

COMPARISON OF DOUBLE RANDOM FOREST AND LONG SHORT-TERM MEMORY METHODS FOR ANALYZING ECONOMIC INDICATOR DATA

Andika Putri Ratnasari¹, Budi Susetyo^{2*}, Khairil Anwar Notodiputro³

^{1,2,3}Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University
Raya Dramaga Street., Bogor, West Java, 16680, Indonesia

Corresponding author's e-mail: ^{2*} budisu@apps.ipb.ac.id

ABSTRACT

Article History:

Received: 5th December 2022

Revised: 10th April 2023

Accepted: 13th April 2023

Keywords:

DRF;

LSTMs;

Non-underfitting;

Nonlinear;

Under-fitting.



This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License.

How to cite this article:

A. P. Ratnasari, B. Susetyo and K. A. Notodiputro., "COMPARISON OF DOUBLE RANDOM FOREST AND LONG SHORT-TERM MEMORY METHODS FOR ANALYZING ECONOMIC INDICATOR DATA," *BAREKENG: J. Math. & App.*, vol. 17, iss. 2, pp. 0757-0766, June, 2023.

Copyright © 2023 Author(s)

Journal homepage: <https://ojs3.unpatti.ac.id/index.php/barekeng/>

Journal e-mail: barekeng.math@yahoo.com; barekeng_journal@mail.unpatti.ac.id

Research Article • Open Access

1. INTRODUCTION

Time series data consists of sequences of observations collected over time [1]. Methods to analyze time series data have been widely developed, including several algorithms in machine learning [2]. Machine learning is an alternative method in analyzing time series data and it has been reported as powerful when the data is nonlinear [3]. Machine learning is categorized into two different types, supervised and unsupervised learning. Supervised learning creates a statistical model to predict or estimate an outcome based on one or more inputs, while unsupervised learning finds useful patterns in a dataset without output variables [4].

Some research compared machine learning and ARIMA as well as regression models in modeling and forecasting new Covid-19 cases in Nigeria [5]. Mualifah, et al. compared the performances of GARCH, LSTM, and hybrid GARCH–LSTM models for analyzing the dynamical pattern of stock price volatility of PT Bumi Resources Minerals Tbk in 2022 [6]. These researches showed that machine learning based methods outperform other methods. Another research was also conducted using tree-based methods applied to time series data. The research compared ARIMA, Decision Tree (DT), Random Forest (RF), and Gradient Boosted Trees (GBT) to predict monthly gold prices. The result of this study found that RF is superior to the other methods [7].

Sunwoo Han, Hyunjoong Kim, and Yung-Seop Lee developed a new supervised learning method called Double Random Forest (DRF) in 2020. This method was developed to overcome the limitation of Random Forest (RF) when the data is under-fitting. While RF has the best performance using minimum node size, it may have a chance to be under-fitting, or the size of the tree is not large enough. The previous study compared DRF and RF using 34 datasets with the best performance on minimum node size or under-fitting in RF. The result of this research found that DRF outperforms RF [8].

The other machine learning methods, more specifically Neural Networks (NNs), have the ability to model nonlinear data with high forecasting accuracy and minimal initial assumptions [2]. Long Short-Term Memory Networks (LSTMs) are one of the NNs methods which has the capability to analyze time series data. The main advantage of LSTMs is the ability to remember information for long time periods [9]. The previous research used LSTMs to forecast COVID-19 outbreaks in Egypt [10]. In addition, a comparative study found that LSTMs outperform ARIMA in forecasting cryptocurrency prices [11].

DRF and LSTMs have abilities to analyze under-fitting and nonlinear data, respectively. Therefore, this study aims to present a comparative study between DRF and LSTMs using under-fitting, non-underfitting, and nonlinear data from Indonesia's economic indicators. The indicators used in this study are Export, Import, Official Reserves Assets, and Exchange Rates. These indicators were used because they have fluctuating patterns that tend to be nonlinear and will be suitable if these data are analyzed using machine learning methods. The best performance method was chosen using the least value of Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Error (MAE).

2. RESEARCH METHODS

This study used Indonesia's economy indicators. The indicators used in this study are Export, Import, Official Reserve Assets in a million USD, and Exchange Rates obtained from Bank Indonesia's and International Monetary Fund's websites. These are monthly time series data from January 2010 to December 2021. DRF and LSTMs methods use output and input variables in the analysis process. The input variable used in this study is the data on previous lag ($t-1$), and the data at time t was used as the output variable in this study. The hyperparameters optimized in DRF are node size, defined as the minimum number of samples a terminal node should hold, and ntree, defined as the number of trees that will be created [8]. The hyperparameters optimized in LSTMs are epoch, defined as the number of times the algorithm will be run for the entire data [12], and learning rate, defined as a parameter that controls the amount of weight change that is updated during the training process in response to the error value generated during the training process. It can affect the convergence time of the gradient descent [4].

The analysis stages in this study are as follows.

1. Examining each indicator used in this study, whether the data is considered under-fit, non-under-fit, or nonlinear.

2. Carrying out the training process on the training data on all split groups of time series cross-validation using each hyperparameter combination of DRF and LSTMs.
3. Calculating RMSE, MAPE, and MAE from test data on all split groups of time series cross-validation data for each parameter combination. The model with optimal parameter is selected based on the least average value of RMSE, MAPE, and MAE from the combination of hyperparameters used on all split groups of time series cross-validation data.
4. The best model selected from step 3 is used to create a model using DRF and LSTMs with a proportion of 83% for training data and 17% for testing data.
5. Comparing the performance of DRF and LSTMs.

2.1 Double Random Forest (DRF)

The ensemble method is a method that combines prediction results from some baseline models to obtain enhanced performance [13]. Double Random Forest is a new ensemble method developed by Sunwoo Han, Hyunjoong Kim, and Yung-Seop Lee in 2020. This method can increase the can performance of Random Forest (RF) while RF is under-fitting because it can create a bigger tree which can reduce the bias in prediction. Under-fitting in RF can be identified using relative test MAPE, defined as (the MAPE value of RF under the given node size) / (the MAPE value of RF when node size is set to its default). Therefore, if the relative MAPE is greater than 1, RF with the largest trees is more accurate than that with the smaller trees. After all, this means that RF may under-fit. Here is the algorithm of DRF [8].

1. Each regression tree in DRF is created by using all the training data (D) with size n .
2. Steps of selecting the best splitting:
 - a. For a given node t , a random sampling of size n_t is carried out with bootstrap, if the number of samples at node t is greater than $n \times 0.1$. If it does not meet the requirements, then bootstrap will not be done.
 - b. Randomly select $p \approx \frac{m}{3}$ or \sqrt{p} features.
 - c. Find the best split features and cut points using the random feature subset.
 - d. Send down the data using the best split features and cut points.
 - e. Steps a to d are repeated until the stopping rules are met to obtain the estimation result from a regression tree.
3. Steps 1-2 are repeated until DRF creates b regression trees.
4. The prediction result from DRF is obtained using each tree's average prediction.

2.2 Long Short-Term Memory Networks (LSTMs)

Long Short-Term Memory Networks were introduced by Sepp Hochreiter and Jurgen Schmidhuber in 1997 [14]. LSTMs are special kinds of Recurrent Neural Networks (RNNs) and have capability to learn long-term temporal or sequential dependencies [15]. LSTMs contain an input gate, an output gate, a forget gate, and internal state (cell memory) [16]. Figure 1 shows the architecture of LSTMs [14].

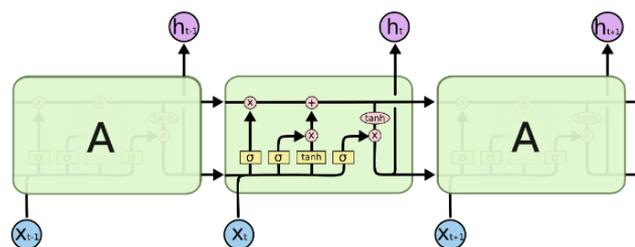


Figure 1. The architecture of LSTMs

According to the architecture in Figure 1, LSTMs have several steps in the learning process. The learning stages of LSTMs are as follows [16].

1. Determining the LSTMs output using Equations (1-5) (forward learning). The equations can be written as follows.

$$a(t_i) = \sigma(w_\alpha x(t_i) + w_{h\alpha} h(t_{i-1}) + b_\alpha) \quad (1)$$

$$f(t_i) = \sigma(w_f x(t_i) + w_{hf} h(t_{i-1}) + b_f) \quad (2)$$

$$c(t_i) = f_t \times c(t_{i-1}) + a_t \times \tanh(w_c x(t_i) + w_{hc} h(t_{i-1}) + b_c) \quad (3)$$

$$o(t_i) = \sigma(w_o x(t_i) + w_{ho} h(t_{i-1}) + b_o) \quad (4)$$

$$h(t_i) = o(t_i) \times \tanh(c(t_i)) \quad (5)$$

with,

$x(t_i)$: the input values

$a(t_i)$: the output results for input gate

σ and \tanh : the activation functions

$f(t_i)$: the output results for forget gate

$c(t_{i-1})$ and $c(t_i)$: the cell states at time $t-1$ and t

$o(t_i)$: the output results for output gate

$h(t_{i-1})$ and $h(t_i)$: the output value at time $t-1$ and t

w_α, w_o, w_f, w_c : the weight matrixes of input gate, output gate, forget gate, and internal state

$w_{ha}, w_{hf}, w_{hc}, w_{ho}$: the recurrent weights

2. Computing the error between the resulted data and input data of each layer.
3. The error is reversely propagated to the input gate, cell, and forget gate.
4. The optimization algorithm based on the error term is used to update the weight of each gate.
5. Stages 1– 4 are reiterated for a set number of iterations until the biases and weights optimal value can be obtained.

2.3 Measure of Forecast Accuracy

The model performance of DRF and LSTMs is assessed by three measures of forecast accuracy, those are RMSE, MAPE, and MAE. The model with the least RMSE, MAPE, and MAE is considered the best performance. RMSE is the standard deviation of the model prediction results [17]. The formula of RMSE can be written as follows [18].

$$e_t = y_t - \hat{y}_t \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (e_t)^2} \quad (7)$$

MAPE represents the mean absolute percentage error function for the prediction and the eventual outcomes, this error measures express error as a percentage [19]. The formula of MAPE can be written as follows [18].

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{e_t}{y_t} \right| \times 100 \quad (8)$$

with e_t is the error value at time t , y_t is the actual value at time t , and \hat{y}_t is the forecasted value at time t . The MAPE values interpretations are shown in Table 1 [20].

Table 1. MAPE Values Interpretation

MAPE Value	Interpretation
>50%	Inaccurate forecasting
20% - 50%	Reasonable forecasting
10% - 20%	Good forecasting
<10%	High accurate forecasting

MAE is obtained from the average of absolute difference between the forecasted value and the actual value [17]. The formula of MAE can be written as follows [18].

$$MAE = \frac{1}{n} \sum_{t=1}^n |e_t| \quad (9)$$

3. RESULTS AND DISCUSSION

3.1. Under-fitting/Non-underfitting Data

Under-fitting data can be detected using Random Forest. If RF gives the best performance using the default node size = 5 or the largest tree is more accurate than the smaller one, it means that RF may under-fit. **Table 2** shows the MAPE value of each node size in Export, Import, Official Reserve Assets, and Exchange Rates data.

Table 2. MAPE Value of Each Node size

Node size	Export	Import	Official Reserve Assets	Exchange Rates
5	11.93790	11.54937	4.324446	1.414259
6	11.84959	11.53789	4.331469	1.301333
7	11.87713	11.50116	4.387146	1.245586
8	11.85902	11.45797	4.427928	1.211501
9	11.83365	11.50479	4.533176	1.180476
10	11.88511	11.51324	4.577068	1.172965
11	11.89959	11.51269	4.643693	1.146187
12	11.97062	11.51991	4.681113	1.120810

Furthermore, the relative MAPE for each data is also calculated and shown in **Figure 2**. If the relative MAPE is greater than 1, it means that RF may under-fit because the largest tree gives a smaller MAPE value. The result of relative MAPE shown in **Figure 2** shows that RF under-fit in Official Reserve Assets data, and RF does not under-fit in Export, Import, and Exchange Rates data.

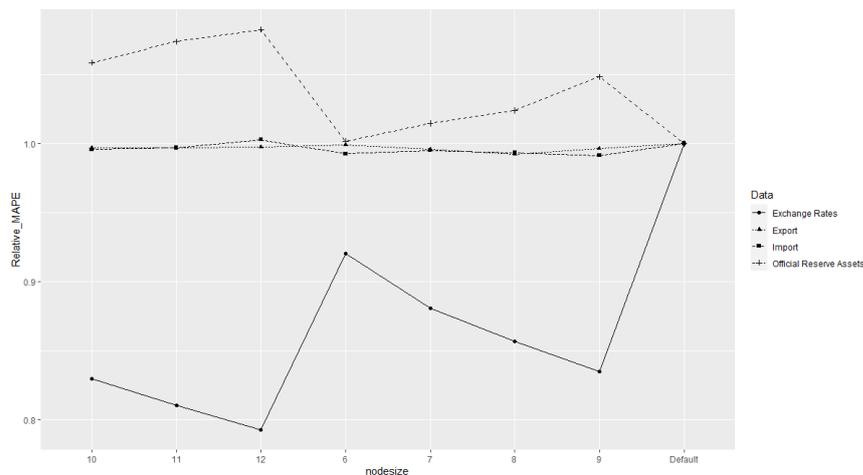


Figure 2. Relative MAPE

3.2. Linear/Nonlinear Data

Terasvirta test was used in this study to detect nonlinearity in the data. The significance level used in this test is 0.05, with hypothesis null states that data has a linear relationship with its lag. **Table 3** shows the result of the Terasvirta test for Export, Import, Official Reserves Assets, and Exchange Rates data.

Table 3. The result of the Terasvirta test

Data	p-value	Conclusion
Export	0.0000	Nonlinear
Import	0.4752	Linear
Official Reserve Assets	0.5218	Linear
Exchange Rates	0.0000	Nonlinear

3.3. The Results of DRF and LSTMs

Before comparing the performance of DRF and LSTMs, the optimal hyperparameter for each model was selected using time series cross-validation. **Table 4** shows the list of hyperparameters which were used in this study.

Table 4. The Hyperparameters for Each Model

Model	Hyperparameters	
DRF	Node size	5, 6, 7, 8, 9, and 10
	ntree	500 and 700
LSTMs	Epoch	50 and 100
	Learning Rates	0.001, 0.0025, and 0.005

Time series cross-validation (TSCV) splits the time series data in sequence into K complementary partitions. The model is validated on a test set in one round of TSCV and then trained on other $K-1$ partitions (referred to as training sets). The validation processes are repeated for $K-1$ rounds using different $K-1$ time series partitions to decrease the model variability. Then the performances of the model across different $K-1$ validation partitions are averaged [21]. **Table 5** shows the data partition for time series cross-validation with $K = 6$.

Table 5. Data Partition

K	Training Data Period	Testing Data Period
1	January 2010 – July 2020	August 2020 – December 2021
2	January 2010 – August 2020	September 2020 – December 2021
3	January 2010 – September 2020	October 2020 – December 2021
4	January 2010 – October 2020	November 2020 – December 2021
5	January 2010 – November 2020	December 2020 – December 2021
6	January 2010 – December 2020	January 2021 – December 2021

This study used RMSE, MAPE, and MAE resulted from time series cross-validation to evaluate the optimal hyperparameters. **Table 6** shows the optimal hyperparameters of DRF and LSTMs for Export, Import, Official Reserves Assets, and Exchange Rates data.

Table 6. The Optimal Hyperparameters for Each Data

Data	Model	Hyperparameters	
Export	DRF	Node size	6
		ntree	700
	LSTMs	Epoch	50
		Learning Rates	0.0025
Import	DRF	Node size	10
		ntree	700
	LSTMs	Epoch	100
		Learning Rates	0.001
Official Reserve Assets	DRF	Node size	5
		ntree	500
	LSTMs	Epoch	50
		Learning Rates	0.005
Exchange Rates	DRF	Node size	10
		ntree	500
	LSTMs	Epoch	100
		Learning Rates	0.0025

After the optimal hyperparameters of DRF and LSTMs were selected, this study evaluated the performance of DRF and LSTMs for each data by splitting the data into training and testing data with proportions 83% and 17% respectively. According to the result in 3.1 and 3.2 Export data is considered non-underfitting and nonlinear data, import data is considered as non-underfitting and linear data, Official Reserve Assets data is considered under-fitting and linear data, and Exchange Rates data is considered non-

underfitting and nonlinear data. The performance comparison of DRF and LSTMs for each data is shown in **Table 7**, **Table 8**, and **Table 9**.

Table 7. RMSE Value of DRF and LSTMs

Method	Data	RMSE Export	RMSE Import	RMSE Official Reserve Assets	RMSE Exchange Rates
DRF	Train	765.182	1058.891	1759.239	178.418
	Test	2423.838	2206.910	9374.651	532.652
LSTMs	Train	15.751	317.101	543.212	62.174
	Test	15.751	317.101	554.851	68.228

Table 8. MAPE Value of DRF and LSTMs

Method	Data	MAPE Export	MAPE Import	MAPE Official Reserve Assets	MAPE Exchange Rates
DRF	Train	4.473	6.005	1.354	1.057
	Test	11.147	11.843	5.868	2.076
LSTMs	Train	0.111	2.337	0.430	0.473
	Test	0.100	2.359	0.329	0.423

Table 9. MAE Value of DRF and LSTMs

Method	Data	MAE Export	MAE Import	MAE Official Reserve Assets	MAE Exchange Rates
DRF	Train	621.157	804.934	1455.395	128.155
	Test	1921.524	1679.684	8150.797	310.769
LSTMs	Train	15.751	317.101	460.703	54.776
	Test	15.751	317.101	449.796	61.119

The visualization of actual data and forecasted data for training and testing data are shown in **Figures 3-6**. The black color in the figure represents the actual data, the green color in the figure represents the forecasted value of training data, and the red color in the figure represents the forecasted value of testing data.

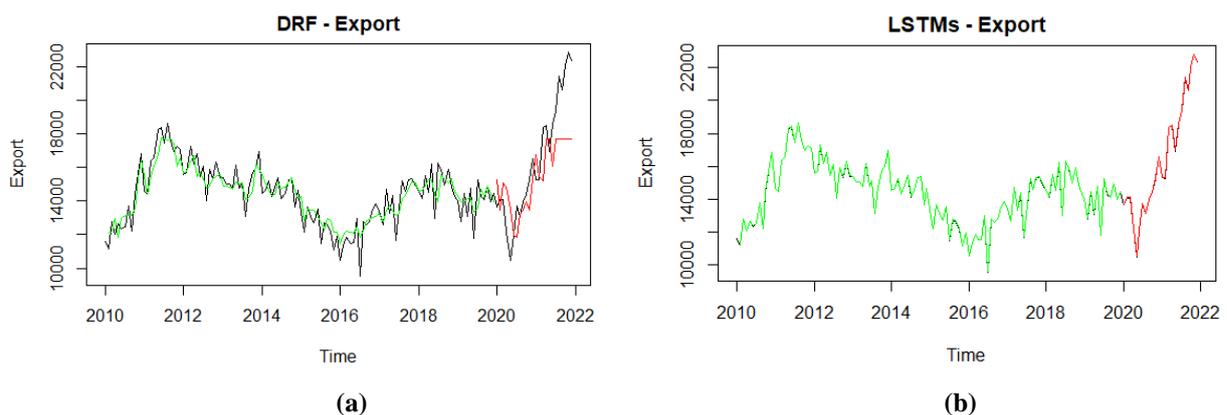


Figure 3. The comparison of actual and forecasted export data (a) The comparison of actual and forecasted export data resulted from DRF, (b) The comparison of actual and forecasted export data resulted from LSTMs

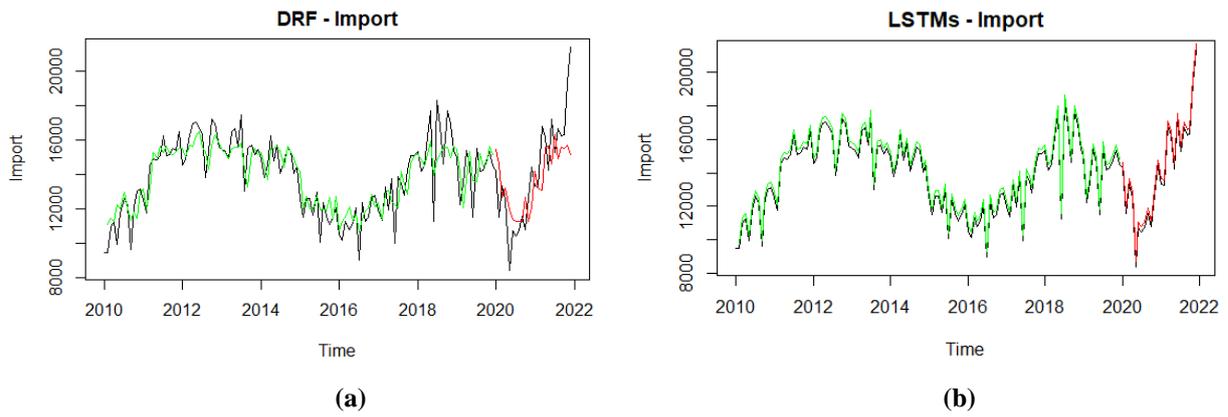


Figure 4. The comparison of actual and forecasted Import data (a) The comparison of actual and forecasted Import data resulted from DRF, (b) The comparison of actual and forecasted Import data resulted from LSTMs

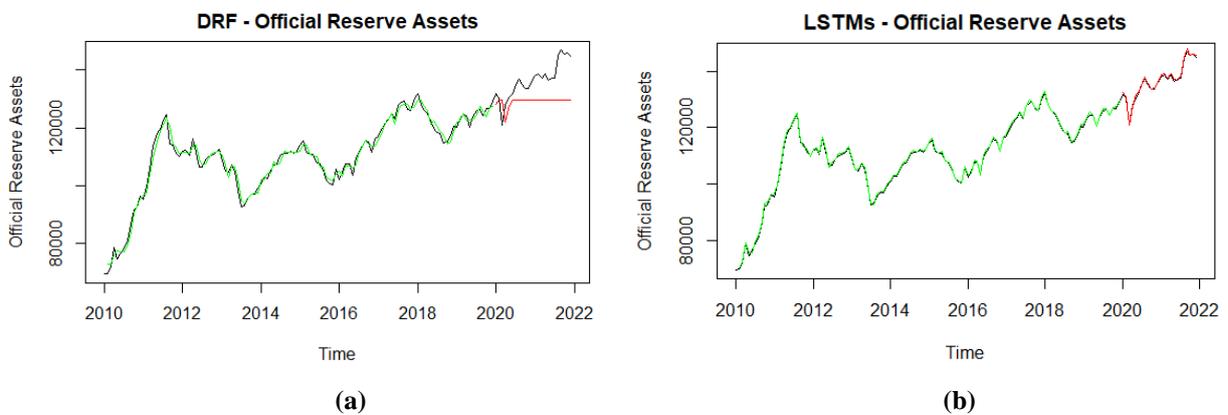


Figure 5. The comparison of actual and forecasted official reserves assets data (a) The comparison of actual and forecasted official reserves assets data resulted from DRF, (b) The comparison of actual and forecasted official reserves assets data resulted from LSTMs

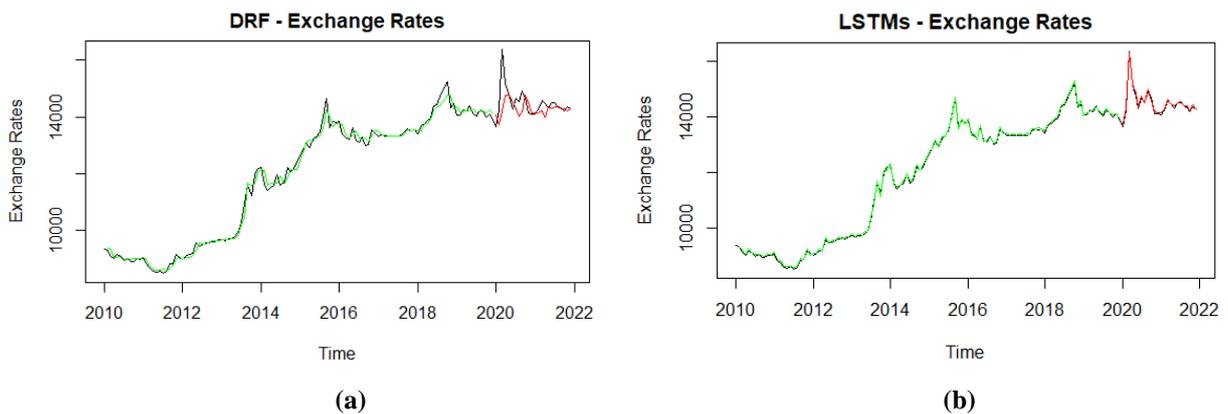


Figure 6. The comparison of actual and forecasted Exchange Rates data (a) The comparison of actual and forecasted Exchange Rates data resulted from DRF, (b) The comparison of actual and forecasted Exchange Rates data resulted from LSTMs

According to the result of the measure of forecast accuracy and the visualization of actual and forecasted data in **Figures 3-6**, LSTMs still outperform DRF whether the data is under-fitting, non-underfitting, or nonlinear.

4. CONCLUSIONS

This study used Indonesia's economy indicators considered to under-fitting, non-underfitting, and nonlinear data. These indicators are Export, Import, Official Reserve Assets, and Exchange Rates. The performance comparison results using RMSE, MAPE, and MAE showed that LSTMs have higher performance than DRF in each data.

ACKNOWLEDGMENT

This research was supported by the Education Fund Management Institutions of the Republic of Indonesia (LPDP), grant numbers KEP-732/LPDP/LPDP.3/2021.

REFERENCES

- [1] M. Larrea, A. Porto, E. Irigoyen, A. J. Barragán, and J. M. Andújar, "Extreme learning machine ensemble model for time series forecasting boosted by PSO: Application to an electric consumption problem," *Neurocomputing*, vol. 452, pp. 465–472, 2021.
- [2] V. Plakandaras, T. Papadimitriou, and P. Gogas, "Forecasting the USD / EUR daily and monthly rate with machine learning techniques," *J. Forecast.*, vol. 34, pp. 560–573, 2015.
- [3] C. Katris, "A time series-based statistical approach for outbreak spread forecasting: Application of COVID-19 in Greece," *Expert Systems with Applications*, vol. 166, 2021.
- [4] M. Khalil, A. S. McGough, Z. Pourmirza, M. Pazhoohesh, and S. Walker, "Machine Learning, Deep Learning and Statistical Analysis for forecasting building energy consumption — A systematic review," *Eng. Appl. Artif. Intell.*, vol. 115, 2022.
- [5] S. I. Busari and T. K. Samson, "Modelling and forecasting new cases of Covid-19 in Nigeria: Comparison of regression, ARIMA and machine learning models," *Scientific African*, vol. 18, 2022.
- [6] L. N. A. Mualifah, A. M. Soleh, and K. A. Notodiputro, "A comparative study of GARCH, LSTM, and HYBRID GARCH–LSTM models for analyzing the dynamical pattern of stock price volatility," Master Thesis, IPB University, Bogor, West Java, Indonesia, 2022.
- [7] E. H. A. Rady, H. Fawzy, and A. M. A. Fattah, "Time series forecasting using tree based methods," *J. Stat. Appl. Probab.*, vol. 10, no. 1, pp. 229–244, 2021.
- [8] S. Han, H. Kim, and Y. S. Lee, "Double random forest," *Mach. Learn.*, vol. 109, no. 8, pp. 1569–1586, 2020.
- [9] A. H. Elsheikh, V. P. Katekar, O. L. Muskens, S. S. Deshmukh, M. A. Elaziz, and S. M. Dabour, "Utilization of LSTM neural network for water production forecasting of a stepped solar still with a corrugated absorber plate," *Process Safety and Environmental Protection*, vol. 148, pp. 273–282, 2021.
- [10] M. Marzouk, N. Elshaboury, A. A. Latif, and S. Azab, "Deep learning model for forecasting COVID-19 outbreak in Egypt," *Process Safety and Environmental Protection*, vol. 153, pp. 363–375, 2021.
- [11] I. Nasirtafreshi, "Forecasting cryptocurrency prices using Recurrent Neural Network and Long Short-term Memory," *Data Knowl. Eng.*, vol. 139, 2022.
- [12] R. Supakar, P. Satvaya, and P. Chakrabarti, "A deep learning based model using RNN-LSTM for the Detection of Schizophrenia from EEG data," *Computers in Biology and Medicine*, vol. 151, 2022.
- [13] M.A Ganaie, Hu. Minghui, A.K. Malik, M. Tanveer, and P.N. Suganthan, "Ensemble deep learning: A review," *Eng. Appl. Artif. Intell.*, vol.115, 2022.
- [14] C. Olah, "Understanding LSTM Networks," 27 August 2015, [Online]. Address: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/> (Accessed 16 November 2022).
- [15] S. D. Khan, L. Alarabi, and S. Basalamah, "Toward smart lockdown: A novel approach for covid-19 hotspots prediction using a deep hybrid neural network," *Computers*, vol. 9, no. 4, pp. 1–16, 2020.
- [16] H. Abbasimehr, M. Shabani, and M. Yousefi, "An optimized model using LSTM network for demand forecasting," *Comput. Ind. Eng.*, vol. 143, 2020.
- [17] Z. Yan, J. Wang, L. Sheng, and Z. Yang, "An effective compression algorithm for real-time transmission data using predictive coding with mixed models of LSTM and XGBoost," *Neurocomputing*, vol. 462, pp. 247–259, 2021.
- [18] W. Zha, Y. Liu, Y. Wan, R. Luo, D. Li, S. Yang, and Y. Xu, "Forecasting monthly gas field production based on the CNN-LSTM model," *Energy*, vol. 260, 2022.
- [19] Y. Ensafí, S. H. Amin, G. Zhang, and B. Shah, "Time-series forecasting of seasonal items sales using machine learning – A comparative analysis," *International Journal of Information Management Data Insights*, vol. 2, no. 1, 2022.
- [20] J. J. Montañó Moreno, A. Palmer Pol, A. Sesé Abad, and B. Cajal Blasco, "Using the R-MAPE index as a resistant measure of forecast accuracy," *Psicothema*, vol. 25, no. 4, pp. 500–506, 2013.
- [21] S. X. Lv, L. Peng, H. Hu, and L. Wang, "Effective machine learning model combination based on selective ensemble strategy for time series forecasting," *Inf. Sci.*, vol. 612, pp. 994–1023, 2022.

