Gold Price Prediction Based On Yahoo Finance Data Using Lstm Algorithm

1st Windha Mega Pradnya Dhuhita
Informatics
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
windha@amikom.ac.id

2nd Muhammad Farhan Al Farid
Informatics
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
m.farhanalfarid@students.amikom.ac.id

3th Ainul Yaqin
Informatics Engineering
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
ainulyaqin@amikom.ac.id

4th Haryoko
Information Technology
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
haryoko@amikom.ac.id

5th Arif Akbarul Huda
Informatics
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
arif.akbarul@amikom.ac.id

2023 International Conference on Informatics, Multimedia, Cyber and Information Systems (ICIMCIS) | 979-8-3503-0730-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIMCIS60089.2023.10349035

***Abstract*— Gold, a highly valued precious metal, has significant intrinsic value in contemporary society. Therefore, more and more people are embarking on precious metal investments known as gold. Individuals who wish to invest in gold should be vigilant in monitoring the fluctuations in this precious metal's buying and selling prices. Yahoo Finance is a leading online platform that can be used as a reliable reference for monitoring the fluctuations in gold's buying and selling prices. These factors mainly include changes in the closing price, opening price, highest value, and lowest value of gold.**

The observed phenomenon shows that gold prices exhibit high volatility, mainly due to frequent and repeated fluctuations. The LSTM (Long Short-Term Memory) technique, coupled with applying hyperparameter optimization through grid search, enables accurate gold price prediction using historical gold price data. This result is achieved after the grid search procedure, where the lowest error result is 0.00033, generated by the LSTM parameter unit of 200 with a dropout of 0.1.

Keywords— gold price, error, lstm, prediction, grid search

I. INTRODUCTION

Gold is a widely traded commodity in global economic and currency markets. Gold, a highly versatile and sought-after precious metal, is widely used in various industries. These industries include, but are not limited to, electronics, aerospace, medicine, and jewelry[1]. Additionally, it should be noted that this particular entity is commonly used as an investment tool, thus serving as a tool for individuals looking to allocate their financial resources in a strategic manner. Gold is widely recognized as a highly valuable metal due to its inherent properties, which makes it desirable both as a commodity and as a form of currency[2]. The subject in question has unique characteristics as it is classified as both a commodity, a valuable material that can be bought and sold, and a precious metal, a rare and highly sought-after substance. In addition, gold also functions as a form of currency, serving as a medium of exchange in various economic transactions[3]. Gold has an important function as a raw material in various industries, apart from being used as jewelry and jewelry refinement. The importance of gold makes it a worthy choice for investment purposes. Based on existing research, gold investment has shown its efficacy as a safeguard against economic volatility [4].

The role of gold prices in the global economy is very important. In times of bad equity and bond markets, investors often turn to gold as an alternative for financial protection and investment purposes. Various factors, including the US dollar exchange rate, gold production costs, inflation, monetary policy and geopolitical factors can influence gold price fluctuations. Accurate gold price estimates have significant relevance for investors in developing their investment strategies and policies[4][5].

Therefore, a prediction system is needed to estimate gold price movements based on historical data. One approach that can be applied in this prediction is the LSTM (Long Term Memory) method. LSTM is suitable for predicting gold prices because of its ability to process long-term information, learn complex patterns, handle sequential data, and handle multivariate data [5][6].

In the context of gold price predictions, long-term information is very important because macroeconomic factors, long-term trends and global market changes influence gold prices. LSTM can recognize and utilize this long-term information to provide more accurate predictions. In addition, LSTM can capture non-linear relationships and long-term dependencies between input variables, which are important in financial asset price prediction[5][7].

The Long Short-Term Memory (LSTM) model has many favorable characteristics that make it suitable for predicting gold price movements. One of the important characteristics of the system being investigated is the hyperparameters possessed by LSTM [6][7]. Therefore, researchers will use the Grid Search technique to maximize important parameters in the LSTM model. Grid Search is a widely used methodology in Machine Learning model development, where a comprehensive exploration of various parameter values is carried out to identify the optimal combination that produces the best performance. In this context, parameter adjustments, including the number of LSTM layers and units in each layer, dropout will be carried out within a predefined range.

II. EASE OF USE

In the research, several stages will be carried out starting from data collection on the Yahoo Finance website, data preprocessing, splitting the dataset into 3 parts, namely 70% for training data, 20% for testing data and 10% for validation as in previous research[8][9], then continuing with creating an LSTM model and hyperparameter tuning using a CV search grid. The model is tested on testing and validation data; the final stage is evaluation with RMSE. The research stages can be seen in Fig 1 below.

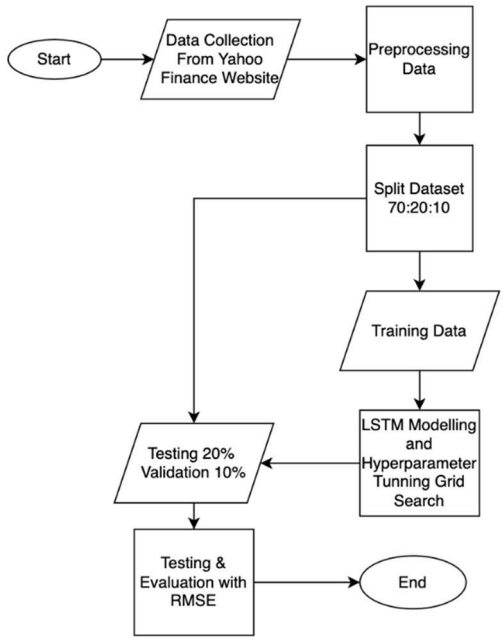

Fig. 1. Research Stages

A. Data collection

This research uses gold price data from Yahoo Finance with the code Gold Aug 23. The dataset was obtained from 2000 to 2023 with a daily period. The data contains features such as Date, Open, High, Low, Close*, Adj Close**, and Volume. The data has a total of 5294 data or rows.

TABLE I. DATASETS

Date	Open	High	Low	Close*	Adj Close**	Volume
Jun 02, 2023	1,977.10	1,982.50	1,947.40	1,952.40	1,952.40	750
Jun 01, 2023	1,963.20	1,983.00	1,954.30	1,978.00	1,978.00	750

The columns or features above have the following meanings:

1. Date is the date when the price was recorded
2. Close* is the Closing Price of Gold in USD
3. Volume is the number of purchases and sales of Gold Commodities
4. Open is the opening price of Gold on that day
5. High is the highest price of gold on that day
6. Low is the lowest price of gold on a particular day
7. Adj Close** is the average closing price

The researcher chose the Close* column to be the target based on the information above. The closing price is the last price of the day and is an important price often used by traders and analysts to make investment decisions.

B. Data Preprocessing

Data preprocessing that will be carried out includes:

1. Check the data type in the dataset.
2. Change the data type of the Date feature to datetime and change the Open, High, Low, Close*, Adj Close**, and Volume features to the float data type.
3. Sort the Date feature from 2000 to 2023.
4. Change the names of the previous features: Date, Open, High, Low, Close*, Adj Close**, and Volume to date, open, high, low, close, adj close, and volume.
5. Look at the correlation between the features, to look for features that will be input for the prediction.
6. Normalize the data using the Min Max Scaler by changing concrete or factual values into various values. The mathematical expression that shows the normalization process is represented by equation 1[10].

$$\acute{x} = \frac{x - xmin}{xmax - xmin} \quad (1)[10] \qquad (1)$$

C. Split Dataset

Divide the dataset into 3 parts, namely training, testing, and validation data. The training data will be used to train the LSTM model, while the testing data will be used to objectively test the model's performance. and validation data are used to optimize model parameters. In this study, researchers divided the data into 70% training, 20% testing, and 10% validation data [11].

D. Creating LSTM Models and Hyperparameter Tuning Grid Search

The design of the Long Short-Term Memory (LSTM) model begins with initializing the parameters required by the Long Short-Term Memory (LSTM) architecture. The LSTM network is a further development of the RNN. It is suggested as a potential solution to the vanishing gradient problem [12]. One widely used method for capturing long-term temporal dependencies in raw time series data is the LSTM scheme [13]. Put simply, LSTM tackles the challenge of long-term dependencies by enhancing its memory cells and employing a gating mechanism to regulate the information flow [12][6]. The parameters covered in this research include the number of LSTM units or number of memory cells (neurons), dropout rate, number of epochs (iterations), and batch size (number of data samples). The use of robust regressors, early stopping, and ADAM optimization algorithms have been incorporated into the Long Short-Term Memory (LSTM) model design. The system's output will be evaluated using the Root Mean Square Error (RMSE) metric, which measures the closeness between predicted and actual values. Next, the hyperparameters of LSTM will be optimized using grid search, which will produce mean_test_score (Average Test Score), std_test_score (Standard Deviation of Test Score), and params (Parameters used) [11][14].

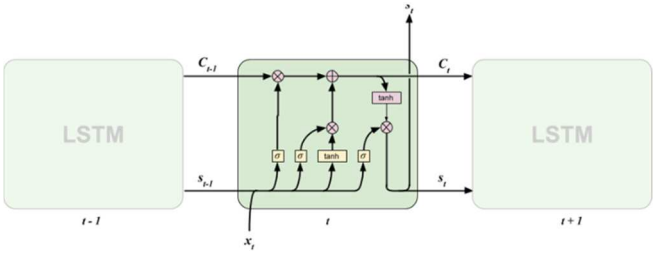

Fig. 2. LSTM Architecture [5]

E. Testing and Evaluation

Once a prediction model is developed, it undergoes a series of rigorous testing and evaluation procedures to assess its performance. The evaluation involves calculating the Root Mean Square Error (RMSE) metric. Root mean square error (RMSE) is a mathematical metric derived from the mean square error (MSE) obtained during the evaluation of a particular method. Root mean square error (RMSE) is a widely used metric in predictive modelling. This metric serves as a means to assess the accuracy of model predictions by comparing them with corresponding observed values. RMSE provides a single numerical value that measures the overall difference between model predictions and actual results by calculating the square root of the average of the squared differences between predicted and observed values. To get small MSE and RMSE values, researchers will use several parameters such as LSTM_units, dropout. For the mathematical equation, mean square error (MSE) can be seen in equation 2 and the mathematical equation Root Mean Square Error (RMSE) can be seen in equation 3 below.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} \quad (P_i - A_i)^2 \qquad (2)[15] \qquad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \quad (P_i - A_i)^2} \qquad (3)[14] \qquad (3)$$

III. RESULTS AND DISCUSSION

The following are the experimental results of testing and evaluation of grid search which can be seen in Table II below.

TABLE II. HYPERPARAMETER GRID SEARCH RESULTS

Lstm Unit	Dropout	MSE	RMSE	Mean Score (std Score)
16	0.1	0.0031	0.0554	-0.000699 (0.001185)
		1.0888 e-04	0.0104	
		1.5224 e-04	0.0123	
		1.5091 e-04	0.0123	
		1.6131 e-05	0.0040	
32	0.1	6.4643 e-04	0.0254	-0.000210 (0.000242)
		3.0150 e-05	0.0055	
		6.2482 e-05	0.0079	
		3.0028 e-04	0.0173	
		8.3352 e-06	0.0029	
64	0.1	5.9298 e-04	0.0244	-0.000153 (0.000222)
		9.1988 e-05	0.0096	
		2.0399 e-05	0.0045	
		4.1289 e-05	0.0064	
		1.8458 e-05	0.0043	
128	0.1	1.1753 e-05	0.0034	-0.000165 (0.000293)
		6.5984 e-06	0.0026	
		1.1133 e-05	0.0033	
		7.5017 e-04	0.0274	
		4.4313 e-05	0.0067	
200	0.1	6.1841 e-05	0.0079	-0.000033 (0.000020)
		1.7453 e-05	0.0042	
		2.0028 e-05	0.0045	
		5.3273 e-05	0.0073	
		1.2946 e-05	0.0036	
16	0.2	0.0025	0.0505	-0.000577 (0.000989)
		2.1620 e-04	0.0147	
		3.6819 e-05	0.0061	
		5.5713 e-05	0.0075	
		2.7644 e-05	0.0053	
32	0.2	0.0013	0.0354	-0.000331 (0.000470)
		2.7695 e-05	0.0053	
		2.8711 e-04	0.0169	
		5.3540 e-05	0.0073	
		3.6213 e-05	0.0060	
64	0.2	8.8812 e-04	0.0298	-0.000284 (0.000321)
		5.5011 e-05	0.0074	
		3.3703 e-04	0.0184	
		1.1001 e-04	0.0105	
		3.1055 e-05	0.0056	
128	0.2	5.9383 e-04	0.0244	-0.000193 (0.000233)
		1.3060 e-05	0.0036	

		2.3312 e-05	0.0048	
		3.2414 e-04	0.0180	
		1.2810 e-05	0.0036	
200	0.2	1.5277 e-04	0.0124	-0.000072 (0.000046)
		2.4453 e-05	0.0049	
		6.5730 e-05	0.0081	
		3.3735 e-05	0.0058	
		8.3668 e-05	0.0091	
16	0.3	0.0020	0.0450	-0.000710 (0.000699)
		1.1957 e-04	0.0109	
		7.5564 e-04	0.0275	
		1.5808 e-04	0.0126	
		4.8904 e-04	0.0221	
32	0.3	0.0023	0.0483	-0.000800 (0.000784)
		2.1079 e-04	0.0145	
		5.5600 e-04	0.0236	
		6.0766 e-04	0.0247	
		2.8682 e-04	0.0169	
64	0.3	4.0926 e-04	0.0202	-0.000247 (0.000125)
		1.0357 e-04	0.0102	
		3.8377 e-04	0.0196	
		1.8181 e-04	0.0135	
		1.5689 e-04	0.0125	
128	0.3	3.4077 e-04	0.0185	-0.000104 (0.000124)
		1.3749 e-05	0.0037	
		2.1143 e-05	0.0046	
		1.1596 e-04	0.0108	
		2.8130 e-05	0.0053	
200	0.3	1.4777 e-04	0.0122	-0.000158 (0.000196)
		1.8929 e-05	0.0044	
		1.9822 e-05	0.0045	
		5.3747 e-04	0.0232	
		6.4219 e-05	0.0080	

The research uses 5 folds of cross validation so that each hyperameter will be carried out 5 times in iterations. Then the mean is the average of the RMSE of each hyperparameter. While std is the standard deviation. So based on the results of the grid search above, it can be seen that the best hyperparameter is in 'dropout': 0.1, lstm_units: 200, because it has the smallest error of 0.000033. The following is the mse graph which can be seen in Fig 3. From the mse graph it can be seen that the model does not experience overfitting and the training loss is close to validation.

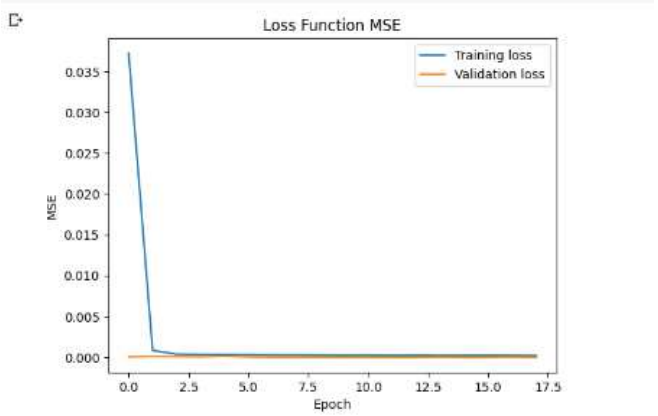

Fig. 3. MSE 'dropout': 0.1, 'lstm_units': 200

Then the RMSE graph can be seen in Fig 4 below. There is no overfitting in the RMSE graph, and the training loss line is close to validation. The RMSE graph better describes the actual value than MSE because it is the root of the MSE value.

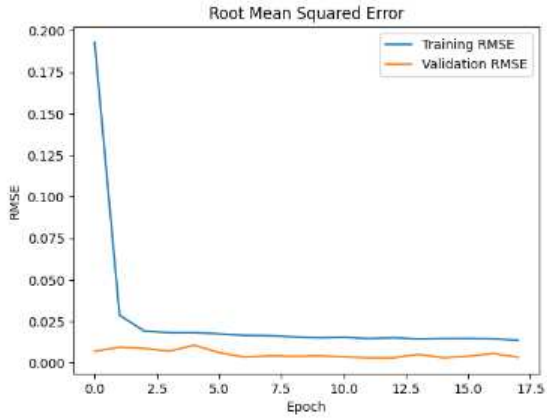

Fig. 4. RMSE 'dropout': 0.1, 'lstm_units': 200

Then in Fig 5 you can see that the predicted value graph follows the actual value.

Fig. 5. Actual & Predicted Value Graph

For more details, see Fig 6 which displays the actual and predicted values in table form. It can also be seen that the predicted value is not much different from the actual value.

	Date	Actual	Predicted
3705	2016-12-27	1137.3	1139.371216
3706	2016-12-28	1139.4	1142.259644
3707	2016-12-29	1156.4	1152.944946
3708	2016-12-30	1150.0	1158.239868
3709	2017-01-03	1160.4	1157.885010
3710	2017-01-04	1163.8	1162.710327
3711	2017-01-05	1179.7	1179.011353
3712	2017-01-06	1171.9	1176.582642
3713	2017-01-09	1183.5	1180.859253
3714	2017-01-10	1184.2	1187.355225
3715	2017-01-11	1195.6	1191.304810
3716	2017-01-12	1198.9	1199.492188
3717	2017-01-13	1195.3	1195.973999
3718	2017-01-17	1212.0	1210.575928
3719	2017-01-18	1211.3	1212.962158
3720	2017-01-19	1200.9	1202.686768
3721	2017-01-20	1204.3	1208.006470
3722	2017-01-23	1215.0	1215.444702
3723	2017-01-24	1210.3	1216.332275
3724	2017-01-25	1197.3	1201.622314
3725	2017-01-26	1189.5	1195.873413
3726	2017-01-27	1188.1	1190.522705
3727	2017-01-30	1193.2	1194.996460
3728	2017-01-31	1208.6	1205.296021
3729	2017-02-01	1205.6	1209.059814
3730	2017-02-02	1216.7	1217.333252
3731	2017-02-03	1218.5	1217.337891
3732	2017-02-06	1230.0	1228.849609
3733	2017-02-07	1234.2	1235.684692
3734	2017-02-08	1237.6	1238.636963

Fig. 6. Actual & Predicted Value Table

IV. CONCLUSION

1. The suitable hyperparameter in the lstm model combines lstm units 200 and dropout of 0.1 which is determined using Grid Search.
2. The lowest error rate resulting from predicting gold prices using LSTM is 0.000033.

ACKNOWLEDGMENT

The author would like to thank AMIKOM Yogyakarta University for providing support in the form of research data, so that this research can be carried out well until the results the researchers want to achieve are obtained.

REFERENCES

[1] I. E. Livieris, E. Pintelas, and P. Pintelas, "A CNN–LSTM model for gold price time-series forecasting," *Neural Comput. Appl.*, vol. 32, no. 23, pp. 17351–17360, 2020, doi: 10.1007/s00521-020-04867-x.

[2] F. D. S. A. G. I. M. C. S. K. Aditya, "Prediksi Harga Emas Menggunakan Metode Time Series Long Short - Term Memory Neural Network," *J. Repos.*, no. Vol 3 No 4 (2021): Agustus 2021, pp. 375–386, 2021, [Online]. Available: https://repositor.umm.ac.id/index.php/repositor/article/view/1378/pdf

[3] M. Owen, V. Vincent, R. Br Ambarita, and E. Indra, "Implementasi Metode Long Short Term Memory Untuk Memprediksi Pergerakan Nilai Harga Emas," *J. Tek. Inf. dan Komput.*, vol. 5, no. 1, p. 96, 2022, doi: 10.37600/tekinkom.v5i1.507.

[4] I. W. Krisna Gita Santika, S. Sa'adah, and P. E. Yunanto, "Gold price prediction using Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM)," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, vol. 4, no. 3, pp. 6–13, 2021, doi: 10.22219/kinetik.v6i3.1253.

[5] V. Riandaru Prasetyo *et al.*, "Prediksi Harga Emas Berdasarkan Data gold.org menggunakan Metode Long Short Term Memory Gold Price Prediction Based on Gold.org Data using the Long Short Term Memory Method," vol. 11, no. September, pp. 623–629, 2022, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[6] A. R. S. Parmezan, V. M. A. Souza, and G. E. A. P. A. Batista, "Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model," *Inf. Sci. (Ny).*, vol. 484, pp. 302–337, 2019, doi: 10.1016/j.ins.2019.01.076.

[7] H. D. Nguyen, K. P. Tran, S. Thomassey, and M. Hamad, "Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management," *Int. J. Inf. Manage.*, vol. 57, no. December 2019, p. 102282, 2021, doi: 10.1016/j.ijinfomgt.2020.102282.

[8] N. Noiyoo and J. Thutkawkornpin, "A Comparison of Machine Learning and Neural Network Algorithms for An Automated Thai Essay Quality Checking," in *2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2023, pp. 482–487. doi: 10.1109/JCSSE58229.2023.10201941.

[9] M. E. Karabulut, K. Vijay-Shanker, and Y. Peng, "CU-UD: text-mining drug and chemical-protein interactions with ensembles of BERT-based models," no. Table II, pp. 1–4, 2021, [Online]. Available: http://arxiv.org/abs/2112.03004

[10] R. Julian and M. R. Pribadi, "Peramalan Harga Saham Pertambangan Pada Bursa Efek Indonesia (BEI) Menggunakan Long Short Term Memory (LSTM)," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 3, pp. 1570–1580, 2021, doi:

10.35957/jatisi.v8i3.1159.

[11] F. I. Sanjaya and D. Heksaputra, "Prediksi Rerata Harga Beras Tingkat Grosir Indonesia dengan Long Short Term Memory," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 7, no. 2, pp. 163–174, 2020, doi: 10.35957/jatisi.v7i2.388.

[12] L. Munkhdalai *et al.*, "An end-to-end adaptive input selection with dynamic weights for forecasting multivariate time series," *IEEE Access*, vol. 7, pp. 99099–99114, 2019, doi: 10.1109/ACCESS.2019.2930069.

[13] S. Du, T. Li, Y. Yang, and S. J. Horng, "Multivariate time series forecasting via attention-based encoder–decoder framework," *Neurocomputing*, vol. 388, pp. 269–279, 2020, doi: 10.1016/j.neucom.2019.12.118.

[14] Z. Alameer, M. A. Elaziz, A. A. Ewees, H. Ye, and Z. Jianhua, "Forecasting gold price fluctuations using improved multilayer perceptron neural network and whale optimization algorithm," *Resour. Policy*, vol. 61, no. September 2018, pp. 250–260, 2019, doi: 10.1016/j.resourpol.2019.02.014.

[15] H. Gunduz, "An efficient stock market prediction model using hybrid feature reduction method based on variational autoencoders and recursive feature elimination," *Financ. Innov.*, vol. 7, no. 1, 2021, doi: 10.1186/s40854-021-00243-3.