JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Attention-Augmented GRU for Stock Forecasting: A Trade-Off Between Directional Accuracy and Price Prediction Error

R. Daniel Hartanto [1, 2,*], Guruh Fajar Shidik [3], Farrikh Alzami [3], Ahmad Zainul Fanani [3], Aris Marjuni [3], and Abdul Syukur [3]

[1] Dinas Pengendalian Penduduk dan Keluarga Berencana Kota Semarang, Semarang, 50269, Jawa Tengah, Indonesia; e-mail : daniel_hartanto@semarangkota.go.id
[2] Master of Informatics Engineering Program, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Jawa Tengah, Indonesia
[3] Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Jawa Tengah, Indonesia; e-mail : guruh.fajar@research.dinus.ac.id; alzami@dsn.dinus.ac.id; a.zainul.fanani@dsn.dinus.ac.id; aris.marjuni@dsn.dinus.ac.id; abah.syukur01@dsn.dinus.ac.id
* Corresponding Author : R. Daniel Hartanto

Abstract: Attention mechanisms have been widely incorporated into recurrent neural network architectures for financial time series forecasting, with most prior work reporting improvements in price-level error metrics. This study revisits that claim through a controlled empirical comparison of four deep learning architectures on nearly two decades of Telkom Indonesia (TLKM) closing price data from the Indonesia Stock Exchange (IDX). The models evaluated are a three-layer Gated Recurrent Unit (GRU) baseline, a comparable Long Short-Term Memory (LSTM) network, a Bahdanau end-attention GRU (Attn-GRU-V2), and a multi-head self-attention GRU hybrid (Attn-GRU-V3). Each architecture is trained over 30 independent runs with distinct random seeds, and performance is reported as 95% confidence intervals derived from the t-distribution. Statistical comparisons employ the Wilcoxon signed-rank test, a nonparametric paired test appropriate given the confirmed non-normality of residuals. The main finding is a consistent trade-off: the plain GRU achieves the lowest RMSE (94.02 ± 1.22 IDR) across all 30 runs, while Attn-GRU-V2 achieves the highest directional accuracy (45.91 ± 0.09%), surpassing GRU in every independent run. Bahdanau attention weights are nearly uniform across the 30-day lookback window (coefficient of variation: 3.21%), indicating that the mechanism cannot identify selectively informative timesteps in this univariate price series. This finding is consistent with the weak-form Efficient Market Hypothesis for the Indonesian market. An ablation study reveals that a 20-day lookback window maximizes directional accuracy (47.72 ± 0.21%) for the Attn-GRU-V2 model. These results suggest that Bahdanau end-attention consistently and significantly improves directional accuracy relative to a plain GRU baseline, providing an architecturally attributable advantage for direction-based applications, even when absolute price-level error is not reduced. The directional accuracy values remaining below 50% across all models are consistent with a weak-form efficiency characterization of the Indonesian market.

Keywords: Attention mechanism; Deep learning; Directional accuracy; Efficient market hypothesis; Financial time series; Gated recurrent unit; Stock forecasting; Sustainable economic systems.

Received: March, 15th 2026
Revised: March, 30th 2026
Accepted: April, 4th 2026
Published: April, 5th 2026

1. Introduction

Stock price forecasting is a canonical problem at the intersection of financial economics and machine learning. Accurate price prediction has clear practical value for portfolio construction, risk management, and automated trading. At the same time, the Efficient Market Hypothesis (EMH) [1] posits that in weak-form efficient markets, historical price sequences do not contain systematically exploitable information, implying a theoretical upper bound on the predictability achievable from price data alone. The Indonesia Stock Exchange (IDX) represents a particularly relevant context for this investigation. As one of the largest emerging market exchanges in Southeast Asia, the IDX operates under conditions of information

asymmetry and higher return volatility compared to developed markets [2], characteristics that are theoretically associated with weak-form inefficiencies and therefore with a non-trivial, albeit bounded, predictive signal in historical price data. At the same time, the large-capitalization segment of the IDX, represented by constituents such as Telkom Indonesia (TLKM), may approach weak-form efficiency more closely than small-cap or illiquid issuers.

The central hypothesis of this study is that incorporating a Bahdanau end-attention mechanism into a Gated Recurrent Unit (GRU) network produces a systematic trade-off in forecasting behavior: the averaging property of near-uniform attention weights reduces price-level accuracy while simultaneously smoothing prediction trajectories, thereby improving directional accuracy relative to the plain GRU baseline. Recurrent neural networks (RNN), particularly Long Short-Term Memory (LSTM) [3], [4] and GRU architectures, have demonstrated the capacity to capture nonlinear temporal dependencies in financial sequences and consistently outperform classical statistical baselines on price-level error metrics [5]. The subsequent introduction of attention mechanisms into these architectures [6]–[8] has yielded further reported improvements, with attention layers theoretically enabling models to selectively weight the most informative historical timesteps.

However, much of this literature shares a methodological limitation: performance is typically evaluated over a single training run, on a single dataset, using price-level error metrics exclusively. This practice conflates model quality with random seed effects and overlooks directional prediction accuracy, a metric arguably more relevant for practical trading decisions, where the correct identification of price movement direction determines profitability independently of prediction magnitude [9]. Furthermore, despite the growth of the IDX as one of the largest emerging capital markets in Southeast Asia, research specifically evaluating deep learning architectures on IDX-listed stocks remains limited relative to studies on developed markets [2], [10], [11]. The question of whether attention mechanisms offer genuine benefits in this context has therefore not been rigorously examined.

This study addresses these gaps through the following research questions: (1) Does the incorporation of attention mechanisms into GRU networks improve forecasting performance on TLKM stock data as measured by both price-level error and directional accuracy? (2) Do attention weights reflect meaningful temporal selectivity, or do they remain near-uniform across the lookback horizon? (3) What lookback window length optimizes directional accuracy for attention-augmented GRU models? In this study, GRU is selected as the primary baseline over LSTM based on its parameter efficiency and comparable sequential modeling performance documented in the financial forecasting literature. TLKM is selected as a large-capitalization, liquid, state-owned telecommunications equity whose stable sector fundamentals reduce confounding from macro-financial shocks, providing a controlled environment for evaluating attention mechanism behavior.

The contributions of this work are as follows:

- First, a statistically rigorous comparison of four deep learning architectures is conducted over 30 independent experimental runs with 95% confidence intervals and Wilcoxon signed-rank tests. Single-run evaluations without variance reporting have been documented as the dominant practice in the financial deep learning literature [5], and the present protocol directly addresses this reproducibility gap.
- Second, an empirical accuracy-direction trade-off is identified, in which attention-augmented GRU achieves consistently higher directional accuracy than plain GRU across all independent runs, despite inferior price-level accuracy.
- Third, Bahdanau attention weights are analyzed quantitatively across the full test set, with the coefficient of variation (3.21%) providing a concrete measure of temporal selectivity. This analysis establishes a mechanistic link between near-uniform attention and the observed trade-off, constituting an empirical test of weak-form efficiency through neural attention analysis rather than conventional statistical tests.
- Fourth, a lookback ablation study identifies the window length that maximizes directional accuracy, providing actionable guidance for practitioners.

The remainder of this paper is organized as follows. Section 2 reviews related work on deep learning and attention mechanisms for stock forecasting. Section 3 describes the proposed methodology, including data preprocessing, model architectures, and evaluation protocol. Section 4 presents the experimental setup. Section 5 reports the results and provides a

detailed discussion. Finally, Section 6 concludes the paper and outlines its limitations and directions for future work.

2. Related Work

Research on deep learning for stock price forecasting has expanded substantially since the foundational work of Hochreiter and Schmidhuber [4] on LSTM networks for sequential data. Systematic reviews consistently show that recurrent architectures outperform shallow regressors and ARIMA-class models on standard price-level metrics, particularly RMSE and MAE, although the magnitude of improvement varies across datasets, evaluation protocols, and market contexts [5]. A bibliometric analysis of Scopus-indexed studies from 2014 to 2023 further confirms the rapid adoption of deep learning and hybrid architectures, particularly after 2020, with GRU and LSTM remaining dominant sequential modeling approaches [3].

The GRU architecture, introduced by Cho et al. [12] as a streamlined alternative to LSTM [4], achieves comparable performance with fewer parameters by consolidating the cell and hidden states into a unified representation governed by reset and update gates. Empirical studies confirm that GRU often matches or exceeds LSTM accuracy while converging faster, especially on moderately sized datasets. For instance, Lawi et al. [13] reported that a single-layer GRU achieved the highest accuracy among multiple GRU and LSTM variants, while Chen et al. [14] demonstrated improved performance using a reconstructed GRU dataset incorporating peer-company information. These findings support the use of GRU as a strong baseline for evaluating the incremental contribution of attention mechanisms.

Attention mechanisms represent a complementary advancement in sequential modeling by enabling models to weight temporally informative observations rather than relying solely on the final hidden state. The additive attention formulation proposed by Bahdanau et al. [15] computes alignment scores as $e_t = v^\top \tanh(W_a h_t)$, from which normalized weights α_t are obtained via softmax. This approach has been widely adopted in financial forecasting. Azman et al. [7] combined bidirectional GRU with attention to forecast global stock indices, achieving improved RMSE and MAE. Louisa et al. [6] applied a CNN-BiGRU-Attention model to Indonesian stock data, while Yang [16] demonstrated improved directional prediction using attention-augmented GRU with technical indicators. Similarly, Zhang et al. [17] reported improved performance using a CNN-BiLSTM-Attention model on multivariate OHLCV inputs, with attention weights showing temporal selectivity. Liu et al. [18] further showed that attention weight concentration improves when noise is reduced via CEEMD decomposition, highlighting the dependence of attention effectiveness on signal quality.

Multi-head self-attention, introduced in the Transformer architecture [19], models dependencies across multiple representation subspaces. Its integration with recurrent models has also shown promising results. Tiwari et al. [20] incorporated Temporal Fusion Transformer components into a CNN-LSTM framework, while B.H.C. and Jeena Jacob [8] reported RMSE improvements using a CNN-LSTM-Attention hybrid. Odion et al. [21] demonstrated the effectiveness of hybrid statistical–deep learning models, and Vanguri et al. [22] highlighted stable convergence behavior using optimized deep recurrent architectures. Thach et al. [23] further showed improved forecasting performance using Transformer-based models with multivariate inputs. These studies indicate that attention mechanisms can enhance performance, particularly in multivariate settings.

Despite these advances, several methodological limitations remain. Many studies rely on single-run evaluations, which do not account for variability due to random initialization [5]. Confidence intervals and statistical significance testing are rarely reported, limiting reproducibility. Moreover, most studies focus exclusively on price-level error metrics, while directional accuracy—arguably more relevant for trading applications—is often neglected [24]. Attention mechanisms are also seldom analyzed quantitatively, leaving unclear whether they truly learn temporally selective representations.

Research on the Indonesian market remains relatively limited. Haryono et al. [2] conducted a large-scale study across 727 IDX-listed companies, while Budiharto [25] and Sartono et al. [26] highlighted the challenges posed by structural breaks during the COVID-19 period. Other studies [10], [11], [27] show that multivariate inputs and technical indicators can improve prediction accuracy. However, attention-based models on IDX stocks are still underexplored, and existing work (e.g., Louisa et al. [6]) lacks multi-run evaluation and statistical validation. Finally, the relationship between predictability and market efficiency provides an

important theoretical perspective. Bustos [24] found that machine learning accuracy is inversely related to market efficiency across multiple markets. Butet et al. [1] further noted that emerging markets often exhibit weak-form inefficiencies due to higher volatility and information asymmetry, creating a complex environment for predictive modeling.

Despite extensive work on deep learning and attention mechanisms for stock forecasting, three key gaps remain. First, most studies rely on single-run evaluations without reporting variability or statistical significance. Second, directional accuracy is rarely treated as a primary evaluation metric. Third, attention mechanisms are seldom analyzed quantitatively to verify whether they exhibit meaningful temporal selectivity. In addition, evidence from the Indonesian market remains limited, particularly for attention-based GRU models evaluated under rigorous experimental protocols. This study addresses these gaps through a multi-run statistical evaluation, joint analysis of price-level and directional metrics, and quantitative examination of attention weights.

3. Proposed Method

Figure 1 summarizes the complete experimental pipeline.

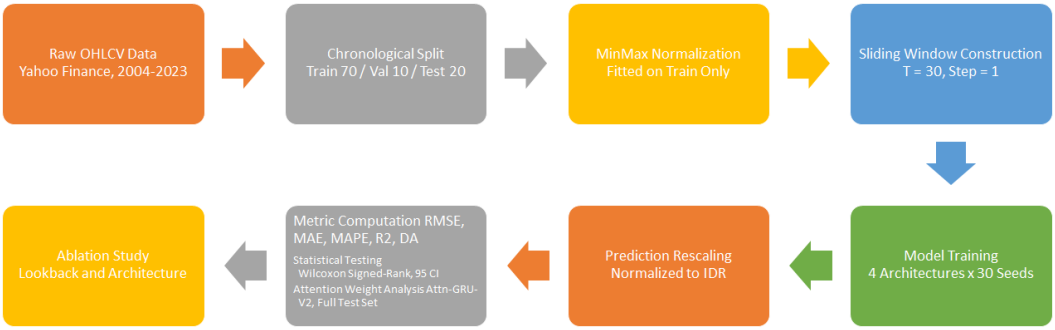

Figure 1. Experimental pipeline of the study.

Raw daily OHLCV data for TLKM are retrieved from Yahoo Finance and subjected to chronological partitioning, followed by univariate close price extraction and MinMax normalization fitted exclusively on the training partition. Sliding window sequences of length $T = 30$ are constructed and fed independently to each of the four model architectures. Each architecture is trained 30 times with distinct random seeds, and the resulting predictions are rescaled to IDR before metric computation. Statistical comparisons are performed using the Wilcoxon signed-rank test over the 30-run distributions, and attention weight analysis is conducted on the Attn-GRU-V2 model using the full test set.

3.1. Exploratory Data Analysis

The dataset contains 4,768 daily observations of TLKM closing prices spanning September 2004 to December 2023. TLKM is the largest telecommunications company in Indonesia, a state-owned enterprise with a dominant position in fixed-line, mobile, and broadband markets. As a constituent of the IDX Composite and one of the highest-capitalization equities on the IDX, TLKM exhibits relatively stable long-run fundamentals compared to sectors with direct exposure to monetary policy cycles, such as banking.

This characteristic makes TLKM a suitable subject for studying the intrinsic temporal modeling behavior of attention mechanisms in a univariate price setting, without strong confounding effects from sector-specific macroeconomic shocks. The selection is further supported by its relevance during structural regime changes, including the COVID-19 market disruption that falls within the test partition of this study.

Figure 2 presents the exploratory data analysis of the TLKM series. The close price exhibits a sustained upward trend from 2004 to approximately 2017, followed by a prolonged decline and partial recovery through 2023. Daily trading volume shows no systematic trend but contains several high-volume episodes coinciding with major market events. Daily log returns display volatility clustering, with the largest shocks concentrated around the 2008 global financial crisis and the 2020 COVID-19 onset. The log return series has a mean of

0.052%, a standard deviation of 1.95%, skewness of 0.376, and excess kurtosis of 4.23, confirming a heavy-tailed distribution typical of financial return data.

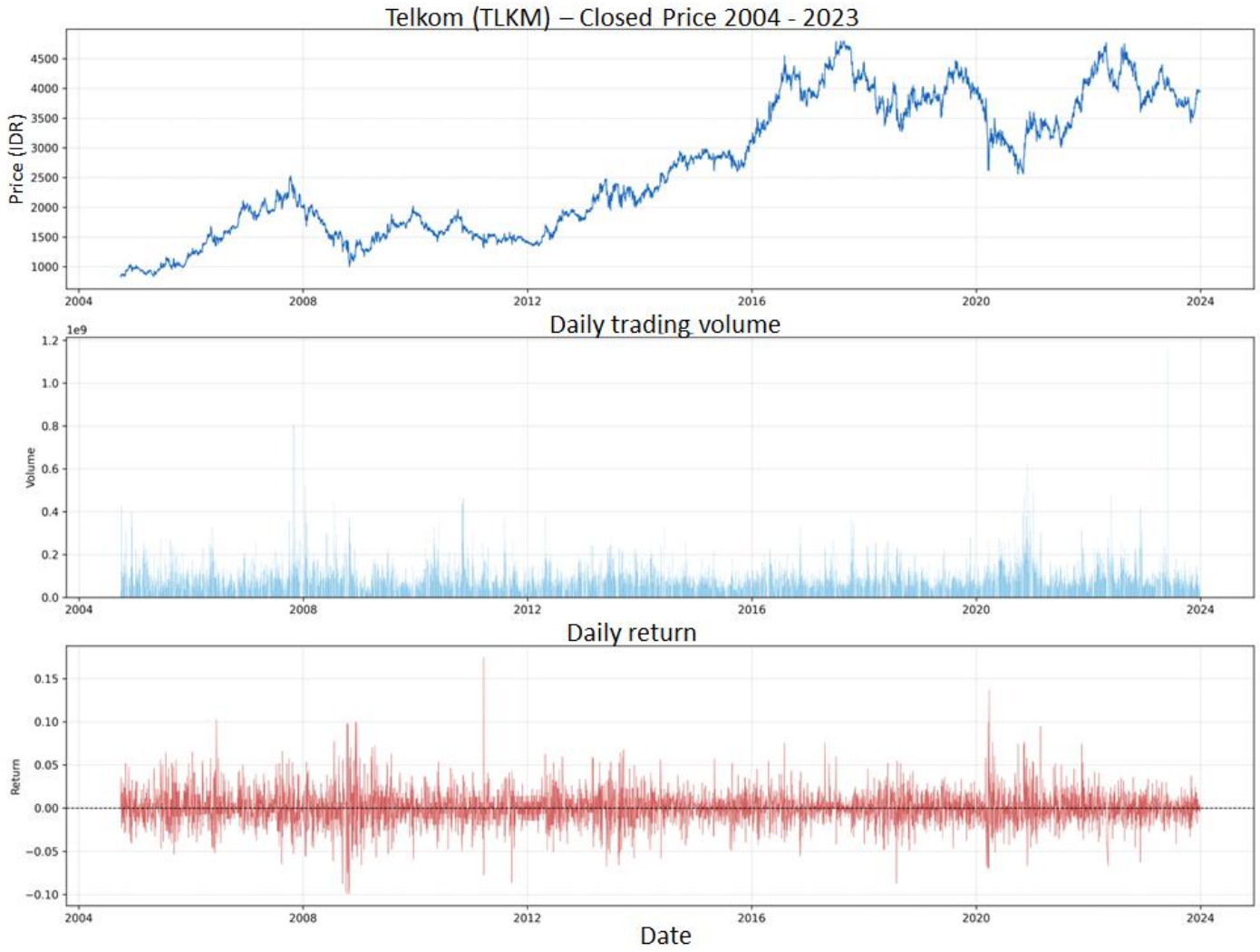

Figure 2. Exploratory data analysis of TLKM (2004–2023): close price trajectory (top), daily trading volume (middle), and daily log returns (bottom).

3.2. Dataset and Preprocessing

The dataset comprises daily adjusted closing prices of TLKM, a large-cap telecommunications company and constituent of the IDX Composite index, covering the period from 28 September 2004 to 29 December 2023. Data were obtained from Yahoo Finance. After removing non-trading days, the final dataset contains 4,768 observations spanning approximately 19 years. Summary statistics for all available OHLCV fields are reported in Table 1. Although all six fields (Open, High, Low, Close, Adjusted Close, and Volume) are available and reported for descriptive completeness, only the close price sequence is used as model input. This univariate formulation isolates the contribution of temporal modeling from multivariate feature engineering.

The daily log return is defined as $r_t = \ln(P_t/P_{t-1})$, computed over the full series, and exhibits a mean of 0.052%, a standard deviation of 1.95%, skewness of 0.376, and excess kurtosis of 4.23. The positive excess kurtosis confirms a heavy-tailed distribution typical of financial returns, motivating the use of nonparametric statistical tests rather than normality-assuming paired t-tests. Only the close price is used as input, which also facilitates clearer interpretation of attention weights, since multivariate inputs would confound temporal and feature-level selectivity.

Table 1. Descriptive statistics of TLKM OHLCV data (4,768 trading days).

Statistic	Adj. Close	Close	High	Low	Open	Volume ($\times 10^6$)
Mean	1984.35	2657.17	2687.14	2627.25	2657.16	100.37
Std	1080.41	1126.69	1135.89	1117.56	1126.82	65.41
Min	426.27	825.00	825.00	810.00	820.00	0.00
25%	1005.58	1600.00	1620.00	1580.00	1600.00	60.84
Median	1730.86	2465.00	2497.50	2430.00	2465.00	86.23
75%	2985.83	3780.00	3820.00	3740.00	3780.00	122.35
Max	4295.70	4800.00	4850.00	4780.00	4850.00	1155.86

The dataset is partitioned chronologically into three non-overlapping splits to prevent temporal leakage, as shown in Table 2. A MinMaxScaler is fitted exclusively on the training set and applied to validation and test sets.

Table 2. Chronological train/validation/test data partition.

Split	Period	Days	Proportion
Train	28 Sep 2004 – 26 Mar 2018	3,337	70%
Validation	27 Mar 2018 – 28 Jan 2020	476	10%
Test	29 Jan 2020 – 29 Dec 2023	955	20%

Sequences are constructed using a sliding window of length $T = 30$. Each sample (X, y) consists of $X \in \mathbb{R}^{30\times1}$ input sequence of 30 normalized prices, y is input sequence of 30 normalized prices $(t + 1)$. This corresponds to a one-step-ahead forecasting task for next-day price prediction. The intended application is short-term forecasting for daily trading signals, where directional movement is the primary decision variable.

The test period (January 2020–December 2023) includes the COVID-19 market shock and subsequent recovery, representing a challenging out-of-sample evaluation with structural regime changes. The data split is illustrated in Figure 3.

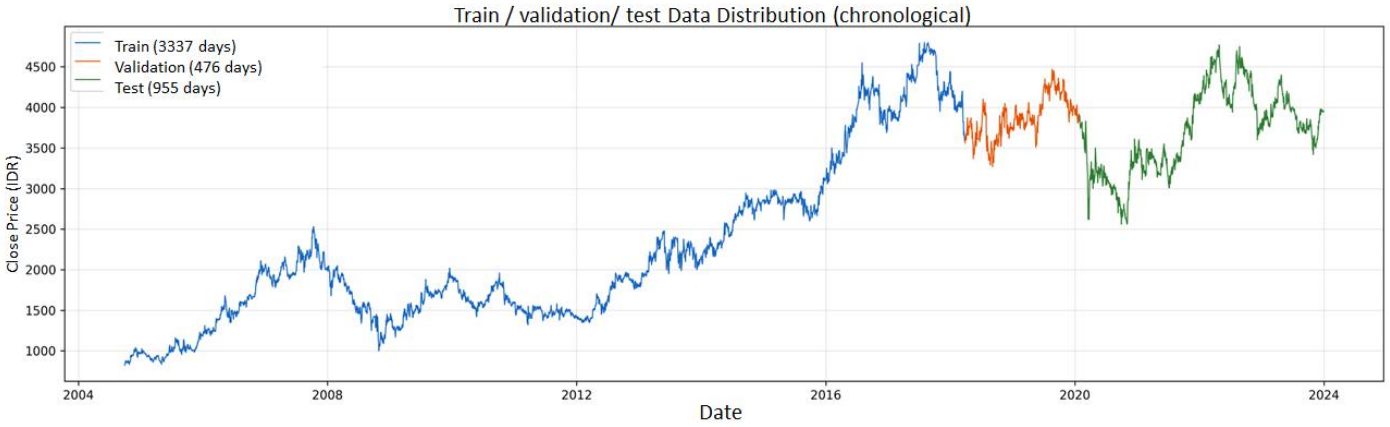

Figure 3. Chronological partitioning of the TLKM price series into training, validation, and test sets.

3.3. Model Architectures

The designations V2 and V3 reflect the internal development versioning of the proposed variants: V2 denotes the Bahdanau end-attention formulation and V3 denotes the multi-head self-attention hybrid, both developed and evaluated as original architectural variants in this study. Four architectures are evaluated. Two serve as baselines, and two incorporate attention mechanisms as proposed variants.

1. GRU Baseline: The first baseline is a three-layer stacked GRU network. Hidden unit counts follow a contracting schedule (128, 64, 32) to produce a progressive repre-senta-tion compression. Each layer is followed by a dropout layer with rate 0.2. The final GRU

layer returns only the terminal hidden state, which ispassed directly to a linear output unit

$$\hat{y} == W_o\, h_T^{(3)} + b_o \tag{1}$$

where $h_T^{(3)} \in \mathbb{R}^{32}$ is the terminal hidden state of the third GRU layer and $T = 30$ is the sequence length

2. LSTM Baseline: The LSTM baseline [4] uses an identical topological structure (128-64-32 units, dropout 0.2 after each layer) with LSTM cells replacing GRU cells. This comparison isolates the effect of cell type independently of architectural complexity.
3. Attn-GRU-V2 (Bahdanau End-Attention): The first proposed variant applies Bahdanau additive attention [7], [15] at the end of a two-layer GRU stack (128, 64 units). Both GRU layers return the full hidden state sequence $(h_t^{(2)})_{t=1}^{T}$, $h_t^{(2)} \in \mathbb{R}^{64}$. Attention scores are computed as:

$$e_t = v^{\top} \tanh\left(W_a\, h_t^{(2)}\right), \quad \alpha_t = \frac{\exp(e_t)}{\sum_{j=1}^{T} \exp(e_j)}, \quad c = \sum_{t=1}^{T} \alpha_t\, h_t^{(2)} \tag{2}$$

where $W_a \in \mathbb{R}^{64\times64}$ and $v \in \mathbb{R}^{64}$ are learned parameters, and $c \in \mathbb{R}^{64}$ is the context vector. This context vector is then passed through a dense hidden layer (32 units, ReLU) with dropout 0.2 before the linear output unit. The key architectural constraint where attention operating at the terminal position with no subsequent recurrent layer overwriting the context is strictly observed, in contrast to architectures that insert attention between intermediate recurrent layers.
4. Attn-GRU-V3 (Multi-Head Self-Attention Hybrid): The second proposed variant embeds a Transformer-style encoder block [19], [28] within a GRU-to-GRU pipeline. A single GRU layer (64 units, return sequences) produces the input sequence to a multi-head self-attention layer with 4 heads and key dimension 16. The attention output is combined with the GRU sequence via a residual connection and layer normalization. A position-wise feed-forward network (Dense 128 → Dense 64, ReLU) with a second residual connection follows. The resulting sequence is then processed by a final GRU layer (32 units, no return sequences), followed by a linear output unit. Dropout rates of 0.1 are applied after the initial GRU and after attention, with 0.2 after the final GRU.

3.4. Training Configuration

All four models share the same optimizer (Adam with initial learning rate 0.001), loss function (mean squared error), and regularization callbacks: early stopping with patience 10 (monitoring validation loss), and learning rate reduction on plateau with patience 5 and factor 0.5. The maximum epoch count is 100 and batch size is 32.

3.5. Evaluation Metrics

Five metrics are computed on the rescaled (IDR) predictions and targets:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{3}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{4}$$

$$\text{MAPE} = \frac{100}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{|y_i|} \tag{5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{6}$$

$$\text{DA} = \frac{100}{N}\sum_{i=1}^{N} 1\left[\text{sgn}(\Delta y_i) = \text{sgn}(\Delta \hat{y}_i)\right] \tag{7}$$

where $\Delta y_i = y_i - y_{i-1}$ and $\Delta \hat{y}_i = \hat{y}_i - \widehat{y_{i-1}}$. RMSE, MAE, and MAPE are price-level error metrics (lower is better), R^2 measures variance explained (higher is better), and DA measures the proportion of correctly predicted price movement directions (higher is better; 50% represents random guessing).

3.6. Statistical Testing Protocol

Each model is trained and evaluated 30 times using distinct random seeds $s_k = 42 + k$ for $k \in \{0, 1, \ldots, 29\}$. Each seed controls Python's random module, NumPy, TensorFlow's global random state, the PYTHONHASHSEED environment variable, and the TF_DETERMINISTIC_OPS flag to ensure full reproducibility within each run. The 95% confidence interval for a metric M across the 30 runs is computed as:

$$\text{CI} = \bar{M} \pm t_{0.975,\,29} \cdot \frac{s_M}{\sqrt{30}} \tag{8}$$

where $\bar{M}$ and s_M are the sample mean and standard deviation, and $t_{0.975,29} \approx 2.045$.

Pairwise comparisons employ the Wilcoxon signed-rank test, which is nonparametric and does not assume normally distributed differences. For price-level metrics (RMSE, MAE, MAPE), the null hypothesis is H_0: the competing model and GRU produce equal metric distributions, and the one-tailed alternative is H_1: GRU produces lower values. For directional accuracy, the null hypothesis is H_0: the competing model and GRU produce equal DA distributions, and the one-tailed alternative is H_1: the competing model produces higher DA. Each test is conducted at $\alpha = 0.05$.

The Shapiro-Wilk test on GRU residuals yields $W = 0.964$, $p < 0.001$, and the Kolmogorov-Smirnov test yields $D = 0.055$, $p = 0.007$, both rejecting normality at the 5% level and confirming the suitability of nonparametric comparison. All significance tests use a one-tailed alternative at $\alpha = 0.05$.

4. Experimental Setup

The complete experimental pipeline is implemented in Python 3.10 using TensorFlow 2.x/Keras for model construction and training, scikit-learn for preprocessing and statistical utilities, and SciPy for Wilcoxon signed-rank tests. All computations are performed on a Kaggle compute instance equipped with dual NVIDIA T4 GPUs. The sequence construction, model training, and evaluation pipelines are encapsulated in a reproducible Jupyter notebook that generates all reported figures and tables directly from raw data, with raw per-run results preserved in structured CSV files to facilitate independent verification.

The 30-run protocol produces 120 evaluation records (4 models × 30 seeds). Each seed controls Python's random module, NumPy, TensorFlow global state, PYTHONHASHSEED, and TF_DETERMINISTIC_OPS to ensure full within-run reproducibility. Sliding windows are constructed with a step size of one trading day, and no overlap is permitted between the final window of the training partition and the first window of the validation partition, consistent with the strict temporal isolation described in Section 3.2. For each of the 120 records, five scalar metrics are computed on test-set predictions rescaled from the normalized space to the original IDR scale. The total computational budget across all 120 runs is approximately three GPU-hours.

Key architectural and training hyperparameters are consolidated in Tables 3 and 4 for reproducibility. These parameters are held constant across all 30 seeds for each model; only the random seed varies between runs.

Table 3. Architectural configuration of the four models.

Parameter	GRU	LSTM	Attn-GRU-V2	Attn-GRU-V3
GRU/LSTM layers	3	3	2	2
Hidden units (per layer)	128,64,32	128,64,32	128,64	64,32
Dropout rate	0.2	0.2	0.2	0.1/0.2
Attention type	None	None	Bahdanau additive	Multi-head self
Attention heads / key dim	-	-	-	4/16
Dense hidden layer	None	None	32 units, ReLU	FFN 128-64, ReLU
Residual + LayerNorm	No	No	No	Yes

Table 4. Training configuration shared across all models.

Parameter	Value
Optimizer	Adam
Initial learning rate	0.001
Loss function	Mean Squared Error
Batch size	32
Maximum epochs	100
Early stopping patience	10 (monitor: val loss)
LR reduction patience	5 (factor: 0.5)
Number of runs	30
Seeds	42, 43, ..., 71

5. Results and Discussion

5.1. Main Performance Comparison

Table 5 reports the mean and 95% confidence interval for each metric across 30 independent runs. The GRU baseline achieves the lowest RMSE (94.02 ± 1.22 IDR), lowest MAE (68.73 ± 0.90 IDR), lowest MAPE (1.90 ± 0.03%), and highest R^2 (0.9658 ± 0.0009). Attn-GRU-V2 achieves the highest directional accuracy (45.91 ± 0.09%) while also exhibiting the highest price-level error among the four models (RMSE: 182.08 ± 0.88 IDR). Attn-GRU-V3 occupies an intermediate position: its RMSE (131.85 ± 3.96 IDR) and R^2 (0.9325 ± 0.0041) are better than Attn-GRU-V2 but worse than GRU, while its DA (42.84 ± 0.21%) exceeds GRU but does not reach the level of V2.

Table 5. Model performance over 30 independent runs (mean ± 95% CI).

Model	RMSE (IDR)	MAE (IDR)	MAPE (%)	R^2	DA (%)
GRU	**94.02 ± 1.22**	**68.73 ± 0.90**	**1.90 ± 0.03**	**0.9658 ± 0.0009**	41.71 ± 0.23
LSTM	115.78 ± 2.36	83.86 ± 1.63	2.33 ± 0.05	0.9481 ± 0.0021	43.93 ± 0.15
Attn-GRU-V2	182.08 ± 0.88	140.82 ± 0.71	3.90 ± 0.02	0.8720 ± 0.0012	**45.91 ± 0.09**
Attn-GRU-V3	131.85 ± 3.96	97.99 ± 3.37	2.68 ± 0.09	0.9325 ± 0.0041	42.84 ± 0.21

A key observation is the inverse relationship between price-level accuracy and directional accuracy across models. The model with the best RMSE (GRU) exhibits the lowest DA, while the model with the worst RMSE (Attn-GRU-V2) achieves the highest DA. This trade-off is not attributable to random variation: in 30 out of 30 paired runs, Attn-GRU-V2 achieves higher DA than GRU, and in 27 out of 30 runs, Attn-GRU-V3 also exceeds GRU in DA. Importantly, different metrics provide complementary perspectives on model performance. RMSE and MAE emphasize absolute price deviation, favoring models that closely track price levels, while DA captures directional correctness, which is more directly aligned with trading decisions. The observed divergence between these metrics indicates that improving directional signals may come at the cost of reduced price-level precision.

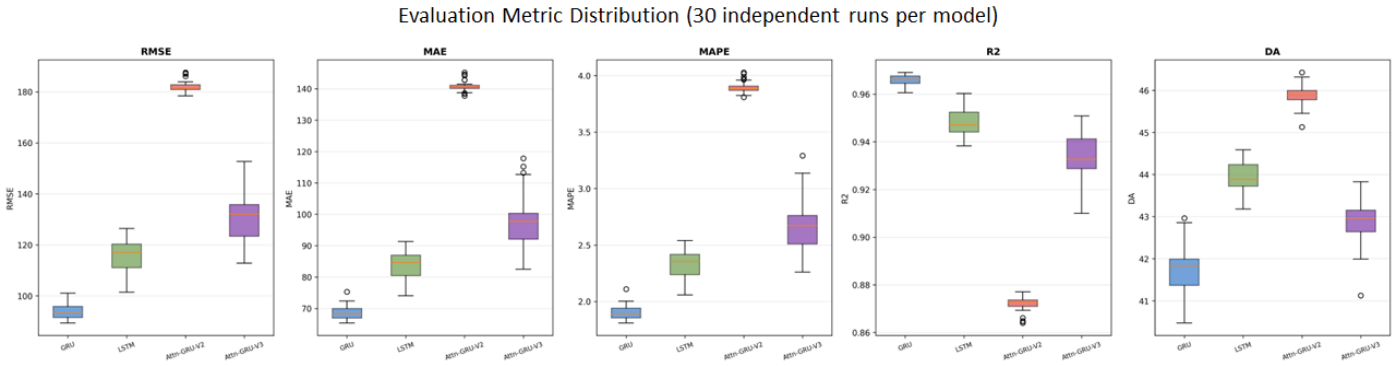

Figure 4. Distribution of evaluation metrics across 30 independent runs for each model. Each box represents the interquartile range, and whiskers extend to 1.5×IQR.

The boxplot distributions confirm that variability across runs is smallest for Attn-GRU-V2 on RMSE (coefficient of variation: 1.30%) and largest for Attn-GRU-V3 (CV: 8.04%), indicating that the multi-head self-attention model is more sensitive to random initialization.

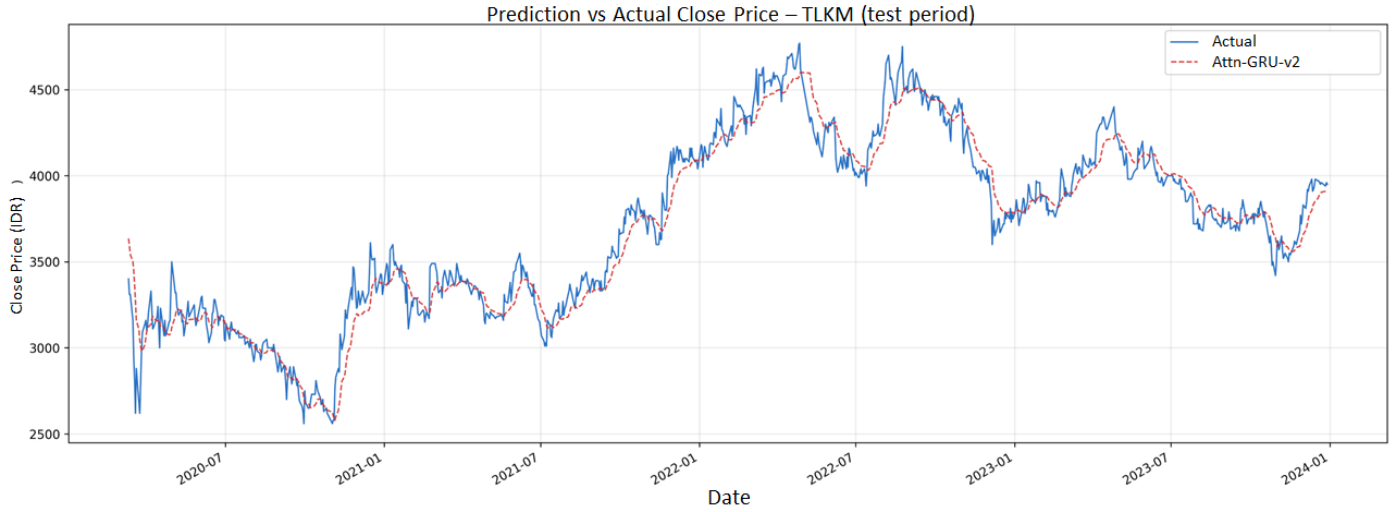

Figure 5. Predicted versus actual TLKM close prices over the full test period (955 days) for Attn-GRU-V2 (seed 42).

Visual inspection shows that all models capture the general trend. At local turning points, Attn-GRU-V2 more frequently predicts the correct direction, even when its magnitude deviates from the true price. This behavior is consistent with its higher DA but larger RMSE. This pattern suggests that the attention mechanism introduces a smoothing effect on predictions, reducing sensitivity to short-term fluctuations while preserving directional movement. As a result, the model sacrifices pointwise accuracy but gains robustness in capturing trend direction, which is critical in decision-oriented forecasting tasks.

5.2. Statistical Significance Test

Table 6 reports Wilcoxon signed-rank test results for pairwise comparisons against the GRU baseline at significance level $\alpha = 0.05$. For RMSE, MAE, and MAPE, all competing models yield $W = 465$ (maximum for n=30) and $p < 0.001$, confirming that GRU significantly outperforms other models on price-level metrics. The DA comparison for Attn-GRU-V2 vs GRU requires clarification. The statistic $W = 0$ indicates that all 30 paired differences are strictly positive, yielding an exact one-tailed p-value of $1/2^{30} \approx 9.3 \times 10^{-10}$. This confirms that the directional advantage of Attn-GRU-V2 is both consistent and statistically significant.

Table 6. Wilcoxon signed-rank test results (n = 30, α = 0.05, one-tailed)

Comparison	Metric	W	p-value	Significant	Interpretation
LSTM vs GRU	RMSE	465.0	< 0.001	Yes	GRU significantly better
LSTM vs GRU	MAE	465.0	< 0.001	Yes	GRU significantly better
LSTM vs GRU	MAPE	465.0	< 0.001	Yes	GRU significantly better
LSTM vs GRU	DA	0.0	1.0	No	No significant difference
Attn-GRU-V2 vs GRU	RMSE	465.0	< 0.001	Yes	GRU significantly better
Attn-GRU-V2 vs GRU	MAE	465.0	< 0.001	Yes	GRU significantly better
Attn-GRU-V2 vs GRU	MAPE	465.0	< 0.001	Yes	GRU significantly better
Attn-GRU-V2 vs GRU	DA	0.0	< 0.001	Yes	V2 significantly better
Attn-GRU-V3 vs GRU	RMSE	465.0	< 0.001	Yes	GRU significantly better
Attn-GRU-V3 vs GRU	MAE	465.0	< 0.001	Yes	GRU significantly better
Attn-GRU-V3 vs GRU	MAPE	465.0	< 0.001	Yes	GRU significantly better
Attn-GRU-V3 vs GRU	DA	17.5	1.0	No	No significant difference

This result reinforces that the observed improvement in directional accuracy is not incidental but is structurally induced by the attention mechanism. In contrast, Attn-GRU-V3 does not exhibit a statistically significant improvement in DA, indicating that the directional benefit is architecture-specific rather than a general property of attention-based models.

5.3. Attention Weight Analysis

Figure 6 presents the mean Bahdanau attention weights $\bar{\alpha}_t$, averaged over the entire test set for the Attn-GRU-V2 model, along with a heatmap of attention weight patterns across 50 randomly selected test samples.

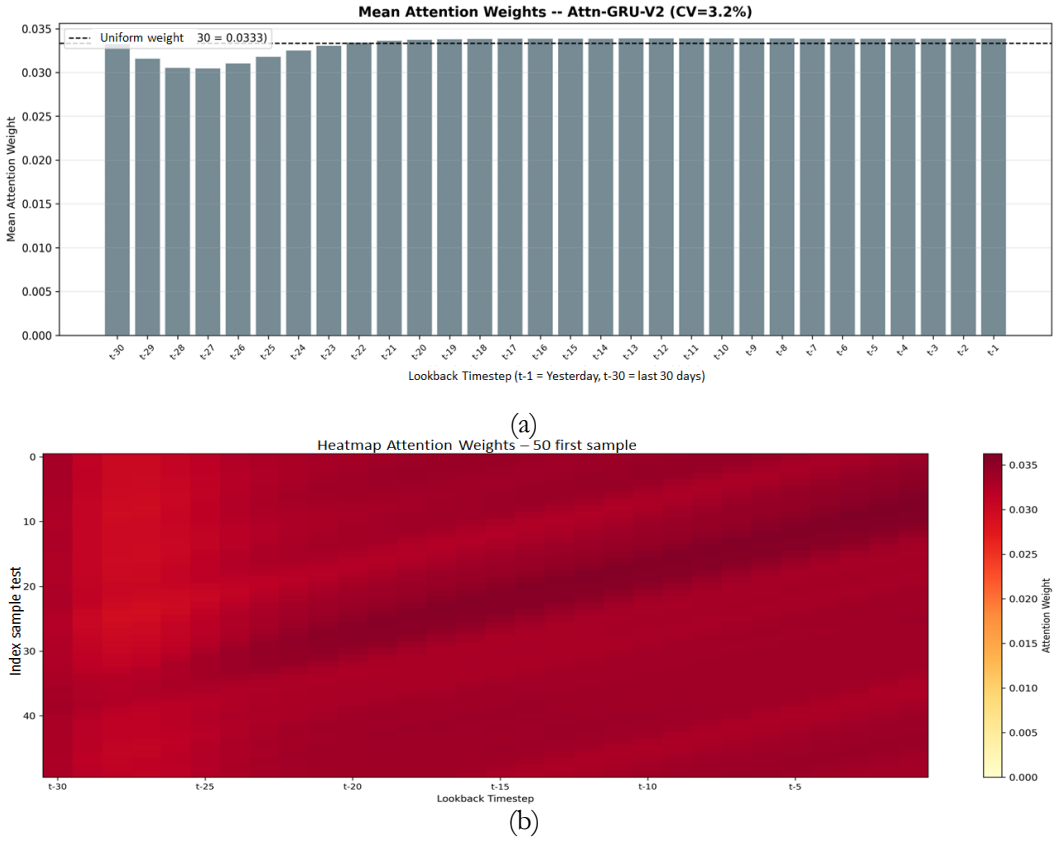

Figure 6. Bahdanau attention weight distributions for Attn-GRU-V2. (a) Mean attention weights $\bar{\alpha}_t$ across the 30-day lookback window, with standard deviation bands. Timestep $(t-1)$ denotes the observation immediately preceding the prediction date (most recent), while $(t-30)$ denotes the oldest observation in the window. (b) Attention weight heatmap for 50 test samples.

The mean weights across the 30-day lookback window are remarkably uniform. The most recent timestep $(t-1)$ receives a mean weight of 0.03392, while the most distant timestep $(t-30)$ receives 0.03355, a ratio of approximately 1.11×. The coefficient of variation across all 30 timestep mean weights is 3.21%, indicating near-complete uniformity. A mild gradient is observable: weights are slightly elevated for mid-range timesteps ($t-9$ to $t-13$) and slightly reduced for the most distant observations ($t-25$ to $t-30$), but deviations from uniform weighting $(1/30 \approx 0.0333)$ remain negligible. Standard deviations per timestep range from 0.000315 to 0.000524, with slightly higher variance at the most recent timestep, suggesting that while the model is generally uniform, it occasionally concentrates attention on recent observations for individual predictions. The heatmap confirms that this uniformity is consistent across individual test samples: no clear clusters of samples exhibit selective attention toward specific temporal regions.

The near-uniform weight distribution has a direct mechanistic interpretation under weak-form EMH. In such a setting, historical price sequences do not contain temporally localized windows of systematically higher predictive information. Consequently, the attention mechanism finds no basis for assigning higher weights to specific timesteps, and optimization converges to a near-uniform weighting. The coefficient of variation (3.21%) provides a quantitative measure of this lack of temporal selectivity. This result contrasts with findings by Liu et al. [18], who observed concentrated attention weights in GRU-attention models after CEEMD-based noise decomposition of intraday series, suggesting that attention selectivity is sensitive to the signal-to-noise ratio. In the present univariate daily close price setting, the signal-to-noise characteristics are consistent with weak-form efficiency, leading the attention mechanism to approximate a simple averaging operation over hidden states. This behavior also explains the performance trade-off observed earlier: a near-uniform attention mechanism effectively smooths temporal information, which reduces sensitivity to local noise and improves directional consistency, but at the cost of diminished price-level precision.

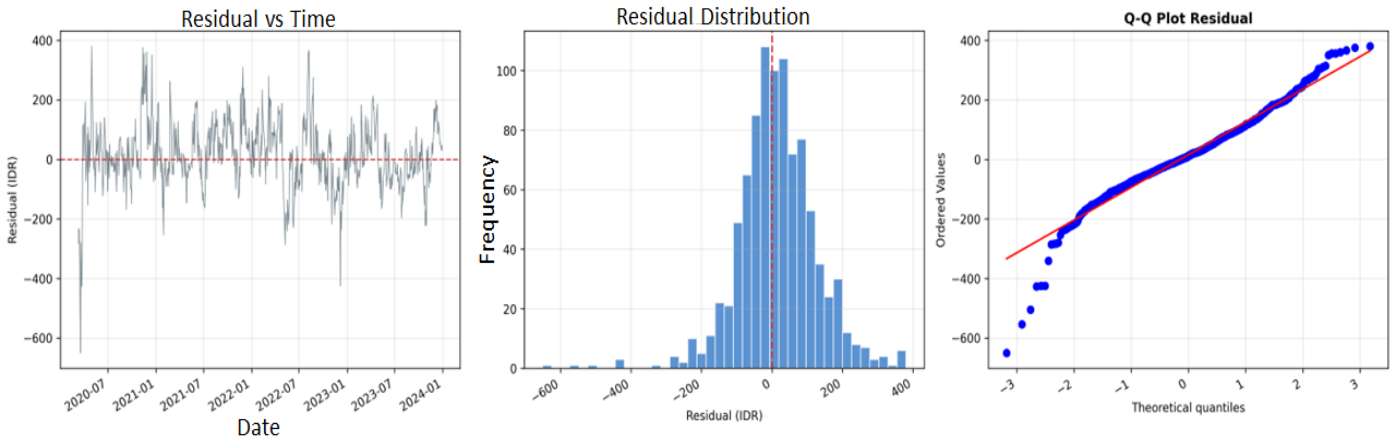

Figure 7. Residual analysis for the GRU baseline (seed 42): residuals over time (left), histogram with normal overlay (center), and Q–Q plot (right).

Residual analysis for the GRU model is shown in Figure 7. The results confirm a non-normal error distribution (Shapiro–Wilk $W = 0.964$, $p < 0.001$; Kolmogorov-Smirnov $D = 0.055$, $p = 0.007$), with heavier tails than a Gaussian distribution. This is consistent with the known properties of financial return residuals and supports the use of nonparametric statistical testing throughout this study. The heavy-tailed distribution implies that prediction errors are not uniformly distributed around zero but occasionally exhibit large deviations corresponding to market shocks. This behavior reflects volatility clustering: periods of low volatility produce small residuals, while events such as the COVID-19 onset generate extreme residuals that disproportionately affect RMSE relative to MAE. This distinction highlights that RMSE is more sensitive to rare extreme errors, whereas MAE better reflects typical model performance under normal market conditions, reinforcing the importance of evaluating multiple complementary metrics.

5.4. Ablation Study

5.4.1. Effect of Lookback Window Length

Table 7 and Figure 8 report the performance of Attn-GRU-V2 across five lookback window lengths: 10, 20, 30, 50, and 60 trading days. Each configuration is evaluated over 30 independent runs under identical training conditions.

Table 7. Ablation study: effect of lookback window length on Attn-GRU-V2 performance (mean ± 95% CI over 30 runs)

Lookback	RMSE (IDR)	MAE (IDR)	MAPE (%)	R^2	DA (%)
10	126.82 ± 1.81	92.41 ± 1.56	2.57 ± 0.05	0.9060 ± 0.0019	44.09 ± 0.15
20	155.25 ± 1.60	115.81 ± 1.05	3.20 ± 0.03	0.9060 ± 0.0019	**47.72 ± 0.21**
30	182.83 ± 1.65	141.49 ± 1.33	3.91 ± 0.04	0.8709 ± 0.0023	45.95 ± 0.14
50	206.80 ± 2.25	171.78 ± 1.82	4.67 ± 0.05	0.8317 ± 0.0037	45.69 ± 0.24
60	200.95 ± 27.47	166.96 ± 24.75	4.53 ± 0.68	0.8351 ± 0.0347	44.88 ± 0.91

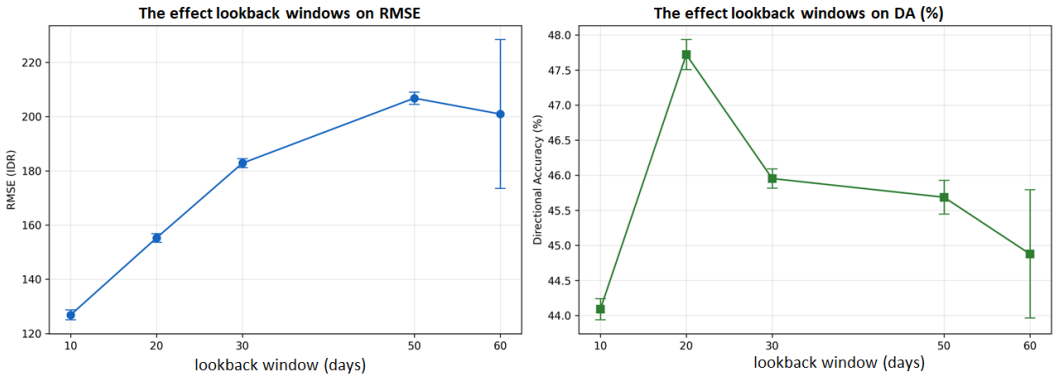

Figure 8. Effect of lookback window length on RMSE (left axis) and DA (right axis) for Attn-GRU-V2 across 30 runs. Error bars represent 95% confidence intervals.

RMSE exhibits a monotonically increasing trend from lookback 10 to 50, indicating that longer history consistently degrades price-level accuracy for the attention model. Directional accuracy follows a non-monotonic pattern, peaking at lookback 20 (47.72 ± 0.21%) before declining at longer windows.

This result suggests that a 20-day lookback provides the most useful temporal context for direction prediction, while extending the window introduces additional noise rather than signal. Notably, the confidence interval for lookback 60 is substantially wider (RMSE ± 27.47 IDR, MAPE ± 0.68%) than for all other configurations, indicating that very long lookback windows introduce training instability. This divergence further reinforces that optimal temporal context depends on the target objective: shorter windows favor stable price estimation, while intermediate windows balance noise reduction and directional signal extraction.

5.4.2. Architectural Contribution of Attention

Table 8 isolates the contribution of the attention mechanism by comparing Attn-GRU-V2 against a GRU model with identical topology but without the attention layer (i.e., the context vector is replaced by the terminal hidden state of the second GRU layer).

Table 8. Ablation study: contribution of attention mechanism (mean ± 95% CI over 30 runs, lookback = 30)

Configuration	RMSE (IDR)	MAE (IDR)	MAPE (%)	R^2	DA (%)
GRU (no attention)	92.68 ± 1.56	67.77 ± 1.05	1.88 ± 0.03	0.9668 ± 0.0011	41.44 ± 0.37
Attn-GRU-V2	182.83 ± 1.65	141.49 ± 1.33	3.91 ± 0.04	0.8709 ± 0.0023	**45.95 ± 0.14**

The ablation confirms that the attention component directly increases RMSE while improving directional accuracy. The GRU without attention achieves RMSE = 92.68 ± 1.56

IDR and DA = 41.44 ± 0.37%, which are comparable to, but distinct from, the three-layer GRU baseline in Table 5 (RMSE = 94.02 ± 1.22 IDR, DA = 41.71 ± 0.23%). The small difference reflects architectural variation (two-layer vs. three-layer GRU), confirming that the ablation isolates the effect of attention within a controlled setting. Adding the Bahdanau context vector increases RMSE by approximately 97% while improving DA by 4.5 percentage points. The confidence intervals of the DA values do not overlap ([41.07, 41.81] vs. [45.81, 46.09]), providing strong evidence that the directional improvement is a direct consequence of the attention mechanism. This result highlights that attention does not universally improve predictive accuracy, but instead reshapes the error profile, favoring directional consistency over pointwise precision.

5.5. Discussion

The central empirical finding is a systematic trade-off between price-level accuracy and directional accuracy in attention-augmented recurrent networks applied to univariate stock price data. The GRU baseline achieves the lowest RMSE across all 30 runs, while Attn-GRU-V2 consistently achieves the highest DA. These performance dimensions do not improve simultaneously, indicating that different modeling choices prioritize distinct predictive objectives.

The near-uniform attention weights observed in Section 5.3 provide a mechanistic explanation for this behavior. Because the attention mechanism does not identify selectively informative timesteps, the resulting context vector approximates an average of hidden states rather than a selective summary. This contrasts with Liu et al. [18], who reported concentrated attention weights after CEEMD-based noise reduction, suggesting that attention selectivity depends on the signal-to-noise ratio. In the present setting, the averaging effect acts as a low-pass filter, smoothing prediction trajectories and reducing sensitivity to short-term fluctuations. As a result, directional stability improves, but deviations from actual price levels increase.

Directional accuracy for all models remains below the 50% random baseline, consistent with weak-form EMH. This suggests that historical price sequences do not contain exploitable directional signals. The presence of structural breaks during the 2020–2023 period, as reported in prior IDX studies [25], [26], further increases forecasting difficulty. These findings should therefore be interpreted as reflecting market characteristics rather than model failure.

This interpretation aligns with weak-form EMH theory [1], [24]: in the absence of temporally localized predictive structure, attention mechanisms converge to near-uniform weighting. The observed behavior is thus consistent with the underlying data-generating process. The results also clarify discrepancies in prior studies. Reported improvements of attention-based models [6], [7], [16], [17] are typically observed in multivariate settings, where attention can exploit cross-feature interactions. In contrast, the present univariate formulation isolates temporal selectivity and shows that attention provides no advantage for price-level accuracy in this regime.

From a practical perspective, the trade-off has clear implications. GRU remains preferable for applications requiring precise price estimation, while Attn-GRU-V2 provides a consistent advantage for direction-based decision-making. The ablation study further indicates that a 20-day lookback offers the most effective balance for directional prediction. The intermediate performance of Attn-GRU-V3 suggests that the placement of attention within the architecture is critical. The additional GRU layer following attention in V3 may attenuate the smoothing effect observed in V2. This architectural sensitivity highlights the importance of carefully designing attention integration in recurrent models. Finally, the lookback analysis supports the smoothing interpretation. At shorter windows, temporal context is limited, while longer windows introduce excessive averaging. The observed peak in directional accuracy at 20 days reflects a balance between noise reduction and signal preservation..

6. Conclusions

This study presents a statistically rigorous evaluation of four deep learning architectures for stock price forecasting on TLKM data from the Indonesia Stock Exchange (IDX), covering 19 years of observations and assessed through 30 independent runs with confidence intervals and Wilcoxon signed-rank tests. The results consistently show a statistically significant trade-off between price-level accuracy and directional accuracy when attention

mechanisms are incorporated into GRU models. The GRU baseline achieves the lowest RMSE, while Attn-GRU-V2 consistently achieves higher directional accuracy across all runs.

These findings directly address the research objective of evaluating whether attention mechanisms improve forecasting performance in a univariate price setting. The results support the hypothesis that attention does not enhance absolute prediction accuracy in this context, but instead alters model behavior toward improved directional consistency. This demonstrates that different evaluation metrics capture distinct aspects of predictive performance, and that gains in directional accuracy may occur alongside degradation in price-level precision.

The main contribution of this work lies in providing a mechanistic and statistically validated explanation of this trade-off. By linking near-uniform attention weights to weak-form market efficiency [1], [24], the study clarifies why attention mechanisms may fail to provide benefits in univariate financial time series. This shifts the focus from model comparison toward understanding model behavior, offering insight into when attention-based architectures are appropriate. From a practical perspective, the findings suggest that model selection should be aligned with the intended application: GRU is preferable for accurate price estimation, while attention-augmented GRU is more suitable for direction-based decision-making tasks.

Several limitations should be acknowledged. The study is restricted to a single stock (TLKM), limiting generalizability across assets with different volatility profiles, sector characteristics, and market conditions. In addition, the univariate input setting excludes commonly used features such as technical indicators, trading volume, sentiment signals, and macroeconomic variabless [8], [26]. which may influence attention behavior. The evaluation is limited to one-step-ahead prediction $(t + 1)$ and does not account for transaction costs or execution constraints in real trading environments. Furthermore, the nonstationarity of the TLKM price series introduces a domain shift between training and test periods, which may affect predictive performance.

Future work should extend this analysis to multiple stocks with diverse characteristics and incorporate multivariate inputs to evaluate whether attention mechanisms exhibit stronger selectivity under richer feature representations. Alternative validation strategies, such as rolling or walk-forward evaluation, should be considered to assess robustness across market regimes. Additional investigation into attention behavior in different signal-to-noise conditions, as well as architectural variations such as adaptive lookback windows or ensemble models, may further clarify the conditions under which attention mechanisms provide practical advantages in financial forecasting.

Author Contributions: Conceptualization and Methodology: RDH.; Software: RDH; Validation: GFS., FAZ; Formal analysis and Investigation: RDH.; Resources: FAZ.; Data curation: FAZ.; Writing—original draft preparation: RDH; Writing—review and editing: RDH; Visualization: RDH.; Supervision: GFS, FAZ, AZF, AM, AS.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset used in this study is publicly available at: https://doi.org/10.5281/zenodo.19324867. The data were originally retrieved from Yahoo Finance using the yfinance Python library. To ensure reproducibility and avoid potential inconsistencies due to data updates or retrieval differences, a mirrored and preprocessed version of the dataset used in this study has been provided in the above repository.

Acknowledgments: The authors would like to thank the Master's Program in Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, where the first author completed his master's degree.

Conflicts of Interest The authors declare no conflict of interest.

References

[1] R. Butet and S. A. Kesuma, "Efficient Market Hypothesis: A Systematic Literature Review," *RIGGS J. Artif. Intell. Digit. Bus.*, vol. 4, no. 4, pp. 2127–2132, Nov. 2025, doi: 10.31004/riggs.v4i4.3549.

[2] A. T. Haryono, R. Sarno, and K. R. Sungkono, "Stock price forecasting in Indonesia stock exchange using deep learning: a comparative study," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 1, p. 861, Feb. 2024, doi: 10.11591/ijece.v14i1.pp861-869.

[3] F. Furizal, A. B. Fawait, H. Maghfiroh, A. Ma'arif, A. A. Firdaus, and I. Suwarno, "Long Short-Term Memory vs Gated Recurrent Unit: A Literature Review on the Performance of Deep Learning Methods in Temperature Time Series Forecasting," *Int. J. Robot. Control Syst.*, vol. 4, no. 3, pp. 1506–1526, Sep. 2024, doi: 10.31763/ijrcs.v4i3.1546.
[4] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[5] G. Sonkavde, D. S. Dharrao, A. M. Bongale, S. T. Deokate, D. Doreswamy, and S. K. Bhat, "Forecasting Stock Market Prices Using Machine Learning and Deep Learning Models: A Systematic Review, Performance Analysis and Discussion of Implications," *Int. J. Financ. Stud.*, vol. 11, no. 3, p. 94, Jul. 2023, doi: 10.3390/ijfs11030094.
[6] M. Louisa, G. Darmawan, and B. Tantular, "Enhancing Stock Price Forecasting with CNN-BiGRU-Attention: A Case Study on INDY," *Mathematics*, vol. 13, no. 13, p. 2148, Jun. 2025, doi: 10.3390/math13132148.
[7] S. Azman, D. Pathmanathan, and V. Balakrishnan, "A two-stage forecasting model using random forest subset-based feature selection and BiGRU with attention mechanism: Application to stock indices," *PLoS One*, vol. 20, no. 5, p. e0323015, May 2025, doi: 10.1371/journal.pone.0323015.
[8] B. H. C. and I. Jeena Jacob, "A Hybrid CNN-LSTM Attention-Based Deep Learning Model for Stock Price Prediction Using Technical Indicators," *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 5, pp. 28012–28017, Oct. 2025, doi: 10.48084/etasr.12685.
[9] Y. Li, S. Lv, X. Liu, and Q. Zhang, "Incorporating Transformers and Attention Networks for Stock Movement Prediction," *Complexity*, vol. 2022, no. 1, Jan. 2022, doi: 10.1155/2022/7739087.
[10] S. Joddy, "Comparative Analysis of CNN, LSTM, and CNN–LSTM for Indonesian Stock Prediction," *Eng. Math. Comput. Sci. J.*, vol. 7, no. 3, pp. 283–289, Sep. 2025, doi: 10.21512/emacsjournal.v7i3.14326.
[11] S. Agusta, F. Rakhman, J. H. Mustakini, and S. Wijayana, "Enhancing the accuracy of stock return movement prediction in Indonesia through recent fundamental value incorporation in multilayer perceptron," *Asian J. Account. Res.*, vol. 9, no. 4, pp. 358–377, Aug. 2024, doi: 10.1108/AJAR-01-2024-0006.
[12] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734. doi: 10.3115/v1/D14-1179.
[13] A. Lawi, H. Mesra, and S. Amir, "Implementation of Long Short-Term Memory and Gated Recurrent Units on grouped time-series data to predict stock prices accurately," *J. Big Data*, vol. 9, no. 1, p. 89, Dec. 2022, doi: 10.1186/s40537-022-00597-0.
[14] C. Chen, L. Xue, and W. Xing, "Research on Improved GRU-Based Stock Price Prediction Method," *Appl. Sci.*, vol. 13, no. 15, p. 8813, Jul. 2023, doi: 10.3390/app13158813.
[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *arXiv*, May 2016, pp. 1–15. [Online]. Available: http://arxiv.org/abs/1409.0473
[16] M.-C. Lee, "Research on the Feasibility of Applying GRU and Attention Mechanism Combined with Technical Indicators in Stock Trading Strategies," *Appl. Sci.*, vol. 12, no. 3, p. 1007, Jan. 2022, doi: 10.3390/app12031007.
[17] J. Zhang, L. Ye, and Y. Lai, "Stock Price Prediction Using CNN-BiLSTM-Attention Model," *Mathematics*, vol. 11, no. 9, p. 1985, Apr. 2023, doi: 10.3390/math11091985.
[18] Y. Liu, X. Liu, Y. Zhang, and S. Li, "CEGH: A Hybrid Model Using CEEMD, Entropy, GRU, and History Attention for Intraday Stock Market Forecasting," *Entropy*, vol. 25, no. 1, p. 71, Dec. 2022, doi: 10.3390/e25010071.
[19] A. Vaswani *et al.*, "Attention Is All You Need," *arXiv*, vol. 30, Aug. 2023, [Online]. Available: http://arxiv.org/abs/1706.03762
[20] A. Tiwari, C.-S. Shieh, M. P. Kantipudi, and S. Shilpa, "Hybrid CNN–LSTM Integrated with Temporal Fusion Transformer for Accurate and Interpretable Stock Market Forecasting," *Ingénierie des systèmes d Inf.*, vol. 30, no. 11, pp. 3045–3054, Nov. 2025, doi: 10.18280/isi.301122.
[21] P. O. Odion, M. M. Lawal, and A. Abdulrauf, "A Comparative Analysis of an Enhanced Hybrid Model for Predicting Dollar Against Naira Exchange Rate Using Deep Learning and Statistical Methods," *J. Comput. Theor. Appl.*, vol. 2, no. 4, pp. 511–522, Apr. 2025, doi: 10.62411/jcta.12513.
[22] N. Y. Vanguri, S. Pazhanirajan, and T. A. Kumar, "Competitive feedback particle swarm optimization enabled deep recurrent neural network with technical indicators for forecasting stock trends," *Int. J. Intell. Robot. Appl.*, vol. 7, no. 2, pp. 385–405, Jun. 2023, doi: 10.1007/s41315-022-00250-2.
[23] T. T. Thach, "Forecasting Stock Market Indices Using Integration of Encoder, Decoder, and Attention Mechanism," *Entropy*, vol. 27, no. 1, p. 82, Jan. 2025, doi: 10.3390/e27010082.
[24] O. Bustos, A. Pomares-Quimbaya, and R. Stellian, "Machine learning, stock market forecasting, and market efficiency: a comparative study," *Int. J. Data Sci. Anal.*, vol. 20, no. 7, pp. 6815–6839, Nov. 2025, doi: 10.1007/s41060-025-00854-4.
[25] W. Budiharto, "Data science approach to stock prices forecasting in Indonesia during Covid-19 using Long Short-Term Memory (LSTM)," *J. Big Data*, vol. 8, no. 1, p. 47, Dec. 2021, doi: 10.1186/s40537-021-00430-0.
[26] B. Sartono, T. S. Elenaputri, Y. Angraini, and G. A. Dito, "Long Short-Term Memory-Based Prediction of Indonesian Composite Stock Index Returns for Early Identification of Market Crises," *Appl. Comput. Intell. Soft Comput.*, vol. 2025, no. 1, Jan. 2025, doi: 10.1155/acis/6174081.
[27] B. Y. Dwiandiyanta, R. Hartanto, and R. Ferdiana, "Harnessing Deep Learning and Technical Indicators for Enhanced Stock Predictions of Blue-Chip Stocks on the Indonesia Stock Exchange (IDX)," *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 1, pp. 20348–20357, Feb. 2025, doi: 10.48084/etasr.9850.
[28] M. N. Aisy, S. A. Wulandari, and D. R. I. M. Setiadi, "A Probabilistic Feature-Augmented GRU-Attention Model for Chronic Disease Prediction on Imbalanced Data," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 2, pp. 282–293, Jul. 2025, doi: 10.62411/faith.3048-3719-100.