, depth, tubing internal diameter, condensate content, water content, wellhead pressure (WHFP) and bottom hole pressure (BHP). During the training process, the data were separated into two sets, training data and testing data, where the training data occupied 70% and testing data occupied 30%.

The results for variations of critical gas rate predicted by SVM and observed in the field are indicated in Figure 9 and Figure 10 for training and testing data, respectively. These results are presented based on the association between the critical gas rate predicted by the algorithms (SVM) and the one which is observed in the field. This relationship is determined by evaluating the R^2. When there is the best prediction of critical gas rate, the R^2 is 1, and the poorest results is when the R^2 is 0. The results indicate that R^2 in the training data was 0.9737, and in the testing data was 0.8444. These results indicate that the model had good results in training and predicting the liquid loading (critical gas rate). Therefore, these results indicate the good ability of this method to forecast the critical gas rate for liquid loading.

Copyright© Mgimba M.

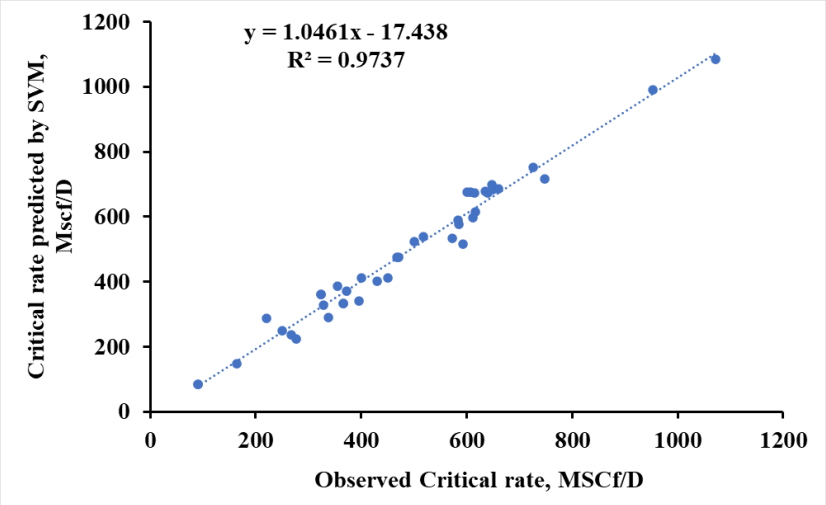


Figure 9: The Association Between the Critical Gas Rate Predicted by SVM and the One Observed in the Field (Training Data).

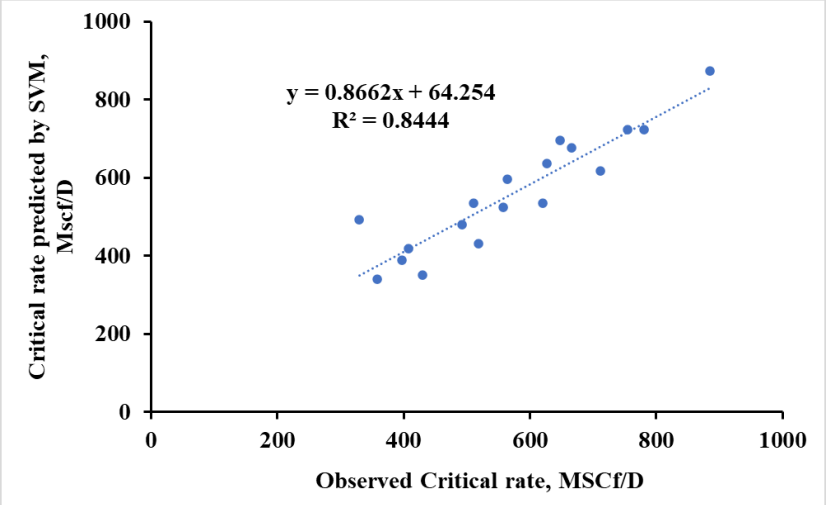


Figure 10: The Association Between the Critical Gas Rate Predicted By SVM and the One Observed in the Field (Testing Data).

Forecasting of Critical Gas Rate for Liquid Loading by Utilising ANN Algorithms

Another algorithm which was used to predict the critical gas rate liquid loading was the ANN. In this technique, 70% of the data were used as training data, and the rest 30% were applied as testing data. The results were found when the model run after 12 iterations in 10 seconds; other parameters used are illustrated in Table 5. Moreover, 10 hidden layers and 1 output layer were used in this case.

Parameter	Initial Value	Final Value	Target Value
Epoch	0	12	1000
Gradient	2.2e6	5399e3	1.0e-7
Performance	6.34e5	2.84e3	0
Elapsed time	-	0:0:10	0
Validation checks	0	6	6

Table 5: The ANN Model Parameters Which were used in Predicting Liquid Loading.

Copyright© Mgimba M.

The ANN algorithm was trained in the training data with the outcomes illustrated in Figure 11. The figure displays that the critical gas rate forecast by the ANN in the training data was approaching the critical gas rate observed in the field. The linear relationship shows that the R^2 between the critical gas rate forecasted by ANN and the one observed in the site was 0.9685. These results are less accurate than that of SVM, which has an R^2 of 0.9737 in the training data.

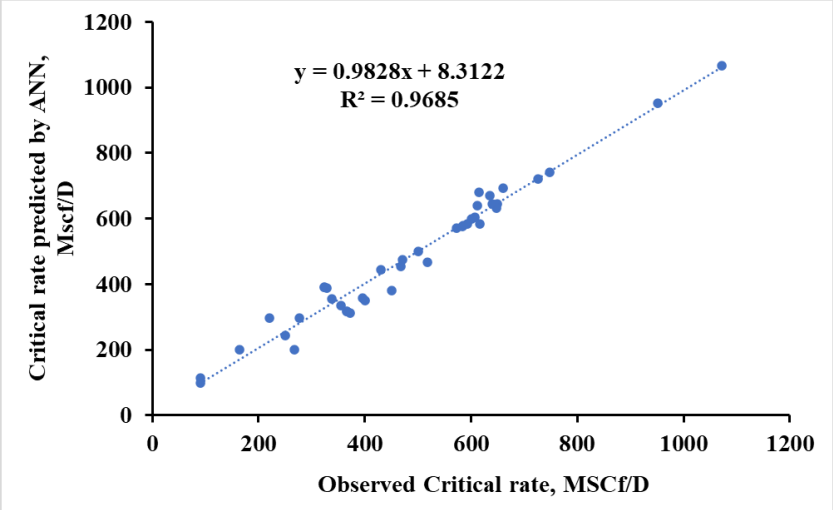


Figure 11: The Association Between the Critical Gas Rate Predicted by ANN and the One Observed in the Field (Training Data).

Also, the model was applied in the test data after training the algorithms. As shown in Figure 12, the coefficient of correlation (R^2) was 0.8921, indicating that the critical gas rate predicted by ANN was close to that observed in the field. The value of R^2 was also higher than that of SVM (0.8444), which indicates a better prediction of critical gas rate than to the SVM technique.

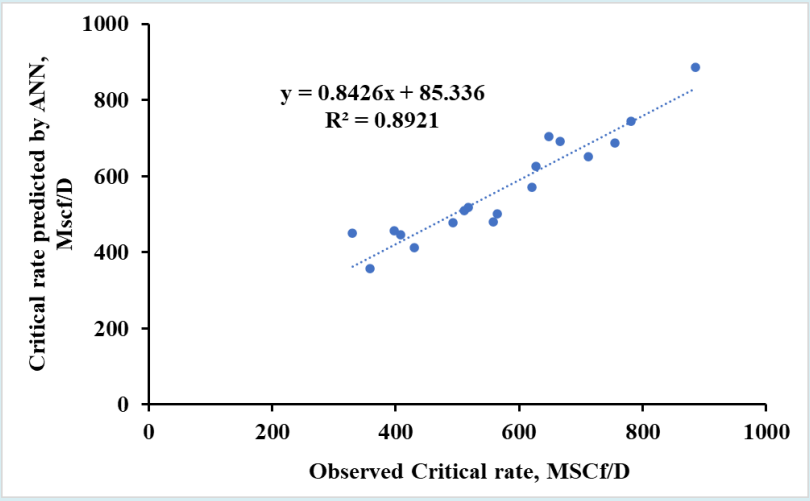


Figure 12: The Association Between the Critical Gas Rate Predicted by ANN and the One Observed in the Field (Testing Data).

Forecasting of Critical Gas Rate for Liquid Loading by Utilising GMDH Algorithms

The GMDH algorithm was employed to forecast the critical gas rate for liquid loading to happen in the gas well. This algorithm was employed in the field data obtained from the article of Coleman, et al. [13]. In these data, the liquid loading (critical gas rate) was predicted from various data, including condensate content, gas specific gravity, water content, depth, tubing internal diameter, wellhead pressure (WHFP) and bottom hole pressure (BHP).

Copyright© Mgimba M.

At the beginning, the algorithms employed 20 neurons, 50 layers, and a ratio of 0.7, which separate the training data (70%) and testing data (30%). However, the results were obtained after 20 layers with 20 neurons. The GMDH algorithm was trained, and the results are presented in Figure 13. The results indicate that the critical gas rate predicted by the GMDH in the training data was approaching the critical gas rate observed in the field. The linear relationship suggests that the coefficient of correlation (R^2) between the critical gas rate predicted by GMDH and the one observed in the field is 0.9725.

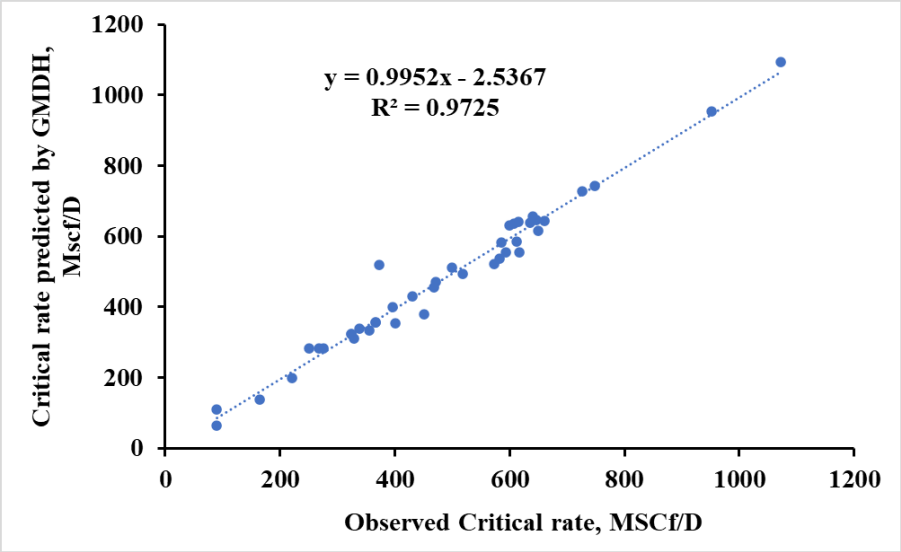


Figure 13: The Association Between the Critical Gas Rate Predicted by GMDH and the One Observed in The Field (Training Data).

After training the model, then the model was ultilised to the testing data. 30% of all data were used as the testing data, whereas their prediction findings, as illustrated in Figure 14. The findings indicate a good correlation between the critical gas rate predicted by GMDH and the one observed in the field. Whereas the R^2 in the linear graph between the critical gas predicted by GMDH and the one observed in the field is 0.9161. These findings indicate that the model has received proper training, but also its application on the testing data shows better results. When compared to other two algorithms (SVM and ANN), the GMDH shows the best accuracy because it has the highest value of R^2.

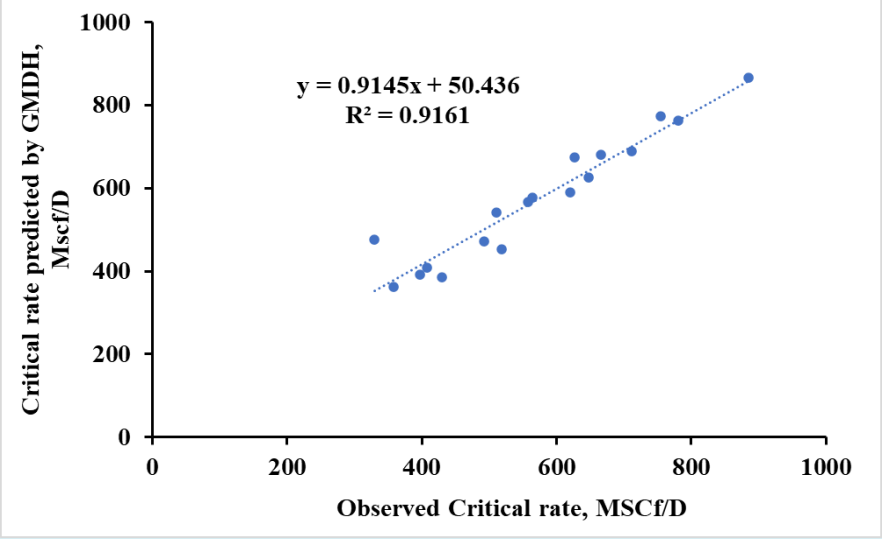


Figure 14: The Association Between the Critical Gas Rate Predicted by GMDH and the One Observed in the Field (In Testing Data).

Copyright© Mgimba M.

The performance of the algorithms in predicting the critical gas rate

Apart from the R^2, the performance of different methods in forecasting the critical gas rate for liquid loading was determined based on the root mean square error (RMSE). The RMSE represents the difference between the observed critical gas rate and the one predicted by the algorithms. Mathematically, the RMSE is evaluated by using Equation (24). The minimum value of RMSE specifies good performance of the model, whereas the higher value of the RMSE shows poor performance.

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(P_i-Y_i\right)^2} \quad (24)$$

Generally, the performance of the algorithms in predicting the critical gas rate for liquid loading is presented in Table 6**.** The results show that both SVM, ANN and GMDH have good performances in predicting critical gas rates for liquid loading. But, GMDH shows the best performance compared to other algorithms since it has an RMSE of 43.7472 Mscf/D, which is the lowest compared with the remaining algorithms. ANN is observed to be the second one in its performance, whereas it has an RMSE of 50.3773 Mscf/D, and SVM shows the worst performance by having an RMSE of 60.8203 Mscf/D. The RMSE results are supported by the values of R^2, whereby the GMDH has the highest value of R^2 (0.9161), followed by ANN (0.8921) and lastly SVM (0.8444).

	Training data		Testing data	
Methods	**R^2**	**RMSE (Mscf/D)**	**R^2**	**RMSE (Mscf/D)**
SVM	0.9797	37.0798	0.8444	60.8203
ANN	0.9685	36.7301	0.8921	50.3773
GMDH	0.9725	34.9119	0.9161	43.7472

Table 6: The Performance of Different Algorithms in Predicting the Critical Gas Rate for Liquid Loading

Discussions

The findings of this study show that the GMDH technique has more accuracy in predicting the liquid loading and critical gas rate for liquid loading. These results agreed with other scholars' findings, who utilised the GMDH algorithms, including Mgimba, et al. [41], who found the best performance of GMDH on forecasting the pore pressures when compared to ANN and polynomial classifier (POL) techniques. Also, Mulashani, et al. [44] found that the GMDH method provided the best performance in forecasting the TOC compared to ANN and logR methods. Furthermore, Gao, et al. [71], Mathew Nkurlu, et al. [35], Mkono, et al. [37], Parsa, et al. [72], Zhang and Zhang [73] found that GMDH surpassed other algorithms, including the DT algorithm, MELM and RBFNN in predicting permeability, formation pressure and TOC. Thus, the GMDH algorithm looks better at forecasting different parameters in the oil and gas industry.

The accuracy of predicting the liquid loading in this work was 0.956 for GMDH, 0.922 for ANN and 0.844 for SVM. These findings show a good ability of ML to forecast liquid loading in the gas wells and agree with the findings provided by Ansari et al. (2018). In that work, they developed the smart model, and through that model, they generated several production data. Then, they used the production data to predict the liquid loading. The findings of that study show that the application of ANN to predict the liquid loading has an accuracy of 99.9%. The results were better than our results, but first, the actual site data were not used, and also, the classification of the well with liquid loading was done based on Turner's critical rate, which other scholars say this rate underestimates the critical flow rate [14]. But also, few input data, including casing pressure, tubing pressure and production rate, were employed to forecast liquid loading in the gas wells.

Lastly, the GMDH provided its best performance in forecasting the critical gas rate compared to other models. For instance, the coefficient of correlation (R^2) between the critical gas rate predicted by the GMDH and the observed one was 0.9161, and the RMSE was 43.7472 Mscf/D. These results are better when compared with other models as reported by Enos [14], whereby Turner, et al. [13] model provided an RMSE of 56 Mscf/D, the Barnea model provided an RMSE of 168 Mscf/D, Luo, et al (2014) provided an RMSE of 72 Mscf/D and Shekhar, et al. provided RMSE of 94 Mscf/D on the same kind of data.

Therefore, the application of the GMDH method on the prediction of loading is more accurate than other mathematical models which have been used previously. Moreover, the use of GMDH in forecasting the liquid loading will save the cost for direct measurement of critical gas rate for liquid loading. Therefore, we recommend the application of GMDH in forecasting liquid loading/ critical gas rate in the gas wells.

Copyright© Mgimba M.

Limitations of the Study

The main limitation of this study is that few data were applied in the forecasting of both liquid loading and critical gas rate for liquid loading; thus, more data are recommended for more accuracy of the results. Also, the input data considered in this study are specific gravity of gas, depth, tubing internal diameter, condensate content, water content, wellhead pressure (WHFP) and bottom hole pressure (BHP); thus, other data, including orientation of the well and flow pattern were not considered. Therefore, we recommend that other scholars consider those data.

Conclusions

The application of ML algorithms in predicting liquid loading has shown some success. Five algorithms, namely GMDH, ANN, SVM, Naïve Bayes, and Decision tree, were employed to forecast the liquid loading in the gas wells. These methods were employed in the data issued in the article by Turer, et al. [13]. The findings display that the GMDH algorithm has the best performance on forecasting the liquid loading with an accuracy of 0.956, precision of 0.971 and recall of 0.917. Then, the ANN was the second technique for good prediction of liquid loading, whereas it has an accuracy of 0.922, precious of 0.892 and recall of 0.917. Other algorithms, including SVM and decision tree, show better results, but Naïve Bayes shows worse results compared to other algorithms with an accuracy of 0.789, recall of 0.472 and precision of 1. Thus, the GMDH algorithm performs best in forecasting liquid loading in the Turner data. Also, the poor performance in forecasting liquid loading was demonstrated by the Naïve Bayes algorithm.

Apart from direct predictions of liquid loading by the classification methods, other field data were presented based on the critical gas rate for liquid loading. Thus, on the other part, the regressions algorithms for GMDH, ANN and SVM were applied to predict the critical gas rate for liquid loading in the data presented in the article of Coleman, et al. [15]. The critical gas rates were predicted from input data, including specific gravity of gas, depth, tubing internal diameter, condensate content, water content, wellhead pressure (WHFP) and bottom hole pressure (BHP). The results indicate that GMDH has the best performance on predicting the critical gas rates for liquid loading with the smallest RMSE (43.7472 Mscf/D), followed by ANN with an RMSE of 50.3773 Mscf/D. However, SVM shows the worst results compared to GMDH and ANN, whereas its RMSE was 60.8203 Mscf/D. Also, the correlation of the coefficient of GMDH was highest (0.9161) when compared to ANN (with RMSE of 0.8921) and SVM (with RMSE of 0.8444).

GMDH shows the best performance in predicting liquid loading and critical gas rates for liquid loading to happen. But this method also has other advantages, including the capacity to identify the data structure, the automatic selection of pertinent input data, and direct approximation. Therefore, GMDH is recommended as the best technique for predicting the liquid loading in the gas wells.

Conflict of Interest

We declare no conflict of interest

Data Availability

All data supporting the findings of this study are available within the paper and its Supplementary Information. Most data used are available in the article published by Coleman et al. [15] (DOI: https://doi.org/10.2118/20280-PA) and Turner, et al. [13] (https://doi.org/10.2118/123657-PA).

Competing Interest

The authors state that none of their known financial conflicts or interpersonal connections might have had an impact on the work presented in this paper.

Ethical Statement

This material is the authors' own original work, which has not been previously published elsewhere.

Funding Declaration

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Zwickl-Bernhard S, Rodgarkia-Dara A, Gatzen C, Sonnen L, Lane A, et al. (2024) Modeling insights from the Austrian national gas grid under declining natural gas demand and increasing domestic renewable gas generation by 2040. Energy Rep 11: 1302-1317.

2. Nallaparaju YD (2012) Prediction of liquid loading in gas wells. SPE Annual Technical Conference and Exhibition.

3. Fan D, Gong J, Zhang S, Shi G, Kang Q, et al. (2021) A transient composition tracking method for natural gas pipe networks. Energy 215: 119131.

4. Whitson CH, Brule MR (2000) Phase behavior. Vol 20, Henry L. Doherty Memorial Fund of AIME, Society of Petroleum Engineers.

Copyright© Mgimba M.

5. Adaze E, Al-Sarkhi A, Badr H, Elsaadawy E (2019) Current status of CFD modeling of liquid loading phenomena in gas wells: a literature review. J Pet Explor Prod Technol 9(2): 1397-1411.

6. Hong B, Li X, Di G, Li Y, Liu X, et al. (2019) An integrated MILP method for gathering pipeline networks considering hydraulic characteristics. Chem Eng Res Des 152: 320-335.

7. Shi G, Song S, Shi B, Gong J, Chen D (2021) A new transient model for hydrate slurry flow in oil-dominated flowlines. J Pet Sci Eng 196: 108003.

8. Landjobo Pagou A, Han G, Peng L, Dehdah O, Gueyap Kamdem V, et al. ( ) Liquid loading prediction and identification model for vertical and inclined gas wells. J Nat Gas Sci Eng 84: 103641.

9. Alsanea M, Mateus-Rubiano C, Karami H (2022) Liquid loading in natural gas vertical wells: a review and experimental study. SPE Prod Oper 37(03): 554-571.

10. He G, Chen D, Liao K, Sun J, Nie S (2019) A methodology for the optimal design of gathering pipeline system in old oilfield during its phased development process. Comput Ind Eng 130: 14-34.

11. Chen S, Webb GI, Liu L, Ma X (2020) A novel selective naive Bayes algorithm. Knowl Based Syst 192: 105361.

12. Barnea D (1987) A unified model for predicting flow-pattern transitions for the whole range of pipe inclinations. Int J Multiphase Flow 13(1): 1-12.

13. Turner RG, Hubbard MG, Dukler A (1969) Analysis and prediction of minimum flow rate for the continuous removal of liquids from gas wells. J Pet Technol 21(11): 1475-1482.

14. Enos MK (2018) Evaluation of Models to Predict Liquid Loading in Gas wells. NTNU.

15. Coleman SB, Clay HB, McCurdy DG, Norris LH III (1991) A new look at predicting gas-well load-up. J Pet Technol 43(03): 329-333.

16. Luo S, Kelkar M, Pereyra E, Sarica C (2014) A new comprehensive model for predicting liquid loading in gas wells. SPE Prod Oper 29(04): 337-349.

17. Nosseir M, Darwich T, Sayyouh M, Sallaly ME (2000) A New Approach for Accurate Prediction of Loading in Gas Wells Under Different Flowina Conditions. SPE Prod Facil 15(04): 241-246.

18. Van't Westende JMC (2008) Droplets in annular-dispersed gas-liquid pipe-flows.

19. Veeken K, Hu B, Schiferli W (2010) Gas-well liquid-loading-field-data analysis and multiphase-flow modeling. SPE Prod Oper 25(03): 275-284.

20. Duns H, Ros N (1963) Vertical flow of gas and liquid mixtures from boreholes. Bataafse Internationale Petroleum Maatschappij, NV.

21. Abhulimen KE, Abhulimen K, Oladipupo A (2023) Modelling of liquid loading in gas wells using a software-based approach. J Pet Explor Prod Technol 13(1): 1-17.

22. Beggs HD, Brill JP (1973) A Study of Two-Phase Flow in Inclined Pipes. J Pet Technol 255.

23. Dukler A, Taitel Y (2017) Measurement and modeling of flow pattern transitions in gas-liquid systems. Multiph Sci Technol 29(1-4).

24. Guo B, Ghalambor A, Xu C (2005) A systematic approach to predicting liquid loading in gas wells. SPE Oklahoma City Oil and Gas Symposium/Production and Operations Symposium.

25. Ansari A, Fathi E, Belyadi F, Takbiri-Borujeni A, Belyadi H (2018) Data-based smart model for real time liquid loading diagnostics in Marcellus Shale via machine learning. SPE Canada Unconventional Resources Conference.

26. Kabir CS, Hasan AR (1990) Performance of a two-phase gas/liquid flow model in vertical wells. J Pet Sci Eng 4(3): 273-289.

27. Riza MF, Hasan AR, Kabir CS (2016) A pragmatic approach to understanding liquid loading in gas wells. SPE Prod Oper 31(03): 185-196.

28. Belfroid S, Schiferli W, Alberts G, Veeken CA, Biezen E (2008) Prediction onset and dynamic behaviour of liquid loading gas wells. SPE Annual Technical Conference and Exhibition.

29. Pagou AL, Wu X, Zhu Z, Peng L (2020) Liquid film model for prediction and identification of liquid loading in vertical and inclined gas wells. J Pet Sci Eng 188: 106896.

30. Hutomo P, Rosid M, Haidar M (2019) Pore pressure prediction using eaton and neural network method in carbonate field "X" based on seismic data. IOP Conference Series: Materials Science and Engineering.

31. Qi C, Fourie A, Chen Q (2018) Neural network and particle swarm optimization for predicting the unconfined compressive strength of cemented paste backfill. Constr

Copyright© Mgimba M.

Build Mater 159: 473-478.

32. Ahmed A, Elkatatny S, Ali A, Mahmoud M, Abdulraheem A (2019) New model for pore pressure prediction while drilling using artificial neural networks. Arab J Sci Eng 44(6): 6079-6088.

33. Hu L, Deng J, Zhu H, Lin H, Chen Z, et al. (2013) A new pore pressure prediction method-back propagation artificial neural network. Electron J Geotech Eng 18: 4093-4107.

34. Keshavarzi R, Jahanbakhshi R (2013) Real-time prediction of pore pressure gradient through an artificial intelligence approach: a case study from one of middle east oil fields. Eur J Environ Civ Eng 17(8): 675-686.

35. Mathew Nkurlu B, Shen C, Asante-Okyere S, Mulashani AK, Chungu J, et al. (2020) Prediction of permeability using group method of data handling (GMDH) neural network from well log data. Energies 13(3): 551.

36. Nyakilla EE, Silingi SN, Shen C, Jun G, Mulashani AK, et al. (2022) Evaluation of source rock potentiality and prediction of total organic carbon using well log data and integrated methods of multivariate analysis, machine learning, and geochemical analysis. Nat Resour Res 31(1): 619-641.

37. Mkono CN, Chuanbo S, Mulashani AK, Mwakipunda GC (2023) Deep learning integrated approach for hydrocarbon source rock evaluation and geochemical indicators prediction in the Jurassic-Paleogene of the Mandawa basin, SE Tanzania. Energy 284: 129232.

38. Nadege MN, Jiang S, Mwakipunda GC, Kouassi AKF, Harold PK, et al. (2024) Brittleness index prediction using modified random forest based on particle swarm optimization of Upper Ordovician Wufeng to Lower Silurian Longmaxi shale gas reservoir in the Weiyuan Shale Gas Field, Sichuan Basin, China. Geoenergy Sci Eng 233: 212518.

39. Kanin EA, Osiptsov AA, Vainshtein AL, Burnaev EV (2019) A predictive model for steady-state multiphase pipe flow: Machine learning on lab data. J Pet Sci Eng 180: 727-746.

40. Hong BY, Liu SN, Li XP, Fan D, Ji SP, et al. (2022) A liquid loading prediction method of gas pipeline based on machine learning. Pet Sci 19(6): 3004-3015.

41. Mgimba MM, Jiang S, Nyakilla EE, Mwakipunda GC (2023) Application of GMDH to Predict Pore Pressure from Well Logs Data: A Case Study from Southeast Sichuan Basin, China. Nat Resour Res 32(4): 1711-1731.

42. Najafzadeh M, Barani GA, Hessami-Kermani MR (2015) Evaluation of GMDH networks for prediction of local scour depth at bridge abutments in coarse sediments with thinly armored beds. Ocean Eng 104: 387-396.

43. Mulashani AK, Shen C, Nkurlu BM, Mkono CN, Kawamala M (2022) Enhanced group method of data handling (GMDH) for permeability prediction based on the modified Levenberg Marquardt technique from well log data. Energy 239: 121915.

44. Mulashani AK, Shen C, Asante-Okyere S, Kerttu PN, Abelly EN (2021) Group Method of Data Handling (GMDH) Neural Network for Estimating Total Organic Carbon (TOC) and Hydrocarbon Potential Distribution (S1, S2) Using Well Logs. Nat Resour Res 30(5): 3605-3622.

45. Youcefi MR, Hadjadj A, Boukredera FS (2022) New model for standpipe pressure prediction while drilling using Group Method of Data Handling. Petroleum 8(2): 210-218.

46. Mesbah M, Habibnia S, Ahmadi S, Saeedi Dehaghani AH, Bayat S (2022) Developing a robust correlation for prediction of sweet and sour gas hydrate formation temperature. Petroleum 8(2): 204-209.

47. Korbicz J, Mrugalski M (2008) Confidence estimation of GMDH neural networks and its application in fault detection systems. Int J Syst Sci 39(8): 783-800.

48. Liu GP (2001) Nonlinear identification and control: a neural network approach. Springer Science & Business Media.

49. Murphy KP (2002) Dynamic bayesian networks: representation, inference and learning. University of California, Berkeley.

50. Rahim NA, Taib M, Adom A, Mashor M (2006) The NARMAX model for a dc motor using mlp neural network. Proceeding of the First International Conference on MAN-MACHINE SYSTEMS (ICoMMS).

51. Abu-Kheil YMZ (2009) System Identification using group method of data handling (GMDH).

52. Xiang Z, Lv X, Zhang K (2014) An image classification method based on multi-feature fusion and multi-kernel SVM. 2014 Seventh International Symposium on Computational Intelligence and Design.

53. Bressan TS, de Souza MK, Girelli TJ, Junior FC (2020) Evaluation of machine learning methods for lithology classification using geophysical data. Comput Geosci 139: 104475.

Copyright© Mgimba M.

54. Chandra MA, Bedi S (2021) Survey on SVM and their application in image classification. Int J Inf Technol 13: 1-11.

55. Li Z, Kang Y, Feng D, Wang XM, Lv W, et al. (2020) Semi-supervised learning for lithology identification using Laplacian support vector machine. J Pet Sci Eng 195: 107510.

56. Yu L, Porwal A, Holden EJ, Dentith MC (2012) Towards automatic lithological classification from remote sensing data using support vector machines. Comput Geosci 45: 229-239.

57. Bhattacharya S, Carr TR (2019) Integrated data-driven 3D shale lithofacies modeling of the Bakken Formation in the Williston basin, North Dakota, United States. J Pet Sci Eng 177: 1072-1086.

58. Wang G, Carr TR, Ju Y, Li C (2014) Identifying organic-rich Marcellus Shale lithofacies by support vector machine classifier in the Appalachian basin. Comput Geosci 64: 52-60.

59. Song YY, Ying L (2015) Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry 27(2): 130.

60. Yu Z, Haghighat F, Fung BC, Yoshino H (2010) A decision tree method for building energy demand modeling. Energy Build 42(10): 1637-1646.

61. Duan Y, Xie J, Su Y, Liang H, Hu X, et al. (2020) Application of the decision tree method to lithology identification of volcanic rocks-taking the Mesozoic in the Laizhouwan Sag as an example. Sci Rep 10(1): 19209.

62. Mollas I, Tsoumakas G, Bassiliades N (2019) Lionforests: Local interpretation of random forests through path selection. arXiv preprint arXiv:1911.08780.

63. Timofeev R (2004) Classification and regression trees (CART) theory and applications. Humboldt University, Berlin.

64. Aldino AA, Sulistiani H (2020) Decision Tree C4.5 Algorithm For Tuition Aid Grant Program Classification (Case Study: Department Of Information System, Universitas Teknokrat Indonesia). J Ilm Edutic 7(1): 40-50.

65. Jin C, De-Lin L, Fen-Xiang M (2009) An improved ID3 decision tree algorithm. 2009 4th International Conference on Computer Science & Education.

66. Hssina B, Merbouha A, Ezzikouri H, Erritali M (2014) A comparative study of decision tree ID3 and C4.5. Int J Adv Comput Sci Appl 4(2): 13-19.

67. Janssens D, Wets G, Brijs T, Vanhoof K (2007) The Development of a Segmentation Scheme for the Evaluation and Prediction of Activity-Travel Sequences.

68. Ren Q, Zhang H, Zhang D, Zhao X, Yan L, et al. (2022) A framework of active learning and semi-supervised learning for lithology identification based on improved naive Bayes. Expert Syst Appl 202: 117278.

69. Swarnkar M, Hubballi N (2016) OCPAD: One class Naive Bayes classifier for payload based anomaly detection. Expert Syst Appl 64: 330-339.

70. Li K, Xi Y, Su Z, Zhu J, Wang B (2021) Research on reservoir lithology prediction method based on convolutional recurrent neural network. Comput Electr Eng 95: 107404.

71. Gao G, Hazbeh O, Rajabi M, Tabasi S, Ghorbani H, et al. (2023) Application of GMDH model to predict pore pressure. Front Earth Sci 10: 1043719.

72. Parsa M, Carranza EJM, Ahmadi B (2022) Deep GMDH neural networks for predictive mapping of mineral prospectivity in terrains hosting few but large mineral deposits. Nat Resour Res 31(1): 37-50.

73. Zhang S, Zhang R (2023) Prediction of the thermal conductivity of rocks using group method of data handling (GMDH). Geothermics 115: 102823.

Copyright© Mgimba M.