MIKAILALSYS
Journal of Multidisciplinary Sciences

e-ISSN : 2987-2286
p-ISSN : 2987-3924
Index: Harvard, Boston, Sydney, Copernicus, Sinta 4, Scilit, Semantic, Google, etc

https://doi.org/10.58578/mikailalsys.v4i3.12708

A Comparative Analysis of Temporal Fusion Transformer and Long Short-Term Memory for the Forecasting of the Indonesia Composite Stock Price Index (IDX Composite)

Dzikra Rahmi Deni, Nonong Amalita, Yenni Kurniawati, Zamahsari Martha
Universitas Negeri Padang, Indonesia
rahmidenidzikra@gmail.com; nongmat@fmipa.unp.ac.id

Submitted:	Revised:	Accepted:	Published:
Oct 9, 2026	Nov 7, 2026	Nov 18, 2026	Nov 23, 2026

Abstract

Forecasting the Indonesia Composite Stock Price Index (IHSG) is challenging because of nonlinear price movements and changing market conditions. This study compared the forecasting performance of the Temporal Fusion Transformer (TFT) and Long Short-Term Memory (LSTM) models using daily observations from 4 January 2000 to 30 December 2025. Both models used the IHSG and five external variables as inputs, including the S&P 500, global gold prices, West Texas Intermediate crude oil prices, the USD/IDR exchange rate, and the CBOE Volatility Index. The models used a 252-session input window to forecast the next seven trading sessions. Data were partitioned chronologically into training (70%), validation (15%), and test (15%) sets. Forecast accuracy and bias were assessed on 6,601 matched date–horizon pairs, supplemented by analyses of variable importance and temporal attention. TFT achieved a mean absolute error of 580.63, a root mean squared error of 712.65, and a mean absolute percentage error of 7.82%, compared with 962.28, 1,055.30, and 13.16%, respectively, for LSTM. Both models underestimated the index on average. TFT's nominal 80% prediction intervals achieved only 8.44% empirical coverage, with an average width of 409.50 index points and no quantile crossings. TFT produced more accurate point forecasts than the implemented LSTM baseline. Its low empirical coverage indicates that the prediction intervals require calibration.

Mikailalsys is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

Keywords: Temporal Fusion Transformer; Long Short-Term Memory; Indonesia Composite Stock Price Index; Stock Market Forecasting; Probabilistic Forecasting

INTRODUCTION

Stock market indices serve as useful benchmarks for examining equity-market performance and changes in market conditions. In Indonesia, the Indonesia Stock Exchange (IDX) publishes various indices designed to represent different segments of the listed equity market. The Indonesia Composite Stock Price Index, commonly known as the Indeks Harga Saham Gabungan (IHSG), tracks the price performance of companies listed on the Main Board and Development Board of the IDX. Owing to its broad coverage, the index provides a relevant basis for studying aggregate stock-market movements and developing forecasting models for the Indonesian equity market (Bursa Efek Indonesia, 2010).

Predicting the future movement of a stock market index is difficult because financial prices evolve over time and may respond to new information, changing expectations, and market conditions. The efficient market hypothesis provides a theoretical foundation for understanding this challenge by examining the extent to which available information is incorporated into security prices. Fama (1970) explains that market efficiency is considered in relation to different information sets, which has important implications for the predictability of financial markets. Accordingly, historical observations may contain useful patterns for forecasting, but their predictive value should be assessed empirically rather than presumed to remain stable over time (Fama, 1970).

In addition to historical index values, information from other financial markets may provide useful context for understanding IHSG movements. This study considers five external financial variables: the S&P 500 index, global gold prices, global crude oil prices, the USD/IDR exchange rate, and the CBOE Volatility Index (VIX). These variables represent international equity-market performance, commodity-market conditions, currency movements, and market uncertainty. Previous studies have examined the relationship between these factors and the Indonesian stock market. Hidayat and Sudjono (2022),for example, investigated the effects of global gold prices, oil prices, exchange rates, and inflation on the IHSG. Fauziyah et al. (2021) examined IHSG forecasting using an ARIMAX-TARCH approach. Other research has explored LSTM-based forecasting of the IHSG (Patriya et al.,

2023) and the use of deep learning methods to predict the Indonesian composite stock index (Rasyid et al., 2021). Andriyani and Pimada (2025) also examined the influence of global risks on the composite index. These studies provide a basis for considering external financial information in IHSG forecasting, although the predictive contribution of each variable still needs to be evaluated within the model and dataset used in this study.

Among neural-network approaches to sequential data, Long Short-Term Memory (LSTM) was developed to address the difficulty of learning dependencies over extended time intervals. Hochreiter and Schmidhuber (1997) introduced an architecture built around memory cells and gating mechanisms that regulate how information is retained, updated, and passed through recurrent processing. These mechanisms allow the network to preserve relevant information across time steps and help overcome some of the learning limitations of conventional recurrent neural networks. The broader principles of deep learning, as discussed by Goodfellow et al (2016), provide further context for understanding how neural networks learn patterns from data. In financial time-series forecasting, LSTM is therefore a relevant approach because observations from earlier periods may contain information useful for predicting subsequent market movements. However, its forecasting performance is not determined by the architecture alone; it also depends on the characteristics of the dataset, the model configuration, and the design of the forecasting experiment.

Another approach to sequential forecasting is the Temporal Fusion Transformer (TFT), an architecture developed for interpretable multi-horizon time-series prediction. Lim et al. (2021) introduced TFT to handle different types of inputs, including static variables, known future inputs, and time-varying observations that are available only from past periods. Its architecture combines recurrent layers, which process local temporal patterns, with attention mechanisms that help capture relationships across longer sequences. The use of attention is also central to the Transformer architecture introduced by Vaswani et al. (2017),which provides a way to model relationships between elements in a sequence. In addition, Bengio et al. (2013) discussed the importance of representing and learning temporal dependencies in sequence-based tasks. TFT extends these ideas through variable-selection networks and gating components, which help manage multiple input features and identify information relevant to the forecasting process. These characteristics make TFT a suitable model to examine in financial forecasting, particularly when predictions are required across several future time steps. However, its performance still needs to be assessed using the specific dataset and forecasting setting considered in this study.

The growing use of sophisticated neural architectures does not remove the need to compare their forecasting performance empirically. In their study of long-term time-series forecasting, Zeng et al. (2023) found that simple linear models outperformed several more complex Transformer-based approaches across the benchmark datasets they examined. This finding suggests that a more complex architecture does not necessarily produce more accurate forecasts. Model performance may also depend on how its configuration is selected. Bergstra and Bengio (2012) discussed the role of hyperparameter optimization in machine learning, highlighting that the choice of model settings can influence learning outcomes. This issue is relevant when comparing TFT and LSTM, as the two architectures use different mechanisms to learn temporal relationships and may respond differently to their respective configurations. A fair assessment should therefore examine their forecasting errors under a consistent evaluation framework, rather than assume that either architecture is inherently superior.

The reliability of a model comparison also depends on the way forecasting accuracy is evaluated. Different error metrics describe different aspects of prediction performance, so relying on a single measure may not provide a complete picture of a model's accuracy. Hyndman and Koehler (2006) examined several commonly used forecasting accuracy measures and highlighted their limitations under certain conditions. In this study, four evaluation metrics are used to compare the forecasting performance of TFT and LSTM: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and signed bias. MAE measures the average absolute difference between predicted and actual values, while RMSE gives greater weight to larger forecasting errors. MAPE expresses prediction errors as percentages of the actual values, making it easier to assess errors relative to the scale of the observed index. Signed bias, on the other hand, indicates whether a model tends to overestimate or underestimate actual values. Tashman (2000) emphasized the importance of appropriate forecasting evaluation procedures, particularly when assessing model performance over time. Therefore, using these four metrics together provides a more comprehensive basis for comparing the forecasting performance of TFT and LSTM.

Another important consideration is the forecast horizon, since a model that performs well for the nearest future observation may not maintain the same accuracy over longer prediction intervals. To examine this issue, the present study evaluates both TFT and LSTM across seven forecasting horizons, designated H1–H7. The comparison uses the same target

index, the same five lagged external financial predictors, and the same set of evaluation metrics. This design is intended to make differences in forecast performance more directly comparable and to reveal whether the relative performance of the two architectures changes across horizons. The use of a consistent evaluation procedure is essential because forecasting conclusions can depend on the evaluation period, the selected error measures, and the experimental setup (Hyndman & Koehler, 2006).

Recent studies have begun to examine TFT in Indonesian equity forecasting. Hartanto et al. (2024) investigated multivariate time-series forecasting of Indonesian stock prices, while Nurhasan and Ghufron (2026) compared TFT and LSTM for three LQ45 constituents—BBRI, TLKM, and ADRO—using technical and fundamental features over a seven-trading-day horizon. Their results showed that TFT performed better for BBRI and TLKM, whereas LSTM performed better for ADRO, indicating that relative model performance may depend on the characteristics of the asset being forecast. These studies provide relevant evidence for Indonesian equities, but their analyses focus on selected individual stocks rather than the aggregate IHSG. The present study addresses this contextual gap by evaluating the two models for the broad market index across seven forecast horizons and by examining point-forecast accuracy, TFT prediction-interval coverage, variable importance, and temporal attention. Its contribution is therefore a multi-aspect evaluation of the implemented forecasting systems for IHSG, rather than a claim that TFT is universally superior to LSTM.

Based on these considerations, this study compares the forecasting performance of TFT and LSTM for the Indonesia Composite Stock Price Index. The analysis uses 6,325 synchronized daily observations before lag construction and five one-session-lagged external financial variables: the S&P 500 index, global gold prices, global crude oil prices, the USD/IDR exchange rate, and the VIX. Both architectures are evaluated across horizons H1–H7 using MAE, RMSE, MAPE, and signed bias. In addition, the study examines the empirical coverage of TFT's nominal 80% prediction interval and its variable-importance and temporal-attention outputs. The objective is to determine how the implemented forecasting systems differ in accuracy and directional error across the selected horizons, while separately assessing the reliability and interpretability of TFT's probabilistic forecasts.

METHODS

This study adopts a quantitative forecasting approach to compare the performance of the Temporal Fusion Transformer (TFT) and Long Short-Term Memory (LSTM) models in predicting the Indonesia Composite Stock Price Index (IHSG). This section outlines the research design, data selection and collection procedures, data preprocessing, model development, and evaluation methods used to conduct the comparative analysis.

Type of Research

This study employs a quantitative approach with a predictive research design based on secondary financial time-series data. The study aims to compare the forecasting performance of TFT and LSTM in predicting the IHSG, also known as the IDX Composite. The analysis focuses on multi-horizon forecasting using historical IHSG observations and five external financial market variables. Model performance is assessed using out-of-sample forecasts to evaluate the ability of each model to predict observations that were not used during model training. This approach follows established time-series forecasting principles, which emphasize evaluating predictive accuracy on data subsequent to the training period (Tashman, 2000).

Research Design

The research procedure comprises compilation of historical daily market data, data-quality checks, calendar synchronization, lag construction, chronological data partitioning, standardization, model development, model training and validation, out-of-sample forecasting, and comparative evaluation. The daily closing value of the IHSG is designated as the target variable. The external predictor variables consist of the S&P 500, gold futures, WTI crude oil futures, the USD/IDR exchange rate, and the CBOE Volatility Index (VIX). Both models use historical time-series information, although their input structures differ according to their respective implementations.

The dataset is divided chronologically into training, validation, and testing subsets, representing 70%, 15%, and 15% of the observations, respectively. The training subset is used to estimate model parameters, the validation subset supports model selection and training control, and the testing subset is reserved for the final evaluation. Maintaining the temporal order of the observations helps prevent future data from being incorporated into model development for earlier periods. Chronological evaluation is therefore appropriate for

forecasting applications in which the objective is to predict future observations using information available at the forecast origin (Tashman, 2000).

Both forecasting models use a historical encoder length of 252 observations and a forecast horizon of seven trading days. The encoder provides the historical information used to generate forecasts, while the forecast horizon specifies the number of future trading-day values to be predicted. Forecast accuracy is evaluated across the seven horizons and, where applicable, separately for each horizon. The first forecast horizon is also assessed through the rolling one-step evaluation implemented in the research notebook. To ensure comparability, the evaluation aligns predictions by target date and forecast horizon.

Participants and Sampling Technique

This study does not involve human participants. The units of analysis are daily financial time-series observations comprising the IHSG closing value and the associated external market variables. The analytical sample is determined by data availability, chronological coverage, and the availability of observations that can be synchronized across the relevant series.

The synchronized dataset contains 6,325 daily observations before the one-session lag is created. After the lag-1 transformation, the first row has no preceding-session external values and is excluded, leaving 6,324 complete observations for model development and evaluation. The IHSG serves as the target series, while the five external financial variables serve as predictors. The observations are partitioned chronologically rather than sampled probabilistically, preserving the time order required for forecasting.

Instruments and Data Collection

This study uses secondary daily time-series data obtained from Investing.com (*https://id.investing.com*). covering the period from 1 January 2000 to 31 December 2025. Daily closing values are used for all six financial-market series. The target variable is the daily closing value of the Indonesia Composite Stock Price Index (IHSG), while the external predictors are the S&P 500, Gold Futures, Crude Oil WTI Futures, the USD/IDR exchange rate, and the CBOE Volatility Index (VIX). The initial IHSG trading-day dataset contains 6,325 observations, with the observed records extending from 4 January 2000 to 30 December 2025. The IHSG trading calendar is used as the reference calendar when synchronizing the series.

Table 1. Input Variables and Their Roles in IHSG Forecasting

Variable	Instrument or indicator	Data type	Unit	Role
IHSG	IDX Composite (JKSE)	Daily closing value	Index points	Target
S&P 500	S&P 500	Daily closing value	Index points	External predictor
Gold	Gold price series	Daily market value	As recorded in the source data	External predictor
Oil	WTI crude oil price series	Daily market value	As recorded in the source data	External predictor
USD/IDR	US dollar–Indonesian rupiah exchange rate	Daily exchange-rate value	Rupiah per US dollar, subject to source convention	External predictor
VIX	CBOE Volatility Index	Daily index value	Index points	External predictor
Time index	time_idx	Temporal index	Time steps	TFT feature
Month	month	Calendar feature	Month	TFT feature
Day of week	day_of_week	Calendar feature	Day	TFT feature

Table 1 summarizes the financial variables and temporal features used in the forecasting framework. The IHSG closing value is the target, while the five external market series provide historical predictor information. The TFT input structure also includes calendar features comprising day of the week, month, quarter, and year, together with time_idx as the sequential time index and series_id as the series identifier. These temporal and identification fields are modeling features rather than additional external financial-market variables.

Data processing and model development are conducted using Python. The computational environment includes pandas and NumPy for data manipulation, scikit-learn for data standardization, PyTorch for neural network computation, PyTorch Forecasting for time-series forecasting implementation, and Lightning for model training management. These libraries support the overall workflow, including data preparation, sequential input construction, model training, and forecast evaluation.

Data Analysis Technique

Data Preprocessing

Data preprocessing is performed to prepare the financial time series for model training and evaluation. Initial checks cover date formats, data types, chronological ordering, duplicate observations, missing values, and consistency across observation dates. The IHSG trading calendar is used as the reference calendar. For each IHSG trading date, each external series is aligned to the latest available observation dated on or before that date. Each external predictor is then shifted by one IHSG trading session to create a lag-1 input, so that the model uses external information from the preceding session. The synchronized dataset contains 6,325 date-indexed observations before lag construction; removing the first observation without a preceding-session value leaves 6,324 complete observations for modeling.

Following synchronization, the dataset is divided chronologically into training, validation, and testing subsets. Numerical standardization is performed using StandardScaler from scikit-learn. The scaling parameters are estimated from the training data and then applied to the validation and testing data. This procedure prevents information from the validation and testing subsets from influencing the scaling parameters used during model development.

$$z_{j,t}^{\text{lag}} = z_{j,t-1} \quad (1)$$

$$Z_{t,j} = \frac{X_{t,j} - \mu_j^{\text{train}}}{\sigma_j^{\text{train}}} \quad (2)$$

Equation (1) defines the one-session lag, and Equation (2) gives the training-based standardization. Where $z_{j,t}$ is the observed value of external variable j at time t, and $z_{j,t}^{\text{lag}}$ is its one-session lag. The standardized value $Z_{t,j}$ is calculated using the training-period mean μ_j^{train} and standard deviation σ_j^{train}. The same training-derived parameters are applied unchanged to validation and test observations.

$$N_{\text{window}} = n - k - h + 1 \quad (3)$$

In Equation (3), n denotes the number of synchronized observations, k is the encoder length (252 sessions), and h is the forecast horizon (7 sessions). The expression gives the number of possible complete sliding windows before restrictions imposed by the

chronological train, validation, and test partitions. The data are split chronologically into training, validation, and testing subsets at proportions of 70%, 15%, and 15%, respectively, without shuffling. Predictions are inverse-transformed to the original IHSG scale before the forecast metrics are calculated.

Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a recurrent neural network (RNN) architecture developed by Hochreiter and Schmidhuber (1997) to address the difficulty of learning long-term dependencies in sequential data. Unlike conventional recurrent networks, LSTM incorporates a memory cell and gating mechanisms that regulate how information is retained, updated, and exposed over time. These mechanisms enable the model to learn temporal patterns from historical observations and make LSTM suitable as a baseline for financial time-series forecasting.

The LSTM architecture consists of three primary gates: the forget gate, input gate, and output gate. Together, these gates regulate the cell state and hidden state at each time step. The forget gate determines which information from the previous cell state should be retained, the input gate controls the incorporation of new information, and the output gate determines which information is exposed through the hidden state. The principal operations are formulated in Equations (4) – (9).

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (4)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (5)$$

$$\tilde{C}_t = tanh(W_f[h_{t-1}, x_t] + b_c) \quad (6)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (7)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (8)$$

$$h_t = o_t \odot \tanh(c_t) \quad (9)$$

where x_t denotes the input vector at time step t, while h_{t-1} represents the hidden state from the previous time step. The variables c_{t-1} and c_t denote the previous and updated cell states, respectively. The forget gate f_t determines the amount of information retained from the previous cell state, whereas the input gate i_t controls how much new information is incorporated into the cell state through the candidate cell state $\tilde{c}_t$. The output gate o_t regulates the information transferred from the updated cell state to the hidden state h_t.

The matrices W_f, W_i, W_c and W_o represent learnable weight matrices, while b_f, b_i , b_c and b_o denote their corresponding bias vectors. The sigmoid function $\sigma(\cdot)$ maps

values into the interval $[0,1]$, while the hyperbolic tangent function $\tanh(\cdot)$ maps values into the interval $[-1,1]$. The notation $[h_{t-1}, x_t]$ indicates vector concatenation, and $\odot$ denotes element-wise multiplication. Together, these operations enable the LSTM to retain relevant historical information, incorporate new input information, and generate an updated hidden representation at each time step.

In this study, LSTM is implemented as a forecasting baseline using historical IHSG values and external predictor information. Sequential input windows are constructed from historical observations to predict IHSG values over a seven-trading-day forecast horizon. The implemented architecture uses a hidden size of 64, two recurrent layers, a dropout rate of 0.2, a batch size of 64, and a learning rate of 0.001. The model is trained for a maximum of 100 epochs, with early stopping configured using a patience of 10 epochs.

The model parameters are optimized by minimizing the mean squared error (MSE) between the predicted and actual standardized target values. Validation performance is monitored during training to support model selection and determine when further optimization is no longer beneficial. In the recorded training run, the best epoch was epoch 2, with a validation MSE of approximately 0.0481 on the standardized target. Training terminated at epoch 12 under the early-stopping procedure. These values describe the training process and do not represent the model's final out-of-sample forecasting performance.

Unlike the TFT implementation, the LSTM baseline does not use the same known-future calendar inputs. Consequently, the comparison evaluates the forecasting performance of the two implemented systems under their respective input configurations rather than assuming that both models receive identical information.

Temporal Fusion Transformer (TFT)

The Temporal Fusion Transformer (TFT) is an attention-based deep learning architecture developed for interpretable multi-horizon time-series forecasting (Lim et al., 2021). It integrates recurrent processing, variable selection, gating mechanisms, and attention to model temporal dependencies in multivariate data. TFT is designed to accommodate static covariates, time-varying observed variables, and time-varying known future variables. Observed variables are available only for historical time steps, whereas known future variables are available over the forecasting horizon.

The main components of TFT include Gated Residual Networks (GRNs), Variable Selection Networks (VSNs), static covariate encoders, LSTM-based temporal processing, interpretable multi-head attention, and quantile output layers. GRNs regulate information flow through gated transformations and residual connections, while VSNs assign relative weights to input variables. The LSTM layers capture sequential temporal patterns, and the attention mechanism models relationships among temporal representations. Finally, the quantile output layers generate forecasts at specified quantile levels.

The general TFT forecasting function for quantile q and forecast horizon τ can be expressed as:

$$\hat{y}_i(q,t,\tau) = f_q\left(\tau, y_{i,t-k:t}, \mathbf{z}_{i,t-k:t}, \mathbf{x}_{i,t-k:t+\tau}, \mathbf{s}_i\right) \quad (10)$$

where $\hat{y}_i(q,t,\tau)$ denotes the forecast at quantile (q) and horizon (τ); $(y_{i,t-k:t})$ represents historical target values; $(\mathbf{z}_{i,t-k:t})$ denotes time-varying observed variables; $(\mathbf{x}_{i,t-k:t+\tau})$ represents time-varying known variables; and (s) denotes static covariates. Here, (k) is the encoder length, and (τ) identifies the forecast horizon. In this study, $k = 252$ and $\tau = 1, \ldots, 7$. The notation distinguishes past-only observed inputs from covariates available in advance over the decoder horizon.

Variable Selection and Gated Residual Networks

TFT uses Variable Selection Networks (VSN) to estimate the relative relevance of input variables within the learned forecasting representation. The variable selection weights can be expressed as:

$$\mathbf{v}_{\chi t} = \text{Softmax}\left[\text{GRN}_{v_\chi}(\Xi_t, \mathbf{c}_s)\right] \quad (11)$$

where $\mathbf{v}_{\chi t}$ denotes the combined representations of input variables at time (t), and $(\Xi_t, \mathbf{c}_s)$ represents a static context vector. The $softmax$ function normalizes the weights across the input variables, allowing the model to construct a weighted representation of the available information.

A Gated Residual Network (GRN) supports this process by applying nonlinear transformations while regulating the flow of information through gating and residual connections. In general, a GRN can be represented as:

$$GRN_\omega(\boldsymbol{a}, \boldsymbol{c}) = LayerNorm[\boldsymbol{a} + GLU_\omega(\boldsymbol{\eta}_1)] \quad (12)$$

where a is the primary input, c is an optional context vector, η_1 is the transformed intermediate representation, GLU denotes a Gated Linear Unit, and LayerNorm denotes layer normalization. The residual term a is added to the gated transformation before normalization, allowing the network to regulate transformed information while retaining the original input pathway.

Temporal Processing and Interpretable Attention

TFT combines LSTM-based temporal processing with interpretable multi-head attention. The LSTM component processes sequential representations to capture local temporal patterns, while the attention mechanism learns relationships between temporal representations over the input sequence. This combination supports the modeling of both sequential dependencies and longer-range temporal relationships.

The scaled dot-product attention operation is formulated as:

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_{attn}}}\right)\boldsymbol{V} \quad (13)$$

where Q, K, and V denote the query, key, and value matrices, respectively, and d_{attn} is the attention-key dimension used for scaling. The product QK^{T} measures pairwise compatibility between temporal representations; Softmax converts these scores into attention weights, which are then used to form a weighted combination of V. The weights describe the model's learned allocation of attention and should not be interpreted as causal relationships.

Quantile Forecasting

TFT produces forecasts for multiple future horizons and can estimate conditional quantiles to represent predictive uncertainty. In this study, the quantile levels are:

$$q \in \{0.10, 0.50, 0.90\} \quad (14)$$

The 0.50 quantile serves as the point forecast used to calculate point-forecast accuracy metrics, while Q10 and Q90 define the lower and upper bounds of the nominal 80% prediction interval. The nominal level follows from $0.90 - 0.10 = 0.80$. However, nominal quantile levels do not guarantee empirical calibration; coverage must be checked against held-out observations. The overall TFT architecture is illustrated in Figure 1.

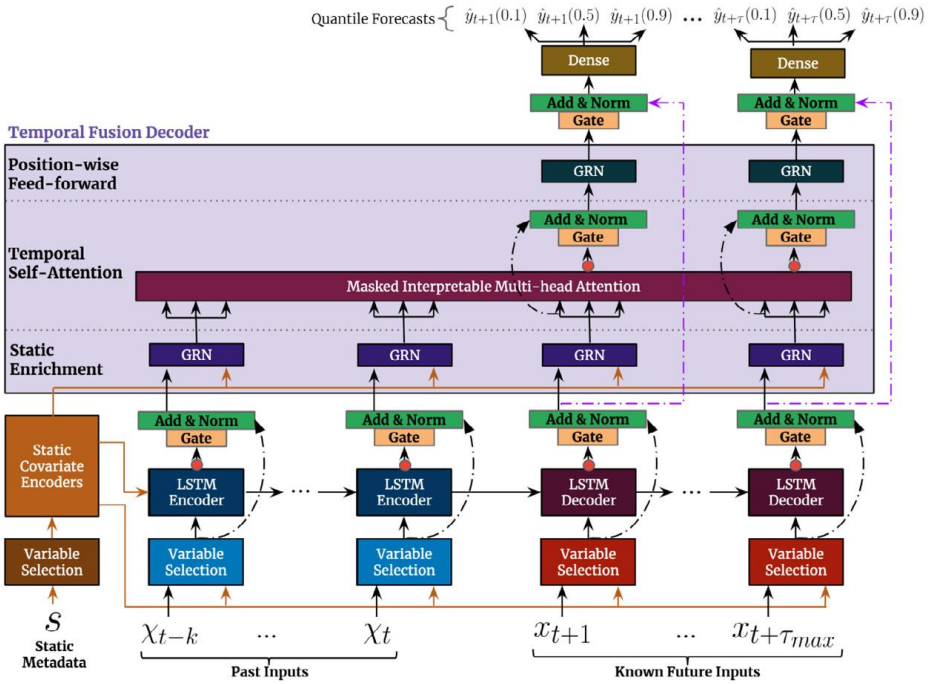


Figure 1. Temporal Fusion Transformer architecture

The figure presents the main TFT components, including variable selection, gated transformations, LSTM-based temporal processing, interpretable attention, and quantile forecasting. In this study, TFT is used to generate seven-step-ahead forecasts of the Indonesia Composite Stock Price Index (IHSG). Its known-future calendar inputs differ from those used by the LSTM baseline; therefore, the comparison evaluates the implemented forecasting systems rather than models receiving identical input information.

Model Evaluation and Comparison

The forecasting performance of the LSTM and TFT models is assessed using the mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and mean forecast bias, where applicable to the reported evaluation results. MAE measures the average magnitude of absolute errors, RMSE gives greater weight to larger errors, MAPE expresses absolute errors relative to observed values, and bias indicates the direction of systematic forecast error. These measures provide complementary perspectives on forecast accuracy (Hyndman & Koehler, 2006).

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \qquad (15)$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (16)$$

$$\text{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (17)$$

where N denotes the total number of evaluated forecast cases, y_i represents the observed IHSG value, and $\hat{y}_i$ denotes the corresponding predicted value. MAE measures the average absolute difference between observed and predicted values in index points. RMSE measures the square root of the average squared prediction errors and assigns greater weight to larger deviations. MAPE expresses the average absolute prediction error as a percentage of the observed values and is calculated only for nonzero y_i. Bias measures the average signed prediction error, calculated as the predicted value minus the observed value. A positive bias indicates systematic overprediction, whereas a negative bias indicates systematic underprediction.

The models are compared using predictions aligned by target date and forecast horizon. Performance is assessed across the combined seven-day horizon and separately for individual horizons where applicable. The first forecast horizon is also evaluated through the rolling one-step procedure implemented in the notebook. Matching target dates and forecast horizons ensure that the comparison is based on corresponding forecasting cases. All point-forecast metrics are calculated after the predictions have been returned to the original IHSG scale.

Lower MAE, RMSE, and MAPE values indicate better performance under the respective metrics, whereas bias is used to examine the direction and magnitude of systematic forecast error. The comparative findings are interpreted in consideration of the models' architectures and input structures. In particular, because the TFT uses known-future calendar features that are not used by the LSTM baseline, differences in forecasting performance are interpreted as differences between the implemented forecasting systems rather than as effects attributable solely to model architecture.

RESULTS

The effective modeling dataset contained 6,324 complete observations after the external variables were synchronized with the IHSG calendar and shifted by one session. The final test comparison covered 18 January 2022 to 30 December 2025. Each test origin generated seven forecasts, one for each horizon H1–H7. After matching predictions by target date and horizon, 6,601 date–horizon pairs were retained for the combined comparison, comprising 943 pairs at each individual horizon. Thus, 6,324 refers to complete daily input observations after lag construction, whereas 6,601 refers to the total number of matched test predictions across seven horizons.

Table 2. Overall forecasting performance on matched test predictions

Model	Matched predictions	MAE (points)	RMSE (points)	MAPE (%)	Bias (prediction − actual)
TFT	6,601	580.63	712.65	7.82	−569.81
LSTM	6,601	962.28	1,055.30	13.16	−961.09

Table 2 indicates a consistent advantage for TFT across the three point-forecast accuracy criteria, rather than an improvement limited to one error measure. The reduction in RMSE alongside MAE suggests that TFT also mitigated the influence of larger deviations. The negative bias for both models reveals a shared directional tendency: forecasts were generally below the observed index. This distinction matters because smaller absolute errors do not, by themselves, remove systematic underestimation.

Table 3. Forecast errors by horizon

Horizon	N	TFT MAE	TFT MAPE (%)	LSTM MAE	LSTM MAPE (%)
H1	943	561.49	7.57	967.97	13.26
H2	943	570.81	7.70	946.92	12.96
H3	943	574.44	7.74	923.63	12.63
H4	943	579.94	7.81	935.34	12.79
H5	943	586.23	7.90	980.34	13.41
H6	943	592.65	7.98	1,019.07	13.95
H7	943	598.84	8.06	962.70	13.15

The horizon-wise pattern in Table 3 shows that TFT's advantage is maintained throughout the full seven-step forecast, rather than being driven only by the nearest horizon. Its gradually increasing error suggests a modest deterioration as the forecast lead extends. In

contrast, LSTM's errors fluctuate more across horizons, indicating less uniform performance over the forecast window. The practical implication is that TFT provides the more consistent point-forecast baseline in this experiment, although neither model should be assumed to track every abrupt market movement.

The probabilistic results show an important distinction between correct quantile ordering and useful uncertainty quantification. TFT achieved an empirical coverage of 8.44% for its nominal 80% prediction interval, with an average interval width of 409.50 index points and a quantile crossing rate of 0%. The absence of quantile crossing means that the lower, median, and upper forecasts remained in the expected order; it does not establish that the interval captures the intended proportion of observations. The observed coverage is far below its nominal level, indicating severe under-coverage and limiting the use of these intervals as calibrated measures of forecast uncertainty.

TFT Interpretability Results

The variable-importance output contained no static-variable importance values for this single-series setup. The encoder results suggest that the model relies most strongly on the historical target sequence, with VIX_lag1 and the calendar quarter also receiving notable relative weight. This pattern is consistent with the model drawing on both recent IHSG history and selected external or calendar context; it does not establish that these variables cause subsequent index movements. For the decoder, the time index dominates the known-future inputs, suggesting that the model makes substantial use of temporal position when constructing forecasts.

Table 4. Encoder variable importance

Rank	Variable	Importance
1	IHSG	314.900
2	VIX_lag1	154.070
3	quarter	117.300
4	USDIDR_lag1	99.892
5	Oil_lag1	96.436
6	month	68.810
7	SP500_lag1	32.293
8	Gold_lag1	31.568
9	time_idx	15.259
10	day_of_week	12.474

Table 5. Decoder variable importance

Rank	Variable	Importance
1	time_idx	766.741
2	day_of_week	65.714
3	quarter	63.084
4	month	47.461

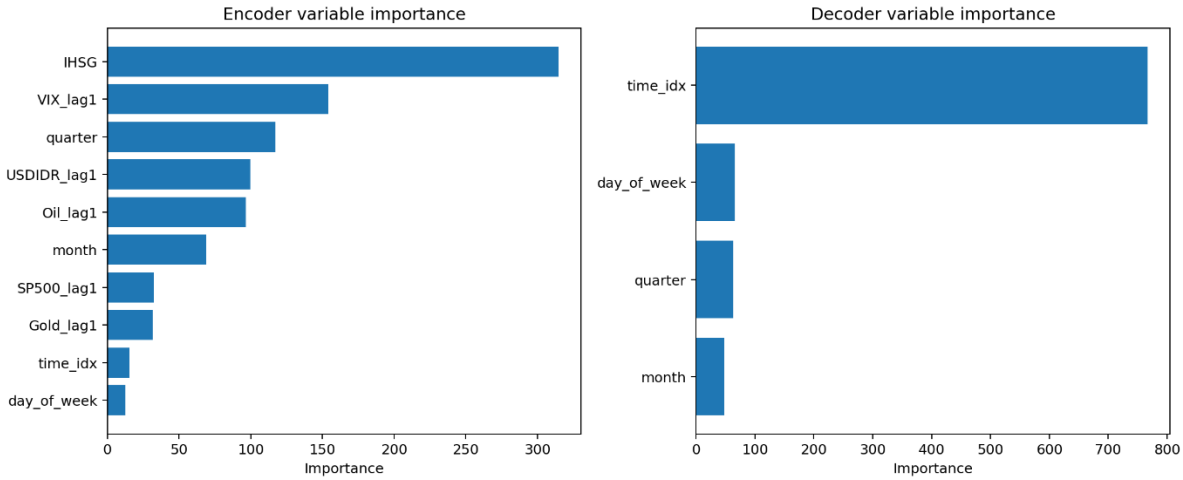


Figure 2. Variable importance for the TFT encoder and decoder

Figure 2 highlights a difference between the encoder and decoder roles: the encoder importance is distributed across the historical IHSG series and selected lagged or calendar inputs, whereas the decoder importance is concentrated strongly in the time index. This indicates that the model uses different information sources when representing past conditions and known future time positions. Importance scores are internal model weights and should be read as relative contribution measures, not as measures of statistical significance or causality.

The temporal-attention output was examined as a separate interpretability diagnostic. The analysis indicated that the strongest reported attention weights were assigned to the earliest encoder positions, with the next position also receiving comparatively high weight. Because these positions are relative locations within the 252-session look-back window, they should not be interpreted as fixed calendar dates or as evidence that a particular historical date causes a forecast. The attention profile is described here but is not displayed as a separate figure in the current manuscript.

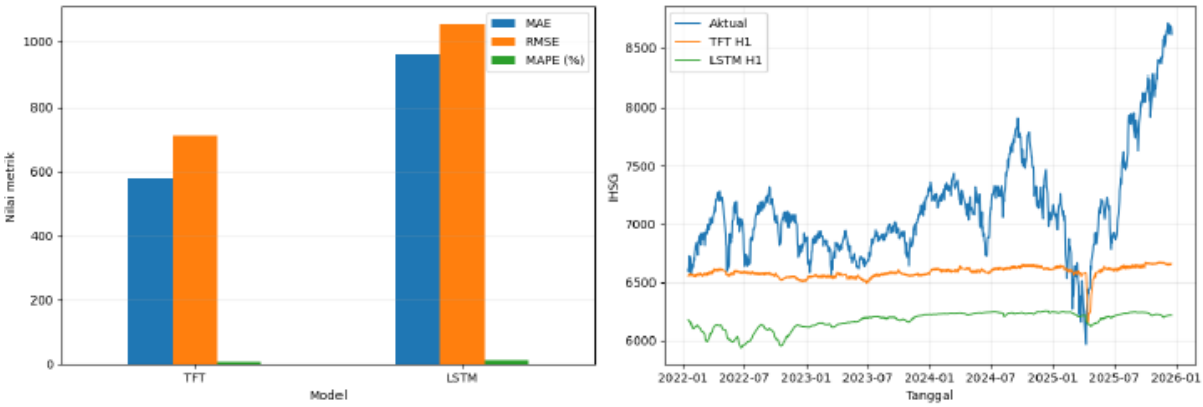


Figure 3. Combined H1–H7 error metrics and actual versus H1 forecasts on matched dates.

Figure 3 links the numerical comparison to the time-series behavior of the forecasts. The visual separation between the actual series and the smoother prediction paths is most apparent around pronounced upward movements, where the forecasts do not fully follow the amplitude of the observed rise. This helps explain why both models exhibit negative bias and why point-forecast metrics should be read together with the time-path plot: average error summarizes performance over all matched cases, whereas the plot reveals where deviations occur and how the models respond to turning points.

DISCUSSION

The experimental results demonstrate that TFT achieved better point-forecast accuracy than the implemented LSTM baseline for seven-horizon IHSG forecasting. The lower MAE, RMSE, and MAPE values indicate that TFT produced smaller prediction errors across the evaluated test observations. This finding is consistent across the aggregate evaluation and individual forecasting horizons.

The comparison with earlier Indonesian stock-market studies provides useful context for this result. Hartanto et al. (2024) applied TFT to multivariate forecasting of Indonesian stock prices, and Nurhasan and Ghufron (2026) reported that TFT achieved lower MAPE for BBRI and TLKM, while LSTM performed better for ADRO. The latter finding is important because it shows that the relative performance of TFT and LSTM is not necessarily uniform across assets. In the present study, TFT achieved lower aggregate errors for IHSG across all seven evaluated horizons. This extends the comparison to a broad

market index, although the numerical results should not be compared directly with the issuer-level MAPE values because the targets, data, predictors, and evaluation settings differ.

One possible explanation for this performance difference is the ability of TFT to integrate historical information, variable selection mechanisms, recurrent temporal processing, and attention-based components within a unified forecasting architecture. Furthermore, TFT incorporated known future calendar information, which provided additional temporal context during prediction. However, the observed performance difference should be interpreted within the experimental configuration of this study. TFT and LSTM did not have completely identical input configurations or hyperparameter optimization procedures. TFT underwent a Random Search procedure, whereas LSTM used a predefined configuration with early stopping. Therefore, the results demonstrate the superiority of the implemented TFT configuration over the implemented LSTM baseline rather than establishing the universal superiority of TFT over LSTM.

The horizon-wise results further indicate that TFT maintained relatively consistent performance across the seven forecasting steps. Its MAPE increased gradually from 7.57% at H1 to 8.06% at H7. This pattern is consistent with the possibility that longer prediction leads involve greater uncertainty because the target observations are increasingly distant from the historical information available to the model. Nevertheless, the magnitude and pattern of forecasting errors may also depend on market conditions and the characteristics of the evaluation period.

An additional finding concerns the negative forecast bias exhibited by both models. TFT recorded a bias of approximately -569.81, while LSTM recorded approximately -961.09. These results indicate that both models tended to underestimate actual IHSG values during the test period. Although TFT reduced the magnitude of this bias relative to LSTM, its negative bias suggests that lower overall error does not necessarily eliminate systematic prediction deviations. This characteristic should be considered when interpreting the practical implications of the forecasting results, particularly when predictions are used to support market monitoring and risk assessment.

The probabilistic forecasting results provide a further perspective on TFT performance. Despite producing a quantile crossing rate of 0%, the model achieved an empirical coverage of only 8.44% for its nominal 80% prediction interval. This substantial difference indicates that the predicted interval did not adequately represent the observed

uncertainty in IHSG movements. A zeroquantile crossing rate only confirms that the predicted quantiles remained correctly ordered; it does not demonstrate that the interval is sufficiently wide or statistically calibrated. Therefore, the results suggest that the model's point-forecast performance and probabilistic reliability should be assessed as separate dimensions.

The use of quantile-based outputs is also consistent with broader applications of TFT beyond financial markets. Akmal and Pratama (2026), for example, evaluated TFT for short-term sales forecasting using MAE and q-Risk and examined variable-selection and attention outputs. Nilasari et al. (2026) also compared TFT and LSTM in multivariate financial forecasting, although their target was gold prices rather than IHSG. These studies support the relevance of evaluating both predictive accuracy and model interpretation, but they do not establish that the prediction intervals in the present study are calibrated. The observed 8.44% coverage therefore remains a study-specific limitation that must be addressed through calibration and further out-of-sample testing.

The limited empirical coverage may indicate that the predicted intervals were too narrow relative to the variability of the actual observations. It may also reflect limitations in the model's representation of financial market uncertainty. Further investigation using calibration techniques, alternative quantile objectives, and coverage evaluation across different market regimes would be valuable. Such investigations could improve the reliability of prediction intervals without necessarily changing the point-forecast architecture.

The interpretability analysis also provides insights into the internal forecasting behavior of TFT. The encoder variable importance results identified historical IHSG information as an important component of the model's input representation. Meanwhile, the decoder importance results indicated a prominent contribution from the time index. These importance values describe the relative contribution of input variables within the trained model and should not be interpreted as evidence of causal relationships. Similarly, temporal attention provides information about the distribution of attention across historical positions, rather than establishing direct causal effects on future stock index movements.

Overall, the findings indicate that TFT offers a promising approach for multivariate multi-horizon IHSG forecasting under the experimental conditions examined in this study. Nevertheless, its negative forecast bias and limited probabilistic interval coverage demonstrate that improvements in point-forecast accuracy do not automatically guarantee

reliable uncertainty estimates. Future research should therefore focus not only on reducing forecasting errors but also on improving calibration, evaluating performance across different market regimes, and conducting more controlled comparisons between forecasting architectures.

CONCLUSION

This study compared Temporal Fusion Transformer (TFT) and Long Short-Term Memory (LSTM) for multivariate seven-horizon forecasting of the Indonesia Composite Stock Price Index (IHSG). Using historical IHSG observations and five external financial variables, the models were evaluated through a chronological test procedure consisting of 6,601 matched window–horizon predictions.

The experimental results showed that TFT outperformed the implemented LSTM baseline in terms of point-forecast accuracy. TFT achieved an MAE of 580.63, RMSE of 712.65, and MAPE of 7.82%, compared with 962.28, 1,055.30, and 13.16%, respectively, for LSTM. TFT also maintained lower MAPE values across all seven forecasting horizons. However, both models exhibited negative forecast bias, indicating a tendency to underestimate actual IHSG values.

The probabilistic evaluation revealed that TFT achieved an empirical prediction interval coverage of 8.44%, substantially below the nominal 80% level. Although the quantile crossing rate was 0%, the low coverage indicates that the prediction intervals were not adequately calibrated. The interpretability analysis further identified the relative importance of historical variables and temporal positions within the TFT architecture, providing additional insights into its forecasting behavior.

Despite its stronger point-forecast performance, the results should be interpreted within the limitations of the experimental configuration, particularly the differences in input information and hyperparameter optimization between the two models. Future studies should investigate more comparable optimization strategies, improve probabilistic calibration, and evaluate model performance across different financial market conditions. These developments are important for improving both the predictive accuracy and uncertainty reliability of deep learning-based stock market forecasting systems.

REFERENCES

Akmal, M. A., Trimono, & Pratama, A. R. (2026). Implementation of temporal fusion transformer (TFT) for short-term sales prediction of Telkomsel data packages in East Java. *Jurnal Teknologi Informatika dan Komputer, 12*(1), 399–413. https://doi.org/10.37012/jtik.v12i1.3268

Andriyani, R., & Pimada, L. (2025). The influence of global risks on the composite index. *Journal of Development Economic and Social Studies, 4*(4), 1111–1124.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798–1828.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*, 281–305. https://jmlr.org/papers/v13/bergstra12a.html

Bursa Efek Indonesia. (2010). *Buku Panduan Indeks Harga Saham Bursa Efek Indonesia.*

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance, 25*(2), 383–417. https://doi.org/10.1111/j.1540-6261.1970.tb00518.x

Fauziyah, E., Ispriyanti, D., & Tarno. (2021). Pemodelan dan Peramalan Indeks Harga Saham Gabungan (IHSG) Menggunakan ARIMAX-TARCH. *Jurnal Gaussian, 10*(4), 595–604. https://ejournal3.undip.ac.id/index.php/gaussian/article/view/33102

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press. https://mitpress.mit.edu/9780262035613/deep-learning/

Hartanto, S., & Gunawan, A. A. S. (2024). Temporal fusion transformers for enhanced multivariate time series forecasting of Indonesian stock prices. *International Journal of Advanced Computer Science and Applications, 15*(7), 140–148. https://thesai.org/Publications/ViewPaper?Code=ijacsa&Issue=7&SerialNo=13&Volume=15

Hidayat, R., & Sudjono. (2022). The effect of world gold price, world oil price, USD/IDR exchange rate, and inflation on the joint stock price index (JCI) on the Indonesia Stock Exchange (IDX). *Scholars Bulletin, 8*(1), 33–41. https://saudijournals.com/articles/6203/

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688.

Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting, 37*(4), 1748–1764.

Nilasari, Eko Saputro, R., & Karyono, G. (2026). Perbandingan Model LSTM dan Temporal Fusion Transformer untuk Prediksi Harga Emas. *Building of Informatics, Technology and Science (BITS), 7*(4), 2811–2820. https://ejurnal.seminar-id.com/index.php/bits/article/view/8204

Nurhasan, & Ghufron. (2026). Perbandingan Metode Temporal Fusion Transformer (TFT) dan Long Short-Term Memory (LSTM) dalam Prediksi Harga Saham Indonesia

Berbasis Data Teknikal dan Fundamental. *Applied Information Technology and Computer Science (AICOMS), 5*(1), 118–125. https://aicoms.jurnal.politap.ac.id/index.php/aicoms/article/view/2234

Patriya, E., Latif, A., & Handayani. (2023). Peramalan Harga Saham Penutupan Indeks Harga Saham Gabungan (IHSG) Menggunakan Algoritma Long Short Term Memory (LSTM). *Jurnal Ilmiah Ekonomi Bisnis, 28*(2), 304–314. https://ejournal.gunadarma.ac.id/archive/index.php/ekbis/article/view/7964

Rasyid, A. F., R., D. A., & Ediraras, D. T. (2021). Deep learning methods in predicting Indonesia composite stock price index (IHSG). *International Journal of Computer and Information Technology, 10*(5), 209–217. https://doi.org/10.24203/ijcit.v10i5.153

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting, 16*(4), 437–450.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2023). Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(9), 11121–11128.