Volume XI Issue 2 Year 2026 | Page 520-526 | ISSN: 2527-9866
Received: ***19-04-2026*** | *Revised:* ***02-05-2026*** | *Accepted:* ***26-05-2026***

A Comparative Analysis of Machine Learning Models in Bitcoin Price Prediction Based on Training Data Variations

Farrel Amri Naufal Sandio[1], Renny Sari Dewi[2]
[1,2] State University of Surabaya, Jl. Ketintang, Ketintang, Gayungan District, Surabaya City, East Java 60231
e-mail: farrel.23163@mhs.unesa.ac.id;[1] , rennydewi@unesa.ac.id[2]

Abstract**:** *The rapid growth of cryptocurrency, particularly Bitcoin, has introduced high-return investment opportunities accompanied by extreme price volatility, posing challenges for accurate forecasting. Previous studies have applied various machine learning models for Bitcoin price prediction; however, limited attention has been given to how different training data horizons affect model performance and generalisation. This study addresses this gap by comparing three machine learning algorithms: Linear Regression (LR), XGBoost, and Long Short-Term Memory (LSTM). The analysis examines different training periods, with a primary focus on a 3-year training scenario. Historical Bitcoin data (1-minute intervals) from Kaggle was aggregated into daily observations and processed using strict chronological splitting (80:20) without data leakage. Feature engineering was applied using lag-based variables, moving averages, and volatility indicators, whilst LSTM utilised sequence windowing with 30–60 time steps. Empirical results from the 3-year training scenario show that LR and XGBoost achieve strong predictive performance (R^2 = 0.9757 and 0.9667), whilst LSTM performs moderately (R^2 = 0.72) with higher prediction errors. Additional exploratory experiments on shorter training horizons (e.g., 6 months) indicate a decline in performance across models, reflected in unstable generalisation and negative R^2 values on test data, suggesting overfitting. However, directional accuracy remains above 55% in the primary scenario. These findings suggest that model performance is sensitive to the length and stability of historical data. Whilst simpler models such as linear regression and tree-based methods demonstrate consistent performance in the evaluated setting, conclusions regarding model superiority should be interpreted within the scope of the experiment.*

Keywords: *cryptocurrency, Bitcoin, price prediction, machine learning, Linear Regression, XGBoost, LSTM, time series.*

1. Introduction

Advances in digital technology have driven significant transformation in the financial sector, one example being the emergence of cryptocurrency as a decentralised digital asset. Bitcoin, as the leading cryptocurrency, has demonstrated rapid growth in terms of both market capitalisation and user adoption; however, it is characterised by extremely high price volatility[1], [2]. This makes predicting cryptocurrency prices a complex challenge, as price movements are influenced by non-linear factors such as market sentiment, demand dynamics, and global macroeconomic conditions.

Various approaches have been used to model cryptocurrency price movements, ranging from conventional statistical methods to machine learning. Models such as linear regression, decision tree-based methods, and deep learning techniques like Long Short-Term Memory (LSTM) have demonstrated an ability to capture patterns in time-series data[3], [4]. However, previous research findings remain inconsistent, particularly regarding the relative performance of different models. Some studies report the superiority of LSTM in capturing long-term

dependencies, whilst others suggest that simpler models can deliver competitive performance, particularly under conditions of limited data or unstable distributions.

Nevertheless, there is a significant limitation in previous research, namely the lack of comparative analyses that systematically evaluate model performance across different training horizons using consistent evaluation protocols. Most studies have used only a single data period without exploring how variations in the length of historical data affect the model's generalisation ability. Furthermore, methodological issues such as data leakage, lag-based feature selection, and the use of chronological data splitting are often not strictly accounted for, which may potentially affect the validity of the results.

Given this gap, this study aims to evaluate the performance of three main approaches Linear Regression (LR), XGBoost, and LSTM in predicting Bitcoin prices, taking into account variations in the length of the training data. Specifically, this study focuses on a 3-year training scenario as the main experiment, with additional exploration of shorter time horizons. Historical Bitcoin data was obtained from the Kaggle platform in high resolution and processed into daily data, then analysed using a lag-based feature engineering approach and chronological data splitting without data leakage.

The main contribution of this study lies in the development of a structured comparative experimental framework to evaluate the performance of three key machine learning approaches namely, linear models (Linear Regression), ensemble tree-based models (XGBoost), and deep learning models LSTM in the context of Bitcoin price prediction. Unlike most previous studies, which tend to evaluate models within a single specific data configuration, this study systematically compares model performance within a consistent experimental framework, thereby enabling a fairer and more controlled analysis of the relative strengths of each approach. Furthermore, this study also contributes to examining the influence of training data length (training horizon) on prediction accuracy and model generalisation ability, a topic that has been relatively rarely discussed explicitly in the previous literature.

This study also employs a rigorous evaluation protocol to ensure the validity of the results, including the use of lag-based features to avoid data leakage, as well as a chronological split that aligns with the characteristics of time series data. This approach aims to replicate real-world conditions in price prediction, where models have access only to historical information when making predictions. Consequently, the research findings are expected not only to make an academic contribution but also to have practical relevance for the development of more reliable cryptocurrency price prediction systems.

Based on these objectives, the research question in this study is: how do Linear Regression, XGBoost and LSTM perform in predicting Bitcoin prices under specific training scenarios using historical features, to what extent variations in training data length affect the accuracy, stability, and generalisation ability of each model, and which model demonstrates the most consistent and robust performance in dealing with the volatile, non-linear, and not entirely stationary characteristics of Bitcoin data. Furthermore, this study also implicitly explores whether model complexity is always directly proportional to improved performance, or whether simpler models can actually deliver more optimal results under certain conditions.

2. Literature Review

Cryptocurrency is an innovation in the digital financial system that utilises blockchain technology to enable decentralised, transparent and secure transactions without intermediaries[5]. Among the various cryptocurrencies, Bitcoin has been the primary focus of much research due to its market dominance and the availability of extensive historical data. However, Bitcoin's characteristics including high volatility, non-normal return distributions, and sensitivity to market sentiment make the process of price prediction a complex challenge[6].

In the context of price modelling, traditional statistical approaches such as linear regression have limitations in capturing non-linear relationships and complex dynamics in financial time-series data. Nevertheless, linear models are still frequently used as a baseline due to their advantages in terms of interpretability, stability, and relatively lower data requirements. Under certain conditions, particularly when data is limited or non-stationary, linear models can deliver competitive performance compared to more complex models.

Advances in machine learning have led to the use of more flexible models for predicting cryptocurrency prices. One widely used approach is ensemble-based methods such as XGBoost, which are capable of capturing non-linear relationships and interactions between features through boosting and regularisation mechanisms. These models are also known to be relatively robust against outliers and changes in data distribution, and are therefore frequently applied to various tabular data-based prediction problems, including in the field of finance[3], [7].

Long Short-Term Memory, as part of the recurrent neural network architecture, is designed to model long-term dependencies in sequential data. LSTM has the ability to retain historical information through gating mechanisms, making it theoretically superior to static models in capturing complex temporal patterns. Consequently, LSTM is widely used in time series forecasting, including cryptocurrency price forecasting.

Nevertheless, the results of empirical research on the performance of LSTMs in Bitcoin price prediction remain inconsistent. Some studies report that LSTMs are capable of improving prediction accuracy by utilising long-term temporal dependencies[8]. However, other studies indicate that LSTM performance is highly dependent on the quantity and quality of training data, and is vulnerable to changes in data distribution. Under conditions of limited or unstable data, simpler models such as linear regression or tree-based models may actually yield more stable and competitive results.

Furthermore, most previous studies have tended to evaluate models on a single training data configuration without systematically exploring the influence of historical data length on model performance. Yet, in the context of time series, the length of the training data plays a crucial role in determining the model's ability to capture long-term patterns and generalise to new data. The lack of research in this area has resulted in a limited understanding of the optimal conditions for using each model.

Based on this review, this study focuses on a comparative analysis of Linear Regression, XGBoost and LSTM in predicting Bitcoin prices, taking into account variations in the length of the training data. Furthermore, this study also applies a rigorous evaluation protocol through the use of lag-based features and chronological data splitting without data- -leakage. This approach is expected to provide a more comprehensive understanding of the relationship between data characteristics and model performance.

3. Methods

This study employs a quantitative approach with a comparative experimental design to evaluate the performance of machine learning models in predicting Bitcoin prices across various training data horizons. In general, this study considers three training data length scenarios, namely 12 years, 3 years, and 6 months, with the main focus of the analysis on the 3-year scenario, whilst the other scenarios are used as supplementary evaluations to observe the model's sensitivity to the availability of historical data. The data used consists of historical Bitcoin price data obtained from Kaggle ("Bitcoin Historical Data Bitstamp Exchange") with a 1-minute resolution, covering the period from 1 January 2012 to 12 April 2026, comprising a total of 7,507,845 rows consisting of the attributes Timestamp, Open, High, Low, Close, and Volume.

The pre-processing stage begins with converting timestamps to the datetime format and setting them as the time index. Next, data quality is checked by identifying and removing duplicate data based on timestamp similarity, whilst missing values are not found in the dataset. The data is then converted from minute-by-minute resolution to daily data through a resampling process using OHLC aggregation, where the Open value is taken as the first value, High as the maximum value, Low as the minimum value, Close as the last value, and Volume as the total daily transactions. No explicit outlier handling was performed as extreme fluctuations are considered an inherent characteristic of the cryptocurrency market. For modelling purposes, particularly for neural network-based models, normalisation was carried out using the Min-Max scaling method within the range [0,1], whilst the Linear Regression and XGBoost models used the data in its original scale. The result of this stage yielded 5,216 daily data observations.

During the feature engineering stage, all features were constructed with the primary principle of avoiding data leakage, namely by using only information from previous time periods (t-1 or earlier). In total, there are 23 features reflecting various aspects of price movements, including historical values (lag features), volatility indicators, momentum returns, moving average-based trends, price-to-moving-average ratios, volume indicators, and calendar features such as the day of the week and the month. The selection of these features is based on a common approach in financial time series analysis that emphasises the importance of historical information, trend patterns, and volatility in predicting price movements.

Data splitting is performed chronologically without randomisation to preserve the time series structure and avoid data leakage. In each scenario, the training data is determined based on the horizon length used namely 12 years, 3 years, or 6 months whilst the test data uses the six-month period following the training data. Consequently, the data split does not follow a fixed ratio such as 80:20, but is time-based. The 3-year training scenario is used as the main experiment, with approximately 1,000 training observations and around 180 test observations.

This study utilised three main algorithms: Linear Regression (LR) as a linear baseline model, XGBoost as an ensemble tree-based model capable of capturing non-linear relationships, and Long Short-Term Memory (LSTM) as a deep learning model for sequential data. The XGBoost model was implemented with standard parameters such as 100 estimators, a maximum depth of 6, a learning rate of 0.1, and a subsample of 0.8. , the LSTM model was built with a single LSTM layer containing 50 units, using the tanh activation function, the Adam optimiser, and the Mean Squared Error loss function, and was trained for 50 epochs with a batch size of 32. To prevent overfitting, a dropout rate of 0.2 and an early stopping mechanism were applied, with an input sequence length of 30 timesteps.

All models were trained without extensive hyperparameter tuning; consequently, the research focused on comparing model performance and the impact of training data length on prediction results. Model evaluation was conducted using several metrics, namely Mean Absolute Error

(MAE), Root Mean Squared Error (RMSE), the coefficient of determination (R^2), and Directional Accuracy (DA) to measure the accuracy of price movement direction. A negative R^2 value on the test data is interpreted as an indication that the model is unable to generalise well compared to a simple baseline, which in this context relates to overfitting or model instability due to limitations in the training data.

4. Results and Discussion

The characteristics of Bitcoin data, based on preliminary analysis, indicate very significant and exponential growth, rising from approximately $4.38 at the start of 2012 to reach $124,728 in April 2026. Bitcoin's daily returns have an average of 0.2688% with a standard deviation of 4.05%, as well as a rather extreme range, from -53.84% to +35.81%. The return distribution also shows negative skewness and high kurtosis, indicating the presence of *fat tails* or extreme risk. These characteristics are consistent with the literature stating that the cryptocurrency market has high volatility and a non-normal distribution, making the prediction process difficult[9]. Furthermore, the non-linear and non-stationary nature of Bitcoin data poses a major challenge in machine learning-based time series modelling[10].

Table 3. Model Results – 3-Year Training

Model	MAE (USD)	RMSE (USD)	R	DA (%)
Linear Regression	1,595.84	2,210.21	0.9757	55.87
XGBoost	1,918.82	2,589.89	0.9667	57.54
LSTM	6,192.53	7,510.16	0.7200	43.02

In the 3-year training scenario as the main experiment, the results showed that Linear Regression (LR) and XGBoost were able to achieve high performance with R^2 values of 0.9757 and 0.9667, respectively. LR produced lower prediction errors, whilst XGBoost demonstrated superiority in Directional Accuracy. Meanwhile, LSTM demonstrated lower performance with an R^2 of 0.72 and a higher error rate. These findings are consistent with previous research indicating that models such as XGBoost are effective in handling non-linear relationships in financial data.

However, the superiority of linear regression in this study demonstrates that under certain conditions particularly when lag-based features and short-term trends are used the relationship between the input variables and the target variable can be fairly stable and close to linear. This enables simple models to deliver competitive results with a lower risk of overfitting. Similar findings have also been reported in comparative studies showing that simple models can outperform complex models under conditions of limited data or unstable distributions[11].

On the other hand, the lower performance of LSTMs can be attributed to the model's requirement for a large and stable dataset to capture temporal dependencies optimally. LSTMs are highly sensitive to data quality and feature engineering; consequently, under conditions of limited data, this model tends to experience a decline in performance[14]. Furthermore, several studies have also shown that the performance of LSTMs can improve significantly when combined with other models in a hybrid approach[13].

Feature contribution analysis indicates that historical price variables, particularly the *high* and *low* values from the previous period, have a dominant influence on price prediction. This suggests that information regarding extreme price levels contains important signals regarding

future price movements. These findings are consistent with studies emphasising the importance of technical and historical features in cryptocurrency price prediction[10].

Furthermore, the residual analysis shows that Linear Regression has a relatively stable error distribution, whilst XGBoost exhibits greater error variation, and LSTM has the highest deviation. This suggests that model complexity does not always correlate directly with improved performance, particularly in the context of financial data, which has a high level of noise.

Additional experiments on shorter data horizons revealed a significant decline in performance across all models. This reinforces the view that the length of the training data is a crucial factor in determining a model's generalisation ability, as also found in previous research on time series forecasting for cryptocurrencies[9]. However, this study has not yet conducted formal statistical significance tests or robustness checks; therefore, the results obtained should be understood as empirical findings within the context of the experiments conducted.

5. Conclusions

This study evaluates the performance of three machine learning algorithms Linear Regression, XGBoost and Long Short-Term Memory (LSTM) in predicting Bitcoin prices, taking into account variations in the length of the training data. Overall, the results indicate that the length of the training data horizon is a key factor influencing the models' ability to generalise. The models tested demonstrated varying levels of performance depending on the availability and characteristics of the historical data used.

Over a medium-term data horizon (3 years), Linear Regression and XGBoost delivered high and stable performance, whilst LSTM showed more moderate performance. However, over a shorter data horizon, all models experienced a significant decline in performance, indicating limitations in capturing patterns when data volume is insufficient. These findings confirm that model complexity does not always correlate directly with improved performance, particularly under conditions of limited or unstable data.

The main contribution of this study lies in providing a consistent comparative evaluation framework for comparing linear, tree-based and deep learning models in the context of Bitcoin price prediction, as well as in demonstrating how variations in training data length affect the performance of each model. Furthermore, this study highlights the importance of using rigorous evaluation protocols, such as chronological data splitting and the use of features based on past information, to avoid bias in time series modelling.

Although the results of the study indicate that simple models such as Linear Regression can deliver competitive performance under certain conditions, these findings must be interpreted within the context of the experiments conducted and cannot be broadly generalised without further testing. This study also has limitations, including the lack of extensive hyperparameter tuning and the absence of statistical significance tests to formally compare the performance of different models.

Therefore, further research is recommended to explore more comprehensive approaches, such as the use of rolling window-based validation techniques, testing under various market conditions, and the integration of trading strategy-based evaluations to assess the practical implications of the prediction results. Furthermore, exploring advanced feature engineering techniques and hybrid models also has the potential to improve predictive performance on complex cryptocurrency time series data.

References

[1] N. Rahma, A. N. Nebras, F. W. S. Suhaeb, and I. I. I. Idrus, "Cryptocurrency dan Masa Depan Keuangan Global: Tantangan dan Peluang," *J-CEKI: Jurnal Cendekia Ilmiah*, vol. 4, no. 4, pp. 772–782, 2025.

[2] Julianto, "Analisis Teknologi Blockchain pada Pengembangan Mata Uang Digital (Cryptocurrenncy)," vol. 02, no. 01, 2025.

[3] A. Fauzi, N. Maulidah, R. Supriyadi, H. Nalatissifa, and S. Diantika, "Prediksi Harga Properti Di Indonesia Menggunakan Algoritma Random Forest," *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 4, no. 1, pp. 43–49, 2025.

[4] A. A. Prayoga, M. Hasanuddin, S. Khodijah, and C. A. Rizki, "Analisis Penerapan Machine Learning dalam Sistem Prediksi dan Pengambilan Keputusan," *Journal of Electrical Engineering Research*, vol. 1, no. 3, pp. 84–90, 2025.

[5] M. Tanley, A. Benneres Roberto, F. Nata, and N. Livanio, "Analisis Potensi Dan Tantangan Teknologi Blockchain Dalam Mendukung Digitalisasi Ekonomi Di Indonesia," *Indonesian Journal of Education And Computer Science*, vol. 2, no. 3, 2024.

[6] T. W. E. Suryawijaya, "Memperkuat keamanan data melalui teknologi blockchain: Mengeksplorasi implementasi sukses dalam transformasi digital di Indonesia," *Jurnal Studi Kebijakan Publik*, vol. 2, no. 1, pp. 55–68, 2023.

[7] M. Alfarizi and D. Lestarini, "Predicting Cryptocurrency Prices Using Machine Learning: A Case Study on Bitcoin," *Journal of Applied Informatics and Computing*, vol. 9, no. 6, pp. 3612–3621, 2025.

[8] M. S. R. A. S. Dapubeang and E. U. Malahina, "Prediksi harga mata uang kripto XRP menggunakan metode deep learning LSTM dan GRU," *Jurnal Manajemen Informatika & Teknologi*, vol. 5, no. 2, pp. 107–124, 2025.

[9] E. Ngai, S. Abdullah, M. Z. A. Nazri, N. S. Sani, and Z. Othman, "Time series prediction of Bitcoin cryptocurrency price based on machine learning approach," *Data Science: Journal of Computing and Applied Informatics*, vol. 7, no. 2, pp. 81–95, 2023.

[10] M.-C. Lee, "Bitcoin trend prediction with attention-based deep learning models and technical indicators," *Systems*, vol. 12, no. 11, p. 498, 2024.

[11] N. Urooj, L. Asif, and Z. Jabin, "Bitcoin price forecasting: a comparative study of machine learning, statistical and deep learning models," *Int J Innov Sci Technol*, vol. 6, pp. 396–412, 2024.

[12] P. S. T. P. Purnama, "Optimizing Bitcoin Price Prediction with LSTM: A Comprehensive Study on Feature Engineering and the April 2024 Halving Impact," *Elinvo (Electronics, Informatics, and Vocational Education)*, vol. 9, no. 1, pp. 165–177, 2024.

[13] M. Gautam, "crypto price prediction using lstm+ xgboost," *arXiv preprint arXiv:2506.22055*, 2025.

[14] Tedyyana, A., and O. Ghazali. "W. Purbo, O.(2024). Enhancing intrusion detection system using rectified linear unit function in pigeon inspired optimization algorithm." *IAES International Journal of Artificial Intelligence (IJ-AI)* 13.2: 1526-1534.

Acknowledgements

The author would like to thank the data providers via the Kaggle platform for supplying the historical Bitcoin dataset. Thanks are also extended to the institutions and colleagues who provided input, support, and facilities throughout the research process and the preparation of this article.