eduvest

Eduvest – Journal of Universal Studies
Volume 6 Number 6, June, 2026
p- ISSN 2775-3735- e-ISSN 2775-3727

Predicting Flight Delays Using LSTM and BILSTM Models with Shap Interpretation

Ifayanti Rohmatul Hidayah*, Irhamah, Santi Puteri Rahayu
Institut Teknologi Sepuluh Nopember, Indonesia
Email: ifayantii2301@gmail.com*, irhamah@its.ac.id, santi_pr@its.ac.id

Keywords
prediction, flight delay, lstm, bilstm, shap, interpretability

Abstract
Flight delays represent a critical challenge in air transportation, affecting passenger satisfaction, operational efficiency, and financial outcomes. This study develops predictive models for flight delay duration at Juanda International Airport, Surabaya, using Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) networks integrated with a Shapley Additive Explanations (SHAP) interpretability method. The research utilized 242,638 flight observations spanning January 2023 to October 2025, incorporating flight operational and meteorological variables. The dataset was partitioned into training (62.35%), validation (9.01%), and testing (28.64%) subsets. After Min-Max normalization and preprocessing, models were designed with varying hyperparameters through grid search optimization. Performance evaluation employed Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Results demonstrated that LSTM with two hidden layers and sixteen neurons achieved superior performance, with an MAE of 15.657 minutes and an RMSE of 20.859 minutes on test data, slightly outperforming BiLSTM (MAE of 16.705 minutes and RMSE of 21.709 minutes), establishing LSTM as the optimal model. SHAP interpretation revealed that operational factors, particularly the association with major airports and routes (Jakarta and Surabaya), flight type, and airline, represent the dominant predictors of delays, whereas meteorological factors such as wind speed and temporal factors such as scheduling time have a relatively minor effect. Although the model's predictive power remains limited, this research provides valuable interpretability insights into delay determinants, enabling data-driven decision-making for airport management and airlines to enhance punctuality and operational efficiency.

INTRODUCTION

. Transportation is the movement of people and goods from one place to another using various types of vehicles. In the modern era, transportation has become a primary necessity, and one of the fastest modes for moving people is air transportation. Air transportation facilitates connectivity between regions, even reaching areas that are difficult to access by other modes. The development of air transportation in Indonesia has increased significantly; however, this growth is also accompanied by the persistent challenge of flight delays in the industry (Aini et al., 2023; International Air Transport Association, 2021).

Flight delays impact not only passengers through travel inconvenience and disruption but also airlines through reduced operational efficiency and significant financial losses (Zámková *et al.* 2022; Ain 2024). Airlines are required to compensate passengers depending on the duration of delays, including meal provision, accommodation, and alternative flights. In addition, delays also affect airport slot management authorities in allocating runway and gate capacity efficiently (Britto *et al.* 2012; Wu 2016; Anupkumar 2023). Accurate delay prediction can help airlines reduce operational costs and enable airport authorities to optimize flight slot allocation, thereby improving overall airport utilization (Goodfellow et al., 2016; Jatavallabha et al., 2024; Li et al., 2023).

To address flight delay problems, an analytical approach is required to produce more accurate predictions of potential delays. Machine learning approaches, particularly deep learning, have proven effective in capturing temporal dependencies in time-series data. Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) are neural network architectures specifically designed to overcome the vanishing gradient problem in long sequential datasets (Ballakur & Arya, 2020; Choi, 2018; Cui et al., 2020; Depuru et al., 2024). However, a major limitation of deep learning models is their "black box" nature, which makes it difficult to interpret the contribution of each variable to predicted outcomes.

This study has four main objectives: (1) to develop and analyze LSTM and BiLSTM models and apply the SHAP method for interpretability; (2) to implement the LSTM model for predicting flight delay duration and evaluate its performance using MAE and RMSE metrics; (3) to implement the BiLSTM model for predicting flight delay duration and evaluate its performance; and (4) to identify the main variables contributing to flight delays through global and local interpretability analysis using SHAP. This research is expected to contribute to the advancement of deep learning-based prediction methods and improve model transparency through interpretability.

LSTM (Long Short-Term Memory) is an extension of Recurrent Neural Networks (RNNs) designed to address the vanishing gradient problem in long time-series data (Chen et al., 2020). LSTM introduces a cell state mechanism that functions as long-term memory and three primary gates: the forget gate (controlling information removal), the input gate (regulating new information storage), and the output gate (controlling the output information). This architecture enables LSTM to retain important information over long sequences and learn complex temporal dependencies between observations.

BiLSTM (Bidirectional Long Short-Term Memory) is an extension of LSTM that processes sequential data in two directions simultaneously: forward (from past to future) and backward (from future to past). This bidirectional structure allows the model to capture contextual information from both directions, improving its ability to model complex temporal patterns. In the literature, BiLSTM has demonstrated strong performance across various prediction tasks, although its higher complexity may increase the risk of overfitting, particularly in smaller datasets.

The SHAP (Shapley Additive Explanations) method is used to explain the contribution of each feature to model predictions in a fair and consistent manner. SHAP is based on cooperative game theory, specifically Shapley values, where each feature is treated as a "player" contributing to the final prediction. This approach decomposes predictions into additive contributions from each feature, thereby enhancing the interpretability of otherwise opaque deep learning models

METHOD

Data Sources and Research Variables

This study used secondary data consisting of flight operational and weather data from Juanda International Airport, Surabaya, covering the period from January 1, 2023, to October 31, 2025 (Hanke & Wichern, 2014). The dataset comprised 242,638 flight observations with 14 predictor variables and one target variable (delay duration in minutes) (Han et al., 2011). The predictors included operational factors such as flight type, airline, origin and destination airports, and scheduled departure time, as well as meteorological variables including temperature, precipitation, wind speed, and air pressure. The target variable represented the difference between actual arrival or departure time and the scheduled time, expressed in minutes..

Preprocessing and Data Preparation

The preprocessing stage included several steps: (1) merging operational and weather data based on observation timestamps; (2) encoding categorical variables using one-hot encoding; (3) removing missing values using the dropna method; (4) applying Min-Max normalization with a range of -1 to 1; and (5) splitting the dataset chronologically into training (January 1, 2023–September 30, 2024; 151,302 observations), validation (October 1, 2024–December 31, 2024; 21,853 observations), and testing sets (January 1, 2025–October 31, 2025; 69,483 observations).

Model Design and Hyperparameters

The LSTM and BiLSTM models were trained using grid search to identify optimal architectures. The hyperparameters included the number of hidden layers (1-3), neurons (8, 16, 32), batch size (64), learning rate (0.001), optimizer (Adam), dropout rate (0.2), and a maximum of 100 epochs. Early stopping with a patience of 10 epochs was applied based on validation loss to prevent overfitting and improve training efficiency.

RESULTS AND DISCUSSION

Descriptive Data Analysis

Descriptive statistical analysis shows that delays dominate flight operations at Juanda Airport. Of the 242,638 total flights, about three-quarters were delayed and one-quarter were on time. For arrivals, 89,306 flights (36.81%) were delayed and 32,100 (13.23%) were on time, while for departures 94,438 flights (38.92%) were delayed and 26,794 (11.04%) were on time. The delay duration distribution is zero-inflated and tends to be right-skewed, indicating a large number of low delay values with some much higher extremes reaching up to about 184 minutes. These data characteristics are important in the selection of data preprocessing and transformation strategies.

Table 1. Descriptive Statistics of Flight Delays

Flight Type	Status	N	Average (minutes)	Std Dev
Arrival	Delay	89,306	-28.066	18.935
Arrival	On Time	32,100	13.932	18.202
Departure	Delay	94,438	-25.661	18.664
Departure	On Time	26,794	23.844	18.919

LSTM Modeling Results

The best LSTM model was obtained with a hidden 2-layer architecture and 16 neurons per layer, reaching its best validation loss at epoch 62. The training process showed that train loss decreased consistently from about 0.095 to 0.068, while validation loss decreased from about 0.120 to around 0.080 with some fluctuation, indicating a proper early stop without sharp overfitting. On the testing data, the LSTM model produced an MAE of 15.657 minutes and an RMSE of 20.859 minutes, the lowest errors among all subsets.

BiLSTM Modeling Results

The best BiLSTM model was obtained with a hidden 2-layer architecture and 16 neurons per layer, reaching its best validation loss earlier, at epoch 29. On the testing data, the BiLSTM model produced an MAE of 16.705 minutes and an RMSE of 21.709 minutes. The performance of BiLSTM is slightly lower than that of LSTM across all subsets, suggesting that the additional complexity of bidirectional processing does not improve prediction in this case and tends to overfit earlier, producing more fluctuating predictions. The performance gap between the two models is nonetheless relatively small.

Table 2. Comparison of LSTM and BiLSTM Model Performance in Data Testing

Model	MAE (minutes)	RMSE (minutes)	Best Epoch
LSTM (2-16)	15.657	20.859	62
BiLSTM (2-16)	16.705	21.709	29

SHAP Interpretation

Interpretability analysis using SHAP on the best LSTM model identifies the features with the most influence on the predictions. Note that SHAP values are expressed in the model's output space (log-transformed and normalized scale), not in raw minutes. Globally, the association of a flight with Soekarno-Hatta Airport, Jakarta was the most dominant factor, occupying the two top ranks both as origin airport (mean |SHAP| of 0.0749) and as destination airport (0.0619), far exceeding other features and highlighting the role of the main Jakarta-Surabaya route. The next most important feature was flight type (0.0597), indicating that arrivals and departures have systematically different delay characteristics. Airline identity also showed a significant influence, with high-frequency carriers such as Lion Air (0.0572), Citilink (0.0568), and Batik Air (0.0446) contributing more than other airlines. Temporal factors such as scheduling time within the day and month (0.0312) contributed less, while the weather factor represented by wind speed (0.0290) ranked in the middle with a non-linear, relatively minor influence.

Table 3. Top 5 Most Influential Features Based on SHAP Value

Features	Average SHAP Score	Ranking	Category
origin_airport_WIII (Jakarta)	0.0749	1	Operational
destination_airport_WIII (Jakarta)	0.0619	2	Operational
flight_type	0.0597	3	Operational
acid_LNI (Lion Air)	0.0572	4	Operational
acid_CTV (Citilink)	0.0568	5	Operational

CONCLUSION

This study successfully developed and compared LSTM and BiLSTM models for predicting flight delay duration at Juanda International Airport, Surabaya. The LSTM model with two hidden layers and sixteen neurons achieved the best performance, with a mean absolute error (MAE) of 15.657 minutes and a root mean squared error (RMSE) of 20.859 minutes on the test dataset, slightly outperforming BiLSTM (MAE of 16.705 minutes and RMSE of 21.709 minutes). The integration of SHAP provided meaningful interpretability by identifying the association with major airports and routes (Jakarta and Surabaya), flight type, and airline as the dominant factors influencing delays, whereas weather factors such as wind speed had a relatively minor effect.

This research contributes to understanding the determinants of flight delays and provides a practical framework for implementing predictive models in the aviation industry. Future research is recommended to: (1) explore advanced data preprocessing techniques, including denoising methods to improve data quality; (2) evaluate the model on more diverse datasets across different airports and time periods; (3) address zero-inflated and right-skewed data distributions more explicitly; (4) enhance hyperparameter optimization strategies to improve predictive performance; and (5) incorporate additional features such as air traffic conditions and more detailed operational variables.

The findings of this study may provide strategic input for Juanda International Airport and airlines in improving flight punctuality, operational efficiency, and passenger satisfaction.

REFERENCE

Ain, N. (2024). Navigating passenger compensation: Implications for airlines and consumers. *Journal of Air Law & Commerce, 89*(3), 507.

Aini, H., School, N., Tekhnologi, T., & Yogyakarta, K. (2023). Analysis of delay management due to weather related to operational technicalities at Citilink Airlines at Komodo Labuan Bajo Airport. *Journal of General Studies and Research, 1*(4), 71–83.

Anupkumar, A. (2023). Investigating the costs and economic impact of flight delays in the aviation industry and the potential strategies for reduction.

Ballakur, A. A., & Arya, A. (2020). Empirical evaluation of gated recurrent neural network architectures in aviation delay prediction. *IEEE Transactions on Aerospace and Electronic Systems, 56*(3), 1890–1902.

Britto, R., Dresner, M., & Voltes, A. (2012). The impact of flight delays on passenger demand and societal welfare. *Transportation Research Part E: Logistics and Transportation Review, 48*(2), 460–469.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*(3), 1247–1250.

Chen, H., Jiang, L., Yang, H., Lu, Z., Fu, Y., Li, L., & Yu, Z. (2020). An efficient hardware architecture with adjustable precision and extensible range to implement sigmoid and tanh functions. *Electronics, 9*(10), 1739.

Choi, C. (2018). Time series prediction with recurrent neural networks in presence of missing data. *Journal of Machine Learning Research, 19*(1), 1–25.

Cui, Z., Ke, R., Pu, Z., & Wang, Y. (2020). Stacked bidirectional and unidirectional LSTM recurrent neural network for predicting network-wide traffic state with missing values. *Transportation Research Part C: Emerging Technologies, 118*, 102674.

Depuru, S., Sirisala, S., Akuthota, K., Suresh Reddy, B. V., Amala, K., & Sivanantham, S. (2024). Enhancing flight delay prediction and classification using a hybrid Bi-LSTM. *Communications on Applied Nonlinear Analysis, 31*(6S), 15–28.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Elsevier.

Hanke, J. E., & Wichern, D. W. (2014). *Business forecasting* (9th ed.). Pearson Education Limited.

International Air Transport Association. (2021). *Economic performance of the airline industry*. International Air Transport Association.

Jatavallabha, A., Gerlach, J., & Naresh, A. (2024). Deciphering air travel disruptions: A machine learning approach. *arXiv*. https://arxiv.org/abs/2408.02802

Li, J., Ji, R., Li, C., Yang, X., Li, J., Li, Y., Xiong, X., Fang, Y., Ding, S., & Cui, T. (2023). Prediction of flight arrival delay time using U.S. Bureau of Transportation Statistics. In

2023 IEEE Symposium Series on Computational Intelligence (pp. 603–608). IEEE.

Wu, C.-L. (2016). *Airline operations and delay management: Insights from airline economics, networks and strategic schedule planning*. Routledge.

Zámková, M., Rojík, S., Prokop, M., & Stolín, R. (2022). Factors affecting international flight delays and their impact on airline operation and management and passenger compensation fees in air transport industry: Case study of selected airlines in Europe. *Sustainability, 14*(22), 14763.