

The Power of Interpolation Effects on SARIMA and LSTM: An Indonesia Domestic Forecasting Case Study

Muhammad Mahdy Yandra¹, Dwi Rantini^{1,2*}, Arip Ramadan³, Chris Matthew Widjaja¹, Adinda Syarifatul Muna¹, Alhassan Sesay⁴

¹Data Science Technology Study Program, Department of Engineering, Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Surabaya, East Java, 60115, Indonesia

²Research Group of Data-Driven Decision Support System, Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Surabaya, East Java, 60115, Indonesia

³Information Systems Study Program, School of Industrial and Systems Engineering, Telkom University, Surabaya Campus, Jl. Ketintang No. 156, Surabaya, East Java, 60231, Indonesia

⁴Faculty of Transformative Education, the United Methodist University, Freetown, Sierra Leone

*Corresponding author: dwi.rantini@ftmm.unair.ac.id

Abstract

In regional economic planning, the accurate prediction of domestic tourism demand presents a critical challenge because limited access to historical time series data often causes advanced forecasting models to perform poorly. This study aims to demonstrate how linear interpolation can effectively serve as a data augmentation technique to overcome data scarcity when developing Long Short Term Memory (LSTM) networks to forecast domestic tourism arrivals. The methodology involves evaluating four distinct forecasting models using 80 months of Indonesian domestic tourism visitation data spanning from January 2019 to August 2025. These models include a statistical benchmark SARIMA (2,0,1)(0,1,1)[3] model selected via grid search out of 216 combinations and three LSTM variants run with identical random seeds. Performance evaluation is rigorously completed using MAE, RMSE, MAPE, sMAPE, and Theil's U statistics. Key findings reveal that while unaugmented LSTM models perform poorly relative to a simple average baseline (Theil's U = 1.38, MAPE = 24.6%), LSTMs enhanced with triple linear interpolation achieve superior predictive accuracy (MAPE = 2.6%, sMAPE = 2.6%, Theil's U = 0.14). This approach significantly outperforms the benchmark SARIMA model (MAPE = 8.0%, Theil's U = 0.63), which otherwise demonstrated adequate statistical validity through satisfactory residual diagnostics, including Ljung Box and Shapiro Wilk tests. In conclusion, the study demonstrates that utilizing linear interpolation as an augmentation technique is highly appropriate and effective, enabling deep learning models like LSTMs to accurately capture underlying temporal patterns from limited tourist arrival.

Keywords

Sustainable Tourism, SARIMA, LSTM, Interpolation, Economic Growth

Received: 20 April 2026, Accepted: 9 July 2026

<https://doi.org/10.26554/sti.2026.11.4.1311-1322>

1. INTRODUCTION

Tourism is a sector that plays a vital role in a region's economic growth (Vu and Hartley, 2022). Tourist visits have been demonstrated to drive the local economy through various expenditures, ranging from accommodation and food to transport (Boto-García and Baños Pino, 2024; Ramadan et al., 2025a). Furthermore, it has been evidenced that tourism creates a multiplier effect on related sectors such as agriculture, the creative industries, and retail. The development of this sector also creates employment opportunities for the local community and contributes to local government revenue through various taxes and fees (Alamineh et al., 2023). In Indonesia, the volume of domestic tourist visitation is indicative of the domestic economy's current state. Consequently, the ability

to accurately forecast these volumes has become a strategic necessity for local governments and tourism industry stakeholders (Li and Gao, 2022; Olukoya, 2023). The formulation of effective marketing strategies, infrastructure planning, and human resource allocation is contingent upon the availability of reliable forecasting data, a prerequisite for minimising the risks of both over tourism and under tourism (Adekunle et al., 2021).

To produce reliable predictions of tourist destinations, several well formed predictive methods need to be applied and quality and quantity data sources must be considered (Wu et al., 2025). In addition to having previous tourist records, those same tourist record have contextual variables that will assist in providing accuracy to current forecasts, i.e., seasonal

calendars, economic indicators, and external factors such as health pandemics and governmental regulations (Hajisafi et al., 2023; Ordóñez-Martínez et al., 2023). Advanced statistical or machine learning models will be much better suited to identifying these dimensions if they have enough complete and well formed data (Supriyadi et al., 2025; Choudhury et al., 2021; Ramadan et al., 2025b). Unfortunately, in practice, the availability of national tourism data (e.g., monthly) will often be very limited and will usually be subjected to changes that further limit the predictive models that can be utilized.

The patterns of tourist visitation data display seasonal elements, including the impact of holidays, cultural and seasonal events, and changes to climate, which has resulted in a predictable set of seasonal visitation records (Torres et al., 2021; Lim and Zohren, 2021). Two modelling methods are well suited to analysing these types of data: Seasonal Autoregressive Integrated Moving Average (SARIMA) adds seasonal components (P, D & Q) to the ARIMA methodology, allowing for an understandable and reliable way of identifying trend and periodicity in relatively short time series (Perez-Guerra et al., 2023). Long Short Term Memory (LSTM), uses a recurrent neural network architecture with gating mechanisms that determine what temporal information to keep and what to forget. This means that LSTM can learn intricate non linear relationships and longer term relationships that are not easily modelled using standard linear statistical techniques (Ahmed et al., 2023; Rantini et al., 2026). LSTMs have been shown to be more accurate in their predictive capabilities than SARIMA when there is sufficient training data available and for seasonal data that is highly irregular or complex (Saputra et al., 2025).

In the wider forecasting literature, many alternative architectures have been explored, such as Prophet and Gated Recurrent Units (GRU). Prophet is an additive model (decomposable) and explicitly defines components in order to capture seasonality and holidays therefore being appropriate to use by practitioners, however, this makes it less flexible when the underlying temporal dynamics differ from the underlying structural assumptions (Taylor and Letham, 2018). GRU is a simplified version of LSTM as it does not utilize a separate cell state, but retains much of the sequential learning capability of LSTM with a less computationally burdensome design (Cho et al., 2014). Comparisons in tourism forecasting environments consistently show that standard LSTM has achieved equal or better predictive accuracy than GRU for moderate length (seasonal) time series, and that SARIMA remains a viable benchmark due to its interpretability and strong historical performance on short structured time series (He et al., 2021; Ahmed et al., 2023). In this research, SARIMA and LSTM form the primary modelling framework, and the main analytical focus is on data augmentation, rather than an exhaustive model comparison.

Both SARIMA and LSTM models have distinct advantages to offer but both can only perform well if there is enough data available for the model. SARIMA will produce reasonable estimates when used with relatively few data points; however, it is able to provide a much better fit for data that have long

term (historical) seasonal patterns (Ampountolas, 2021; Okere and Balyan, 2025). On the other hand, LSTM requires a sufficiently large amount of training data in order to properly optimise its parameters, avoid problems with overfitting, and learn non linear patterns in time series data that represent the greatest benefit of using LSTM as a forecasting technique as compared to traditional statistical forecasting methods (He et al., 2021; Deif et al., 2025). Therefore, when considering national tourism data, this is especially important because these statistics often have very limited historical records with many observed "discontinuities" and comparatively shorter historical records due to events such as the COVID-19 pandemic.

The use of linear interpolation as a solution for data sparse conditions in time series models has been documented and is based on the premise that smooth transitions exist over a period of time in which two known data values occur (Rantini et al., 2021). Interpolation provides SARIMA and LSTM algorithms a continuous input when creating more frequent observations of the training dataset and, as a result, conforms the temporality or chronological order by increasing the density or frequency of the training dataset. In addition, since interpolated values represent estimates of what would have been a 'true' measurement of a time period not represented by a known observation, the use of interpolated values to improve predictive accuracy should be empirically validated prior to being implemented to avoid generating false patterns due to the interpolation (Ali et al., 2025; Ismail and Al-Gounmmeen, 2022). In this study, we addressed those issues by comparing the accuracy of predicting the arrival of domestic tourists to Indonesia using both SARIMA and LSTM algorithms with and without interpolation at a frequency of either 1 or 3 interpolations.

2. EXPERIMENTAL SECTION

2.1 Research Methodology

This research presents a comparative forecast framework to predict the number of domestic tourists in Indonesia using traditional statistical modelling such as SARIMA and deep learning techniques such as LSTM. The research variable used is the number of monthly domestic tourist visits in Indonesia from January 2019 to August 2025 which is obtained directly from the official website of the Indonesian Central Statistics Agency about Tourism. The total observations are 80. One major challenge that emerged during the exploratory stage was the absence of observations. It is a constraint that causes considerable difficulties for deep learning models, which typically require large datasets to reliably identify temporal patterns.

The modelling process commences with data partitioning, whereby the dataset is divided into training and testing subsets. The final six months of observations are reserved as a test set to evaluate model performance, while the remaining historical data is used as a training set for parameter estimation. The stages of analysis in this study are delineated in the flowchart shown in Figure 1.

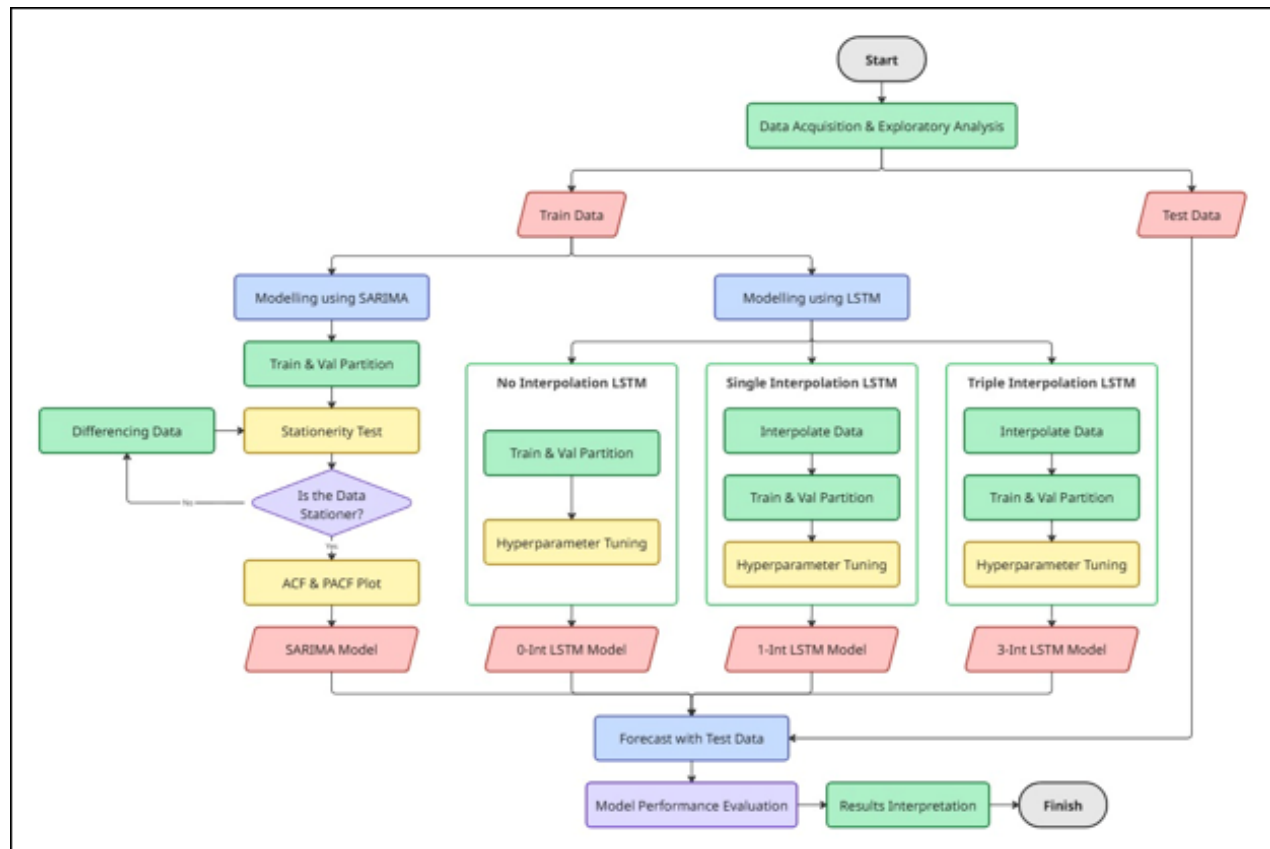


Figure 1. Research Flowchart

2.2 SARIMA Modelling Approach

The SARIMA (Seasonal Autoregressive Integrated Moving Average) model served as a statistical baseline to forecast monthly domestic tourist arrivals. The stationarity of the time series was tested using the Augmented Dickey Fuller (ADF) test prior to creating the model. There is an obvious seasonal structure in the data seen in the preliminary graphic analysis (i.e., line graphs). In light of this, both regular and seasonal differencing were included in the specifications of the model. After comparing the two strategies for adding seasonal differencing to this dataset, we decided to use seasonal differencing at lag 3 ($D=1$, $s=3$) due to the fact that there are clear seasonality patterns shown by both the autocorrelation function (ACF) and partial autocorrelation (PACF) of this series.

After determining that the time series is stationary based on the stationarity tests mentioned previously, we used the ACF and PACF of the differenced format of the data to select p , d , and q for our initial model specifications. We examined the ACF in order to determine the MA order based on the lag where autocorrelation coefficients had no statistically significant lag, and determined the AR order from the PACF. The same diagnostic approach was applied to seasonal lag intervals to analyse the seasonal component (Ajobo et al., 2024). Finally, in order to conduct a systematic and objective model selection process, we performed a grid search through the parameters

for p {0,1,2}, d {0,1}, q {0,1,2}, seasonal order P {0,1,2,3}, $D=1$, Q {0,1,2} and $s=3$ yielding a total of 216 combinations of parameters, and each of these 216 models was assessed using either the AIC or BIC.

2.3 LSTM Modelling Approach

As a complement to traditional statistical methods, this study uses a Long Short Term Memory (LSTM) deep learning architecture to identify non linear patterns and long term dependencies in domestic tourist visitation data. Given the limited number of observations, the study examines three different LSTM model versions: one without interpolation, one with single interpolation (assuming data is collected bi-monthly), and one with triple interpolation (assuming data is collected weekly). These interpolation methods are designed to address the common issues caused by insufficient data, which is a challenge when using deep learning models on time series data with few observations.

The LSTM modeling process begins with a data preprocessing stage that includes normalization and transforming the data structure. The domestic tourist data is normalized first to allow quicker and more stable training. Next, the time series data is changed into a format suitable for supervised learning using the sliding window technique. This technique creates input sequences and target outputs based on a set look back period.

The look back period refers to the number of past time steps used to predict the next time step. The division of training and testing data follows the same rule as the SARIMA model, with the last six months used as the test set. Adjusting the look back period matches the seasonal value in the SARIMA model, ensuring consistency between the two models compared.

The model consists of two stacked LSTM layers. A dropout layer, placed after the first LSTM layer, helps reduce the risk of overfitting by randomly deactivating neurons during training. A dense layer with ReLU activation processes the features from the LSTM layers, concluding with a single output neuron to predict tourist values. The model is compiled using the Adam optimizer and mean squared error (MSE) as the loss function. The tuning process examines various combinations of hyperparameters within a set search space, such as the number of units in the LSTM layer, the number of units in the dense layer, the dropout rate, and the learning rate of the optimizer.

2.4 Interpolation for Data Augmentation

To surmount the limitations associated with the number of observations that affect the training of deep learning models, linear interpolation will be used as a method of augmenting data through data generation. The process of interpolation entails the generation of synthetic data points between each original observation. The strategy is predicated on the assumption that all data points exhibit a smooth transition from one period to another. The employment of two distinct methods in implementing this strategy may facilitate a comparison of its effectiveness by simulating multiple data collection methods for a single category monthly.

The implementation of a single interpolation method will result in the generation of a single additional data point (to be located between 2) for each pair of sequential original observations. The purpose of this function is to create a new index of $2n-1$ (number of overall original observations) for n original observations. The subsequent step involves the estimation of the value at each new index through linear interpolation methodology. To illustrate this point, consider the following example: The mid-point between the two original observations of 1000 and 1200 would yield a mid-point estimate of 1100 tourists. The underlying principle of this technique is predicated on the assumption that data will be collected at a 15-day frequency (or alternatively, collected twice a month).

Using triple interpolation, the addition of three interpolation points will produce an additional total of $4n-3$ new data points for every pair of surrounding observations and therefore greatly increases the total amount of training data. This process creates evenly distributed new indices and uses linear interpolation to determine the values at all the new indices based on the values of the original indices. The basis for this technique assumes that data has been collected on a weekly basis, and so will have a total of four data points in a 1-month period.

While using interpolation does increase the number of data points available for training, only time points that correspond to the original observations will be the basis for the final evalu-

ations. This is done through filtering of the predictive results, whereby the predictive result from every second data point is needed for single interpolation and the predictive results derived from every fourth data point are used for triple interpolation. Therefore, the evaluation of the interpolated data points to the actual data and other models will be a fair and valid comparison.

2.5 Model Evaluation and Comparison

To assess model performance five metrics were used: MAE (Mean Absolute Error), RMSE (Root Mean Square Error), MAPE (Mean Absolute Percentage Error), sMAPE (Symmetric Mean Absolute Percentage Error), and Theil's U. The MAE quantifies an average absolute error in the original unit of measurement (also referred in the literature as forecast deviation). The RMSE applies a greater penalty for large errors due to the squared error principle used to compute the error and thus will be more sensitive to outliers than the MAE. While MAPE values the percentage of the actual value that represents an error in the measurement unit, it is asymmetric, and therefore undefined when the actual value approaches zero, and the sMAPE is provided as a more robust alternative for MAPE. The Theil's U statistic is included to provide context for model performance in comparison to a naïve benchmark (the assumption would be that no prior knowledge existed of future values). These metric choices, taken collectively, will help ensure that a model comparison will not depend solely on one metric, especially given the high volatility associated with the test period.

All four model variants were evaluated equally on the same test set consisting of the final six months of observations. This comparative approach is justified by two reasons: first, to determine which of the four models produces the most accurate forecasts of the number of Indonesian domestic tourists, and second, to quantitatively evaluate the effectiveness of interpolation based data augmentation as a viable solution to the problem of scarcity of data when employing deep learning techniques to forecast short time series data.

3. RESULTS AND DISCUSSIONS

3.1 Exploratory Data Analysis

An exploratory analysis was conducted to understand the characteristics and temporal patterns of data on domestic tourist visits in Indonesia. Figure 2 presents a visualisation of the monthly time series of domestic tourist visits from January 2019 to August 2025.

As demonstrated in Figure 2, several noteworthy patterns emerge when examining this data. In early 2019, the number of tourist arrivals fluctuated between 45 and 80 million per month. By mid-2020, there was a sharp decline to a low of around 25 million tourists per month, reflecting the significant impact of the pandemic on the domestic tourism sector. Since June 2021, a recovery has been observed, with an upward trend continuing through to 2025.

In the year 2024 to 2025, domestic tourist visits reached their highest level during the observation period, ranging from

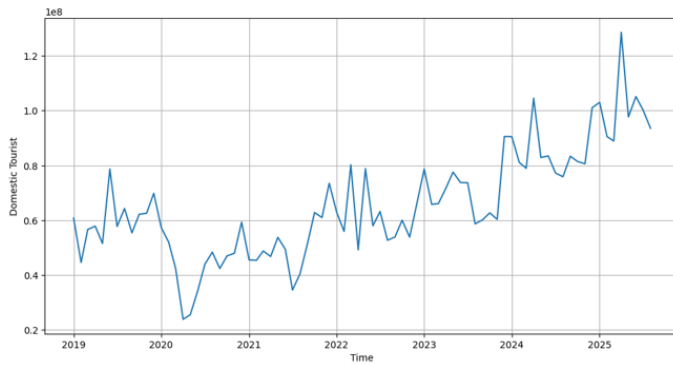


Figure 2. Time Series Plot of Domestic Tourist Visits in Indonesia (January 2019 - August 2025)

80 to 130 million visits per month and even exceeding pre-pandemic figures. The data also displays a relatively consistent pattern of fluctuation from year to year, indicating the presence of a seasonal component in the data. This component must be given due consideration during the subsequent modelling process.

3.2 SARIMA Model Results

The initial stage in constructing a SARIMA model is to assess the stationarity of the data, to ascertain whether differencing is necessary. Stationarity is a fundamental assumption in ARIMA modelling, whereby the data must have a constant mean and variance over time (Schaffer et al., 2021). This test is conducted by using the Augmented Dickey Fuller (ADF) test. Table 1 presents the results of the stationarity test on the original data.

Table 1. Results of the ADF Stationarity Test on Original Data

Metric	Value
ADF Statistic	-1.6417
<i>P</i> -value	0.4614
Decision	Not Stationer

As illustrated in Table 1, the outcomes of the ADF test demonstrate a statistic of -1.6417, accompanied by a *p*-value of 0.4614. When 0.05 significance level is employed, the *p*-value obtained is significantly greater than alpha. Consequently, the null hypothesis cannot be rejected. This finding suggests that the data concerning domestic tourist visits, in its original form, is non-stationary and still contains a unit root. This condition is consistent with the visual pattern in Figure 2, which shows an upward trend, particularly during the post-pandemic recovery period. It is imperative that a differencing process is applied to transform the data into a stationary form. Only then can the SARIMA model be identified and estimated.

In the context of SARIMA modelling, two distinct types of differencing can be employed: regular differencing and seasonal differencing (Montgomery et al., 2015). In view of the limited

number of observations in this study (80 monthly observations), the choice of differencing strategy must be made with great care, as each differencing process reduces the amount of data available for modelling. Consequently, the exploration was constrained to combinations of differencing that minimise data reduction, namely quarterly seasonal differencing with a lag of 3 ($D=1, s=3$) and half-yearly seasonal differencing with a lag of 6 ($D=1, s=6$). Table 2 presents a comparison is presented of the stationarity test results for these two differencing strategies.

Table 2. ADF Stationarity Test Results with Differencing

Metric	$D=1, s=3$	$D=1, s=6$
ADF Stats	-3.1500	-3.1700
<i>P</i> -value	0.0230	0.0210
Decision	Stationer	Stationer

The selection of the optimal differencing strategy is not only based on the achievement of stationarity via the ADF test, but also takes into account the clarity of patterns in the ACF and PACF plots, particularly in the seasonal component. It should be noted that SARIMA models integrate seasonal AR and MA components; therefore, the differenced data must exhibit clear and identifiable seasonal patterns at the seasonal lags. Figure 3 displays the ACF and PACF plots for both differencing results.

The seasonal differencing strategy with a lag of 6 ($D=1, s=6$) reveals an issue in the PACF plot in the form of a large spike at lag 7, which creates ambiguity in identifying the seasonal AR order (*P*) and complicates the specification of an appropriate model. Conversely, seasonal differencing with a lag of 3 ($D=1, s=3$) produces the clearest and most consistent seasonal pattern. The ACF plot shows a significant spike at lag 3, whilst the PACF plot reveals spikes at lags 3, 6, and 9, which are consistent multiples of the quarterly period. This pattern indicates the presence of a strong and regular seasonal AR structure, making it highly suitable for SARIMA modelling.

Based on the analysis, the most appropriate approach for the SARIMA model formation would be to utilise seasonal differencing with a lag of 3 ($D=1, s=3$). This method was selected on the basis that it produced the most clear, consistent, and easily interpretable seasonal pattern. Subsequent to the establishment of the differencing strategy, the selection of the SARIMA model order was conducted systematically via a grid search, utilising the AIC and BIC information criteria. The grid search was run over the range of non-seasonal orders $p \in 0,1,2$, $d \in 0,1$, $q \in 0,1,2$ and seasonal orders $P \in 0,1,2,3$, $D=1$, $Q \in 0,1,2$ with a period of 3, resulting a total of 216 candidate models being evaluated. Table 3 presents the ten best models from this grid search process.

Of all the models that were subjected to testing, SARIMA-(2,0,1)(0,1,1)[3] was determined to be the optimal model, as it exhibited the lowest AIC value of 2520.1200 and third lowest BIC of 2531.4300. The model does not require a seasonal AR component, indicating that the seasonal structure of the

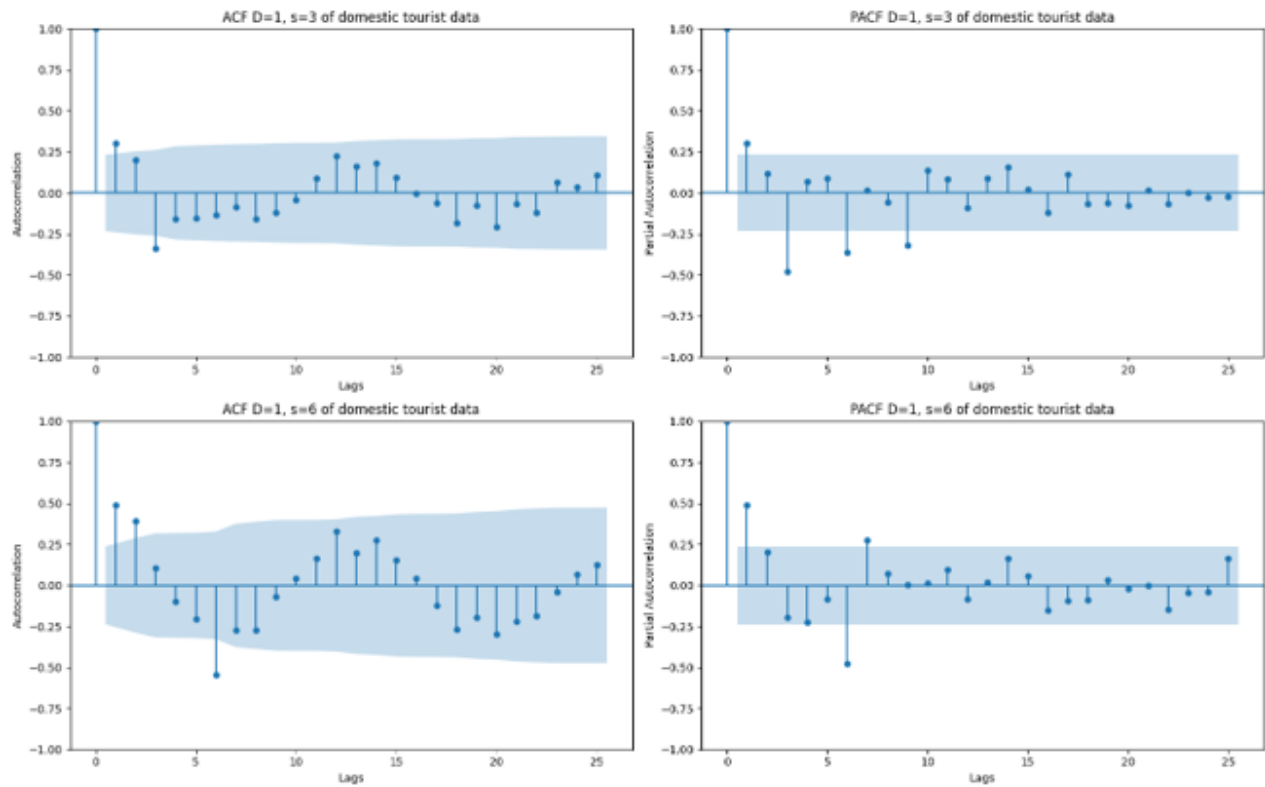


Figure 3. Comparison of ACF and PACF Plots for 2 Differencing Strategies

Table 3. Best 10 SARIMA Model Candidates Based on AIC/BIC Values

Model	AIC Values	BIC Values
SARIMA(2,0,1)(0,1,1)[3]	2520.1200	2531.4300
SARIMA(2,0,1)(0,1,2)[3]	2521.7100	2535.2900
SARIMA(2,0,1)(1,1,1)[3]	2521.9000	2535.4700
SARIMA(1,0,2)(0,1,1)[3]	2521.9900	2533.3100
SARIMA(1,0,2)(0,1,0)[3]	2522.0900	2531.1400
SARIMA(0,0,2)(0,1,1)[3]	2522.1600	2531.2100
SARIMA(0,0,2)(0,1,2)[3]	2522.2700	2533.5800
SARIMA(2,0,1)(2,1,1)[3]	2522.3200	2538.1600
SARIMA(1,0,2)(0,1,2)[3]	2522.3600	2535.9300
SARIMA(2,0,2)(0,1,0)[3]	2522.5300	2533.8400

data is sufficiently captured by a single seasonal MA component. Despite the earlier ACF/PACF analysis suggesting the potential for seasonal AR up to a lag of 9, the AIC/BIC-based approach revealed that incorporating such a component did not result in a proportional enhancement in fit relative to the model’s complexity penalty. The estimated parameters of the selected model are presented in Table 4.

As demonstrated in Table 4, the parameter estimation process results demonstrate that the AR(1) term is significant statistically, with a p-value of 0.0040, thereby indicating a strong

Table 4. Parameter Results of SARIMA(2,0,1)(0,1,1)[3]

Parameter	Coefficient	Std Err	P-value
AR(1)	0.9272	0.3200	0.0040
AR(2)	0.0636	0.3080	0.8310
MA(1)	-0.4671	0.2870	0.1040
MA-Seasonal(3)	-0.9677	0.2720	0.0000

dependence on the value from one period earlier. The seasonal MA component at lag 3 is also highly significant, with a p-value of less than 0.0010, reflecting the strong influence of the seasonal moving average component. However, the AR(2) and MA(1) parameter was determined to be statistically insignificant, yet it was retained on the basis that the AIC-based selection process had taken into account the complexity penalty. Moreover, this model was determined to provide the optimal fit when compared to all other candidates that were evaluated.

In order to validate the suitability of the selected model, a diagnostic analysis was conducted on the residuals from the SARIMA(2,0,1)(0,1,1)[3]. The analysis encompassed a residual autocorrelation test utilising the Ljung-Box test at multiple lags, in addition to normality tests employing the Jarque Bera and Shapiro Wilk tests. It is important to note that the residuals analysed begin from the fourth observation, given that the first three observations constitute the model’s initialisation period

due to the application of seasonal differencing with a lag of 3. Table 5 and Figure 4 present the complete results of the diagnostic tests and their visualisations.

Table 5. Results of the Diagnostic Tests for the SARIMA(2,0,1)(0,1,1)[3]

Test	Statistic	P-value	Summary
Ljung Box (Lag 3)	1.3594	0.7150	No Autocorrelation
Ljung Box (Lag 6)	3.9571	0.6820	No Autocorrelation
Ljung Box (Lag 9)	12.1434	0.4340	No Autocorrelation
Jarque Bera	0.8771	0.6450	Normal Residuals
Shapiro Wilk	0.9880	0.7380	Normal Residuals

The outcomes of the Ljung-Box test demonstrate the non-existence of significant autocorrelation in the residuals across all lags examined, with p-values of 0.7150 at lag 3, 0.6830 at lag 6, and 0.4340 at lag 12, all of which exceed the significance threshold of 0.05. This finding suggests that the model residuals are white noise, thereby indicating that the model has effectively captured all the patterns that can be modelled from the data. With regard to the normality tests, both the Jarque Bera (p -value = 0.6450) and Shapiro Wilk (p -value = 0.7380) tests failed to reject the null hypothesis, thus indicating that the residuals are normally distributed. This hypothesis is reinforced by a skewness value close to zero (0.2440) and low kurtosis (0.3690).

In addition to validation via residual diagnostics, this study also conducted a sensitivity analysis to test the model's stability against variations in the forecast horizon. This analysis was carried out by evaluating the performance of the SARIMA(2,0,1)(0,1,1)[3] model over four different forecast horizons, namely 3, 6, 9, and 12 periods ahead, using a fixed split approach where the training data was truncated according to the length of the horizon being tested. The results of the analysis are presented in Table 6.

Based on Table 6, the model demonstrates stable performance over short forecast horizons. At 3 periods, the model achieves a MAPE of 6.62% and an sMAPE of 6.33%, whilst at 6 periods (the primary forecast horizon in this study) the MAPE of 8.02% and sMAPE of 8.57% remain within an acceptable range. A decline in performance begins to be evident when using 9 and 12 periods, with MAPE reaching 19.48% and 20.28% respectively. This pattern is consistent with the general characteristics of ARIMA models, where forecast accuracy tends to decline as the forecast horizon increases due to the accumulation of estimation errors (Box et al., 2015). Thus, the selection of 6 periods as the forecast horizon in this study can be said to fall within a stable and reliable performance zone

for the SARIMA model used.

3.3 LSTM Model Results

Subsequent to the development of the SARIMA model, this study proceeded with the implementation of the LSTM architecture in three variants: without interpolation, with single interpolation, and with triple interpolation. A consistently set random seed was utilised to ensure the reproducibility of the results. The initial variant was constructed utilising the original data without augmentation as an evaluation baseline. The preprocessing stage entailed the normalisation of data using a MinMaxScaler within the range of (0,1). This was followed by the transformation of the data into a format suitable for supervised learning via a sliding window technique with a look-back period of 3, which was selected to align with the seasonal period in the SARIMA model. From the original 80 observations, the process yielded 77 sequence samples. These samples were then divided into 71 training samples and 6 test samples. The last six months were allocated as the test set.

The optimisation of the LSTM model architecture was conducted utilising the Random Search hyperparameter tuning technique by using the Keras Tuner library, which explored various combinations within a predefined search space. This search space encompassed the number of units in the two LSTM layers (32-128 units), the number of units in the dense layer around 16-64 units, the dropout rate from 0.0-0.5, and three learning rate optimiser (0.01, 0.001, 0.0001). The tuning process is subject to a maximum of 10 trials, with the objective function is to minimise validation loss. Early stopping is also employed, with a patience of 20 epochs in order to prevent overfitting.

The subsequent two LSTM variants were developed by implementing a data augmentation strategy that employed linear interpolation to address the limitation of the number of observations. In the variant with single interpolation, an additional data point was inserted between each pair of consecutive observations by using the formula $2n-1$. This process resulted from the original of 80 observations to a total of 159 interpolated data points, thereby effectively simulating bi-weekly data collection. The temporal scope of the data set was expanded from quarterly into termly to accommodate the augmented data frequency, using 142 training samples and 11 test samples, which is twice the number of samples by the model devoid of interpolation. Conversely, the variant with triple interpolation inserts three data points between each pair of original observations, employing the formula $4n-3$. This results in 317 data points that simulate weekly data collection. The temporal window during which data was collected was expanded from a quarterly period into yearly period. This modification produced a total of 284 training samples and 21 test samples, which is four times the amount collected during the initial phase of the study. This increase was implemented to ensure that the model is endowed with substantially more temporal information during the training process.

Despite the substantial disparities in the quantity of data

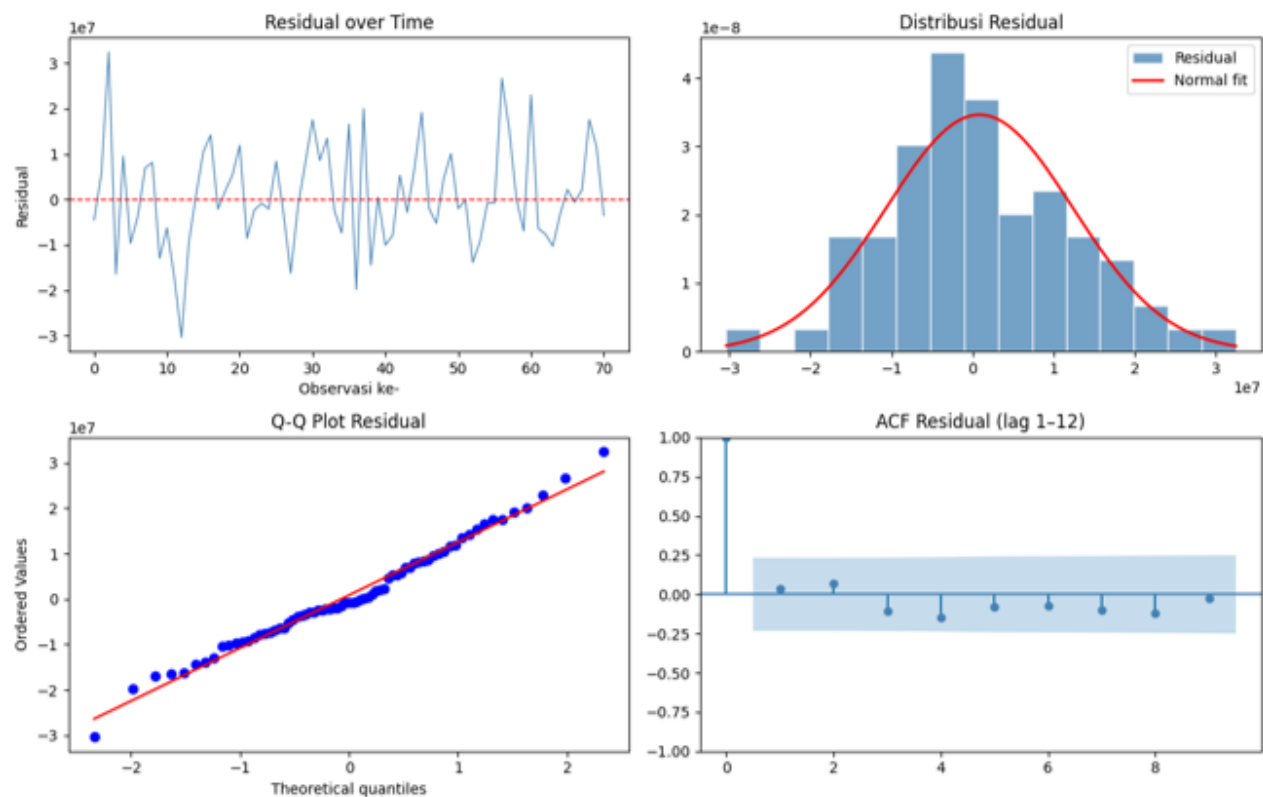


Figure 4. Diagnostic Plots for the SARIMA(2,0,1)(0,1,1)[3]: (a) Residuals Over Time, (b) Residual Distribution, (c) Q-Q Plot, (d) Residual ACF

Table 6. Results of the Sensitivity Analysis for the SARIMA(2,0,1)(0,1,1)[3]

Horizon (n)	Train Size	MAE	RMSE	MAPE	sMAPE
3	77	6,493,336	7,593,801	0.0660	0.0630
6	74	9,042,744	13,778,090	0.0800	0.0850
9	71	20,441,545	22,929,657	0.1940	0.2190
12	68	20,683,773	24,234,881	0.2020	0.2320

and the duration of the look back period, the LSTM model architecture and the hyperparameter tuning process for these two interpolation variants are analogous to the baseline model devoid of interpolation. All three models utilise the same configuration, comprising two stacked LSTM layers, a dropout layer, a dense layer with ReLU activation, and a Random Search tuning process with an identical search space. Table 7 provides a comparison of the outputs generated after the hyperparameter tuning process has been completed, highlighting the disparities between the optimal model selected from the three LSTM schemes.

As illustrated in Table 7, there are substantial discrepancies in the optimal configuration and performance of the three LSTM model variants. The interpolation strategy exerts a direct influence on validation loss, whereby the model employing triple interpolation attains the lowest validation loss

(0.0004), followed by single interpolation (0.0024), and no interpolation (0.0100). The findings demonstrate that as the quantity of training data is augmented, there is a decrease in validation loss. This suggests that the augmentation of data through interpolation is an effective strategy for enhancing the model’s capacity for learning on this type of data. However, the duration of the training process exhibited a range from 2 minutes and 6 seconds to about 3 minutes and 50 seconds. This result is partially normal because the triple interpolation model necessitated the greatest training time duration due to its processing of quadruple the amount of data.

3.4 Forecasting Performance Comparison

Subsequent to the development and optimisation of all models, a performance evaluation of the predictions was conducted on a test set comprising the final six months of the observation

Table 7. Comparison of Hyperparameters from the Optimal LSTM Model

Parameter	Interpolation		
	No	Single	Triple
Training Samples	71	142	284
Look Back Period	3	6	12
LSTM Layer 1 Units	64	80	64
LSTM Layer 2 Units	32	96	32
Dense Layer Units	32	64	32
Dropout Rate	0.2000	0.0000	0.2000
Learning Rate	0.0100	0.0100	0.0100
Best Validation Loss	0.0100	0.0024	0.0004
Total Parameters	30,401	100,481	30,401
Training Time	2m 24s	2m 06s	3m 50s

period, namely March to August 2025. In view of the disparities in the volume of output data across the LSTM variants, a filtration mechanism was implemented to guarantee a fair and valid comparison. In the single interpolation LSTM, every second data point was utilised for prediction, whereas in the triple interpolation model, every fourth data point was selected. This process ensures that all predictions are evaluated at the same time points as the original observations, thus facilitating an objective comparison between models. As illustrated in the following table and figure, a comparison is presented of the prediction results of the four models against actual data on domestic tourist visits in Indonesia.

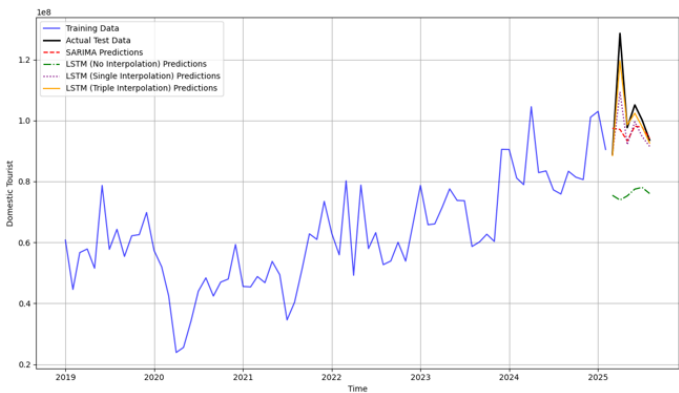


Figure 5. Comparison of Predictions vs Actual Data on Domestic Tourist Visits for SARIMA and LSTM

Table 8 and Figure 5 show a comparison of the prediction results from the four models against actual data for the period from March to August 2025. A sharp spike is visible in April 2025, when the number of tourist visits reached 128.5 million, far exceeding that of other months. This spike is not an isolated event, as three external factors coincided during that period: Eid al-Fitr 1446H fell on 31 March 2025 with an 11-day national holiday period until 7 April, a government policy

reducing domestic airfares by 13 to 14 per cent from 24 March to 7 April 2025, and official BPS data recording a 23.02 per cent year-on-year increase in domestic tourist visits during that month. It is these conditions that have created an anomaly not previously recorded in the training data, presenting a unique challenge for all models.

The SARIMA(2,0,1)(0,1,1)[3] model produces forecasts that tend to be conservative, ranging from 87 to 98 million visits, and predicts only 98.3 million for April 2025. The LSTM model without interpolation performed the worst, with very flat predictions in the range of 73-78 million, confirming that 71 training samples are insufficient for a deep learning model to learn complex temporal patterns. Conversely, the LSTM with single interpolation showed a significant improvement, with an April forecast of 109.3 million, whilst the LSTM with triple interpolation recorded the best performance, with an April forecast reaching 122.7 million visits and consistent accuracy across other months. To measure performance more objectively, a quantitative evaluation was conducted using five metrics: MAE, RMSE, MAPE, sMAPE, and Theil's U, the results of which are presented in Table 9.

As demonstrated in Table 9, it reinforces the findings of the previous visual analysis and reveals quite striking differences in the prediction accuracy of the four models. The LSTM model with triple interpolation demonstrated the best performance across all evaluation metrics, with an MAE of 2,791,815 tourists, an RMSE of 3,062,580 tourists, a MAPE of 2.6%, an sMAPE of 2.6%, and a Theil's U of 0.14. A Theil's U value well below 1.0 indicates that this model substantially outperforms the naive forecast, whilst the MAPE of 2.6% indicates a very low average prediction error despite the test data covering periods of high volatility, such as the April 2025 surge.

The LSTM with single interpolation ranks second with an MAE of 6,322,780, an RMSE of 8,797,673, a MAPE of 5.6%, an sMAPE of 5.9%, and a Theil's U of 0.42. Although its error values are approximately twice as large as those of the triple interpolation model, this performance still far surpasses the baseline model, and a Theil's U value below 1.0 confirms that this model is still capable of producing better predictions than a naive forecast.

The SARIMA(2,0,1)(0,1,1)[3] exhibited moderate performance, with an MAE of 9,042,675, an RMSE of 13,778,011, a MAPE of 8.0%, an sMAPE of 8.6%, and a Theil's U of 0.63. Despite its reduced accuracy in comparison to the two LSTM variants with interpolation, this conventional statistical model nevertheless exhibits superior performance to the LSTM without interpolation, consistently generating more accurate predictions than the naive forecast. The LSTM model, devoid of interpolation, exhibited the poorest performance among the four models, with an MAE of 26,302,000, an RMSE of 29,529,421, a MAPE of 24.6%, an sMAPE of 28.7%, and a Theil's U of 1.38. A Theil's U value exceeding 1.0 is a significant finding, as it indicates that the model performs even worse than simply using the previous period's value as a forecast. This finding serves to substantiate the assertion that the

Table 8. Comparison of Predicted Domestic Tourist Visits on the Test Data

Month	Actual Data	SARIMA	LSTM with No Int	LSTM with Single Int	LSTM with Triple Int
March 2025	88,909,481	97,453,238	75,459,320	88,766,376	88,501,296
April 2025	128,587,806	97,087,229	73,917,864	109,388,408	119,519,472
May 2025	97,673,863	93,321,589	75,257,976	92,094,248	98,569,512
June 2025	105,115,881	98,106,766	77,499,208	99,548,526	102,396,096
July 2025	100,197,984	97,878,439	78,063,328	94,756,936	97,972,712
August 2025	93,566,867	94,098,062	76,042,200	91,560,752	92,611,632

Table 9. Comparison of Model Evaluation Metrics

Model	MAE	RMSE	MAPE	sMAPE	Theil's U
SARIMA(2,0,1)(0,1,1)[3]	9,042,675	13,778,011	0.0800	0.0860	0.6338
LSTM No Interpolation	26,302,000	29,529,421	0.2460	0.2870	1.3797
LSTM Single Interpolation	6,322,780	8,797,673	0.0560	0.0590	0.4183
LSTM Triple Interpolation	2,791,815	3,062,580	0.0260	0.0260	0.1439

71 training samples in question represent a critical constraint for deep learning models in the context of learning complex temporal dependencies. Moreover, it is demonstrated that in the absence of data augmentation strategies, LSTM is unable to achieve optimal performance on limited datasets.

In addition to the accuracy evaluation, a comparison of computational efficiency was also carried out to provide a more comprehensive picture of the trade-off between performance and computational cost for each model. As illustrated in Table 10, a comparison of execution times is presented.

As demonstrated in Table 10, the SARIMA model is significantly more computationally efficient, requiring a mere 0.20 seconds to train in comparison to the 122 to 217 seconds required by the three LSTM models. This discrepancy is indicative of the fundamental divergence between analytically estimated statistical models and neural networks that have been optimised via backpropagation. The LSTM model with triple interpolation requires the most time (217.89 seconds) as it processes four times as much data. However, the resulting enhancement in accuracy, from an 8.0% MAPE in SARIMA to 2.6%, renders this trade-off worthwhile in contexts where predictive accuracy is prioritised.

The results of this study indicate that data augmentation using linear interpolation improves the performance of LSTM models for time series forecasting with few observations. The substantial improvement in performance of the triple interpolation model can be attributed to two main mechanisms. First, the increase in the length of the time series used for training data, due to the addition of interpolated data points, gives deep learning models such as LSTM more context to learn long term dependencies. Second, from the perspective of data structure analysis, the interpolation method itself further enhances the persistence of autocorrelation, which is reflected in

the augmented data having a larger ACF pattern, thus enriching the temporal signals that the LSTM can learn through the gate mechanism. Despite the alterations in data variance post-interpolation, the Levene test reveals no statistically significant disparities ($p = 0.925$), thereby suggesting that the fundamental distribution of the data remains intact.

The fact that the LSTM model without interpolation failed (Theil's U value of 1.38) proved that the 71 training samples were not enough for the LSTM architecture to converge to a meaningful pattern. It can be shown that if Theil's U is greater than 1.0, the model performs worse than a simple naive forecast. This is a common problem when training deep learning models with too little data where the model ends up learning the noise instead of the signal. This result lends support to the argument that data augmentation strategies are not optional, but a necessity, when the amount of available time-series data is extremely small.

The LSTM model with triple interpolation outperforms the SARIMA with a MAPE of 2.6% compared to 8.0%, representing a threefold improvement in accuracy. However, this improvement comes at a significantly higher computational cost of 206 seconds compared to 0.20 seconds for SARIMA, which is roughly 1215 times longer. The best model depends on the purpose of the model. SARIMA is more suitable when the model is needed for fast prediction or when there is a lack of computational power. On the contrary, the LSTM with triple interpolation is better for applications where the main target is to achieve the best prediction accuracy.

This study has some limitations that need to be considered. The evaluation was performed using only six original data points as the test set, which is due to the scarcity of data. This figure is in line with the forecast horizon considered and it is supported by the results of the sensitivity analysis, which show

Table 10. Comparison of Model Evaluation Metrics

Model	Training Time	Parameters	Train Size
SARIMA(2,0,1)(0,1,1)[3]	0.20s	5	74 obs
LSTM No Interpolation	122.22s	1,000+	71 obs
LSTM Single Interpolation	123.25s	1,000+	142 obs
LSTM Triple Interpolation	217.89s	1,000+	284 obs

the model stability in the short run. Future studies should consider increasing the size of the test set to improve the overall reliability of the evaluation.

Furthermore, the developed model does not incorporate exogenous variables that have been shown to play a significant role in tourist visitation patterns. These variables are the national holiday calendar, the price of transport tickets and macroeconomic conditions. The exclusion of these variables has a significant impact on the accuracy of all models in predicting the surge in April 2025, although the LSTM with triple interpolation managed to approximate the actual value. The augmentation strategy also employed linear interpolation, which assumes gradual change between two observation points, and may not always hold for data with sudden changes or strong non-linear patterns. Finally, the models were developed and evaluated using Indonesian domestic tourist visitation data only; therefore, generalising the findings to other countries or tourism sectors requires further validation. Further research could explore more complex augmentation methods, the inclusion of exogenous variables, and application to datasets with more diverse characteristics.

4. CONCLUSIONS

This research offers multiple contributions to the area of study. The first contribution is that it provides empirical evidence concerning how well interpolation strategies work in improving the performance of LSTM’s for tourism forecasting when there are limited amounts of available data. The second contribution is that it creates a comprehensive framework for comparing traditional statistical methods to deep learning models, along with diagnostic evaluation of SARIMA model residuals using the Ljung Box test and residual normality. The third contribution is that the introduction of the sMAPE and Theil’s U metrics to the framework provides a broader evaluation perspective on performance relative to other models. The Theil’s U value greater than 1.0 for the LSTM model without interpolation indicates data augmentation is required when working with limited datasets, rather than just being optional. The fourth contribution is that all results from this study can be replicated and built upon by follow up researchers with the use of the same random seed.

5. ACKNOWLEDGEMENT

We would like to thank Badan Pusat Statistik (BPS) Indonesia for the dataset. The authors extend the gratitude to all the

reviewers who have contributed to the peer review process of the manuscript in this issue.

REFERENCES

Adekunle, B. I., E. C. Chukwuma-Eke, E. D. Balogun, and K. O. Ogunsola (2021). Predictive Analytics for Demand Forecasting: Enhancing Business Resource Allocation Through Time Series Models. *Journal of Frontiers in Multidisciplinary Research*, **2**(01); 32–42

Ahmed, S. F., M. S. B. Alam, M. Hassan, M. R. Rozbu, T. Ish-tiak, N. Rafa, M. Mofijur, A. B. M. Shawkat Ali, and A. H. Gandomi (2023). Deep Learning Modelling Techniques: Current Progress, Applications, Advantages, and Challenges. *Artificial Intelligence Review*, **56**(11); 13521–13617

Ajobo, S., O. O. Alaba, and A. Zaenal (2024). Generalised Space-Time Seasonal Autoregressive Integrated Moving Average Seemingly Unrelated Regression Modelling of Seasonal and Non-Stationary Data. *Scientific African*, **24**; e02189

Alamineh, G. A., J. W. Hussein, Y. Endaweke, and B. Tad-desse (2023). The Local Communities’ Perceptions on the Social Impact of Tourism and Its Implication for Sustainable Development in Amhara Regional State. *Heliyon*, **9**(6)

Ali, M., R. Jain, R. Joshi, and K. D. Gautam (2025). Dynamic Bayesian KNN with Confidence Intervals: A Reliable Imputation Method for High-Dimensional and Sparse Data. *Engineering Letters*, **33**(12)

Ampountolas, A. (2021). Modeling and Forecasting Daily Hotel Demand: A Comparison Based on SARIMAX, Neural Networks, and GARCH Models. *Forecasting*, **3**(3); 580–595

Boto-García, D. and J. F. Baños Pino (2024). The Economics of Second-Home Tourism: Are There Expenditure Reallocation Effects from Accommodation Savings? *Tourism Economics*, **30**(4); 969–995

Box, G. E. P., G. M. Jenkins, G. C. Reinsel, and G. M. Ljung (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons

Cho, K., B. Van Merriënboer, Ç. Gulçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio (2014). Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 1724–1734

Choudhury, P., R. T. Allen, and M. G. Endres (2021). Machine

- Learning for Pattern Discovery in Management Research. *Strategic Management Journal*, **42**(1); 30–57
- Deif, M. A., A. Bounceur, S. Mouakket, M. A. Hafez, and M. Elhoseny (2025). A Multi-Agent Framework for Detecting and Forecasting Silent Resistance in Organizational Communication: Implications for Open Innovation Dynamics. *Journal of Open Innovation: Technology, Market, and Complexity*; 100646
- Hajisafi, A., H. Lin, S. Shaham, H. Hu, M. D. Siampou, Y.-Y. Chiang, and C. Shahabi (2023). Learning Dynamic Graphs from All Contextual Information for Accurate Point-of-Interest Visit Forecasting. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*. pages 1–12
- He, K., L. Ji, C. W. D. Wu, and K. F. G. Tso (2021). Using SARIMA–CNN–LSTM Approach to Forecast Daily Tourism Demand. *Journal of Hospitality and Tourism Management*, **49**; 25–33
- Ismail, M. T. and R. S. Al-Gounmmeen (2022). Overview of Long Memory for Economic and Financial Time Series Dataset and Related Time Series Models: A Review Study. *IAENG International Journal of Applied Mathematics*, **52**(2)
- Li, B. and J. Gao (2022). Regional Tourism Economic Forecasting Model Based on GM Grey Forecasting Method. *Mathematical Problems in Engineering*, **2022**(1); 3477246
- Lim, B. and S. Zohren (2021). Time-Series Forecasting with Deep Learning: A Survey. *Philosophical Transactions of the Royal Society A*, **379**(2194); 20200209
- Montgomery, D. C., C. L. Jennings, and M. Kulahci (2015). *Introduction to Time Series Analysis and Forecasting*. John Wiley & Sons
- Okere, E. E. and V. Balyan (2025). A Deep Learning-Based Prediction and Forecasting of Tomato Prices for the Cape Town Fresh Produce Market: A Model Comparative Analysis. *Forecasting*, **7**(2); 19
- Olukoya, O. (2023). Time Series-Based Quantitative Risk Models: Enhancing Accuracy in Forecasting and Risk Assessment. *International Journal of Computer Applications Technology and Research*, **12**(11); 29–41
- Ordóñez-Martínez, D., J. M. Seguí-Pons, and M. Ruiz-Pérez (2023). Conceptual Framework and Prospective Analysis of EU Tourism Data Spaces. *Sustainability*, **16**(1); 371
- Perez-Guerra, U. H., R. Macedo, Y. P. Manrique, E. A. Condori, H. I. Gonzáles, E. Fernández, N. Luque, M. G. Pérez-Durand, and M. García-Herreros (2023). Seasonal Autoregressive Integrated Moving Average (SARIMA) Time-Series Model for Milk Production Forecasting in Pasture-Based Dairy Cows in the Andean Highlands. *PLoS One*, **18**(11); e0288849
- Ramadan, A., D. Rantini, B. D. Putri, Y. Widiyastusi, F. Amaliyah, N. Putu, E. K. Prayascita, I. I. Hakim, Z. Salsabila, M. M. Yandra, et al. (2025a). Economic Activity as a Catalyst for Tourism Development: A Spatial Analytical Perspective from Indonesia. *IAENG International Journal of Applied Mathematics*, **56**(1); 402–414
- Ramadan, A., M. A. Syahputra, D. Rantini, R. A. Ningrum, M. N. Fakhruzzaman, A. Fajar, M. M. Yandra, N. A. Alya, M. F. A. Hillaby, A. Sesay, et al. (2025b). Predicting Surabaya's Rainfall: A Comparative Study of Naïve Bayes, K-Nearest Neighbor, and Random Forest. *Data and Metadata*, **4**; 1075–1075
- Rantini, D., N. Iriawan, and Irhamah (2021). On the Reversible Jump Markov Chain Monte Carlo (RJMCMC) Algorithm for Extreme Value Mixture Distribution as a Location-Scale Transformation of the Weibull Distribution. *Applied Sciences*, **11**(16); 7343
- Rantini, D., R. Kruba, Y. J. Wijaya, M. M. Yandra, H. Rahman, A. Ramadan, F. Othman, et al. (2026). Predicting Upwelling Dynamics in the South Sea of Java, Indonesia: A Deep Learning Approach with ConvLSTM and 3D-CNN. *MethodsX*; 103802
- Saputra, D. E., D. Suandi, E. Tanuar, B. Siswanto, A. H. Rangkuti, N. W. Prasetya, S. E. D. Putra, P. D. Kusuma, et al. (2025). Comparative Analysis of Machine Learning Models for Predicting Water Salinity Levels. *IAENG International Journal of Computer Science*, **52**(12)
- Schaffer, A. L., T. A. Dobbins, and S.-A. Pearson (2021). Interrupted Time Series Analysis Using Autoregressive Integrated Moving Average (ARIMA) Models: A Guide for Evaluating Large-Scale Health Interventions. *BMC Medical Research Methodology*, **21**(1); 58
- Suprihadi, E., N. Danila, and Z. Ali (2025). Enhancing Financial Product Forecasting Accuracy Using EMD and Feature Selection with Ensemble Models. *Journal of Open Innovation: Technology, Market, and Complexity*, **11**(2); 100531
- Taylor, S. J. and B. Letham (2018). Forecasting at Scale. *The American Statistician*, **72**(1); 37–45
- Torres, J. F., D. Hadjout, A. Sebaa, F. Martínez-Álvarez, and A. Troncoso (2021). Deep Learning for Time Series Forecasting: A Survey. *Big Data*, **9**(1); 3–21
- Vu, K. and K. Hartley (2022). Drivers of Growth and Catch-Up in the Tourism Sector of Industrialized Economies. *Journal of Travel Research*, **61**(5); 1156–1172
- Wu, D. C., S. Zhong, J. Wu, and H. Song (2025). Tourism and Hospitality Forecasting with Big Data: A Systematic Review of the Literature. *Journal of Hospitality & Tourism Research*, **49**(3); 615–634