

Comparison of Linear Regression and LSTM (Long Short-Term Memory) in Cryptocurrency Prediction

Marisa Istaltofa^{1*}, Sarwido², Adi Sucipto³

^{1*,2,3}Information Engineering, Faculty of Science and Technology, Nahdlatul Ulama Islamic University Jepara

¹*211240001144@unisnu.ac.id, ²sarwido@unisnu.ac.id, ³adisucipto@unisnu.ac.id

Abstract

Cryptocurrency, particularly Bitcoin, has become a major topic in the financial and digital trading sectors due to its ability to facilitate direct transactions without intermediaries and the transparency offered by blockchain technology. However, the high volatility of Bitcoin prices necessitates accurate prediction methods to support better investment decisions. This research aims to compare the accuracy of Linear Regression and Long Short-Term Memory (LSTM) methods in predicting Bitcoin prices using historical data from Yahoo Finance. The research process begins with the collection of historical Bitcoin price data from September 17, 2014, to July 15, 2024, followed by data processing that includes cleaning and splitting the dataset into training and test data. Linear Regression and LSTM models are applied to the training data and tested to evaluate their performance in price prediction. The findings indicate that the LSTM model significantly outperforms the Linear Regression model in terms of prediction accuracy, achieving much lower Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and a perfect R^2 score of 1.00 on both datasets, alongside an impressive F1 Score of 0.99. In contrast, the Linear Regression model demonstrates higher errors and an F1 Score of 0.88, indicating its limitations in capturing the complexities of Bitcoin price dynamics. These findings suggest that LSTM is more effective in modeling temporal patterns and fluctuations in Bitcoin prices, providing better accuracy and guidance for investors in this highly dynamic market.

Keywords: *Cryptocurrency, Bitcoin, Linear Regression, LSTM, Prediction, Comparison.*

© 2024 Journal of DINDA

1. Introduction

Cryptocurrency has become a frequently discussed topic in recent years. As the first digital currency to use cryptographic systems for direct transactions between two parties without intermediaries, cryptocurrency has seen rapid growth in the financial, business, and trading sectors [1]. Cryptocurrency represents the first implementation of blockchain technology, utilizing a distributed system and consensus-based database with high cryptographic security and transparency. This enables the use of a distributed and immutable ledger, ensuring that every transaction cannot be manipulated, thereby eliminating the need for a trusted third party [2].

One of the most famous cryptocurrencies is Bitcoin, introduced by Satoshi Nakamoto in January 2009. Bitcoin is governed by an open-source software system

that allows anyone to modify it [3]. Since its introduction, Bitcoin has shown remarkable value

growth with significant price fluctuations, peaking in November 2021 at \$68,000 per coin. However, the Bitcoin market exhibits very high volatility, up to 10 times higher than the volatility of foreign exchange rates [4]. An illustration of Bitcoin's growth can be seen below:



Figure 1. Bitcoin Growth ¹, highlighting significant price peaks and fluctuations. This contextualizes the volatility and the need for accurate predictive models in cryptocurrency trading.

In the context of cryptocurrency price research and analysis, various methods have been used to predict price movements. Linear regression is one of the commonly used methods due to its simplicity [5]. Linear Regression is used to build a model that identifies the linear relationship between independent variables (such as opening price, highest price, and trading volume) and the dependent variable (closing price) [6].

Linear Regression is chosen for its simplicity and ability to establish a linear relationship between historical prices and future prices. However, it has limitations in capturing non-linear patterns and temporal dependencies inherent in financial time series data. Alternatively, the Long Short-Term Memory (LSTM) method, which is part of artificial neural networks, has gained significant attention. LSTM is a type of artificial neural network method that has the capability to handle sequential data, such as stock or cryptocurrency price data. This method is designed to model temporal patterns in Bitcoin price data, which can aid in future price predictions[7].

LSTM, a type of recurrent neural network, is selected due to its capability to model sequential data and capture long-term dependencies, making it well-suited for predicting highly volatile and temporally dependent cryptocurrency prices. The combination of these models allows us to compare a traditional statistical approach with a more advanced machine learning method[8].

A previous study conducted by Khalis Sofi, Aswan, Supriyadi Sunge, Sasmitoh Rahmad Riady, and Antika Zahrotul Kamalia in 2021 compared the linear regression, LSTM, and GRU algorithms for predicting stock prices. The results demonstrated that LSTM had an advantage in stock price prediction. The study reported an RMSE value of 0.048, an MSE of 0.002, and an MAE of 0.038 for LSTM, whereas for Linear Regression, the RMSE value was 4.621, the MSE was 2.136, and the MAE was 2.890[9].

This research aims to improve the accuracy of cryptocurrency price predictions, particularly Bitcoin, by utilizing historical data from Yahoo Finance. The methods employed, namely Linear Regression and Long Short-Term Memory (LSTM), are expected to address the limitations of previous predictions by identifying temporal patterns and enhancing prediction accuracy. This, in turn, provides more accurate information regarding cryptocurrency price fluctuations to support more effective investment decision-making and increase investor confidence in this dynamic market.

Based on the problem formulation above, the research questions that arise are as follows:

1. How does the performance of linear regression compare to LSTM in predicting Bitcoin prices?
2. Which method provides higher prediction accuracy and reliability in the context of the high volatility often seen in cryptocurrency prices?

2. Research Methodology

In preparing this research report, several stages were carried out. The research began with problem identification, literature review, data collection, data processing, model implementation using Linear Regression and LSTM, analysis and comparison, and was concluded with results and conclusions. The stages involved in the research process are illustrated in Figure 2 below:

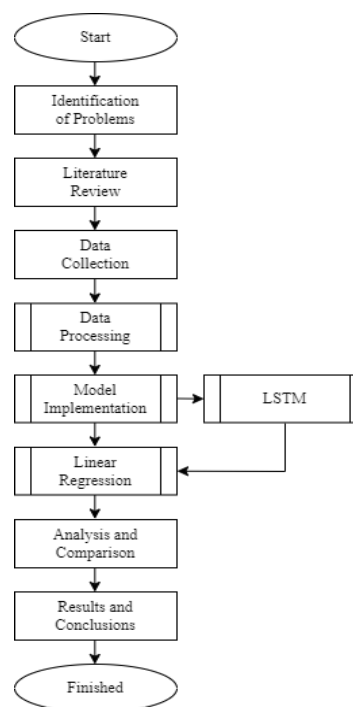


Figure 2. Process Flow

2.1 Problem Identification

The research began by identifying the main problem to be addressed, which is to improve the accuracy of Bitcoin price predictions by comparing the performance of Linear Regression and LSTM.

2.2 Literature Review

This stage involves reviewing relevant literature to understand the context and theories underlying the research. The literature review helps identify gaps in previous research and forms the basis for the approach used in this study.

2.3 Data Collection

Historical Bitcoin price data was obtained from Yahoo Finance, covering the period from September 17, 2014, to July 15, 2024. The collected data includes variables such as date, opening price, highest price, lowest price, closing price, adjusted price, and trading volume.

2.4 Data Processing

The collected data will be processed through the following stages:

2.4.1 Reading and Processing Data

The data is read from a CSV file containing columns Date, Open, High, Low, Close, and Volume. The Date column is converted to a date format and set as the index in the data frame. Subsequently, the data is sorted by date to ensure the correct order and displayed for an initial check to verify the accuracy and consistency of the information.

Journal Program

```
# Load data
file_path = 'btc 7 sept 14-15 juli 24.csv'
data = pd.read_csv(file_path)

# Display the first few rows of the data
print(data.head())
```

2.4.2 Data Selection and Cleaning

The Close column is selected as the target variable to be predicted, while all other columns are used as features. Next, the data is cleaned by removing rows containing missing values (NaN) to ensure the quality and consistency of the dataset.

Journal Program

```
# Select relevant column for prediction ('Close' column)
data = data[['Close']]

# Handling missing values (if any)
data = data.dropna()
```

2.4.3 Data Splitting

The dataset is divided into training and testing data using `train_test_split()` from scikit-learn, with 80% of the data used for training and 20% for testing. This splitting

ensures that the model is trained on the training data and validated on unseen testing data.

Journal Program

```
# Split the data into training and testing sets

X_train_lr, X_test_lr, Y_train_lr, Y_test_lr =
train_test_split(X_lr, Y_lr, test_size=0.2,
random_state=42)
```

2.5 Model Implementation

Linear Regression and Long Short-Term Memory (LSTM) models are applied to the processed data. The models are trained using the training set and tested using the testing set to evaluate their performance.

2.5.1 Linear Regression

The process of predicting Bitcoin prices using linear regression involves several important steps. First, Bitcoin price data is collected and prepared, including normalization to ensure consistent feature scaling. Next, the data is split into two sets: a training set to build the model and a testing set for evaluation. The linear regression model is built using the formula:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where Y is the predicted Bitcoin price, X represents independent variables such as historical prices, β_0 is the intercept, β_1 is the regression coefficient, and ϵ is the prediction error. After the model is trained using the training data, predictions are made on the testing data, and the results are compared with the actual prices to evaluate the model's accuracy.[10].

2.5.2 LSTM

The process of predicting Bitcoin prices using LSTM begins with the collection and normalization of Bitcoin price data to ensure consistent feature scaling. The data is then split into training and testing sets. The LSTM model is built using Keras, which involves components such as the forget gate to determine which information to discard, the input gate to decide which new information to store, and the output gate to determine the next hidden state value. After training the model, predictions are made on the testing data, and the predicted results are compared with the actual prices to evaluate the model's performance.[11].

The main formulas used in LSTM (Long Short-Term Memory) include[12]:

- Forget Gate: Determines which part of the previous cell state to discard.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

Where σ is the sigmoid function, W_f is the weight matrix for the forget gate, h_{t-1} is the previous hidden state, x_t is the input at time, and b_f is the bias term for the forget gate.

- Input Gate: Decides the extent of new information to add to the cell state.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

Where W_i is the weight matrix for the input gate and b_i is the bias term for the input gate.

- Candidate Cell State: Generates candidate updates for the new cell state, with values between -1 and 1.

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

Where \tanh is the hyperbolic tangent function, W_c is the weight matrix for the cell state, and b_c is the bias term for the cell state.

- Cell State Update: Combines old and new information to update the cell state.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Where C_{t-1} is the previous cell state.

- Output Gate: Regulates the information output from the cell state.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

Where W_o is the weight matrix for the output gate and b_o is the bias term for the output gate.

- Hidden State: Generates the output based on the updated cell state and output gate.

$$h_t = o_t * \tanh(C_t)$$

2.6 Analysis and Comparison

The next step is the analysis and comparison, where the results of applying linear regression and LSTM in cryptocurrency prediction are evaluated. At this stage, the analysis presents a detailed assessment of each model's performance in predicting cryptocurrency prices, as well as a comparison between the results obtained from the two methods. The findings from this

analysis will be used as a basis to conclude the advantages and disadvantages of each model in the context of cryptocurrency prediction.

3. Results and Discussion

3.1. Dataset Management

Historical Bitcoin price data was obtained from a CSV file covering the period from September 7, 2014, to July 15, 2024. The collected data includes date and closing price ('Close') variables. The data processing starts with loading the data from the CSV file, followed by converting the 'Date' column to datetime format and setting it as the index of the DataFrame. Next, the relevant columns for analysis, specifically the closing price, are selected, and rows with missing values are removed using `dropna()`. The data is then prepared for the Linear Regression model by creating feature and target variables and splitting the dataset into training and test sets using `train_test_split`. Model evaluation is conducted by calculating Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R² Score, and F1 Score for both the training and test data. Finally, the prediction results are compared with the actual data in a plot for model performance visualization.

3.2 Implementasi Model

3.2.1 Linear Regression

The linear regression model is trained using the training data and tested with the test data. Evaluation is performed by calculating the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R² Score, and F1 Score to measure the model's performance. The evaluation results are presented in Table 1.

3.2.2 LSTM

The LSTM model is built using the Keras framework. This model is trained with the training data and tested with the test data. Evaluation is conducted by calculating MSE, RMSE, R² Score, and F1 Score. The evaluation results are presented in Table 2.

3.3. Model Evaluation

3.3.1 Linear Regression

Table 1. Evaluation Table for the Linear Regression Model

Measurement	Training Data	Test Data
MSE	130,148,318.66	127,306,850.20
RMSE	11,408.26	11,283.03
R ² Score	0.63	0.66
F1 Score	0.88	0.89

In the analysis of the Linear Regression model, the evaluation results indicate a reasonably good performance but with some limitations. On the training data, the model produced a Mean Squared Error (MSE) of 130,148,318.66 and a Root Mean Squared Error (RMSE) of 11,408.26. The R² Score on the training data was 0.63, suggesting that the model explains about 63% of the variability in the training data, leaving 37% unexplained. On the test data, the model showed an MSE of 127,306,850.20 and an RMSE of 11,283.03, with a slightly improved R² Score of 0.66. This indicates that the linear regression model performs well in predicting test data, with slightly lower error compared to the training data. The F1 Score on the training data was 0.88, and on the test data, it improved slightly to 0.89. These scores reflect a balance between precision and recall, suggesting that the model is effective in minimizing false positives and false negatives, although further refinements are needed to enhance its overall predictive accuracy. However, there is still room for improvement in accuracy and precision of predictions for both the training and test data, as the variability explanation remains suboptimal.

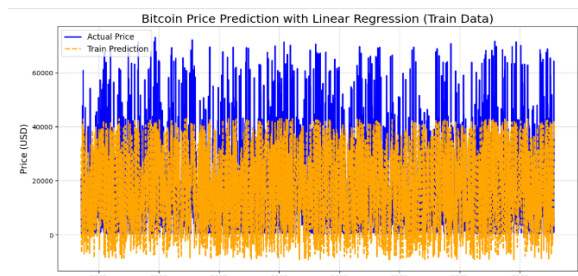


Figure 3. Training Data Prediction Graph for Linear Regression

Figure 3 shows the predicted Bitcoin prices using the Linear Regression model on the training data. In the graph, the x-axis represents the date, and the y-axis represents the Bitcoin price in USD. The blue line displays the actual Bitcoin prices, while the orange dashed line indicates the predicted prices on the training data. This graph provides an overview of how well the Linear Regression model can predict Bitcoin prices based on the trained data.

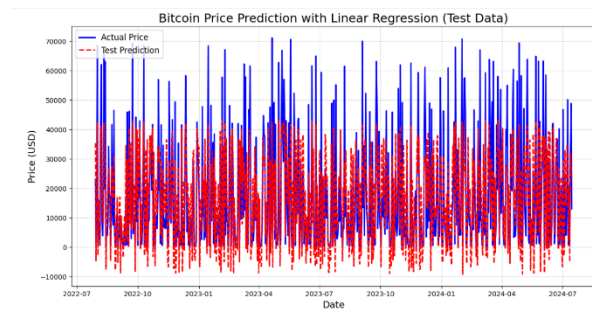


Figure 4. Test Data Prediction Graph for Linear Regression

Figure 4 shows the predicted Bitcoin prices using the Linear Regression model on the test data. As in the previous graph, the x-axis represents the date, and the y-axis represents the Bitcoin price in USD. The blue line depicts the actual Bitcoin prices, while the red dashed line illustrates the predicted prices on the test data. This graph provides a visualization of the performance of the Linear Regression model in predicting Bitcoin prices based on data that was not used in training the model, thereby indicating the model's ability to generalize.

3.3.2 LSTM:

Table 2. Evaluation Table for the LSTM Model

Measurement	Training Data	Test Data
MSE	1,192,873.76	949,865.30
RMSE	1,092.19	974.61
R ² Score	1.00	1.00
F1 Score	0.99	0.99

In the analysis of the Long Short-Term Memory (LSTM) model, the evaluation results show impressive performance on both datasets, training and test data. On the training data, the model achieved a Mean Squared Error (MSE) of 1,192,873.76 and a Root Mean Squared Error (RMSE) of 1,092.19, with a perfect R² Score of 1.00. This indicates that the LSTM model can predict the training data with extremely high accuracy and very low prediction error. On the test data, the model demonstrated an MSE of 949,865.30 and an RMSE of 974.61, with an R² Score remaining at 1.00. This success suggests that the LSTM model is not only highly effective in learning from the training data but also in generalizing to unseen data, with optimal prediction accuracy and very minimal error. The F1 Score for both the training and test data was an impressive 0.99, indicating an exceptional balance between precision and recall. This high F1 Score reinforces the model's capability to minimize false positives and false negatives, affirming its robustness in capturing the underlying patterns and trends in the data exceptionally well, resulting in highly accurate predictions on both datasets.

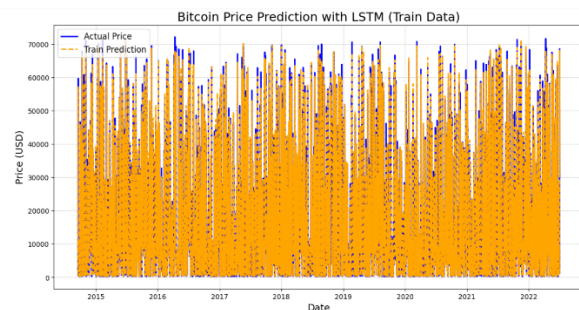


Figure 5. Training Data Prediction Graph for LSTM

Figure 5 shows the predicted Bitcoin prices using the LSTM model on the training data. In the graph, the x-axis represents the date, and the y-axis represents the Bitcoin price in USD. The blue line displays the actual Bitcoin prices, while the orange dashed line indicates the predicted prices on the training data. This graph provides an overview of how well the LSTM model can predict Bitcoin prices based on the trained data.

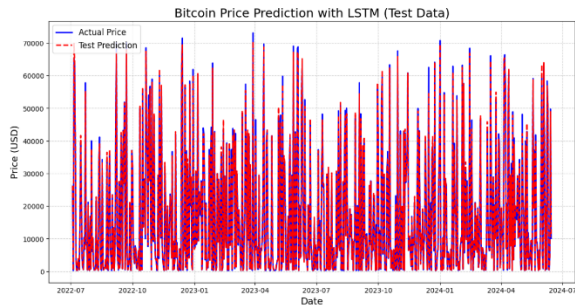


Figure 6. Test Data Prediction Graph for LSTM

Figure 6 shows the predicted Bitcoin prices using the LSTM model on the test data. In the graph, the x-axis represents the date, and the y-axis represents the Bitcoin price in USD. The blue line displays the actual Bitcoin prices in the test data, while the red dashed line represents the predictions made by the LSTM model for the test data. This graph provides an overview of how well the LSTM model performs in predicting Bitcoin prices based on data that the model has not previously seen.

3.4 Model Performance Comparison

Table 3. Training Data Comparison

Measurement	Training Data Linear Regression	Training Data LSTM
MSE	130,148,318.66	1,192,873.76
RMSE	11,408.26	1,092.19
R ² Score	0.63	1.00
F1 Score	0.88	0.99

Table 4. Test Data Comparison

Measurement	Test Data Linear Regression	Test Data LSTM
MSE	130,148,318.66	949,865.30
RMSE	11,408.26	974.61
R ² Score	0.66	1.00
F1 Score	0.99	0.99

The comparison between the linear regression and LSTM models shows that LSTM significantly outperforms linear regression in terms of prediction

accuracy. As illustrated in Table 3, the LSTM model demonstrates considerably lower Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) compared to linear regression, with values of 1,192,873.76 and 1,092.19 for training data, while linear regression reports MSE of 130,148,318.66 and RMSE of 11,408.26. Additionally, LSTM achieves a perfect R² Score of 1.00, in contrast to the 0.63 achieved by the linear regression model. In terms of F1 Score, which balances precision and recall, the LSTM model excels with a score of 0.99 on the training data, compared to 0.88 for linear regression.

Table 4 further reinforces these findings, showing that on the test data, the LSTM model continues to perform exceptionally well with an MSE of 949,865.30 and RMSE of 974.61, while linear regression records higher errors with MSE of 130,148,318.66 and RMSE of 11,408.26. Both models achieve a perfect R² Score of 1.00 for LSTM and 0.66 for linear regression, but the F1 Scores highlight that the LSTM maintains its superior predictive capability with a score of 0.99, while linear regression achieves an F1 Score of 0.99. This indicates that while linear regression provides reasonable results, it struggles to capture complex price dynamics, particularly in the context of Bitcoin's high volatility.

3.5 Limitations and Recommendations

The linear regression model may not adequately capture the nonlinear patterns in Bitcoin price data, which can affect prediction accuracy. Additionally, the size of the dataset and feature selection can also impact model performance. While the LSTM model has demonstrated strong performance, its effectiveness can also be influenced by the chosen hyperparameters, and the amount of data used. We recommend further tuning of LSTM parameters and considering additional features to enhance model performance in the future.

4. Conclusion

The evaluation of the Linear Regression and Long Short-Term Memory (LSTM) models reveals a clear difference in prediction accuracy. The Linear Regression model yields higher Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values, with 130,148,318.66 and 11,408.26 on the training data, and 127,306,850.20 and 11,283.03 on the test data. The R² Scores are 0.63 for the training data and 0.66 for the test data, indicating that while the model explains a substantial portion of the data variability, significant errors remain. Additionally, the F1 Score for the Linear Regression model is 0.88 on the training data and 0.89 on the test data, reflecting its ability to balance precision and recall but also highlighting its limitations in capturing complex patterns.

In contrast, the LSTM model reports significantly lower MSE and RMSE values, with 1,192,873.76 and 1,092.19 on the training data, and 949,865.30 and 974.61 on the test data. The perfect R² Score of 1.00 on both datasets shows that the LSTM model can explain the entire variability of the data with very high accuracy and minimal error. The F1 Score for the LSTM model stands at an impressive 0.99 on both training and test datasets, underscoring its exceptional performance in maintaining a balance between precision and recall.

Overall, the LSTM model demonstrates superior performance in terms of prediction accuracy compared to Linear Regression. While Linear Regression still provides reasonable results, LSTM offers better accuracy and may be more suited to capturing complex temporal patterns in the data, making it a more effective choice for predicting Bitcoin prices.

Acknowledgments

The author wishes to express their deepest gratitude for the assistance and support provided throughout the research process. Special thanks to:

1. Mr. Ahmad Soleh, the author's father, for his unwavering financial and moral support.
2. Mrs. Faiqotun, the author's mother, for her constant encouragement, affection, and invaluable support.
3. The author's family, for providing continuous motivation and the necessary resources to complete this research.
4. The author's family, for providing continuous motivation and the necessary resources to complete this research.
5. Teachers and Lecturers, for the knowledge and motivation provided.
6. The author's friends, for their help, advice, and encouragement during the research process.

Without their support and assistance, this research would not have been completed successfully.

References

- [1] B. Zohuri, H. T. Nguyen, and M. Moghaddam, "International Journal of Theoretical & Computational Physics What is the Cryptocurrency? Is it a Threat to Our National Security, Domestically and Globally?," vol. 3, no. 1, pp. 1–14.
- [2] V. Marella, B. Uperti, and J. Merikivi, "Understanding the creation of trust in cryptocurrencies: Bitcoin," *32nd Bled eConference Humaniz. Technol. a Sustain. Soc. BLED 2019* -

Conf. Proc., pp. 839–859, 2020.

- [3] S. Saadah and H. Salsabila, "Prediksi Harga Bitcoin Menggunakan Metode Random Forest," *J. Komput. Terap.*, vol. 7, no. 1, pp. 24–32, 2021.
- [4] C. Hung, J. F. Wijaya, V. Victor, I. A. Pardosi, and F. M. Sinaga, "Prediksi Fluktuasi Harga Bitcoin Dengan Menggunakan Random Forest Classifier," *J. SIFO Mikroskil*, vol. 24, no. 2, pp. 95–108, 2023..
- [5] J. S. Putra, R. D. Ramadhani, and A. Burhanuddin, "Prediksi Harga Saham Bank Bri Menggunakan Algoritma Linear Regresion Sebagai Strategi Jual Beli Saham," *Journal of Dinda: Data Science, Information Technology, and Data Analytics*, vol. 2, no. 1. pp. 1–10, 2022.
- [6] M. F. Arfa, M. R. AlFathan, H. B. Lumbantobing, and R. Rahmadenni, "Prediksi Harga Cryptocurrency Dengan Metode Linier Regresi," *SENTIMAS Semin. Nas. Penelit. dan Pengabd. Masy.*, vol. 1, no. 1, pp. 8–15, 2023, [Online]. Available: <https://journal.irpi.or.id/index.php/sentimas/article/view/609%0Ahttps://journal.irpi.or.id/index.php/sentimas/article/download/609/332>
- [7] H. Utama, "Pendekatan Deep Learning Menggunakan Metode Lstm Untuk Prediksi Harga Bitcoin," *Indones. J. Comput. Sci. Res.*, vol. 2, no. 2, pp. 43–50, 2023.
- [8] N. Latif, J. D. Selvam, M. Kapse, V. Sharma, and V. Mahajan, "Comparative Performance of LSTM and ARIMA for the Short-Term Prediction of Bitcoin Prices," *Australas. Accounting, Bus. Financ. J.*, vol. 17, no. 1, pp. 256–276, 2023.
- [9] Khalis Sofi, Aswan Supriyadi Sunge, Sasmitoh Rahmad Riady, and Antika Zahrotul Kamalia, "Perbandingan Algoritma Linear Regression, Lstm, Dan Gru Dalam Memprediksi Harga Saham Dengan Model Time Series," *Seminastika*, vol. 3, no. 1, pp. 39–46, 2021.
- [10] I. Zulfikar, M. A. Saputra, N. Prasetyo, T. Rijanandi, and F. D. Adhinata, "Prediksi Gaji Berdasarkan Pengalaman Bekerja Menggunakan Metode Regresi Linear," *J. Dinda Data Sci. Inf. Technol. Data Anal.*, vol. 2, no. 2, pp. 58–63, Jul. 2022.
- [11] M. D. Hilmawan, "Deteksi Sarkasme Pada Judul Berita Berbahasa Inggris Menggunakan Algoritme

Bidirectional LSTM,” *J. Dinda Data Sci. Inf. Technol. Data Anal.*, vol. 2, no. 1, pp. 46–51, Feb. 2022.

M. P. K. Reddy, and T. R. Gadekallu, “A Multidirectional LSTM Model for Predicting the Stability of a Smart Grid,” *IEEE Access*, vol. 8, pp. 85454–85463, 2020.

[12] M. Alazab, S. Khan, S. S. R. Krishnan, Q. V. Pham,