Forecasting Tourist Arrivals in Bali: A Grid Search-Tuned Comparative Study of Random Forest, XGBoost, and a Hybrid RF-XGBoost Model

Kadek Jemmy Waciko[1*], Leni Anggraini Susanti[1], Muayyad[1], Rifqi Nur Fakhrurozi[1]
[1]Department of Business Administration, Politeknik Negeri Bali, Badung, Indonesia
*Corresponding author: jemmywaciko@pnb.ac.id

Received: 30 July 2025
Revised: 3 November 2025
Accepted: 3 November 2025

ABSTRACT – Tourism planning, infrastructure growth, and economic stability. This study presents an extensive comparative evaluation of Random Forest (RF), Extreme Gradient Boosting (XGBoost), Long Short-Term Memory (LSTM), and a novel Hybrid RF-XGBoost model in the prediction of monthly international tourist arrivals. A full time series dataset of a ten-year period (2014–2024) from the Central Bureau of Statistics of Bali was used for training and testing the models. Hyperparameter optimization using Grid Search with cross-validation (Grid Search CV) was used for all the machine learning models to obtain best predictive performance. Two robust metrics, Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE), were used to assess forecasting accuracy. Results show that the Random Forest model outperforms all competitors with lowest RMSE (41,772.68) and MAPE (6.30%), indicating high forecasting precision and robustness, especially during structural breaks such as the COVID-19 pandemic. The hybrid model also performs well, with LSTM indicating higher error rates, illustrating its shortcomings on small-to-medium-scale tourism time series. Besides, the study provides six-month ahead predictions (January–June 2025) with 95% prediction intervals, showing an ongoing trend of recovery. The findings affirm the superiority of bagging-based ensemble methods over polynomial-based methods in capturing nonlinearity, seasonality, and exogenous shocks in tourist demand. The study plugs the growing amount of data-driven tourism analytics by offering a reproducible, high-precision forecasting model for developing countries and seasonally driven destinations.

Keywords–Grid Search CV, LSTM, Machine Learning, Random Forest, XGBoost

I. INTRODUCTION

Tourism is a significant contributor to Bali's economy, with international tourist arrivals influencing various economic sectors. Reliable forecasting of tourist arrivals enables strategic planning and enhances the effectiveness of policy formulation. Traditional time series forecasting methods, such as ARIMA or Holt-Winters, often struggle with capturing nonlinear patterns in complex datasets. In contrast, machine learning models such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost) have shown superior performance in modeling complex relationships in time series data [1].

Recent studies over the past five years have increasingly applied machine learning algorithms for tourism forecasting with notable success. For example, [2] have been employed to predict tourism demand by incorporating various predictors such as international COVID-19 cases, tourist arrivals, and quarantine policies. These models outperform traditional models like ARIMA and neural networks in accuracy. In Italy, RF and Gradient Boosting models have been used to predict international tourists' Length of Stay (LoS. These models excel in identifying complex data patterns and provide actionable insights for tourism policymakers, enhancing strategic planning and optimizing services [3]. For tourist emotions analysis, XGBoost has been utilized to analyze tourists' emotional changes in natural forest landscapes across different seasons. This model, combined with SHapley Additive exPlanations (SHAP), helps in understanding the nonlinear impact of landscape indicators on tourist emotions, aiding in sustainable tourism development [4].

While there have been significant improvements in using machine learning for tourism demand prediction, empirical studies on Indonesia, Bali, a globally significant island tourism destination, have been limited and not methodologically sound. Most prior studies employ individual models such as RF or XGBoost, omit hyperparameter tuning, or work with data from non-Indonesian contexts with different tourism features. This study bridges these critical gaps by conducting a comprehensive, grid search-optimized comparison of four advanced models: Random Forest, XGBoost, LSTM, and a novel Hybrid RF-XGBoost model, using a high-quality, decade-long time series of monthly international tourist arrivals to Bali. All models are subjected to systematic hyperparameter tuning via GridSearchCV with cross-validation to ensure reproducibility, fairness, and maximum predictive performance.

The initial aim of this research is to establish the most accurate and dependable machine learning model of forecasting tourist arrivals in Bali under optimal available conditions. By measuring model performance on training and test data with RMSE and MAPE, this study not only contrasts algorithmic performance but also examines the capability for generalization under the presence of structural breaks—such as the COVID-19 pandemic. The study provides a replicable, data-driven forecasting model to suit seasonally volatile and shock-prone tourism economies that can provide added practical application for regional policy makers and add to the body of work on machine learning applications in emerging tourism markets.

II. LITERATURE REVIEW

A. Random Forest

Random Forest in machine learning is categorized as an ensemble method. This algorithm consists of a collection of regression trees built from a number of input variables and data sampled randomly from the original dataset [5]. Random Forest is well-regarded for its ability to process complex, non-linear data effectively. It constructs multiple decision trees and aggregates their predictions, which enhances its capability to manage diverse datasets, including those with high dimensionality and missing values [6].

The general principle of Random Forest is to combine a collection of randomly generated decision trees. Random Forest reduces variance by averaging the results of multiple trees, which decreases the model's sensitivity to the specific data on which it was trained. This averaging process also helps in reducing bias, leading to more accurate and stable predictions [7]. In addition, Random Forest is particularly effective in handling large datasets. It can manage high-dimensional data and datasets with missing values without requiring extensive preprocessing, making it well-suited for applications in fields such as bioinformatics, finance, and crime prediction [8].

The prediction calculation in Random Forest for the regression case is obtained by averaging the predictions from each regression tree, as expressed in Equation 1.

$$\widehat{h}\,(x) = \frac{1}{q}\sum_{\ell=1}^{q} \widehat{h}\,(x, \Theta_\ell) \tag{1}$$

where q is the number of regression trees generated, and $\widehat{h}\,(x, \Theta_\ell)$ represents the prediction from the ℓ regression tree.

B. Extreme Gradient Boosting (XGBoost)

XGBoost is a powerful machine learning algorithm widely used for both regression and classification tasks, including time series forecasting. However, to apply XGBoost effectively to time series data, it is essential to first transform the sequential data into a supervised learning format. This transformation begins with feature extraction, where relevant statistical features such as moving averages, seasonal components, and trend indicators are computed to reflect the distribution and temporal structure of the data. These features provide meaningful inputs that help the model understand time-dependent patterns [9].

The predictive model in this study uses the XGBoost algorithm. XGBoost is well-regarded for its robustness and generalization capabilities. It handles noisy and incomplete data efficiently, making it highly suitable for real-world forecasting scenarios such as power quality disturbance detection and GNSS time series modeling. Altogether, these capabilities make XGBoost a highly versatile and effective tool for time series forecasting [9]. For a dataset consisting of n observations and m variables, the ensemble tree model can be expressed as follows:

$$\hat{y_l} = \phi(x_i) = \sum_{k=1}^{k} f_k(x_i), f_k \in \mathcal{F} \tag{2}$$

In Equation (2), $\hat{y_l}$ represents the predicted value, and f_k denotes the kkk-th regression tree in the set of regression trees $\mathcal{F}$. Next, the objective function is calculated.

$$\mathcal{L}(\phi) = \sum_i l(\hat{y_l}, y_i) + \sum_k \Omega f_k \tag{3}$$

l represents the loss function, which indicates the difference between the predicted value and the actual value. Ωf_k denotes the regularization term that reflects the model complexity. At each step of the XGBoost training process, the objective function is minimized by generating a new regression tree model based on the previously existing models. The optimal objective function, which displays the predicted values at each leaf node, is presented in the following equation:

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\epsilon I_j} g_i\right)^2}{\sum_{i\epsilon I_L} h_i + \lambda} + \frac{\left(\sum_{i\epsilon I_j} g_i\right)^2}{\sum_{i\epsilon I_R} h_i + \lambda} + \frac{\left(\sum_{i\epsilon I_j} g_i\right)^2}{\sum_{i\epsilon I} h_i + \lambda}\right] - \gamma \tag{4}$$

C. Model Goodness Measures

The accuracy level of a forecasting model can be assessed by comparing the projected values with the actual data. The accuracy of a forecasting model is determined by the smallest error value from each accuracy measurement method—the smaller the value, the more accurate the model is in making predictions. The Root Mean Square Error (RMSE) is used to evaluate the developed model [10], while the Mean Absolute Percentage Error (MAPE) is used to evaluate the forecast results. RMSE and MAPE provide insights into model performance: RMSE gives an overall picture of the prediction error magnitude and is more sensitive to outliers. It is calculated using the following formula [11].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

MAPE is the percentage that reflects the forecasting results relative to the actual data over a specific period, calculated by taking the absolute error in each period divided by the actual value, then averaging it over the total number of periods [12].

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - F_i}{y_i}\right| \times 100\% \quad (6)$$

Explanation:

y_i = actual time series value

F_i = forecasted value

n = number of time series data points

The interpretation of MAPE calculation results [28] is as follows: 1). < 10 % : *highly accurate forecasting*, 2). 10% - 20% : *good forecasting*, 3). 20% - 50% : *reasonable forecasting*, dan 4). >50%: *weak and inaccurate predictability*

III. METHODOLOGY

A. Data Sources and Research Variables

The data used in this study is secondary data obtained from the official publications of the Bali Provincial Statistics Agency (Badan Pusat Statistik Provinsi Bali, 2025). The dataset contains information on the number of international tourists visits to Bali Province, recorded on a monthly basis over the period from 2014 to 2024. Structurally, the data consists of three variables: month, year, and number of tourists. The month and year variables represent the time dimension indicating the visit period, while the number of tourists variable indicates the total number of international tourist arrivals in Bali each month.

All data is presented on a ratio scale, with the exception of the time variables—month, which is ordinal, and year, which is interval. The data is time series in nature, making it highly relevant for analysis aimed at identifying trend patterns, seasonal fluctuations, and the impact of special events on the dynamics of tourism in Bali. The data source originates from the official website of BPS Bali Province, which consistently provides valid and methodologically accountable sectoral statistical data. Therefore, this dataset is considered highly reliable to support the analytical needs of this study.

B. Data Analysis Stages

The data analysis phase in this study took place through a thorough, multi-step process that included effective methods in machine learning time series forecasting. Each step was planned carefully to ensure strong analysis, reproducibility, and validity of the findings. The stages are outlined in detail below [14]:

1. Data Collection
   This study began with the collection of monthly tourist arrival data to Bali for the period 2014 to 2024. The data was obtained from the official publications of the Bali Provincial Statistics Agency (BPS) and served as the main dataset for the modeling process.
2. Determining Characteristics and Descriptive Statistics
   Before modeling, the data was analyzed descriptively to identify general characteristics such as minimum, maximum, average, and standard deviation. This analysis also aimed to evaluate the data distribution and detect seasonal patterns and trends within the time series data.
3. Data Splitting (Train-Test Split)
   The dataset was divided into two parts: 80% for the training set and 20% for the testing set. The purpose of this split is to train the model using historical data and to test the model's performance on unseen data.
4. Hyperparameter Tuning
   In this stage, optimal parameter tuning (hyperparameter tuning) was carried out for the three main models:
   a. Random Forest, a bagging ensemble method
   b. XGBoost, a boosting-based gradient tree algorithm
   c. LSTM
   The tuning was performed on the training set using the GridSearchCV technique to obtain the best parameter combination for each model. The goal of this process is to improve model accuracy and generalizability.
5. Hybrid Model Development
   After each individual model was successfully tuned, a hybrid model was built. This hybrid model was created by combining the predictions of Random Forest and XGBoost using a simple averaging approach. The aim is to leverage the strengths of each model and enhance the stability of the prediction results.
6. Forecasting on Training Data
   Each model (Random Forest, XGBoost, LSTM, and Hybrid RF + XGB) was used to forecast the training data. This step was conducted to evaluate the initial performance of the models and to assess how well each model fits the historical data patterns.
7. Model Performance Evaluation
   The predictions from each model were evaluated using the following performance metrics:
   a. Root Mean Squared Error (RMSE)

b. Mean Absolute Percentage Error (MAPE)
This evaluation assesses the level of error and accuracy of each model in forecasting the training data.

8. Selection of the Best Model
The model with the smallest error values (RMSE and MAPE) was selected as the best model. This model was then used to make predictions on the testing data, as it was considered to have the most accurate predictive capability.
9. Prediction Using Testing Data
The best model was applied to the testing data (20%) to predict the number of tourist arrivals not used during model training. This aims to evaluate the model's performance on new, unseen data.
10. Final Results
The prediction results on the testing data serve as the main indicator to assess the model's generalization capability. The accuracy at this stage forms the basis for concluding the model's effectiveness in forecasting future tourist arrivals.

The entire analytical procedure described above can be visually represented through the following flowchart.

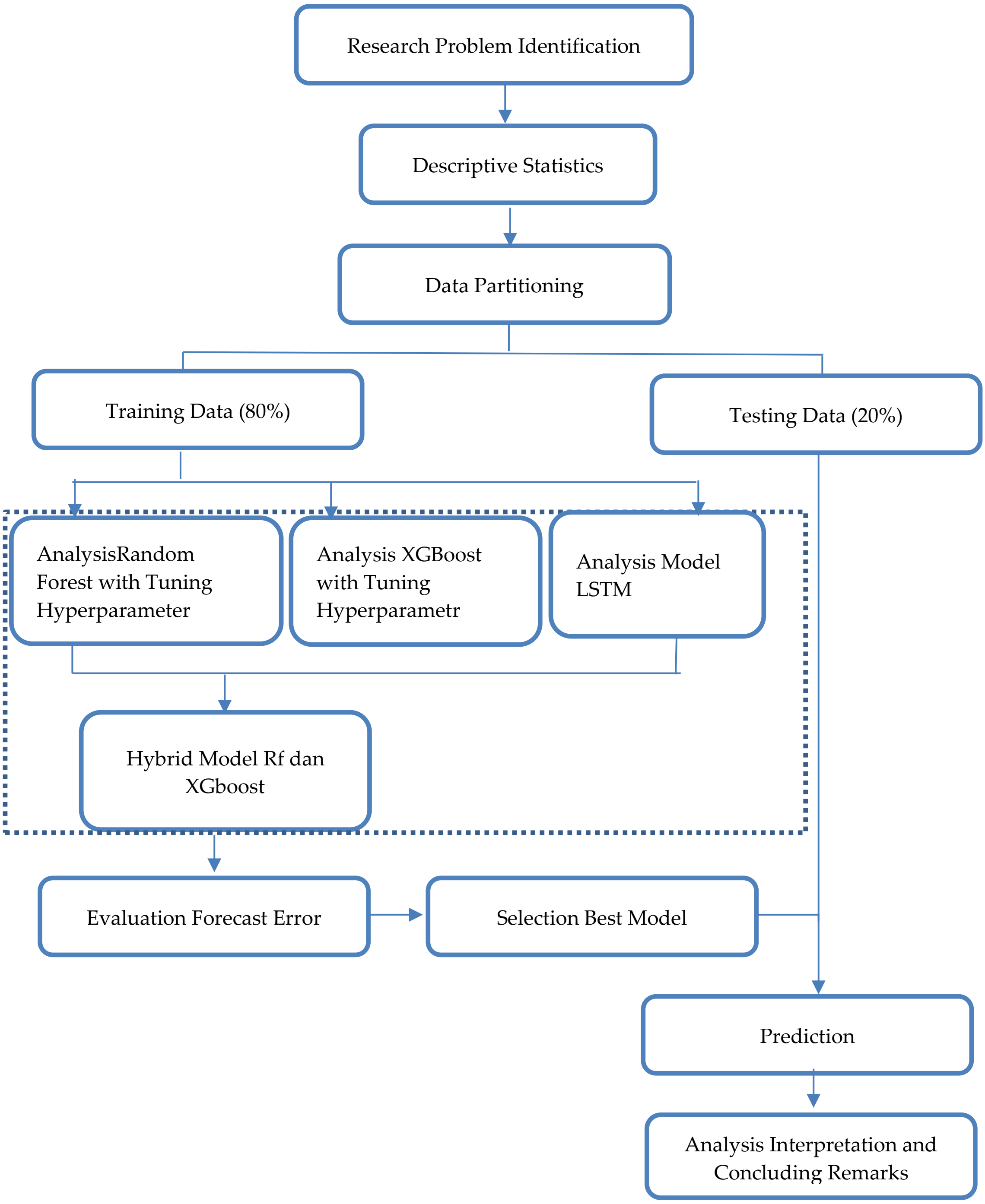

Figure 1. Flowchart of the Research Process

IV. RESULTS AND DISCUSSIONS

A. Descriptive Statistics

Descriptive statistics serve to provide a comprehensive overview of the dataset. In this study, descriptive analysis was performed through the construction of time series plots to identify and illustrate underlying trends in the data. Figure 2 illustrates the monthly arrivals of international tourists to Bali, covering the period from January 2014 to December 2024. The graph exhibits a repeating seasonal oscillation, characterized by predictable peaks aligned with global holiday calendars. A steep, extended drop is evident from early 2020 to mid-2021, clearly indicating the severity of the COVID-19 pandemic on international travel. Yet, a consistent upward trajectory begins in late 2021, with arrivals progressively climbing and projected to approach the levels seen before the pandemic by late 2024. Such observations underscore the importance of employing advanced machine learning techniques that can accommodate both nonlinear growth patterns and sudden structural breaks.

Figure 2 Monthly international tourist arrivals to Bali from 2014 to 2024

Table 1 presents the results of descriptive analysis on the number of international tourist visits to Bali during the period from January 2014 to December 2024. The average number of visits was 345,240 people per month. The maximum value recorded was 625,665 people, while the minimum was only 1 person, indicating the presence of extreme disruptions in tourist arrivals—particularly during the COVID-19 pandemic, which led to the closure of entry access for international travelers. The standard deviation of 193,321 reflects a high level of dispersion in the data, indicating that tourist arrivals fluctuated significantly from month to month. Meanwhile, the median (50th percentile) was 379,021, with the first quartile (25%) and third quartile (75%) at 278,607 and 487,295, respectively. This suggests that the data distribution is relatively wide, with the majority of values falling within that range.

These findings indicate that the dynamics of tourist visits are non-linear and are heavily influenced by seasonal patterns as well as external factors such as global health crises. Therefore, this study recommends the use of advanced analytical approaches capable of capturing these characteristics more comprehensively, such as nonlinear time series models or machine learning algorithms that can accommodate both seasonal patterns and unexpected structural events.

Table 1 Descriptive Analytics

Descriptive Analysis	Monthly Tourist Arrivals to Bali
Mean	345240
Std	193321
Min	1
25%	278607
50%	379021
75%	487295
Max	625665

B. Model Performance Comparison

In this study, four machine learning algorithms were employed as predictive models for time series data, namely Random Forest, Extreme Gradient Boosting (XGBoost), Long Short-Term Memory (LSTM), and a Hybrid Random Forest–XGBoost model. Prior to the modeling stage, the dataset was divided into two parts: 80% for training and 20% for testing. This proportion was chosen to ensure that the models could learn sufficient information from historical data to build representative predictive patterns while still providing adequate test data for performance evaluation. The training process began with two primary models, Random Forest and XGBoost, selected for their ability to capture non-linear patterns and their effectiveness in various time series prediction tasks. To achieve optimal configurations, hyperparameter tuning was conducted using the GridSearchCV method, which systematically searches for the best parameter combinations through a cross-validation scheme. The hyperparameter configurations for each model are presented in Table 2.

Table 2. Parameter Random Forest dan XGBoost

Model	Parameter	Value
Random Forest	n-estimators	100, 200
	max_depth	5, 10, None
	min_samples_split	2, 5
XGBoost	n-estimators	100, 200
	learning_rate	0.05, 0.1, 0.2
	max_depth	3, 5, 7
LSTM	Learning rate	0.003, 0.004, 0.001
	Batch size	8, 16, 32, 64

In the Random Forest algorithm, tuning was performed on three main parameters. The first is n_estimators, which refers to the number of trees in the ensemble, with values set at 100 and 200. The second is max_depth, representing the maximum depth of each tree, set at 5, 10, and None to control the model's ability to learn complex patterns. The third parameter is min_samples_split, the minimum number of samples required to split an internal node, varied at 2 and 5, which helps regulate the granularity of node splits to maintain efficiency and avoid excessively deep trees.

Meanwhile, in the XGBoost model, tuning was also carried out on several key parameters. The n_estimators parameter was set at 100 and 200, indicating the number of boosting rounds. The learning_rate parameter, tested at values of 0.05, 0.1, and 0.2, was used to control the convergence speed of the model toward the global minimum and to avoid error spikes from overly aggressive learning. Lastly, max_depth was set at 3, 5, and 7 to adjust the complexity of each tree, aiming to strike a balance between model bias and variance.

For the LSTM (Long Short-Term Memory) model, hyperparameter tuning was conducted on two primary parameters. The first is the learning rate, tested at values of 0.003, 0.004, and 0.001. This parameter determines the step size used by the optimizer when updating network weights during training. A smaller learning rate allows for more stable convergence but requires more epochs, whereas a larger value accelerates learning but risks overshooting the optimal point. The second parameter is the batch size, varied at 8, 16, 32, and 64, which defines the number of samples processed before the model updates its internal parameters. Smaller batch sizes typically provide more frequent updates and can help capture fine-grained temporal dynamics in time series data, while larger batches improve computational efficiency and gradient stability. These hyperparameters were optimized to achieve a balance between model accuracy, training stability, and computational cost.

Table 3 Best Parameters from Grid Search

Model	Parameter	Value
Random Forest	n-estimators	100
	max_depth	10
	min_samples_split	5
XGBoost	n-estimators	100
	learning_rate	0.1
	max_depth	3
LSTM	Learning rate	0.001
	Batch size	16

Table 3 presents the results of hyperparameter tuning using the GridSearchCV technique for the Random Forest and XGBoost models. For Random Forest, the best-performing parameters were found to be n_estimators = 100, max_depth = 10, and min_samples_split = 5. Meanwhile, for XGBoost, the optimal configuration was n_estimators = 100, learning_rate = 0.1, and max_depth = 3. These parameter combinations were selected because they provided the best predictive performance on the training data and were subsequently used to build the final models.

Once the optimal configurations of both primary models were obtained, the process continued with the development of the LSTM model, a deep learning algorithm specifically designed to recognize sequential patterns and long-term dependencies in time series data. The fourth model developed was the Hybrid Random Forest–XGBoost, which combines the strengths of both ensemble models with the goal of improving predictive accuracy and reducing the weaknesses of individual models.

For the LSTM model, which was optimized separately through grid search due to its deep learning architecture, the best parameters were determined to be a learning rate of 0.001 and a batch size of 16. This configuration achieved the most stable convergence and lowest validation loss among all tested combinations. The selected learning rate allowed the optimizer to update the model weights gradually, avoiding oscillations in loss values, while the batch size of 16 provided a balanced trade-off between computational efficiency and gradient stability. These optimal hyperparameters were then applied in the final LSTM model to capture temporal dependencies and nonlinear relationships in the sequential data effectively.

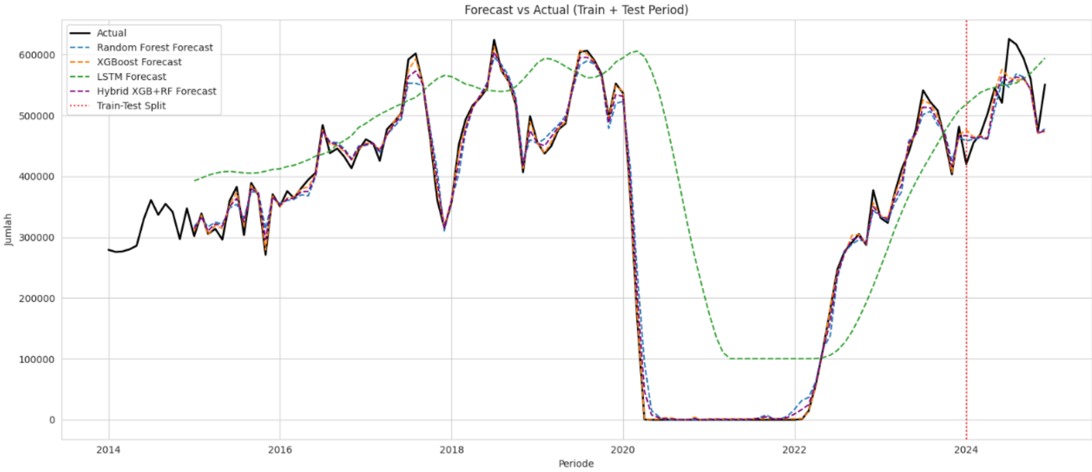

Figure 3. Actual vs. Forecasted Tourist Arrivals Using Multiple Models (Train + Test Period)

Figure 3 compares actual international tourist arrivals to Bali with predictions from four models: Random Forest (RF), XGBoost (XGB), LSTM, and the Hybrid XGB+RF. This covers the full time period from 2014 to 2024, including both training and testing sets. This timeline allows us to assess each model's reliability through changing data dynamics: the steady pre-pandemic period, the sharp decline during the pandemic, and the subsequent recovery.

1. Pre-COVID-19 Stability (2014-2019)
   During the pre-pandemic period, all models perform well. Random Forest, XGBoost, and the Hybrid version track actual arrivals nearly interchangeably. They effectively follow the overlapping seasonal peaks and long-term growth, showing their strength in handling non-linear, cyclical time series. Random Forest and the Hybrid model slightly outperform XGBoost by providing tighter predictions at each peak, especially during the busiest months. LSTM, while consistent, shows a slight lag. This model's design, which focuses on learning from longer dependencies, has difficulty quickly adjusting to the tighter seasonal variations.
2. Structural Break and Crisis Phase (2020–2021)
   In early 2020, the COVID-19 pandemic brought an unprecedented shock that almost stopped international arrivals to Bali. All models capture this change because of the sharp drop in historical values. However, XGBoost performs poorly during this phase and fails to adjust its predictions effectively. This issue arises from XGBoost's use of residual-based additive modeling, which does not handle sudden, one-time structural changes well. In contrast, Random Forest and the Hybrid XGB+RF models show better adaptability and closely follow the actual data. Their ensemble bagging structure likely makes them stronger against outliers and sudden shifts, allowing them to be more resilient during crises. The LSTM model struggles during this time, overestimating tourist numbers because its memory-based prediction method may overfit trends from before the crisis [2], [15].
3. Post-Pandemic Recovery (2022–2024)
   As tourism picked up after COVID, thanks to relaxed travel rules and recovery policies, the differences between the models became clearer. Random Forest and the Hybrid model consistently track the strong upward trend and cyclical recovery accurately. The Hybrid model shows a slightly better ability to follow sharp reversals. This improvement likely comes from its combination of error correction and variance reduction. LSTM again struggles to capture the speed of recovery, producing forecasts that are delayed and muted. This issue highlights a major drawback of deep learning methods in small to medium-sized time series data. They need long training sequences and a lot of variability in training signals to adjust well to new data patterns [16].

C. Quantitative Evaluation of Forecasting Performance

To complement the visual inspection of model performance, Table 4 presents a numerical comparison based on two widely accepted evaluation metrics: Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). These metrics serve as important indicators of prediction accuracy and generalization ability across.

Table 4. Comparative Forecasting Performance of Machine Learning Models.

MODEL	RMSE	MAPE
Random Forest	41772.68	6.30
XGBoost	43605.77	6.51
LSTM	66859.49	12.81
Hybrid XGB+RF	42202.32	6.41

1. Random Forest: The Best Performing Model

   Among all the models tested, Random Forest (RF) achieved the lowest RMSE of 41,772.68 and MAPE of 6.30%. This shows it has better predictive accuracy and consistency across the dataset. RF uses bagging to handle high variance and prevent overfitting. This is important for datasets that contain noise, non-linearity, and seasonal patterns, which are common in tourism time series. Its ability to divide the feature space and average multiple decorrelated decision trees adds to its stability and precision. This performance matches previous research that Random Forest stands out as a highly effective predictive model across various fields, including healthcare, education, business, and safety. Its high accuracy, ability to handle complex data, and flexibility make it a preferred choice for many predictive tasks. However, considerations around computational resources and model complexity should be taken into account to optimize its use [17], [18], [19], [20].
2. XGBoost: Competitive but Less Stable

   Extreme Gradient Boosting (XGBoost) showed a slightly higher RMSE of 43,605.77 and MAPE of 6.51% compared to RF. While the difference seems small, it indicates that XGBoost's boosting setup can make it sensitive to data changes, especially during structural shocks like the COVID-19 pandemic. Unlike RF, XGBoost builds trees one after another, which makes it more likely to pass on errors if the residuals are not consistent. Still, its performance is highly competitive algorithm with exceptional predictive capabilities across various domains. However, its stability can be less reliable due to issues like overfitting and the need for extensive hyperparameter tuning. Practitioners should consider these factors and possibly employ hybrid models or additional regularization techniques to mitigate stability concerns[21], [22].
3. LSTM: Poor Generalization Despite High Model Complexity

   Long Short-Term Memory (LSTM), a deep learning model designed for sequential data, had the highest RMSE of 66,859.49 and MAPE of 12.81%. This shows it did not fit the dataset well. This result highlights a common issue with deep learning in low-to-medium scale time series forecasting. Without enough training data, extensive tuning, and a clear signal, while LSTMs are powerful tools for predictive tasks, their high model complexity often leads to poor generalization. Addressing these challenges requires a combination of regularization techniques, optimized architectures, hybrid models, and advanced optimization methods [23], [24], [25], [26]
4. Hybrid XGB + RF: Trade-Off Between Accuracy and Complexity

   The hybrid model that combines XGBoost and Random Forest (XGB+RF) produced an RMSE of 42,202.32 and MAPE of 6.41%. This is slightly below RF but better than both XGBoost and LSTM. This model takes advantage of XGBoost's strength in capturing residual patterns while using RF's capability to handle non-linearities and variance. The hybrid XGB + RF model offers significant improvements in prediction accuracy and robustness across diverse applications. However, these benefits come with increased complexity, including longer training times, higher computational costs, and the need for extensive hyperparameter tuning. Balancing these trade-offs is crucial for effectively leveraging hybrid models in practical scenarios [27].

Based on the evaluation results using the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) metrics, the Random Forest model demonstrated the most optimal predictive performance compared to the other models. Considering the accuracy and stability of the predictions produced, this model was subsequently selected as the basis for forecasting the number of tourist visits to Bali for the next six periods.

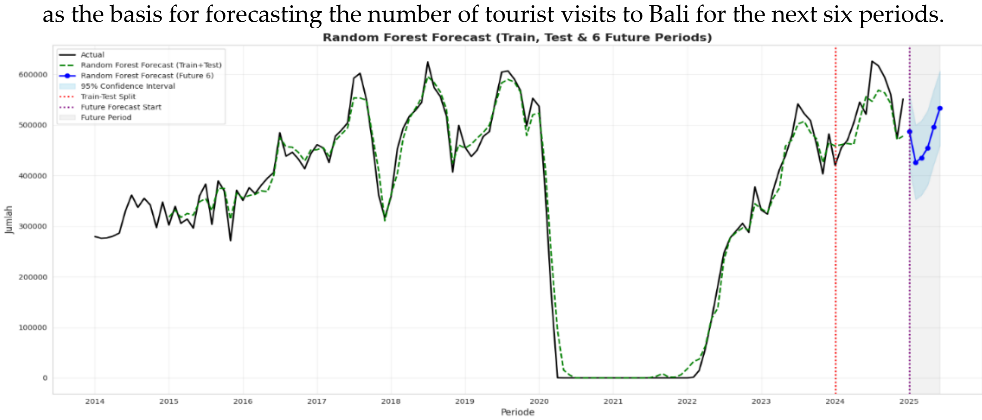

Figure. 4 Random Forest Forecast (training, testing and 6 future periods)

Figure. 4 Performance of the Random Forest in predicting a time series data set in three different ranges (training, testing and future prediction (×6 periods ahead). The black solid line is the true data, while the Random Forest predictions for both training and test are in green dashed line. The six periods ahead of future forecast is shown in blue, the corresponding 95% confidence interval is shown in the shaded light blue. During training and testing, Random Forest shows a good fit with the actual data, reflecting good in-sample fitting and out-of-sample accuracy. A high fit between the true (black) and predicted (green dashed) lines demonstrates the ability of the model to incorporate the underlying non-linear trends, seasonality, and chamges in structure of the data. The generalization is evidenced by the small variance and spread of the prediction error during testing. The time series exhibits a number of structural breaks: Steep decline between 2020 and early 2021, probably due to exogenous factors such as the COVID-19 pandemic. This disturbance is reflected in the actual and forecast lines, demonstrating the model's responsiveness to abrupt structural changes. The fast post-pandemic recovery implied by the rapid increase in 2022 is also consistent with the model. The six-step forecasts, shown in blue, predict a gradual rise in the observed variable. The point estimates range from about 450,000 to 510,000. The fan-shaped confidence interval shows growing uncertainty over time, a typical aspect of time series forecasting. However, the relatively narrow band indicates stable model variance and confidence in its short-term predictive ability. The train-test split is marked by the red vertical dotted line. This line separates the model training phase, which is pre-2024, from the evaluation phase. The future forecasting window begins in late 2024 and continues into early 2025. It is marked by a purple vertical dotted line and a light gray background. These markers create a clear visual structure to assess time periods and ensure reproducibility in future studies.

Table 5. Random Forest Forecast Results and 95% Confidence Intervals for Six Future Periods

Period	Random Forest Forecast	Lower 95% CI	Upper 95% CI
2025-01-01	486870.6875	413385.5	560355.875
2025-02-01	426060.125	352574.9375	499545.3125
2025-03-01	434418.8438	360933.6563	507904.0313
2025-04-01	454332.0625	380846.875	527817.25
2025-05-01	495273.375	421788.1875	568758.5625
2025-06-01	532949.5	459464.3125	606434.6875

The six-step forecast generated by the Random Forest model shows a gradual upward trend in the observed variable from January to June 2025. It starts at about 486,871 in January. The values drop slightly in February before steadily increasing to a peak of 532,950 in June. This pattern suggests a potential short-term fluctuation, likely caused by seasonal effects or structural changes, followed by a strong recovery path. Each forecast point comes with a 95% confidence interval, showing the model's estimate of uncertainty for each prediction. The confidence bands widen slightly over time, creating a fan-shaped structure often seen in time series forecasting. Even with this widening, the intervals stay relatively narrow, averaging ±73,000. This shows that the model keeps stable variance and strong predictive reliability in the short term. The steady increase from March through June shows that the model understands the timing factors involved, including nonlinear growth patterns and possible hidden seasonal effects. The strength of the Random Forest algorithm in this situation is clear, as it can identify these trends and create precise forecast intervals. This improves our confidence in the accuracy of short-term forecasts. In general, these results confirm that the model can deliver dependable, data-driven predictions, even with structural changes. These insights are vital for planning and decision-making, especially in areas where timely and accurate forecasts are important.

V. CONCLUSIONS AND SUGGESTIONS

This study assessed the forecasting capabilities of Random Forest, XGBoost, LSTM, and a Hybrid RF-XGBoost model to predict monthly tourist arrivals to Bali, using a decade-long dataset. Random Forest achieved the most accurate predictions, reflected in the lowest RMSE of 41,772.68 and a MAPE of 6.30%. The Hybrid RF-XGBoost model delivered competitive results, while LSTM lagged, reinforcing the necessity of matching model choice to the dataset's characteristics. The findings highlight the robustness of ensemble methods in tourism forecasting, especially in destinations marked by seasonality and complex visitor patterns. Future research should:

1. Broaden the forecasting scope to a multivariate framework, incorporating external drivers such as flight schedules, currency fluctuations, and geopolitical events.
2. Examine advanced deep hybrid architectures, for instance, CNN-LSTM or Transformer combined with Random Forest, to extend the accuracy of long-term predictions.
3. Leverage explainable AI techniques to clarify model reasoning, thereby providing actionable insights that support evidence-based policymaking.

REFERENCES

[1] Y. Qian and Y. Zhang, "Long-term forecasting in asset pricing: Machine learning models' sensitivity to macroeconomic shifts and firm-specific factors," *North American Journal of Economics and Finance*, vol. 78, May 2025, doi: 10.1016/J.NAJEF.2025.102423.

[2] M. A. Afrianto and M. Wasesa, "The impact of tree-based machine learning models, length of training data, and quarantine search query on tourist arrival prediction's accuracy under COVID-19 in Indonesia," *Current Issues in Tourism*, vol. 25, no. 23, pp. 3854–3870, 2022, doi: 10.1080/13683500.2022.2085079.

[3] F. Antolini and S. Cesarini, "Predicting Domestic Tourists' Length of Stay in Italy leveraging Regression Decision Tree Algorithms," *Electronic Journal of Applied Statistical Analysis*, vol. 17, no. 3, pp. 621–635, 2024, doi: 10.1285/I20705948V17N3P621.

[4] Z. Chen, C. Ye, H. Yang, P. Ye, Y. Xie, and Z. Ding, "Exploring the impact of seasonal forest landscapes on tourist emotions using Machine learning," *Ecol Indic*, vol. 163, Jun. 2024, doi: 10.1016/J.ECOLIND.2024.112115.

[5] Z. Marzak, R. Benabbou, S. Mouatassim, and J. Benhra, "Forecasting Multivariate Time Series with Trend and Seasonality: A Random Forest Approach," *Communications in Computer and Information Science*, vol. 2373 CCIS, pp. 128–144, 2025, doi: 10.1007/978-3-031-80775-6_9.

[6] J. Zhao, C. Der Lee, G. Chen, and J. Zhang, "Research on the Prediction Application of Multiple Classification Datasets Based on Random Forest Model," *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems, ICPICS 2024*, pp. 156–161, 2024, doi: 10.1109/ICPICS62053.2024.10795875.

[7] L. Liu, J. Liu, and X. Zhu, "An augmentation of random forest through using the principle of justifiable granularity," p. 218, Jul. 2024, doi: 10.1117/12.3031228.

[8] S. Bhadula *et al.*, "Optimizing Random Forest Algorithms for LargeScale Data Analysis," *Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2024*, pp. 1673–1678, 2024, doi: 10.1109/IC3I61595.2024.10829145.

[9] Z. Li and T. Lu, "Prediction of Multistation GNSS Vertical Coordinate Time Series Based on XGBoost Algorithm," *Lecture Notes in Electrical Engineering*, vol. 910 LNEE, pp. 275–286, 2022, doi: 10.1007/978-981-19-2576-4_24.

[10] F. A. Nahid, M. N. Jahangir, H. M. Chowdhury, and K. Akter, "Evaluation and Performance Metrics for Forecasting Renewable Power Generation, Demand, and Electricity Price," *Forecasting Methods for Renewable Power Generation*, pp. 173–218, Jan. 2025, doi: 10.1002/9781394249466.CH7.

[11] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geosci Model Dev*, vol. 15, no. 14, pp. 5481–5487, Jul. 2022, doi: 10.5194/GMD-15-5481-2022.

[12] Z. Schwartz, J. Ma, and T. Webb, "The MSapeMER: a symmetric, scale-free and intuitive forecasting error measure for hospitality revenue management," *International Journal of Contemporary Hospitality Management*, vol. 36, no. 6, pp. 2035–2048, Apr. 2024, doi: 10.1108/IJCHM-01-2023-0088.

[13] "Badan Pusat Statistik Provinsi Bali." Accessed: Jul. 28, 2025. [Online]. Available: https://bali.bps.go.id/id

[14] W. Limpornchitwilai, P. Suksompong, C. Charoenlarpnopparut, L. O. Kovavisaruch, T. Sanpechuda, and M. Nakatani, "Improvement of the RF-Based and XGBoost-Based Visitor Data Prediction for Colocated Museums," *19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2022*, 2022, doi: 10.1109/ECTI-CON54298.2022.9795605.

[15] D. T. Andariesta and M. Wasesa, "Machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic: a multisource Internet data approach," *Journal of Tourism Futures*, 2022, doi: 10.1108/JTF-10-2021-0239.

[16] E. Brilliandy, H. Lucky, A. Hartanto, D. Suhartono, and M. Nurzaki, "Using Regression to Predict Number of Tourism in Indonesia based of Global COVID-19 Cases," *2022 3rd International Conference on Artificial Intelligence and Data Sciences: Championing Innovations in Artificial Intelligence and Data Sciences for Sustainable Future, AiDAS 2022 - Proceedings*, pp. 310–315, 2022, doi: 10.1109/AIDAS56890.2022.9918731.

[17] Y. Manzali and M. Elfar, "Random Forest Pruning Techniques: A Recent Review," *Operations Research Forum*, vol. 4, no. 2, Jun. 2023, doi: 10.1007/S43069-023-00223-6.

[18] A. Khanna, D. Goyal, N. Chaurasia, and T. H. Sheikh, "Forecasting Financial Success App: Unveiling the Potential of Random Forest in Machine Learning-Based Investment Prediction," *Lecture Notes in Networks and Systems*, vol. 788 LNNS, pp. 279–292, 2023, doi: 10.1007/978-981-99-6553-3_22.

[19] A. Sharma, A. M. Gupta, and S. Das, "Enhancing liver cirrhosis diagnosis: A comparative analysis of XGBoost, SVM, and random forest classifiers for optimal predictive analysis," *Computational Methods in Science and Technology - Proceedings of the 4th International Conference on Computational Methods in Science and Technology, ICCMST 2024*, vol. 1, pp. 130–135, 2025, doi: 10.1201/9781003501244-22.

[20] S. K. Khaleelullah, H. Kumar, G. Rahul, R. Naik, and S. Teja, "DietRx: Machine Learning Enhanced Disease Specific Nutrition and Precautions," *Proceedings - 2024 8th International Conference on Inventive Systems and Control, ICISC 2024*, pp. 340–344, 2024, doi: 10.1109/ICISC62624.2024.00065.

[21] K. Maji, S. Gupta, and P. K. Dutta, "Enhancing Heart Disease Prediction Accuracy: Comprehensive Analysis of XGBoost and AdaBoost," *IET Conference Proceedings*, vol. 2024, no. 37, pp. 130–136, 2024, doi: 10.1049/ICP.2025.0833.

[22] R. Sibindi, R. W. Mwangi, and A. G. Waititu, "A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices," *Engineering Reports*, vol. 5, no. 4, Apr. 2023, doi: 10.1002/ENG2.12599.

[23] H. Xiao, "Enhanced separation of long-term memory from short-term memory on top of LSTM: Neural network-based stock index forecasting," *PLoS One*, vol. 20, no. 6 June, Jun. 2025, doi: 10.1371/JOURNAL.PONE.0322737.

[24] J. Peeperkorn, S. vanden Broucke, and J. De Weerdt, "Can recurrent neural networks learn process model structure?," *J Intell Inf Syst*, vol. 61, no. 1, pp. 27–51, Aug. 2023, doi: 10.1007/S10844-022-00765-X.

[25] S. Boriratrit and R. Chatthaworn, "Improvement of Long Short-Term Memory via CEEMDAN and Logistic Maps for the Power Consumption Forecasting," *2023 15th International Conference on Advanced Computational Intelligence, ICACI 2023*, 2023, doi: 10.1109/ICACI58115.2023.10146172.

[26] Q. Kang, D. Yu, K. H. Cheong, and Z. Wang, "Deterministic convergence analysis for regularized long short-term memory and its application to regression and multi-classification problems," *Eng Appl Artif Intell*, vol. 133, Jul. 2024, doi: 10.1016/J.ENGAPPAI.2024.108444.

[27] Y. Ying, H. Qi, and L. Na, "Feature Selection and Efficient Disease Early Warning Based on Optimized Ensemble Learning Model:Case Study of Geriatric Depression and Anxiety," *Data Analysis and Knowledge Discovery*, vol. 7, no. 7, pp. 74–88, Jul. 2023, doi: 10.11925/INFOTECH.2096-3467.2022.0718.

[28] Y.A Jatmiko, R.A, Rahayu and G. Darmawan.Perbandingan Keakuratan Hasil Peramalan Produksi Bawang merah Metode Holt-Winters dengan Singular Spectrum Analysis (SSA).Jurnal Matematika"Mantik" Vol.03 No.01:13-23.