Volume XI Issue 3 Year 2026 | Page 1031-1041 | ISSN: 2527-9866
Received: ***30-06-2026*** | *Revised:* ***20-07-2026*** | *Accepted:* ***15-08-2026***

Comparing LSTM, Bi-LSTM, and Transformer Across Feature Scenarios for Rupiah Exchange Rate Forecasting

Titis Chusnul Mahrom[1], Andy Prasetyo Utomo[2], Soni Adiyono[3]
[1,2,3] Muria Kudus University, Kudus, Indonesia
e-mail: 202253129@std.umk.ac.id[1], andy.prasetyo@umk.ac.id[2], soni.adiyono@umk.ac.id[3]
*Correspondence: 202253129@std.umk.ac.id

Abstract: *This study compares LSTM, Bi-LSTM, and Transformer models for one-step-ahead JISDOR forecasting across eight feature scenarios using historical JISDOR, IHSG, Brent oil prices, and inflation. The dataset comprised 1,300 chronological observations from January 2021 to May 2026. Evaluation used chronology-safe preprocessing, four expanding-window walk-forward folds, validation-based model selection, five stochastic seeds, a naïve persistence benchmark, statistical testing, and Integrated Gradients. Among the deep-learning models, LSTM under S1 achieved the lowest mean MAPE of 0.6662%, followed by Bi-LSTM at 0.6899% and Transformer at 1.8135%. However, naïve persistence achieved a lower mean MAPE of 0.2618%, and all model–scenario combinations were significantly worse after Holm-adjusted Diebold–Mariano testing. External variables did not improve forecasting accuracy over historical JISDOR alone. Integrated Gradients showed that historical JISDOR received the largest attribution in all three architectures. These results highlight the importance of temporal validation and simple benchmarks in exchange-rate forecasting.*

Keywords: *Bi-LSTM, exchange-rate forecasting, Integrated Gradients, JISDOR, walk-forward validation*

1. Introduction

The Indonesian Rupiah-United States dollar exchange rate is an important economic indicator associated with trade, investment, production costs, and financial conditions. This study focuses on the Jakarta Interbank Spot Dollar Rate (JISDOR), the Indonesian rupiah reference exchange rate against the United States dollar. The nonlinear and time-dependent characteristics of foreign-exchange data have encouraged the application of deep-learning approaches for forecasting [1], [2]. LSTM and Bi-LSTM are widely used for sequential financial data, whereas Transformer architectures provide an attention-based alternative. Previous studies have reported varying performance across recurrent and Transformer-based models, indicating that no architecture can be assumed to be universally superior [3]-[7]. Exchange-rate forecasting may also benefit from external financial and macroeconomic information. Therefore, this study evaluates IHSG, Brent oil prices, and Indonesian inflation in addition to historical JISDOR [8].

The forecasting origin was defined as the end of trading day, with the JISDOR level on the next available trading day as the target. All predictors were restricted to information available at or before the forecast origin, including inflation values aligned according to their official publication dates. This chronology-safe design was used to reduce potential information leakage. Existing studies differ substantially in forecasting targets, feature sets, architectures, and validation procedures. Limited evidence is available from controlled comparisons of LSTM, Bi-LSTM, and Transformer for JISDOR forecasting that simultaneously consider external-feature scenarios, walk-forward validation, repeated stochastic runs, persistence benchmarking, statistical testing, and model interpretation. Integrated Gradients (IG) was therefore also employed to examine feature and temporal attribution while avoiding causal interpretation [9], [10].

Accordingly, this study addresses three research questions: **RQ1**, how do LSTM, Bi-LSTM, and Transformer compare for one-step-ahead JISDOR forecasting under repeated walk-forward evaluation and persistence benchmarking? **RQ2**, do IHSG, Brent oil prices, and inflation improve forecasting accuracy beyond historical JISDOR? **RQ3**, how do the three architectures differ in their feature-level and temporal attribution patterns? The study contributes a controlled empirical comparison of three deep-learning architectures across eight feature scenarios using chronology-safe preprocessing, four walk-forward folds, five stochastic seeds, a naïve persistence benchmark, statistical testing, residual and volatility-regime diagnostics, and repeated Integrated Gradients analysis. The contribution is therefore evaluative rather than the proposal of a new forecasting architecture.

2. Literature Review

Previous studies have applied recurrent, statistical, and attention-based models to financial and exchange-rate forecasting. Sumargo and Wasito compared RNN, LSTM, and GRU for Rupiah exchange rates and found that performance varied with data availability [11]. García et al. compared LSTM and Bi-LSTM on foreign-exchange series and reported benefits from bidirectional processing under their experimental conditions [5]. Septiani et al. compared GARCH, LSTM, GRU, and CNN for USD/IDR volatility forecasting, with LSTM producing the lowest testing RMSE [12]. However, these studies did not jointly compare LSTM, Bi-LSTM, and Transformer for next-day JISDOR level forecasting.

Transformer and feature-based financial forecasting have also been investigated. Verianto and Shimbun compared Transformer and LSTM using time-series cross-validation on BBCA and USD/IDR data and showed that performance varied across temporal folds [13]. Mushliha applied CNN-BiLSTM to Indonesian Islamic-bank stock prices, while Saputra et al. showed that LSTM forecasting performance varied across different feature combinations [14], [15]. These studies indicate that both architecture and input selection can affect forecasting performance. However, previous studies differ in targets, feature sets, validation procedures, benchmarks, and interpretability. A direct comparison under the same chronology-safe preprocessing, feature scenarios, walk-forward folds, repeated stochastic runs, persistence benchmark, and explanation protocol remains limited. Table 1 summarizes these methodological differences.

Table 1. Comparison with Previous Research

Study	Target / Period	Model & Features	Validation / Horizon	Main Gap
García et al. [5]	Multiple FX pairs and BTC/USD	LSTM, Bi-LSTM; historical prices	Train-test; daily forecasting	No external economic features or XAI
Sumargo and Wasito [11]	Rupiah exchange rates	RNN, LSTM, GRU; historical rates	Multiple historical-data settings	No Transformer, walk-forward evaluation, or XAI
Septiani et al. [12]	USD/IDR volatility; 2015-2024	GARCH, LSTM, GRU, CNN	Train-test; volatility forecasting	Target was volatility, not JISDOR level
Verianto and Shimbun [13]	BBCA and USD/IDR	LSTM, Transformer; lagged features	Five-fold time-series CV	No macroeconomic-feature analysis or XAI
Mushliha [14]	Islamic-bank stocks; 2020-2024	CNN-BiLSTM; historical prices	Model evaluation	Different target; no direct architecture comparison or XAI
Saputra et al. [15]	BCA stock price	LSTM; technical and external variables	Feature-combination evaluation	No architecture comparison or model attribution

The reviewed studies reveal four relevant gaps: limited direct comparison of LSTM, Bi-LSTM, and Transformer under identical conditions; limited systematic evaluation of external variables for JISDOR; inconsistent temporal-validation and benchmarking procedures; and limited integration of forecasting evaluation with model interpretation. This study addresses these gaps through eight feature scenarios, chronology-safe walk-forward evaluation, repeated stochastic runs, naïve persistence benchmarking, statistical testing, and Integrated Gradients.

3. Methods

The study compared LSTM, Bi-LSTM, and Transformer across eight feature scenarios for one-step-ahead JISDOR forecasting. The experimental protocol included chronology-safe preprocessing, four expanding-window walk-forward folds, validation-based model selection, five stochastic seeds, naïve persistence benchmarking, statistical testing, residual diagnostics, volatility-regime analysis, and Integrated Gradients.

A. Dataset and Data Sources

The dataset contained 1,300 JISDOR trading-day observations from January 4, 2021, to May 29, 2026. JISDOR was obtained from Bank Indonesia, IHSG from Yahoo Finance, Brent oil prices from EIA/FRED, and Indonesian inflation from Statistics Indonesia. JISDOR was the forecasting target, while IHSG, Brent oil, and inflation were evaluated as external predictors.

JISDOR trading dates were used as the primary calendar. Missing IHSG and Brent observations were forward-filled using only previously available values; backward filling was not applied. Inflation was aligned according to its official publication date using a backward as-of procedure, ensuring that only information available at the forecast origin was used. After integration, the dataset contained no missing values. A 30-observation window from $t - 29$ to t was used to predict JISDOR on the next available trading day $t + 1$. Min-Max scaling was fitted exclusively on the fitting subset of each fold and then applied unchanged to validation and testing data. Predictions were inverse-transformed to IDR per USD before evaluation.

B. Chronology-Safe Data Preprocessing and Walk-Forward Partitioning

All data were standardized by date and aligned to the JISDOR trading calendar. Missing IHSG and Brent oil observations were handled using forward filling with the most recently available value; backward filling was not applied to avoid future-information leakage. Monthly inflation data were aligned using their official publication dates through a backward as-of merge, ensuring that each observation used only information available at or before the corresponding forecast origin. The final integrated dataset contained 1,300 observations with no missing values; three IHSG and 28 Brent oil observations required forward filling. A 30-observation sliding window was used for one-step-ahead forecasting. For a forecast produced after trading day, observations from $t - 29\ to\ t$ were used to predict the JISDOR value on the next available trading day $t + 1$:

$$X_t = [x_{t-29}, x_{t-28}, \dots, x_t], \quad y_t = \text{JISDOR}_{t+1}$$

Model evaluation used four expanding-window walk-forward folds. The initial fitting window contained 650 observations, followed by 130 validation and 130 testing observations, while the fitting window expanded by 130 observations in each subsequent fold. The testing periods were March 14-October 2, 2024; October 3, 2024-April 23, 2025; April 24-November 5, 2025; and November 6, 2025-May 29, 2026.

Min-Max scaling was performed independently within each fold using parameters estimated only from the fitting subset [16]. The same transformation was then applied to validation and testing data without clipping values outside $[0,1]$. Predictions were subsequently inverse-transformed to the original IDR-per-USD scale before evaluation.

C. Feature Scenarios

Eight feature scenarios were constructed to evaluate whether IHSG, Brent oil prices, and inflation improved one-step-ahead JISDOR forecasting. Historical JISDOR was included in every scenario, while the external variables were added individually and in combination, as shown in Table 2.

Table 2. Feature Scenarios

Scenario	JISDOR	IHSG	Brent oil	Inflation	Number of features
S1	✓	-	-	-	1
S2	✓	✓	-	-	2
S3	✓	-	✓	-	2
S4	✓	-	-	✓	2
S5	✓	✓	✓	-	3
S6	✓	✓	-	✓	3
S7	✓	-	✓	✓	3
S8	✓	✓	✓	✓	4

S1 represented the univariate setting, whereas S2-S8 evaluated individual and combined external variables under the same preprocessing, forecasting horizon, walk-forward folds, and evaluation protocol.

D. Model Development and Fair Configuration Selection

Three architectures were evaluated: LSTM, Bi-LSTM, and Transformer. Four candidate configurations were defined for each architecture and evaluated using anchor scenarios S1 and S8 across four walk-forward validation folds with tuning seed 42, resulting in 96 tuning runs. Testing data were excluded from configuration selection. The candidate with the lowest mean validation MAPE was selected, with RMSE and parameter count used as tie-breaking criteria. The selected configurations were LSTM_C1, BILSTM_C1, and TRANSFORMER_C4, as summarized in Table 3.

Table 3. Selected Model Architectures and Training Parameters

Parameter	LSTM	Bi-LSTM	Transformer
Selected configuration	LSTM_C1	BILSTM_C1	TRANSFORMER_C4
First recurrent layer	48 units	Bidirectional 32 units	-
Second recurrent layer	24 units	Bidirectional 16 units	-
Input projection	-	-	48 dimensions
Attention heads	-	-	4
Key dimension	-	-	12
Feed-forward dimension	-	-	96
Encoder blocks	-	-	2
Positional information	-	-	Learned positional embedding
Temporal pooling	-	-	GlobalAveragePooling1D
Dense hidden layer	24 units	24 units	32 units

Dropout	0.10	0.10	0.20
Learning rate	0.001	0.001	0.0005
Parameter count, S1-S8	17,233-17,809	19,889-20,657	41,057-41,201
Loss function	MSE	MSE	MSE
Optimizer	Adam	Adam	Adam
Maximum epochs	150	150	150
Batch size	32	32	32
Early-stopping patience	15	15	15
Training shuffle	False	False	False

The Transformer used learned positional embeddings, residual connections, layer normalization, and global average pooling. All models used Adam optimization, MSE loss, a maximum of 150 epochs, batch size 32, early-stopping patience of 15, ReduceLROnPlateau with patience 7 and factor 0.5, and shuffle=False. Five seeds (42, 52, 62, 72, and 82) were used for every model-scenario-fold combination in the final experiment.

E. Performance Evaluation and Statistical Testing

Forecasting performance was evaluated using both numerical-error and directional metrics. All numerical metrics were calculated after predictions were transformed back to the original JISDOR scale in IDR per USD. Let (y_i) denote the actual JISDOR value, ($\hat{y}_i$) the predicted value, and (n) the number of testing observations. Mean Absolute Error (MAE) was calculated as:

$$\left[MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|\right]$$

$$\left[RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}\right]$$

$$\left[MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|\right]$$

$$\left[DA = \frac{100}{n}\sum_{i=1}^{n} I\left(d_i = \hat{d}_i\right)\right]$$

where ($I(\cdot)$) is an indicator function equal to one when the predicted and actual directions are identical and zero otherwise. Macro-averaged precision and macro-averaged F1-score were also calculated for the Up and Down classes. Metrics were first calculated separately for each direction and then averaged with equal weight across the two classes. This macro-averaging procedure reduced the influence of unequal class frequencies and complemented Directional Accuracy.

A naïve persistence forecast was included as the primary numerical benchmark. The benchmark assumed that the next JISDOR level would remain equal to the most recently observed level:

$$\left[\hat{y}_{t+1}^{naive} = y_t\right]$$

While the directional benchmark assumed continuation of the most recent movement direction. Each model-scenario combination was evaluated across four walk-forward folds and five seeds, producing 20 testing runs, summarized using mean, standard deviation, and 95% t-confidence intervals. For statistical comparison, absolute errors from the five seeds were first averaged for each testing date, yielding 520 unique date-level observations per model-scenario. Two-sided Diebold-Mariano tests with the Harvey-Leybourne-Newbold correction compared each model-scenario with naïve persistence and compared architectures within each feature scenario. Holm correction was applied at $\alpha = 0.05$, while Wilcoxon signed-rank tests were used as a sensitivity analysis.

F. Residual and Volatility-Regime Analysis

Residual and volatility-regime analyses were conducted as post-hoc diagnostics using the best-performing scenario of each architecture. Because S1 produced the lowest mean testing MAPE for LSTM, Bi-LSTM, and Transformer, predictions from the five seeds were averaged for each testing date, yielding 520 unique observations per model. Residuals were defined as:

$$\left[e_t = y_t - \bar{\hat{y}}_t\right]$$

and were evaluated using mean and median residual, residual standard deviation, MAE, RMSE, MAPE, underprediction and overprediction proportions, lag-1 autocorrelation, and the Ljung-Box test at lag 10. Volatility regimes were constructed from the 20-observation trailing standard deviation of JISDOR percentage returns. To avoid target-date leakage, returns were shifted by one observation before calculating volatility. For each fold, the 33.33rd and 66.67th percentile thresholds were estimated exclusively from the fitting subset and used to classify testing observations into Low, Medium, and High volatility regimes. MAE, RMSE, and MAPE were then compared across regimes for the deep-learning models and naïve persistence.

G. Model Interpretation Using Integrated Gradients

Integrated Gradients (IG) was applied under scenario S8 because it contained all four predictors and therefore enabled cross-feature attribution. The analysis followed the same four walk-forward folds and five stochastic seeds as the forecasting experiment, resulting in 20 attribution runs per architecture. For each fold, 100 approximately evenly spaced testing samples were explained using a baseline defined as the mean S8 fitting sequence in the corresponding fold-specific scaled space. IG was approximated using 50 interpolation steps. Absolute attribution values were aggregated separately across features and temporal positions and normalized within each run. Feature- and temporal-attribution results were summarized using mean, standard deviation, and 95% t-confidence intervals across the 20 runs. A completeness check was used to assess numerical approximation quality. IG was interpreted strictly as model attribution relative to the selected baseline and not as evidence of economic causality.

4. Results and Discussion

This section presents the results obtained after applying chronology-safe preprocessing, expanding-window walk-forward evaluation, repeated stochastic runs, a naïve persistence benchmark, and validation-based model configuration selection. The analysis covers numerical forecasting accuracy, directional performance, the effect of external variables, statistical comparison of forecast errors, residual behavior, volatility-regime performance, and model interpretation using Integrated Gradients.

A. Data Preparation and Final Experimental Protocol

The final dataset contained 1,300 chronological JISDOR observations from January 4, 2021, to May 29, 2026. Three missing IHSG observations and 28 missing Brent oil observations were handled using chronology-safe forward filling, while inflation was aligned according to its official publication dates. The resulting dataset contained no missing values. The final evaluation comprised eight feature scenarios, three architectures, four walk-forward folds, and five stochastic seeds, resulting in 480 deep-learning training runs. Each run generated 130 one-step-ahead testing predictions, producing 62,400 predictions in total. All runs were completed successfully and produced finite predictions.

B. Comparative Forecasting Performance and Naïve Benchmark

Table 4 summarizes the mean MAPE and Directional Accuracy (DA) across the eight feature scenarios. Among the deep-learning models, LSTM under S1 achieved the lowest numerical error, with MAE of IDR 109.08 ± 37.46, RMSE of IDR 132.89 ± 41.21, and MAPE of 0.6662% ± 0.2373. Bi-LSTM and Transformer also achieved their lowest MAPE under S1 at 0.6899% and 1.8135%, respectively. Across all scenarios, average MAPE was 1.0477% for LSTM, 1.0984% for Bi-LSTM, and 2.5009% for Transformer.

Table 4. Final Forecasting Performance Across Model-Scenario Combinations

Scenario	Model	MAE (IDR)	RMSE (IDR)	MAPE (%)	DA (%)	Macro Precision (%)	Macro F1 (%)
S1	LSTM	109.08 ± 37.46	132.89 ± 41.21	0.6662 ± 0.2373	46.73 ± 3.00	49.62 ± 4.29	42.43 ± 5.59
S1	Bi-LSTM	113.11 ± 44.93	137.50 ± 49.26	0.6899 ± 0.2798	46.77 ± 3.78	51.00 ± 6.60	42.10 ± 6.86
S1	Transformer	298.56 ± 133.10	359.48 ± 146.52	1.8135 ± 0.8276	46.54 ± 5.59	44.51 ± 15.98	40.70 ± 8.81
S2	LSTM	165.54 ± 39.23	196.90 ± 48.71	1.0090 ± 0.2438	46.23 ± 4.66	51.56 ± 10.11	40.76 ± 8.01
S2	Bi-LSTM	191.85 ± 81.47	229.93 ± 89.23	1.1595 ± 0.4781	46.81 ± 5.49	46.15 ± 14.68	41.96 ± 9.05
S2	Transformer	392.65 ± 106.13	451.41 ± 120.77	2.3818 ± 0.6299	45.50 ± 4.89	39.90 ± 17.73	38.33 ± 8.68
S3	LSTM	138.80 ± 35.17	169.20 ± 37.62	0.8469 ± 0.2257	47.35 ± 3.72	50.54 ± 4.52	42.66 ± 6.71
S3	Bi-LSTM	165.67 ± 64.99	198.39 ± 68.31	1.0085 ± 0.4061	46.62 ± 3.31	49.78 ± 11.58	40.69 ± 6.41
S3	Transformer	332.84 ± 158.65	421.22 ± 221.06	2.0038 ± 0.9319	48.00 ± 5.94	46.82 ± 14.88	43.53 ± 9.26
S4	LSTM	165.42 ± 82.05	200.59 ± 89.44	1.0033 ± 0.4989	46.31 ± 4.19	50.27 ± 12.90	41.11 ± 7.53
S4	Bi-LSTM	162.28 ± 62.15	193.93 ± 67.54	0.9862 ± 0.3826	46.27 ± 3.81	49.63 ± 12.66	40.82 ± 7.28
S4	Transformer	433.40 ± 117.64	512.70 ± 130.71	2.6242 ± 0.7212	46.38 ± 5.36	48.37 ± 19.36	38.07 ± 7.78
S5	LSTM	183.24 ± 33.28	224.55 ± 49.91	1.1119 ± 0.1929	47.38 ± 5.17	49.13 ± 11.47	43.07 ± 8.00
S5	Bi-LSTM	207.60 ± 65.87	259.78 ± 80.01	1.2586 ± 0.4096	48.62 ± 3.63	53.61 ± 4.19	45.16 ± 5.42

S5	Transformer	428.48 ± 170.76	510.35 ± 204.67	2.5844 ± 0.9866	46.12 ± 5.27	42.72 ± 14.38	41.04 ± 9.02
S6	LSTM	162.16 ± 81.09	203.79 ± 95.27	0.9844 ± 0.5012	48.38 ± 4.79	51.69 ± 8.91	45.28 ± 7.34
S6	Bi-LSTM	174.71 ± 56.68	214.22 ± 62.24	1.0620 ± 0.3586	48.58 ± 4.49	52.58 ± 10.43	44.60 ± 7.60
S6	Transformer	467.11 ± 155.81	540.99 ± 169.59	2.8364 ± 0.9560	46.15 ± 5.32	40.68 ± 18.06	39.01 ± 9.36
S7	LSTM	173.49 ± 92.11	208.96 ± 107.33	1.0515 ± 0.5598	46.15 ± 4.95	52.86 ± 11.28	40.93 ± 8.88
S7	Bi-LSTM	194.45 ± 72.93	228.42 ± 79.46	1.1803 ± 0.4448	45.15 ± 4.62	42.63 ± 12.67	39.40 ± 8.18
S7	Transformer	475.31 ± 172.21	578.39 ± 217.40	2.8707 ± 1.0342	46.15 ± 5.00	43.19 ± 14.32	40.49 ± 8.33
S8	LSTM	281.59 ± 156.24	341.82 ± 173.63	1.7079 ± 0.9639	47.00 ± 4.56	47.16 ± 12.11	41.11 ± 7.73
S8	Bi-LSTM	238.24 ± 90.58	297.12 ± 116.42	1.4424 ± 0.5552	46.73 ± 4.20	49.92 ± 8.61	42.95 ± 6.64
S8	Transformer	479.65 ± 197.11	569.85 ± 229.82	2.8926 ± 1.1499	46.54 ± 5.66	44.81 ± 18.44	39.79 ± 9.28

The naïve persistence benchmark achieved substantially lower numerical errors, with MAE of IDR 42.65 ± 8.42, RMSE of IDR 56.21 ± 11.59, and MAPE of 0.2618% ± 0.0584. None of the 24 deep-learning model-scenario combinations achieved a lower mean MAPE than the naïve benchmark. Paired Diebold-Mariano tests confirmed that all 24 model-scenario combinations produced significantly larger absolute forecast errors than naïve persistence after Holm correction. Pairwise comparisons showed that LSTM significantly outperformed Bi-LSTM in S1, S2, S3, S5, S6, and S7; no significant difference was observed in S4, while Bi-LSTM performed better in S8. Thus, LSTM provided the strongest overall deep-learning performance, although none of the evaluated architectures outperformed persistence. The higher Transformer errors should be interpreted within the evaluated dataset, 30-observation window, selected configurations, and walk-forward protocol rather than as evidence of general architectural inferiority.

C. Effect of External Variables

Adding external variables did not improve mean forecasting accuracy over the univariate S1 scenario for any architecture. LSTM achieved its best multivariate result under S3 (MAPE 0.8469%), Bi-LSTM under S4 (0.9862%), and Transformer under S3 (2.0038%), but all remained worse than their corresponding S1 results of 0.6662%, 0.6899%, and 1.8135%, respectively. The use of all predictors in S8 further increased MAPE to 1.7079% for LSTM, 1.4424% for Bi-LSTM, and 2.8926% for Transformer. Thus, IHSG, Brent oil, and inflation did not provide incremental predictive accuracy under the evaluated one-step-ahead setting. This finding is consistent with Saputra et al. [15], who showed that adding more predictors does not necessarily produce lower forecasting error. However, the result should not be interpreted as evidence that these variables have no economic relationship with JISDOR .

D. Directional Prediction Performance

Directional performance remained limited across the evaluated configurations. The highest mean DA was obtained by Bi-LSTM under S5 at 48.62% ± 3.63%, followed by Bi-LSTM under S6 at 48.58% ± 4.49% and LSTM under S6 at 48.38% ± 4.79%. The best Transformer result was 48.00% ± 5.94% under S3. Notably, the numerically best configuration, LSTM under S1,

achieved a DA of only 46.73% ± 3.00%, showing that lower level-forecasting error did not necessarily correspond to better directional prediction.

The highest macro precision was 53.61% ± 4.19% for Bi-LSTM under S5, while the highest macro F1-score was 45.28% ± 7.34% for LSTM under S6. In comparison, the last-direction persistence benchmark achieved DA of 54.81% ± 3.40% and macro precision and F1-score of 54.03% ± 2.58%. Thus, none of the deep-learning configurations exceeded the directional persistence benchmark, indicating that next-day JISDOR direction remained difficult to predict under the evaluated models and feature scenarios.

E. Integrated Gradients Interpretation

Integrated Gradients analysis under S8 showed that historical JISDOR received the largest mean attribution in all architectures, as summarized in Table 5.

Table 5. Normalized Integrated Gradients Feature Attribution Under Scenario S8

Model	JISDOR (%)	IHSG (%)	Brent Oil (%)	Inflation (%)
LSTM	79.85 ± 12.24	10.10 ± 9.77	3.65 ± 2.44	6.40 ± 7.03
Bi-LSTM	78.86 ± 4.68	12.52 ± 4.74	3.02 ± 2.49	5.59 ± 3.64
Transformer	56.87 ± 15.05	15.65 ± 10.85	13.13 ± 8.85	14.35 ± 12.06

LSTM and Bi-LSTM assigned approximately 79.85% and 78.86% of attribution to historical JISDOR, respectively, whereas Transformer assigned 56.87% and distributed a larger proportion across the external variables. However, this greater external-variable attribution did not correspond to improved forecasting accuracy, since S1 remained the best-performing scenario for all architectures.

Temporal attribution also differed across architectures. The ten most recent observations accounted for 71.97% of LSTM and 62.60% of Bi-LSTM attribution, compared with 33.88% for Transformer, indicating stronger recency-oriented attribution in the recurrent models. Mean completeness errors were low (0.000090 for LSTM, 0.000171 for Bi-LSTM, and 0.001321 for Transformer). These attribution results describe model sensitivity relative to the selected baseline and should not be interpreted as economic causality.

F. Residual and Volatility-Regime Analysis

Residual diagnostics focused on S1, the best-performing scenario for all three architectures, using seed-averaged predictions across 520 testing dates. Figure 5 compares the actual JISDOR series with the LSTM S1 mean prediction and naïve persistence. The LSTM generally followed the overall movement but produced smoother forecasts, whereas persistence remained closer to the observed values during short-term changes.

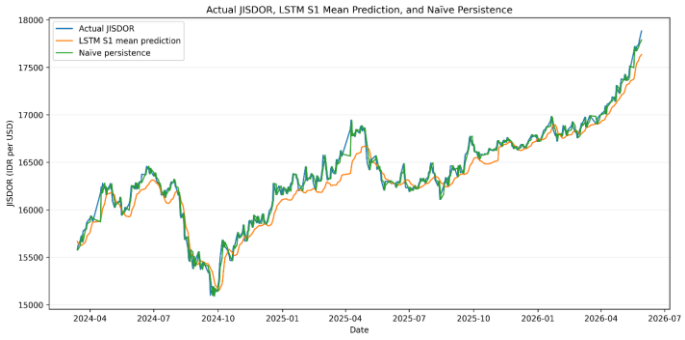

Figure 1. Actual JISDOR, LSTM S1 Mean Prediction, and Naïve Persistence

Table 6. Residual Diagnostics for the Best Scenario of Each Architecture

Model	Mean Residual (IDR)	Residual SD	Lag-1 Autocorrelation	Underprediction (%)	Ljung-Box p-value (lag 10)
LSTM	76.14	108.72	0.8474	78.46	< 0.001
Bi-LSTM	80.60	112.58	0.8574	78.46	< 0.001
Transformer	255.66	273.57	0.9749	83.85	< 0.001

All models showed positive mean residuals and substantial lag-1 autocorrelation, with significant Ljung-Box tests (), indicating systematic underprediction and remaining temporal structure in the forecast errors. Transformer exhibited the largest bias and residual dependence. Volatility-regime analysis classified the 520 testing dates into 116 Low, 150 Medium, and 254 High volatility observations using fitting-subset thresholds. Table 7 shows that errors generally increased with volatility, while naïve persistence retained the lowest MAPE in every regime.

Table 7 MAPE Across Volatility Regimes for S1 Models and Naïve Persistence

Model	Low (%)	Medium (%)	High (%)
LSTM	0.4520	0.5515	0.7799
Bi-LSTM	0.4371	0.5674	0.8137
Transformer	1.3639	1.8488	1.9083
Naïve Persistence	0.1886	0.2395	0.3084

The increase in error from low- to high-volatility periods indicates greater forecasting difficulty under stronger short-term fluctuations. However, the persistence benchmark remained superior across all three regimes, showing that its advantage was not confined to a particular volatility condition.

5. Conclusions

This study compared LSTM, Bi-LSTM, and Transformer for one-step-ahead JISDOR forecasting across eight feature scenarios using chronology-safe preprocessing, four expanding-window walk-forward folds, five stochastic seeds, and a naïve persistence benchmark. Among the deep-learning models, LSTM under S1 achieved the lowest mean MAPE of 0.6662%, followed by Bi-LSTM at 0.6899% and Transformer at 1.8135%. However, naïve persistence achieved a substantially lower mean MAPE of 0.2618%, and all 24 model-scenario combinations produced significantly larger forecast errors after Holm-adjusted Diebold-Mariano testing. External variables did not improve accuracy over S1, while the best directional accuracy of 48.62% also remained below the directional persistence benchmark of 54.81%. Residual analysis showed systematic underprediction and temporal dependence, whereas volatility-regime analysis confirmed that persistence remained superior under low-, medium-, and high-volatility conditions. Integrated Gradients further showed that historical JISDOR received the largest attribution in all architectures.

These findings indicate that, under the evaluated dataset, 30-observation window, model configurations, and one-step-ahead forecasting setting, increased architectural complexity and additional predictors did not provide incremental predictive value over simple persistence. The results should therefore not be interpreted as establishing a universally superior architecture or as evidence that external variables have no economic relationship with JISDOR. Future research may examine longer datasets, alternative input windows, return-based or multi-step targets, additional chronology-safe predictors, and alternative statistical, deep-learning, or hybrid benchmarks.

References

[1] A. Saleh and S. Al, "Analyzing Rupiah-USD Exchange Rate Dynamics : A Study with ARCH and GARCH Models," *Int. J. INFORMATICS Vis.*, vol. 8, no. November, pp. 1802-1809, 2024.

[2] M. A. Junior, P. Appiahene, O. Appiah, and C. N. Bombie, "Forex market forecasting using machine learning : Systematic Literature Review and meta - analysis," *J. Big Data*, vol. 10, no. 9, pp. 1-40, 2023, doi: 10.1186/s40537-022-00676-2.

[3] N. N. D. Ardriani, P. Sugiartawan, and G. A. Santiago, "Deep Learning Approach for USD to IDR Forecasting with LSTM," *JSIKTI J. Sist. Inf. dan Komput. Terap. Indones.*, vol. 8, no. 1, pp. 91-99, 2025.

[4] A. Poernomo, "Rupiah Exchange Rate Prediction with Long Short-Term Memory Algorithm," Syntax Literate: Jurnal Ilmiah Indonesia, vol. 10, no. 1, pp. 122-130, 2025.

[5] F. García, F. Guijarro, J. Oliver, and R. Tamoši, "Foreign Exchange Forecasting Models : LSTM and BiLSTM Comparison," *Eng. Proc.*, vol. 68, no. 1, pp. 1-7, 2024.

[6] T. Fischer, M. Sterling, and S. Lessmann, "Fx-spot predictions with state-of-the-art transformer and time embeddings," *Expert Syst. Appl.*, vol. 249, no. PB, p. 123538, 2024, doi: 10.1016/j.eswa.2024.123538.

[7] L. Zhao and W. Q. Yan, "Prediction of Currency Exchange Rate Based on Transformers," *J. Risk Financ. Manag.*, vol. 17, no. 8, pp. 1-22, 2024.

[8] S. Safi, S. Aliyu, K. S. Ibrahim, and O. I. Sanusi, "Can Oil Price Predict Exchange Rate? Empirical Evidence from Deep Learning," *Int. J. Energy Econ. Policy*, vol. 12, no. 4, pp. 482-493, 2022.

[9] W. J. · W. V. D. H. · R. M. · E. C. · R. S. · G. M. Yeo, "A comprehensive review on financial explainable AI," *Artif. Intell. Rev.*, vol. 58, no. 6, pp. 1-49, 2025.

[10] H. Jang, C. Kim, and E. Yang, "TIMING : Temporality-Aware Integrated Gradients for Time Series Explanation," *Proc. 42nd Int. Conf. Mach. Learn.*, vol. 267, pp. 26877-26895, 2025.

[11] R. Sumargo and I. Wasito, "Deep Learning for Exchange Rate Prediction Within Time Constrain," *Sink. J. dan Penelit. Tek. Inform.*, vol. 8, no. 3, pp. 1259-1271, 2024.

[12] A. V. Septani, F. M. Afendi, and A. Kurnia, "Perbandingan Metode GARCH , LSTM , GRU , dan CNN pada Peramalan Volatilitas Kurs," *Limits J. Math. Its Appl.*, vol. 22, no. 1, pp. 149-169, 2025.

[13] E. Verianto and A. F. Shimbun, "TRANSFORMER WITH LAGGED FEATURES FOR HANDLING LONG-TERM DATA DEPENDENCY IN TIME SERIES FORECASTING," *JIKO (Jurnal Inform. dan Komputer)*, vol. 7, no. 3, pp. 232-241, 2024, doi: 10.33387/jiko.v7i3.9247.

[14] Mushliha, "Implementasi CNN-BiLSTM untuk Prediksi Harga Saham Bank Syariah di Indonesia," *JAMBURA J. Math.*, vol. 6, no. 2, pp. 195-203, 2024.

[15] M. I. Saputra, E. Nurmawati, and R. Abyasa, "Predicting Stock Price Movements with Technical Fundamental , and Sentiment Analysis Using the LSTM Model," *J. Inform.*, vol. 12, no. 1, pp. 1-9, 2025.

[16] Tedyyana, A., Ratnawati, F., Syam, E., & Putra, F. P. (2022). Threat modeling in application security planning citizen service complaints. *Indonesian Journal of Electrical Engineering and Computer Science, 28*(2), 1020.