

## Preparing Better Data for Oil Price Prediction Using Long Short-Term Memory

Raymond Sunardi Oetama<sup>1✉</sup>

<sup>1</sup> Department of Information System, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Indonesia

### Article Information

#### Article History

**Received** : October 5, 2024

**Revised** : October 25, 2024

**Accepted** : October 31, 2024

#### Keywords:

LSTM;  
Machine Learning;  
Oil price prediction;  
Split Data;  
Window Size.

#### Kata Kunci:

LSTM;  
Machine Learning;  
Prediksi harga minyak;  
Rasio Data;  
Ukuran Window.

### ABSTRACT

*Fluctuating oil prices require a prediction model that can capture complex patterns more accurately than traditional methods. This study aims to apply the Long Short-Term Memory (LSTM) model to predict crude oil prices by assessing the effect of the training-test data ratio and window size on model performance. Daily data from 2000 to 2023 were taken from Yahoo Finance, which was then trained and tested on five data ratios and various window sizes. The evaluation was carried out using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and  $R^2$ . The results show that the 90:10 ratio with a window size of 3 provides the best performance, with an MSE of 6.2100, RMSE of 2.4920, MAE of 1.8430, MAPE of 2.1363%, and  $R^2$  of 0.9606. These findings confirm that LSTM can effectively capture temporal dependencies and outperform traditional statistical methods.*

### ABSTRAK

Harga minyak yang fluktuatif memerlukan model prediksi yang dapat menangkap pola kompleks dengan lebih akurat daripada metode tradisional. Penelitian ini bertujuan untuk menerapkan model Long Short-Term Memory (LSTM) untuk memprediksi harga minyak mentah dengan menilai pengaruh rasio data pelatihan-uji dan ukuran jendela terhadap kinerja model. Data harian dari tahun 2000 hingga 2023 diambil dari Yahoo Finance, yang kemudian dilatih dan diuji pada lima rasio data dan berbagai ukuran jendela. Evaluasi dilakukan menggunakan Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), dan  $R^2$ . Hasil penelitian menunjukkan bahwa rasio 90:10 dengan ukuran jendela 3 memberikan kinerja terbaik, dengan MSE 6,2100, RMSE 2,4920, MAE 1,8430, MAPE 2,1363%, dan  $R^2$  0,9606. Temuan ini mengonfirmasi bahwa LSTM dapat secara efektif menangkap ketergantungan temporal dan mengungguli metode statistik tradisional.

#### Corresponding Author :

Raymond Sunardi Oetama

Department of Information Systems, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Indonesia

Boulevard Raya Gading Serpong, Tangerang

E-mail: [raymond@umn.ac.id](mailto:raymond@umn.ac.id)

## INTRODUCTION

Oil, or crude oil, is a liquid fossil fuel formed from ancient organic matter such as plankton and algae buried under high pressure and temperature for millions of years (Hedgpeth et al., 2021). Oil is one of the primary sources of energy in the world and plays an essential role in various sectors, such as vehicle fuels and the petrochemical industry. As a global commodity, oil prices significantly influence the world economy, from fuel prices to inflation rates (Kilian & Zhou, 2022). Oil prices refer to the cost of purchasing a barrel of crude oil, usually expressed in US dollars per barrel. This price is determined in the global market through futures contracts on commodity exchanges such as the New York Mercantile Exchange. Oil prices fluctuate depending on various factors, including supply and demand dynamics, geopolitical events, natural disasters, global economic conditions, and production decisions from organizations such as the Organization of the Petroleum Exporting Countries (Amer, 2022).

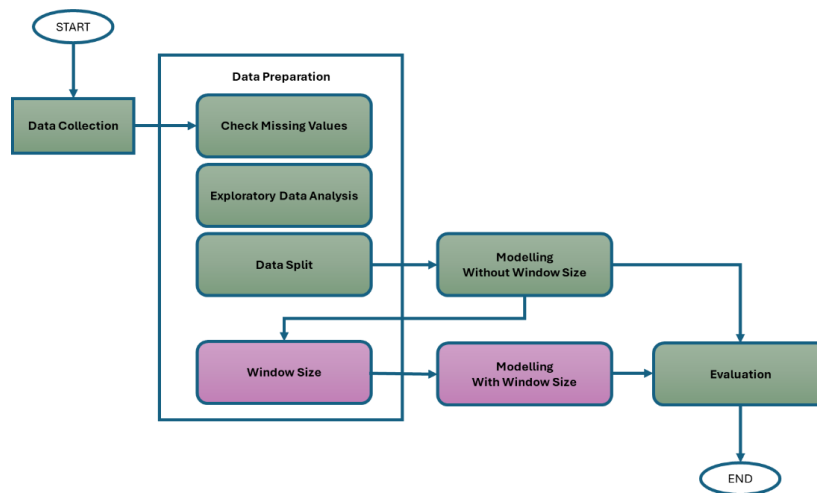
Oil prices are known for their volatility, driven by geopolitical tensions and shifts in supply and demand (Alqahtani & Klein, 2021). While traditional methods have been practical, they often face challenges capturing these fluctuations' complex and unpredictable nature (Li et al., 2021). Predicting such prices requires models that can account for both short-term changes and long-term trends. In this regard, Long Short-Term Memory (LSTM) models offer a promising solution to handle sequential data and retain critical information over time. By addressing these complexities, LSTM models can provide more accurate predictions in a dynamic market like oil.

Several studies discussing the performance of LSTM models in predicting oil prices show significant variations in results. The choice of LSTM for oil price prediction is highly relevant due to its ability to effectively capture complex temporal dependencies in sequential data, such as financial time series (Preeti et al., 2022). Güleriyüz & Özden's (2020) study reported an  $R^2$  of 89%, which means the model could explain 89% of oil price variability, demonstrating competitive performance in capturing price movement patterns. On the other hand, the study of Rusman et al. (2023) shows that their LSTM model produces RMSE values of 6.62 and MAE of 4.96, which indicates that the prediction error is quite large, possibly due to extreme oil price fluctuations or different preprocessing approaches. Another untitled study embedded LSTM model recorded an MSE of 5.877, MAE of 1.629, RMSE of 2.068, and  $R^2$  of 0.843 (Asher, 2023). These differences in results underline the importance of choosing the correct data preprocessing method, especially when splitting data and selecting the appropriate window size. One of the leading causes of inconsistency in these results is the lack of proper window size selection. A window size that is too small may not be sufficient to capture more extended patterns, while a window size that is too large may result in a loss of sensitivity to short-term changes (Guo et al., 2021).

This study focuses on oil price prediction using LSTM. The uniqueness of this study lies in exploring the impact of various data splits and window sizes on the performance of the LSTM model. Data splits refer to the ratio between training and testing data, which affects the model's generalization ability. The window size techniques are also applied, determining how much past information the LSTM uses at each step. This study seeks to provide a deeper understanding of how these data approaches affect the performance of the LSTM model in predicting oil prices, which can provide insights for future studies. These findings contribute to the literature on commodity price prediction, especially in the energy sector, and offer relevant applications for market analysts and policymakers in dealing with global energy market uncertainties.

## RESEARCH METHODS

Figure 1 outlines the data modeling process, starting with data collection and preparation, including checking for missing values and performing EDA to understand trends and correlations. The data is then split into training and testing sets. Two models are built: one without a window size and another incorporating it. Both models are evaluated, and the best is selected for improved predictions.



**Figure 1.** Research Flow

## Data Collection

This study begins with the collection of oil price data from Yahoo Finance. The Yahoo Finance platform was chosen because it offers free access to historical commodity price data such as crude oil with a broad period and reliable data quality. In addition, Yahoo Finance provides an easy-to-process data format for analysis and prediction purposes, making it an efficient and reliable source for academic and professional research (Rosenquist et al., 2024). Data is collected from 2000 to 2023 daily.

## Data Preparation

The data from Yahoo Finance is processed to create a clean dataset. First, missing values are checked and handled if necessary, though none were found. Exploratory data analysis is then performed to understand patterns, visualizing price density, correlations, and trends. The data is split into training and testing sets with ratios ranging from 50:50 to 90:10 to evaluate model performance across different splits. This study emphasizes data preparation, including cleaning, feature engineering, and selecting window sizes, as critical factors for improving model performance, often more impactful than hyperparameter tuning.

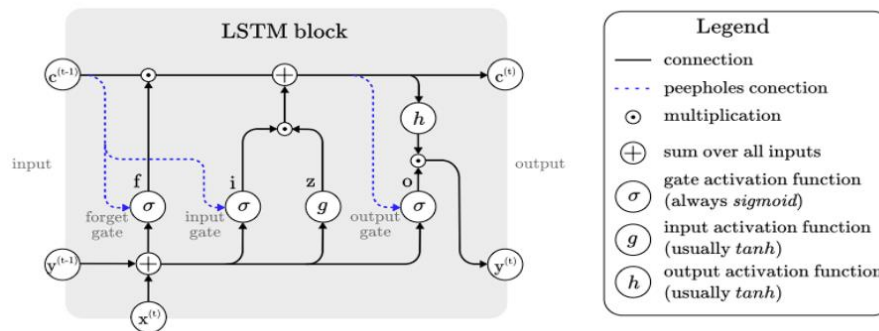
## Modeling

For each data split composition, an LSTM model was built. With its gate architecture consisting of forget, input, and output gates, as shown in Figure 2, LSTM can effectively control the flow of information, deciding which should be kept or discarded (El Bourakadi et al., 2022). It makes LSTM very reliable in processing sequential data such as time series and provides robust performance even when fluctuating data. Its flexibility and ability to capture long-term patterns make it superior in applications such as price prediction (Bukhari et al., 2020). The LSTM model is trained on the training data, and predictions are made using the testing data. Equation (1) to (3) is applied to calculate the forget, input, and output gates (Safa & Oetama, 2024).

$$\text{Forget Gate: } f_t = \sigma(W_f \cdot [y_{t-1}, x_t] + b_f) \quad (1)$$

$$\text{Input Gate: } i_t = \sigma(W_i \cdot [y_{t-1}, x_t] + b_i) \quad (2)$$

$$\text{Output gate: } o_t = \sigma(W_o \cdot [y_{t-1}, x_t] + b_o) \quad (3)$$



**Figure 2.** LSTM architecture (Bagastio et al., 2023)

After the LSTM model is built, the split data is analyzed to identify the best configuration based on prediction error metrics. The model with the best performance across these metrics is selected. Using this optimal split, we apply different window sizes to the data, rebuild the LSTM model, and analyze its performance. The model's performance before and after applying window sizes is compared to determine which configuration performs best. In time series modeling, such as LSTM, the window size refers to the number of previous time points used as input to predict the next time step. It determines how much past data the model uses to capture patterns and make more accurate predictions. Let a set of input  $\{X_{t-n}, X_{t-n+1}, \dots, X_{t-1}\}$  is used to predict  $X_t$ . Where  $n$  is the window size, and  $X_t$  is a value at the time  $t$ . For example, if the window size is 3, then the set of data that is used to predict  $X_t$  is  $\{X_{t-3}, X_{t-2}, X_{t-1}\}$ .

## Evaluation

The evaluation is based on several prediction error metrics, such as MSE, RMSE, MAE, MAPE, and  $R^2$ , computed using formula (4) to (8) (Kristiyanti et al., 2024). MSE measures the average squared difference between predicted and actual values, with smaller values indicating better predictions. RMSE, as the square root of MSE, indicates the prediction error in the same units as the data and tends to be more sensitive to more significant errors. MAE provides the average of absolute errors without distinguishing whether the predictions are over or under the actual values. MAPE represents this error as a percentage, showing how much predictions deviate on average.  $R^2$  reflects how well the model explains the variation in the data, with values closer to 1 suggesting more robust performance.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - f_i)^2 \quad (4)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - f_i)^2} \quad (5)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - f_i| \quad (6)$$

$$MAPE = \frac{100}{N} \sum_{i=1}^N \left| \frac{y_i - f_i}{y_i} \right| \quad (7)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (8)$$

## RESULTS AND DISCUSSION

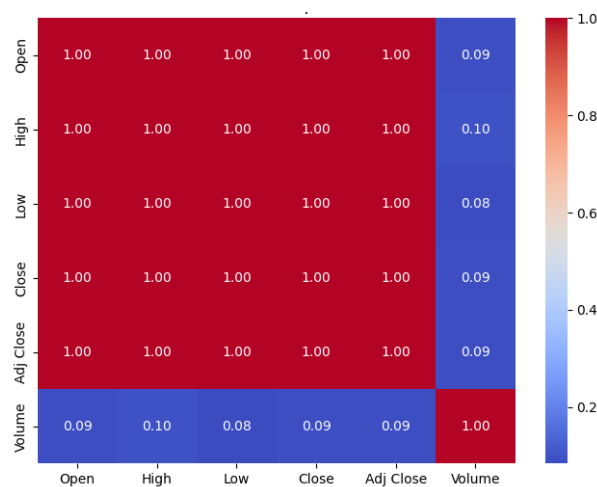
### Exploratory Data Analysis

Oil price and trading volume statistics, with 5,863 data points, are shown in Table 1. The average oil price (Open, High, Low, Close, and Adj Close) is 64-65 USD, while the average trading volume is 303,648. The standard deviation is around 25 for price and over 213,000 for Volume, indicating a significant variation. The minimum price reaches negative values, with a Low of -40.32 and a Close of -37.63, possibly related to extraordinary events such as the 2020 oil crisis. The median value for the price is around 62 USD, and the median Volume is 271,094. The

maximum price reaches 145.29 USD, with the highest Volume of 2,288,230, indicating significant spikes on certain days.

**Table 1. Statistics Information of Oil Dataset**

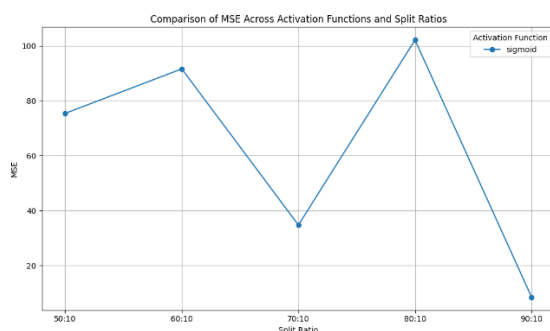
| Statistics | Open     | High     | Low      | Close    | Adj Close | Volume       |
|------------|----------|----------|----------|----------|-----------|--------------|
| count      | 5,863.00 | 5,863.00 | 5,863.00 | 5,863.00 | 5,863.00  | 5,863.00     |
| mean       | 64.19    | 65.19    | 63.12    | 64.18    | 64.18     | 303,648.08   |
| std        | 25.46    | 25.73    | 25.15    | 25.47    | 25.47     | 213,831.05   |
| min        | -14.00   | 13.69    | -40.32   | -37.63   | -37.63    | 0.00         |
| 25%        | 44.87    | 45.91    | 43.84    | 44.79    | 44.79     | 116,576.50   |
| 50%        | 62.10    | 63.09    | 61.25    | 62.10    | 62.10     | 271,094.00   |
| 75%        | 83.69    | 84.99    | 82.26    | 83.74    | 83.74     | 400,408.00   |
| max        | 145.19   | 147.27   | 143.22   | 145.29   | 145.29    | 2,288,230.00 |



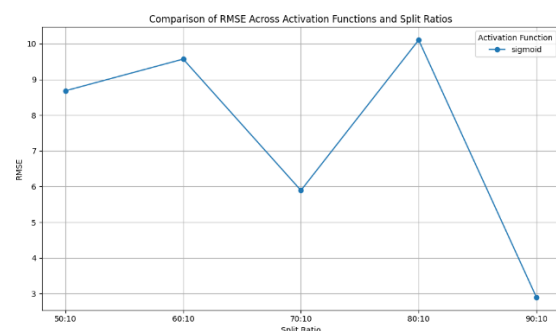
**Figure 3. Correlation Heatmap of Oil Data**

The correlation heatmap in Figure 3 shows that price-related variables such as Open, High, Low, Close, and Adjusted Close are perfectly correlated, meaning they move together. Volume, however, has a weak correlation with prices, indicating it is mainly independent of price movements. Since the Close price perfectly correlates with the others, using it alone simplifies the model without losing important information. Volume adds little value to price prediction.

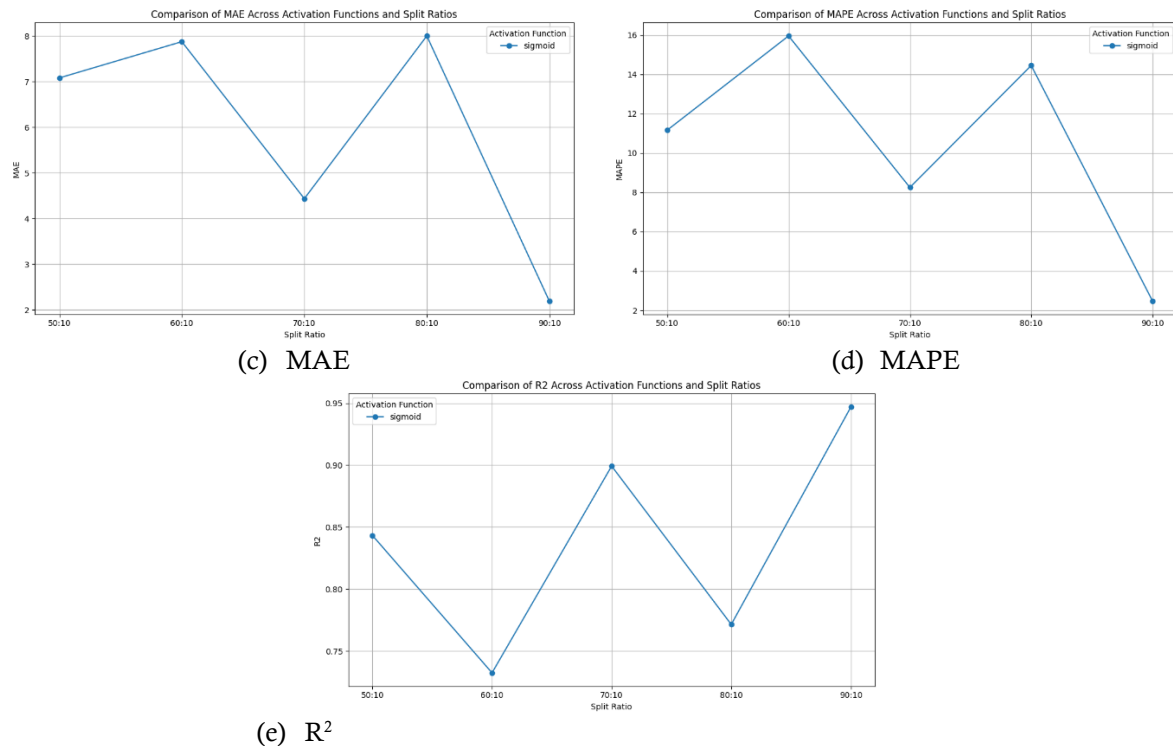
## Modeling and Evaluation



(a) MSE



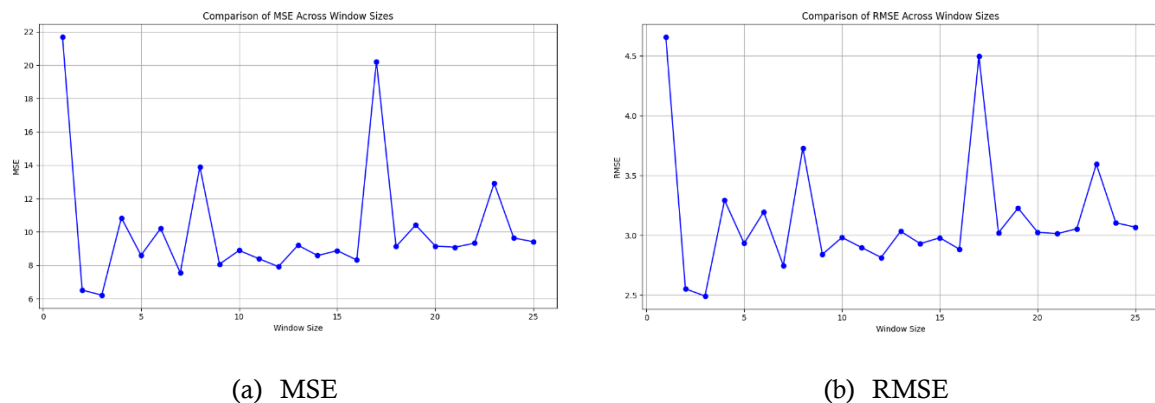
(b) RMSE



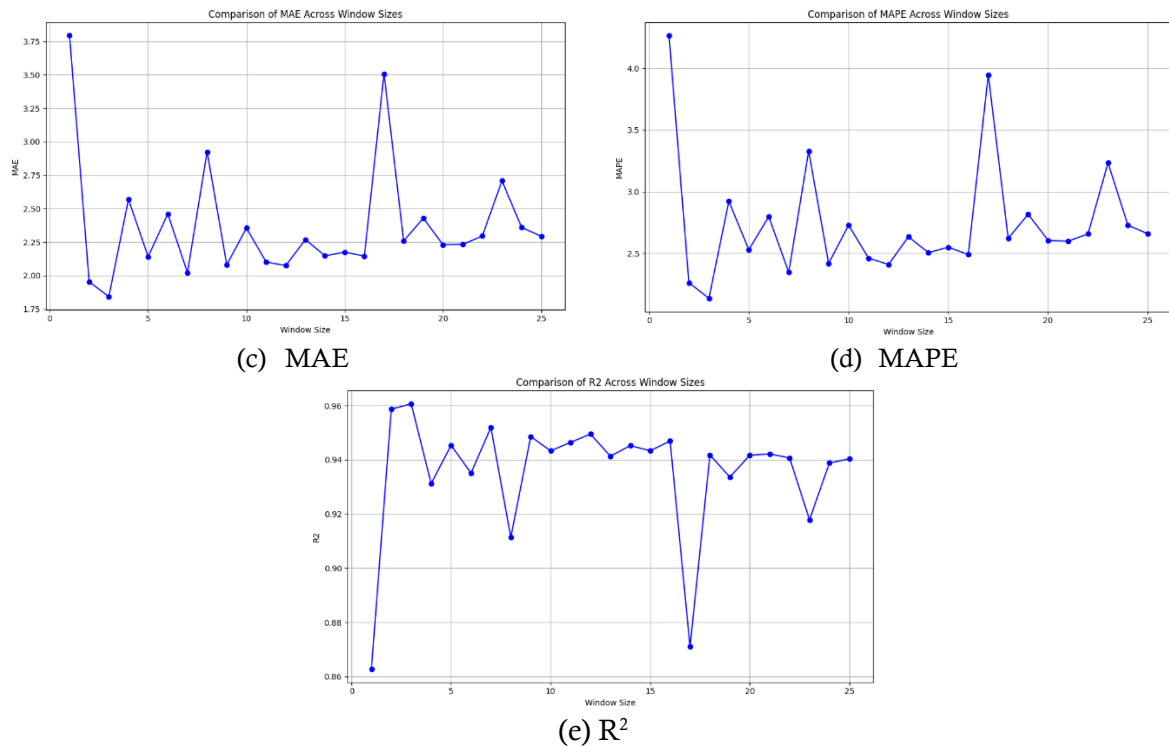
**Figure 4.** Comparison of prediction performance across split data based on (a) MSE, (b) RMSE, (c) MAE, (d) MAPE, and (e) R<sup>2</sup>

The LSTM model has one LSTM layer with 50 units and a sigmoid activation function, followed by a Dense layer for predicting the next closing price. It uses the Adam optimizer, MSE loss function, and a sliding window approach to predict future prices. Trained with a 90:10 data split, the model achieves the best performance across all evaluation metrics in Figure 4. MSE, RMSE, MAE, MAPE, and R<sup>2</sup> fluctuate across data splits, with 90:10 consistently showing the best results.

The graphs in Figure 5 compare error metrics across various window sizes, from 1 to 25. Based on this graph, generally, window size 3 seems to give the best results, with one of the lowest MSE values. The prediction model with window size three can provide more minor prediction errors than other window sizes. All evaluation metrics show good model performance. MSE of 6.2100, RMSE of 2.4920, MAE of 1.8430, and MAPE of 2.1363% indicate that the average prediction error of the model is relatively small, and the error is not too large, even when measured as a percentage error. In addition, the R<sup>2</sup> value of 0.9606 indicates that the model can explain 96.06% of the data variability, which strongly suggests the model has high performance in predicting targets.

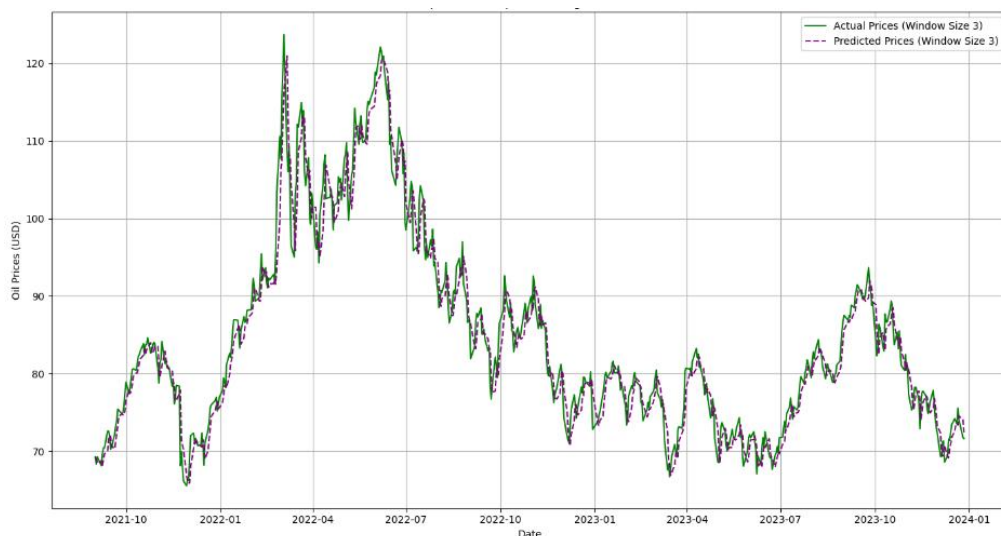




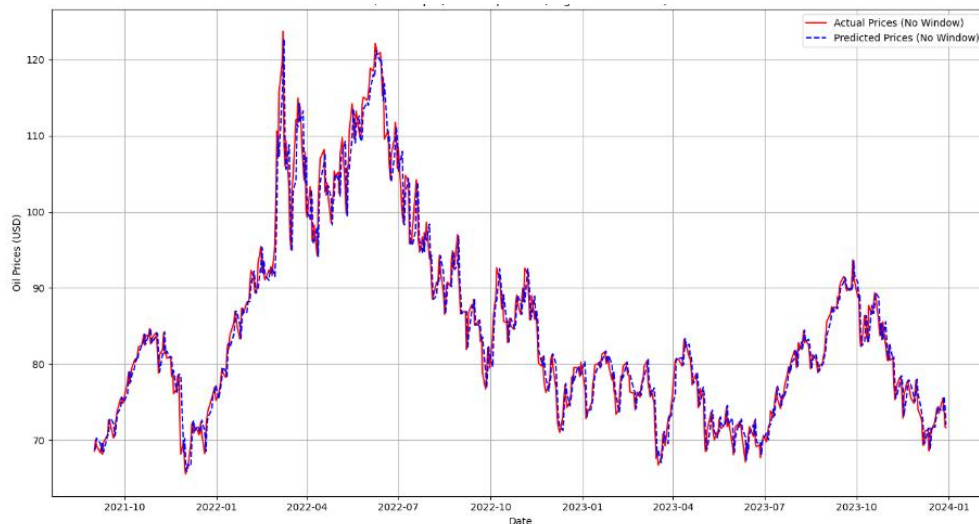


**Figure 5.** Comparison of prediction performance across Windows Size on Split Data 90:10 based on (a) MSE, (b) RMSE, (c) MAE, (d) MAPE, and (e)  $R^2$

The first chart (Figure 6a) shows the predictions using a window size of 3 (green for actual prices and purple for predicted prices). Here, the model's predictions follow the actual price pattern quite well, especially for large fluctuations such as the sharp rise and fall in mid-2022. However, there are slight differences between the actual and predicted prices at specific points, but overall, the predictions are quite consistent. The second chart (Figure 6b) shows the predictions without a window size (red for actual prices and blue for predicted prices). This model also follows the price trend quite well. Still, the predictions appear to be slightly more separated from the actual prices than the model with window size, especially during sharp fluctuations in mid to late 2022.



(a) using windows size 3



(b) without windows size 3

**Figure 6.** Comparison of prediction performance across (a) using windows size three and (b) without windows size.

### Comparison with Other Studies

Gülyüz & Özden (2020) reported an  $R^2$  of 89%, while Rusman et al. (2023) showed their LSTM model produced RMSE of 6.62 and MAE of 4.96. In comparison, this study demonstrates improved performance with a higher  $R^2$  and lower error metrics, likely due to several factors. First, the more comprehensive dataset, covering daily oil prices from 2000 to 2023, provided richer data for the LSTM model to capture both short-term and long-term dependencies effectively. The optimal selection of window size three also contributed to balancing recent trends with historical patterns, enhancing the model's ability to predict price movements accurately. Additionally, improved data preprocessing—such as careful handling of missing values and experimentation with different data splits—helped the model generalize better. In contrast, models without optimized windowing, as noted by Fong et al. (2020), often struggle to capture the complexity of the data. Furthermore, the lower RMSE (6.2100) and MAE (1.8430) in this study compared to Rusman et al. are likely the result of better data handling and model optimization, allowing for better prediction of complex price patterns (Dong et al., 2020).

Asher (2023) shows an MSE of 5.877, MAE of 1.629, RMSE of 2.068, and  $R^2$  of 0.843, while this study results in worse MSE, MAE, and RMSE but better  $R^2$ . A higher  $R^2$  in this study indicates that the LSTM model better captures the data's overall variability and long-term trends. The model effectively learned the broader patterns and relationships in oil price movements, leading to better trend predictions. Since LSTM models are designed to capture patterns in sequential data, a window size allows the model to recognize these patterns more effectively (Rokhsatyazdi et al., 2020). However, the worse MSE, MAE, and RMSE values could indicate that the model struggled with precise point-to-point accuracy, especially in highly volatile. These error metrics measure absolute prediction errors, and higher values suggest that this model had more significant deviations from the actual values in specific instances. This could be due to differences in data preprocessing, feature selection, or model configurations. Asher may have used more refined tuning or handled short-term fluctuations better, leading to lower error values. This study's focus on capturing broader trends resulted in a better  $R^2$  but higher individual prediction errors.

### CONCLUSION

This study successfully shows that the LSTM model is very effective in predicting crude oil prices using variations in window size. The choice of window size and data split significantly impact the performance of an LSTM model. A larger window size can capture broader trends.



Likewise, a more extensive training dataset can help the model generalize better. From the evaluation results, window size 3 provides the best performance based on several prediction error metrics. It is reinforced by the MSE of 6.2100, RMSE of 2.4920, MAE of 1.8430, MAPE of 2.1363%, and  $R^2$  of 0.9606, all of which indicate that the model with window size three can produce the most accurate predictions. In addition, using a 90:10 data split ratio between training data and testing data has proven optimal in maintaining a balance between model training and validation of prediction results.

The limitation of the solution here is that it is focused only on data preparation optimization. For future studies, to enhance the model's performance, it may be helpful to explore advanced LSTM variants, such as Bi-directional LSTM or LSTM with attention mechanisms, as these could improve the model's ability to capture essential patterns in the data. Additionally, considering the inclusion of external factors like macroeconomic indicators or geopolitical events might offer the model a better context for making predictions. These adjustments could provide a more comprehensive understanding of oil price fluctuations and contribute to more accurate forecasting.

## ACKNOWLEDGEMENTS

The author sincerely thanks Universitas Multimedia Nusantara for the support and facilities provided throughout this research. The assistance and resources from the university were greatly appreciated and contributed significantly to the completion of this study.

## REFERENCES

- Alqahtani, A., & Klein, T. (2021). Oil price changes, uncertainty, and geopolitical risks: On the resilience of GCC countries to global tensions. *Energy*, 236, 121541. <https://doi.org/10.1016/j.energy.2021.121541>
- Amer, A.-K. A. (2022). World Oil Prices: Dynamics Of Markets And Politics. *International Journal of Social Sciences*, 4(1).
- Asher, M. (2023). Improving Crude Oil Price Prediction with CEEMDAN-Hilbert Transform and LSTM Models. *Journal of Computer Technology and Software*, 2(1).
- Bagastio, K., Oetama, R. S., & Ramadhan, A. (2023). Development of stock price prediction system using Flask framework and LSTM algorithm. *Journal of Infrastructure, Policy and Development*, 7(3). <https://doi.org/10.24294/jipd.v7i3.2631>
- Bouazizi, T., Lassoued, M., & Hadhek, Z. (2021). Oil price volatility models during coronavirus crisis: Testing with appropriate models using further univariate garch and monte carlo simulation models. *International Journal of Energy Economics and Policy*, 11(1), 281–292. <https://doi.org/10.32479/ijee.10374>
- Bukhari, A. H., Raja, M. A. Z., Sulaiman, M., Islam, S., Shoaib, M., & Kumam, P. (2020). Fractional neuro-sequential ARFIMA-LSTM for financial market forecasting. *IEEE Access*, 8, 71326–71338. <https://doi.org/10.1109/ACCESS.2020.2985763>
- Dong, L., Fang, D., Wang, X., Wei, W., Damaševičius, R., Scherer, R., & Woźniak, M. (2020). Prediction of streamflow based on dynamic sliding window lstm. *Water (Switzerland)*, 12(11), 1–11. <https://doi.org/10.3390/w12113032>
- El Bourakadi, D., Yahyaouy, A., & Boumhidi, J. (2022). Intelligent energy management for micro-grid based on deep learning LSTM prediction model and fuzzy decision-making. *Sustainable Computing: Informatics and Systems*, 35, 100709. <https://doi.org/10.1016/j.suscom.2022.100709>
- Fong, S. J., Li, G., Dey, N., Crespo, R. G., & Herrera-Viedma, E. (2020). Finding an accurate early forecasting model from small dataset: A case of 2019-nCoV novel coronavirus outbreak.

- International Journal of Interactive Multimedia and Artificial Intelligence*, 6(1), 132–140. <https://doi.org/10.9781/ijimai.2020.02.002>
- Guo, Q., Lei, S., Ye, Q., & Fang, Z. (2021). MRC-LSTM: A Hybrid Approach of Multi-scale Residual CNN and LSTM to Predict Bitcoin Price. *Proceedings of the International Joint Conference on Neural Networks*, 2021-July. <https://doi.org/10.1109/IJCNN52387.2021.9534453>
- Hedgpeth, B. M., McFarlin, K. M., & Prince, R. C. (2021). Crude Oils and their Fate in the Environment. *Petrodiesel Fuels*, 891–910. <https://doi.org/10.1201/9780367456252-5>
- Kilian, L., & Zhou, X. (2022). The impact of rising oil prices on US inflation and inflation expectations in 2020–23. *Energy Economics*, 113, 106228. <https://doi.org/10.1016/j.eneco.2022.106228>
- Kristiyanti, D. A., Pramudya, W. B. N., & Sanjaya, S. A. (2024). How can we predict transportation stock prices using artificial intelligence? Findings from experiments with Long Short-Term Memory based algorithms. *International Journal of Information Management Data Insights*, 4(2), 100293. <https://doi.org/10.1016/j.ijime.2024.100293>
- Li, R., Hu, Y., Heng, J., & Chen, X. (2021). A novel multiscale forecasting model for crude oil price time series. *Technological Forecasting and Social Change*, 173, 121181. <https://doi.org/10.1016/j.techfore.2021.121181>
- Makalesi, A., Güleriyüz, D., & Özden, E. (2020). The Prediction of Brent Crude Oil Trend Using LSTM and Facebook Prophet. *Avrupa Bilim ve Teknoloji Dergisi*, 20, 1–9. <https://dergipark.org.tr/tr/pub/ejosat/issue/56357/759302>
- Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Van Le, H., Tran, V. Q., Prakash, I., & Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, 2021(1), 4832864. <https://doi.org/10.1155/2021/4832864>
- Preeti, Bala, R., & Singh, R. P. (2022). A dual-stage advanced deep learning algorithm for long-term and long-sequence prediction for multivariate financial time series. *Applied Soft Computing*, 126, 109317. <https://doi.org/10.1016/j.asoc.2022.109317>
- Rokhsatyazdi, E., Rahnamayan, S., Amirinia, H., & Ahmed, S. (2020). Optimizing LSTM Based Network for Forecasting Stock Market. *2020 IEEE Congress on Evolutionary Computation, CEC 2020 - Conference Proceedings*. <https://doi.org/10.1109/CEC48606.2020.9185545>
- Rosenquist, H., Hasselquist, D., Arlitt, M., & Carlsson, N. (2024). On the Dark Side of the Coin: Characterizing Bitcoin Use for Illicit Activities. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 14538 LNCS, 37–66. [https://doi.org/10.1007/978-3-031-56252-5\\_3](https://doi.org/10.1007/978-3-031-56252-5_3)
- Rusman, J. A., Chunady, K., Makmud, S. T., Setiawan, K. E., & Hasani, M. F. (2023). Crude Oil Price Forecasting: A Comparative Analysis of ARIMA, GRU, and LSTM Models. *2023 IEEE 9th International Conference on Computing, Engineering and Design, ICCED 2023*, 1–6. <https://doi.org/10.1109/ICCED60214.2023.10425576>
- Safa, R. P., & Oetama, R. S. (2024). Hybrid LSTM Model for Predicting Indonesian Telecommunication Companies Stock Price. *2024 International Conference on Artificial Intelligence, Blockchain, Cloud Computing, and Data Analytics (ICoABCD)*, 184–189.