International Journal on Advanced Technology, Engineering, and Information System
https://ojs.transpublika.com/index.php/IJATEIS
Online ISSN 2828-5425

https://doi.org/10.55047/ijateis.v4i3.1881

Original Article

Forecasting Tourism Visitor Numbers Using a Recurrent Neural Network with a Long Short-Term Memory Algorithm

Ibnu Fallah Rosyadi[1*], Nurul Arifin Subandi[2], Rusdah[3]

[1-3]Master Program in Computer Science, Faculty of Information Technology, Universitas Budi Luhur, Indonesia
Email: [1)] ibnufallah@gmail.com, [2)] arifins.nurul@gmail.com, [3)] rusdah@budiluhur.ac.id

Received : 10 July - 2025
Accepted : 12 August - 2025
Published online : 14 August - 2025

Abstract

Accurate forecasting of visitor numbers is essential in tourism management to ensure service quality and visitor satisfaction, especially during peak seasons such as holidays and weekends. This study addresses the lack of a predictive tool at PT Taman Impian Jaya Ancol (TIJA), a major recreational destination in Indonesia, by developing a forecasting model for visitor numbers. The research utilized monthly time series data of visitor numbers from January 2012 to December 2022. A Deep Learning approach was applied using the Recurrent Neural Network (RNN) architecture with the Long Short-Term Memory (LSTM) algorithm. The dataset was split with an 80:20 ratio for training and testing, normalized using the RobustScaler technique, and optimized with the ADAM optimizer. The model achieved a minimum Mean Squared Error (MSE) of 0.3095 and a prediction accuracy of 94.85%. These results indicate that the LSTM model can effectively predict visitor trends. The findings are expected to support TIJA and other tourism operators in preparing resources and facilities in advance, improving operational planning, and enhancing the overall visitor experience.

Keywords: Deep Learning, Long Short-Term Memory, Recurrent Neural Network, Forecasting, Time Series.

1. Introduction

PT Pembangunan Jaya Ancol Tbk (PJA) is a publicly listed company operating in tourism, real estate, and trade & services sectors. Its tourism operations are managed by the subsidiary PT Taman Impian Jaya Ancol (TIJA), which oversees major attractions such as Dunia Fantasi, SeaWorld, Ocean Dream Samudra, Atlantis Water Adventure, and Hotel Putri Duyung. TIJA also manages various supporting businesses including food & beverages, merchandise, and transportation services. These recreational facilities have long served as a cornerstone of domestic tourism in Jakarta. The company's official profile and offerings are accessible via its website: www.ancol.com. In the post-pandemic era, digital engagement particularly through platforms like TikTok has significantly influenced tourists' decision-making processes, revitalizing interest in visiting destinations like Ancol (Oktavia et al., 2023). This trend highlights the growing importance of understanding visitor behavior patterns in response to digital and contextual triggers.

Prior to the Covid-19 pandemic, the company's performance at the end of 2019 was as illustrated in Figure 1 below.

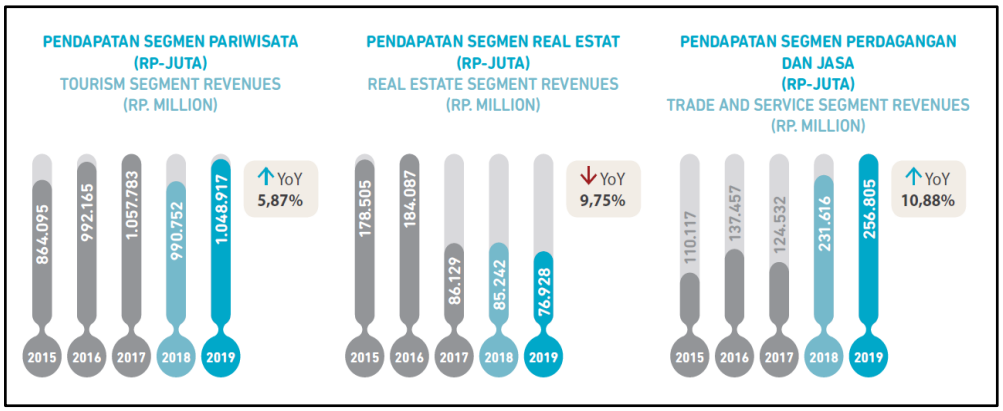

Figure 1. Annual Report 2019

Based on the figure 1, it can be seen that the total business revenue amounted to IDR 1.38 trillion, comprising the tourism segment at IDR 1 trillion (75.86%), the real estate segment at IDR 76.9 billion (5.56%), and the trade and services segment at IDR 256.8 billion (18.57%). A more detailed breakdown shows that the primary source of revenue within the tourism segment comes from ticket sales to the various recreational units mentioned earlier. This indicates a direct correlation between the number of visitors and the revenue generated from the tourism segment. Therefore, information related to visitor numbers becomes highly significant, such as reports on visitor counts over specific time periods, reports by recreational units, visitor trend reports, and others. A study by (Yamin et al., 2020) found that *"the number of tourist visits and hotel stays had a significant influence on tourism sector revenue and economic growth in Indonesia"*.

However, the current reporting system at TIJA primarily relies on descriptive historical data rather than predictive analytics. This poses a technical problem: without an accurate forecasting mechanism, the company faces challenges in efficiently planning human resources (e.g., security, cleaning, service staff) and operational logistics (e.g., shuttle services, ride maintenance, digital infrastructure). Forecasting models can greatly enhance strategic planning and resource readiness, especially during peak periods. Recent research emphasizes the value of forecasting tourism volumes using machine learning models that integrate historical and contextual data to optimize operational performance (Bollenbach et al., 2024).

The urgency of this issue becomes apparent when considering public safety responsibilities shared with local police, particularly during high-traffic events such as Eid holidays or year-end celebrations. Inadequate visitor management can lead to overcrowding, service failure, and security vulnerabilities.

Despite the abundance of tourism studies, a clear research gap remains in the integration of predictive analytics and contextual variables to forecast daily visitor numbers in the context of urban recreation areas like Ancol. Existing models tend to focus on long-term tourism projections or macroeconomic trends, leaving a gap in short-term, high-accuracy forecasting tailored for operational decision-making at the unit level.

The novelty of this study lies in the development of a forecasting model using a machine learning approach that incorporates internal and external variables, such as historical footfall data, weather conditions, holiday calendars, and social media sentiment. This approach is

expected to provide more dynamic, real-time, and operationally useful insights than traditional statistical models.

Therefore, the objective of this research is to design and evaluate a predictive model for daily visitor numbers at Taman Impian Jaya Ancol using machine learning techniques, aiming to support decision-making for resource allocation, service optimization, and public safety preparedness.

2. Literature Review

2.1. Forecasting

Forecasting is defined as the process of estimating future needs in terms of quantity, quality, time, and location to meet the demand for goods or services (Luthfianto, 2017). According to (Evans, 2003), forecasting is more of an art than a science, requiring even the most robust econometric models to be regularly adjusted for optimal accuracy. Evans further emphasizes that all forecasts inherently contain errors, and it is crucial to acknowledge this reality from the outset. Forecasting involves studying historical data to identify systematic relationships, trends, and patterns in order to predict future events.

Forecasting can be categorized into several types. (Heizer and Render, 2014)classify forecasts into economic, technological, and demand forecasts. Economic forecasts focus on business cycles and macroeconomic indicators such as inflation rates and capital availability. Technological forecasts deal with advancements that may lead to innovative product launches and the need for new facilities and equipment. Demand forecasts project future demand for a company's products or services. In terms of time horizon, forecasting is further divided into three categories: long-term (more than three years), used for strategic planning; medium-term (one to three years), used for budgeting and production planning; and short-term (up to one year), used for operational decisions such as overtime scheduling and workforce allocation (Luthfianto, 2017).

2.2. Time Series

Time series is defined as a sequence of observations recorded over time, whether at regular or irregular intervals. These observations can take the form of numerical values, labels, or other attributes, and they may reflect trends, seasonality, or cyclic behaviors (Mahmoud & Mohammed, 2021). Seasonal time series such as increases in tourism during school holidays or rising ice cream sales during summer are influenced by periodic factors. A standard time series consists of four components: trend, seasonal, cyclical, and irregular patterns (Tsaur et al., 2002). An essential condition in time series data is stationarity, where the mean and variance remain stable over time, ensuring valid forecasting results.

Time series analysis is a statistical approach used to model and forecast future values based on past observations, assuming a correlation exists among the data points. The main objectives include understanding underlying mechanisms, forecasting future values, and optimizing control systems. Time series forecasting methods are broadly divided into classical and modern approaches (Mahmoud & Mohammed, 2021). Classical methods, such as Naive, Moving Average, Weighted Moving Average, and Exponential Smoothing, have been widely used since the early 20th century. In contrast, modern methods leverage Deep Learning techniques, which fall under the broader field of Machine Learning and Artificial Intelligence, offering higher flexibility and improved accuracy in handling complex and non-linear time series data (Hanke & Wichren, 2005).

2.3. Artificial Intelligence

Artificial Intelligence (AI) emerged in the 1950s when pioneers from the nascent field of computer science began to question whether machines could be made to "think" a question whose implications continue to be explored today. A concise definition of AI is the effort to automate intellectual tasks that are typically performed by humans. As such, AI is a broad domain that encompasses Machine Learning and Deep Learning, but also includes a wide range of approaches that do not involve learning processes at all (Chollet, 2018). For instance, early chess programs were based solely on hardcoded rules designed by programmers and thus did not qualify as examples of Machine Learning.

For a considerable period, many researchers believed that human-level intelligence could be achieved by developing a sufficiently extensive set of explicit rules to manipulate knowledge. This approach, known as symbolic AI, was the dominant paradigm from the 1950s through the late 1980s and reached peak popularity during the expert systems boom of the 1980s. While symbolic AI proved effective for solving well-defined logical problems, such as playing chess, it struggled with more complex and ambiguous tasks, such as image classification, speech recognition, and language translation, for which it was difficult to manually specify rules. This limitation gave rise to a new paradigm: Machine Learning, which now underpins many modern AI systems.

3. Methods

The research methodology employed in this study can be illustrated through the process flowchart presented in Figure 2.

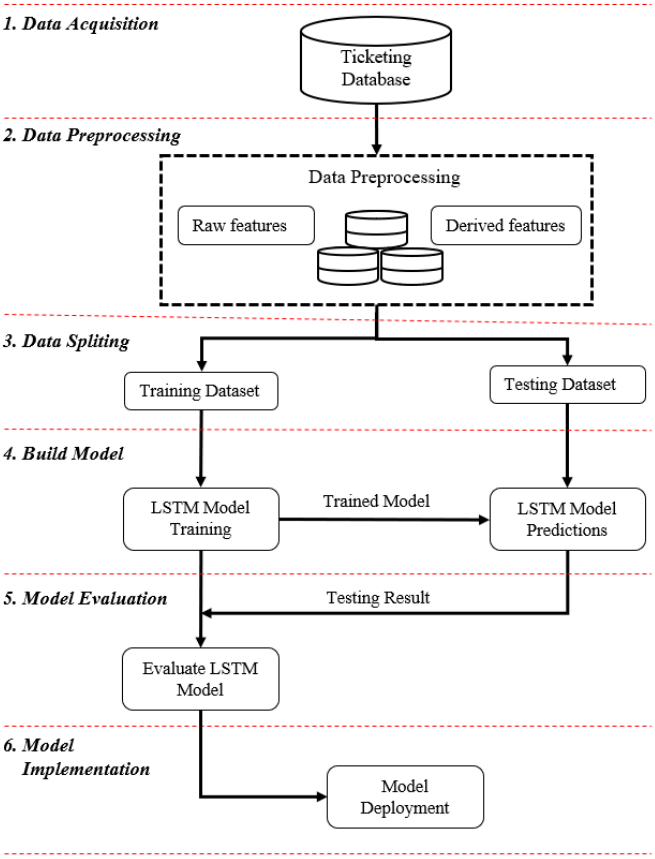

Figure 2. Research Process Flowchart

The data acquisition process in this study involved sourcing time series visitor data from Ancol's operational Ticketing database, which integrates various platforms such as desktop, web, mobile apps, and mobile POS systems, along with third-party integrations via APIs (e.g., OTA platforms and retail chains). This phase included in-depth interviews with IT and business support staff from PT Pembangunan Jaya Ancol Tbk to understand the system architecture. The acquired raw data was compiled into a reporting database, serving as the foundation for data extraction and further analysis. Data preprocessing then played a critical role in ensuring data validity, involving cleaning (e.g., handling missing values due to pandemic-related closures), integration of multiple tables, transformation (standardizing categorical variables), and reduction to optimize storage and processing time. RobustScaler was chosen for normalization due to its resilience against outliers, transforming visitor counts into a uniform range of 0–1 (Megariansyah, 2019).

After preprocessing, the dataset was split using the holdout method into training and testing sets with a ratio of 80:20, enabling model training and performance evaluation (Genç & Tunç, 2019). The model built was a Recurrent Neural Network (RNN) using Long Short-Term Memory (LSTM) architecture implemented via PyTorch. Key hyperparameters, such as batch size, learning rate, and epoch count, were initialized prior to training. The model was trained using backpropagation with Mean Squared Error (MSE) as the loss function. Model evaluation employed both MSE and Mean Absolute Percentage Error (MAPE) to assess accuracy on unseen data. Finally, the trained model was deployed using both a local device (Intel Core i7, 8GB RAM) and Google Colaboratory for broader accessibility and scalability. This deployment facilitated real-time predictions of visitor volumes for operational planning and decision-making purposes.

4. Results and Discussion

4.1. Research Results

4.1.1. Data Collection

The data used in this study were sourced from the Ancol Ticketing System, which comprises several applications operating on different platforms, including:

1. A desktop-based ticketing application,
2. A web-based ticketing application,
3. A mobile ticketing application (available for both Android and Ios),
4. An Android-based ticketing application used on Mobile Point of Sales (M-POS) devices,
5. A gate entry (turnstile) ticketing application, and
6. Application Programming Interfaces (APIs) that integrate the Ancol Ticketing System with systems owned by Online Travel Agents (OTAs) such as Traveloka, Lakupon, Blibli, Tiket.com, among others, as well as APIs that integrate with convenience stores including Indomaret, Alfamart, and AncolMart.

4.1.2. Data Preprocessing

Data preprocessing on raw data includes the following steps:

1. Visitor Count Aggregation:

This process involves summarizing the total number of visitors by grouping the data based on KODE_UNIT and TGL_TRAN. The aggregation is performed using a Structured Query Language (SQL) command as follows:

```
SELECT KODE_UNIT, TGL_TRAN, SUM(JUMLAH) AS JUMLAH FROM
T_PENJUALAN_TIKET
GROUP BY KODE_UNIT, TGL_TRAN
```

2. This step entails selecting relevant attributes to be used as the data source (dataset). For the purpose of time series analysis in this study, the chosen attributes are TGL_TRAN (transaction date) and JUMLAH (number of visitors).

As previously explained, Ancol was closed to visitors during April–May 2020, resulting in the absence of visitor data for that period. To address this, the missing data were normalized by filling the gaps with the median value calculated from the entire dataset. Based on the computation, the median value is 719,172, which was subsequently inserted into the dataset, as shown in Table 1.

TGL_TRAN	JUMLAH
2019-12	1463083
2020-01	847006
2020-02	726130
2020-03	320535
2020-04	719.172
2020-05	719.172
2020-06	30923
2020-07	160926
2020-08	223763
2020-09	111163
2020-10	228456

Table 1. Missing Value Normalisation

4.1.3. Visitor Count Prediction Model Using ARIMA

The author previously developed a visitor count prediction model for Ancol using a traditional approach, namely the Autoregressive Integrated Moving Average (ARIMA) method. The testing process began by visualizing the time series data through various stages including plotting the data graphically, calculating the rolling mean and standard deviation, performing decomposition, and conducting the Dickey-Fuller test.

Figure 3 presents the time series graph of Ancol visitor numbers from 2012 to 2022. In the graph, the X-axis represents the year, while the Y-axis indicates the number of visitors (in millions). Visually, 388 tis evident that the time series data exhibits trend characteristics, showing both upward and downward trends, as well as seasonal patterns.

Figure 3. Visitor Count Prediction Model Using ARIMA

To identify trends and clarify patterns that may not be immediately visible in the raw data, rolling mean and rolling standard deviation were applied to smooth the time series data. The results of this smoothing process are presented in Figure 4.

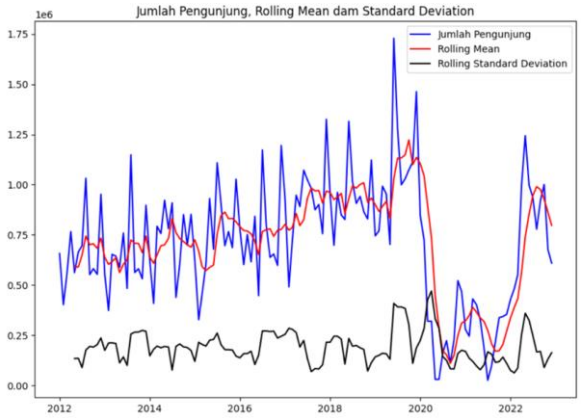

Figure 4. Rolling Mean & Rolling Standard Deviation

To further clarify the trend and seasonal patterns in the Ancol visitor data, time series decomposition was performed using the seasonal_decompose function from the statsmodels.tsa.seasonal library. The results of the decomposition are presented in Figure 5 below.

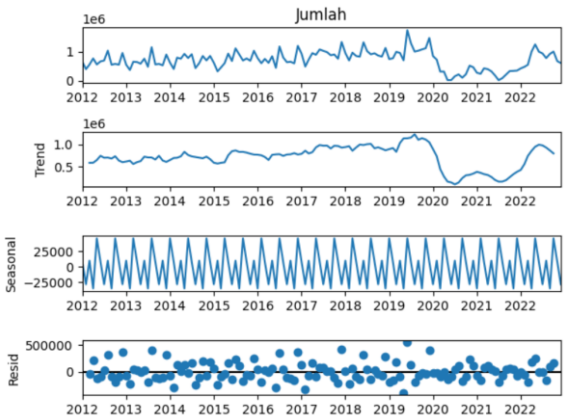

Figure 5. Decomposition Results

Based on the decomposition results, it can be observed that the number of visitors to Ancol from 2012 to 2022 exhibited an upward trend. The highest increase occurred in 2019, followed by a sharp decline in 2020 due to the Covid-19 pandemic, with a gradual recovery in 2022.

Each year, the data reveals a consistent seasonal pattern characterized by fluctuations an increase, a dip, and then a sharp rise at the end of the year. This pattern aligns with actual conditions, as during the New Year's Eve celebration (December 31), the number of visitors to Ancol under normal circumstances typically exceeds 100,000 in a single day.

The next step in the ARIMA experiment involves splitting the data into two sets: training data and testing data, using an 80:20 ratio. Figure 6 presents the graph of both datasets, where the blue line represents the training data and the red line indicates the testing data. The X-axis corresponds to the months, ranging from Month 1 to Month 132, which covers the monthly data period from 2012 to 2022. The Y-axis represents the number of visitors, measured in millions of people.

Figure 6. Training Data & Testing Data

Using the pmdarima.arima library, the ARIMA model was executed to perform forecasting based on the testing data relative to the training data. The results are presented in Figure 7 below.

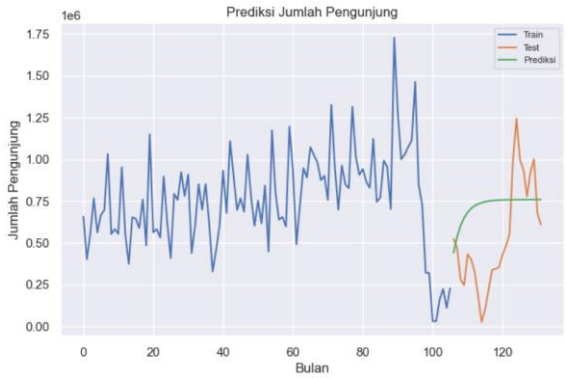

Figure 7. Prediction Results of Visitor Numbers Using the ARIMA Model

The visitor forecast is represented by the green line. It is evident that the model fails to accurately predict future values, as the forecasted results deviate significantly from the testing data (orange line). This observation is supported by error metrics, where the Root Mean Square Error (RMSE) is 317,079.89 and the Mean Absolute Percentage Error (MAPE) is 36.29%. Although the MAPE value still falls within the range considered acceptable for

implementation, it remains far from being accurate. A forecasting model is generally considered highly accurate when the MAPE is less than 10%.

Given the low accuracy of the ARIMA-based forecast, the author subsequently employed a Deep Learning approach, specifically a Recurrent Neural Network (RNN) architecture using the Long Short-Term Memory (LSTM) algorithm, with the aim of achieving more precise forecasting results.

4.2. Forecasting Model of Visitor Numbers Using RNN-LSTM

The data used in this model is the same as that employed in the ARIMA prediction model, which comprises visitor data to Ancol from January 2012 to December 2022. A visualization of the data is presented in Figure 8.

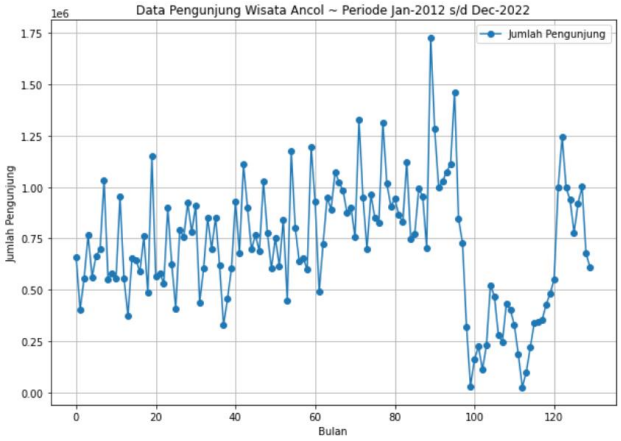

Figure 8. Ancol Visitor Data from 2012 to 2022

After obtaining the Ancol visitor data, the next step is to split the data into training and testing datasets with an 80:20 ratio. This composition is chosen based on previous studies, which have shown that the 80:20 split yields accurate prediction results. The training dataset is used to train the LSTM model, which will then be utilized to predict the number of visitors based on the testing dataset.

The next step in data preprocessing is data scaling. The figures below presents the results of scaling the visitor data using four different types of data scalers: MinMaxScaler, StandardScaler, MaxAbsScaler, and RobustScaler.

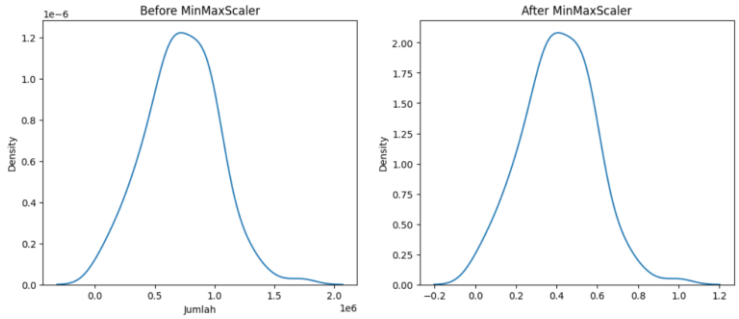

Figure 9. Scaled Data Results Using MinMaxScaler

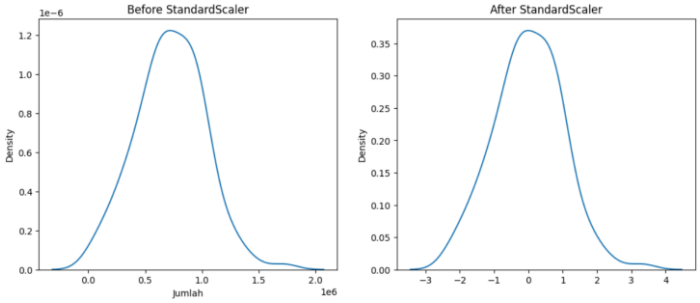

Figure 10. Scaled Data Results Using StandardScaler

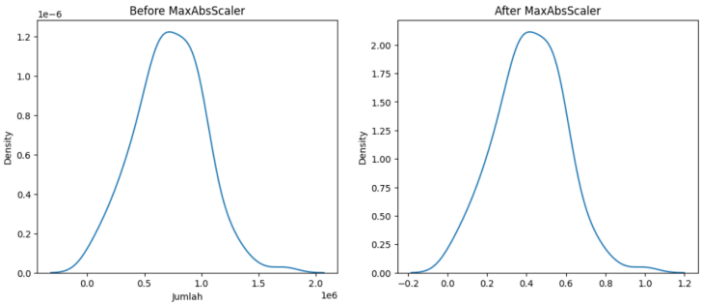

Figure 11. Scaled Data Results Using MaxAbsScaler

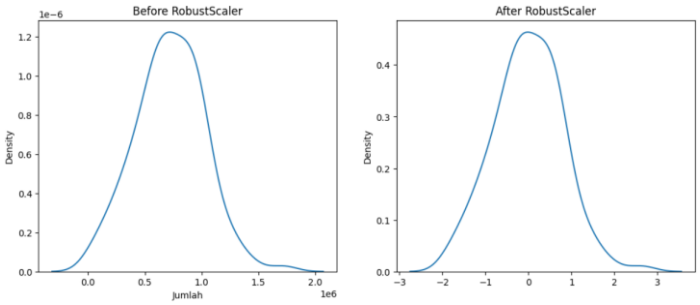

Figure 12. Scaled Data Results Using RobustScaler

Among the four types of scalers presented above, data scaling using the RobustScaler demonstrates results that most closely resemble a normal distribution. In addition, RobustScaler is reliable in handling outliers. For these reasons, this study employs RobustScaler for data scaling.

After constructing the LSTM network model, the next step is to train it using the predetermined dataset, which consists of Ancol visitor data from the 2012–2022 period. The training process aims to evaluate the performance of the network model. Prior to model training, the optimizer used must also be specified. In this study, Adaptive Moment Estimation (ADAM) was employed, which is an improvement over the Stochastic Gradient Descent (SGD) algorithm.

The evaluation of the LSTM model on the training dataset primarily aims to assess the magnitude of error and the model's prediction accuracy. In this study, the model's error (loss function) is measured using the Mean Squared Error (MSE) criterion. Below are the results of three training sessions using the ADAM optimizer with three different maximum epoch settings: 100, 500, and 1000, respectively.

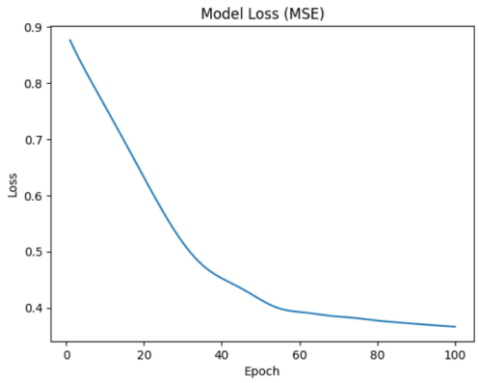

Figure 13. Model Loss (MSE) with Maximum Epoch of 100

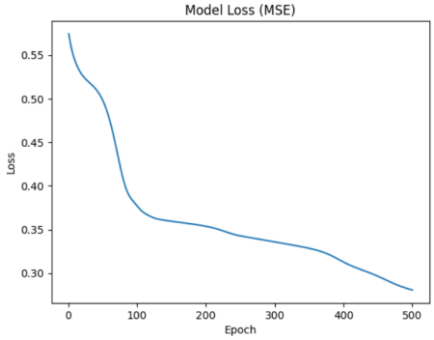

Figure 14. Model Loss (MSE) with Maximum Epoch of 500

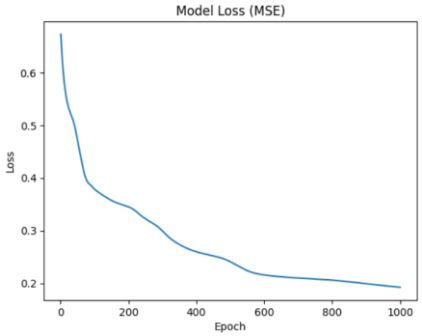

Figure 15. Model Loss (MSE) with Maximum Epoch of 1000

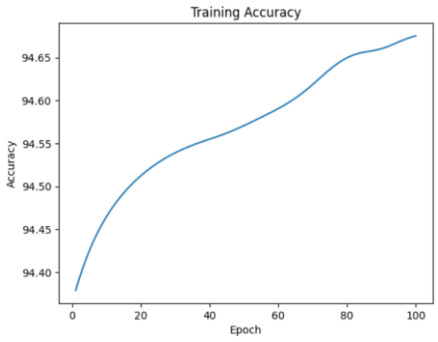

Figure 16. Model Accuracy with Maximum Epoch of 100

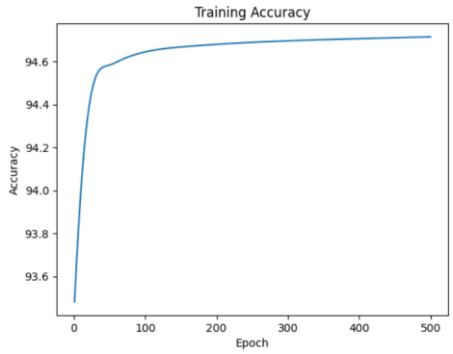

Figure 17. Model Accuracy with Maximum Epoch of 500

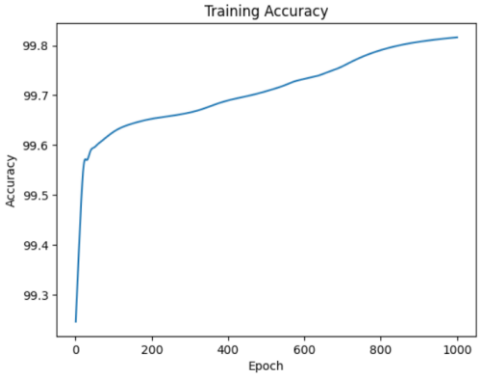

Figure 18. Model Accuracy with Maximum Epoch of 1000

After training the model using the training dataset, the final step is to evaluate the model using the testing dataset to predict future visitor numbers based on the test data. Figure 13 - Figure 18 presents the prediction results using the LSTM model. Visually, the graph shows that the LSTM model's predictions closely align with the actual data, both for the training and testing datasets.

4.3. Discussion

The study demonstrates that the use of the RNN-LSTM model significantly improves the accuracy of visitor forecasting compared to traditional models like ARIMA. The ARIMA model yielded a Mean Absolute Percentage Error (MAPE) of 36.29%, which is relatively high, indicating poor forecasting performance. Conversely, the LSTM model achieved a much lower Mean Squared Error (MSE) of 0.3095 and a prediction accuracy of 94.85%. This finding supports recent advancements in tourism forecasting using deep learning, which have proven effective in capturing nonlinear time series patterns (Munawar et al., 2024).

The LSTM model's superior performance can be attributed to its ability to learn long-term dependencies and temporal patterns within time series data. Unlike ARIMA, which assumes linear relationships, LSTM adapts to fluctuations and seasonality in visitor data such as spikes during holidays and weekends. Moreover, the integration of RobustScaler normalization contributed to better outlier handling, a crucial factor in real-world datasets often affected by anomalies like pandemic closures. This is consistent with findings by (Liu et al., 2021), who emphasized the importance of preprocessing and deep learning synergy in tourism demand forecasting.

Data collection and preprocessing stages were essential to ensure the reliability of the forecasting model. The dataset was compiled from multiple ticketing platforms integrated via API with various vendors, including Traveloka and Alfamart. Gaps in data due to pandemic-induced closures were filled using the median value, enabling continuity in the time series.

Similar approaches were validated by (Wang et al., 2023), who showed that appropriate imputation techniques significantly enhance the stability of time series forecasting models.

5. Conclusion

This study concludes that the application of the Recurrent Neural Network using Long Short-Term Memory (RNN-LSTM) provides a highly accurate model for forecasting visitor numbers to Ancol. The model was trained and tested using an 80:20 data split, with robust preprocessing through RobustScaler and optimization via the ADAM algorithm. Among several configurations, the model achieved a minimum Mean Squared Error (MSE) of 0.3095 and a maximum prediction accuracy of 94.85%. These results confirm the LSTM model's superior capability compared to the previously used ARIMA model.

The success of the LSTM-based model demonstrates its effectiveness in capturing complex seasonal and nonlinear trends inherent in tourism data. Accurate forecasting is vital for supporting resource allocation, infrastructure planning, and visitor safety especially during peak seasons. This approach has great potential for operational decision-making at PT Taman Impian Jaya Ancol and similar recreational destinations. Consequently, the implementation of this model can significantly enhance visitor management strategies in the tourism sector.

It is recommended that tourism operators adopt the RNN-LSTM forecasting model to optimize resource planning and enhance service delivery. Future studies may integrate additional variables, such as weather patterns, public holidays, or social media trends, to improve prediction accuracy. A web-based dashboard for real-time prediction output could further assist management in daily operations. Additionally, extending this model to other tourist destinations in Indonesia would allow comparative analysis and broader applicability. Finally, collaboration with local governments and emergency services should be considered to align forecasting with public safety strategies.

6. References

Bollenbach, J., Neubig, S., Hein, A., Keller, R., & Krcmar, H. (2024). Enabling active visitor management: local, short-term occupancy prediction at a touristic point of interest. *Information Technology & Tourism*, *26*(3), 521-552.

Chollet, F. (2018). Deep Learning with Python. In 2018 21st International Conference on Information Fusion, FUSION 2018. Manning Shelter Islan. https://doi.org/10.23919/ICIF.2018.8455530

Evans, M. K. (2003). Practical Business Forecasting. Blackwell Publishers Ltd.

Hanke, J. E., & Wichern, D. W. (2005). *Business forecasting*. Pearson Educación.

Heizer, J., & Render, B. (2014). Operations Management: Sustainability and Supply Chain

Liu, Z., Zhu, Z., Gao, J., & Xu, C. (2021). Forecast methods for time series data: A survey. *Ieee Access*, *9*, 91896-91912.

Luthfianto, S. (2017). Perencanaan Dan Pengendalian Produksi. In Universitas Pancasakti Tegal,.

Mahmoud, A., & Mohammed, A. (2021). A Survey on Deep Learning for Time-Series Forecasting. In Studies in Big Data (Vol. 77, Issue February). Springer International Publishing. https://doi.org/10.1007/978-3-030-59338-4_19

Megariansyah, T. S. (2019). Prediksi Debit Aliran menggunakan Long Short-Term Memory (LSTM). Zenodo, 0–6.

Munawar, M., McNeil, R., Jani, R., Nur, E. M., & McNeil, D. (2024). Variation and Forecasting of Land Surface Temperature in Malaysia. *Pertanika Journal of Science &*

Technology, *32*(6).

Oktavia, R. C. D., Nurbaeti, Ratnaningtyas, H., & Rachmadhita, M. A. (2023). The influence of TikTok and destination image on the decision to visit Taman Impian Jaya Ancol post-pandemic and visitor satisfaction.

Tsaur, R. C., Wang, H. F., & Yang, J. . (2002). Fuzzy Regression For Seasonal Time Series Analysis. International Journal of Information Technology & Decision Making.

Wang, S., Li, J., Shi, X., Ye, Z., Mo, B., Lin, W., ... & Jin, M. (2024). Timemixer++: A general time series pattern machine for universal predictive analysis. *arXiv preprint arXiv:2410.16032*.

Yamin, M., Muthalib, A. A., Tin, R., & Rahim, M. (2020). Influence of the number of tourism visits, and hotel occupancy on tourism sector revenue and economic growth in Indonesia. *International Journal of Economics and Management Studies*, *7*(8), 205-209.