Article

Field Telemetry Drilling Dataset Modeling with Multivariable Regression, Group Method Data Handling, Artificial Neural Network, and the Proposed Group-Method-Data-Handling-Featured Artificial Neural Network

Amir Mohammad * and Mesfin Belayneh (iD)

Department of Electrical Engineering & Computer Science, University of Stavanger, 4021 Stavanger, Norway; mesfin.a.belayneh@uis.no
* Correspondence: amirmoh4321@gmail.com; Tel.: +47-951-29-664

Abstract: This paper presents data-driven modeling and a results analysis. Group method data handling (GMDH), multivariable regression (MVR), artificial neuron network (ANN), and new proposed GMDH-featured ANN machine learning algorithms were implemented to model a field telemetry equivalent mud circulating density (ECD) dataset based on surface and subsurface drilling parameters. Unlike the standard GMDH-ANN model, the proposed GMDH-featured ANN utilizes a fully connected network. Based on the considered eighteen experimental modeling designs, all the GMDH regression results showed higher R-squared and minimum mean-square error values than the multivariable regression results. In addition, out of the considered eight experimental designs, the GMDH-ANN model predicts about 37.5% of the experiments correctly, while both algorithms have shown similar results for the remaining experiments. However, further testing with diverse datasets is necessary for better evaluation.

Keywords: swab; surge; GMDH; MVR; ANN; GMDH-featured ANN; ECD

1. Introduction

Swab and surge pressures are created during the drill string tripping out and into the wellbore, respectively. The magnitude of the hydrostatic pressure fluctuation due to the swab/negative and surge/positive depends on, among others, the tripping speed. An unoptimized tripping speed may induce wellbore instability, leading to well collapse or formation fracturing, as well as control issues, which increase the nonproductive time (NPT) and hence the overall drilling budget. The precise prediction of well pressures is crucial, especially in deep-water and horizontal drilling, where a narrow well stability margin poses significant challenges. In addition to tripping speed, fluid properties, geometry, eccentricity, and flow rates also control the well pressure. Moreover, the prediction of the optimum maximum and minimum tripping speeds reduces undesired nonproductive times.

Over the years, researchers have conducted numerous studies of swab and surge phenomena during tripping in and out of the wellbore. They have conducted several experiments and also developed various models based on assumptions such as steady-state and transient conditions, non-slip at the wall, different flow scenarios, fluid rheological properties, well configurations, and operational parameters. Amir et al. (2022) [1,2] extensively reviewed existing swab and surge models, including contributions from Burkhardt (1961) [3], Schuh (1964) [4], Fontenot and Clark (1974) [5], Mitchell (1988) [6], Ahmed (2008) [7], Crespo (2010) [8], Srivastav (2012) [9], Gjerstad (2013) [10], Tang (2016) [11], Fredy (2012) [12], Erge (2015) [13], He (2016) [14], Evren M. (2018) [15], Ettehadi (2018) [16], Shwetank (2020) [17,18], Zakarya (2021) [19], and Amir et al. (2023) [20]. However, these

models did not consider all the parameters that affect the swab and surge, and their applicability to estimate experimental data is limited to the specific assumptions and setup conditions.

During drilling and tripping operations, the well pressure is normally determined by the hydrostatic pressure and by the pressure loss due to fluid flow. The equivalent circulation density in specific gravity (sg) is given as (Mitchel et al., 2011) [21]:

$$ECD = \rho_{static} + \frac{\Delta P_{annulus}}{0.0981.TVD} \tag{1}$$

where $\Delta P_{annulus}$ is the pressure loss (bar) in the annulus due to fluid flow and ρ_{static} is the static drilling fluid density (sg). TVD is the true vertical depth (m).

The pump pressure is also determined from the pressure losses across circulation flowlines. The pressure loss in the annulus is given as (Mitchel et al., 2011) [21].

$$\Delta P_{\ annulus} = \frac{2f\rho V_Q^2}{D_H}L \tag{2}$$

where $V_Q = Q/A$ is the velocity of the fluid flow, D_H is the hydraulic fluid flow through the annulus (D_{Well}–D_{Pipe}), L is the length of the flow line, and f is the friction factor.

The friction factor f is a function of the Reynolds number, and the surface roughness is given by Haaland (1983) [22]:

$$\frac{1}{\sqrt{f}} = -1.8 \log \left[\left(\frac{\varepsilon/D}{3.7} \right)^{1.11} + \frac{6.9}{Re} \right] \tag{3}$$

where ε is the surface roughness coefficient ($\varepsilon = k/d$), k is the surface roughness, and D is the diameter of the pipe.

The friction factor parameter is sensitive, and its prediction is difficult as it is a profile. The theoretical calculation of pressure losses in a wellbore requires knowledge of fluid properties at various temperatures and the shear rates as the fluid flows through each interval of a borehole.

To determine the rheological properties of the drilling fluids and the ECD, there are several models available in the industry's commercial software. Despite the availability of various mathematical, empirical, and physics-based models currently used in the drilling and well construction sector, incidents related to swab and surge pressures continue to occur. The comparisons of field-measured data with the hydraulic well-flowing models showed discrepancies, and the model required a calibration factor based on measured data (Lohne et al., 2008) [23]. Simulation studies conducted by Amir et al., 2023 [24] showed that the swab and surge prediction of the models were inconsistent and deviated from each other for the considered experimental setup.

In recent years, the application of data-driven modeling has been employed in various sectors, including petroleum drilling. There are several machine learning modeling algorithms. For instance, Amir et al. (2022, 2023) [24,25] utilized machine learning techniques (i.e., linear regression, multivariable regression (MVR), Random Forest, ANN, long-short-term memory (LSTM), and XGboost models) to predict the tripping and drilling operation's equivalent circulating mud density (ECD), and the results showed satisfactory performance.

In the petroleum industry, among others, ANN algorithms have been applied for prediction such as ROP (Reda Abdel Azim (2020) [26], Ramin Aliyev (2019) [27]), ECD (Husam H. Alkinani (2020) [28], Amir et al., 2021 [24,25]), drilling speed (Ahmad Al-Abduljabbar et al. (2020) [29]), and drilling-fluid-rheological-parameter real-time prediction (Khaled Al-Azani et al. (2018) [30]). In addition, A. Alnmnr (2024) implemented machine learning to investigate Swell Mitigation [31]. RP Ray (2023) studied the importance of data integration in Geotechnical Engineering [32]. E. Gurina (2022) deployed machine learning techniques to predict dysfunctional events in drilling and wells [33].

A literature review indicates the application of the Group Method of Data Handling (GMDH) technique across diverse fields. GMDH is an extended version of multivariable regression that contains non-linear interacting terms. Among others, GMDH has been utilized for accurate log interval value estimation (Mohammed Ayoub (2014) [34]), permeability prediction by Alvin K. Mulashani (2019) [35] and Lidong Zhao (2023) [36], as well as permeability modeling and pore pressure analysis by Mathew Nkurlu (2020) [37]. Additionally, GMDH finds applications in cement compressive strength design (Edwin E. Nyakilla, 2023 [38]), rock deformation prediction (Li et al., 2020 [39]), bubble point pressure estimation by Fahd Saeed Alakbari (2022) and Mohammad Ayoub (2022) [40,41], gas viscosity determination, CO_2 emission modeling (Rezaei et al., 2020 and 2018 [42,43]), the prediction of CO_2 adsorption by Zhou L. (2019) [44] and Li (2017) [45], forecasting stock indices, and modeling power and torque as demonstrated by Ahmadi (2015) [46] and Gao Guozhong (2023) [47], and the prediction of pore pressure by Mgimba (2023) [48].

The GMDH neural network architecture is not as fully connected as that of the commonly used ANN. The modeling performance of the GMDH network in comparison with that of the ANN model is presented in several publications including André et al., 2012 [49]; Bernard et al., 2020 [50]; Ahmadi et al., 2015 [46]; Rezaei et al., 2015 [43].

This study aimed to implement four machine learning algorithms on field drilling data. The first study compared MVR with GMDH to assess the impact of the non-linear interacting features that GMDH has on model prediction. The second study proposed new GMDH-method-generated features that can be utilized as inputs for deep-learning (ANN) modeling, and the networks were fully connected. Then, the newly proposed GMDH-ANN method was compared with a standard ANN that did not include interacting terms. Finally, empirical models were derived from field drilling data by using GMDH and MVR methods.

2. Methodology

The typical machine learning modeling workflow comprises three parts, namely data processing, modeling, and model accuracy performance evaluations. This section presents the details of the description of the dataset and the machine modeling algorithms.

2.1. Description of the Dataset

In this study, therefore, field drilling data acquired through a high-speed (wired drill pipe) telemetry system was used for modeling equivalent circulating density (ECD) based on drilling parameters. Data quality determined the model accuracy performance of the data-driven machine learning modeling. Therefore, the raw data was preprocessed to ensure cleanliness and appropriate feature selection using a correlation heat map to meticulously identify the most suitable features for the modeling process.

2.2. Description of the Machine Learning Algorithms

This section presents the description of the machine learning algorithms implemented in this study that included multivariable regression (MVR), GMDH, ANN, and the proposed fully connected GMDH-featured ANN.

2.2.1. Multivariate Regression

In statistical analysis, researchers actively employ MVR to generate a relationship between the multiple independent variables $x_1, x_2, x_3, \ldots, x_n$, and a single dependent variable y_i (Prentice, 1981) [51]. This technique holds significance in fields where several factors concurrently influence an outcome. MVR has been implemented in several fields and in the petroleum industry. Among others, Amir et al. (2022) [21] applied the method on the field tripping-out dataset.

In this paper, MVR related independent drilling variables/features to predict the target variable, y (ECD). The multiple linear regression model represented the linear combination of weighted features and is expressed as (Anderson T.W., 2003) [52]:

$$y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon \tag{4}$$

Here, y_i is a dependent variable representing ECD, while $x_1, x_2, x_3, \ldots x_n$ different surface and downhole parameters. In Equation (1), β_0 represents the y-intercept (the value of y when all other independent variables are set to 0), β_1 denotes the regression coefficient of the first independent variable x_1, β_n represents the regression coefficient of the last independent variable x_n, and ε is the model error (that describes the degree of variation from the estimate of y_i).

The MVR equation can be represented in a concise framework that relates the target, regression variables, regression coefficients, and random errors in the form:

$$y_i = X\beta + \varepsilon \tag{5}$$

The model in matrix form can be represented as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots & \ldots & x_{1M} \\ 1 & x_{21} & \ldots & \ldots & x_{2M} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & \ldots & \ldots & x_{NM} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_N \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} \tag{6}$$

Here y_i represents a column vector of the observed values (ECD), X denotes a matrix of independent variables (each column represents a different drilling variable), β represents a column vector of coefficients, and ε a column vector of errors.

The optimized set of coefficients. $(\beta_0, \beta_1, \ldots, \beta_n)$ that minimizes the sum of squared differences between the observed y_i and predicted values from the regression function is described in Equation (4).

$$SSR = \min_{\beta_0, \beta_1, \ldots, \beta_n} \sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_n X_{ni}))^2 \tag{7}$$

The solution involves minimizing the sum of squared errors, leading to the estimated coefficients:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y_i \tag{8}$$

Here, $\hat{\beta}$ is the vector of estimated coefficients $\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_n\right)$, X^T is the transpose of matrix X, and $\left(X^T X\right)^{-1}$ is the inverse of the product of the transpose of X and X.

The estimated coefficients are plugged back into the original equation, yielding the final multiple regression equation:

$$\widehat{ECD} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_n X_n \tag{9}$$

2.2.2. Group Method Data Handling (GMDH) Algorithm

Alexey G. Ivakhnenko [53] developed GMDH in 1971 as a self-organizing artificial neural network (ANN) for modeling complex systems with multiple variables and nonlinear relationships between inputs and outputs. The relationship between the output and

input variables is mathematically described using a process called the Kolmogorov–Gabor polynomial [53]:

$$y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} a_{ijk} x_i x_j x_k + \ldots \tag{10}$$

In Equation (7), n represents the number of input variables, $(x_1, x_2, \ldots, x_n)$ stands for the input features, and $(a_0, a_1, \ldots, a_n)$ the coefficients. The Kolmogorov–Gabor polynomial arrangement involving only two parameters can be expressed as [53]:

$$\widehat{y}_i = a_o + a_1 x_i + a_2 x_j + + a_3 x_i^2 + a_4 x_j^2 + a_5 x_i x_j \tag{11}$$

For two features having an M dataset, using Equation (11), the following matrix can be generated by relating the input with the output [53].

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{1p} & x_{1q} & x_{1p}x_{1q} & x_{1p}^2 & x_{1q}^2 \\ 1 & x_{2p} & x_{2q} & x_{2p}x_{2q} & x_{2p}^2 & x_{2p}^2 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 1 & x_{Mp} & x_{Mp} & x_{Mp}x_{Mq} & x_{Mp}^2 & x_{Mq}^2 \end{bmatrix} \begin{bmatrix} a_o \\ a_1 \\ \vdots \\ a_N \end{bmatrix} \tag{12}$$

Representing the y_i column by Y, the input features part of the matrix by X, and the coefficient matrix as A. Equation (12) can be written in short form as:

$$Y = XA \tag{13}$$

By using matrix inversion, the coefficient can be computed from X and Y as [53]:

$$A = \left(X^T X\right)^{-1} X^T . Y \tag{14}$$

2.2.3. Artificial Neural Network

An artificial neural network (ANN), also known as a neuron network, is a systems mathematical model that simulates biological neural networks that operate in the human brain and are capable of learning, prediction, and recognition (Agatonovic-Kustrin and Beresford, 2000) [54]. ANN uses nodes, like neurons, building the same sorts of complex interconnections between them (synapses).

Figure 1 shows the standard multilayer perceptron (ANN) neural network algorithms that use two features, two hidden layers, and one target. The input features are any two drilling parameters (x_1 and x_2) and the target (y, ECD). These input features and targets are used for the multivariable and ANN algorithms.

The ANN computation is built based on forward- and backpropagation networks. During the forward feed, the neurons compute the sum of the weighted inputs and bias. The transfer function will then convert the computed signal to an output signal. In this study, for the ANN and proposed GMDH-ANN algorithms' comparison purpose, the Rectified Linear Unit (ReLu) transfer function was selected in the hidden layers. The forward feed process finally computes the model prediction and loss function as the mean squared error between the model prediction ($\hat{y}_j$) and the actual output (y_j).
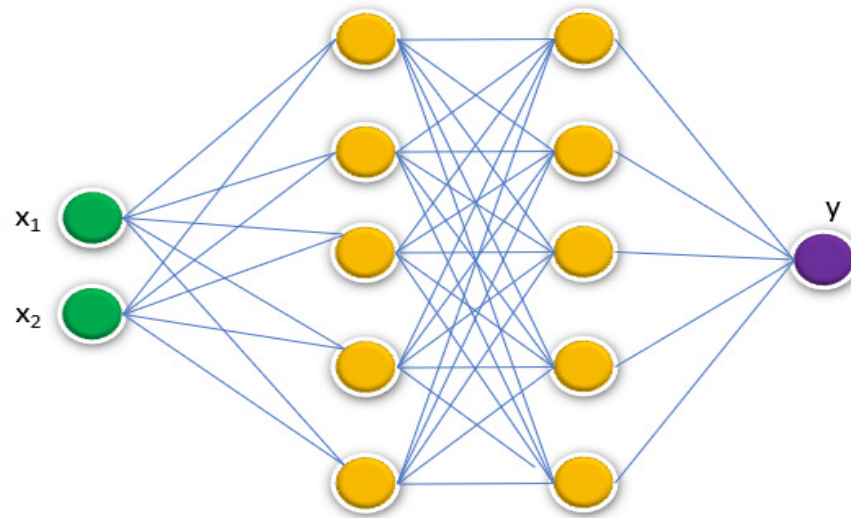
Figure 1. GMDH-uncoupled-normal-input-feature-ANN model.

During backpropagation, training neural network algorithms uses the loss function to improve/update the weights and bias values of the model, which minimizes errors between target values and actual outputs. The process of optimization was carried out until reaching the considered epochs that resulted in a satisfactory model accuracy. The standard method of backpropagation was performed based on a gradient descent method. There are several gradient-descent-based optimizers. However, an Adam optimizer was utilized to evaluate and compare the proposed GMDH-ANN performance with the ANN.

2.2.4. Proposed GMDH-Featured Artificial Neural Network Modeling

Figure 2 illustrates the contrast in architecture between the ANN and GMDH networks. It shows that ANN is a fully connected network, while the GMDH neural network is not. Researchers have compared the performance of the GMDH network modeling with that of the ANN model. Among others, readers may refer to these references (André et al., 2012 [45]; Bernard et al., 2020 [49]; Ahmadi et al., 2015 [42]; Rezaei et al., 2015 [38]).
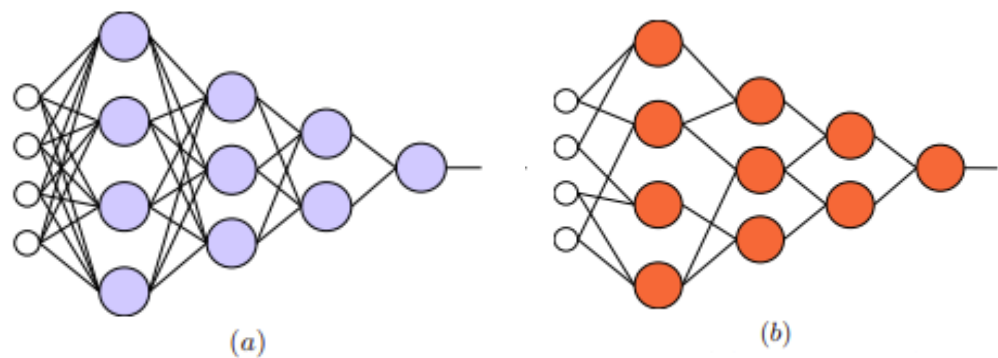

Figure 2. Comparisons between (**a**) ANN and (**b**) GHDH neural networks [André et al. 2012] [48].

This article proposes a fully connected network new GMDH-featured ANN model, as illustrated in Figure 3. The standard ANN (Figure 1) uses two drilling parameters (x_1 and x_2). To include the nonlinear and features' interacting effects, the proposed GMDH used five input features that were generated from the two features (x_1 and x_2). These are x_1, x_2, x_1^2, x_2^2, and x_{1x} x_2. Figure 3 illustrates the architecture along the input features to the proposed GMDH-featured ANN model.
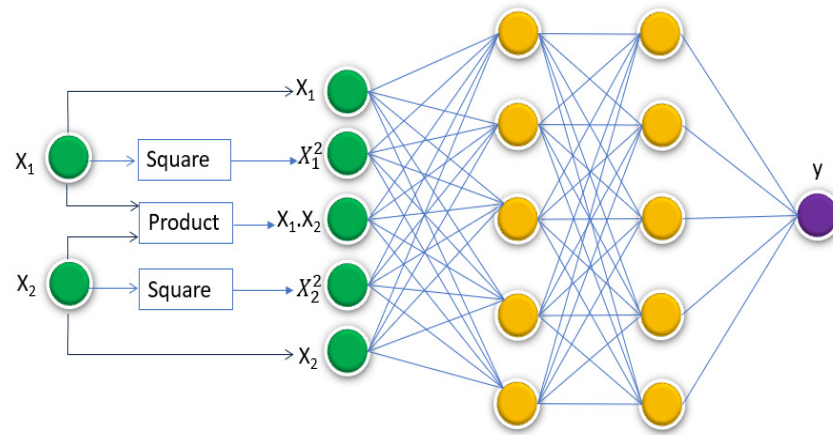
Figure 3. Proposed GMDH-featured network algorithm with inputs (green) and two hidden layers (dark yellow) and output (purple).

2.3. Model Accuracy Performance Evaluation

To assess the accuracy of the model performance, the commonly used statistical parameters, namely mean squared error (MSE) and regression coefficients (R^2) were employed (Montgomery (2019) [55].

2.3.1. Mean Square Error (MSE)

The mean square error (MSE) evaluates the average squared difference between observed and predicted values. When a model exhibits no error, the MSE equals zero. As the model error increases, its value increases. Recognizing the mean square error as the mean square deviation (MSD) denotes the portion of variation that the regression model does not explain.

$$\text{MSE} = \frac{1}{\text{N}} \sum_{i=1}^{\text{N}} \left(y_i^{\text{predicted}} - y_i^{\text{Actual}} \right)^2 \tag{15}$$

2.3.2. Regression Coefficient (R^2)

R-squared (R^2), also referred to as the coefficient of determination, quantifies the extent of variance in the dependent variable (output/target) that can be explained by the independent variable (input features). R^2 values range from 0 to 1. Scoring the R^2 value 1 indicates that the input features correlate with the target 100% and provide the best fit. In contrast, 0 indicates no correlation between the input and target. Mathematically, R^2 is computed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{\text{N}} \left(y_i^{\text{predicted}} - y_i^{\text{Actual}} \right)^2}{\sum_{i=1}^{\text{N}} \left(y_{\text{Actual}}^{\text{Mean}} - y_i^{\text{Actual}} \right)^2} \tag{16}$$

3. Results

This section provides detailed information about the data preprocessing, experimental designs, modeling and performance evaluations, comparisons, and the presentation of the results with accompanying illustrations.

3.1. Data Preprocessing

The measured drilling data was obtained from one of the Norwegian Oil and Gas Exploration operators. Employing the Pandas library, the necessary data preprocessing and feature selection were performed. The final input feature selection was based on a medium correlation factor, which is a correlation factor higher than 0.35. This was just for model performance evaluation purposes. Applying the selection criteria, out of the fifteen available features, the number of input features was reduced to eight and labeled as

D1–D8. The target ECD was labeled as D9, as shown in Table 1. Table 2 details the naming conventions of the labeled features D1–D9. Figure 4 displays the Min–Max normalized dataset scaled according to Equation (17) (Yanchang et al., 2014) [56].

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{17}$$

where, x represents the measured values, while min(x) and max(x) denote the minimum and maximum values, respectively.

Table 1. Correlation coefficients of the selected dataset.

	D1	D2	D3	D4	D5	D6	D7	D8	D9
D1	1								
D2	0.661	1							
D3	0.267	0.567	1						
D4	−0.219	−0.572	−0.995	1					
D5	0.849	0.548	0.189	−0.152	1				
D6	−0.774	−0.782	−0.317	0.300	−0.424	1			
D7	0.569	0.125	−0.203	0.269	0.274	−0.530	1		
D8	0.647	0.184	−0.187	0.260	0.411	−0.524	0.898	1	
D9	0.835	0.686	0.405	−0.378	0.467	−0.925	0.535	0.535	1

Table 2. Input features (D1–D8) and target variable (D9).

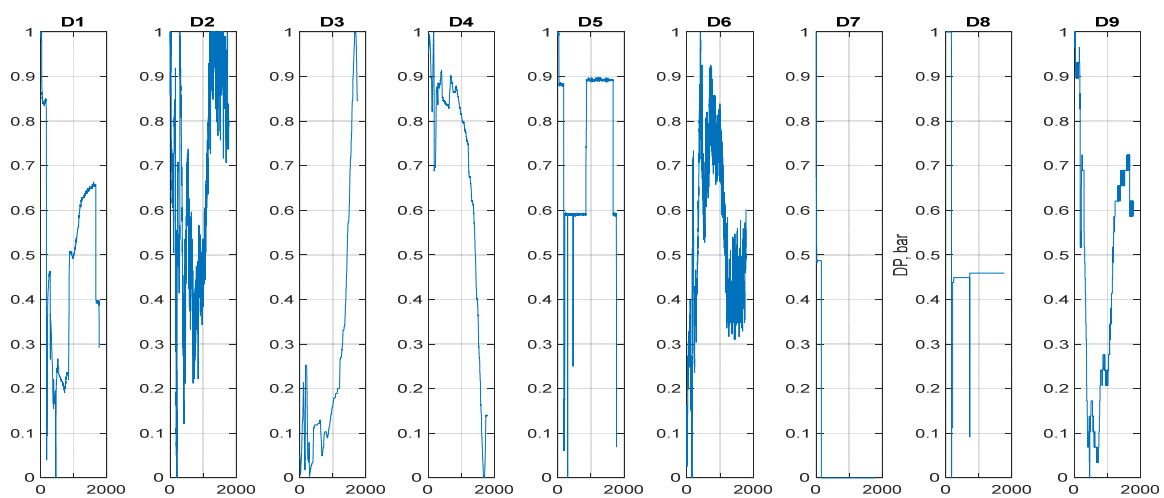| Data Label	Name	Input/Target
D1	Pump Pressure	
D2	Bit Weight	
D3	Block Height	
D4	Bit Position	
D5	Flow Injection Rate	Input Features
D6	Hook Load	
D7	Rate of Penetration (ROP)	
D8	Revolution per Minute (RPM)	
D9	DHT001 EMW	Target Parameter

Figure 4. The normalized drilling input features (D1 to D8) and target dataset (D9).

3.2. Experimental Design and Results

As indicated in the correlation in Table 1, D1 and D6 demonstrated a higher correlation with the target parameter (D9). Hence, utilizing D1 and D6 as input features, as well as the

combination of either D1 or D6 with the other features, Experiment A (#1–7) and Experiment B (#8–13) were designed to assess the performance of MVR and GMDH. Furthermore, the design of Experiment C (#14–18) involved the selection of two features that did not include D1 or D6. Table 3 details the designed experiments.

Table 3. Experiment A, B, and C designs used for GMDH and MVR modeling.

Experiment A	Input Features		Output
#1	D1	D6	
#2	D2	D6	
#3	D3	D6	
#4	D4	D6	D9
#5	D5	D6	
#6	D7	D6	
#7	D8	D6	
Experiment B	**Input Features**		**Output**
#8	D2	D1	
#9	D3	D1	
#10	D4	D1	
#11	D5	D1	D9
#12	D7	D1	
#13	D8	D1	
Experiment C	**Input Features**		**Output**
#14	D2	D8	
#15	D3	D8	
#16	D4	D8	D9
#17	D5	D8	
#18	D4	D5	

3.2.1. Comparison of GMDH vs. MVR

Figures 5–7 show the results obtained from the experiments. The model performance assessment was performed by calculating the regression coefficient. Comparing the three experiments, GMDH modeling achieved a higher R^2 value than MVR modeling. This could be due to the interacting terms.
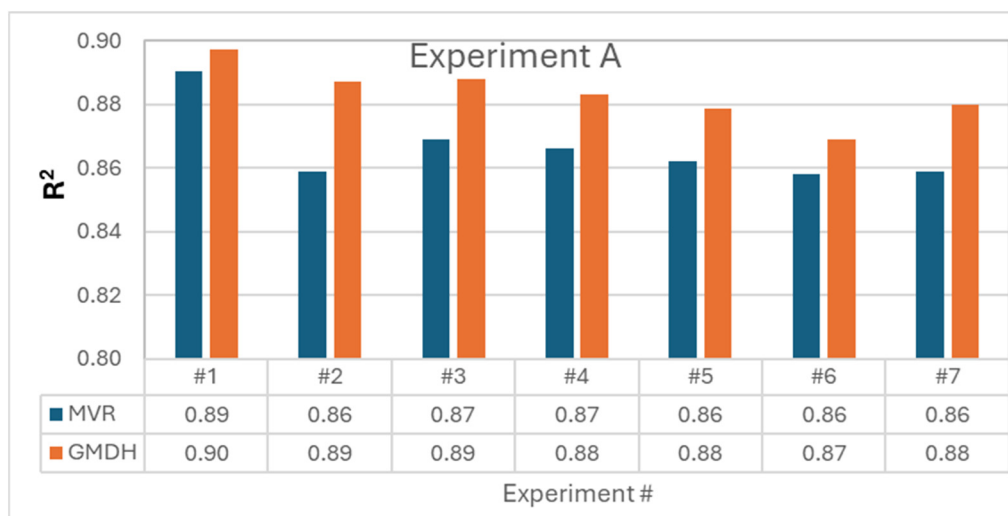

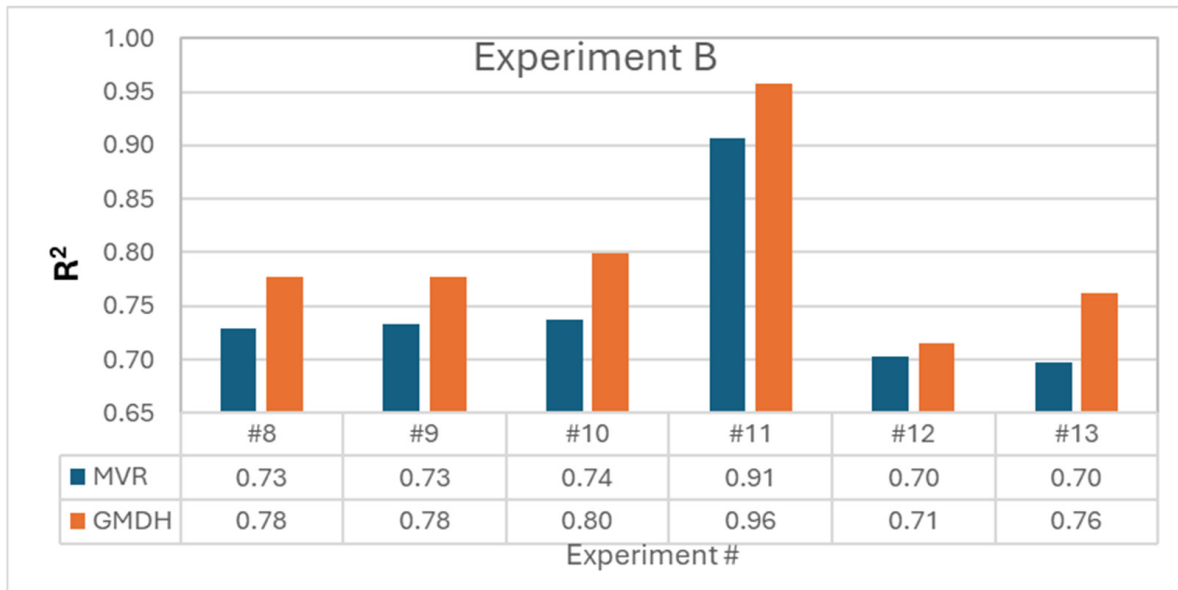**Figure 5.** Experiment A's modeling regression coefficient result.

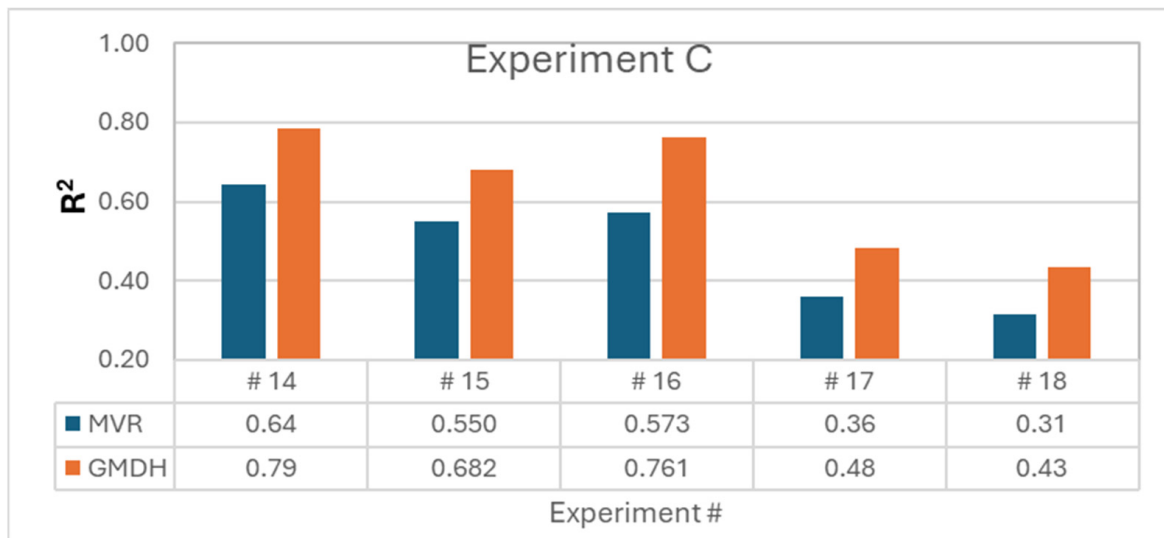**Figure 6.** Experiment B's modeling regression coefficient result.

Figure 7. Experiment C's modeling regression coefficient result.

Figures 8–11 display the selected experimental designs that compared GMDH predictions with measured ECD values. The model captured the data, with Experiment #11 performing exceptionally well. This experimental features (standpipe pressure and flow injection rate) strongly correlated with ECD. Since the flow rate influenced ECD, this correlation likely contributed to the accurate prediction. Standpipe pressure and flow were measured at the surface, whereas ECD was measured downhole. The ECD was calculated/measured based on the pressure losses in the annulus. The presence of the input feature allowed a good prediction of the downhole ECD. More modeling and testing are required to verify the application of these features for downhole ECD estimation.
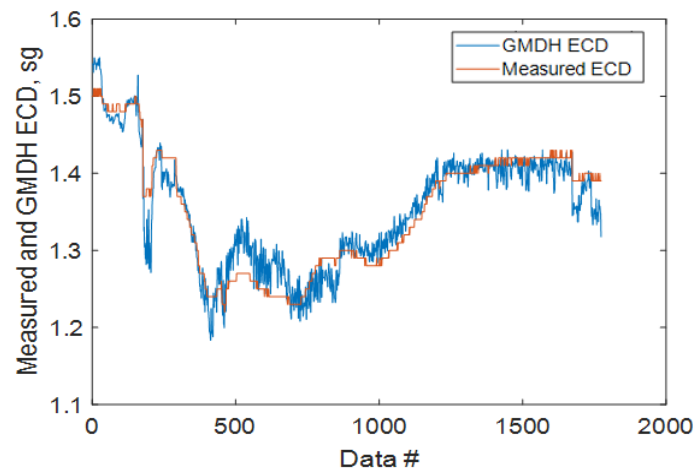
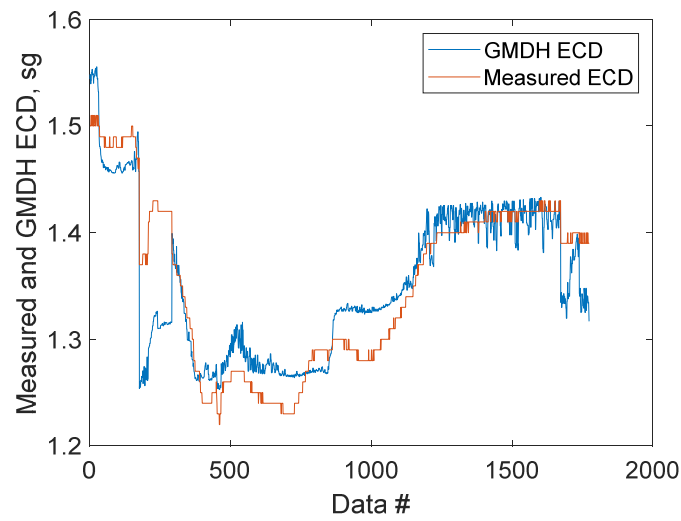**Figure 8.** Experiment A (Input D1, D6).
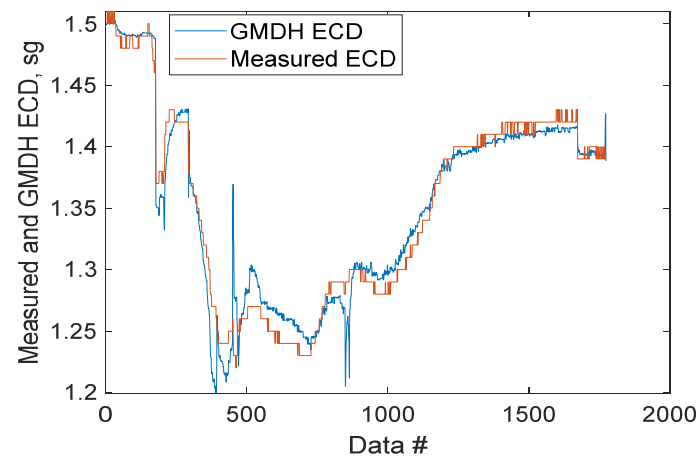

Figure 9. Experiment B (Input D1, D2).

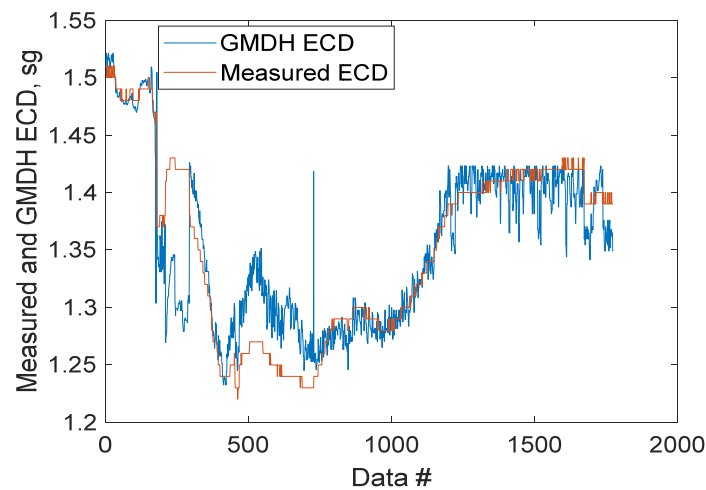**Figure 10.** Experiment B (Input D1, D5).

Figure 11. Experiment C (Input D2, D8).

3.2.2. Comparison of GMDH-Featured ANN vs. Normal-Featured ANN

Figures 1 and 3 depict the two different networks and their modeling performances were evaluated. The standard ANN model used two input parameters (e.g., D1 and D2), while the proposed GMDH-ANN model modified the standard features by incorporating three additional nonlinear feature terms (D1, D2, D1 × D2, D1^2, and D2^2).

The networks were built with two hidden layers comprising five nodes. The Adam optimizer and ReLU activation function were selected without performing detailed hyperparameter tuning.

Table 4 provides the experimental designs for the two networks. Furthermore, given the unsatisfactory model results observed in Experiment C as displayed in Figure 7, four additional experiments were designed (#5a,b to #8a,b) to evaluate the two networks' modeling performances.

Table 4. Input feature selection for the normal ANN and GMDH-featured ANN.

| Experiment	Feature of ANN Model	Features				
#1a	Normal-featured ANN	D1	D2			
#1b	GHDH-featured ANN	D1	D2	D1 × D2	(D1)^2	(D2)^2
#2a	Normal-featured ANN	D1	D6			
#2b	GHDH-featured ANN	D1	D6	D1 × D6	(D1)^2	(D6)^2
#3a	Normal-featured ANN	D1	D5			
#3b	GHDH-featured ANN	D1	D5	D1 × D5	(D1)^2	(D5)^2
#4a	Normal-featured ANN	D2	D8			
#4b	GHDH-featured ANN	D2	D8	D2 × D8	(D2)^2	(D8)^2
#5a	Normal-featured ANN	D3	D8			
#5b	GHDH-featured ANN	D3	D8	D3 × D8	(D3)^2	(D8)^2
#6a	Normal-featured ANN	D4	D8			
#6b	GHDH-featured ANN	D4	D8	D4 × D8	(D4)^2	(D8)^2
#7a	Normal-featured ANN	D5	D8			
#7b	GHDH-featured ANN	D5	D8	D5 × D8	(D5)^2	(D8)^2
#8a	Normal-featured ANN	D4	D5			
#8b	GHDH-featured ANN	D4	D5	D4 × D5	(D4)^2	(D5)^2

The summary of the modeling results is provided in Table 5. Among the eight considered test designs, the GMDH-modified ANN had an improved performance compared with the regular ANN for the (#1b, #5b, and #8b) features. Both networks showed similar results for the remaining designs. The comparisons demonstrated that including interacting terms as part of the input features improved the model performance. Figure 12a,b and Figure 13a,b illustrate the ANN and GMDH-featured ANN results for experiment designs #1 and #5.

Table 5. Modeling result summary of normal ANN and GMDH-featured ANN.

Experiment	Feature of ANN Model	R^2	MSE
#1a	Normal-featured ANN	0.83	0.164
#1b	GHDH-featured ANN	0.96	0.04
#2a	Normal-featured ANN	0.97	0.026
#2b	GHDH-featured ANN	0.98	0.018
#3a	Normal-featured ANN	0.98	0.018
#3b	GHDH-featured ANN	0.98	0.019
#4a	Normal-featured ANN	0.87	0.125
#4b	GHDH-featured ANN	0.87	0.123
#5a	Normal-featured ANN	0.78	0.213
#5b	GHDH-featured ANN	0.87	0.121
#6a	Normal-featured ANN	0.82	0.179
#6b	GHDH-featured ANN	0.82	0.179
#7a	Normal-featured ANN	0.57	0.432
#7b	GHDH-featured ANN	0.56	0.437
#8a	Normal-featured ANN	0.69	0.303
#8b	GHDH-featured ANN	0.75	0.247
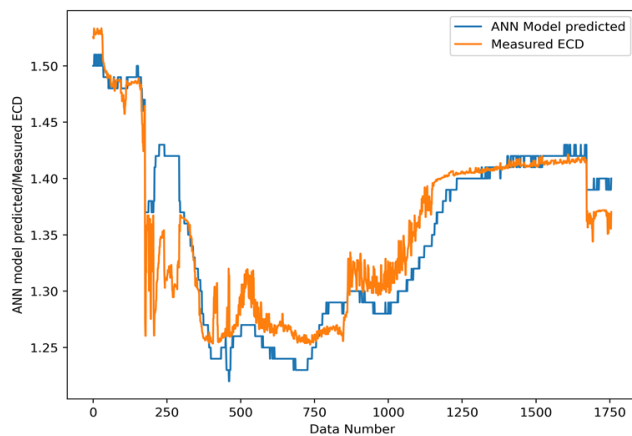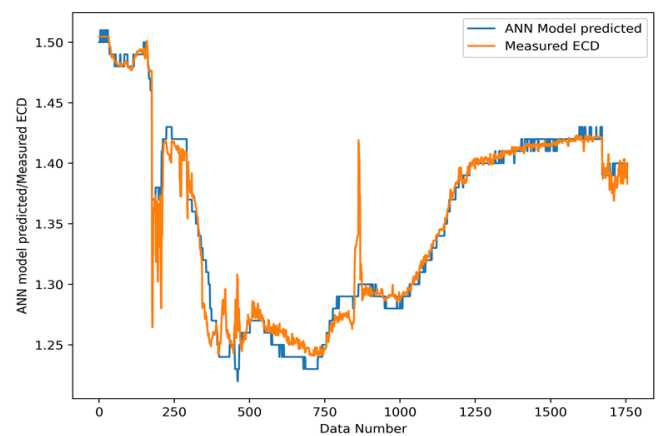
Figure 12. (**a**) Experiment 1a: normal-featured ANN D1–D2 (R^2 = 0.83, MSE = 0.16492). (**b**) Experiment 1b: GMDH-featured ANN D1–D2 (R^2 = 0.96, MSE = 0.0403).
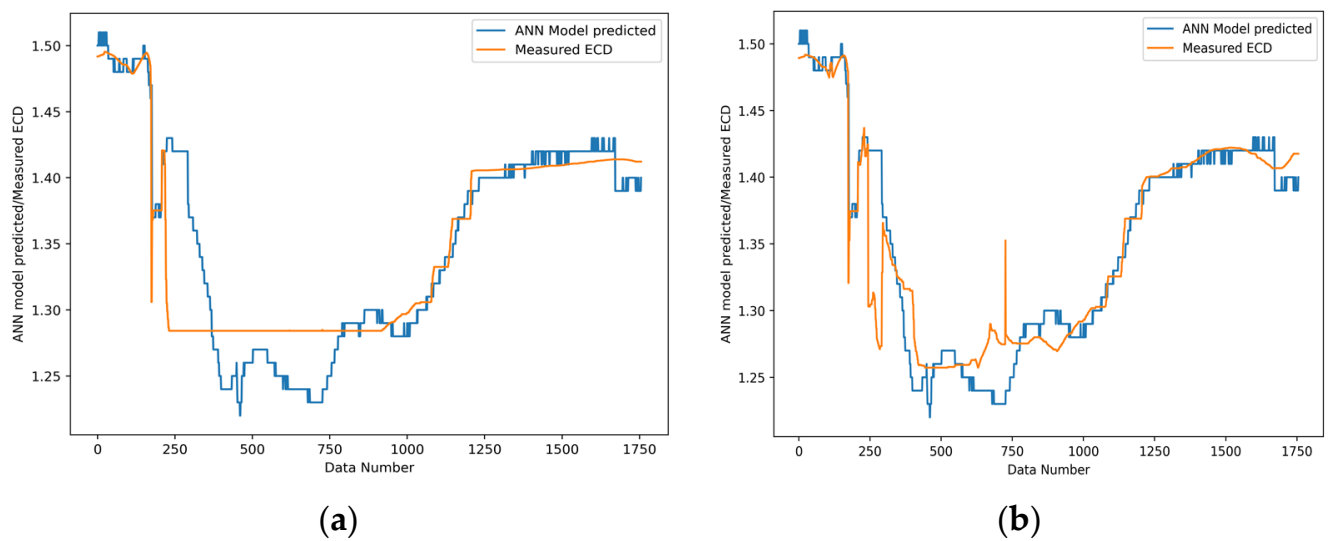
(**a**)   (**b**)

Figure 13. (**a**) Experiment 5a: normal-featured ANN D3–D8 (R^2 = 0.788, MSE = 0.213). (**b**) Experiment 5b: GMDH-featured ANN D3–D8 (R^2 = 0.87, MSE = 0.121).

4. Discussion

Accurately predicting the equivalent circulating density (ECD) during tripping in/out and drilling operations is crucial in ensuring safe and cost-effective well drilling. There are several empirical and physics-based hydraulics models available in the literature [3–20]. However, the application of the models is limited to the considered assumptions and model controlling parameters. Therefore, it is common to practice calibrating the model with a measured dataset (Lohne et al., 2008) [23]. Recent research has focused on using data-driven modeling techniques applied in diverse fields including the petroleum industry.

This study explored the performance of multivariable regression (MVR), the Group Method of Data Handling (GMDH), a standard Multilayered Perceptron (ANN), and the proposed GMDH-featured ANN model to predict ECD based on drilling parameters.

Before implementing the machine learning algorithms, the field drilling dataset was preprocessed to make it clean and to select the appropriate features. The data used in this study included surface and downhole parameters. Appropriate Python libraries were employed for data preprocessing and feature selection.

The first part of the study compared multivariable regression with the GMDH regression. The multivariable regression related two or more independent variables with the target variable. The nature of the regression was an independent linear combination of the variables. However, input features may have nonlinearly varied with the target parameter. Moreover, the input parameters may have had an interaction effect on the target parameters. To study these effects, the GMDH regression was considered and compared with the multivariable regression. The application of the GMDH network has been implemented in several fields [34–44]. The GMDH algorithm, as described in several references [34,35,46,53], utilizes multiple inputs to identify the best combination and generates a quadratic polynomial. An external criterion determines the selection of this optimal combination of two features. Two input features were selected to compare the GMDH method (Equation (8)) with MVR (Equation (1)). A total of 18 experiments were designed (Table 3 in Section 3). Figures 5–7 offer a comparison of the results obtained from the experiments. The results showed that all the GMDH models predicted a higher R^2 compared to the multivariable regression (MVR). This indicated that the nonlinear and interaction terms had a significant effect on the ECD prediction. The degree of the model accuracy performance depended on the correlation of the features with the target variable.

The second part of this study involved comparing the performance of the standard ANN and the proposed GMDH-featured ANN. To ensure consistency in the comparison, both the proposed GMDH-featured ANN (Figure 3) and the fully connected ANN models

(Figure 1) were developed, unlike the standard GMDH network where neurons are not fully connected (Figure 2b). The ANN used only two features, and the proposed GMDH-featured ANN had five inputs generated from the two selected features. Out of the eight experimental designs (Table 4), three of the designs showed that the proposed GMDH-featured ANN exhibited a higher model performance as compared with the ANN, whereas the remaining five designs were the same as shown in Table 5. This could have been due to the insignificant impact of the nonlinear and interacting terms.

Table 6 displays the input features of the standard ANN and the modeling results obtained from the selected three experiments #1, #5, and #8. Table 5 provides a summary of the model results, showing that the proposed GMDH-featured ANN achieved the R^2 values of 0.96, 0.87, and 0.75, respectively, compared to the regular ANN with performance accuracies of 0.83, 0.78, and 0.69, respectively. These results indicated that the proposed GMDH-feature ANN had an enhanced performance compared to the regular ANN. However, more experiments need to be performed to examine the model's performance for other transfer/optimizers and hyperparameters. This will be studied in future work.

Table 6. Comparisons of the four ML algorithms results.

| Experiment | Output/Input Features | | ML Algorithms | R^2 |
	Output	Input		
#1	D9	D1 D2	MVR	0.73
		D1 D2	GMDH	0.78
		D1 D2	Normal-featured ANN	0.83
		D1 D2 D1 $\times$ D2 $(D1)^2$ $(D2)^2$	GHDH-featured ANN	0.96
#5	D9	D3 D8	MVR	0.55
		D3 D8	GMDH	0.68
		D3 D8	Normal-featured ANN	0.78
		D3 D8 D3 $\times$ D8 $(D3)^2$ $(D8)^2$	GHDH-featured ANN	0.87
#8	D9	D4 D5	MVR	0.31
		D4 D5	GMDH	0.43
		D4 D5	Normal-featured ANN	0.69
		D4 D5 D4 $\times$ D5 $(D4)^2$ $(D5)^2$	GHDH-featured ANN	0.75

Based on the GMDH and MVR, the mathematical model derived from Tests #1, #5, and #8 shown in Table 6 is summarized in Tables 7–9. The GMDH model given in Equation (11) is a function of input data (x_i and x_j) and has six coefficients a_0 to a_5. Similarly, the MVR model shown in Equation (4) has input data (x_i and x_j) with three coefficients β_0 to β_2.

Table 7. Test #1 GMDH and MVR coefficients with input (D1 = pump pressure, and D2 = weight on the bit).

GMDH	x_i = D1 and x_j = D2	MVR	x_i = D1 and x_j = D2
Coefficient	Values	Coefficient	Values
a_0	1.494609	β_0	1.234896
a_1	0.001905	β_1	0.001227
a_2	0.056985	β_2	0.007881
a_3	0.000009		
a_4	0.003414		
a_5	-0.000188		
R^2	0.78	R^2	0.72

Table 8. Test #5 GMDH and MVR coefficients with input (D3 = block height, D8 = RPM).

GMDH	x_i = D3 and x_j = D8	MVR	x_i = D3 and x_j = D8
Coefficient	$10^3 \times$	Coefficient	Values
a_0	−0.497074	β_0	0.958565
a_1	0.160636	β_1	0.013871
a_2	0.010814	β_2	0.003009
a_3	−0.002787		
a_4	0.000033		
a_5	−0.000809		
R^2	0.68	R^2	0.55

Table 9. Test #8 GMDH and MVR coefficients with inputs (D4 = bit position and D5 = flow in).

GMDH	x_i = D4 and x_j = D5	MVR	x_i = D4 and x_j = D5
Coefficient	$10^3 \times$	Coefficient	Values
a_0	6.812333	β_0	19.234220
a_1	−0.005875	β_1	−0.007818
a_2	−0.000025	β_2	0.000106
a_3	0.0000013		
a_4	0.0000000002		
a_5	0.00000001		
R^2	0.44	R^2	0.31

The regression coefficients obtained from the GMDH and MVR were rounded to decimal digits and presented in Tables 7–9. Moreover, during MVR modeling, the *p*-test values of the coefficients of all the experiments showed less than 5%. Hence, the MVR model was statistically significant.

5. Conclusions

The reviewed literature studies indicated that the applications of data-driven modeling have shown satisfactory performance in diverse fields. In this study, a total of four machine learning algorithms were employed to model field drilling datasets and to compare their performances.

Based on the considered modeling setup, the results showed that

➢ the GMDH model exhibited a higher model prediction as compared to the MVR. This could have been due to the impact of the interacting features.
➢ the proposed GMDH-featured ANN model prediction was better than the ANN for about 37.5% of the experiments. For the remaining, both methods' predictions were the same.

The proposed GMDH-featured ANN model demonstrated exemplary performance based on the considered dataset. However, in future work, additional modeling with diverse datasets will be conducted to evaluate the model performance and limitations. Furthermore, the commonly used fully-unconnected-GMDH-based network will be compared with the proposed methodology.

Author Contributions: A.M.: conceptualization, methodology, data processing, ML modeling, testing, result analysis, interpretation, and writing. M.B.: methodology, draft manuscript preparation, supervision, review, and editing. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data used in this study is owned by 3rd party and we don't have permission to share it. We have no permission either to mentioned about the data provider's 3rd part name.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mohammad, A.; Davidrajuh, R. Modeling of Swab and Surge Pressures: A Survey. *Appl. Sci.* **2022**, *12*, 3526. [CrossRef]
2. Mohammad, A.; Davidrajuh, R. Review of Four Prominent Works on Swab and Surge Mathematical Models. *Int. J. Simul. Syst. Sci. Technol.* **2022**, *23*. [CrossRef]
3. Burkhardt, J.A. Wellbore pressure surges produced by pipe movement. *J. Pet. Technol.* **1961**, *13*, 595–605. [CrossRef]
4. Schuh, F.J. Computer makes surge-pressure calculations useful. *Oil Gas J.* **1964**, *31*, 96.
5. Fontenot, J.E.; Clark, R. An improved method for calculating swab, surge, and circulating pressures in a drilling well. *Soc. Pet. Eng. J.* **1974**, *14*, 451–462. [CrossRef]
6. Mitchell, R.F. Dynamic surge/swab pressure predictions. *SPE Drill. Eng.* **1988**, *3*, 325–333. [CrossRef]
7. Ahmed, R.M.; Miska, S.Z. Experimental study and modeling of yield power-law fluid flow in annuli with drill pipe rotation. In Proceedings of the IADC/SPE Drilling Conference, Orlando, FL, USA, 4–6 March 2008.
8. Crespo, F.; Ahmed, R.; Saasen, A. Surge and swab pressure predictions for yield-power-law drilling fluids. In Proceedings of the SPE Latin American & Caribbean Petroleum Engineering Conference, Lima, Peru, 1–3 December 2010; p. SPE-138938.
9. Srivastav, R.; Ahmed, R.; Saasen, A. Experimental Study and Modeling of Surge and Swab Pressures in Horizontal and Inclined Wells. In Proceedings of the AADE-17-NTCE-075 Technical Conference and Exhibition, Houston, TX, USA, 11–12 April 2017.
10. Gjerstad, K.; Time, R.W.; Bjorkevoll, K.S. A Medium-Order Flow Model for Dynamic Pressure Surges in Tripping Operations. In Proceedings of the SPE/IADC Drilling Conference, Amsterdam, The Netherlands, 5–7 March 2013.
11. Tang, M.; Ahmed, R.; Srivastav, R.; He, S. Simplified surge pressure model for yield power law fluid in eccentric annuli. *J. Pet. Sci. Eng.* **2016**, *145*, 346–356. [CrossRef]
12. Crespo, F.; Ahmed, R.; Enfis, M.; Saasen, A.; Amani, M. Surge-and-Swab Pressure Predictions for Yield-Power-Law Drilling Fluids. *SPE Drill. Complet.* **2012**, *27*, 574–585. [CrossRef]
13. Erge, O.; Ozbayoglu, E.M.; Miska, S.Z.; Yu, M.; Takach, N.; Saasen, A.; May, R. The Effects of Drillstring-Eccentricity, -Rotation, and -Buckling Configurations on Annular Frictional Pressure Losses While Circulating Yield-Power-Law Fluids. *SPE Drill. Complet.* **2015**, *30*, 257–271. [CrossRef]
14. He, S.; Tang, M.; Xiong, J.; Wang, W. A numerical model to predict surge and swab pressures for yield power law fluid in concentric annuli with open-ended pipe. *J. Pet. Sci. Eng.* **2016**, *145*, 464–472. [CrossRef]
15. Ozbayoglu, E.M.; Erge, O.; Ozbayoglu, M.A. Predicting the pressure losses while the drill string is buckled and rotating using artificial intelligence methods. *J. Nat. Gas Sci. Eng.* **2018**, *56*, 72–80. [CrossRef]
16. Ettehadi, A.; Altun, G. Functional and practical analytical pressure surges model through Herschel Bulkley fluids. *J. Pet. Sci. Eng.* **2018**, *171*, 748–759. [CrossRef]
17. Krishna, S.; Ridha, S.; Vasant, P. Prediction of Bottom-Hole Pressure Differential during Tripping Operations Using Artificial Neural Networks (ANN) in Lecture Notes in Networks and Systems. In Proceedings of the Intelligent Computing and Innovation on Data Science (ICTIDS 2019), Petaling Jaya, Malaysia, 11–12 October 2019; Springer: Singapore, 2021; Volume 118, pp. 379–388.
18. Krishna, S.; Ridha, S.; Vasant, P.; Ilyas, S.U.; Irawan, S.; Gholami, R. Explicit flow velocity modeling of a yield power-law fluid in concentric annulus to predict surge and swab pressure gradient for petroleum drilling applications. *J. Pet. Sci. Eng.* **2020**, *195*, 107743. [CrossRef]
19. Belimane, Z.; Hadjadj, A.; Ferroudji, H.; Rahman, M.A.; Qureshi, M.F. Modeling surge pressures during tripping operations in eccentric annuli. *J. Nat. Gas Sci. Eng.* **2021**, *96*, 104233. [CrossRef]
20. Mohammad, A.; Belayneh, M. New Simple Analytical Surge/Swab Pressure Model for Power-Law and Modified Yield-Power-Law Fluid in Concentric/Eccentric Geometry. *Appl. Sci.* **2023**, *13*, 12867. [CrossRef]
21. Mitchell, R.F.; Miska, S. *Fundamentals of Drilling Engineering*; Society of Petroleum Engineers: Richardson, TX, USA, 2011; ISBN 978-1-61399-951-6.
22. Haaland, S.E. Simple and Explicit Formulas for the Friction Factor in Turbulent Pipe Flow. *J. Fluids Eng.* **1983**, *105*, 89–90. [CrossRef]
23. Lohne, H.P.; Gravdal, J.E.; Dvergsnes, E.W.; Nygaard, G.; Vefring, E.H. Automatic calibration of real-time computer models in intelligent drilling control systems-results from a north sea field trial. In Proceedings of the IPTC 2008: International Petroleum Technology Conference, Kuala Lumpur, Malaysia, 3–5 December 2008; p. cp-148.
24. Mohammad, A.; Belayneh, M.; Davidrajuh, R. Physics-Based Swab and Surge Simulations and the Machine Learning Modeling of Field Telemetry Swab Datasets. *Appl. Sci.* **2023**, *13*, 10252. [CrossRef]
25. Mohammad, A.; Karunakaran, S.; Panchalingam, M.; Davidrajuh, R. Prediction of Downhole Pressure while Tripping. In Proceedings of the 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN), Al-Khobar, Saudi Arabia, 4–6 December 2022; IEEE: Piscataway, NJ, USA, 2023; pp. 505–512.

26. Abdel Azim, R. Application of artificial neural network in optimizing the drilling rate of penetration of western desert Egyptian wells. *SN Appl. Sci.* **2020**, *2*, 1177. [CrossRef]

27. Aliyev, R.; Paul, D. A novel application of artificial neural networks to predict rate of penetration. In Proceedings of the SPE Western Regional Meeting, San Jose, CA, USA, 23–26 April 2019.

28. Alkinani, H.H.; Al-Hameedi, A.T.T.; Dunn-Norman, S.; Lian, D. Application of artificial neural networks in the drilling processes: Can equivalent circulation density be estimated prior to drilling? *Egypt. J. Pet.* **2020**, *29*, 121–126. [CrossRef]

29. Al-Abduljabbar, A.; Gamal, H.; Elkatatny, S. Application of artificial neural network to predict the rate of penetration for S-shape well profile. *Arab. J. Geosci.* **2020**, *13*, 784. [CrossRef]

30. Al-Azani, K.; Elkatatny, S.; Abdulraheem, A.; Mahmoud, M.; Al-Shehri, D. Real time prediction of the rheological properties of oil-based drilling fluids using artificial neural networks. In Proceedings of the SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition, Dammam, Saudi Arabia, 23–26 April 2018; OnePetro: Richardson, TX, USA, 2018.

31. Alnmr, A.; Ray, R.; Alzawi, M.O. A Novel Approach to Swell Mitigation: Machine-Learning-Powered Optimal Unit Weight and Stress Prediction in Expansive Soils. *Appl. Sci.* **2024**, *14*, 1411. [CrossRef]

32. Ray, R.P.; Alnmr, A.N. The Significance of Data Integration in Geotechnical Engineering: Mitigating Risks and Enhancing Damage Assessment of Expansive Soils. *Chem. Eng. Trans.* **2023**, *107*, 541–546.

33. Gurina, E.; Klyuchnikov, N.; Antipova, K.; Koroteev, D. Forecasting the abnormal events at well drilling with machine learning. *Appl. Intell.* **2022**, *52*, 9980–9995. [CrossRef]

34. Ayoub, M.A.; Mohamed, A.A. Estimating the Lengthy Missing Log Interval Using Group Method of Data Handling (GMDH) Technique. *Appl. Mech. Mater.* **2014**, *695*, 850–853. [CrossRef]

35. Mulashani, A.K.; Shen, C.; Nkurlu, B.M.; Mkono, C.N.; Kawamala, M. Enhanced group method of data handling (GMDH) for permeability prediction based on the modified Levenberg Marquardt technique from well log data. *Energy* **2021**, *239*, 121915. [CrossRef]

36. Zhao, L.; Guo, Y.; Mohammadian, E.; Hadavimoghaddam, F.; Jafari, M.; Kheirollahi, M.; Rozhenko, A.; Liu, B. *Modeling Permeability Using Advanced White-Box Machine Learning Technique: Application to a Heterogeneous Carbonate Reservoir*; ACS Omega: Washington, DC, USA, 2023.

37. Mathew Nkurlu, B.; Shen, C.; Asante-Okyere, S.; Mulashani, A.K.; Chungu, J.; Wang, L. Prediction of Permeability Using Group Method of Data Handling (GMDH) Neural Network from Well Log Data. *Energies* **2020**, *13*, 551. [CrossRef]

38. Nyakilla, E.E.; Jun, G.; Charles, G.; Ricky, E.X.; Hussain, W.; Iqbal, S.M.; Kalibwami, D.C.; Alareqi, A.G.; Shaame, M.; Ngata, M.R. Application of Group Method of Data Handling via a Modified Levenberg-Marquardt Algorithm in the Prediction of Compressive Strength of Oilwell Cement with Reinforced Fly Ash Based on Experimental Data. *SPE Drill. Complet.* **2023**, *38*, 452–468. [CrossRef]

39. Li, D.; Armaghani, D.J.; Zhou, J.; Lai, S.H.; Hasanipah, M. A GMDH Predictive Model to Predict Rock Material Strength Using Three Non-destructive Tests. *J. Nondestruct. Eval.* **2020**, *39*, 81. [CrossRef]

40. Alakbari, F.S.; Mohyaldinn, M.E.; Ayoub, M.A.; Muhsan, A.S.; Hussein, I.A. An Accurate Reservoir's Bubble Point Pressure Correlation. *ACS Omega* **2022**, *7*, 13196–13209. [CrossRef]

41. Ayoub, M.A.; Elhadi, A.; Fatherlhman, D.; Saleh, M.; Alakbari, F.S.; Mohyaldinn, M.E. A new correlation for accurate prediction of oil formation volume factor at the bubble point pressure using Group Method of Data Handling approach. *J. Pet. Sci. Eng.* **2021**, *208*, 109410. [CrossRef]

42. Rezaei, F.; Jafari, S.; Hemmati-Sarapardeh, A.; Mohammadi, A.H. Modeling viscosity of methane, nitrogen, and hydrocarbon gas mixtures at ultra-high pressures and temperatures using group method of data handling and gene expression programming techniques. *Chin. J. Chem. Eng.* **2021**, *32*, 431–445. [CrossRef]

43. Rezaei, M.H.; Sadeghzadeh, M.; Alhuyi Nazari, M.; Ahmadi, M.H.; Astaraei, F.R. Applying GMDH artificial neural network in modeling CO_2 emissions in four nordic countries. *Int. J. Low Carbon Technol.* **2018**, *13*, 266–271. [CrossRef]

44. Zhou, L. Prediction of CO_2 adsorption on different activated carbons by hybrid group method of data-handling networks and LSSVM. *Energy Sources Part A Recover. Util. Environ. Eff.* **2019**, *41*, 1960–1971. [CrossRef]

45. Li, R.Y.M.; Fong, S.; Chong, K.W.S. Forecasting the REITs and stock indices: Group Method of Data Handling Neural Network approach. *Pac. Rim Prop. Res. J.* **2017**, *23*, 123–160. [CrossRef]

46. Ahmadi, M.H.; Ahmadi, M.-A.; Mehrpooya, M.; Rosen, M.A. Using GMDH Neural Networks to Model the Power and Torque of a Stirling Engine. *Sustainability* **2015**, *7*, 2243–2255. [CrossRef]

47. Gao, G.; Hazbeh, O.; Rajabi, M.; Tabasi, S.; Ghorbani, H.; Seyedkamali, R.; Shayanmanesh, M.; Radwan, A.E.; Mosavi, A.H. Application of GMDH model to predict pore pressure. *Front. Earth Sci.* **2023**, *10*, 1043719. [CrossRef]

48. Mgimba, M.M.; Jiang, S.; Nyakilla, E.E.; Mwakipunda, G.C. Application of GMDH to Predict Pore Pressure from Well Logs Data: A Case Study from Southeast Sichuan Basin, China. *Nat. Resour. Res.* **2023**, *32*, 1711–1731. [CrossRef]

49. Braga, A.L.; Llanos, C.H.; dos Santos Coelho, L. Comparing artificial neural network implementations for prediction and modeling of dynamical systems. *ABCM Symp. Ser. Mechatron.* **2012**, *5*, 602–609.

50. Kumi-Boateng, B.; Ziggah, Y.Y. Feasibility of using Group Method of Data Handling (GMDH) approach for horizontal coordinate transformation. *Geod. Cartogr.* **2020**, *46*, 55–66. [CrossRef]

51. Prentice, R.L.; Williams, B.J.; Peterson, A.V. On the regression analysis of multivariate failure time data. *Biometrika* **1981**, *68*, 373–379. [CrossRef]

52. Anderson, T.W. *An Introduction to Multivariate Statistical Analysis*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2003; ISBN 978-0-471-36091-9.

53. Ivakhnenko, A.G. Polynomial theory of complex systems. *IEEE Trans. Syst. Man Cybern.* **1971**, *4*, 364–378. [CrossRef]

54. Agatonovic-Kustrin, S.; Beresford, R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J. Pharm. Biomed. Anal.* **2000**, *22*, 717–727. [CrossRef] [PubMed]

55. Douglas, C.M. *Introduction to Statistical Quality Control*, 8th ed.; Wiley: Hoboken, NJ, USA, 2019; ISBN 978-1-119-39930-8.

56. Zhao, Y.; Cen, Y. *Data Mining Applications with R*; Academic Press: Cambridge, MA, USA, 2014; ISBN 978-0-12-411511-8.