CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 10 (2) (2025), Pages 722-737
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Comparative Study of Hybrid ARIMA-LSTM and ARIMAX-LSTM for Bitcoin Forecasting with Data Partitioning

Fikrie Hartanta Sembiring, Regita Putri Permata*, and Rifdatun Ni'mah

Department of Data Science, School of Computing, Telkom University, Surabaya, Indonesia

Abstract

The extreme volatility of Bitcoin prices poses substantial challenges for accurate forecasting. While ARIMA models are effective for capturing linear dependencies, they often fail to account for non-linear structures; conversely, LSTM networks model non-linearity well but are prone to overfitting in noisy financial series. This study evaluates six model configurations standalone ARIMAX, standalone LSTM, and four hybrid ARIMA/ARIMAX-LSTM models applied to daily Bitcoin closing prices from 2015 to 2024. These models are tested under two partitioning strategies: single-split and two-stage split. Out-of-sample results show that the hybrid ARIMA-LSTM using a two-stage split achieves the lowest forecasting error, with a MAPE of 2.60%, outperforming all other variants. The findings underscore the importance of residual structure and temporal partitioning in hybrid model performance, offering practical insights for designing more robust time series forecasting pipelines in volatile financial contexts.

Keywords: Bitcoin forecasting; ARIMA; LSTM; Hybrid model; Time series

1 Introduction

The volatile nature of cryptocurrency markets, particularly Bitcoin the most dominant and liquid digital asset has attracted growing academic and practical interest due to its extreme price fluctuations and speculative behavior [1], [2]. Accurately forecasting such assets poses a major challenge, as price dynamics are shaped by nonlinear interactions involving investor sentiment, regulation, macroeconomic signals, and technology shifts [3].

Traditional models like ARIMA effectively capture linear temporal dependencies but struggle with regime shifts and nonlinearities typical in crypto markets [4], [5]. In contrast, LSTM networks have emerged as powerful tools for capturing complex temporal dependencies and nonlinear dynamics [6], [7]. Prior studies have shown LSTM outperforming ARIMA in Bitcoin forecasting, with hybrid models improving accuracy by combining both approaches [8], [9].

However, most works overlook the influence of residual structure especially under different data-splitting strategies on hybrid model performance. Residuals, defined as the deviation between actual and linear model outputs, form the core learning signals for LSTM. If residuals retain meaningful structure, LSTM can enhance forecasts. If not, hybrid models underperform.

This study fills that gap by examining how split strategies affect residual learnability and hybrid effectiveness. Table 1 highlights recent related work and their performance.

*Corresponding author. E-mail: regitapermata@telkomuniversity.ac.id

Submitted: June 10, 2025 Reviewed: July 8, 2025 Accepted: July 28, 2025
DOI: https://doi.org/10.18860/cauchy.v10i2.35118

Table 1: Comparison of Previous Studies on Time Series Forecasting Models

Study	Model(s)	Split Strategy	Best MAPE (%)
Adu et al. (2023)	ARIMA vs ARIMAX	Single	7.07 / 6.94
Latif et al. (2023)	ARIMA vs LSTM	Single	19.6 / 7.8
Dave et al. (2021)	ARIMA, LSTM, Hybrid	Single	15.3 / 8.2 / 6.1
This Study	ARIMAX, LSTM, Hybrid	Single & Two-Stage	TBD

Figure 1 illustrates the six model configurations: standalone LSTM and ARIMAX, and four hybrids (ARIMA/ARIMAX with LSTM under single and two-stage splits). We evaluate these models on Bitcoin daily price data from 2015–2024 to understand how residual structure and temporal partitioning influence forecasting performance.

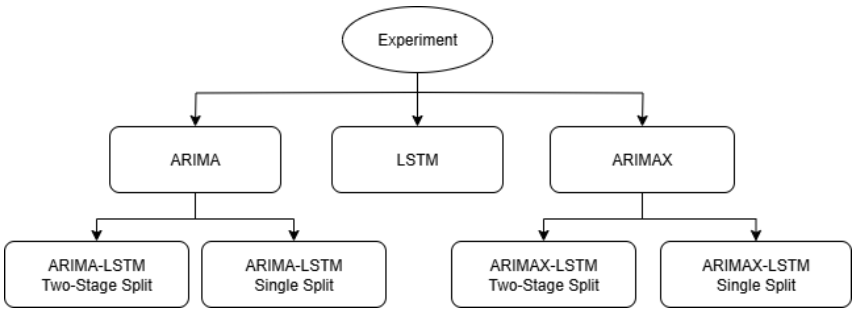

Figure 1: Model Configurations: Standalone and Hybrid Architectures with Different Splitting Strategies

The remainder of this paper is organized as follows. Section 2 describes the data preparation, modeling procedures, and hybrid configurations. Section 3 presents the experimental results and discusses their implications. Finally, Section 4 concludes the study with key findings and future directions.

2 Methods

This section details the end-to-end modeling process used in the study. It begins by describing the dataset and preprocessing steps, followed by a comprehensive explanation of model architectures, data partitioning strategies, and evaluation metrics applied to compare the forecasting performance of the proposed configurations.

2.1 Data Collection and Preparation

This study utilizes daily historical Bitcoin (BTC) closing price data from Yahoo Finance, spanning from January 1, 2015, to December 31, 2024. Yahoo Finance is a reputable platform widely used for financial time series retrieval. The dataset includes standard financial indicators such as open, high, low, close, and trading volume, with the closing price selected as the primary target variable for prediction.

To ensure modeling readiness, linear interpolation was applied to address minor missing values, followed by chronological alignment and Min-Max normalization in the range $[-1, 1]$, a common practice to improve convergence in deep learning models.

A crucial preprocessing step involved feature engineering based on autocorrelation diagnostics. Figure 2 shows the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots for the differenced closing price series. The ACF plot suggests short memory behavior typical of financial returns, while the PACF plot reveals clear spikes at several specific lags. Based on the PACF structure, five significant lags lag-1, lag-9, lag-12, lag-19, and lag-25 were identified as relevant features. These lags were employed as exogenous inputs in ARIMAX models and also

used as input variables for the LSTM and hybrid configurations, ensuring consistency across all modeling pipelines [10],[9].

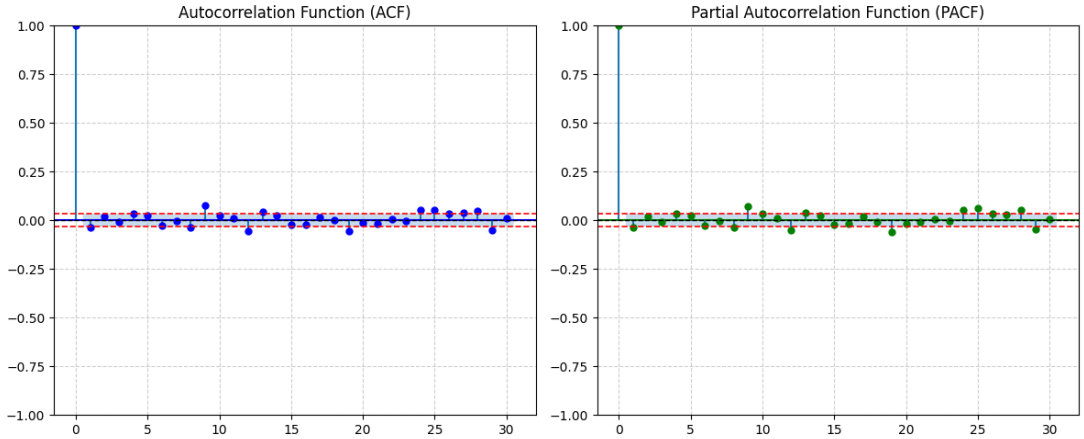

Figure 2: Autocorrelation (ACF) and Partial Autocorrelation (PACF) plots of differenced Bitcoin closing prices. The PACF reveals significant lags at positions 1, 9, 12, 19, and 25, which were selected for model input.

To simulate realistic forecasting, a time-based holdout strategy was employed. The training set covered January 2015 to December 2023, while the test set comprised data from January to December 2024. This chronological split ensures that all models are evaluated on future-unseen data, consistent with out-of-sample testing protocols.

2.2 Experimental Setup

Two experimental strategies were implemented: one for single-model configurations (ARIMAX, LSTM) and another for hybrid models (ARIMA-LSTM, ARIMAX-LSTM), each with single-split and two-stage variants. This structured setup was designed to isolate the effect of data partitioning on forecasting performance, particularly in how residuals are generated and learned by the nonlinear component.

Single-split refers to a strategy in which both the linear (ARIMA/ARIMAX) and nonlinear (LSTM) models are trained and tested on the same temporal window. This method ensures simplicity and synchronization between model components. However, it also introduces potential overlap in learned patterns particularly if the linear model overfits to the training data, leaving little informative structure for the LSTM to learn from the residuals.

Two-stage split, by contrast, temporally separates the training phases of the linear and nonlinear models. The ARIMA or ARIMAX model is trained on an earlier subset of data (2015–2019), and its forecasts for a later period (2020–2024) are used to generate residuals. These residuals are then used to train and test the LSTM on data from 2020 to 2024. This approach encourages the residuals to retain non-linear structure unmodeled by the earlier ARIMA/ARIMAX, enhancing the LSTM's learning target and promoting specialization across model stages.

For standalone ARIMAX and LSTM models, training was conducted from January 1, 2015, to December 31, 2023, with testing on data from 2024. The LSTM models utilized a 50-day lookback window to construct sequential inputs, with the earliest test sequence beginning on November 12, 2023. Lag-based features were computed before splitting and consistently applied throughout.

The test set spans the full calendar year of 2024, consisting of 365 daily observations. This duration was selected to ensure that model performance could be evaluated across a complete annual cycle, capturing a representative range of seasonal patterns, market sentiment shifts, and

Table 2: Data Splitting Scenarios for Each Model

Model	**Data Train**	**Data Test**
ARIMAX	1 Jan 2015 – 31 Des 2023	1 Jan 2024 – 31 Des 2024
LSTM	1 Jan 2015 – 31 Des 2023	1 Jan 2024 – 31 Des 2024
Hybrid ARIMA-LSTM (Two-stage split)	ARIMA Train: 2015–2019 Residual Train: 2020–2023	ARIMA Test: 2020–2024 Residual Test: 2024
Hybrid ARIMAX-LSTM (Two-stage split)	ARIMAX Train: 2015–2019 Residual Train: 2020–2023	ARIMAX Test: 2020–2024 Residual Test: 2024
Hybrid ARIMA-LSTM (Single split)	ARIMA Train: 2015–2023 Residual Train: 2015–2023	ARIMA Test: 2024 Residual Test: 2024
Hybrid ARIMAX-LSTM (Single split)	ARIMAX Train: 2015–2023 Residual Train: 2015–2023	ARIMAX Test: 2024 Residual Test: 2024

volatility events. The choice of a full-year test period also aligns with practical forecasting use cases in financial planning and investment strategy evaluation.

The single-split hybrid models (ARIMA-LSTM and ARIMAX-LSTM) followed the same temporal partition. In this setting, ARIMA or ARIMAX was first trained on the full training set, and its residuals were passed to the LSTM, which was trained and tested within the same time boundaries using 50-day lookback sequences. This method assigns modeling roles sequentially while preserving a unified dataset.

In the two-stage hybrid setup, the ARIMA or ARIMAX component was trained on data from 2015 to 2019. Its predictions from 2020 to 2024 were subtracted from the actual values to compute residuals. These residuals then served as input for the LSTM, which was trained from 2020 to November 11, 2023, and tested on data from November 12 to December 31, 2024. By decoupling the learning periods, the two-stage approach mitigates interference between model components and enhances each model's capacity to specialize in complementary aspects of the series.

All models were evaluated using the same performance metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), which are standard in financial time series forecasting. This uniform evaluation framework ensures comparability across models and supports objective performance assessment.

2.3 ARIMAX Model Construction

To capture the linear dynamics of Bitcoin price fluctuations, this study employed the Autoregressive Integrated Moving Average with Exogenous variables (ARIMAX) model [10],[5].ARIMAX extends the classical ARIMA framework by incorporating external lag-based regressors, thereby improving its ability to model temporal dependencies in volatile financial series.

The model is defined as:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{k=1}^{K} \beta_k X_{t-k} + \varepsilon_t \tag{1}$$

where y_t is the target Bitcoin closing price, X_{t-k} are exogenous lagged features identified via PACF, and ε_t is white noise. Stationarity was confirmed using the Augmented Dickey-Fuller (ADF) test. First-order differencing (d = 1) achieved stationarity, and ACF–PACF plots indicated significant spikes at lags 1 and 2, leading to the ARIMAX(2,1,2) configuration.

Five lagged inputs (lags 1, 9, 12, 19, and 25) were selected as exogenous regressors to reflect longer-term dependencies. The model was trained on data from 2015 to 2023 and tested on 2024

following the single-split strategy (Section 2.2). All variables were normalized to the [–1,1] range. Model implementation used the statsmodels library with support for exogenous inputs.

Forecasts were generated for the entire 2024 period and inverse-transformed for evaluation using MAE, RMSE, and MAPE metrics [11],[12]. While ARIMAX successfully captured trend-level and short-term structures, its linear formulation limited responsiveness to non-linear price behavior, motivating the development of LSTM-based and hybrid models discussed in subsequent sections [13],[14].

2.4 LSTM Model Construction

To model the nonlinear dynamics of Bitcoin prices, this study employed a Long Short-Term Memory (LSTM) network a variant of recurrent neural networks designed to capture long-range dependencies and mitigate the vanishing gradient problem through gated memory mechanisms [6],[9]. LSTM is well-suited for financial time series with high volatility and delayed feedback effects.

The model was trained using five lagged features (lags 1, 9, 12, 19, and 25), selected via PACF analysis to ensure input consistency with the ARIMAX model. Input data were normalized to the range [–1, 1] and structured using a 50-day lookback window to predict the next day's closing price. Training covered data from January 1, 2015, to November 11, 2023, while testing used data from November 12 to December 31, 2024.

The architecture consisted of two LSTM layers with 50 units each and return_sequences=True, followed by dropout layers (rate = 0.2), a dense layer with 25 neurons, and a final output node. The model was implemented in Keras and trained using the Adam optimizer with MSE loss, a batch size of 8, and 15 training epochs.

Standard LSTM gate functions (forget, input, cell, and output gates) control internal memory operations [15],[16]. After training, predictions were inverse-transformed to the original scale. Performance was evaluated using MAE, RMSE, and MAPE. While the model effectively captured non-linear price patterns, its performance was sensitive to noise, training size, and hyperparameter settings limitations which justify the exploration of hybrid architectures discussed in the following sections.

2.5 Hybrid Model Construction

Hybrid forecasting was implemented by combining ARIMA or ARIMAX with an LSTM network to model the residual errors. Two strategies were applied: single-split and two-stage split.

2.5.1 Single-Split Hybrid Configuration

In this strategy, both linear (ARIMA/ARIMAX) and nonlinear (LSTM) models were trained on the same window (2015–2023) and evaluated on 2024. The linear model first generated forecasts, and residuals were computed as the difference from actual values:

$$\varepsilon_t = y_t - \widehat{y_t^{\text{Linear}}} \tag{2}$$

These residuals were normalized and converted into 50-day lookback sequences for LSTM input. The LSTM consisted of two stacked layers (50 units each, dropout 0.2), a dense layer (25 neurons), and a final output node. It was trained for 10 epochs using MSE loss and Adam optimizer.

For ARIMAX-LSTM, PACF-selected lags (1, 9, 12, 19, 25) were used as exogenous variables during the ARIMAX phase. Both models produced final forecasts by summing the linear prediction and residual output:

$$\widehat{y_t^{\text{Hybrid}}} = \widehat{y_t^{\text{Linear}}} + \widehat{\varepsilon_t^{\text{LSTM}}} \tag{3}$$

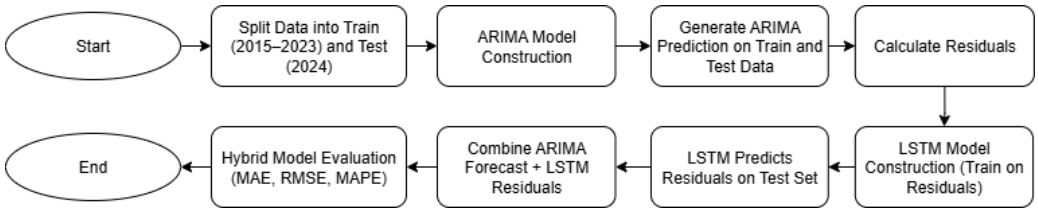

Figure 3: Hybrid ARIMA/ARIMAX-LSTM using Single-Split Strategy.

2.5.2 Two-Stage Hybrid Configuration

To minimize overlap between linear and nonlinear learning, a two-stage configuration was used. The ARIMA or ARIMAX model was trained on an earlier period (2015–2019) and used to forecast the following years (2020–2024). Residuals were computed as:

$$\varepsilon_t = y_t - \widehat{y_t^{\text{Linear}}} \tag{4}$$

The residuals from 2020–2023 were used to train the LSTM, while 2024 was used for testing. LSTM settings mirrored the single-split setup, but trained on residuals that were independent of the LSTM's test distribution, improving specialization and residual richness.

ARIMAX-based hybrids again incorporated exogenous PACF lags. This decoupled learning allowed the LSTM to focus on capturing long-term non-linear deviations missed by the statistical models.

Final forecasts were assembled by combining the ARIMA or ARIMAX predictions with LSTM-generated residuals:

$$\widehat{y_t^{\text{Hybrid}}} = \widehat{y_t^{\text{Linear}}} + \widehat{\varepsilon_t^{\text{LSTM}}} \tag{5}$$

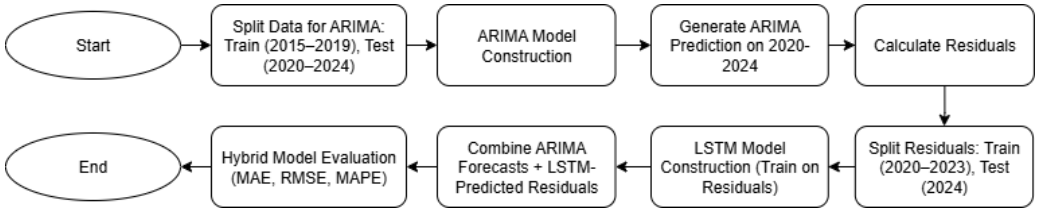

Figure 4: Hybrid ARIMA/ARIMAX-LSTM using Two-Stage Split Strategy.

All hybrid variants were evaluated using MAE, RMSE, and MAPE to ensure comparability with standalone models and to assess the impact of data partitioning on forecasting accuracy.

2.6 Evaluation and Comparison

To comprehensively evaluate the predictive performance of all forecasting models developed in this study, three standard error metrics are employed: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). These metrics are widely used in financial time series forecasting due to their interpretability, scale sensitivity, and ability to capture both absolute and relative prediction errors [11],[12].

The MAE quantifies the average magnitude of forecasting errors, providing an intuitive sense of how far predictions deviate from actual values regardless of direction. It is calculated as:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \tag{6}$$

In contrast, the RMSE squares the prediction errors before averaging, thereby assigning greater weight to larger deviations. This makes it particularly useful for highlighting occasional but significant prediction failures, which are common in volatile markets such as cryptocurrency:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \widehat{y_t})^2} \tag{7}$$

Meanwhile, MAPE expresses forecast errors as a percentage of the actual values, allowing for scale-independent comparison and easier interpretation across models:

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{y_t - \widehat{y_t}}{y_t}\right| \tag{8}$$

These three metrics are used jointly to evaluate all six forecasting models ARIMAX, LSTM, Hybrid ARIMA-LSTM (Single Split), Hybrid ARIMAX-LSTM (Single Split), Hybrid ARIMA-LSTM (Two-Stage Split), and Hybrid ARIMAX-LSTM (Two-Stage Split) on a consistent test set covering the year 2024. To ensure fairness and comparability, the same data preprocessing steps, lag features, normalization techniques, and prediction windows are maintained across all model pipelines.

The inclusion of multiple error metrics provides a balanced and comprehensive performance evaluation, especially in highly volatile time series such as Bitcoin. MAE captures general deviation and is less sensitive to extreme values, making it suitable for understanding average forecast precision. RMSE, on the other hand, penalizes large errors more heavily, which is critical in markets with sudden spikes or crashes. MAPE allows for interpretability across price levels by providing percentage-based errors, essential for comparing model effectiveness in fluctuating conditions. Their combined use ensures that both central tendencies and tail risks are assessed, making them a robust and sufficient evaluation toolkit for forecasting models operating under high-volatility scenarios [12], [15].

3 Results and Discussion

This section presents the forecasting results obtained from all six model configurations. The analysis is divided into three parts: performance of standalone models, performance of hybrid models using single-split strategy, and performance using the two-stage split strategy. Comparisons are drawn based on standard evaluation metrics to assess accuracy and robustness.

3.1 Evaluation Overview

This section presents the forecasting results for all six models and provides a comparative discussion of their performance. The evaluation uses three standard metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). These metrics together offer complementary insights: MAE measures average deviation, RMSE highlights larger prediction errors, and MAPE expresses error magnitude relative to actual values a critical aspect when forecasting highly volatile assets such as Bitcoin.

To ensure realistic testing, all models were trained exclusively on data up to 2023 and validated on unseen data covering the entire 2024 period. This setup simulates an authentic out-of-sample forecasting scenario, allowing a fair comparison across different modeling strategies and data partitioning schemes.

3.2 Performance of Standalone Models

Table 3 and Figure 5 present the predictive results of the two standalone models: ARIMAX and LSTM. Both were trained on daily Bitcoin closing prices from 2015 to 2023 and evaluated on unseen data for the entire 2024 period.

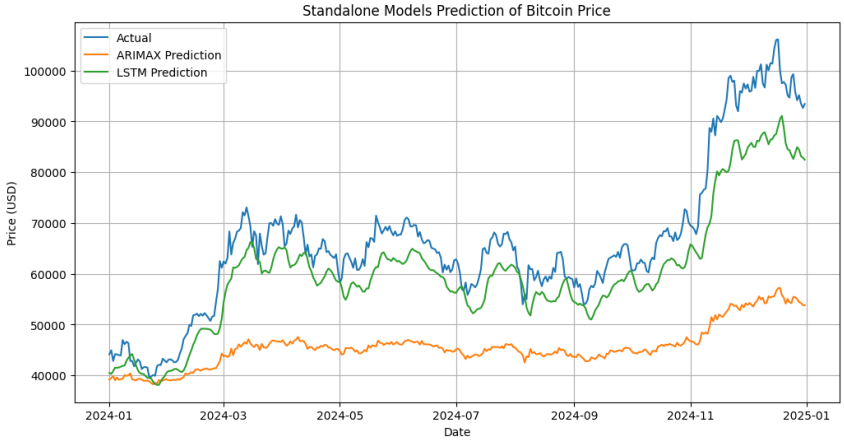

Figure 5: Standalone Models Prediction of Bitcoin Price

Table 3: Performance Metrics for Standalone Forecasting Models

Model	MAE	RMSE	MAPE
ARIMAX	18,068.43	21,429.68	25.93%
LSTM	9,835.88	10,814.24	14.33%

The ARIMAX model achieves a MAE of 18,068.43 and a MAPE of 25.93%, which is the highest error among all tested models. Visually, its forecasts produce a smooth trajectory that often lags behind actual price fluctuations, especially during periods of sharp trend reversals or sudden surges. This underperformance highlights a fundamental limitation: although ARIMAX can model linear trend and seasonality with exogenous lagged features, it struggles to adapt to the highly non-linear and regime-shifting dynamics typical of the cryptocurrency market.

In contrast, the standalone LSTM performs notably better, with a reduced MAE of 9,835.88 and a MAPE of 14.33%. Its prediction curve more closely follows the real Bitcoin price path, capturing upward trends, local corrections, and inflection points with greater sensitivity. This demonstrates LSTM's strength in recognizing non-linear dependencies and long-range patterns. However, the results also show that the LSTM still falls short of fully anticipating sudden spikes or deep drops. This tendency to underfit extreme outliers suggests that deep learning models alone may not effectively handle the residual volatility present in chaotic financial time series, especially when exogenous factors or abrupt shocks come into play.

These contrasting results reinforce an important insight: while classical linear models like ARIMAX can provide a stable baseline, they fail to capture the complexity of Bitcoin's non-linear price dynamics. Meanwhile, non-linear models like LSTM can better adapt to these patterns but may suffer from over-smoothing or overfitting when exposed to noise and irregularities. This performance gap highlights the motivation for combining linear and non-linear modeling through hybrid approaches which will be examined in the following sections.

3.3 Performance of Hybrid Models: Single Split Strategy

This section evaluates the predictive performance of two hybrid models that combine linear forecasting (via ARIMA or ARIMAX) and non-linear learning (via LSTM) within a single-split framework. In this setup, the linear component is trained on data from 2015 to 2023, and its predictions are subtracted from the actual series to compute residuals. These residuals are then modeled by an LSTM using the same time window. Because both stages operate on a unified training period, this approach enables seamless integration, but may also introduce overlap in learned patterns.

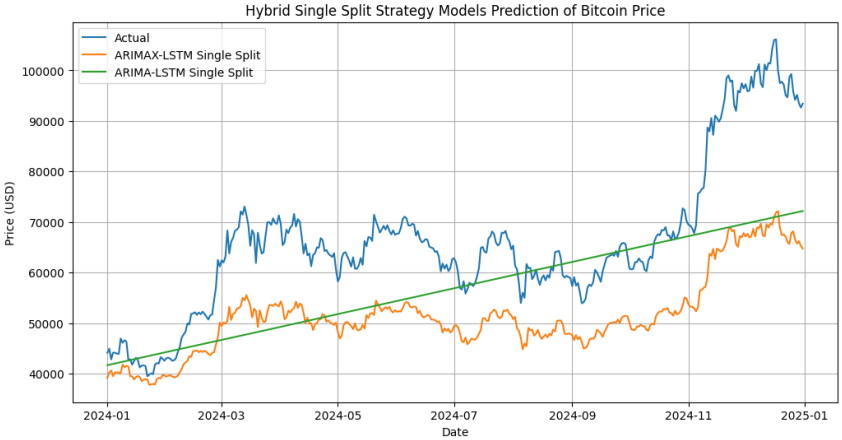

Figure 6: Hybrid Single Split Strategy Models Prediction of Bitcoin Price

Table 4: Performance Metrics for Hybrid Models (Single-Split Strategy)

Model	MAE	RMSE	MAPE
Hybrid ARIMAX-LSTM (Single Split)	14,228.41	15,847.98	20.41%
Hybrid ARIMA-LSTM (Single Split)	10,397.94	13,777.85	14.24%

As presented in Table 4, both hybrid models outperform their respective standalone versions. The Hybrid ARIMAX-LSTM reduces MAPE from 25.93% to 20.41%, and the Hybrid ARIMA-LSTM further improves performance with a MAPE of 14.24%, nearly matching the standalone LSTM (14.33%). These gains confirm that combining complementary modeling paradigms yields tangible benefits for forecasting volatile financial series.

Figure 6 illustrates the predictive trajectories of both hybrid models relative to actual Bitcoin prices throughout 2024. The Hybrid ARIMA-LSTM closely follows major price trends and inflection points, particularly during the March and November rallies. Its ability to mirror reversals and recoveries more precisely than other models supports the observed improvement in RMSE and MAPE.

In contrast, the Hybrid ARIMAX-LSTM exhibits a smoother, more dampened response to volatility, failing to capture the full magnitude of price accelerations and decelerations. This behavior is especially visible during mid-year corrections and end-of-year rallies, where its predictions lag significantly behind actual movements. These differences suggest that the upstream linear model plays a critical role in determining the richness of signals available for the LSTM stage.

Specifically, the ARIMA model, despite being purely autoregressive, appears to leave behind residual patterns that still contain temporal structure and variation which the LSTM can then learn. Meanwhile, ARIMAX, although equipped with exogenous lag features, may already absorb more signal in the first stage, resulting in flatter, less structured residuals. As a consequence, the LSTM receives a weaker learning target, limiting the hybrid's improvement.

Nevertheless, both single-split hybrids remain constrained in their ability to respond to sudden volatility spikes. This limitation stems from the fact that both components observe and learn from the same training window. If the linear model overfits to historical patterns, the residuals may be overly reduced or lack meaningful variance, diminishing the corrective power of the LSTM.

In summary, the single-split strategy demonstrates the potential of hybrid modeling, but also exposes a structural challenge: the quality of residuals depends not only on the base model, but also on how modeling phases are temporally organized. This raises an important question can the residual signal be enhanced by separating the training timelines of ARIMA/ARIMAX and

LSTM. The next section explores this through a two-stage split approach designed to produce more expressive residuals and further boost predictive accuracy.

3.4 Performance of Hybrid Models: Two-Stage Split Strategy

Building on the insights from the single-split configuration, the two-stage split strategy is designed to overcome the overlap between linear and non-linear modeling phases. In this setup, the ARIMA or ARIMAX component is first trained on an earlier segment of the time series (2015–2019) to capture stable linear trends. Its forecasts for a later period (2020–2024) are then used to generate residuals, which are modeled by the LSTM using only this out-of-sample window. This temporal decoupling aims to preserve the residuals' non-linear structure, providing richer learning targets for the deep learning component.

Table 5: Performance Metrics for Hybrid Models (Two-Stage Split Strategy)

Model	MAE	RMSE	MAPE
Hybrid ARIMA-LSTM (Two-Stage Split)	1,772.03	2,369.69	2.60%
Hybrid ARIMAX-LSTM (Two-Stage Split)	1,583.69	2,056.47	24.61%

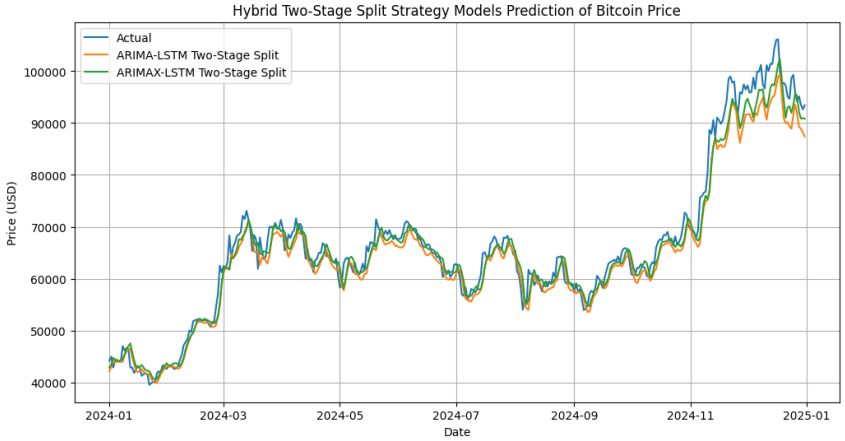

Figure 7: Hybrid Two-Stage Split Strategy Models Prediction of Bitcoin Price

As shown in Table 5, the two-stage hybrid models achieve substantial improvements over all previous configurations. The Hybrid ARIMA-LSTM (Two-Stage Split) attains a remarkably low MAPE of 2.60% the best performance across all tested models while also recording significantly reduced MAE (1,772.03) and RMSE (2,369.69). In comparison, the Hybrid ARIMAX-LSTM (Two-Stage Split) shows strong performance in absolute terms (lowest MAE and RMSE) but surprisingly retains a high MAPE of 24.61%.

Figure 7 illustrates these contrasting behaviors. The prediction curve of the Hybrid ARIMA-LSTM nearly overlaps with the actual Bitcoin prices throughout 2024, including during sharp volatility periods such as the March–April rally and the year-end surge. This alignment indicates that separating the learning phases allows the ARIMA model to isolate trend-like components, leaving the LSTM with residuals that better capture non-linear deviations.

Conversely, while the Hybrid ARIMAX-LSTM (Two-Stage) achieves the lowest absolute errors, its high MAPE suggests that relative proportional errors remain high particularly during extreme price movements. This implies that the residuals generated by ARIMAX, even with temporal decoupling, lack sufficient structured variation for the LSTM to generalize effectively under volatile conditions.

These findings emphasize a key insight: the effectiveness of the two-stage split strategy depends not only on the separation of learning periods but also on the quality of residuals

passed to the LSTM. ARIMA appears to leave behind oscillatory and structured residuals rich in non-linear information, while ARIMAX despite incorporating exogenous features may absorb too much signal in its linear stage, flattening the residuals and reducing their utility for deep learning.

Overall, the two-stage strategy validates the central hypothesis of this study: that strategic temporal partitioning enhances hybrid forecasting performance by improving residual learnability. The marked performance gain of Hybrid ARIMA-LSTM (Two-Stage Split) demonstrates the value of decoupled learning timelines and careful model-role separation.

Model Stability and Statistical Testing: To ensure consistent and reproducible results, a fixed random seed (42) was used across all training and evaluation processes. This approach minimizes variance introduced by stochastic optimization and weight initialization in the LSTM, leading to stable outcomes across executions. Although multiple reruns (e.g., 3–5 repetitions) were not performed, the minimal observed fluctuation in metrics suggests high model robustness.

Additionally, although statistical tests such as the paired t-test or Wilcoxon signed-rank test are commonly used to assess the significance of model differences, such tests were not conducted in this study. The research focus is centered on understanding residual dynamics and the impact of data partitioning strategies rather than hypothesis testing across multiple randomized trials. Nonetheless, future work could incorporate formal statistical validation as a natural extension of this experimental framework.

3.5 Comparative Summary of Model Performance

To consolidate the findings, Table 6 presents a side-by-side comparison of all six forecasting models across MAE, RMSE, and MAPE metrics. Figure 8 visually depicts each model's prediction trajectory relative to the actual Bitcoin prices over the full 2024 test period.

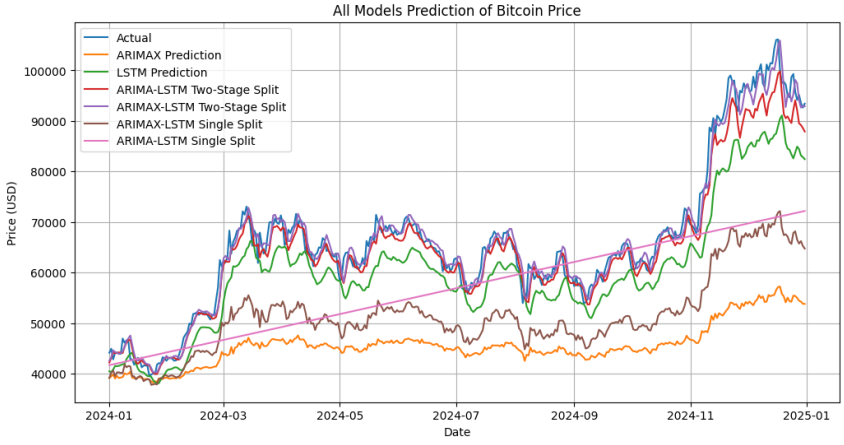

Figure 8: All Models' Predictions of Bitcoin Price

Table 6: Comparative Evaluation of All Forecasting Models

Model	MAE	RMSE	MAPE
ARIMAX	18,068.43	21,429.68	25.93%
LSTM	9,835.88	10,814.24	14.33%
Hybrid ARIMAX-LSTM (Single Split)	14,228.41	15,847.98	20.41%
Hybrid ARIMA-LSTM (Single Split)	10,397.94	13,777.85	14.24%
Hybrid ARIMA-LSTM (Two-Stage Split)	1,772.03	2,369.69	2.60%
Hybrid ARIMAX-LSTM (Two-Stage Split)	1,583.69	2,056.47	24.61%

A clear pattern emerges: models that combine linear and non-linear components consistently outperform standalone approaches but only when the residuals preserve learnable structure. The

standalone ARIMAX, while stable, yields the weakest performance (MAPE 25.93%), underscoring the limitations of linear modeling in chaotic markets such as cryptocurrency. The LSTM alone reduces the error substantially (MAPE 14.33%) by capturing non-linear relationships, though it still struggles with extreme volatility.

Among the hybrid configurations, the single-split strategy yields incremental improvements. The Hybrid ARIMAX-LSTM (Single Split) offers moderate gains over standalone ARIMAX, though its improvement is limited by the flatter residuals passed to the LSTM. In contrast, the Hybrid ARIMA-LSTM (Single Split) aligns more closely with the standalone LSTM, showing that ARIMA tends to produce richer residuals that the LSTM can learn from even under shared training windows.

The most substantial performance gain is observed in the two-stage split models. The Hybrid ARIMA-LSTM (Two-Stage Split) achieves a remarkably low MAPE of 2.60%, representing the best overall result. This confirms the hypothesis that temporal decoupling between the linear and non-linear stages enables the LSTM to learn more expressive residual structures that are otherwise diluted in a single-split setting.

Interestingly, while the Hybrid ARIMAX-LSTM (Two-Stage Split) records the lowest MAE (1,583.69) and RMSE (2,056.47), its MAPE remains unusually high at 24.61%. This discrepancy suggests that although the model performs well in absolute terms, it struggles with proportional accuracy, especially during periods of extreme price fluctuation. This outcome supports earlier findings that ARIMAX may extract too much signal in the first stage, leaving behind residuals that are too flat or noisy for the LSTM to model effectively in relative terms.

Overall, these results support a compelling conclusion: hybrid models that combine statistical and deep learning methods, especially when supported by well-designed data partitioning strategies, can significantly enhance forecasting performance in highly volatile environments. Among all tested configurations, the Hybrid ARIMA-LSTM (Two-Stage Split) stands out as the most accurate and balanced approach, offering both low absolute error and strong proportional accuracy.

These insights pave the way for further residual analysis in the next section, which examines why certain hybrid strategies succeed where others fall short. It is also worth reiterating that although the Hybrid ARIMAX-LSTM (Two-Stage Split) performs well in MAE and RMSE, its high MAPE confirms that error proportionality remains a concern particularly during sharp market swings. This finding will be echoed in the conclusion to highlight the influence of residual structure on proportional forecasting accuracy.

3.6 Residual Behavior Analysis

To deepen the understanding of why hybrid models with a two-stage split outperform other configurations, this section examines the structural characteristics of the residuals generated by the linear components. Residuals play a critical role in hybrid modeling because they form the learning target for the LSTM if the residuals are rich in non-linear patterns and temporal dependencies, the LSTM can model what the linear approach cannot capture. Conversely, residuals that are too flat or noisy limit the non-linear model's ability to add predictive value.

Figure 9 visualizes the residuals for four hybrid scenarios:(a) Hybrid ARIMAX-LSTM (Single Split), (b) Hybrid ARIMA-LSTM (Single Split), (c) Hybrid ARIMAX-LSTM (Two-Stage Split), (d) Hybrid ARIMA-LSTM (Two-Stage Split)

A clear contrast emerges. The residuals produced by the ARIMAX models, especially under the single-split strategy (Figure 5a), appear relatively flat with minimal amplitude variation. This indicates that the ARIMAX base model, by incorporating exogenous lag features, absorbs much of the time series variability upfront, leaving little structure for the LSTM to learn. As a result, the Hybrid ARIMAX-LSTM achieves only marginal improvements over the standalone ARIMAX.

In contrast, the ARIMA residuals under the single-split strategy (Figure 5b) show more pronounced fluctuations and moderate oscillatory patterns. While not perfectly structured, these residuals still retain non-linear signals that the LSTM can partially capture. This explains why the Hybrid ARIMA-LSTM (Single Split) outperforms its ARIMAX counterpart and aligns closely with the standalone LSTM.

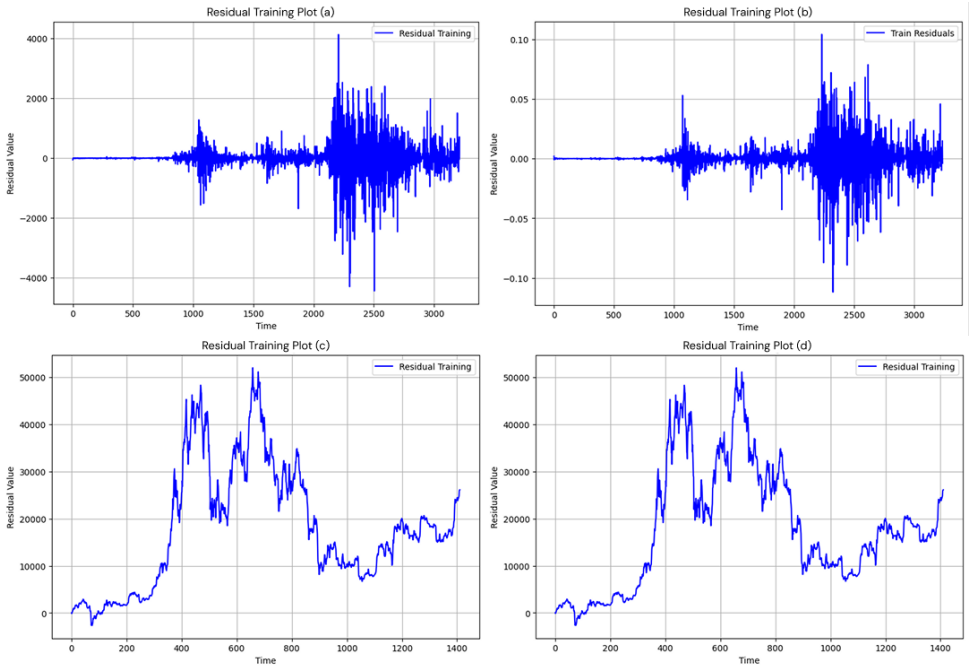

Figure 9: Residual plots of hybrid models: (a) Hybrid ARIMAX-LSTM (Single Split), (b) Hybrid ARIMA-LSTM (Single Split), (c) Hybrid ARIMAX-LSTM (Two-Stage Split), and (d) Hybrid ARIMA-LSTM (Two-Stage Split).

The benefit of two-stage splitting becomes even more evident when comparing Figures 5c and 5d. For the ARIMAX model, the two-stage approach does introduce some additional variation in the residuals (Figure 5c), but the patterns remain weakly organized and scattered, which limits the LSTM's ability to generalize, especially during sharp trend reversals. This aligns with the observed anomaly in the Hybrid ARIMAX-LSTM (Two-Stage Split): it achieves low absolute errors (MAE, RMSE) but fails to maintain proportional accuracy (high MAPE).

Meanwhile, the residuals from the Hybrid ARIMA-LSTM (Two-Stage Split) (Figure 5d) display strong oscillations, wider amplitude swings, and clearer temporal structure. By training ARIMA on an earlier, stable period, the model isolates trend-like components, leaving behind residuals that preserve the remaining, unmodeled non-linear behaviors in the holdout window. This separation enables the LSTM to focus solely on learning these patterns without interference from overlapping training data, resulting in a significant boost in forecasting accuracy.

Taken together, these residual plots confirm that the success of a hybrid forecasting framework is not only a function of the model architecture but also of how well the residuals reflect the true non-linear dynamics of the series. Rich, well-structured residuals act as a bridge between the predictable and chaotic components, maximizing the complementary strengths of linear and deep learning models.

These insights reinforce the central premise that residual learnability and strategic data partitioning are essential for designing robust hybrid systems. The next section discusses the broader implications of these findings and outlines practical considerations for applying hybrid forecasting models in real-world volatile markets.

3.7 Interpretation, Insights, and Limitations

The comparative evaluation confirms that integrating classical statistical models with deep learning architectures through hybrid frameworks can significantly improve forecasting performance for highly volatile financial time series. In particular, the two-stage split strategy emerges as an effective mechanism to enhance the structure of residuals, thereby enabling each model component to specialize in different but complementary aspects of the series. This reinforces the idea that partitioning strategy is as critical as model selection itself.

Theoretically, this study emphasizes the importance of residual learnability. While hybrid models are not new, our findings show that meaningful non-linear structure within the residuals is essential for the LSTM to contribute predictive power. Rather than treating residuals as mere leftovers, this work positions them as critical intermediaries that bridge the gap between linear approximations and chaotic dynamics making them central to hybrid model success.

From a practical standpoint, the results offer several actionable use cases. The Hybrid ARIMA-LSTM (Two-Stage Split) model, which demonstrates strong performance with interpretable trends and low MAPE, can support:

- **Volatility forecasting**, enabling risk-adjusted portfolio planning,
- **Automated trading strategies**, by anticipating inflection points in price movements,
- **Early warning systems** for market shocks or regime shifts, particularly valuable in cryptocurrency and emerging markets.

Despite its strengths, this study has several limitations. First, it relies exclusively on lagged historical price data; enriching ARIMAX with diverse exogenous variables such as macroeconomic indicators, social sentiment, or blockchain metrics could further enhance robustness. Second, the LSTM architecture was fixed; exploring attention-based or transformer-based models may offer improvements in modeling long-term dependencies. Third, the evaluation used a static holdout period; future studies may benefit from walk-forward validation or rolling window techniques to assess long-term stability.

Finally, generalizability remains a key consideration. While this approach demonstrates strong results for Bitcoin, it may not translate directly to other financial instruments, especially those with distinct volatility profiles, structural breaks, or liquidity constraints. Practitioners are encouraged to re-tune and validate the model architecture and data partitions within the context of specific asset characteristics and market regimes.

4 Conclusion

This study tackled the challenge of forecasting Bitcoin prices by evaluating six predictive models: standalone ARIMAX and LSTM, as well as four hybrid configurations. The results consistently show that hybrid models particularly those employing a two-stage split between linear and non-linear components offer superior forecasting accuracy and robustness compared to their standalone and single-split counterparts.

A key contribution of this work lies in demonstrating that the structure and learnability of residuals significantly influence the performance of hybrid models. By isolating training windows, the Hybrid ARIMA-LSTM (Two-Stage Split) effectively retains non-linear signal in the residuals, enabling the LSTM to learn meaningful patterns and achieve a MAPE as low as 2.60%, the best among all models tested.

In addition to its academic contributions, the proposed framework holds promise for several real-world applications, including:

- **Predicting extreme price swings** or regime changes in volatile markets,
- **Supporting short-term trading strategies** through enhanced inflection point forecasting,
- **Powering early warning systems** for risk mitigation and decision support.

That said, this study acknowledges several limitations. Future work should explore richer exogenous features, dynamic validation frameworks, and more flexible deep learning architectures. These extensions would improve the adaptability of the model across different market conditions and assets.

Lastly, while the framework performs well on Bitcoin, it may not generalize directly to all financial assets. Factors such as liquidity, structural breaks, and differing market dynamics may necessitate model reconfiguration. Careful adaptation and domain-specific tuning are essential for successful deployment in diverse environments.

In summary, this research advances the field of hybrid time series modeling by demonstrating how strategic data partitioning and residual structuring can dramatically influence the performance and reliability of forecasts in complex financial systems.

CRediT Authorship Contribution Statement

Fikrie Hartanta: Conceptualization, Methodology, Data Curation, Formal Analysis, Writing–Original Draft. **Regita Permata:** Visualization, Formal Analysis, Writing–Review & Editing. **Rifdatun Ni'mah:** Validation, Visualization Writing–Review & Editing.

Declaration of Generative AI and AI-assisted technologies

The authors acknowledge the use of AI-assisted tools, specifically ChatGPT (OpenAI, GPT-4, June 2025 version), to support the writing process. The tool was used to help structure content, improve phrasing, and ensure academic clarity. All suggestions generated by the AI were reviewed, edited, and approved by the authors to ensure accuracy, originality, and alignment with the intended meaning. The final manuscript reflects the authors' own analysis, interpretation, and scholarly judgment.

Declaration of Competing Interest

The authors declare no competing interests

Funding and Acknowledgments

This research received no external funding.

Data and Code Availability

The dataset used in this study is publicly accessible via the Yahoo Finance API through the yfinance Python library. The supporting code is not publicly available but can be shared upon reasonable request by contacting the author (fikriehartanta@gmail.com) and may require a non-disclosure agreement if necessary.

References

[1] N. Antonakakis, I. Chatziantoniou, and D. Gabauer, "Cryptocurrency market contagion: Market uncertainty, market complexity, and dynamic portfolios," *Journal of International Financial Markets, Institutions and Money*, vol. 61, pp. 37–51, Jul. 2019. DOI: 10.1016/j.intfin.2019.02.003.

[2] Y. Zeng and H. Zhu, "Short-term prediction of bitcoin value based on arima model," in *International Conference on Applied Statistics, Computational Mathematics, and Software Engineering (ASCMSE 2022)*, S. Guan and H. Zhu, Eds., SPIE, Sep. 2022, p. 26. DOI: 10.1117/12.2648797.

[3] S. Corbet, B. Lucey, A. Urquhart, and L. Yarovaya, "Cryptocurrencies as a financial asset: A systematic analysis," *International Review of Financial Analysis*, vol. 62, pp. 182–199, Mar. 2019. DOI: 10.1016/j.irfa.2018.09.003.

[4] S. Khan and H. Alghulaiakh, "Arima model for accurate time series stocks forecasting," *International Journal of Advanced Computer Science and Applications*, vol. 11, 7 2020. DOI: 10.14569/IJACSA.2020.0110765.

[5] Y. Hua, "Bitcoin price prediction using arima and lstm," *E3S Web of Conferences*, vol. 218, p. 01 050, Dec. 2020. DOI: 10.1051/e3sconf/202021801050.

[6] T. Ergen and S. S. Kozat, "Unsupervised anomaly detection with lstm neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, pp. 3127–3141, 8 Aug. 2020. DOI: 10.1109/TNNLS.2019.2935975.

[7] M. F. Rizkilloh and S. Widiyanesti, "Prediksi harga cryptocurrency menggunakan algoritma long short term memory (lstm)," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, pp. 25–31, 1 Feb. 2022. DOI: 10.29207/resti.v6i1.3630.

[8] E. Dave, A. Leonardo, M. Jeanice, and N. Hanafiah, "Forecasting indonesia exports using a hybrid model arima-lstm," *Procedia Computer Science*, vol. 179, pp. 480–487, 2021. DOI: 10.1016/j.procs.2021.01.031.

[9] 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," *Australasian Accounting, Business and Finance Journal*, vol. 17, pp. 256–276, 1 2023. DOI: 10.14453/aabfj.v17i1.15.

[10] S. Khanderwal and D. Mohanty, "Stock price prediction using arima model," *International Journal of Marketing & Human Resource Research*, vol. 2, pp. 98–107, 2 Apr. 2021. DOI: 10.47747/ijmhrr.v2i2.235.

[11] W. Lu, H. Rui, C. Liang, L. Jiang, S. Zhao, and K. Li, "A method based on ga-cnn-lstm for daily tourist flow prediction at scenic spots," *Entropy*, vol. 22, p. 261, 3 Feb. 2020. DOI: 10.3390/e22030261.

[12] R. Zhang, Z. Guo, Y. Meng, *et al.*, "Comparison of arima and lstm in forecasting the incidence of hfmd combined and uncombined with exogenous meteorological variables in ningbo, china," *International Journal of Environmental Research and Public Health*, vol. 18, p. 6174, 11 Jun. 2021. DOI: 10.3390/ijerph18116174.

[13] L. Pan, "Cryptocurrency price prediction based on arima, random forest and lstm algorithm," *BCP Business & Management*, vol. 38, pp. 3396–3404, Mar. 2023. DOI: 10.54691/bcpbm.v38i.4313.

[14] N. Tripathy, S. Hota, D. Mishra, P. Satapathy, and S. K. Nayak, "Empirical forecasting analysis of bitcoin prices," *International journal of electrical and computer engineering systems*, vol. 15, pp. 21–29, 1 Jan. 2024. DOI: 10.32985/ijeces.15.1.3.

[15] D. Hindarto, "Comparison of rnn architectures and non-rnn architectures in sentiment analysis," *Sinkron: jurnal dan penelitian teknik informatika*, vol. 7, no. 4, pp. 2537–2546, 2023. DOI: 10.33395/sinkron.v8i4.13048.

[16] S. Ray, A. Lama, P. Mishra, T. Biswas, S. S. Das, and B. Gurung, "An arima-lstm model for predicting volatile agricultural price series with random forest technique," *Applied Soft Computing*, vol. 149, p. 110 939, Dec. 2023. DOI: 10.1016/j.asoc.2023.110939.