Journal of Natural Sciences and Mathematics Research
J. Nat. Scien. & Math. Res. Vol. 12 No. 1 (2026) 13-22
ISSN: 2460-4453 (Online); 2614-6487 (Printed)
Available online at http://journal.walisongo.ac.id/index.php/jnsmr

Forecasting boarding passenger volume at PT KAI Daop 5 Purwokerto using SARIMA and LSTM

Fajar Tri Wahyuni[1], Atika Ratna Dewi[1*]
[1] Data Science Study Program, Faculty of Informatics, Telkom University, Indonesia.

* Corresponding author's e-mail: atikad@telkomuniversity.ac.id

ABSTRACT
Railway transportation plays a vital role in supporting public mobility, requiring accurate passenger demand forecasting to assist operational planning, particularly during seasonal peak periods. This study aims to apply and compare the Seasonal Autoregressive Integrated Moving Average (SARIMA) and Long Short-Term Memory (LSTM) models for daily passenger boarding volume forecasting at PT Kereta Api Indonesia (KAI) Daop 5 Purwokerto. The dataset consists of daily passenger boarding records from January 1, 2023, to July 15, 2025, treated as a time series. For the SARIMA model, data preprocessing included log transformation, differencing, and seasonal differencing to achieve stationarity. Meanwhile, the LSTM model employed data normalization and sequence construction to capture nonlinear temporal dependencies. The experimental results show that the best SARIMA configuration is SARIMA(3,1,3)(0,1,1)7, while the optimal LSTM architecture uses a window size of 30 with two hidden layers of 64 and 32 units trained for 50 epochs. Performance evaluation on the test data indicates that the LSTM model outperforms SARIMA, achieving lower error values (RMSE = 2632.94; MAPE = 13.2%) compared to SARIMA (RMSE = 3399.48; MAPE = 17.9%). These findings indicate that, in this case study, the LSTM model achieved better forecasting performance than the SARIMA model on the selected hold-out test dataset (882 training observations and 45 testing observations). The obtained MAPE of 13.2% indicates good forecasting accuracy according to the adopted accuracy scale, suggesting that LSTM is a promising approach for supporting passenger demand forecasting at PT KAI Daop 5 Purwokerto.

Keywords:
SARIMA; LSTM; passenger volume forecasting; time series.

Introduction

Railway transportation is one of the mass transportation modes widely used by the Indonesian public because it is considered more efficient, safe, and comfortable. In July 2025, the number of departing train passengers reached 50.1 million, an increase of 9.27% from July 2024 (Badan Pusat Statistik, 2025). The growth in passenger numbers is also evident cumulatively, as during the first semester of 2025, PT Kereta Api Indonesia served approximately 240.90 million passengers, an increase of 8.90% compared to the same period in the previous year (Muhamad, 2025). The increasing passenger volume underscores PT KAI's strategic role in meeting transportation demand nationwide and emphasises the need for effective operational management. Daop 5 Purwokerto is one of the vital operational units that support community mobility in western Central Java.

Daop 5 Purwokerto's need for capacity planning and train allocation makes predicting passenger volume an important variable. This variable directly represents the demand for train services in the operational area. Passenger volume data is historical time-based data whose movements show significant fluctuations, with demand spikes often occurring during specific periods such as national holidays, weekends, or vacation seasons. The accuracy of passenger volume predictions is crucial for efficient train allocation, thereby minimising unnecessary operational costs and ensuring service capacity is ready to maintain customer satisfaction during demand surges.

Traditional statistical time-series models, such as the Seasonal Autoregressive Integrated Moving Average (SARIMA) model, have been widely applied for passenger demand forecasting due to their strong theoretical foundation and interpretability. SARIMA is particularly effective in modelling linear relationships and seasonal patterns commonly observed in transportation data. In a previous study, it was used to predict the number of foreign tourists in Bali, yielding excellent results with a MAPE of 4.97% (Prianda & Widodo, 2021). However, its reliance on linear assumptions may limit its ability to model the more complex, nonlinear dynamics of passenger demand behaviour. In contrast, recent advancements in deep learning have introduced more flexible approaches for time series forecasting. One such approach is the Long Short-Term Memory (LSTM) neural network, which is specifically designed to capture long-term dependencies and nonlinear relationships in sequential data. LSTM was used in a previous study to predict wind speed in Kupang, achieving a MAPE of 19.40%, which is better than the compared model, ARIMA (Magfirrah et al., 2024).

Although both SARIMA and LSTM have been widely used in time-series forecasting, comparative studies evaluating their performance under consistent experimental settings, especially at the regional railway operational level, remain limited. In the context of PT KAI Daop 5 Purwokerto, empirical evidence comparing these two approaches using real operational data is still scarce. This gap highlights the need for a systematic comparison to determine the most suitable forecasting method to support regional railway operations. Therefore, this study aims to compare the forecasting performance of SARIMA and LSTM models in predicting daily passenger boarding volume at PT KAI Daop 5 Purwokerto. The comparison is conducted using the same dataset, consistent evaluation metrics, and a grid-search-based hyperparameter-tuning process for the LSTM model to ensure fairness and reliability. The findings of this study are expected to provide practical insights for railway operators in selecting appropriate forecasting models to support data-driven operational planning and decision-making.

Methods

Analysis Prosedures

The analysis procedure consisted of four main stages: data preparation, modelling, evaluation, and comparative analysis. Two forecasting approaches were applied and compared, namely the SARIMA model and the LSTM neural network. SARIMA was selected as a representative statistical forecasting method due to its strong theoretical basis and ability to model linear and seasonal time-series patterns. In contrast, LSTM was chosen as a deep learning-based approach capable of capturing nonlinear relationships and long-term dependencies in sequential data. The comparison between these two models aims to assess their suitability for passenger demand forecasting at the regional railway operational level. The best model will be selected by the smallest RMSE and MAPE values (Magfirrah et al., 2024).

Dataset

This study uses historical daily passenger data obtained from PT KAI Daop 5 Purwokerto. The dataset consists of 927 observations covering the period from January 1, 2023, to July 15, 2025. Each observation contains two main attributes: the date, which serves as a time index, and volume, which is the number of boarding passengers. The dataset forms a univariate time series suitable for time series forecasting analysis. The dataset was divided into training and testing subsets. The training dataset includes observations from January 1, 2023, to May 31, 2025, while the testing dataset covers the period from June 1, 2025, to July 15, 2025. Table 1 presents the division of the testing and training datasets.

Table 1. Proportion of Data Splitting

Training	882
Testing	45
Total	927

SARIMA Model

The SARIMA model is an extension of the ARIMA model designed to handle seasonal time-series data by combining non-seasonal and seasonal components. The SARIMA model is denoted as SARIMA (p, d, q) (P, D, Q)1, where (p, d, q) represent the non-seasonal AR, differencing, and MA orders. (P, D, Q) represent the corresponding seasonal components, and s represents the seasonal period (Santoso & Widodo, 2024).

$$\phi_p(B)\phi_P(B^s)(1-B)^d(1-B^S)^D X_t = \theta_q(B)\theta_Q(B^S)et \quad (1)$$

where:

X_t	: value of variable X in period t	θ_Q	: seasonal MA parameter
e_t	: error in period t	S	: number of seasonal periods
ϕ_p	: non-seasonal AR parameter	d	: non-seasonal differencing order
ϕ_P	: seasonal AR parameter	D	: seasonal differencing order
θ_q	: non-seasonal MA parameter		

Before obtaining the model, the mean stationarity test was conducted using ADF and KPSS, and the variance stationarity test was conducted using the Box-Cox lambda value (Sirisha et al., 2022). The selected model candidates were evaluated based on AIC and parameter significance (p-value < 0.05) (Magfirrah et al., 2024) (Wang, 2023). The selected model was then validated using a residual diagnostic test to ensure the absence of residual autocorrelation (Wang, 2023).

LSTM Model

Figure 1 show the Long Short-Term Memory (LSTM) model is a specialised RNN architecture (Mahjoub et al., 2022), modified to overcome the main weakness of RNNs, namely the problem of vanishing or exploding gradients (Yao, 2024). LSTM is widely used across various fields, including language modelling, voice recognition, and time-series analysis in finance (Chaiko, 2024). Its ability to use information from previous inputs and outputs allows LSTM to understand patterns and temporal changes in data better.

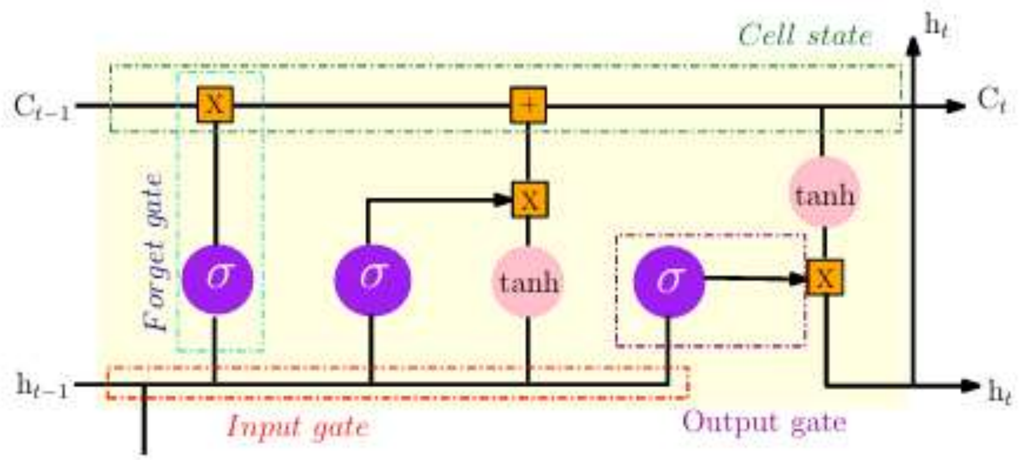

Figure 1. LSTM Architecture (Mahjoub et al., 2022)

The process within a single LSTM unit is controlled by three main gates that regulate cell-state updates and the hidden state. Each gate produces a value between 0 and 1 using the sigmoid activation function, which indicates how much information will be passed on. The tanh activation is used to generate new memory candidates with a range of values between -1 and 1 (Mahjoub et al., 2022) (Wang, 2023). The calculations for LSTM are as follows (Mahjoub et al., 2022):

$$f_t = \sigma(w_f[h_{t-1}, X_t] + b_f \quad (2)$$
$$i_t = \sigma(w_i[h_{t-1}, X_t] + b_i \quad (3)$$
$$O_t = \sigma(w_o[h_{t-1}, X_t] + b_o \quad (4)$$
$$a_t = tanh(w_a[h_{t-1}, X_t] + b_a \quad (5)$$

$$c_t = f_t * c_{t-1} + i_t * a_t \tag{6}$$
$$h_t = O_t * \tanh(c_t) \tag{7}$$

where:

f_t : The output from forget gate, determines how much of the old memory C_{t-1} needs to be retained

i_t : The output from input gate, determines how much new information is added to the memory

a_t : candidate memory, the result of transforming the current input value using the tanh function

c_t : new cell state, the result of combining old memory and new information

h_t : output from the output gate, determining which value will be the output to the hidden state

h_t : new hidden state, which becomes the output at time step t

Hyperparameter Tuning

To obtain the optimal LSTM configuration, a grid search-based hyperparameter tuning approach was employed (Sanhatham, 2024). Several key hyperparameter combinations were evaluated, including input window sizes of 15 and 30 time steps. The LSTM unit configurations consisted of a single-layer model with 32 units and a stacked architecture with 64 and 32 units, with training epochs of 30 and 50, respectively. Each hyperparameter combination was trained and evaluated using the same training and testing data split to ensure consistency. The optimal LSTM configuration was selected based on the lowest prediction error on the validation dataset. This systematic tuning process ensures the reliability of the LSTM model and allows for a fair comparison with the SARIMA model. The hyperparameter combinations were selected based on the predefined hold-out experimental setting. Due to the limited size of the available dataset, a separate validation dataset was not used. Therefore, the reported results should be interpreted within the context of this experimental design. Future studies should employ an independent validation dataset or rolling-origin validation to obtain more robust hyperparameter selection.

Model Accuracy

The model accuracy evaluation stage is conducted to identify the most suitable architecture or parameter configuration and to validate prediction performance. The evaluation is performed using the MAPE (mean absolute percentage error) method and RMSE (root mean square error), with RMSE being one of the most widely used evaluation metrics. Technically, RMSE only takes the square root of the mean squared error. However, its impact is very significant because the double figures produced by MSE return to normal with RMSE (Saputra & Kristiyanti, 2022). A lower RMSE value indicates a more accurate and robust model (Do et al., 2022). The RMSE is given by Equation 8.

$$RMSE = \sqrt{\sum \frac{(\hat{Y} - Y)^2}{n}} \tag{8}$$

where $\hat{Y}$ is the predicted value, Y is the actual value, and n is the number of data points.

MAPE analysis is used to measure the model's predictive accuracy (Saputra & Kristiyanti, 2022). The equation of the RMSE is expressed in Equation 9.

$$MAPE = \frac{1}{N}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \times 100\% \tag{9}$$

where n denotes the total number of observations, Y_i represents the actual data value of the i, and $\hat{Y}_i$ is the predicted value for i.

In MAPE analysis, a value <10% indicates very high accuracy, values ranging from 11%-20% are classified as accurate, values between 12% and 50% are categorized as moderately accurate, and values >50% indicates low accuracy.

Comparative Analysis

The final stage of this study involved a comparative analysis of the SARIMA and LSTM models using the selected evaluation metrics. Both models were trained and tested under comparable conditions to minimise experimental bias. The comparison results were used to identify the most suitable forecasting approach for predicting boarding passenger volume at PT KAI Daop 5 Purwokerto. This methodological design ensures the suitability, clarity, and validity of the results. It supports the objective evaluation of statistical and deep learning-based forecasting models in a real-world railway operational context.

Results and Discussions

Data Exploration

Exploratory data analysis was used to observe passenger boarding patterns at PT KAI Daop 5 Purwokerto from 1 January, 2023 to 15 July, 2025. The daily time series data are illustrated in Figure 2.

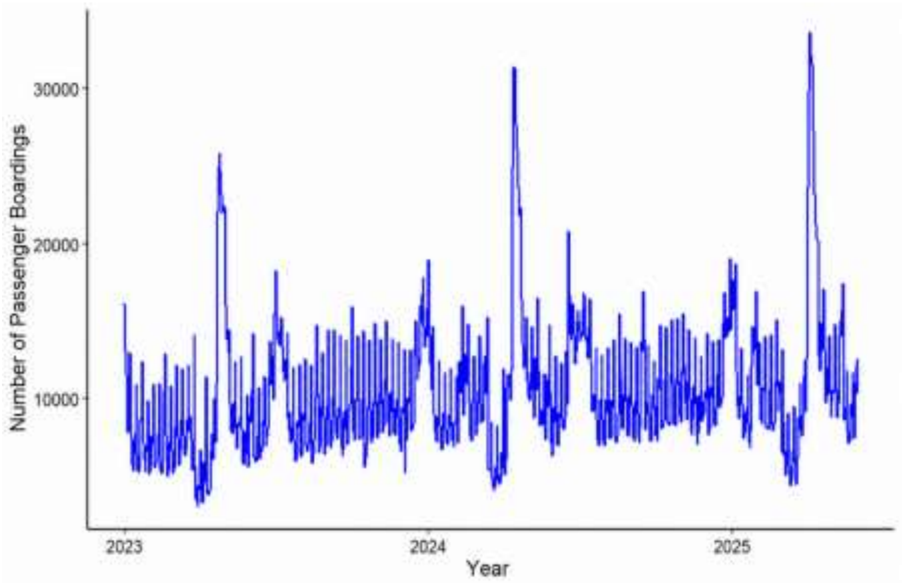

Figure 2. Time series plot

Based on Figure 2, there are three significant spikes in passenger volume. These spikes occurred in April 2023, April 2024, and April 2025, which are likely related to the Eid al-Fitr holiday period. The available data also show a weekly seasonal pattern, with passenger volume tending to be higher on weekends than on regular weekdays. Passenger data for PT KAI Daop 5 Purwokerto shows that the lowest number of passengers was 3045, while the highest reached 33535.

SARIMA Model

The first step before building a SARIMA model is to test the stationarity of the data. Box-Cox lambda, ADF, and KPSS tests are employed to evaluate data stationarity.

Table 2. Lambda Box-Cox, ADF, and KPSS tests on training data

Measure	Value	Interpretation
Box-Cox Lambda	0.114	Not Stationary
ADF	0.01	Stationary
KPSS	0.01	Not Stationary

Based on Table 2, the Box-Cox lambda value obtained is 0.114. Ideally, the lambda value should be close to 1 to indicate that the data is stationary based on the variance. Because the value obtained is far from 1, data transformation steps are applied. In addition, the KPSS test results indicate the data are not stationary. Therefore, after the transformation, differencing was performed to achieve stationarity. The stationarity test results after this process are shown in Table 3.

Table 3. Stationarity after log-transform and differencing

Test	p-value	Description
ADF	0.01	Stationary
KPSS	0.1	Stationary

After the first differencing (d=1), the KPSS test produced a p-value of 0.1 (> 0.05). This means we fail to reject H0, so we conclude that the data are stationary with respect to the mean. Although it is stationary in terms of mean, the ACF and PACF plots were examined to detect seasonal patterns.

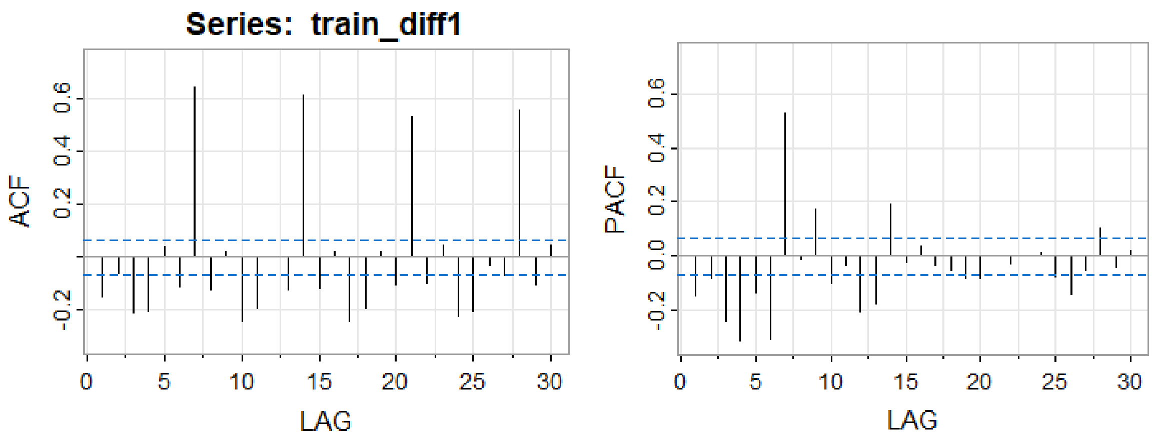

Figure 3. ACF and PACF after transformation and differencing 1

Figure 3 shows a significant and periodic spike in lags that are multiples of 7 (i.e., 7, 14, 21, 28, etc.). This confirms the existence of a strong weekly seasonal pattern in the data. To address this seasonality, seasonal differencing (D = 1) with frequency m = 7 was performed. The ACF/PACF graph after seasonal differencing is shown in Figure 4.

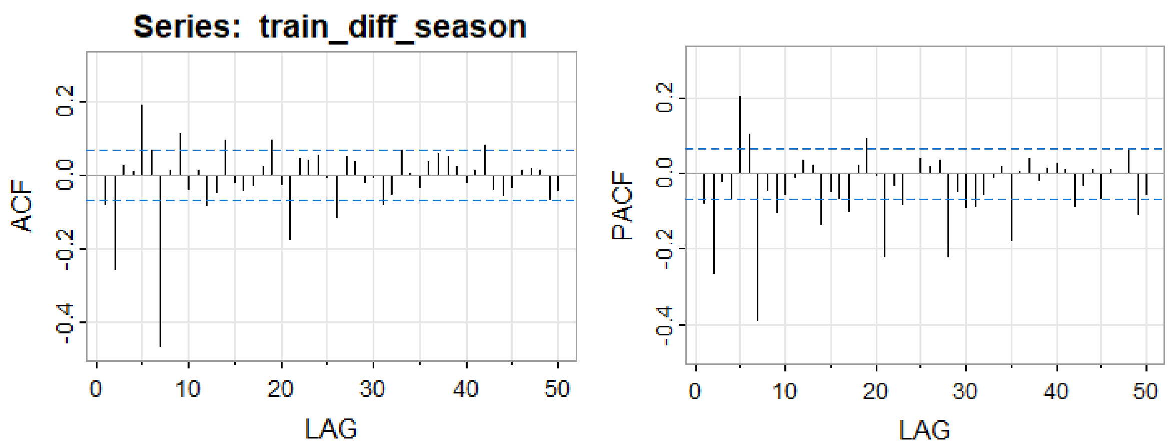

Figure 4. ACF and PACF after differencing 1 seasonal

After verifying the stationarity and seasonal identification, the next step is to identify candidate models. Three models were examined, namely (1,1,5) (0,1,1)7, (3,1,3) (0,1,1)7, and (3,1,4) (0,1,1)7. Table 4 shows the parameter estimation and AIC values for the three candidate models.

Table 4. SARIMA candidates

Model	Parameter	Coefficient	p-value	AIC
SARIMA (3,1,4) (0,1,1)7	AR (1)	0.6297	0.0187	-720.6598
	AR (2)	0.2591	0.0152	
	AR (3)	-0.581	0.0000	
	MA (1)	-0.799	0.0040	
	MA (2)	-0.353	0.0244	
	MA (3)	0.8239	0.0000	
	MA (4)	-0.114	0.2881	
	SMA (1)	-0.965	0.0000	
SARIMA (3,1,3) (0,1,1)7	AR (1)	0.3800	0.0000	-720.5469
	AR (2)	0.3474	0.0000	
	AR (3)	-0.485	0.0000	
	MA (1)	-0.530	0.0000	
	MA (2)	-0.510	0.0000	
	MA (3)	0.6667	0.0000	
	SMA (1)	-0.9648	0.0000	
SARIMA (1,1,5) (0,1,1)7	AR (1)	-0.627	0.0000	-719.1637
	MA (1)	0.4743	0.0000	
	MA (2)	-0.322	0.0000	
	MA (3)	-0.117	0.0087	
	MA (4)	0.0644	0.1000	
	MA (5)	0.1500	0.0000	
	SMA (1)	-0.926	0.0000	

Based on Table 4, although SARIMA (3,1,3) (0,1,1)7 produced a slightly lower AIC value, one of its parameters (MA (4)) was not statistically significant ($p = 0.288$). Therefore, SARIMA (3,1,3) (0,1,1)7 was selected because all estimated parameters were statistically significant and it maintained a comparable AIC value. After obtaining the best model, a diagnostic test will be performed to assess its feasibility. A random residual test will be performed using the Ljung-Box test. The Ljung-Box test result for this model is 0.07657. With a p-value > 0.05, this indicates that the (3,1,3) (0,1,1)7 model meets the assumption of sample randomness.

LSTM Model

Before creating the LSTM model, minmax scaling was applied to normalise the data to the (0,1) range. The modelling process involved systematic exploration of several hyperparameter combinations to obtain the best model configuration. The initial hyperparameter combinations tested are presented in Table 5.

Table 5. Parameter

Parameter	Value
Window Size	15, 30
Unit	32, (64, 32)
Epoch	30, 50

The hyperparameters evaluated include the window size, the number of LSTM units, and the number of training epochs. The LSTM architecture consists of one or two LSTM layers followed by a dense output layer with a single neuron and linear activation. The model is trained using the Adam optimiser with mean squared error loss, a batch size of 32, and early stopping to prevent overfitting. Model performance was evaluated using RMSE and MAPE. Based on the grid search results, the optimal configuration was obtained with a window size of 30, two hidden LSTM layers with 64 and 32 units, and 50 training epochs, which achieved the lowest RMSE and MAPE values.

Model Performance Comparison

In this phase, the SARIMA and LSTM prediction results are evaluated against the observed data using MAPE and RMSE (Table 6).

Table 6. Model evaluation

Model	RMSE train	MAPE train	RMSE test	MAPE test
SARIMA	2103.3	10.51%	3399.48	17.9%
LSTM	2082.1	16.81%	2632.94	13.2%

During training, the SARIMA model achieved a lower MAPE, whereas the LSTM model obtained a slightly lower RMSE. However, on the independent test set, LSTM outperformed SARIMA in both RMSE and MAPE, indicating better generalisation. LSTM was better at capturing patterns of fluctuation in daily passenger volume. The evaluation results of the SARIMA and LSTM models are compared in Figures 5 and 6.

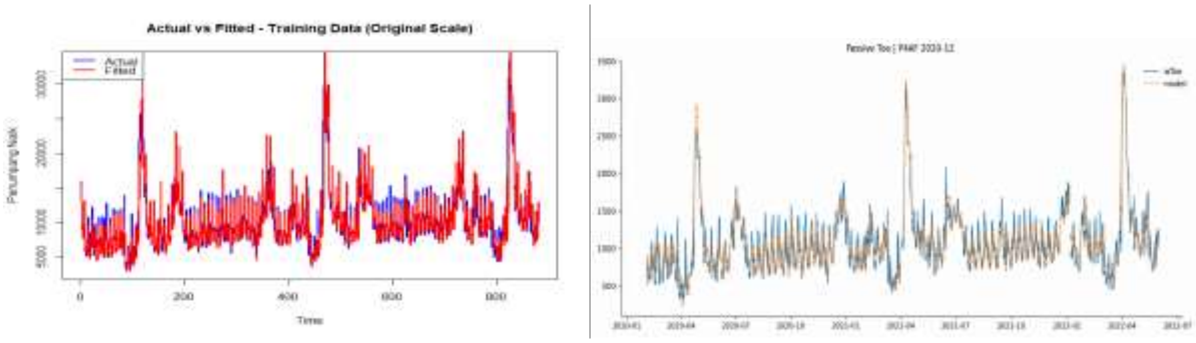

Figure 5. Plot predicted and actual data train SARIMA (left) and LSTM (right)

Figure 5 compares the actual and predicted values on the training dataset for both the SARIMA and LSTM models. Both models capture the overall temporal pattern of the training data, indicating they fit the observed data reasonably well. However, overfitting cannot be assessed solely from training performance. Instead, it should be evaluated by comparing model performance on the training and testing datasets. As shown in Table 6, the LSTM model exhibited a smaller degradation in prediction accuracy from training to testing than the SARIMA model, particularly in terms of RMSE. Although SARIMA achieved a lower training MAPE, LSTM consistently outperformed SARIMA on the test set, suggesting better generalisation under the adopted evaluation setting.

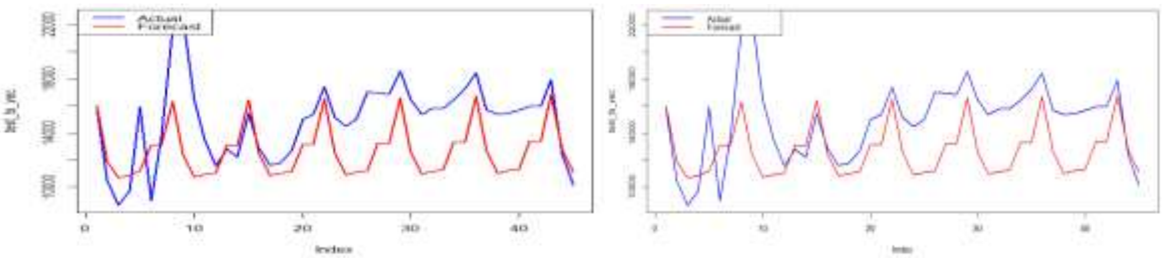

Figure 6. Plot predicted and actual data test SARIMA (left) and LSTM (right)

Figure 6 compares actual data with predictions from both models. The predicted values from the LSTM model are very close to the actual data, indicating that the model performs well. This conclusion is supported by the RMSE and MAPE values for the LSTM test data, which are the lowest among the evaluation results presented in Table 6.

Conclusion

Based on the results of forecasting the daily increase in passenger volume at PT KAI Daop 5 Purwokerto, the LSTM model shows better predictive performance than the SARIMA model. This is evidenced by lower error values in the test dataset, where LSTM achieved an RMSE of 2632.94 and a MAPE of 13.2%, while SARIMA recorded an RMSE of 3399.48 and a MAPE of 17.9%. Although both models showed comparable performance during training, the higher test accuracy of LSTM indicates stronger generalisation capabilities. Within the scope of this case study and the adopted experimental setting, the LSTM model demonstrated better predictive performance than SARIMA on the testing dataset. Nevertheless, further studies involving longer observation periods, multiple evaluation windows, and additional baseline models are required before broader conclusions can be drawn.

Acknowledgments

The authors would like to thank all reviewers who have provided suggestions to improve this article.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

Badan Pusat Statistik. (2025). *Siaran Pers 1 September 2025*.

Chaiko, V. (2024). Comparing ChatGPT And LSTM In Predicting Changes In Quarterly Financial Metrics. *Business & Management Compass*, *68*(2), 35–45. https://doi.org/10.56065/crmc1136

Do, P., Chow, C. W. K., Rameezdeen, R., & Gorjian, N. (2022). Wastewater Inflow Time Series Forecasting at Low Temporal Resolution Using SARIMA Model: a Case Study in South Australia. *Environmental Science and Pollution Research*, *29*(47), 70984–70999. https://doi.org/10.1007/s11356-022-20777-y

Magfirrah, I., Ilma, M., Notodiputro, K. A., Angraini, Y., & Mualifah, L. N. A. (2024). Comparative Analysis of ARIMA and LSTM for Forecasting Maximum Wind Speed in Kupang City, East Nusa Tenggara. *Jambura Journal of Mathematics*, *6*(2), 169–175. https://doi.org/10.37905/jjom.v6i2.25834

Mahjoub, S., Chrifi-Alaoui, L., Marhic, B., & Delahoche, L. (2022). Predicting Energy Consumption Using LSTM, Multi-Layer GRU and Drop-GRU Neural Networks. *Sensors*, *22*(11). https://doi.org/10.3390/s22114062

Muhamad, N. (2025). *KAI Layani 240 Juta Penumpang pada Semester I 2025*. https://databoks.katadata.co.id/transportasi-logistik/statistik/686f2ab013120/kai-layani-240-juta-penumpang-pada-semester-i-2025

Prianda, B. G., & Widodo, E. (2021). Comparison of Seasonal ARIMA and Extreme Learning Machine Methods in Forecasting on the Number of Foreign Tourist to Bali. *Barekeng*, *15*(4), 639–650. https://doi.org/10.30598/barekengvol15iss4pp639-650

Sanhatham, S. (2024). Stock Price Prediction of Thai Oil Public Company Limited (TOP.BK) Using LSTM Model with Grid Search Hyperparameter Optimization. *International Journal of Artificial Intelligent and Informatics*, *2*(1), 33–40. https://doi.org/10.33292/ijarlit

Santoso, A. B., & Widodo, T. (2024). Predicting the Number of Forest and Land Fire Hotspot Occurrences Using the ARIMA and SARIMA Methods. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *13*(1), 119–129. https://doi.org/10.32736/sisfokom.v13i1.2018

Saputra, I., & Kristiyanti, D. A. (2022). *Machine Learning untuk Pemula*. INFORMATIKA.

Sirisha, U. M., Belavagi, M. C., & Attigeri, G. (2022). Profit Prediction Using ARIMA, SARIMA and LSTM Models in Time Series Forecasting: A Comparison. *IEEE Access*, *10*, 124715–124727. https://doi.org/10.1109/ACCESS.2022.3224938

Wang, D. (2023). Prediction Model of Gold Yield Time Series Based on ARIMA and LSTM. *BCP Business & Management*, *38*, 2479–2489. https://doi.org/10.54691/bcpbm.v38i.4123

Yao, Y. (2024). Predicting Stock Prices Using The RF-LSTM Combination Model. *HSET*, *111*, 272–281. https://doi.org/10.54097/ptm7gz19