Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.1, February 2026, pp. 772~781
e-ISSN: 2548-6861

Application of SARIMA, GRU, and Prophet for Capturing Seasonal Patterns in Consumer Price Inflation

Laily Nissa Atul Mualifah[1*]
[1]Statistics and Data Science, School of Data Science, Mathematics, and Informatics, IPB University
lailyatul@apps.ipb.ac.id

Article Info

Article history:

Received 2025-11-19
Revised 2026-01-03
Accepted 2026-01-13

Keyword:

CPI,
GRU,
Prophet,
SARIMA,
Seasonality,
Sliding window.

ABSTRACT

Seasonal dynamics make inflation forecasting challenging in emerging economies where holiday effects, regulated prices, and supply shocks interact. This study models Indonesia's monthly consumer price inflation (CPI) using official data from Statistics Indonesia (May 2006–April 2025) and evaluates three forecasting paradigms: a classical seasonal baseline (SARIMA), a decomposable model with trend–seasonality components (Prophet), and a neural sequence learner (GRU). A 10-fold sliding window design is employed to preserve temporal order. Performance is assessed with RMSE, MAE, and MASE, summarized across folds with boxplots and statistical descriptives (means, standard deviations, and 95% confidence intervals). Across folds and metrics, Prophet consistently achieves the lowest error and the tightest dispersion, GRU ranks second with competitive accuracy and stable variance, and SARIMA remains a transparent yet weaker benchmark. MASE values below one for Prophet (and generally for GRU) indicate improvements over a naïve baseline. Practically, Prophet's decompositions support policy communication by linking forecast movements to interpretable components (e.g., Ramadan/Eid and year-end effects), while GRU is useful during more nonlinear or volatile periods; SARIMA remains valuable for diagnostics in stable regimes.

I. INTRODUCTION

Seasonal patterns in economic time series pose persistent challenges for accurate modeling and forecasting. The interaction between recurring patterns, nonlinear trends, and structural breaks often leads to model instability and biased forecasts, especially in macroeconomic data such as consumer price inflation (CPI). Traditional approaches such as the Seasonal Autoregressive Integrated Moving Average (SARIMA) model have long served as robust benchmarks for seasonally dependent data due to their interpretability and clear parameterization [1], [2]. However, SARIMA's assumptions of fixed periodicity and linear dynamics can limit its performance when the seasonal structure evolves or when exogenous calendar effects play a major role [3], [4].

Recent studies emphasize that seasonal time series often exhibit multiple, overlapping, or time-varying seasonalities, which require more flexible frameworks such as Multiple Seasonal-Trend Decomposition (MSTL) or TBATS [4], [5]. For business and policy applications, decomposable models like Prophet have become increasingly popular due to their ability to capture complex patterns using additive components of trend, seasonality, and holiday effects [6]–[9]. Prophet's explicit inclusion of holiday and calendar regressors makes it especially suitable for economies like Indonesia, where inflation cycles often correspond to festive seasons such as Ramadan and Eid al-Fitr [10], [11]. Comparative studies show that Prophet and its hybrid forms (e.g., Prophet-SVR, Prophet-EMD) can outperform traditional methods when nonlinear or irregular seasonal patterns are present [6], [7], [12], [13].

Parallel to these developments, advances in deep learning have reshaped the landscape of time-series forecasting. Neural-network architectures such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) can learn long-range temporal dependencies without explicit feature engineering [14]–[16]. Among these, GRU has received increasing attention because of its computational efficiency

and comparable accuracy to LSTM, particularly for economic and financial forecasting [17]–[19]. Researchers have also proposed hybrid and enhanced GRU variants—such as eGRU or Bi-GRU architectures—that improve performance under extreme events or multivariate setups [16], [18], [19]. Empirical comparisons consistently demonstrate that deep-learning methods can capture nonlinear dynamics in inflation and energy-price series more effectively than traditional statistical models [13], [20]–[22].

Within the inflation-forecasting literature, results remain mixed. Classical models like SARIMA or Holt-Winters often perform well under stable conditions [23], [24], whereas neural and hybrid models excel when seasonality interacts with policy shocks or high volatility [25]–[27]. Studies in emerging markets show that inflation data frequently contain irregular periodicities, suggesting that flexible decomposable or learning-based models may better adapt to structural breaks and shifts [10], [25]. However, the performance of these models depends heavily on data characteristics, tuning, and cross-validation strategy—issues that remain under-explored for Indonesian inflation.

This study contributes to that gap by comparing SARIMA, Prophet, and GRU models on Indonesia's monthly CPI. The dataset provides an ideal environment to evaluate seasonal modeling because it combines recurring within-year patterns with evolving macroeconomic conditions. By systematically evaluating three contrasting approaches—a statistical benchmark, a decomposable machine-learning model, and a neural-network sequence model—this study assesses how each captures seasonal complexity and temporal dependencies. Similar comparative frameworks have proven valuable in prior forecasting studies across domains including energy, retail, and tourism [5], [12], [13].

Ultimately, this research aims to offer both empirical and methodological contributions. Empirically, it provides an updated evaluation of seasonal forecasting models in the context of Indonesian inflation—a critical variable for monetary and fiscal policy. Methodologically, it extends prior comparative analyses by jointly considering interpretability, stability, and accuracy under seasonal complexity. The findings are expected to guide statistical agencies, financial institutions, and researchers in selecting forecasting tools that balance robustness, transparency, and adaptability in modeling seasonal time series [3], [15], [21].

II. METHOD

A. Data

This study employs monthly CPI data for Indonesia covering the period May 2006 to April 2025, obtained from Statistics Indonesia (Badan Pusat Statistik) through its official website https://www.bps.go.id. The CPI series represents percentage changes in the consumer price index, which reflects the general movement of consumer goods and service prices across the country.

The selected time span provides a sufficiently long horizon to capture multiple economic cycles, structural shifts, and seasonal patterns within the Indonesian economy. It includes periods of major policy adjustments, global economic crises, the COVID-19 pandemic, and post-pandemic recovery phases – allowing the models to learn from diverse macroeconomic conditions and to identify both stable and evolving seasonal structures. Using data up to April 2025 ensures that the analysis incorporates the most recent and comprehensive information available, enhancing the relevance of the forecasting evaluation. The monthly frequency further allows detailed examination of intra-annual variations, which are essential for modeling seasonal inflation dynamics in Indonesia's consumer price behavior.

B. SARIMA

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model extends the traditional ARIMA framework by incorporating both non-seasonal and seasonal components to account for recurring patterns in time-series data. It is particularly suitable for modeling and forecasting economic indicators that exhibit seasonal fluctuations, such as consumer price inflation. The general form of a SARIMA model is denoted as SARIMA(p, d, q)(P, D, Q)$_s$, where (p, d, q) represents the non-seasonal autoregressive, differencing, and moving-average orders, and (P, D, Q)$_s$ corresponds to their seasonal counterparts with a periodicity s (e.g., s = 12 for monthly data). The general equation for a SARIMA process can be expressed as:

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D Y_t = \Theta_Q(B^s)\theta_q(B)\varepsilon_t \quad (1)$$

where y_t denotes the observed time series at time t, B is the backshift operator ($By_t = y_{t-1}$), $\phi_p(B)$ and $\Phi_P(B^s)$ are the non-seasonal and seasonal autoregressive polynomials, $\theta_q(B)$ and $\Theta_Q(B^s)$ are the corresponding moving-average polynomials, and ε_t represents a white-noise error term with zero mean and constant variance [28]–[30].

The SARIMA modeling process generally involves four main stages: (1) data visualization and stationarity assessment, (2) identification of model orders, (3) parameter estimation and model fitting, and (4) diagnostic checking and model selection. First, the data are visually inspected to detect trends and seasonality. Stationarity is then evaluated using unit-root tests such as the Augmented Dickey–Fuller (ADF) or Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests [31], [32]. If the series is non-stationary, differencing (both regular and seasonal) is applied to stabilize the mean and variance.

Next, candidate models are identified by analyzing the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots, which help determine appropriate values of *p*, *q*, *P*, and *Q*. After potential models are specified, parameters are estimated using maximum likelihood, and the best-fitting model is chosen based on statistical criteria such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and diagnostic measures from residual analysis [33]–[35].

In this study, to improve efficiency and avoid manual trial-and-error, the model specification was selected automatically

using the Auto ARIMA algorithm. Auto ARIMA systematically searches through multiple combinations of (p, d, q)(P, D, Q)$_s$ and evaluates each model's fit according to AIC or BIC values [36]. The procedure combines unit-root testing for differencing order selection with a stepwise search algorithm that iteratively adds or removes parameters to identify the optimal configuration [37]. This approach is particularly advantageous for long seasonal economic time series, such as monthly inflation data, because it reduces subjectivity in model identification and ensures computational consistency [38].

C. Gated Recurrent Unit (GRU)

The Gated Recurrent Unit (GRU) is a recurrent neural network (RNN) architecture designed to efficiently model sequential data by overcoming the vanishing gradient problem inherent in traditional RNNs. The GRU offers a simplified, yet powerful structure compared to the Long Short-Term Memory (LSTM) network, making it computationally efficient while maintaining comparable predictive performance [14], [39]. The GRU architecture operates through two gating mechanisms—an update gate and a reset gate—that control how much past information is carried forward and how much is discarded at each time step. These gates are mathematically defined as:

$$z_t = \sigma(W_z[h_{t-1}, x_t]) \quad (2)$$

and

$$r_t = \sigma(W_r[h_{t-1}, x_t]) \quad (3)$$

where z_t and r_t denote the update and reset gates, respectively; h_{t-1} is the previous hidden state; x_t is the current input; and $\sigma(\cdot)$ is the sigmoid activation function. The candidate hidden state is then computed as:

$$\tilde{h}_t = \tanh(W_h[r_t \odot h_{t-1}, x_t]) \quad (4)$$

and the final hidden state at time *t* is updated as:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (5)$$

Through this mechanism, GRU can adaptively retain or forget information over time, allowing them to effectively capture both short- and long-term dependencies that characterize inflation time series [18], [40], [41].

The GRU modeling process followed three main stages: preprocessing, model training, and optimization. The preprocessing stage involved data reconstruction into data that is suitable for supervised learning methods (i.e., data with predictors and respons variables). Next, the GRU network was constructed as a sequential model comprising one GRU layer followed by a dropout layer 10%, and dense output layer. Lastly, hyperparameter tuning was performed to identify the most effective configuration of the GRU model.

The tuning process aimed to balance accuracy, computational cost, and generalization capability. Three key hyperparameters were optimized through a grid search: number of neurons in the GRU layer, batch size, and learning rate. The tested and selected values are summarized in Table 1. The model was trained using the Adam optimizer over 500 epochs.

TABLE 1
THE SELECTED HYPERPARAMETER VALUES FOR TUNING

GRU		PROPHET	
Hyper-parameter	**Tested values**	**Hyper-parameter**	**Tested values**
Number of neurons (n)	32, 64, 128	Changepoint prior scale (cps)	0.05, 0.1, 0.5
Batch size (bs)	2, 4, 6, 8, 10, 12	Seasonality prior scale (sps)	5, 10, 20
Learning rate (lr)	0.01, 0.001, 0.001	Seasonality mode (sm)	Additive, multiplicative
		Yearly seasonality (ys)	5, 10

This systematic hyperparameter tuning process enabled the GRU to effectively generalize across different seasonal regimes and inflation shocks. By adaptively learning long-term dependencies through its gating mechanisms, the final model achieved stable convergence and accurate inflation forecasts on both training and testing sets. These results are consistent with recent findings demonstrating that GRU-based models can outperform traditional statistical methods and even more complex deep learning architectures in terms of forecasting accuracy, computational efficiency, and robustness to structural changes in macroeconomic data [21], [42].

D. Prophet

The Prophet model represents a flexible and interpretable forecasting framework that combines time-series decomposition with additive regression modeling. It is designed to handle complex seasonal structures, missing data, trend shifts, and holiday effects. Prophet decomposes a time series $y(t)$ into three main components—trend $g(t)$, seasonality $s(t)$, and holiday effects $h(t)$—plus an error term ε_t, expressed as:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \quad (6)$$

where $g(t)$ captures long-term growth or decline, $s(t)$ models periodic patterns such as annual or monthly seasonality, and $h(t)$ accounts for known external events that may influence prices. Prophet assumes an additive structure, which makes it intuitive for interpretation and robust to outliers [43], [44].

The trend component $g(t)$ can take either a piecewise linear or logistic growth form (in this study linear form is used). The piecewise linear trend allows changes in slope at specified changepoints, making it well-suited for inflation data characterized by abrupt shifts due to policy changes, global economic shocks, or structural transitions. It is formulated as:

$$g(t) = (k + a(t)^\top \delta)t + (m + a(t)^\top \gamma) \quad (7)$$

where k is the initial growth rate, m is the offset parameter, $a(t)$ is a vector of indicators showing whether each

changepoint has been passed, and δ and γ represent rate and offset adjustments at changepoints, respectively [43]. Prophet automatically determines changepoints and adjusts the flexibility of the trend using a hyperparameter known as the *changepoint prior scale*, which balances model adaptability and generalization [18].

The seasonal term $s(t)$ is modeled using a Fourier series expansion, which captures periodic patterns of arbitrary complexity. The model approximates seasonality with:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right) \quad (8)$$

where P is the seasonal period (e.g., 12 for monthly data), and N determines the number of harmonics used. This representation allows Prophet to flexibly approximate complex seasonal effects that are not strictly sinusoidal, which is crucial for modeling Indonesia's consumer price inflation given its multi-peak patterns linked to Ramadan, Eid, and agricultural cycles [6]–[8]. Meanwhile, the holiday and event component $h(t)$ captures short-term variations associated with specific calendar events.

Hyperparameter tuning was conducted to refine Prophet's performance by adjusting several key parameters (Table 1): the seasonality mode (additive vs. multiplicative), changepoint prior scale, seasonality prior scale, and yearly seasonality. The changepoint prior scale was varied among 0.05, 0.1 and 0.5, with smaller values producing smoother trends and larger values allowing more flexibility. The seasonality prior scale, controlling the smoothness of the seasonal curve, was tuned between 5, 10 and 20 [7], [8]. The yearly seasonality was tuned between 5 and 10, controlling whether and how Prophet models repeating yearly (12-month) patterns in the data.

E. Model Evaluation

To evaluate the predictive performance and generalization ability of the forecasting models, this study employed a 10-fold cross-validation procedure based on a sliding window scheme (Figure 1). In this approach, the time series data were divided sequentially to preserve their temporal structure. For each fold, approximately 90% of the data were used for model training and the remaining 10% for testing, with the window moving forward chronologically across the dataset. This design ensures that each observation is used for both training and validation while preventing data leakage from future to past, thereby mimicking realistic forecasting conditions. The sliding window cross-validation approach is particularly appropriate for non-stationary and seasonal economic data such as inflation, as it allows performance evaluation across different time regimes and structural patterns [45].

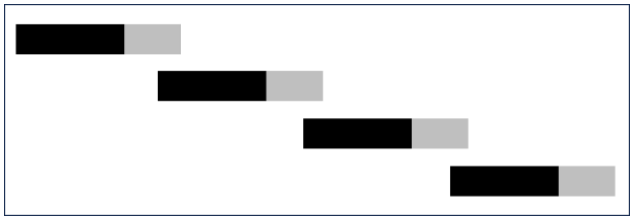

Figure 1. Time series cross validation sliding window illustration.

Model performance was assessed using three complementary error metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Scaled Error (MASE). These metrics capture different aspects of forecasting accuracy and are widely used in time-series evaluation [36]. RMSE measures the square root of the average squared difference between predicted and actual values, placing greater emphasis on large errors and thus reflecting the model's sensitivity to outliers. It is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2} \quad (9)$$

where $y_t$and $\hat{y}_t$represent the actual and forecasted values, respectively, and *n* is the number of test observations. Lower RMSE values indicate better model performance.

The MAE quantifies the average absolute deviation between predicted and actual values, providing a straightforward interpretation of forecast accuracy in the same units as the data:

$$MAE = \frac{1}{n}\sum_{t=1}^{n} | \hat{y}_t - y_t | \quad (10)$$

Unlike RMSE, MAE treats all errors equally and is less sensitive to extreme deviations. Because of its interpretability and robustness, MAE is especially useful for comparing models across different datasets or inflation regimes [46].

The MASE metric was also included to provide a scale-independent measure that facilitates model comparison across datasets or forecasting horizons. MASE is defined as:

$$MASE = \frac{1}{n}\sum_{t=1}^{n} \frac{|\hat{y}_t - y_t|}{\frac{1}{n-1}\sum_{t=2}^{n}|y_t - y_{t-1}|} \quad (11)$$

where the denominator represents the in-sample MAE of a naive one-step-ahead forecast. The MASE value below one indicates that the model outperforms the naive benchmark, while values greater than one suggest inferior performance [47]. Together, RMSE, MAE, and MASE provide a comprehensive evaluation of model accuracy by balancing sensitivity to large deviations, interpretability, and comparability.

III. Results And Discussion

Figure 2 presents the monthly CPI inflation rate of Indonesia from May 2006 to April 2025, figuring the temporal evolution and inherent seasonal fluctuations of consumer prices over nearly two decades. The series exhibits clear cyclical behavior characterized by recurring peaks and

troughs within each year. In the early period between 2006 and 2009, the inflation rate shows substantial volatility, reaching its highest recorded level—exceeding 10%—around 2008, which coincides with the global financial and commodity price crisis that affected Indonesia's domestic price stability. Following this peak, inflation experienced a temporary decline but remained unstable with pronounced annual fluctuations. Between 2010 and 2015, the series continues to show recurrent sharp increases, reflecting structural price adjustments and seasonal shocks, particularly those linked to Ramadan–Eid al-Fitr demand surges and fuel price policy reforms. Notably, there are several spikes around 2013–2014, aligning with government-led fuel subsidy reductions that contributed to sudden price hikes.

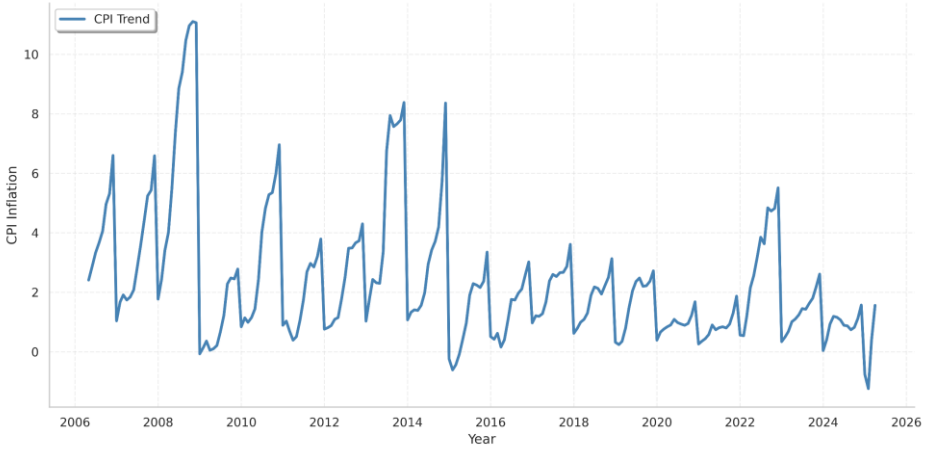

Figure 2. Time series plot of CPI rate at Indonesia.

The evaluation of the forecasting models in this study employed a sliding window cross-validation scheme, which is illustrated in Figure 3. In this approach, the CPI time series was divided sequentially into multiple overlapping training and testing subsets to ensure that the temporal order of the data was preserved throughout the evaluation process. Each panel in Figure 3 represents one of the folds, where the black line indicates the training data used to fit the model and the blue line represents the testing portion used for out-of-sample validation. The procedure follows a 10-fold configuration, where each fold uses approximately 90% of the observations (108) for training and 10% (12) for testing. After each iteration, the window slides forward by one testing segment, discarding the earliest portion of data and incorporating newer observations into the training set. This method ensures that each observation is used for both training and testing across different folds while maintaining the chronological integrity of the time series. The sliding window design also helps assess how well each model adapts to structural changes and evolving seasonal dynamics in Indonesia's inflation pattern, providing a more realistic and comprehensive evaluation of predictive performance.

Across the 10 sliding-window folds, the selected SARIMA specifications show a consistent pattern: seasonal differencing is never required ($D = 0$), and regular differencing is rarely needed (most folds use $d = 0$, with $d = 1$ only in a few windows), indicating stable seasonal means once seasonal dynamics are modeled explicitly. The seasonal component is dominated by autoregressive terms, most frequently SAR (1) or SAR (2) with period 12—sometimes paired with a mild seasonal MA (1)—which reflects strong annual persistence in Indonesia's CPI inflation. On the non-seasonal side, models are typically parsimonious AR (1), with occasional ARMA structures (e.g., ARMA (2,1) or ARMA (2,2)) appearing when short-run volatility is more pronounced. Overall, these outcomes suggest that the inflation process is driven by recurrent yearly inertia plus intermittent transitory shocks, so simple seasonal AR terms capture most of the structure, while MA terms and occasional first differencing are only invoked in folds that include localized drifts or higher short-horizon noise. The detail results are presented in Table 2.

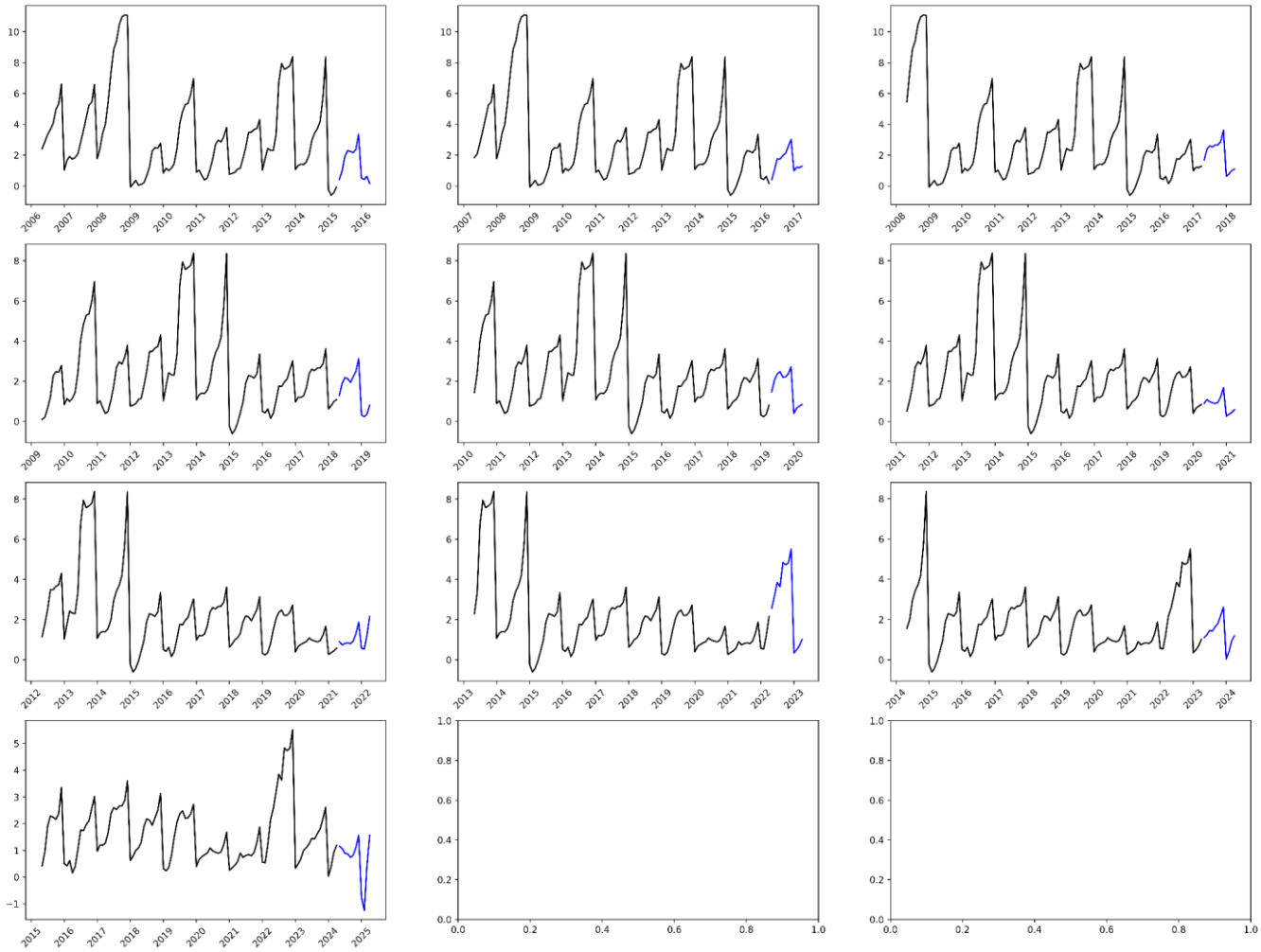

Figure 3. Sliding window scheme of CPI with 10 windows.

For the GRU, the tuned settings display a stable pattern across folds that balances capacity with training stability. Hidden units concentrate at 32 and 128 (each appearing in four folds) with 64 selected twice, suggesting that the series alternates between calmer windows—where a compact 32-unit layer suffices—and more complex regimes requiring a richer 128-unit representation, while 64 units serves as an intermediate capacity. Batch sizes are small to moderate—most often 6 or 12 (with occasional 2, 4, 8, 10)—which injects enough gradient noise to regularize without destabilizing training and fits the relatively short monthly sequence length. The most frequent learning rate is 0.0001 (six folds), indicating that conservative steps are generally needed to achieve smooth convergence on CPI data; 0.01 appears in three folds and 0.001 in one fold, typically paired with smaller batches and/or larger neuron counts when the window contains stronger local nonlinearity and the optimizer benefits from slightly faster progress. Taken together, these results indicate that GRU performance is most reliable with small learning rates and small-to-moderate batches, while model capacity adapts to regime complexity—favoring 32 units in stable periods and 128 units when seasonal patterns interact with short-run shocks.

For Prophet, the tuned settings show a clear pattern across folds: the model most often preferred multiplicative seasonality (8/10 folds), consistent with CPI inflation exhibiting seasonal amplitudes that scale with the level, while additive seasonality appeared only in two calmer windows where seasonal swings were more level-invariant. The yearly seasonality order concentrated at 10 Fourier terms (with occasional 5), indicating that a relatively flexible annual curve is needed to capture Indonesia's multi-peak within-year structure (e.g., Ramadan/Eid, year-end effects), with $N = 5$ chosen when a smoother seasonal shape sufficed. The changepoint prior scale ranged from 0.05 to 0.5: lower values (0.05–0.1) dominated—favoring smoother trends—while 0.5 was selected in a few folds that contained sharper post-shock adjustments, allowing greater trend flexibility. The seasonality prior scale clustered around 10 (spanning 5–20), balancing fit and overfitting risk; higher settings (20) were chosen when the seasonal component needed more "wiggle room," and lower settings (5) when parsimony was

preferable. Overall, these outcomes suggest that Prophet captures CPI dynamics best with a moderately flexible trend and a relatively rich annual seasonal basis, typically in multiplicative mode, while adapting its smoothness parameters to the volatility regime present in each rolling window.

TABLE 2
THE BEST HYPERPARAMETER EACH FOLD

Fold	SARIMA	GRU			PROPHET			
	$(p,d,q)(P,D,Q)_s$	Neuron	Batch size	Learning rate	Changepoint prior scale	Seasonality prior scale	Seasonality mode	Yearly seasonality
1	$(2,0,1)\times(2,0,0)_{12}$	64	10	0.0001	0.5	20	multiplicative	5
2	$(1,0,1)\times(2,0,0)_{12}$	128	2	0.01	0.05	5	multiplicative	10
3	$(1,0,0)\times(2,0,0)_{12}$	128	6	0.001	0.1	10	multiplicative	10
4	$(1,0,0)\times(1,0,0)_{12}$	128	6	0.0001	0.05	10	multiplicative	10
5	$(1,0,0)\times(1,0,1)_{12}$	32	2	0.0001	0.5	5	multiplicative	5
6	$(1,0,0)\times(1,0,0)_{12}$	32	12	0.0001	0.05	10	multiplicative	10
7	$(0,1,0)\times(1,0,0)_{12}$	32	8	0.0001	0.1	20	multiplicative	10
8	$(2,1,2)\times(1,0,1)_{12}$	128	6	0.01	0.5	20	additive	5
9	$(1,0,0)\times(1,0,1)_{12}$	64	4	0.01	0.05	5	additive	10
10	$(2,0,2)\times(2,0,0)_{12}$	32	12	0.0001	0.5	10	multiplicative	10

As the model evaluation, the boxplots in Figure 4 summarize out-of-sample accuracy across folds for the three metrics—RMSE, MAE, and MASE—and they show a consistent ranking of models. Prophet (green) exhibits the lowest central tendency and the tightest interquartile ranges in all three panels, indicating both superior accuracy and high stability across windows; its MASE distribution lies well below 1 for most folds, outperforming the naïve benchmark. GRU (orange) typically sits in the middle, with medians above Prophet but below SARIMA and comparatively compact boxes, suggesting reliable performance and limited sensitivity to regime changes; its MASE cluster hovers around (often just below) the threshold of 1. SARIMA (blue) shows the highest medians and the widest dispersion, including several high outliers—most visible for MASE—implying larger and more variable errors when the window contains shocks or shifts points. The alignment of these patterns across RMSE, MAE, and MASE reinforces the conclusion that Prophet provides the most accurate and stable forecasts in this setting, GRU offers competitive second-best performance with moderate variability, and SARIMA trails with higher error levels and greater fold-to-fold volatility.

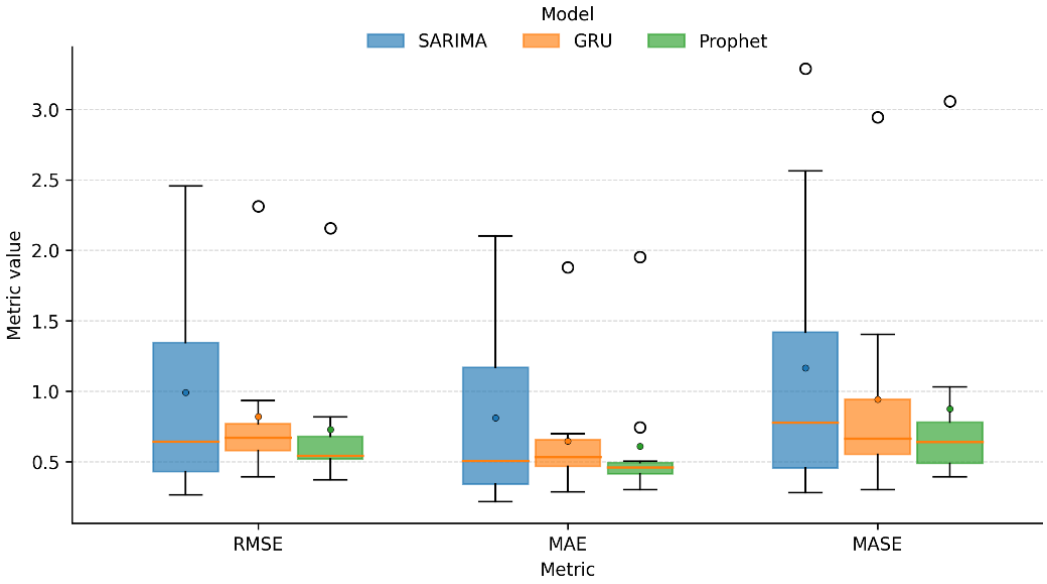

Figure. 4 Boxplot of model performance evaluation.

TABLE 3
STATISTICAL SUMMARY OF METRIC EVALUATION

Model	RMSE	MAE	MASE
SARIMA	0.990 ± 0.763 $(0.444, 1.537)$	0.811 ± 0.632 $(0.358, 1.263)$	1.164 ± 1.015 $(0.438, 1.890)$
GRU	0.820 ± 0.546 $(0.429, 1.210)$	0.646 ± 0.453 $(0.322, 0.970)$	0.942 ± 0.766 $(0.394, 1.490)$
PROPHET	$\mathbf{0.728 \pm 0.516}$ $\mathbf{(0.358, 1.097)}$	$\mathbf{0.609 \pm 0.486}$ $\mathbf{(0.262, 0.957)}$	$\mathbf{0.874 \pm 0.792}$ $\mathbf{(0.308, 1.441)}$

Table 3 synthesizes the fold-level accuracy into means ± SD with 95% confidence intervals and shows a consistent ordering across metrics. Prophet attains the lowest average errors—RMSE 0.728 ± 0.516 (0.358, 1.097), MAE 0.609 ± 0.486 (0.262, 0.957), MASE 0.874 ± 0.792 (0.308, 1.441)—indicating both strong central performance and relatively tight uncertainty bands. GRU ranks second—RMSE 0.820 ± 0.546 (0.429, 1.210), MAE 0.646 ± 0.453 (0.322, 0.970), MASE 0.942 ± 0.766 (0.394, 1.490)—with dispersion comparable to Prophet and slightly smaller SDs for MAE/MASE. SARIMA records the highest means and broadest intervals—RMSE 0.990 ± 0.763 (0.444, 1.537), MAE 0.811 ± 0.632 (0.358, 1.263), MASE 1.164 ± 1.015 (0.438, 1.890)—signaling larger and more variable errors across windows. Relative to SARIMA, Prophet reduces average error by ~26.6% (RMSE), 24.9% (MAE), and 24.9% (MASE); GRU yields ~17.2%, 20.3%, and 19.1% reductions, respectively, while Prophet improves over GRU by ~11.2% (RMSE), 5.7% (MAE), and 7.2% (MASE). Notably, MASE highlights practical relevance: SARIMA's mean > 1 suggests, on average, it underperforms the naïve benchmark, whereas Prophet and GRU means < 1 indicate consistent benchmark-beating forecasts. Although the confidence intervals overlap, the pattern of lower means and tighter spread for Prophet—and, to a lesser extent, GRU—corroborates the boxplot evidence of superior and more stable performance.

To further examine whether the performance differences observed across models are statistically meaningful, the analysis of variance (ANOVA) was conducted. The analysis produced a p-value of 0.2169, indicating that the null hypothesis of equal mean performance across SARIMA, GRU, and Prophet cannot be rejected at conventional significance levels. Accordingly, while Prophet consistently exhibits lower average errors and more stable distributions in descriptive analyses, these advantages do not translate into statistically significant differences. This result reinforces the interpretation that the observed superiority of Prophet is practically relevant and empirically robust across windows, but not statistically conclusive under the ANOVA framework employed in this study.

From a practical standpoint, these findings provide clear guidance for decision making in financial and macroeconomic policy contexts, even in the absence of statistically significant differences under ANOVA. For instance, central banks, fiscal authorities, and financial institutions that rely on short-term CPI forecasts for inflation targeting, budget planning, or interest rate setting can adopt Prophet as a default operational model due to its consistently lower errors, stability across rolling windows, and transparent decomposition of trend and seasonality. The ability to explicitly identify seasonal components—such as Ramadan/Eid effects or year-end demand pressures—enables policymakers to distinguish between transitory and persistent inflation movements, reducing the risk of overreacting to seasonal shocks. In financial markets, asset managers and risk analysts can incorporate Prophet-based CPI projections into inflation-linked bond valuation, real return forecasts, and stress-testing scenarios, while using GRU as a complementary tool during periods of heightened volatility when nonlinear dynamics dominate. Meanwhile, SARIMA remains useful as a benchmark model for model validation and regulatory reporting, where interpretability and methodological familiarity are essential. Overall, the study demonstrates how model selection informed by empirical stability and interpretability can enhance the credibility and effectiveness of inflation-related decisions, even when formal statistical tests do not establish a definitive performance hierarchy.

IV. CONCLUSION

Using nearly two decades of Indonesian CPI data and a 10-fold sliding-window evaluation, this study finds that Prophet delivers the lowest and most stable forecasting errors (RMSE, MAE, MASE) across folds, with GRU a consistent runner-up and SARIMA serving as a transparent but weaker baseline. However, the subsequent ANOVA indicates that these differences are not statistically significant at conventional levels, implying that no single model can be declared superior in a strict inferential sense. Nevertheless, MASE values below 1 for Prophet (and generally for GRU) indicate performance superior to a naïve benchmark. Beyond accuracy, Prophet's decomposable structure enhances interpretability, allowing agencies to trace forecast movements to recognizable events (e.g., Ramadan/Eid, year-end demand) and communicate policy narratives clearly. Overall, the results support adopting Prophet as the default operational forecaster for CPI monitoring, with GRU reserved for periods of heightened nonlinearity and SARIMA maintained as a diagnostic reference.

This study acknowledges the limitations. Future work can build on three priority directions. First, extend from univariate to multivariate models by incorporating macro-financial and

sectoral covariates—e.g., exchange rates (USD/IDR), import prices, core vs. volatile foods, and policy/calendar dummies (VAT changes, fuel price adjustments, Ramadan/Eid, school holidays). Second, evaluate multi-horizon forecasting (1, 3, 6, 12 months ahead) and probabilistic outputs (prediction intervals, CRPS) to quantify risk. Third, investigate hybrid/ensemble strategies that combine Prophet, GRU, and SARIMA or use regime-aware weighting. Collectively, these enhancements would strengthen both explanatory power and operational reliability, yielding forecasts that are more accurate, better calibrated, and easier to trust in policy and market settings.

REFERENCES

[1] J. Banaś and K. Utnik-Banaś, "Evaluating a seasonal autoregressive moving average model with an exogenous variable for short-term timber price forecasting," *For. Policy Econ.*, vol. 131, p. 102564, Oct. 2021, doi: 10.1016/J.FORPOL.2021.102564.

[2] M. G. S. Kenyi and K. Yamamoto, "A hybrid SARIMA-Prophet model for predicting historical streamflow time-series of the Sobat River in South Sudan," *Discov. Appl. Sci.*, vol. 6, no. 9, pp. 1–20, Sep. 2024, doi: 10.1007/S42452-024-06083-X/TABLES/5.

[3] Y. Ensafi, S. H. Amin, G. Zhang, and B. Shah, "Time-series forecasting of seasonal items sales using machine learning – A comparative analysis," *Int. J. Inf. Manag. Data Insights*, vol. 2, no. 1, p. 100058, Apr. 2022, doi: 10.1016/J.JJIMEI.2022.100058.

[4] K. Bandara, R. J. Hyndman, and C. Bergmeir, "MSTL: a seasonal-trend decomposition algorithm for time series with multiple seasonal patterns," *Int. J. Oper. Res.*, vol. 52, no. 1, pp. 79–98, 2025, doi: 10.1504/IJOR.2025.143957.

[5] A. Ampountolas, "Addressing complex seasonal patterns in hotel forecasting: a comparative study," *J. Revenue Pricing Manag.*, vol. 24, no. 2, pp. 143–152, Apr. 2025, doi: 10.1057/S41272-024-00494-6.

[6] L. Guo, W. Fang, Q. Zhao, and X. Wang, "The hybrid PROPHET-SVR approach for forecasting product time series demand with seasonality," *Comput. Ind. Eng.*, vol. 161, p. 107598, Nov. 2021, doi: 10.1016/J.CIE.2021.107598.

[7] H. Abbasimehr, A. Behboodi, and A. Bahrini, "A novel hybrid model to forecast seasonal and chaotic time series," *Expert Syst. Appl.*, vol. 239, p. 122461, Apr. 2024, doi: 10.1016/J.ESWA.2023.122461.

[8] X. Gao, J. Zhou, Y. Ci, and L. Wu, "An improved Prophet emergency traffic-flow prediction model," *Proc. Inst. Civ. Eng. - Transp.*, vol. 178, no. 1, pp. 9–21, Feb. 2025, doi: 10.1680/JTRAN.23.00081.

[9] S. J. Taylor and B. Letham, "Forecasting at Scale," *Am. Stat.*, vol. 72, no. 1, pp. 37–45, Jan. 2018, doi: 10.1080/00031305.2017.1380080.

[10] D. M. PRATIWI, "Perbandingan metode SARIMA, holt winter's exponential smoothing, dan prophet pada peramalan data inflasi di Indonesia," Universitas Telkom, S1 Sains Data - Kampus Purwokerto, Purwokerto, 2025.

[11] T. T. Nguyen, H. G. Nguyen, J. Y. Lee, Y. L. Wang, and C. S. Tsai, "The consumer price index prediction using machine learning approaches: Evidence from the United States," *Heliyon*, vol. 9, no. 10, p. e20730, Oct. 2023, doi: 10.1016/J.HELIYON.2023.E20730.

[12] R. Devi, A. Agrawal, J. Dhar, and A. K. Misra, "Forecasting of Indian tourism industry using modeling approach," *MethodsX*, vol. 12, p. 102723, Jun. 2024, doi: 10.1016/J.MEX.2024.102723.

[13] A. L. M. Serrano *et al.*, "Statistical comparison of time series models for forecasting brazilian monthly energy demand using economic, industrial, and climatic exogenous variables," *Appl. Sci. 2024, Vol. 14, Page 5846*, vol. 14, no. 13, p. 5846, Jul. 2024, doi: 10.3390/APP14135846.

[14] P. T. Yamak, L. Yujian, and P. K. Gadosey, "A comparison between ARIMA, LSTM, and GRU for time series forecasting," *ACM Int. Conf. Proceeding Ser.*, pp. 49–55, Dec. 2019, doi: 10.1145/3377713.3377722.

[15] K. E. ArunKumar, D. V. Kalaga, C. Mohan Sai Kumar, M. Kawaji, and T. M. Brenza, "Comparative analysis of Gated Recurrent Units (GRU), long Short-Term memory (LSTM) cells, autoregressive Integrated moving average (ARIMA), seasonal autoregressive Integrated moving average (SARIMA) for forecasting COVID-19 trends," *Alexandria Eng. J.*, vol. 61, no. 10, pp. 7585–7603, 2022, doi: 10.1016/j.aej.2022.01.011.

[16] H. Balti, A. Ben Abbes, and I. R. Farah, "A Bi-GRU-based encoder–decoder framework for multivariate time series forecasting," *Soft Comput.*, vol. 28, no. 9–10, pp. 6775–6786, May 2024, doi: 10.1007/S00500-023-09531-9/METRICS.

[17] A. Yunita *et al.*, "Performance analysis of neural network architectures for time series forecasting: A comparative study of RNN, LSTM, GRU, and hybrid models," *MethodsX*, vol. 15, p. 103462, Dec. 2025, doi: 10.1016/J.MEX.2025.103462.

[18] Y. Zhang, R. Wu, S. M. Dascalu, and F. C. Harris, "A novel extreme adaptive GRU for multivariate time series forecasting," *Sci. Rep.*, vol. 14, no. 1, pp. 1–10, Dec. 2024, doi: 10.1038/S41598-024-53460-Y;SUBJMETA.

[19] P. Deka, M. Cordeiro-Costas, R. Pérez-Orozco, M. Chabiński, and A. Szlęk, "Novel NSGA-II optimized LSTM and GRU models for short-term forecasting of residential heating load," *Energy Build.*, vol. 344, p. 115999, Oct. 2025, doi: 10.1016/J.ENBUILD.2025.115999.

[20] J. Vitale and J. Robinson, "In-Season Price Forecasting in Cotton Futures Markets Using ARIMA, Neural Network, and LSTM Machine Learning Models," *J. Risk Financ. Manag. 2025, Vol. 18, Page 93*, vol. 18, no. 2, p. 93, Feb. 2025, doi: 10.3390/JRFM18020093.

[21] M. Vilenko, "BiHRNN -- Bi-Directional Hierarchical Recurrent Neural Network for Inflation Forecasting," Feb. 2025, Accessed: Oct. 29, 2025. [Online]. Available: https://arxiv.org/pdf/2503.01893.

[22] L. N. A. Mualifah, A. M. Soleh, and K. A. Notodiputro, "Comparison of GARCH, LSTM, and Hybrid GARCH-LSTM Models for Analyzing Data Volatility," *Int. J. Adv. Soft Comput. its Appl.*, vol. 16, no. 2, pp. 150–165, 2024, doi: 10.15849/IJASCA.240730.10.

[23] A. HASSAN, M. M. ALAM, and A. FAEIQUE, "Forecasting Monthly Inflation in Bangladesh: A Seasonal Autoregressive Moving Average (SARIMA) Approach," *J. Econ. Financ. Anal.*, vol. 7, no. 2, pp. 25–43, 2023, doi: 10.1991/JEFA.V7I2.A61.

[24] G. Alomani, M. Kayid, and M. F. Abd El-Aal, "Global inflation forecasting and Uncertainty Assessment: Comparing ARIMA with advanced machine learning," *J. Radiat. Res. Appl. Sci.*, vol. 18, no. 2, p. 101402, Jun. 2025, doi: 10.1016/J.JRRAS.2025.101402.

[25] T. T. Nguyen, H. G. Nguyen, J. Y. Lee, Y. L. Wang, and C. S. Tsai, "The consumer price index prediction using machine learning approaches: Evidence from the United States," *Heliyon*, vol. 9, no. 10, p. e20730, Oct. 2023, doi: 10.1016/J.HELIYON.2023.E20730.

[26] R. Peirano, W. Kristjanpoller, and M. Minutolo, "Forecasting Inflation in Latin American Countries Using a SARIMA-LSTM Combination," Jun. 2021, doi: 10.21203/RS.3.RS-607554/V1.

[27] O. Barkan, J. Benchimol, I. Caspi, E. Cohen, A. Hammer, and N. Koenigstein, "Forecasting CPI inflation components with Hierarchical Recurrent Neural Networks," *Int. J. Forecast.*, vol. 39, no. 3, pp. 1145–1162, Jul. 2023, doi: 10.1016/J.IJFORECAST.2022.04.009.

[28] W. W. S. Wei, "Time Series Analysis Univariate and Multivariate Methods," *New introduction to Multiple Time Series Analysis*. pp. 1–764, 2006.

[29] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis Forecasting and Control*, 5th ed. Wiley, 2016.

[30] A. S. Azad *et al.*, "Water Level Prediction through Hybrid SARIMA and ANN Models Based on Time Series Analysis: Red Hills Reservoir Case Study," *Sustain. 2022, Vol. 14, Page 1843*, vol. 14, no. 3, p. 1843, Feb. 2022, doi: 10.3390/SU14031843.

[31] Y. M. Maiga, "Empirical Approach to Modelling and Forecasting Inflation using ARIMA Model: Evidence from Tanzania," *J. Agric. Stud.*, vol. 12, no. 2, p. 58, Apr. 2024, doi: 10.5296/JAS.V12I2.21695.

[32] K. A. Notodiputro, Y. Angraini, and L. N. A. Mualifah, *Analisis Data Deret Waktu dengan Python Pendekatan Box-Jenkins dan Machine Learning*. Bogor: IPB Press, 2025.

[33] A. Ahmadpour, S. H. Mirhashemi, and M. Panahi, "Comparative evaluation of classical and SARIMA-BL time series hybrid models in predicting monthly qualitative parameters of Maroon river," *Appl. Water Sci.*, vol. 13, no. 3, pp. 1–10, Mar. 2023, doi: 10.1007/S13201-023-01876-8/TABLES/7.

[34] L. N. A. Mualifah *et al.*, "PERBANDINGAN PERFORMA MODEL ARIMA-GARCH DAN LSTM DALAM MERAMALKAN JUMLAH KUNJUNGAN WISATAWAN DANAU KASTOBA," *J. Gaussian*, vol. 14, no. 2, pp. 314–324, Sep. 2025, doi: 10.14710/J.GAUSS.14.2.314-324.

[35] L. N. A. Mualifah *et al.*, "Forecasting the Number of Passengers for the Jakarta-Bandung High-Speed Rail using SARIMA and SSA Models," *J. Appl. Informatics Comput.*, vol. 9, no. 5, pp. 2443–2449, Oct. 2025, doi: 10.30871/JAIC.V9I5.10720.

[36] Rob J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. 2018.

[37] M. Slimane, N. Bedioui, and M. Besbes, "Assessing integration orders for SARIMA modeling A hypothesis testing approach with information criterion hyperparameter selection, case of predicting gas consumption in central Tunisia," *Energy Strateg. Rev.*, vol. 61, p. 101866, Sep. 2025, doi: 10.1016/J.ESR.2025.101866.

[38] N. A. Fazrina, D. R. Wijaya, B. A. Hadie, and S. D. Budiwati, "Poverty Level Prediction Based on Time Series Data using Auto Arima," *BIO Web Conf.*, vol. 144, p. 04008, Nov. 2024, doi: 10.1051/BIOCONF/202414404008.

[39] K. E. ArunKumar, D. V. Kalaga, C. Mohan Sai Kumar, M. Kawaji, and T. M. Brenza, "Comparative analysis of Gated Recurrent Units (GRU), long Short-Term memory (LSTM) cells, autoregressive Integrated moving average (ARIMA), seasonal autoregressive Integrated moving average (SARIMA) for forecasting COVID-19 trends," *Alexandria Eng. J.*, vol. 61, no. 10, p. 7585, Oct. 2022, doi: 10.1016/J.AEJ.2022.01.011.

[40] H. Balti, A. Ben Abbes, and I. R. Farah, "A Bi-GRU-based encoder–decoder framework for multivariate time series forecasting," *Soft Comput.*, vol. 28, no. 9–10, pp. 6775–6786, May 2024, doi: 10.1007/S00500-023-09531-9/METRICS.

[41] P. Deka, M. Cordeiro-Costas, R. Pérez-Orozco, M. Chabiński, and A. Szlęk, "Novel NSGA-II optimized LSTM and GRU models for short-term forecasting of residential heating load," *Energy Build.*, vol. 344, p. 115999, Oct. 2025, doi: 10.1016/J.ENBUILD.2025.115999.

[42] S. Giantsidi and C. Tarantola, "Deep learning for financial forecasting: A review of recent trends," *Int. Rev. Econ. Financ.*, vol. 104, p. 104719, Dec. 2025, doi: 10.1016/J.IREF.2025.104719.

[43] S. J. Taylor and B. Letham, "Forecasting at Scale," *Am. Stat.*, vol. 72, no. 1, pp. 37–45, Jan. 2018, doi: 10.1080/00031305.2017.1380080;SUBPAGE:STRING:ACCESS.

[44] X. Gao, J. Zhou, Y. Ci, and L. Wu, "An improved Prophet emergency traffic-flow prediction model," *Proc. Inst. Civ. Eng. - Transp.*, vol. 178, no. 1, pp. 9–21, Feb. 2025, doi: 10.1680/JTRAN.23.00081.

[45] X. Mei and P. K. Smith, "A Comparison of In-Sample and Out-of-Sample Model Selection Approaches for Artificial Neural Network (ANN) Daily Streamflow Simulation," *Water 2021, Vol. 13, Page 2525*, vol. 13, no. 18, p. 2525, Sep. 2021, doi: 10.3390/W13182525.

[46] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 Competition: 100,000 time series and 61 forecasting methods," *Int. J. Forecast.*, vol. 36, no. 1, pp. 54–74, Jan. 2020, doi: 10.1016/J.IJFORECAST.2019.04.014.

[47] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *Int. J. Forecast.*, vol. 22, no. 4, pp. 679–688, Oct. 2006, doi: 10.1016/J.IJFORECAST.2006.03.001.