Vol. 8, No. 3, June 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

A Robustness-Oriented Evaluation of LSTM, GRU, and Hybrid LSTM-GRU Models for ANTM.JK Stock Price Forecasting

Khoirudin[1], Prind Triajeng Pungkasanti[2], Nur Wakhidah[3], Vinay Rishiwal[4]

[1]Department of Information Technology, Universitas Semarang, Indonesia

[2,3] Department of CSIT, MJP RohilKhand University, Bareilly, India

Received:
October 12, 2025
Revised:
May 17, 2026
Accepted:
June 24, 2026
Published:
June 27, 2026

Corresponding Author:
Author Name*:
Khoirudin
Email*:
khoirudin@usm.ac.id

DOI:
10.63158/journalisi.v8i3.1660

Abstract. Accurately forecasting stock prices remains challenging because of the nonlinear and volatile nature of financial markets, particularly during periods of heightened uncertainty, such as the COVID-19 pandemic. This study evaluates the robustness of three models, LSTM, GRU, and Hybrid LSTM-GRU, for ANTM.JK stock price forecasting using a volatility-oriented evaluation framework. Historical stock data from September 2005 to May 2022 were transformed into supervised time-series datasets using a 15-lag sliding window. The model performance was evaluated using baseline prediction accuracy, 5-fold chronological cross-validation consistency, and synthetic stress scenarios consisting of controlled price drops, price rises, and high-volatility noise. Evaluation metrics included RMSE, MSE, MAE, R, and R^2. The GRU model delivered the top baseline prediction results, achieving the smallest RMSE of 52.95 and MAE of 28.14. In cross-validation, the LSTM model recorded the lowest average RMSE of 119.41. Meanwhile, the Hybrid LSTM-GRU exhibited the highest prediction consistency and robustness across various synthetic stress scenarios. In contrast to earlier research that mainly focused on prediction precision, this study presents a comprehensive framework for evaluating robustness. This framework combines baseline accuracy, consistency through cross-validation, and an analysis of synthetic stress scenarios. The generated robustness map offers a systematic interpretation of model strengths across diverse evaluation goals, facilitating a more thorough assessment of stock-forecasting models in different market environments.

Keywords: stock price forecasting; deep learning; LSTM; GRU; Hybrid LSTM-GRU; robustness analysis; ANTM.JK

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

The COVID-19 pandemic led to major disruptions in global financial markets, increasing the unpredictability of stock prices. Throughout this time, financial time series data exhibited increased volatility, nonlinearity, and non-stationarity, complicating the task of forecasting stock prices compared to typical market conditions [1], [2], [3], [4], [5]. In this scenario, prediction models are required to deliver low forecasting errors under standard conditions and to sustain consistent performance when market dynamics shift suddenly. Consequently, model robustness, as defined in this study, pertains to the steadiness of the prediction performance and its ability to withstand unusual market conditions, serving as a crucial evaluation criterion in financial forecasting. The Indonesian stock market experienced notable fluctuations during the COVID-19 pandemic. Specifically, PT ANTM.JK experienced significant stock price movements from the onset of the pandemic to the Omicron wave in late 2021 and early 2022 [6].

Figure 1. ANTM.JK Stock Price [7]

This condition makes ANTM.JK a relevant case for examining the robustness of stock price forecasting models under pandemic-related volatility. As illustrated in Figure 1, the stock price pattern of ANTM.JK reflects highly dynamic market behavior, where sharp increases, declines, and unstable movements may reduce forecasting reliability if models are evaluated only under ordinary conditions [8], [9]. Although the motivation for this study is derived from the volatility observed during the COVID-19 pandemic, the dataset used in this research spans a longer historical period from September 2005 to May 2022.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Therefore, the COVID-19 pandemic and Omicron wave are treated as volatility contexts for evaluating forecasting robustness rather than as the exclusive training period. The availability of long-term historical data allows the proposed models to be evaluated across multiple market regimes while retaining the pandemic period as an important reference for market instability.

Several statistical and machine learning methods have been applied to stock price forecasting, including Support Vector Machines (SVM), Genetic Algorithm (GA)-based optimization, and other traditional predictive models [8]. Although these methods are useful for certain prediction tasks, financial time-series data often contain complex temporal dependencies that are difficult to capture using conventional approaches. Deep learning models, particularly Recurrent Neural Network (RNN)-based architectures, offer an alternative because they are specifically designed to process sequential data and learn temporal patterns. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are among the most widely adopted recurrent architectures for time-series forecasting due to their ability to address the limitations of conventional RNNs in capturing long-term dependencies [9], [10], [11]

LSTM employs memory cells and gating mechanisms to preserve relevant historical information, whereas GRU provides a simpler structure with fewer parameters and lower computational complexity [10], [11], [12], [13]. In addition to single-architecture approaches, hybrid LSTM-GRU models have been proposed to combine the complementary strengths of both architectures in modeling short-term and long-term temporal patterns [14], [15]. Recent studies have shown that deep learning and hybrid models can improve forecasting performance in volatile and nonstationary environments. For example, Ge [16] proposed a hybrid MEME-AO-LSTM model for forecasting several global market indices, whereas Mirza et al. [17] introduced a symbolic state deep learning framework to address non-stationary market regimes during the COVID-19 period. Most earlier research has predominantly assessed forecasting models using individual performance indicators, such as RMSE, MAE, MSE, or R^2, within a particular testing setup. Although these indicators are significant, they do not completely reveal whether a model maintains stability across various time segments or withstands unusual market scenarios. Consequently, there is a scarcity of data on the performance of deep learning models when both prediction consistency and robustness are assessed together. Few stock forecasting studies have explicitly investigated the resilience of models during abrupt price drops, swift price surges and significant volatility noise [17]. There is a research gap here, as a model that performs well on average may not maintain its reliability if market conditions shift suddenly. Grasping the robustness of a model is crucial because its appropriateness can differ among various markets.

To address this gap, this study examines the predictive performance of LSTM, GRU, and Hybrid LSTM-GRU models for ANTM.JK stock prices during market fluctuations triggered by COVID-19. Robustness analysis was conducted from two perspectives. First, the prediction consistency was evaluated by applying 5-fold chronological cross-validation to the original dataset. The models were then tested under synthetic stress scenarios, which included controlled price decreases, increases, and high-volatility noise conditions. Consequently, this study extends model evaluation beyond standard prediction accuracy by investigating how different deep learning architectures perform under various market stress conditions.

The primary aim of this research is to assess and contrast the resilience of the LSTM, GRU, and Hybrid LSTM-GRU models in forecasting ANTM.JK stock prices amid the volatility linked to the COVID-19 pandemic. This study specifically seeks to (1) determine the baseline predictive capabilities of the LSTM, GRU, and Hybrid LSTM-GRU models; (2) evaluate the consistency of these models' performance through chronological cross-validation on the original dataset; and (3) examine the models' robustness in synthetic market-stress scenarios. This study contributes to the stock forecasting literature in three ways. First, it introduces a multidimensional framework for evaluating robustness that goes beyond traditional accuracy metrics. Second, it assesses forecasting models by considering both chronological cross-validation consistency and controlled synthetic stress scenarios, offering a more comprehensive evaluation of model behavior under different market conditions. Third, it presents a robustness map that highlights the relative strengths and weaknesses of the LSTM, GRU, and Hybrid LSTM-GRU models, aiding context-aware model selection in volatile financial settings. These contributions provide a more thorough understanding of forecasting model robustness than methods that focus solely on average prediction accuracy.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

2. LITERATURE REVIEW

Stock price forecasting has been extensively studied because financial markets generate complex time-series data that are nonlinear, noisy, and highly sensitive to external events. Traditional forecasting approaches, including statistical models and machine learning techniques, have been widely applied to predict stock-price movements. For example, Ramdhani and Mubarok [8] employed an SVM combined with GA optimization to forecast ANTM.JK stock prices and demonstrated that ML methods can provide useful predictive performances. However, conventional machine learning approaches often rely on extensive feature engineering and may have limitations in capturing the sequential dependencies that naturally occur in financial time-series data. To address these limitations, deep learning architecture has gained increasing popularity in financial forecasting. Among recurrent neural network architectures, LSTM and GRU are the most widely adopted models because they are specifically designed to learn temporal dependencies from sequential observations [10], [18]. Previous studies have shown that recurrent models can effectively capture the nonlinear and time-dependent behaviors of stock market data [19], [20], [21]. Although both architectures have demonstrated strong forecasting capabilities, their relative performance often depends on the characteristics of the dataset and market conditions being analyzed.

Recent research has explored hybrid deep learning architectures that combine different recurrent models to improve forecasting performance. Hybrid LSTM-GRU models attempt to exploit the complementary strengths of LSTM and GRU in learning both long-term and short-term temporal patterns [14], [15]. In addition to hybrid recurrent architecture, more advanced forecasting frameworks have been proposed. Ge [16] introduced a hybrid MEME-AO-LSTM model and reported improved forecasting performance for several global stock indices. Similarly, Mirza et al. [17] developed a symbolic deep learning framework capable of adapting to non-stationary market regimes during the COVID-19 period. These studies demonstrate the continuing evolution of deep learning approaches for addressing complex financial forecasting problems. Despite these advances, most previous studies continue to evaluate forecasting performance primarily using accuracy-oriented metrics such as RMSE, MAE, MSE, and R^2. While these measures provide useful information regarding prediction quality, they do not fully explain whether a model remains stable when evaluated across different temporal partitions or is resilient when

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

exposed to abnormal market conditions. Consequently, model robustness remains relatively underexplored in stock forecasting literature. A forecasting model that performs well on a single testing dataset may not necessarily maintain comparable performance when the market behavior changes abruptly. Therefore, evaluating robustness alongside conventional accuracy metrics is important for obtaining a more comprehensive understanding of model behavior under varying market conditions.

Table 1 summarizes the position of this study in relation to previous research. The comparison focuses on the forecasting model, dataset, evaluation perspective, and whether a robustness assessment was explicitly conducted. Note that the numerical performance values reported across studies are not directly comparable because they were obtained from different datasets, price scales, forecasting horizons, and evaluation periods. Therefore, the comparison is intended to highlight methodological differences rather than rank forecasting performance.

Table 1. Analysis in Relation to Prior Research

Model	**Dataset**	**Evaluation**	**Robustness**
SVM + GA [8]	ANTM.JK	RMSE = 10.495	Not explicitly tested
LSTM [22]	ANTM.JK	LSTM reported as effective for stock prediction	Not explicitly tested yet
LSTM-RNN [23]	BITCOIN	RMSE = 0.14 (Bitcoin scale)	Not explicitly tested
Hybrid MEME-AO-LSTM [16]	S&P 500, CSI 300	Improved forecasting performance on global indices	Addresses non-stationary resilience
Symbolic-State CRNN [17]	Multi-market	Significant MSE reduction across market regimes	Focuses on regime-shift adaptation
LSTM [19]	ANTM.JK	LSTM reported as effective for stock prediction	Not explicitly tested
This Study	**ANTM.JK**	**Baseline accuracy, cross-validation consistency, and synthetic stress-scenario performance**	**Explicitly tested using chronological cross-validation and synthetic stress scenarios**

A review of the literature shows that previous research has greatly improved the accuracy of stock price predictions using machine learning, deep learning, and hybrid forecasting models. However, most studies still focus mainly on prediction accuracy, with little attention paid to robustness. Few reports have evaluated baseline forecasting performance, consistency in cross-validation, and resilience in controlled stress environments simultaneously. This study aims to address this gap by presenting a comprehensive framework for evaluating robustness. It compares the LSTM, GRU, and Hybrid LSTM-GRU models, considering both chronological cross-validation consistency and artificial stress scenarios. This study expands the scope of forecasting evaluation beyond traditional accuracy measures, providing a more complete view of model robustness under different market conditions.

3. METHODS

3.1 Study Framework

This study introduces a prediction framework focused on evaluating the robustness of the LSTM, GRU, and Hybrid LSTM-GRU models for ANTM.JK stock prices. The framework assesses the model's performance from three perspectives: basic prediction accuracy, consistency in cross-validation, and strength under simulated stress conditions. This study examines the robustness of deep learning models for ANTM.JK stock price forecasting during the COVID-19 pandemic's volatility. Three RNN-based models were compared: LSTM, GRU, and Hybrid LSTM-GRU. The research framework includes steps such as data collection, data preparation, supervised sequence generation, chronological data splitting, model development, model evaluation, consistency testing, robustness testing, visualization, and model recommendation, as illustrated in Figure 2.

The framework in Figure 2 was created to assess the model performance in three ways. First, the baseline prediction accuracy was verified using the original test data. Second, the prediction consistency was tested using 5-fold chronological cross-validation on the original dataset. Third, the model resilience was tested using synthetic extreme scenarios, including extreme drops, extreme rises, and high-volatility conditions. Thus, the evaluation considers both the average forecasting accuracy and whether each model can perform reliably under various market conditions.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

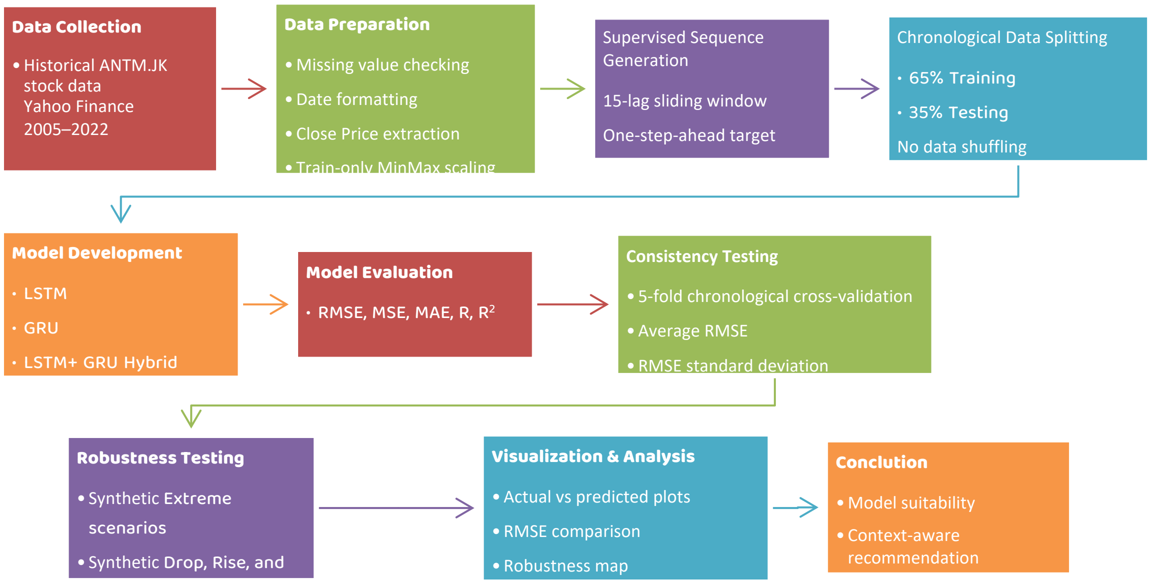

Figure 2. Research Framework

3.2 Data Collection and Preparation

The dataset in Table 2 consists of 4,128 daily observations of ANTM.JK stock data from September 29, 2005, to May 31, 2022. The data were chronologically divided into 2,683 training observations (65%) and 1,445 test observations (35%). Only the Close Price variable was used as the forecasting target. The Close Price was selected because it represents the final market valuation at the end of each trading session and is the most used variable in univariate stock forecasting studies. Although the dataset covers a long historical period, the COVID-19 and Omicron periods are used as the main volatility context because ANTM.JK experienced substantial price fluctuations during this period.

Table 2. Dataset ANTM.JK **[24]**

Date	**Open**	**High**	**Low**	**Close**	**Adj Close**	**Volume**
29/09/2005	432.59	436.79	407.39	432.59	298.01	76180670
30/09/2005	457.79	457.79	432.59	457.79	315.37	105493978
...	...	...	...	...	...	...
31/05/2022	2570	2570	2500	2510	2471.11	168242000

Data preparation included missing value checking, date formatting, Close Price extraction, normalization, and supervised sequence generation. The Close Price data were normalized using $MinMaxScaler$ into the range $[0,1]$. To reduce the risk of data leakage,

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

the scaler fitted only to the training data and then applied to the testing data. The dataset was divided chronologically into 65% training data and 35% testing data without randomized shuffling. This chronological split was applied because stock price data are time-series observations in which future information must not be used to train models to predict earlier observations. The normalized Close Price series was transformed into a supervised learning dataset using a 15-lag sliding window. The 15-lag window represents approximately three trading weeks and is used to providing recent historical information for one-step-ahead forecasting. Let P_t denote the normalized closing price at time t. The input-output formulation is defined as shown in Equation 1 and 2.

$$X_t = [P_{t-15}, P_{t-14}, \dots, P_{t-1}] \quad (1)$$

$$y\_t = P\_t \quad (2)$$

where X_t represents the input sequence and y_t represents the next-day Close Price target. Thus, this study performed one-step-ahead forecasting rather than multi-step forecasting.

3.3 Model Development

Three deep learning models were developed using TensorFlow-Keras: LSTM, GRU, and Hybrid LSTM-GRU. All models used the same input shape, consisting of 15-time steps and one feature. The output layer used a linear activation function because the forecasting task was a regression problem. To ensure a fair comparison, the main experimental configuration was kept consistent across all models, as shown in Table 3. Each model used recurrent layers with 32 units, the hyperbolic tangent activation function, Mean Squared Error (MSE) as the loss function, Adam as the optimizer, 200 training epochs, and a batch size of 5. The LSTM model was constructed using LSTM-based recurrent layers, followed by a dense output layer. The GRU model was constructed using GRU-based recurrent layers, followed by a dense output layer. The Hybrid LSTM-GRU model combined LSTM and GRU recurrent blocks before the final dense output layer. Although all models ended with a dense layer, the hybrid model used a more complex recurrent configuration by combining the LSTM and GRU components to capture complementary temporal patterns before producing the final prediction. Using the same input window, optimizer, loss function, epoch number, and batch size ensured that the comparison focused on the architectural differences between LSTM, GRU, and Hybrid LSTM-GRU. All models were trained under identical experimental conditions to ensure their fairness. Thus, any differences in forecasting performance are mainly due to architectural features and not training setup variations.

Table 3. Model Configuration

Model	Recurrent Architecture	Units	Optimizer	Loss	Epochs
LSTM	LSTM (32), LSTM (32), LSTM (32)	32 per layer	adam	mean_squared_error	200
GRU	GRU (32), GRU (32), GRU (32), GRU (32)	32 per layer	adam	mean_squared_error	200
Hybrid LSTM-GRU	LSTM (32), LSTM (32), GRU (32), GRU (32)	32 per layer	adam	mean_squared_error	200

3.4 Model Evaluation and Consistency Testing

The baseline model performance was evaluated using a chronological testing set. The evaluation metrics included RMSE, MSE, MAE, R, and R^2. These metrics were selected to assess the prediction accuracy, error magnitude, and ability of the models to explain variations in stock prices. To evaluate prediction consistency, a 5-fold chronological cross-validation procedure was implemented using the $TimeSeriesSplit$ strategy. Unlike conventional k-fold cross-validation, $TimeSeriesSplit$ preserves the temporal order of observations by ensuring that earlier data are used for training and later data are used for testing in each fold. This approach is consistent with the best practices for financial time-series forecasting because random shuffling may introduce future information into earlier observations and lead to overly optimistic performance estimates.

To prevent data leakage, a new MinMaxScaler was fitted exclusively to the training partition of each fold and subsequently applied to the corresponding testing partition. The normalized data were then transformed into supervised learning sequences using the same 15-lag sliding window configuration employed in the main experiment. For each fold, the model was trained and evaluated independently, and the resulting RMSE values were recorded. The average RMSE and RMSE standard deviation across the five folds were used to summarize the prediction accuracy and consistency:

$$\overline{RMSE} = \frac{1}{K}\sum_{k=1}^{K} RMSE_k \quad (3)$$

$$SD_{RMSE} = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(RMSE_k - \overline{RMSE}\right)^2} \quad (4)$$

where $K = 5$, $RMSE_k$ is the $RMSE$ obtained in fold k, $\overline{RMSE}$ is the average $RMSE$, and SD_{RMSE} is the standard deviation of $RMSE$ across folds. A lower average $RMSE$ indicates better cross-validation accuracy, while a lower standard deviation indicates stronger prediction consistency.

3.5 Robustness Testing and Visualization

Robustness testing was conducted to evaluate how well each model responded to abnormal market conditions beyond the original testing data. In this study, robustness was assessed using three synthetic stress scenarios: extreme drop, extreme rise, and high-volatility noise. These scenarios were not intended to replicate a specific real crisis event directly but to provide controlled stress conditions for comparing the resilience of the LSTM, GRU, and Hybrid LSTM-GRU models under different types of market shocks. Let P_t denote the original Close Price at time t, and P_t' denote the modified Close Price under a synthetic stress scenario. The extreme drop scenario was generated by reducing the original Close Price using a predefined drop factor:

$$P_t' = P_t(1-\alpha) \quad (5)$$

where α represents the percentage decrease injected into the data. In this study, $\alpha = 0.05$, meaning that the original Close Price was reduced by 5% to simulate a sudden downward market movement. The extreme rise scenario was generated by increasing the original Close Price using a predefined rise factor:

$$P_t' = P_t(1+\beta) \quad (6)$$

where β represents the percentage increase injected into the data. In this study, $\beta = 0.05$, meaning that the original Close Price was increased by 5% to simulate a sudden upward market movement. The high-volatility scenario was generated by adding Gaussian noise to the original Close Price series:

$$P_t' = P_t + \epsilon_t \quad (7)$$

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

$$\epsilon_t \sim N\left(0, \left(\sigma \times SD(P)\right)^2\right) \quad (8)$$

where ϵ_t represents Gaussian noise, $SD(P)$ denotes the standard deviation of the original Close Price series, and σ represents the noise-level factor. Table 4 presents the synthetic robustness scenarios created to test how forecasting models react to unusual market conditions not covered by the original test data.

Table 4 - Synthetic Robustness Scenario Design

Scenario	Parameter	Purpose
Extreme Drop	$\alpha = 0.05$	Simulates a 5% sudden price decline
Extreme Rise	$\beta = 0.05$	Simulates a 5% sudden price increase
High Volatility	$\sigma = 0.01$	Simulates random noise equal to 1% of the standard deviation of the original Close Price series

Values of α = 0.05 and β = 0.05 were chosen to represent moderate synthetic market shocks. These shocks were large enough to change the price path but maintained the overall structure of the original time series. Similarly, σ = 0.01 was chosen to add controlled random variability without completely changing the underlying pricing pattern. These parameters aim to create similar stress conditions rather than mimic a specific past market event. The strength of each model was tested using RMSE, MAE, and R^2 under each synthetic scenario. Lower RMSE and MAE values mean smaller prediction errors under stress, while higher R^2 values show better explanatory performance. Together, these metrics offer a complementary view of the forecasting strength under controlled market disturbances. The final analysis of robustness was supported by actual-versus-predicted plots, RMSE comparison graphs, and a robustness map. The robustness map highlights the strengths and weaknesses of each model in terms of baseline prediction accuracy, cross-validation consistency, and synthetic stress scenarios, assisting in context-aware model selection under different market conditions.

4. RESULTS AND DISCUSSION

In this section, we detail the experimental outcomes for the LSTM, GRU, and Hybrid LSTM-GRU models across three evaluation phases: initial prediction accuracy, consistency in cross-validation, and resilience testing in artificially induced stress situations. Subsequently, a robustness map was employed to summarize the findings, highlighting the comparative strengths of each model across various evaluation contexts.

4.1. Baseline Model Performance

Baseline evaluation was conducted using a chronological testing set. This evaluation aimed to measure the prediction accuracy of each model under the original testing conditions before applying the synthetic stress scenarios. The evaluation metrics include RMSE, MSE, MAE, R, and R^2. Table 5 shows that all models achieved strong predictive performance on the original testing data, as indicated by R^2 values above 0.99. The GRU model produced the lowest RMSE (52.95) and the lowest MAE (28.14), indicating the best baseline test performance among the three models. The LSTM model achieved an RMSE of 53.84 and an MAE of 39.45, whereas the Hybrid LSTM-GRU model obtained an RMSE of 56.21 and an MAE of 29.85. These results indicate that the GRU slightly outperformed the other models under the original testing conditions. However, the performance differences among the three models were relatively small, suggesting that LSTM, GRU, and Hybrid LSTM-GRU were all capable of capturing the general temporal pattern of ANTM.JK stock prices. Nevertheless, baseline accuracy alone is insufficient to evaluate robustness because a model that performs well under the original testing data may not necessarily remain stable under different temporal partitions or synthetic market-stress conditions.

Table 5 - Results of the Basic Model Performance Evaluation

Model	RMSE (Test)	MSE (Test)	MAE (Test)	R (Test)	R^2 (Test)
LSTM	53.84	2898.75	39.45	0.9939	0.9930
GRU	52.95	2804.61	28.14	0.9945	0.9944
Hybrid LSTM-GRU	56.21	3160.44	29.85	0.9938	0.9937

4.2. Cross-Validation Consistency

The second evaluation stage examined the model consistency using 5-fold chronological cross-validation on the original dataset. Because stock price data are time-series observations, the cross-validation process was performed without random shuffling to preserve the temporal order. The consistency of each model was assessed using the average RMSE and standard deviation of the RMSE across the folds.

Table 6. Cross-Validation Consistency Results

Model	Average Test RMSE	Standard Deviation Test RMSE
LSTM	119.41	7.1164
GRU	130.53	7.5080
Hybrid LSTM-GRU	122.30	1.9978

As shown in Table 6, the LSTM model obtained the lowest average cross-validation RMSE of 119.41, indicating the best average prediction accuracy across the folds. The Hybrid LSTM-GRU model achieved the lowest RMSE standard deviation of 1.9978, indicating the strongest cross-fold consistency. Meanwhile, the GRU model recorded the highest average RMSE of 130.53 and the highest standard deviation of 7.5080. The findings indicate that the average prediction accuracy and prediction consistency highlight distinct facets of robustness. While LSTM demonstrated higher average accuracy across chronological folds, the Hybrid LSTM-GRU model exhibited greater stability across these folds. Consequently, the hybrid model was deemed more consistent despite not attaining the lowest average RMSE value. This differentiation is crucial because robustness should not be assessed based solely on a single metric.

4.2.1. Robustness Results under Synthetic Stress Scenarios

The third evaluation stage examined the resilience of each model to synthetic stress scenarios. Following the robustness testing procedure described in Section 3.5, three controlled stress scenarios were used: drop, rise, and noisy high-volatility. The drop scenario reduced the original Close Price by $5\% (\alpha = 0.05)$. In the rise scenario, the original Close Price is increased by $5\% (\beta = 0.05)$. In the noisy scenario, Gaussian noise with a standard deviation equal to 1% of the standard deviation of the original Close Price series$(\sigma = 0.01)$.

Table 7. Robustness Results under Synthetic Stress Scenarios

Model	Scenario	RMSE	MAE	R^2 Score
LSTM	Drop	47.7697	27.0425	0.995049

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Model	Scenario	RMSE	MAE	R^2 Score
LSTM	Rise	52.9865	29.9157	0.995014
LSTM	Noisy	51.2536	30.2506	0.994857
GRU	Drop	49.5838	29.3111	0.994666
GRU	Rise	55.0713	32.2800	0.994614
GRU	Noisy	53.1941	32.3402	0.994460
Hybrid LSTM-GRU	Drop	46.8361	27.0101	0.995241
Hybrid LSTM-GRU	Rise	51.8683	29.9564	0.995222
Hybrid LSTM-GRU	Noisy	50.1911	30.2786	0.995068

Table 7 shows that all models maintained high predictive performance under the synthetic stress scenarios, with R^2 values above 0.994. However, the Hybrid LSTM-GRU model demonstrated the strongest overall robustness across the three scenarios. In the drop scenario, the hybrid model achieved the lowest RMSE (46.8361) and the lowest MAE (27.0101). In the rise scenario, it also obtained the lowest RMSE of 51.8683 and the highest R^2 score of 0.995222. In the noisy scenario, the hybrid model again achieved the lowest RMSE of 50.1911 and the highest R^2 score of 0.995068. The LSTM model exhibited the second-best robustness performance across the stress scenarios. Compared with GRU, LSTM consistently produced lower RMSE and MAE values and higher R^2 scores in drop, rise, and noisy conditions. This indicates that LSTM was more stable than GRU when the test data were modified using controlled synthetic stress. Meanwhile, the GRU model showed slightly higher error values across all three scenarios, suggesting that its simpler structure was less robust than those of LSTM and Hybrid LSTM-GRU in this specific stress-testing setting. Overall, these findings indicate that the Hybrid LSTM-GRU model provides the most stable predictive behavior under controlled synthetic stress scenarios. This result supports the view that combining LSTM and GRU components can improve resilience when the input data are exposed to moderate abnormal market movements. However, because the stress parameters were set at a 5% drop, a 5% rise, and a 1% noise level, these results should be interpreted as robustness under controlled synthetic stress, not as evidence that the hybrid model will always dominate under all real-crisis conditions.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

4.2.2. Robustness Map

To thoroughly interpret the experimental outcomes, a robustness map was created by combining the results of baseline testing, chronological cross-validation, and synthetic stress-scenario assessments. The aim of this map is not to pinpoint a universally superior forecasting model but to highlight the relative advantages of each architecture based on various evaluation criteria. As shown in Table 8, the robustness map indicates that a model's superiority is contingent on the evaluation goals. The GRU model excelled in the baseline prediction on the original testing dataset, achieving the lowest RMSE and MAE values. The LSTM model recorded the lowest average RMSE across chronological cross-validation folds, demonstrating a strong average predictive performance across different time segments. In contrast, the Hybrid LSTM-GRU model exhibited the highest prediction consistency, as evidenced by the lowest RMSE standard deviation, and performed best under synthetic drop, rise, and noisy stress scenarios. These results suggest that no single model outperforms all evaluation dimensions. Instead, each architecture displays distinct strengths depending on the forecasting goals and market conditions. Therefore, forecasting models should not be selected based solely on a single accuracy metric. The robustness map offers a multidimensional view that aids context-aware model selection by simultaneously considering baseline accuracy, prediction consistency, and resilience under controlled market disturbances.

Table 8. Robustness Map of LSTM, GRU, and Hybrid LSTM-GRU

Evaluation Aspect	Best Model	Supporting Result	Interpretation
Baseline test accuracy	GRU	RMSE = 52.95; MAE = 28.14	Best performance on the original testing set
Average cross-validation RMSE	LSTM	Average RMSE = 119.41	Lowest average error across chronological folds
Cross-validation consistency	Hybrid LSTM-GRU	RMSE standard deviation = 1.9978	Most stable performance across folds
Drop stress scenario	Hybrid LSTM-GRU	RMSE = 46.8361; MAE = 27.0101; R^2 = 0.995241	Most robust under 5% downward stress
Rise stress scenario	Hybrid LSTM-GRU	RMSE = 51.8683; R^2 = 0.995222	Most robust under 5% upward stress
Noisy stress scenario	Hybrid LSTM-GRU	RMSE = 50.1911; R^2 = 0.995068	Most robust under Gaussian noise stress

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

4.2.3. Comparison with Previous Research

To position the findings of this study within the existing stock forecasting literature, a comparison was conducted with significant research employing machine learning, deep learning, and hybrid forecasting techniques for stock-price prediction. This comparison focuses on the evaluation perspective rather than direct numerical results, as forecasting outcomes are influenced by differences in datasets, stock characteristics, forecasting periods, and evaluation methods.

Table 9. Comparison with Previous Research

Model	Dataset	Main Findings	Robustness Evaluation
SVM + GA [8]	ANTM.JK	RMSE = 10.495	Not explicitly tested
LSTM [19]	ANTM.JK	LSTM reported as effective for stock prediction	Not explicitly tested
LSTM-RNN [23]	Bitcoin	RMSE = 0.14 (Bitcoin scale)	Not explicitly tested
Hybrid MEME-AO-LSTM [16]	Global market indices	Improved forecasting performance under non-stationary conditions	Partially addressed
Symbolic-State CRNN [17]	Multi-market dataset	Significant MSE reduction across market regimes	Focused on regime adaptation
LSTM, GRU, Hybrid LSTM-GRU	**ANTM.JK**	**GRU achieved the best baseline accuracy, LSTM obtained the lowest average cross-validation RMSE, and Hybrid LSTM-GRU demonstrated the strongest robustness under synthetic stress scenarios**	**Explicitly evaluated through chronological cross-validation consistency and synthetic stress testing**

Table 9 reveals that prior studies have predominantly focused on improving forecast accuracy by employing optimized machine learning algorithms, recurrent neural networks, or hybrid deep learning models. Although these approaches have demonstrated promising predictive capabilities, there is a relative lack of assessments of their robustness. Typically, research evaluates performance using standard error metrics such as RMSE, MAE, MSE, or R^2 within a single test setup, with limited attention to how models perform across different time segments and in atypical market conditions. This study extends previous work by introducing a framework for assessing robustness across multiple dimensions. The results suggest that the effectiveness of the model depends on

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

the evaluation objective. The GRU model provided the highest baseline prediction accuracy, the LSTM achieved the lowest average cross-validation RMSE, and the Hybrid LSTM-GRU exhibited the most consistency and robustness during periods of decline, growth, and noise. These findings suggest that focusing on robustness in evaluations can provide a more comprehensive understanding of forecasting model behavior than relying solely on accuracy metrics.

4.3. Discussion

The experimental findings indicate that the performance of the forecasting model is significantly influenced by the selected evaluation perspective. Although all three deep learning architectures demonstrated high predictive capabilities, their relative advantages differed when baseline testing, cross-validation consistency, and synthetic stress-scenario evaluation were considered. This observation implies that a single performance metric is insufficient to fully describe forecasting robustness. Among the three architectures, the GRU model excelled in the baseline prediction performance on the original testing dataset, achieving the lowest RMSE and MAE values. This could be attributed to the relatively compact gating structure of the GRU, which allows for efficient learning of temporal dependencies while minimizing the number of trainable parameters compared with more complex recurrent architectures. In the context of a univariate forecasting problem, such as predicting the ANTM.JK stock price, this simpler structure may promote effective generalization of unseen testing data without adding unnecessary complexity to the model.

Conversely, the LSTM model achieved the lowest average RMSE during chronological cross-validation. This outcome indicates that the memory cell mechanism of the LSTM is effective in capturing stable temporal dependencies across various chronological segments of the dataset. Because stock price series encompass long-term temporal information that can extend over multiple market regimes, LSTM's ability to retain historical data may contribute to its relatively strong average performance across different folds. The Hybrid LSTM-GRU model exhibited distinct behaviors. Although it did not reach the highest baseline accuracy or lowest average cross-validation RMSE, it showed the greatest prediction consistency and robustness under synthetic drop, rise, and noisy stress conditions. This finding implies that integrating the LSTM and GRU components enables the model to leverage long-term memory retention and adaptive

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

sequence representation. Consequently, the hybrid architecture appears to be less sensitive to controlled disturbances introduced into the test data and is therefore more resilient under stressed conditions.

These findings support the main argument of this study that model superiority depends on the evaluation objective rather than a single accuracy metric. If the evaluation is based exclusively on the baseline prediction accuracy, the GRU is selected as the preferred model. However, when prediction consistency and resilience under stressed conditions are considered, the Hybrid LSTM-GRU becomes a more favorable alternative. This observation underscores the significance of evaluating models from multiple dimensions for financial forecasting applications. Methodologically, this study adds to the stock forecasting literature by illustrating that robustness should be assessed through various complementary dimensions rather than relying solely on traditional error-based metrics. While earlier studies typically present forecasting performance using RMSE, MAE, MSE, or similar measures, the proposed framework combines baseline prediction accuracy, consistency in chronological cross-validation, and evaluation under synthetic stress scenarios into a single assessment. This multidimensional strategy offers a more comprehensive insight into model behavior across different market conditions, thereby broadening the scope of evaluation practices beyond conventional accuracy-centered analyses.

This study further contributes by presenting a robustness map that encapsulates the strengths of the models across various evaluation criteria. Rather than pinpointing a single superior architecture, the robustness map offers a systematic interpretation of the circumstances under which a specific model is favored. This viewpoint is more in tune with practical forecasting scenarios, where model requirements can differ based on market conditions and evaluation goals. The results indicate that distinct forecasting models may be more suitable for various applications. The GRU might be chosen when the main focus is on the baseline-forecasting accuracy of the original test data. In contrast, LSTM may be more appropriate when a consistent average performance across several temporal segments is required. The hybrid LSTM-GRU model may be the best choice when prediction consistency and resilience under challenging market conditions are crucial. Consequently, the choice of forecasting model should be determined by the intended application context rather than relying solely on a single evaluation metric.

This study has several limitations. Initially, it concentrates on a single stock, ANTM.JK, which means that the results might not be applicable to other stocks, sectors, or financial markets. Additionally, the forecasting models were constructed using only the Close Price variable, whereas incorporating other data, such as technical indicators, trading volume, macroeconomic variables, and sentiment indicators, could enhance predictive accuracy. Moreover, synthetic stress scenarios simulate controlled market disruptions but fail to fully reflect the intricacies of actual financial crises in the real world. Finally, the robustness assessment was performed using fixed stress parameters (α=0.05, β=0.05, and σ=0.01), indicating moderate rather than extreme market disturbances. Future studies could expand the proposed framework by including multidimensional multivariate analyses across various stocks and sectors and by applying robustness tests to real crisis event subperiods. Further evaluation methods, such as walk-forward validation, directional accuracy, and risk-adjusted forecasting metrics, may offer more comprehensive insights into the robustness of forecasting models in evolving financial contexts.

5. CONCLUSION

This study assessed the resilience of the LSTM, GRU, and Hybrid LSTM-GRU models in forecasting ANTM.JK stock prices amid the volatility caused by the COVID-19 pandemic. The results revealed that the best forecasting model varied based on the evaluation criteria rather than relying on a single performance measure. The GRU model delivered the most reliable baseline prediction, the LSTM model achieved the smallest average error across chronological cross-validation folds, and the Hybrid LSTM-GRU model exhibited the greatest prediction consistency and robustness in scenarios involving synthetic drops, rises, and noise. These findings suggest that models excelling under standard testing conditions may not be the most resilient when assessed in various market environments. From a methodological standpoint, this study advances stock forecasting research by presenting a multidimensional robustness evaluation framework that combines baseline prediction accuracy, chronological cross-validation consistency, and synthetic stress scenario analysis into a cohesive assessment. The proposed robustness map offers a structured interpretation of model strengths across diverse

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

evaluation goals, providing a more comprehensive view than traditional accuracy-focused evaluations.

These results indicate that the choice of forecasting model should correspond to the specific application context. The GRU may be a better option when the main focus is on achieving a high baseline prediction accuracy. However, the Hybrid LSTM-GRU model seems more appropriate when the emphasis is on maintaining prediction stability and resilience under challenging conditions. Therefore, evaluating models with a focus on robustness can offer a more thorough foundation for assessing forecasting models in dynamic financial environments. This research is confined to a single-stock case study, focusing on univariate Close Price forecasting and controlled synthetic stress scenarios. Future studies should broaden the proposed framework to include multiple stocks and market sectors, integrate multivariate predictors such as technical indicators, macroeconomic variables, and sentiment data, and assess robustness using real crisis event periods along with more comprehensive walk-forward validation methods.

REFERENCES

[1] S. Kumar and D. Ningombam, "Short-Term Forecasting of Stock Prices Using Long Short Term Memory," *Proceedings - 2018 International Conference on Information Technology, ICIT 2018*, pp. 182–186, Dec. 2018, doi: 10.1109/ICIT.2018.00046.

[2] N. Solanki and M. Jha, "A Decade of Stock Data: Predictive Modelling with Hybrid CNN-BiLSTM Technique," in *2024 IEEE Silchar Subsection Conference, SILCON 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/SILCON63976.2024.10910869.

[3] T. Sanboon, K. Keatruangkamala, and S. Jaiyen, "A Deep Learning Model for Predicting Buy and Sell Recommendations in Stock Exchange of Thailand using Long Short-Term Memory," pp. 757–760, Sep. 2019, doi: 10.1109/ccoms.2019.8821776.

[4] Y. Sharma, A. Kumar, V. Dubey, and V. Rai, "Stock Price Prediction Using LSTM," *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*, vol. 2, no. 4, pp. 1–5, Jun. 2023, doi: 10.1109/ICCCNT56998.2023.10307212.

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

[5] Y. Zeng, J. Chen, N. Jin, X. Jin, and Y. Du, "Air quality forecasting with hybrid LSTM and extended stationary wavelet transform," *Build. Environ.*, vol. 213, Apr. 2022, doi: 10.1016/j.buildenv.2022.108822.

[6] Khoirudin, P. T. Pungkasanti, and N. Wakhidah, "ANTM.JK Stock Price Prediction Using Long Short-Term Memory (LSTM) Method During COVID-19 Pandemic," *1st International Conference on Technology, Engineering, and Computing Applications: Trends in Technology Development in the Era of Society 5.0, ICTECA 2023*, pp. 1–5, 2023, doi: 10.1109/ICTECA60133.2023.10490699.

[7] Tradingview, "Aneka Tambang TBK," tradingview. Accessed: May 12, 2026. [Online]. Available: https://id.tradingview.com/symbols/IDX-ANTM/

[8] Y. Ramdhani and A. Mubarok, "Analisis Time Series Prediksi Penutupan Harga Saham Antm.Jk Dengan Algoritma SVM Model Regresi," *Jurnal Responsif: Riset Sains dan Informatika*, vol. 1, no. 1, pp. 77–82, 2019, [Online]. Available: http://ejurnal.univbsi.id/index.php/jti

[9] S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon, and K. P. Soman, "Stock price prediction using LSTM, RNN and CNN-sliding window model," in *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, IEEE, 2017, pp. 1643–1647. doi: 10.1109/ICACCI.2017.8126078.

[10] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1724–1734, Jun. 2014, doi: 10.3115/v1/d14-1179.

[11] N. K. Manaswi, "RNN and LSTM," in *Deep Learning with Applications Using Python*, Apress, 2018, pp. 115–126. doi: 10.1007/978-1-4842-3516-4_9.

[12] A. Saxena and T. R. Sukumar, "Predicting bitcoin price using lstm And Compare its predictability with arima model," *International Journal of Pure and Applied Mathematics*, vol. 119, no. 17, pp. 2591–2600, 2018, Accessed: Jul. 18, 2022. [Online]. Available: http://www.acadpubl.eu/hub/

[13] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, "LSTM network: A deep learning approach for Short-term traffic forecast," *IET Intelligent Transport Systems*, vol. 11, no. 2, pp. 68–75, Mar. 2017, doi: 10.1049/iet-its.2016.0208.

[14] R. Khaldi, A. El Afia, R. Chiheb, and S. Tabik, "What is the best RNN-cell structure to forecast each time series behavior?," *Expert Syst. Appl.*, vol. 215, Apr. 2023, doi: 10.1016/j.eswa.2022.119140.

[15] X. Zhang, C. Zhong, J. Zhang, T. Wang, and W. W. Y. Ng, "Robust recurrent neural networks for time series forecasting," *Neurocomputing*, vol. 526, pp. 143–157, Mar. 2023, doi: 10.1016/j.neucom.2023.01.037.

[16] Q. Ge, "Enhancing stock market Forecasting: A hybrid model for accurate prediction of S&P 500 and CSI 300 future prices," *Expert Syst. Appl.*, vol. 260, p. 125380, 2025, doi: 10.1016/j.eswa.2024.125380.

[17] F. K. Mirza, Ö. Pekcan, M. Hekimoğlu, and T. Baykaş, "Stock price forecasting through symbolic dynamics and state transition graphs with a convolutional recurrent neural network architecture," *Neural Comput. Appl.*, 2025, doi: 10.1007/s00521-025-11325-z.

[18] N. K. Manaswi, "Deep Learning with Applications Using Python," *Deep Learning with Applications Using Python*, 2018, doi: 10.1007/978-1-4842-3516-4.

[19] D. R. Danistya, F. Qaulifa, Y. A. Ramadani, I. Nurma Yulita, M. N. Ardisasmita, and D. Agustian, "Prediction New Cases of COVID-19 in Indonesia Using Vector Autoregression (VAR) and Long-Short Term Memory (LSTM) Methods," *2021 International Conference on Artificial Intelligence and Big Data Analytics, ICAIBDA 2021*, pp. 127–130, 2021, doi: 10.1109/ICAIBDA53487.2021.9689721.

[20] A. El Filali, E. H. Ben Lahmer, S. El Filali, M. Kasbouya, M. A. Ajouary, and S. Akantous, "Machine Learning Applications in Supply Chain Management: A Deep Learning Model Using an Optimized LSTM Network for Demand Forecasting," *International Journal of Intelligent Engineering and Systems*, vol. 15, no. 2, pp. 464–478, Apr. 2022, doi: 10.22266/ijies2022.0430.42.

[21] M. Li, Y. Zhu, Y. Shen, and M. Angelova, "Clustering-enhanced stock price prediction using deep learning," *World Wide Web*, vol. 26, no. 1, pp. 207–232, Jan. 2023, doi: 10.1007/s11280-021-01003-0.

[22] A. K. Badri, J. Heikal, Y. A. Terah, and D. R. Nurjaman, "Decision-Making Techniques using LSTM on Antam Mining Shares before and during the COVID-19 Pandemic in Indonesia," *APTISI Transactions on Management (ATM)*, vol. 6, no. 2, pp. 167–180, 2021, doi: 10.33050/atm.v6i2.1776.

[23] N. G. Ramadhan, N. A. F. Tanjung, and F. D. Adhinata, "Implementation of LSTM-RNN for Bitcoin Prediction," *Ind. Journal on Computing*, vol. 6, no. 3, pp. 17–24, 2021, doi: 10.34818/indojc.2021.6.3.592.

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

[24] "PT Aneka Tambang Tbk (ANTM.JK) Stock Historical Prices Data," Yahoo Finance. Accessed: Oct. 03, 2023. [Online]. Available: https://finance.yahoo.com/quote/ANTM.JK