A hybrid model to mitigate data gaps and fluctuations in tax revenue forecasting

Rahman Taufik, Aristoteles, Igit Sabda Ilman
Department of Computer Science, Faculty of Mathematics and Natural Sciences, University of Lampung, Bandar Lampung, Indonesia

Article Info

Article history:

Received Sep 4, 2024
Revised Apr 8, 2025
Accepted May 25, 1934

Keywords:

Forecasting
Hybrid model
Long short-term memory
Random forest regressors
Tax revenue

ABSTRACT

This study addresses the critical challenge of advancing tax revenue forecasting models to effectively handle distinctive data gaps and inherent fluctuations in tax revenue data. These challenges are evident in Lampung Province, Indonesia, where limited temporal granularity and non-linear variability hinder accurate fiscal planning. Despite advancements in statistical, machine learning, and hybrid approaches, existing models often fall short in simultaneously managing these challenges. A hybrid model integrating random forest regressors for data interpolation and long short-term memory (LSTM) for capturing complex temporal patterns was proposed. The model was evaluated, achieving an R^2 of 0.86, root mean squared error (RMSE) of 9.65 billion, and mean absolute percentage error (MAPE) of 3.49%. Although the model has limitations in generalizing to unseen data, the results demonstrate that it outperforms existing forecasting models regarding accuracy and reliability. Integrating random forest regressors and long short-term memory delivers a tailored solution to the complexities of tax revenue forecasting, contributing to fiscal forecasting and setting a foundation for further exploration into hybrid approaches.

Corresponding Author:

Rahman Taufik
Department of Computer Science, Faculty of Mathematics and Natural Sciences, University of Lampung
35145 Bandar Lampung, Indonesia
Email: rahman.taufik@fmipa.unila.ac.id

1. INTRODUCTION

Tax revenue is an important component of public finance, supporting infrastructure development, public services, and economic growth. In Lampung Province, Indonesia, tax revenue represents a significant component of the regional budget, providing various development initiatives. However, fluctuations in financial conditions, high poverty levels, limited fiscal capacity, and reforms in local taxation systems often lead to discrepancies between projected and actual tax revenues [1], [2]. A preliminary analysis of tax revenue data in Lampung Province from 1995 to 2023 reveals a trend-driven growth pattern, with an average annual growth rate of 21.87% and a variance of 1.17%. Despite this growth, a high standard deviation of 91.28% of the mean indicates substantial variability, compounded by seasonal impacts contributing 24.34% to the overall variation. These findings highlight the complexity of long-term growth trends, seasonal fluctuations, and non-linear variability in tax revenue data. The challenge of forecasting is further compounded by the data gaps, including limited temporal granularity, which refers to the availability of only 28 years of annual intervals. These challenges underscore the need for advanced forecasting models to address data gaps and manage inherent fluctuations in tax revenue data.

Numerous studies have developed tax revenue forecasting models across various regions and explored different model extensions. For instance, Streimikiene *et al.* [3] utilized an autoregressive integrated moving average (ARIMA) model to forecast total tax revenue, identifying significant variables affecting tax revenue in Pakistan. Buxton *et al.* [4] demonstrated that multi-layer perceptron (MLP) outperformed the long short-term memory (LSTM) in predicting tax for the state of Illinois. Lahiri and Yang [5] adopted a mixed-frequency vector autoregression (MF-VAR) model to improve New York State tax revenue forecasts, especially in response to the volatility introduced by the coronavirus disease (COVID-19) pandemic. In Wenzhou City, Xie [6] used multiple linear regression (MLR) and MLP models to forecast tax revenue, while also analyzing taxation factors. Segun [7] examined and compared MLR, seasonal autoregressive integrated moving average (SARIMA), and LSTM, utilizing multiple independent variables to forecast tax revenue in Nigeria. In the same region, Tasi'u *et al.* [8] demonstrated that SARIMA outperformed the Holt-Winters model, also known as triple exponential smoothing (TES). In Lampung Province, Kurniasari and Ramadhani [9] encountered challenges in accurately forecasting a specific tax revenue due to complex, non-linear data patterns. This led them to propose an artificial neural network (ANN) model that achieved high accuracy. However, in a similar case, Infusi *et al.* [10] found that MLR was more accurate than the ANN model. While these studies applied forecasting models to address regional tax revenue challenges, they predominantly relied on prevailing methodologies and lacked the sophistication to enhance accuracy across diverse temporal scales, leaving forecasting reliability unaddressed.

Subsequent advances in tax revenue forecasting have focused on improving the accuracy and robustness of forecasting models through advanced methodologies, including hybrid approaches. Ilic *et al.* [11] proposed the explainable boosted linear regression (EBLR) algorithm that leverages daily data granularity to improve model performance by sequentially incorporating variables, achieving a normalized root mean squared error (RMSE) of 0.1528. Thayyib *et al.* [12] employed the theta trigonometric seasonality Box-Cox transformation ARIMA errors trend seasonal components (TBATS) model for forecasting goods and services tax revenue in India. It was shown to be more accurate than neural network models, with an RMSE of 0.141. Fathoni and Saputra [13] employed the ARIMA Box-Jenkins model to forecast value-added tax (VAT) revenue in Indonesia. They demonstrated that the resulting forecast closely aligned with actual VAT revenue, exhibiting an RMSE of 2.765. Extensive studies on hybrid models demonstrated potential in integrating feature selection and temporal modeling. Ticona *et al.* [14] presented a hybrid model combining genetic algorithms and neural networks to forecast tax revenue in Brazil, achieving more accurate results with a mean absolute percentage error (MAPE) of 2.37% and a significant reduction in relative error. Smyl [15] introduced a dynamic computational neural network system that integrates exponential smoothing with LSTM, addressing issues related to temporal resolution, and forecasting monthly, annual, and quarterly data. However, this system was less effective for daily and weekly data. Ferdoush *et al.* [16] proposed a hybrid model that combines random forest for feature selection with bidirectional long short-term memory (BiLSTM) for forecasting. This approach yielded an RMSE of 0.4090, demonstrating superior performance compared to traditional LSTM models. Hossain and Ismail [17] proposed a hybrid model combining exponential autoregressive with Markov-switching generalized autoregressive conditional heteroskedasticity (MSGARCH) to address the issues of volatility and nonlinearity in financial time series. Their findings indicated that this model outperformed ARIMA and MSGARCH models, particularly in capturing downside risk. Larroussi *et al.* [18] introduced a hybrid deep learning model that integrates autoencoders with stacked LSTM to enhance the accuracy of tourism demand forecasting. This approach resulted in an R-squared (R2) of 96.52 and RMSE of 0.000992, demonstrating the robustness of this model. Despite these advancements, existing studies have primarily focused on improving either modeling temporal patterns [12]–[14], [16]–[18] or data granularity individually [11], [15], leaving a critical gap in unified approaches that simultaneously address irregularities in data and temporal inconsistencies in tax revenue forecasting. This gap underscores the need for a hybrid approach capable of resolving both challenges simultaneously.

The objective of this study is to propose a forecasting approach that simultaneously addresses both data gaps and fluctuations in tax revenue data, which existing studies have yet to address comprehensively. The proposed hybrid model combines the strengths of random forest regressor (RFR) for data interpolation and long short-term memory (LSTM) for forecasting tax revenue, offering a solution to these challenges. RFR is an effective model for addressing complex data interpolation challenges in various fields. For example, Achite *et al.* [19] successfully predicted annual rainfall in northern Algeria, achieving a high R^2 of 0.9524. Sahoo *et al.* [20] demonstrated that RFR outperformed traditional models such as Kriging and inverse distance weighting in estimating crop yields with finer spatial resolution. Song *et al.* [21] also showed that RFR significantly enhanced the resolution of nutrient distribution maps through effective data interpolation. Therefore, RFR is proposed to improve the granularity of interpolated values from annual to monthly data. In contrast, LSTM has been recognized as a prominent approach in time series forecasting for its ability to model long-term dependencies, as demonstrated in studies such as [4], [7], and [15]. The reviewed studies indicate that LSTM is adept at capturing intricate historical data and challenging temporal dynamics [22], making it well-suited for applications in tax revenue forecasting.

This study introduces a hybrid forecasting model that integrates RFR for effective data interpolation and LSTM for capturing complex temporal patterns, addressing both data gaps and fluctuations in tax revenue. By combining these methods, the model enhances the accuracy and reliability of tax revenue predictions, specifically for Lampung Province. This integrated approach overcomes the limitations of previous models, which typically do not comprehensively address the forecasting challenge, and offers a framework applicable to similar fiscal forecasting challenges globally.

2. METHOD

This section describes the methodology used in the study. It outlines the research approach and explains the proposed method. The model integrates RFR for interpolation and LSTM for time series forecasting.

2.1. Research design

Figure 1 illustrates the research design employed in this study, outlining the steps involved in facilitating the proposed RFR-LSTM hybrid model for tax revenue forecasting. The process begins with data collection from the Lampung Province Central Statistics Agency [23], covering tax revenue data from 1995 to 2023. The dataset was sourced from government reports and then converted into CSV format, resulting in 28 rows, with columns representing the year and corresponding tax revenue values.

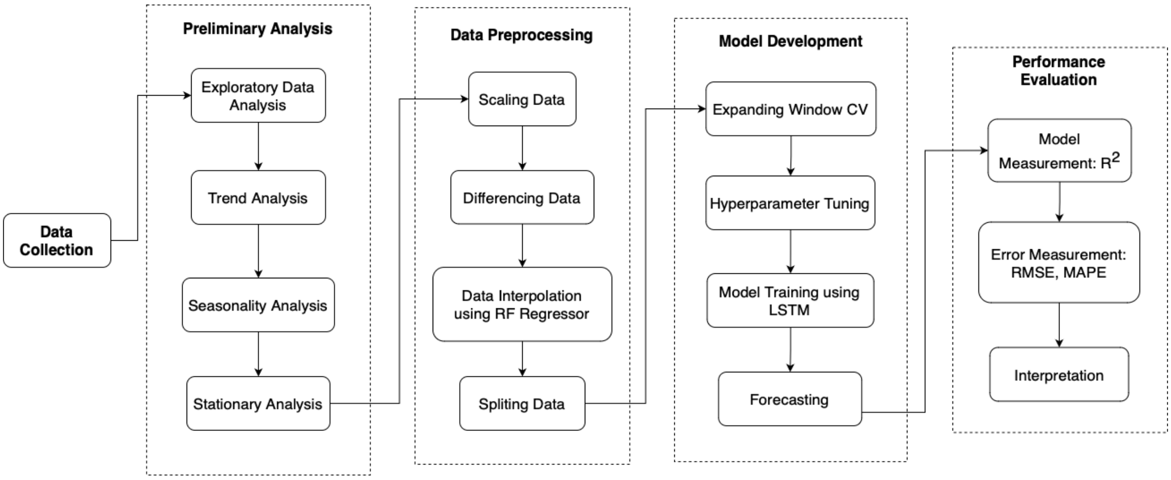

Figure 1. Research design

Preliminary analysis is conducted to explore the characteristics of the data. This includes exploratory data analysis to identify anomalies, trend analysis to uncover long-term patterns, seasonality analysis to detect recurring fluctuations, and stationarity analysis to verify consistent statistical properties over time. Next, data preprocessing is carried out to enhance data quality. Scaling is performed using a robust scaler, which is chosen for its reliability in handling outliers using the median and interquartile range instead of the mean and standard deviation. Differencing addresses non-stationary trends using first-order differencing, with stationarity assessed through the augmented Dickey-Fuller (ADF) test. Then, RFR is applied to interpolate annual data into monthly values, enabling finer temporal resolution [19]–[21]. The dataset is then split into 1995 to 2017 for training (approximately 80%) and 2018 to 2023 for testing (approximately 20%), ensuring that the test set represents unseen data.

Furthermore, the LSTM model is trained on the preprocessed monthly training data. Expanding window cross-validation with five folds is applied to the training data using TimeSeriesSplit. By gradually increasing the size of the training set while keeping a separate validation set, this technique enables the model to capture temporal patterns and adapt to sequential data effectively. Moreover, it maintains the temporal integrity of the dataset and minimizes data leakage, ensuring the validation data remains unseen during training [24], [25]. In addition, hyperparameter tuning is performed using the Keras Tuner library [26] to optimize the model configuration and computational efficiency. Furthermore, the model is evaluated on the test data to assess its generalization ability.

The methodology concludes with the evaluation of the proposed model. The performance of the proposed hybrid model was compared to various benchmark methods, including non-hybrid approaches such as ANN, MLR, LSTM, and ARIMA, as well as hybrid approaches like RFR-ARIMA and RFR-TES. These models were selected based on their extensive use in practical time-series problems [3], [4], [6]–[8] and prior studies related to forecasting Lampung tax revenue [9], [10]. All comparisons were performed using the same dataset. Moreover, these performances are evaluated using MAPE, RMSE, and R2, which are selected for their ability to measure relative accuracy, average error magnitude, and variance explanation, highlighting their relevance to assessing both irregularities in data and temporal inconsistencies in tax revenue forecasting. [27], [28].

2.2. Proposed method

A hybrid research model is proposed to improve the accuracy of tax revenue forecasting by combining two advanced techniques, namely RFR and LSTM. The RFR focuses on transforming low-frequency annual data into high-frequency monthly data, addressing data gaps. While, the LSTM leverages its ability to capture complex temporal dependencies, providing a robust approach for predicting tax revenue patterns with enhanced precision.

RFR effectively addresses data gaps through accurate interpolation, managing non-linear data patterns, and improving stability by reducing overfitting through ensemble learning [19]–[21]. This model combines the predictions of multiple decision trees, each trained on a random subset of the data and features [29]. In the proposed model, the annual tax revenue data x_i serves as input to each decision tree, as expressed in formula (1):

$$\widehat{y_i} = f(x_i; \theta_k) \tag{1}$$

where $\widehat{y_i}$ is the prediction from the k-th tree, and θ_k represents the parameters of that tree. The final prediction from the RFR is obtained by averaging the predictions of all trees in the forest, as shown in formula (2):

$$\hat{y} = \frac{1}{n_{\text{trees}}} \sum_{i=1}^{n_{\text{trees}}} \widehat{y_i} \tag{2}$$

where $\widehat{y_i}$ represents the prediction from each individual tree, n_{trees} represents the number of decision trees in the ensemble, set to 100 based on initial experiments, and $\hat{y}$ represents the monthly interpolated value. These RFR monthly interpolated values serve as input to the LSTM model.

LSTM is well-suited for time series forecasting due to its ability to model long-term dependencies and complex temporal patterns [22]. As outlined by Goodfellow *et al.* [30], this process involves maintaining information from previous time steps, enabling the model to recognize long-term patterns in the data through a network of gates that regulate the flow of information. The following is an explanation of the gates in LSTM:

$$i_t = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i) \tag{3}$$

$$f_t = \sigma\left(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f\right) \tag{4}$$

The input gate, formula (3), determines how much new input should be added to the cell state. While, the forget gate, formula (4), is responsible for deciding how much of the previous cell state should be retained. Here, i_t represents the input gate, f_t is the forget gate, σ is the sigmoid activation function, W and U are the weight matrices associated with the current input x and the previous hidden state h_{t-1}, and b is the bias term. Next, the cell state is updated by combining the candidate cell state $\widetilde{C_t}$ with the results from the input and forget gates, as calculated in formulas (5) and (6):

$$\widetilde{C_t} = \tanh(W_c \cdot x_t + U_c \cdot h_{t-1} + b_c) \tag{5}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C_t} \tag{6}$$

Formula (5) calculates the candidate cell state $\widetilde{C_t}$, where $\tanh$ is the hyperbolic tangent activation function. This candidate cell state represents the new content that could be added to the cell state. Formula (6) then updates the cell state C_t by combining the previous cell state C_{t-1}, modulated by the forget gate f_t, with the candidate cell state $\widetilde{C_t}$, modulated by the input gate i_t.

Finally, the output gate determines the output of the LSTM cell, which also serves as the hidden state for the next time step. This is described in formulas (7) and (8):

$$o_t = \sigma(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o) \tag{7}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{8}$$

In formula (7), o_t is the output gate, which controls how much of the cell state should be exposed to the output. Formula (8) calculates the hidden state h_{t-1}, a function of the current cell state C_t and the output gate o_t. This process enables the LSTM model to process sequential data, uncover temporal relationships, and generate forecasts for subsequent monthly tax revenue values as the final output.

This hybrid approach combines the interpolative capabilities of RFR with the temporal modelling strengths of LSTM to address data gaps and fluctuations. By aligning forecasts with key economic patterns, the method offers a structured framework for reliable tax revenue prediction. These insights provide actionable guidance, ensuring the model is accessible and practical for fiscal policymakers.

3. RESULTS AND DISCUSSION

This section presents the research results along with a comprehensive discussion, organized into sub-sections, including the proposed hybrid model findings and comparisons to other models.

3.1. Interpretation of model findings

The proposed hybrid model integrates RFR for interpolating annual tax revenue data into monthly values, as illustrated in Figure 2. The model was trained using historical annual tax data and derived time-based features within an expanding window framework. The interpolation was performed using specific parameters, namely n_estimators set to 100, a random state of 24, a noise factor of 0.01, and an expanding window size of 4 months. These parameters were selected based on iterative experimentation to optimize accuracy and effectively handle variations in the data while maintaining consistency in capturing temporal dynamics. To ensure accurate interpolation, the performance was evaluated through key metrics, including R², MAPE, and RMSE. An R² of 0.9174 indicates that the model successfully captured 91.74% of the variance in the data, while the RMSE of approximately 9.92 billion reflects a low average error between the interpolated values and the actual data. The MAPE of 0.9% highlights the precision with which the model produces accurate interpolation values. Moreover, the validity of the interpolation is confirmed by the aggregated monthly interpolated data, which aligns closely with the original annual dataset, supporting its use for reliable tax revenue forecasting.

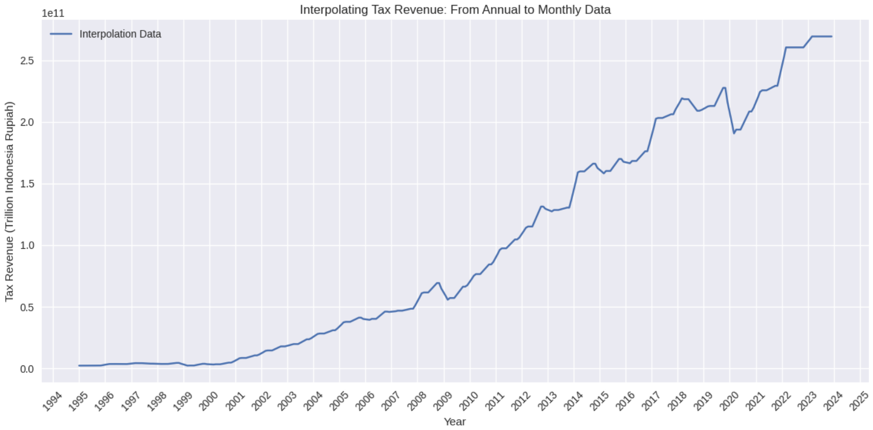

Figure 2. The interpolation data presented in a transformation from annual to monthly data

Subsequently, the monthly interpolated data was modeled using the LSTM model. The optimal configuration comprises 288 units, a dropout rate of 0.4, 176 dense units, the Adam optimizer, and a learning rate of 0.001. Expanding window cross-validation was applied during the training process, with the model trained for 100 epochs. This combination of hyperparameters and cross-validation is particularly effective for time-series analysis. It ensures that the model is progressively trained on more data, enhancing its ability to capture long-term trends and patterns.

Figure 3 demonstrates the prediction performance of the combined training and testing model. The performance of the proposed hybrid model during training was evaluated using several metrics. The R² of 0.897 indicates that approximately 89.7% of the variance in the data was captured, reflecting a strong model fit. However, the relatively high MAPE of 6.29% on the training data suggests that while the model fits well overall, it may still have noticeable prediction errors in specific cases. This is further confirmed by the RMSE of approximately 3.52 billion, representing the average error in the predictions. Notably, the RMSE on the test set increased to 9.66 billion, indicating that the model faces challenges in generalizing to unseen data. The R² for the unseen data dropped to 0.86, still showing a strong fit, but indicating that the performance on new data is slightly less robust. Despite this, the MAPE on the test data improved to 3.49%, suggesting that the model maintains reasonable predictive accuracy. Although the proposed hybrid model performs well on the training data, its ability to generalize to unseen data remains limited. This limitation is likely due to external economic disruptions from 2018 to 2019 that were not captured in the training data. These disruptions caused significant fluctuations, highlighting a potential limitation of the interpolation model, as it struggled to adapt to large-scale changes in the economic landscape and generated inconsistent values.

In addition, Figure 4 demonstrates that the proposed hybrid model is promising in forecasting tax revenue for Lampung Province, with projections showing a 4.08% increase for 2024 and 3.77% for 2025. While this study focuses on data from Lampung Province, the proposed hybrid model can be adapted to regions with similar tax revenue characteristics, requiring modifications only for those with differing tax structures. For instance, the RFR interpolation method needs to be adapted to account for varying tax data formats and regional economic indicators, while the LSTM needs to be reconfigured to reflect local economic cycles and seasonal trends. Differences in tax data quality and granularity across regions affect model performance and necessitate further refinement in data preprocessing and feature selection.

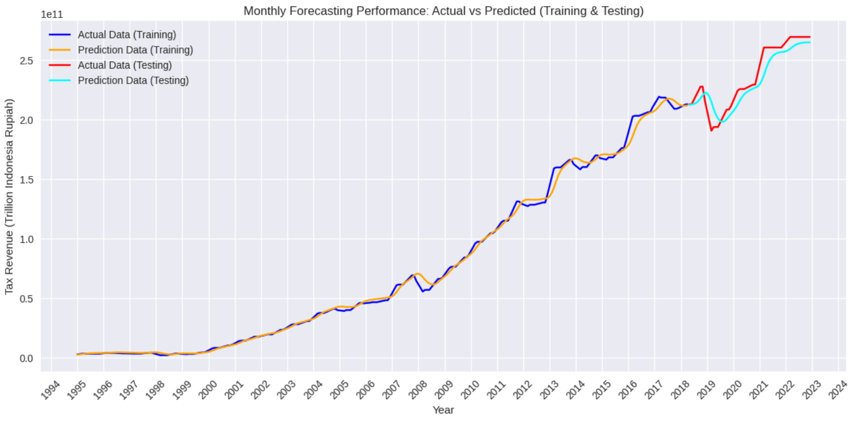

Figure 3. Performance of the proposed model on both the training and testing data sets

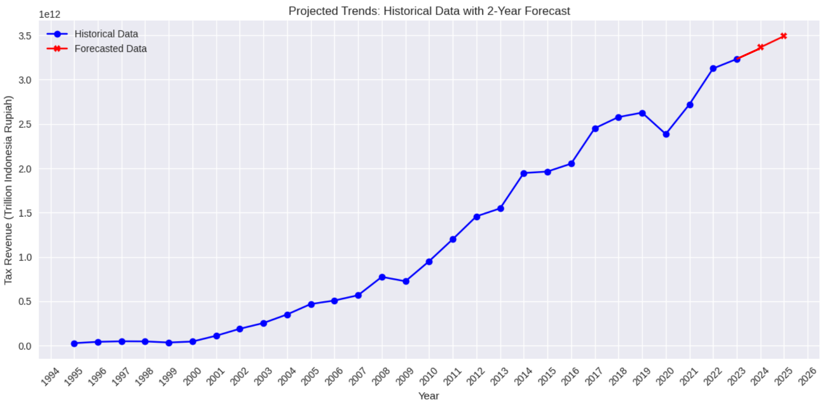

Figure 4. Forecasted tax revenue data for years 2024 and 2025

3.2. Comparative models discussion

The performance of the proposed hybrid model was compared with various benchmark models using insights from existing research and widely used forecasting methods. Benchmark models such as ANN and MLR were selected based on their prior studies in forecasting Lampung tax revenue [9][10]. Additional models, including LSTM and ARIMA, were chosen for their extensive use in practical time-series problems. Hybrid approaches, including RFR-ARIMA and RFR-TES, were tested in this study to explore the potential of combining regression and time-series methods. By comparing these models, this study aims to evaluate the proposed hybrid models effectiveness in addressing data gaps and fluctuations, relative to established approaches and newly explored combinations.

The selected models each offer distinct advantages based on their underlying methodologies. ANN is widely recognized for its ability to capture complex nonlinear patterns in data [9]. MLR is a classical statistical approach for identifying linear relationships between variables, making it a reliable baseline for structured datasets [10]. LSTM is designed to analyze sequential data, allowing it to capture long-term temporal patterns and dependencies critical for time-series forecasting [7], [22]. Hybrid approaches, such as RFR-ARIMA and RFR-TES, combine RFR for data interpolation [19]-[21] with ARIMA for modeling linear temporal structures [13], [17] and TES for addressing seasonal variations through exponential smoothing [8]. The comparative methodology ensures that all models are trained and tested using the same tax revenue dataset from Lampung Province. The dataset was preprocessed using robust scaling and differencing data to enhance data quality and stability. To avoid data leakage, the dataset was divided into 80% for training (1995 to 2017) and 20% for testing (2018 to 2023), maintaining the order of data over time. The training set was further evaluated using Expanding Window Cross-Validation, progressively increasing the size of the training set while reserving a portion for validation. All models were assessed using key metrics, including MAPE, RMSE, and R². This consistent framework ensures consistency and comparability, allowing reliable conclusions about the performance of each model.

The results, summarized in Table 1, demonstrated performance variations across hybrid and non-hybrid models. Among all the models tested, the proposed RFR-LSTM hybrid model demonstrates the best performance, with R² of 0.86, RMSE of 9.65 billion, and MAPE of 3.49%. The high R² indicates that the model can explain most of the variability in the data, while the low RMSE and MAPE suggest that the average prediction error is relatively small. This indicates that the proposed RFR-LSTM hybrid model effectively captures long-term temporal trends and fluctuations, enabling accurate and stable predictions. Moreover, comparisons with other hybrid models, including RFR-ARIMA and RFR-TES, validate the performance of the proposed RFR-LSTM hybrid model. RFR-ARIMA, with R² of 0.61, RMSE of 28.77 billion, and MAPE of 8.79%, shows moderate performance due to its inability to capture the nonlinear patterns in the data. The RFR-TES model, with R² of 0.18, RMSE of 107.02 billion, and MAPE of 41.16%, exhibits the poorest performance. This is primarily attributed to its inability to model the complex, nonlinear trends and fluctuations in the data. These findings highlight that compared hybrid models were less effective in capturing long-term temporal patterns and handling the complexities of Lampung's tax revenue data, underscoring the advantages of the proposed RFR-LSTM hybrid model.

Table 1. Performance comparison of forecasting models

Model	R²	RMSE	MAPE
RFR-LSTM	0.86	9657649496.151793	3.49%
RFR-ARIMA	0.61	28768464073.80	8.79%
RFR-TES	0.18	107019336268.7519	41.16%
LSTM	0.49	19714057348.485973	7.13 %
ANN	0.53	17298174837.55586	6.58 %
MLR	0.71	312386439156.10287	5.1%

In addition, non-hybrid or single models, including LSTM, ANN, and MLR, were evaluated. The LSTM model demonstrated a moderate level of performance, with R² of 0.49, RMSE of 19.71 billion, and MAPE of 7.13%. The ANN model, with R² of 0.53, RMSE of 17.30 billion, and MAPE of 6.58%, performs similarly to LSTM. In contrast, MLR demonstrated better performance in terms of R², achieving R² of 0.71, RMSE of 31.23 billion, and MAPE of 5.1%. However, the relatively high RMSE and MAPE suggest that while MLR performs well in explaining variance, it was less effective in accurately predicting the values. These results confirm that while non-hybrid models provide valuable estimates, they are limited in their ability to capture the fluctuations within the specified data constraints. By combining interpolation and temporal pattern modeling, the proposed RFR-LSTM hybrid model more effectively addresses these complexities and provides more accurate forecasts.

Overall, the comparisons reveal that while single and hybrid models have specific strengths, the RFR-LSTM hybrid model provides a more comprehensive solution by addressing data gaps and fluctuations simultaneously. This highlights its potential to unify different models for dependable and precise tax revenue forecasting. Despite the promising results, further improvements are needed to refine regularization techniques, integrate multiple methodologies, and enhance model generalization to handle the challenges of generalizing to unseen data. Nevertheless, the proposed hybrid model has strong potential for further refinement and broader application in forecasting tax revenue.

4. CONCLUSION

This study developed an RFR-LSTM hybrid model to improve tax revenue forecasting in Lampung Province by integrating the interpolation capabilities of RFR with the temporal pattern-capturing function of LSTM. The model demonstrated robustness in tax revenue forecasting, achieving high predictive accuracy with an R^2 of 0.86, RMSE of 9.65 billion, and MAPE of 3.49%. However, the proposed hybrid model's ability to generalize to unseen data remains a key limitation, as indicated by the increase in RMSE and decrease in R^2 when tested on new data. These findings underscore the importance of improving model robustness and generalization. Future research should explore techniques to enhance generalization, such as incorporating additional data sources or refining interpolation methods. Furthermore, investigating the applicability of this approach to other regions or countries with different tax structures could provide valuable insights on how to adapt the model for broader use. Overall, this study contributes to advancing forecasting methods for tax revenue prediction and provides a foundation for future improvements in predictive modeling.

ACKNOWLEDGMENTS

This work was supported by Universitas Lampung and Regional Development Planning Agency of Lampung Province.

FUNDING INFORMATION

Financial support for this research was provided by Daftar Isian Pelaksanaan Anggaran Badan Layanan Umum (DIPA BLU) Universitas Lampung.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Rahman Taufik	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			
Aristoteles	✓	✓		✓		✓	✓			✓		✓		✓
Igit Sabda Ilman			✓		✓	✓		✓	✓		✓		✓	

C :	**C**onceptualization	I :	**I**nvestigation	Vi :	**Vi**sualization
M :	**M**ethodology	R :	**R**esources	Su :	**Su**pervision
So :	**So**ftware	D :	**D**ata Curation	P :	**P**roject administration
Va :	**Va**lidation	O :	Writing - **O**riginal Draft	Fu :	**Fu**nding acquisition
Fo :	**Fo**rmal analysis	E :	Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study were acquired from publicly accessible sources, specifically the Regional Development Planning Agency of Lampung Province and the Central Bureau of Statistics through their website https://lampung.bps.go.id.

REFERENCES

[1] M. Gumanti, F. Fauzi, and C. Jatiningrum, “The analysis of regional income on economic growth Lampung Province,” *IJEBD (International Journal of Entrepreneurship and Business Development)*, vol. 5, no. 6, pp. 1036–1046, Nov. 2022, doi: 10.29138/ijebd.v5i6.2010.

[2] A. Fitria and A. Ambya, “The Effect of Local Goverment Revenue and Transfer Funds on Fiscal Capacity Ratio in Lampung Province from 2017-2022,” *Sinomics Journal*, vol. 2, no. 6, pp. 1525–1532, 2024, doi: 0.54443/sj.v2i6.246.

[3] D. Streimikiene, R. Raheem Ahmed, J. Vveinhardt, S. P. Ghauri, and S. Zahid, “Forecasting tax revenues using time series techniques–a case of Pakistan,” *Economic Research-Ekonomska Istrazivanja* , vol. 31, no. 1, pp. 722–754, Jan. 2018, doi: 10.1080/1331677X.2018.1442236.

[4] E. Buxton, K. Kriz, M. Cremeens, and K. Jay, “An auto regressive deep learning model for sales tax forecasting from multiple short time series,” in *Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*, Dec. 2019, pp. 1359–1364. doi: 10.1109/ICMLA.2019.00221.

[5] K. Lahiri and C. Yang, “Boosting tax revenues with mixed-frequency data in the aftermath of COVID-19: The case of New York,” *International Journal of Forecasting*, vol. 38, no. 2, pp. 545–566, Apr. 2022, doi: 10.1016/j.ijforecast.2021.10.005.

[6] H. Xie, “Research on the Models for Forecast of Tax Revenue of Wenzhou City,” *Highlights in Science, Engineering and Technology*, vol. 88, pp. 1043–1049, Mar. 2024, doi: 10.54097/dg7x7t56.

[7] A. S. Ajisola, “An efficient time series model for tax revenue forecasting: A case of Nigeria,” M.S. thesis, Department of Computer Science, African University of Science and Technology, 2023.

[8] T. Musa, A. M. Usman, and H. D. Garba, “Modelling and forecasting Nigeria’s tax revenue: A comparative analysis of SARIMA and Holt-Winters models,” *UMYU Scientifica*, vol. 3, no. 3, Jul. 2024, doi: 10.56919/usci.2433.014.

[9] D. Kurniasari, P. S. Ramadhani, W. Wamiliana, and W. Warsono, “Application of the artificial neural network algorithm to predict the realization of the duty tax on the name of motor vehicles in Lampung Province,” *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 7, no. 2, p. 392, Jun. 2024, doi: 10.24014/ijaidm.v7i2.29456.

[10] M. Z. Infusi, G. P. Kusuma, and D. A. Arham, “Prediction of local government revenue using data mining method,” *International Journal of Emerging Technology and Advanced Engineering*, vol. 12, no. 1, pp. 63–74, Jan. 2022, doi: 10.46338/ijetae0122_07.

[11] I. Ilic, B. Görgülü, M. Cevik, and M. G. Baydoğan, “Explainable boosted linear regression for time series forecasting,” *Pattern Recognition*, vol. 120, p. 108144, Dec. 2021, doi: 10.1016/j.patcog.2021.108144.

[12] P. V. Thayyib, M. N. Thorakkattle, F. Usmani, A. T. Yahya, and N. H. S. Farhan, “Research on the models for forecast of tax revenue of Wenzhou City,” *Cogent Economics and Finance*, vol. 11, no. 2, Oct. 2023, doi: 10.1080/23322039.2023.2285649.

[13] M. I. Fathoni and A. Saputra, “Forecasting value-added tax (VAT) revenue using autoregressive integrated moving average (ARIMA) Box-Jenkins method,” *Scientax*, vol. 4, no. 2, pp. 205–218, Apr. 2023, doi: 10.52869/st.v4i2.568.

[14] W. Ticona, K. Figueiredo, and M. Vellasco, “Hybrid model based on genetic algorithms and neural networks to forecast tax collection: Application using endogenous and exogenous variables,” in *Proceedings of the 2017 IEEE 24th International Congress on Electronics, Electrical Engineering and Computing, INTERCON 2017*, Aug. 2017, pp. 1–4. doi: 10.1109/INTERCON.2017.8079660.

[15] S. Smyl, “A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting,” *International Journal of Forecasting*, vol. 36, no. 1, pp. 75–85, Jan. 2020, doi: 10.1016/j.ijforecast.2019.03.017.

[16] Z. Ferdoush, B. N. Mahmud, A. Chakrabarty, and J. Uddin, “A short-term hybrid forecasting model for time series electrical-load data using random forest and bidirectional long short-term memory,” *International Journal of Electrical and Computer Engineering*, vol. 11, no. 1, pp. 763–771, Feb. 2021, doi: 10.11591/ijece.v11i1.pp763-771.

[17] J. Hossain and M. T. Ismail, “Performance of a novel hybrid model through simulation and historical financial data,” *Sains Malaysiana*, vol. 51, no. 7, pp. 2249–2264, Jul. 2022, doi: 10.17576/jsm-2022-5107-25.

[18] H. Laaroussi, F. Guerouate, and M. Sbihi, “A novel hybrid deep learning approach for tourism demand forecasting,” *International Journal of Electrical and Computer Engineering*, vol. 13, no. 2, pp. 1989–1996, Apr. 2023, doi: 10.11591/ijece.v13i2.pp1989-1996.

[19] M. Achite, P. Tsangaratos, G. Pellicone, B. Mohammadi, and T. Caloiero, “Application of multiple spatial interpolation approaches to annual rainfall data in the Wadi Cheliff basin (north Algeria),” *Ain Shams Engineering Journal*, vol. 15, no. 3, p. 102578, Mar. 2024, doi: 10.1016/j.asej.2023.102578.

[20] N. K. G. P, P. M. Sahoo, P. Das, T. Ahmad, and A. Biswas, “Random forest spatial interpolation techniques for crop yield estimation at district level,” *Journal of the Indian Society of Agricultural Statistics*, vol. 78, no. 1, pp. 9–19, May 2024, doi: 10.56093/jisas.v78i1.2.

[21] S. Song *et al.*, “Random forest regression on multi-platform in-situ ocean observations: Investigating high-frequency nutrient dynamics in the Southern Ocean,” May 2024, doi: 10.22541/essoar.171707849.91867565/v1.

[22] D. D. W. Praveenraj, M. Pandey, and M. Victor, “Time series forecasting of stock market volatility using LSTM networks,” in *2023 4th International Conference on Computation, Automation and Knowledge Management, ICCAKM 2023*, Dec. 2023, pp. 1–8. doi: 10.1109/ICCAKM58659.2023.10449596.

[23] “Indonesian Central Bureau of Statistics (BPS), ‘Tax Revenue in Lampung Province.’ 2023. Accessed: Jun. 12, 2024. [Online]. Available: https://lampung.bps.go.id/id/publication/2024/11/22/795b23b04b635b003c9f9fab/financial-statistics-in-lampung-province-2023.html”.

[24] K. Kingphai and Y. Moshfeghi, “On time series cross-validation for deep learning classification model of mental workload levels based on EEG signals,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13811 LNCS, Springer Nature Switzerland, 2023, pp. 402–416. doi: 10.1007/978-3-031-25891-6_30.

[25] K. Hirano and J. H. Wright, “Analyzing cross-validation for forecasting with structural instability,” *Journal of Econometrics*, vol. 226, no. 1, pp. 139–154, Jan. 2022, doi: 10.1016/j.jeconom.2020.10.009.

[26] “Introduction to the Keras Tuner,” *Keras Tuner Library*. https://www.tensorflow.org/tutorials/keras/keras_tuner (accessed Jul. 01, 2024).

[27] M. Shanmugavalli and K. M. J. Ignatia, “Comparative study among MAPE, RMSE and R square over the treatment techniques undergone for PCOS influenced women,” *Recent Patents on Engineering*, vol. 18, no. 1, Jan. 2023, doi: 10.2174/0118722121269786231120122435.

[28] N. K. Rai, D. Saravanan, L. Kumar, P. Shukla, and R. N. Shaw, “RMSE and MAPE analysis for short-term solar irradiance, solar energy, and load forecasting using a Recurrent Artificial Neural Network,” in *Applications of AI and IOT in Renewable Energy*, Elsevier, 2022, pp. 181–192. doi: 10.1016/B978-0-323-91699-8.00010-3.

[29] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[30] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

BIOGRAPHIES OF AUTHORS

Rahman Taufik received his bachelor's degree in computer science education from the University of Education, in 2015, and a master's degree in computer science from Telkom University in 2019 in Bandung, Indonesia. He is currently a lecturer at the Department of Computer Sciences, University of Lampung, Indonesia. His research interests include data analytics, artificial intelligence, and intelligent tutoring system. In addition, he is an associate editor of the journal Informatik: Jurnal Ilmu Komputer at UPN Veteran Jakarta, Indonesia. He can be contacted at email: rahman.taufik@fmipa.unila.ac.id.

Aristoteles received his bachelor's degree in computer science from Universitas Padjadjaran, Indonesia, in 2004. He then pursued his master's degree in computer science at Institut Pertanian Bogor, Indonesia, and graduated in 2011. In 2024, he completed his doctorate at the Faculty of Mathematics and Natural Sciences (FMIPA) at Universitas Lampung, Indonesia. He is currently the vice dean of the Faculty of Mathematics and Natural Sciences, Universitas Lampung, Indonesia. His research has been funded by the Universitas Lampung and the Ministry of Education and Culture of the Republic of Indonesia. He has authored or co-authored more than 200 refereed journal and conference papers, and 4 book chapters. He can be contacted at email: aristoteles.1981@fmipa.unila.ac.id.

Igit Sabda Ilman received his bachelor's degree in information systems from Amikom, Indonesia, in 2016, and his master's degree in computer science from Universitas Gadjah Mada, Indonesia, in 2019. His research interests include information systems, data mining, and databases. He can be contacted at email: igit.sabda@fmipa.unila.ac.id.