PREDICTING PRIMARY COMMODITY PRICES IN THE INTERNATIONAL MARKET: AN APPLICATION OF GROUP METHOD OF DATA HANDLING NEURAL NETWORK

Quang Hung Do, University of Transport Technology, Hanoi, Vietnam
Tran Thi Hoang Yen, VNU University of Economics and Business, Hanoi, Vietnam

ABSTRACT

The fluctuations in primary commodity prices have a significant impact on global economy. Therefore, forecasting price of major commodities prices has been getting much attention both from academic and practitioners' communities. The objective of this study is to develop a model based on group method of data handling (GMDH) technique in one day-ahead forecasting the market prices for major commodities including copper, crude oil, gas and silver. The data on commodities trading were collected from January 2000 to October 2019. In order to validate the effectiveness of the proposed model, other models based on adaptive neuro fuzzy inference system (ANFIS), artificial neural network (ANN), long short-term memory (LSTM) were also developed. The performance indexes including RMSE, MAPE, MAE, R and Theil's U were used to make comparison of the models. The results showed that the proposed model based on GMDH technique outperforms than other methods in prediction of commodity prices. The GMDH-based model provides a promising alternative for price prediction. The GMDH can be a useful tool for economists and practitioners dealing with the forecasting of the commodity price.

Keywords: Primary Commodities, Price Forecasting, GMDH, ANFIS, ANN, LSTM

INTRODUCTION

Key international commodities have been traditionally considered as investment vehicles, traded goods across borders and account for a significant share of international trade. These commodities are considered as potential hedges from inflationary pressures; and monetary substitutes in the event of economic turmoil (Jubinski & Lipton, 2013). The export of primary commodities is an important source of foreign exchange for many nations and regions, especially developing countries, and commodity price movements and volatilities have been associated with investment activities, economic policy planning and formulation. Fluctuations in major commodity prices have a large impact on global economic activities such as commodity-related investments, project appraisals and strategic planning. Therefore, prediction of primary commodity prices with reasonable accuracy has received attentions of economists and practitioners (Husain & Bowman, 2004). Despite of research efforts, commodity price forecasting is still challenging full of volatility, uncertainty and complexity. Primary commodities including agriculture products, raw materials, fuels and base metals, that are

extracted or harvested and also requires very little processing before consumption, account for over 40 per cent of world trade.

Referring to prior studies, several approaches have been proposed for forecasting price. These methods can be grouped into two main categories: (1) methods based on traditional econometric models including time series models, financial models, structural models, conventional regression analysis, and multivariate statistics such as Autoregressive Integrated Moving Average (ARIMA), Generalized Autoregressive Conditional Heteroskedasticity (GARCH) and Seasonal Autoregressive Integrated Moving Average (SARIMA) and (2) methods bases on soft computing techniques including artificial neural network (ANN), support vector machine (SVM), fuzzy logic and heuristic algorithms (Behmiri & Pires Manso, 2013). It is difficult to have a powerful forecasting model using traditional models since these models are derived from strong assumptions and knowledge of input data statistical distributions (Yazdani-Chamzini et al., 2012) and cannot capture the nonlinear patterns hidden in the price series (Yu et al., 2008). Various studies concluded that the artificial intelligence-based models outperform the statistical-based models in time series forecasting. In comparison with traditional econometric techniques, methods based on soft computing techniques (i.e, neural networks) provide a higher degree of robustness and have the capability to forecast volatility (Haidar et al., 2008; Kristjanpoller & Minutolo, 2015). For instance, the ANN is able to forecast with approximately half the mean square error of the econometric method, but both are equally adept at predicting turning points in the time series (Panella et al., 2012). Tapia Cortez et al. (Tapia Cortez et al., 2018) indicated that machine learning techniques have proved to have better performance for forecasting mineral commodity prices compared to all other techniques. The ANN and ANFIS-based models have been developed to predict the oil prices of the OPEC (Lofti & Karimi, 2014). The ANFIS with proposed variables set showed a higher accuracy. The LSTM-based models have been successfully applied to various applications in forecasting prices (Bakir et al., 2018; Jeenanunta et al., 2018; Qu & Zhao, 2019). The mentioned studies have shown the efficiency of the ANN, ANFIS, LSTM in price forecasting. Among computational intelligence techniques, the GMDH is a self-organized system which has the advantages in solving complex nonlinear problems (Amanifard et al., 2008; Ebtehaj et al., 2015).

The GMDH algorithm was first introduced in modeling complex systems in which have a set of data with a multiple inputs and one output (Ivakhnenko, 1971). GMDH has been proved as an effective method in many fields including data mining, knowledge discovery, prediction, optimization and pattern recognition (Teng et al., 2013; Teng et al., 2014). It was showed that GMDH neural network performed better than the traditional forecasting techniques such as single exponential smooth, double exponential smooth, ARIMA and back-propagation neural network (Li et al., 2017).

Our study focuses on forecasting prices of several primary commodities including copper, crude oil, gas and silver. The proposed models utilize historical data and several intelligence techniques to predict price of key international commodities. The best fit model is identified according to several performance criteria including *RMSE, MAPE, MAE, R and Theil's U*. The contribution of this study is twofold. First, we develop an efficient forecasting model based on GMDH technique. The second contribution is to assess the accuracy obtained by different methods namely ANFIS, ANN, GMDH and LSTM in time series forecasting.

GROUP METHOD OF DATA HANDLING (GMDH) NEURAL NETWORK

The GMDH model is a self-organized system which the structure optimized itself according to the data input. The GMDH neural network aims at constructing a function in a feed-forward neural network based on a second-degree transfer function. In the GMDH network, the number of hidden layers and the number of neurons with optimal deterministic transfer function are automatically determined. Accordingly, the optimal model structure is obtained. The relationship between the input and output variables is done by nonlinear functions including Volterra series and Kolmogorov-Gabor polynomial.

If a non-linear system has the input variables $x_1, x_2, \ldots x_n$ (i.e., the lagged time series to be regressed) and the output variable y. The relationship is as follows:

$$y = f(x_1, x_2, \ldots, x_n) \tag{1}$$

The function f can be expressed by the following Kolmogorov–Gabor polynomial:

$$y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} a_{ijk} x_i x_j x_k + \cdots \tag{2}$$

Where n is the number of variables and a_0, a_i, $a_{ij}, a_{ijk}, \ldots$ are coefficients in the polynomial. However, the quadratic form of partial descriptions for two variables is commonly used in the form:

$$G(x_i, x_j) = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2 \tag{3}$$

All pairwise combinations of p lagged time series are considered. Two inputs enter a neuron; the desired output is then obtained. The number of neurons is identified by the number of inputs. Since all pairwise combinations of inputs are considered, the number of neurons is $h = \binom{p}{2}$. An example of GMDH architecture is shown in Figure 1 when there are three layers and four inputs. Since there are four inputs, the number of neurons in a layer is calculated to be six. The coefficients are obtained by the use of equation. In each neuron, the desired output is predicted by the calculated coefficients and inputs. Based on mean square error (MSE) criteria, p neurons are selected and $h - p$ neurons are removed. In Figure 2, four neurons are chosen while two neurons are eliminated from the network. The outputs obtained from selected neurons become the inputs for the successive layer. This step repeats until the last layer. At the last layer, only one neuron is selected. The output at the last layer is the predicted value.

THE MODEL DEVELOPMENT

Data Description

In this study, the research data are daily prices of key international commodities including silver, crude oil (Brent and WTI), copper, and gas from January 3, 2001 to September 30, 2019 (Figure 3). The data is divided into two datasets: the first dataset with 70% of the source data are used for the model development (training) and the other portion (30%) is for testing and evaluating the model.
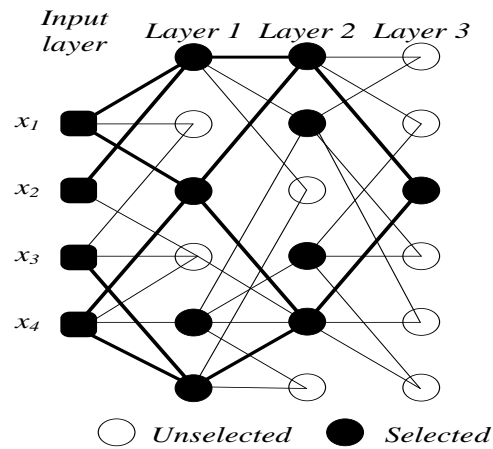
FIGURE 1
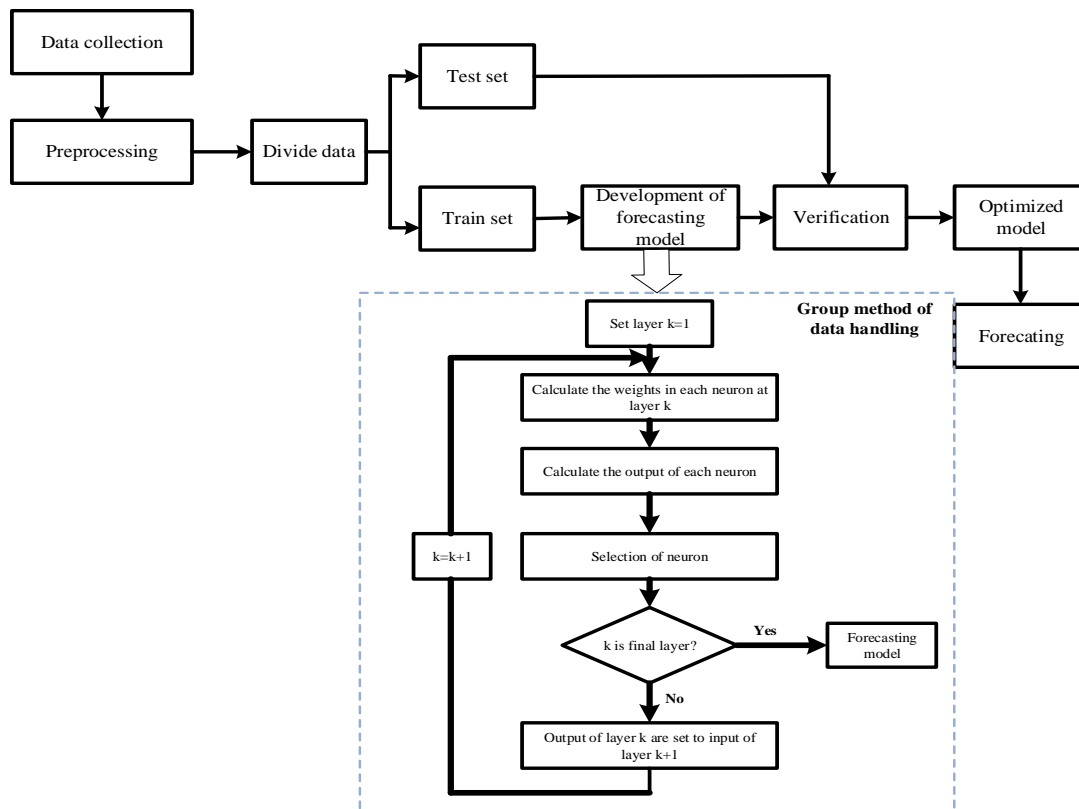**ARCHITECTURE OF GMDH; THE RESEARCH FRAMEWORK**

FIGURE 2
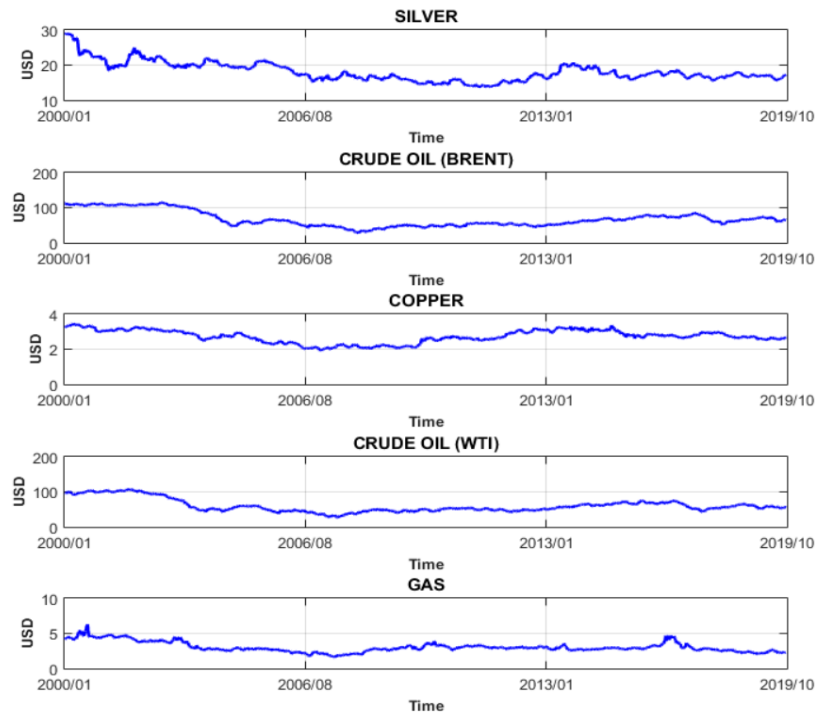**THE DEVELOPMENT STEPS OF FORECASTING MODEL**

FIGURE 3
DAILY PRICES OF KEY INTERNATIONAL COMMODITIES FROM JANUARY 3, 2001 TO SEPTEMBER 30, 2019 ON INTERNATIONAL MARKET

Experimental Structure

In GMDH model development, the maximum number of layers was set to 5; the maximum number of neurons in a layer was equal to 25. When developing GMDH model, one parameter called "Selection pressure" is a proper threshold value to determine the number of neurons in each layer. After calculating the coefficients for all the neurons, those which produce the poorest performance according to the selection criterion (MSE) will be eliminated from the layer. The selection pressure is from 0 (no pressure) to 1 (the maximum pressure of selection). In this study, the selection pressure is set to 0.

Other models were also developed to the research problem. For LSTM model, the number of layer was set equal to 5. The sequential structure with a linear stack of layers was applied to the model development. The main parameters were set as follows: the dimensionality of the output space is 50; the rectified linear unit was used as the activation function. The cost function was mean squared error (MSE). For ANFIS model, the grid partition method was used for FIS (Fuzzy Inference System) generation. In our study, two membership functions were chosen for each input in the model. The Gaussian membership function was used in ANFIS model. The membership function output is linear. Use lagged versions of the variable, five input variables are given $[(x-5), (x-4), (x-3), (x-2), (x-1)]$. For ANN model, the multilayer perceptrons (MLP) was utilized in this study. The used MLP is feed forward fully connected network, with two hidden layers, along with Levenberg-Marquadt training algorithm. Two most commonly nonlinear and linear transfer functions sigmoid, and tangent, were used in first and

second hidden layers, respectively. After a number of trials, the architecture of the MLP network for prediction was determined.

Model Evaluation

To evaluate the performance of the prediction model, several performance indexes were used. These criteria are applied to the developed model to know how well it works. The criteria were used to compare predicted values and actual values. They are as follows:

Root mean squared error ($RMSE$): This index estimates the residual between the actual value and desired value. A model has better performance if it has a smaller$RMSE$. An $RMSE$ equal to zero represents a perfect fit.

$$RMSE = \sqrt{\frac{1}{m}\sum_{k=1}^{m}(t_k - y_k)^2}$$

(4)

Where t_k is the actual (desired) value, y_k is the predicted value produced by the model, and m is the total number of samples.

Mean absolute percentage error ($MAPE$): This index indicates an average of the absolute percentage errors; a model has better performance if it has a smaller $MAPE$.

$$MAPE = \frac{1}{m}\sum_{k=1}^{m}\left|\frac{t_k - y_k}{t_k}\right|$$

(5)

Mean absolute error (MAE): This index indicates how close predicted values are to the actual values. a model with a lower MAE means it has better performance.

$$MAE = \frac{1}{m}\sum_{k-1}^{m}|t_k - y_k|$$

(6)

Correlation coefficient (R): This criterion reveals the strength of relationships between actual values and predicted values. The correlation coefficient has a range from 0 to 1, and a model with a higher R means it has better performance.

$$R = \frac{\sum_{k=1}^{m}(t_k - \bar{t})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}(t_k - \bar{t})^2 \sum_{k=1}^{m}(y - \bar{y})^2}}$$

(7)

Where $\bar{t} = \frac{1}{m}\sum_{k=1}^{m} t_k$ and $\bar{y} = \frac{1}{m}\sum_{k=1}^{m} y_k$ are the average values of t_k and y_k, respectively.

Theil's U-statistic: This index is an accuracy measure that emphasizes the importance of large errors as well as providing a relative basis for comparison with naïve forecasting methods. *Theil's* equation is shown as below:

$$U = \frac{\sqrt{\sum_{k=1}^{m}(t_k - y_k)^2}}{\sum_{k=1}^{m} t_k^2 + \sum_{k=1}^{m} y_k^2}$$

(8)

The *U* value is bound between 0 and 1, with values closer to 0 indicating greater forecasting accuracy.

Through these indexes, the quality (accuracy) of a forecasting model can be estimated by examining the inputs (assumptions) to the model, or by comparing the outputs (forecasts) from the model (Small & Wong, 2002).

RESULTS AND DISCUSSIONS

The performance statistics for each model were calculated and are presented in Table 1. The performance criteria; *RMSE, MAPE, MAE, R* and *Theil's U* obtained by GMDH model were respectively calculated as 0.0368, 0.0098, 0.0265, 0.9949 and 0.0067 for Copper price; 1.1430, 0.0143, 0.8644, 0.9987 and 0.0080 for Brent crude oil; 0.0962, 0.0202, 0.0626, 0.9913 and 0.0154 for Gas; 0.2674, 0.0101, 0.1843, 0.9955 and 0.0073 for Silver; 1.1226, 0.0154, 0.8392, 0.9982 and 0.0089 for WTI crude oil. Theoretically, a forecasting model is regarded as good when *RMSE, MAPE and MAE* are small, R is close to 1 and *Theil's* is close to 0. In Table 1, the value marked in bold and italic indicates the best performance. It can be seen that GMDH archived the best performance at four criteria among five criteria in all forecasts. The performance criteria indicate that the assessed result is highly correlated and precise.

Commodity	Model	RMSE	MAPE	MAE	R	Theil's U-statistic
			TABLE 1			
		PERFORMANCE STATISTICS OF FORECASTING MODELS				
Copper	ANFIS	0.0405	0.0110	0.0297	*0.9942*	0.0073
	ANN	0.0376	0.0100	0.0271	0.9948	0.0068
	GMDH	*0.0368*	*0.0098*	*0.0265*	0.9949	*0.0067*
	LSTM	0.0378	0.0101	0.0274	0.9948	0.0069
Brent crude oil	ANFIS	1.1780	0.0145	0.8775	0.9986	0.0083
	ANN	1.1485	0.0144	0.8664	*0.9987*	0.0081
	GMDH	1.1430	*0.0143*	*0.8644*	*0.9987*	*0.0080*
	LSTM	*1.2169*	0.0156	0.9407	*0.9987*	0.0086
Gas	ANFIS	0.1725	0.0507	0.1482	*0.9908*	0.0271
	ANN	*0.0962*	0.0206	0.0636	0.9914	*0.0154*
	GMDH	*0.0962*	*0.0202*	0.0626	0.9913	*0.0154*
	LSTM	0.1725	0.0545	0.1482	0.9908	0.0271
Silver	ANFIS	0.2798	0.0106	0.1941	*0.9951*	0.0076
	ANN	0.2729	0.0102	0.1866	*0.9951*	0.0075
	GMDH	*0.2674*	*0.0101*	*0.1843*	0.9955	*0.0073*
	LSTM	0.2833	0.0106	0.1950	0.9953	0.0077
WTI crude oil	ANFIS	1.1609	0.0157	0.8601	*0.9981*	0.0092
	ANN	1.1236	0.0155	0.8428	0.9982	*0.0089*
	GMDH	*1.1226*	*0.0154*	0.8392	0.9982	*0.0089*
	LSTM	1.1231	*0.0154*	0.8421	0.9982	*0.0089*

It can be clearly observed from Figures 4-8 that the actual the predicted values obtained by GMDH model are in excellence agreement (Figures 4a to Figure 8a). The corresponding errors between target and predicted values are plotted in Figures 4b to Figure 8b, along with the histogram of errors (Figures 4c to Figure 8c). For silver price forecasting, the values of error were calculated as: $MSE = 0.071478$, $RMSE = 0.26735$, $mean\ error = -0.052876$ and

standard deviation $St.D. = 0.26216$. For Brent price forecasting, the values of error were calculated as: $MSE = 1.3065$, $RMSE = 1.143$, $mean\ error = -0.069199$ and standard deviation $St.D. = 1.1413$. For WTI price forecasting, the values of error were calculated as: $MSE = 1.2602$, $RMSE = 1.1226$, $mean\ error = -0.1082$ and standard deviation $St.D. = 1.1178$. For gas price forecasting, the values of error were calculated as: $MSE = 0.0092529$, $RMSE = 0.096192$, $mean\ error = -0.0067529$ and standard deviation $St.D. = 0.095987$. For copper price forecasting, the values of error were calculated as: $MSE = 0.0013519$, $RMSE = 0.036768$, $mean\ error = -0.0060239$ and standard deviation $St.D. = 0.036283$.

FIGURE 4
SILVER PRICE FORECASTING PERFORMANCE BY GMDH MODEL: (a) FORECASTING AND ACTUAL VALUES, (b) ERROR VALUE AND (c) STANDARD DEVIATION

The comparison between actual values and corresponding output values obtained by the GMDH model are also shown in Figure 9. The figure presents scatter diagrams that illustrate the degree of correlation between predicted values and actual values. In the figure, the 1:1 line was drawn as a reference. In a scatter diagram, the 1:1 line represents that the two sets of data are identical. The more the two data sets agree, the more the points tend to concentrate in the vicinity of the 1:1 line. It may be observed that most predicted values are close to the actual values in Figure 9, and this indicates a good agreement between the forecasting values obtained by the GMDH model and the actual values.
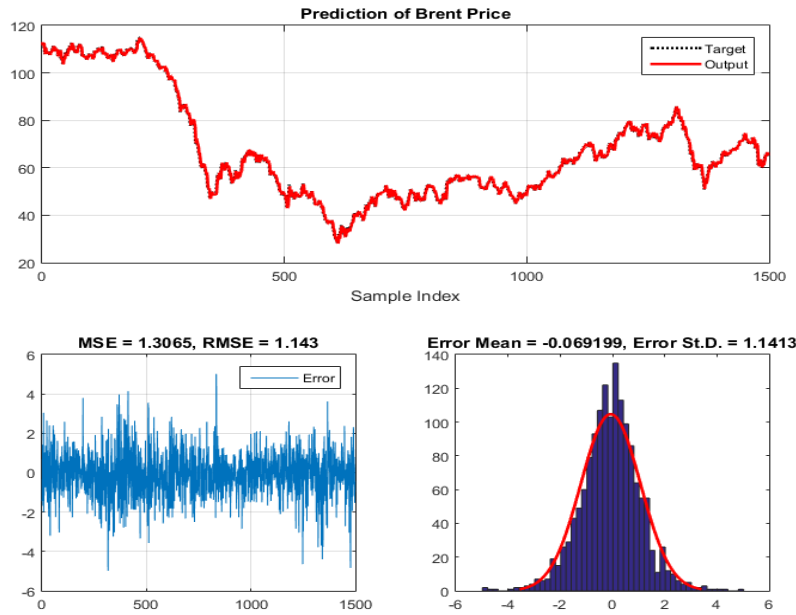
FIGURE 5
BRENT PRICE FORECASTING PERFORMANCE BY GMDH MODEL: (a)
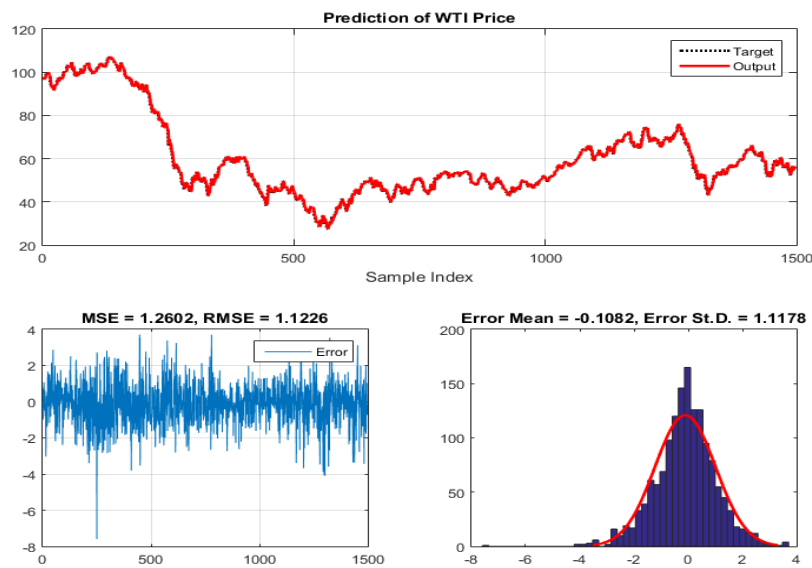**FORECASTING AND ACTUAL VALUES, (b) ERROR VALUE AND (c) STANDARD**
DEVIATION

FIGURE 6
WTI PRICE FORECASTING PERFORMANCE BY GMDH MODEL: (a)
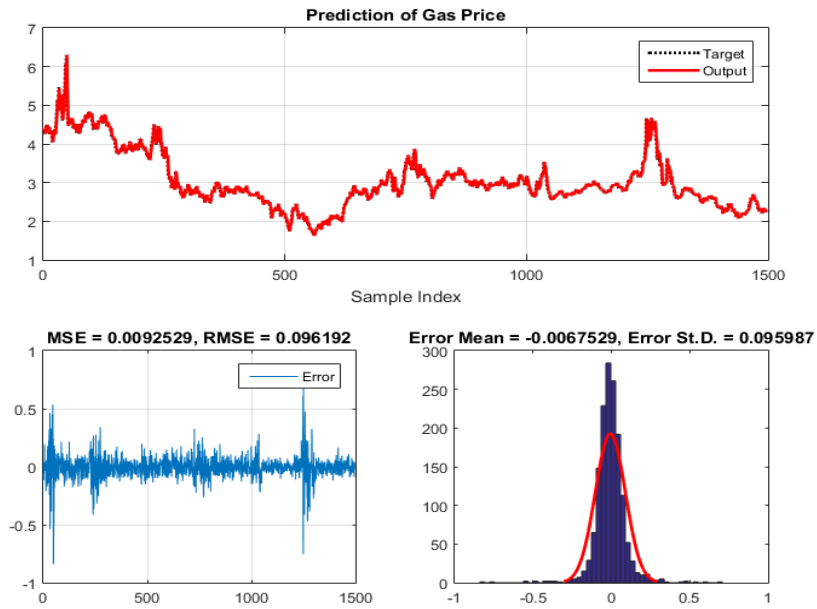**FORECASTING AND ACTUAL VALUES, (b) ERROR VALUE AND (c) STANDARD**
DEVIATION

FIGURE 7
GAS PRICE FORECASTING PERFORMANCE BY GMDH MODEL: (a)
FORECASTING AND ACTUAL VALUES, (b) ERROR VALUE AND (c) STANDARD
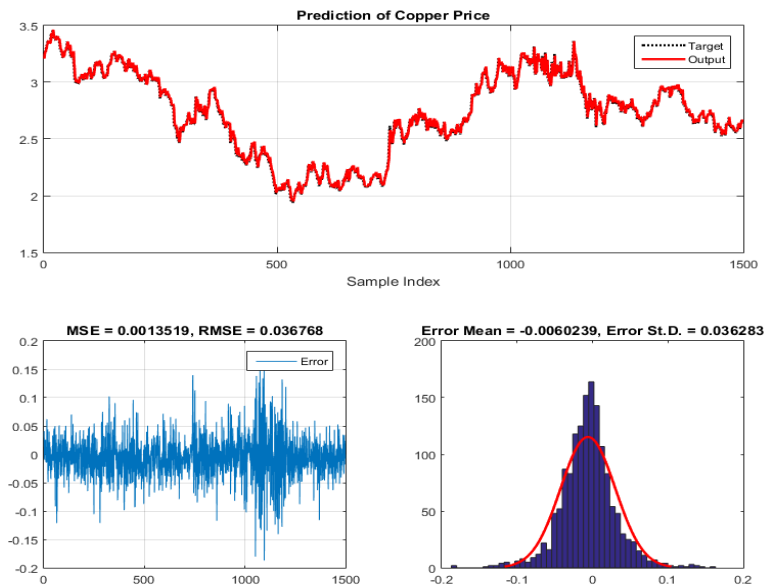**DEVIATION**

FIGURE 8
COPPER PRICE FORECASTING PERFORMANCE BY GMDH MODEL: (a)
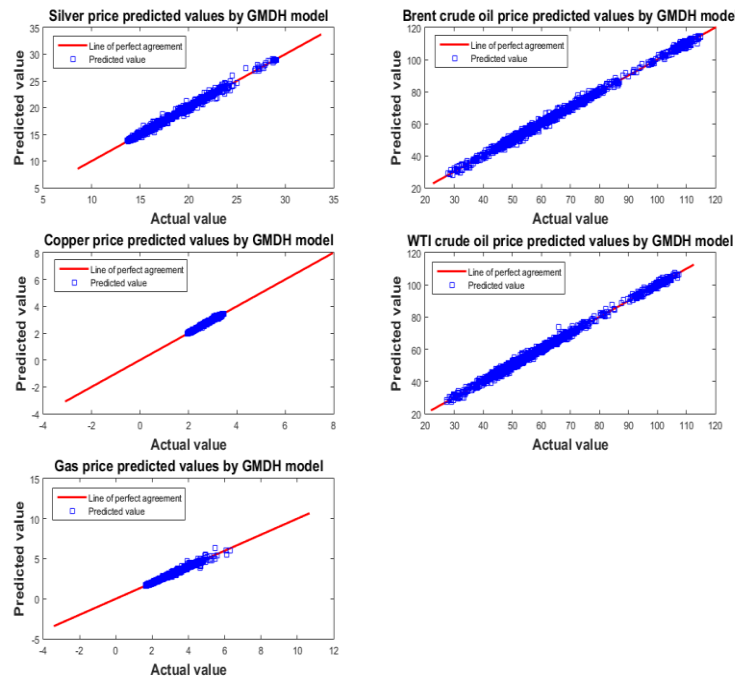**FORECASTING AND ACTUAL VALUES, (b) ERROR VALUE AND (c) STANDARD**
DEVIATION

FIGURE 9
THE SCATTER PLOT OF ACTUAL AND PREDICTED VALUES USING THE GMDH MODEL

Based on the obtained results, it can be concluded that the GMDH model can be used to predict key commodities' price in the market. Regarding forecasting accuracy, the GMDH model is highly appreciated. The GMDH model outperformed the ANFIS, ANN and LSTM models, and the results showed that its prediction outcome is more accurate and reliable. Hence, the GMDH model may be acceptable and good enough to serve as a tool in forecasting commodities' price.

CONCLUSIONS

Financial time series, i.e., commodity price, has characteristics of classical nonlinearity and instability. In this study, we have developed a model based on GMDH technique for forecasting major commodity prices. In order to analyze and compare the ability of the GMDH, other models based on ANFIS, ANN, and LSTM were also applied to the forecasting problem. The simulation results verify the proposed model has better applicability for price time series forecasting. The study findings also demonstrate the forecasting potential of the GMDH in financial applications. The research findings are expected to provide an assistance and forecasting tool for managers and policy makers. Although the model was developed for a specific problem, it will have the potential to be used as a guide for other related problems. We also aspire that our approach can be applied to a great variety of practical problems which need to be solved by artificial intelligence techniques.

REFERENCES

Amanifard, N., Nariman-Zadeh, N., Farahani, M. H., & Khalkhali, A. (2008). Modelling of multiple short-length-scale stall cells in an axial compressor using evolved GMDH neural networks. *Energy Conversion and Management, 49*(10), 2588-2594.

Bakir, H., Chniti, G., & Zaher, H. (2018). E-Commerce price forecasting using LSTM neural networks. *International Journal of Machine Learning and Computing, 8*(2), 169-174.

Behmiri, N. B., & Pires Manso, J. R. (2013). Crude Oil Price Forecasting Techniques: A Comprehensive Review of Literature. *SSRN Electronic Journal, 2013,* 30-48.

Ebtehaj, I., Bonakdari, H., Zaji, A. H., Azimi, H., & Khoshbin, F. (2015). GMDH-type neural network approach for modeling the discharge coefficient of rectangular sharp-crested side weirs. *Engineering Science and Technology, an International Journal, 18*(4), 746-757.

Haidar, I., Kulkarni, S., & Pan, H. (2008). *Forecasting model for crude oil prices based on artificial neural networks*. ISSNIP 2008-Proceedings of the 2008 International Conference on Intelligent Sensors, Sensor Networks and Information Processing, Sydney, NSW, Australia.

Husain, A. M., & Bowman, C. (2004). Forecasting Commodity Prices: Futures Versus Judgment. *IMF Working Papers*, WP/04/41.

Ivakhnenko, A. G. (1971). Polynomial Theory of Complex Systems. *IEEE Transactions on Systems, Man and Cybernetics, SMC-1*(4), 364-378.

Jeenanunta, C., Chaysiri, R., & Thong, L. (2018). *Stock Price Prediction With Long Short-Term Memory Recurrent Neural Network*. International Conference on Embedded Systems and Intelligent Technology & International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES), 1-7. IEEE.

Jubinski, D., & Lipton, A. (2013). VIX, Gold, Silver, and Oil: How do Commodities React to Financial Market Volatility? *Journal of Accounting and Finance, 13*(1), 70-88.

Kristjanpoller, W., & Minutolo, M. C. (2015). Gold price volatility: A forecasting approach using the Artificial Neural Network-GARCH model. *Expert Systems with Applications, 42*(20), 7245-7251.

Li, R. Y. M., Fong, S., & Chong, K. W. S. (2017). Forecasting the REITs and stock indices: group method of data handling neural network approach. *Pacific Rim Property Research Journal, 23*(2), 123-160.

Lofti, E., & Karimi, M. R. (2014). OPEC oil price prediction using ANFIS. *Journal of Mathematics and Computer Science, 10*, 286-296.

Panella, M., Barcellona, F., & D'Ecclesia, R. L. (2012). Forecasting energy commodity prices using neural networks. *Advances in Decision Sciences, 2012*, 1-26.

Qu, Y., & Zhao, X. (2019). Application of LSTM Neural Network in Forecasting Foreign Exchange Price. *Journal of Physics: Conference Series, 1237*(4), 42036.

Small, G. R., & Wong, R. (2002). *The validity of forecasting*. A Paper for Presentation at the Pacific Rim Real Estate Society International Conference, Christchurch, New Zealand.

Tapia Cortez, C. A., Saydam, S., Coulton, J., & Sammut, C. (2018). Alternative techniques for forecasting mineral commodity prices. *International Journal of Mining Science and Technology, 28*(2), 309-322.

Teng, G.-E., He, C.-Z., Xiao, J., & Jiang, X.-Y. (2013). Customer credit scoring based on HMM/GMDH hybrid model. *Knowledge and Information Systems, 36*(3), 731–747.

Teng, G., He, C., & Gu, X. (2014). Response model based on weighted bagging GMDH. *Soft Computing, 18*(2), 2471-2484.

Yazdani-Chamzini, A., Yakhchali, S. H., Volungevičiene, D., & Zavadskas, E. K. (2012). Forecasting gold price changes by using adaptive network fuzzy inference system. *Journal of Business Economics and Management 13*(5), 994-1010, .

Yu, L., Wang, S., & Lai, K. K. (2008). Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. *Energy Economics, 30*(5), 2623-2635.