A comparison of SARIMA and LSTM in forecasting dengue hemorrhagic fever incidence in Jambi, Indonesia

U Khaira¹, P E P Utomo¹, R Aryani¹ and I Weni¹
¹Department of Information System, Universitas Jambi, Jl. Raya Jambi - Muara Bulian KM.15 Mendalo Indah Muaro Jambi, Jambi, Indonesia

*Email: ulfa.ilkom@gmail.com

Abstract. Dengue Hemorrhagic Fever (DHF) is one of the common and fatal diseases in Indonesia. Jambi city is one of the dengue-endemic areas in Jambi province. To reduce the incidence rate of dengue, an early warning based on forecasting is necessary. Time-series forecasting of DHF can provide useful information to support and help public health officers for planning on DHF prevention. This paper compares two methods for Time-series forecasting of DHF incidence, namely seasonal autoregressive integrated moving average (SARIMA) and Long Short-Term Memory (LSTM). The forecasting performance is assessed using the monthly number of DHF incidence data from January 2012 to April 2019 were acquired from the district health offices. To show the effectiveness of the model, the performances are evaluated based on two metrics: mean absolute error (MAE) and root mean square error (RMSE). In the first analysis, we found that the SARIMA ((1,0,0) (1,0,0)12) is the most suitable model to predict the number of monthly DHF incidence with RMSE value of 30.07 and MAE 18.97, and the second one used the LSTM with one hidden layer (1-64-1) architecture with RMSE of 30.41 and MAE 18.27. Based on the experiment between SARIMA and LSTM perform relatively well to predict the future.

1. Introduction
Dengue Hemorrhagic Fever (DHF) is one of the severe and common health problems in Indonesia. Since 1968 the number of cases and the spreads of the disease are increasing. Growing population, rapid urbanization, and modern transportation led to the increasing transmission of dengue. Indonesia is a tropical country with high population density especially in an urban area, potentially as a habitat for dengue viruses. Dengue viruses are transmitted through the bite of the Aedes aegypti and Ae. Albopictus causes a high fever, red spots on the skin, and pain in muscles. There is no vaccine available for preventing DHF, the available treatments for DHF patients are only supportive symptomatic treatment including antipyretics, antiemetics, and intravenous fluids [1].

Jambi Province is one of the dengue-endemic areas, based on data from the Jambi Health Office. During January-February 2019 the number of DHF cases in Jambi Province reached 417 cases with 5 deaths, with a fatality rate of 0.14 per 100,000 population. The case is quite high because not even been two months, the case is already almost 50 percent from the previous year which is 812 cases with a fatality rate (IR) of 0.14 per 100,000 population [2]. DHF is a disease that has a rapid journey from the beginning of being affected until it is diagnosed with Dengue Fever, late treatment often causes death.

Based on reports from the Provincial Health Office, a survey (Epidemiological Investigation) needs to be carried out to find out more about the situation of dengue outbreaks in Jambi Province so
that appropriate countermeasures can be proposed for reducing the number of Dengue cases and preventing or minimizing dengue outbreaks Jambi Province. To lower the incidence rate of this disease, an early warning based on forecasting is necessary. The usage of time series forecasting has been frequently used in many areas especially in the field of epidemiological research of infectious disease [3,4].

Autoregressive Integrated Moving Average (ARIMA) is one of the common traditional prediction methods in health science research. DHF incidence is influenced by seasonal weather. However, the ARIMA method is not suitable for time series data containing seasonality. Therefore, the ARIMA method is modified for seasonal data, which is called SARIMA. The SARIMA model is a combination of an autoregressive model and a moving average model both seasonal and non-seasonal [3,5]. SARIMA is a method for finding patterns from a collection of seasonal time series data for forecasting. This method has better accuracy for predicting seasonal time series data. The SARIMA model has been successfully applied in many areas during the past three decades, but it has limitations. The SARIMA model can only be applied to linear time series data models, can not deal with nonlinear patterns. DHF data contain both linear and nonlinear patterns [6]. A new approach of deep learning algorithms for prediction problems has been introduced, it can deal with non-linearity and complexity in time series forecasting. One deep learning approach is Long Short-Term Memory (LSTM), it allows much longer temporal sequences to be processed. The empirical studies show that LSTM outperforms traditional-based algorithms such as the ARIMA model [7].

In this paper, we developed the SARIMA and LSTM model to forecast the monthly DHF incidence in Jambi city based on reported monthly cases available from January 2012–April 2019 and compare the performance of SARIMA and LSTM algorithms concerning for to minimization achieved in the error rates in prediction.

2. Methods

2.1. Study area

Jambi is the capital city of Jambi province located 1°35′21″S 103°36′36″E, it covers an area 205.38 km². With total population was 591,134 (2017) and the density was 2,877.54/km². Jambi is an endemic area for DHF cases, with the province of Jambi as the area with the highest incidence of dengue.

2.2. Data collection

This study used monthly dengue hemorrhagic cases for the period of January 2012 through April 2019 and was obtained from the Provincial Health Office. The monthly data set contains a total of 88 monthly observations. The plot of DHF case data is presented in Figure 1 below:

![Figure 1](image_url)

Figure 1. The plot of Jambi City DHF case data from 2012 – 2019.
2.3. Seasonal ARIMA (SARIMA)
A group of models known as Autoregressive Integrated Moving Average (ARIMA) models were created by Box and Jenkins. This model can occur because there is a differencing from the original data. In general, the form of the ARIMA model with the order \( (p, d, q) \) with differencing \( d \) as much as shown in the equation (1) [8,9].

\[
\theta_p(B^p)(1 - B)^dY_t = C + \theta_q(B^q)a_t
\]  

(1)

the ARIMA model can be denoted as:

\[
\text{ARIMA}(p,d,q)
\]

(2)

Where \( Y_t \) is a dependent variable on \( t \), \( p \) = order of an autoregressive model, \( q \) = order of moving average model, \( a_t \) = error, \( C \) is a constant number, \( d \) = differencing. While the ARIMA model with seasonal influences, called the Seasonal ARIMA (SARIMA), is expressed in the equation [8].

\[
\phi_p(B^p)\phi_p(B)(1 - B)^d(1 - B^s)Y_t = \theta_q(B)\theta_q(B^s)a_t
\]  

(3)

In general, the ARIMA Seasonal model is notated as follows

\[
\text{SARIMA}(p,d,q)(P, D, Q)_s
\]

(4)

Where \( \phi_p(B) \) = non seasonal Autoregressive, \( \theta_q(B) \) = non seasonal moving average, \( F_s(B^s) \) = Seasonal Autoregressive, \( Q_o(B^s) \) = Seasonal Moving Average. \( p, d, q \) is the non-seasonal part of the model, \( P, D, Q \) the seasonal part of the model and \( s \) sum of periods per season [10].

In this research, The SARIMA method is used to predict time series data from dengue hemorrhagic fever (DHF) cases. DHF Case Variables to predicted based on time series data with monthly periods. Then the DHF case variable is carried out in the modeling process. In forecasting with SARIMA Model, the main idea is to determine the most suitable pattern or model based on time series data. In the first step, start with the stationary checking for variance and mean. Thus, SARIMA forecasting requires full-time series data/historical data (past and current data) to do accurate forecasting processes. An important requirement for SARIMA forecasting which is an important requirement is the stationary data on variance and mean [11].

Stationary data have a variance of data that is not too wide and tends to approach the average value of the data. If the variance and mean are stationary, then the next stage is examining and plotting the Partial Autocorrelation Function (PACF) and Autocorrelation Function (ACF). Several possible models suitable for modeling can be identified from ACF and PACF plots [12]. After these several matching models are established, the next steps have estimated the parameters and diagnostic checks or called "White noise" to get the best SARIMA model. The best SARIMA Model to be used for forecasting the DHF case variable. With this best SARIMA model, it is obtained forecasting results.

2.4. LSTM
LSTM is one of the developments of neural networks that can learn long term dependency. LSTM architecture consists of three gates: forget gates, input gates, and output gates [17].

The forget gate \( F_t \) defines which information to delete from the memory cells (cell state). The activation function used in forget gates is a sigmoid activation function. Where the output is between 0 and 1. If the output is 1 then all data will be stored and if the output is 0 then all data will be discarded. With the following formula:

\[
F_t = \sigma(W_fS_{t-1} + W_fX_t)
\]  

(5)

Where \( W_f \) is a forget gate weight, \( S_{t-1} \) is a previous state or state at time \( t-1 \), \( X_t \) = input at time \( t \), and \( \sigma \) = sigmoid activation function.

The input gate \( I_t \) determines which information is added to the cell state(\( c_t \)). This procedure consists of two operations: Firstly, candidate values \( \tilde{c} \), potentially added to the cell states, are computed. Secondly, the activation values \( I_t \) of the input and gates are calculated. Where \( W_i \) is input gate weight, \( \sigma \) = sigmoid activation function, and \( W_i \) is cell state weight.

\[
I_t = \sigma(W_iS_{t-1} + W_iX_t)
\]  

(6)

\[
\tilde{c} = tanh(W_cS_{t-1} + W_cX_t)
\]  

(7)
The new cell states $c_t$ are calculated based on the results of the previous steps. With the formula as follows:

$$c_t = (F_t \ast c_{t-1} + i_t \ast \hat{C})$$  \hspace{1cm} (8)

The output gate $o_t$ generates output ($h_t$) after the memory cell has passed the forget gate and input gate. At the output gates, two gates will be implemented, the first one using a sigmoid layer which determines which parts of the cell state going to output. A value will be placed in the memory cell using the tanh activation function. Finally, the two gates are multiplied to produce a value that will be issued($h_t$). With formula as follows:

$$o_t = \sigma(W_o S_{t-1} + W_o X_t)$$ \hspace{1cm} (9)

$$h_t = o_t \ast \tanh(c_t)$$ \hspace{1cm} (10)

### 3. Results and discussion

#### 3.1. SARIMA Forecasting

The monthly data set contains a total of 88 monthly observations, from January 2012 to April 2019. Furthermore, based on plot Time Series Data, DHF case data is checked for time series data stationary conditions that the data used in the forecasting process must be stationary in advance both for seasonal and non-seasonal stationary checks. To further ensure the assumption of stationary residual average will be tested stationary test using Unit Root Test [10]. The test was carried out using the Augmented Dickey-Fuller test (ADF) function in R programming at a significance level of 5% [13]. If probability value ≤ 0.05 (critical value). It can be concluded from the results of the Dickey-Fuller test, that yields less than 0.05 probability, that the data is averagely stationary [14]. After getting stationary data on the variance and mean, then identify to get SARIMA models that are possible to be used in forecasting, based on the value of the autocorrelation function (ACF) and partial autocorrelation function (PACF).

![Figure 2. The ACF and PACF plots.](image)

Based on the ACF and PACF plots in figure 3, values of the ACF and PACF plots are shown to identify the SARIMA model that might be formed. After identifying the SARIMA model that might meet the criteria formed, then each of the possible SARIMA models is carried out a process of parameter estimation and diagnostic check is called “white noise” on the residual SARIMA model. Diagnostic check using the Ljung-Box test (Box-test). Based on diagnostic check calculation results show that all models have p-values for all lag>= 0.05 so that it can be concluded that the residual autocorrelation is not significant or there is no correlation between lags then the residual meets the white noise assumption. Because if the model is formed more than 1 best model, it is necessary to take steps to select the model based on the smallest AIC value.

We obtained the SARIMA (1,0,0) (1,0,0)

12 as the best model will be used to back forecast DBD cases. It has an equation $Y_t = 0.8016Y_{t-1} + 0.5021Y_{t-12} - (0.8016 \ast 0.5021Y_{t-13}) + a_t$. This is intended to see the level of error accuracy from the results of forecasting fit compared to the actual value. Figure 3 shows a comparing fit forecast value and actual value.
3.2. LSTM Forecasting
In general, the steps consist of preprocessing data, initialization parameters, LSTM network training, and testing data. To minimize errors, normalization is performed on the dataset, we use a min-max scaling technique. We set the LSTM model has 64 neurons, 1 dense layer with sigmoid activation function, using Adam optimizer with learning rate 0.001, to avoid over-fitting a dropout is set to 0.2, batch size of 1 is used, and epoch 200.

3.3. Evaluation of model
We use 88 records to train and test our LSTM model, and compare the predicted values with the actual values. The results of the SARIMA and LSTM model for predicting the DHF Case are shown in Fig.3. The actual values in blue, the predictions in orange. From the result shown below, It can be seen that both the SARIMA and LSTM model perform relatively well in fitting the datasets.

![DHF Case PredictionWith SARIMA](image1)

![DHF Case PredictionWith LSTM](image2)

**Figure 3.** Comparing fit forecast value and actual value.

To show the effectiveness of the model, the performances are evaluated based on two metrics: mean absolute error (MAE) and root mean square error (RMSE) [16]. The RMSE measures the differences between actual and predicted values and MAE measures how close forecasts and predictions are to the eventual outcomes. The experimental results are shown in Table 1. Both the SARIMA and LSTM achieve the best performance in all cases.

| Model  | RMSE  | MAE   |
|--------|-------|-------|
| SARIMA | 30.07 | 18.97 |
| LSTM   | 30.41 | 18.27 |

**Table 1.** Performance comparison for two models.

Based on the experiment both SARIMA and LSTM perform relatively well to predict the future. We use the model to forecast the number of DHF cases over the next term (Table 2). Figure 3 also represents the forecast values for the period from May to October 2019. Figure 3 shows that the number of DHF case from May to October 2019 tend to decrease. DHF cases increased in rainy seasons, the rainy season occurs in late October to early May.
Table 2. The forecast of the number of DHF cases for the period from May to October 2019.

|       | May | June | July | August | September | October |
|-------|-----|------|------|--------|-----------|---------|
| SARIMA| 40  | 32   | 26   | 21     | 17        | 13      |
| LSTM  | 80  | 69   | 63   | 58     | 56        | 54      |

4. Conclusion
This paper compares two methods for Time-series forecasting of DHF incidence, namely seasonal autoregressive integrated moving average (SARIMA) and Long Short-Term Memory (LSTM). The forecasting performance is assessed using the monthly number of DHF incidence data from January 2012 to April 2019 were obtained from the district health offices. In the first analysis, we found that the SARIMA ((1,0,0) (1,0,0)12) is the most suitable model to predict the number of monthly DHF incidence with RMSE value of 30.07 and MAE 18.97, and the second one used the LSTM with one hidden layer (1-64-1) architecture with RMSE of 30.41 and MAE 18.27. Based on the experiment both SARIMA and LSTM perform relatively well to predict the future.

References
[1] Riaz, M. M., Mumtaz, K., Khan, M. S., Patel, J., Tariq, M., Hilal, H., ... & Shezad, F. (2009). Outbreak of dengue fever in Karachi 2006: a clinical perspective. *Journal of the Pakistan Medical Association*, 59(6), 339.
[2] Kementerian Kesehatan Republik Indonesia [Indonesian Ministry of Health]. Profil kesehatan Indonesia tahun 2018 [Indonesia health profile 2018]. Retrieved October 9, 2019 from http://www.depkes.go.id/resources/download/pusdatin/profil-kesehatan-indonesia/profil-kesehatan-indonesia-2018.pdf.
[3] Siregar, F. A., Makmur, T., & Saprin, S. (2018, January). Forecasting dengue hemorrhagic fever cases using ARIMA model: a case study in Asahan district. In *IOP Conference Series: Materials Science and Engineering* (Vol. 300, No. 1, p. 012032). IOP Publishing.
[4] Sulistiyowati, S., & Winarko, E. (2014). Peramalan KLB Campak Menggunakan Gabungan Metode JST Backpropagationdan CART. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 8(1), 49-58.
[5] Suhartono, S. (2011). Time series forecasting by using seasonal autoregressive integrated moving average: Subset, multiplicative or additive model. *J. Math. Stat.*, 7, 20-27.
[6] Chen, K. Y., & Wang, C. H. (2007). A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan. *Expert Systems with Applications*, 32(1), 254-264.
[7] Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018, December). A Comparison of ARIMA and LSTM in Forecasting Time Series. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1394-1401).
[8] Wei, W.W.S. 2006. Time series Analysis: Univariate and Multivariate Methods Second Edition. Pearson Addison Wesley: USA
[9] Rosadi, D.2014. Analisis Data Runut Waktu dan Aplikasinya dengan R. UGM Press:Yogyakarta
[10] Tadesse K B and Dinka M O 2017 “Application of SARIMA model to forecasting monthly flows in Waterval River, South Africa,” J. Water L. Dev. 35 (1) 23-25
[11] PEP Utomo, SN Azhari, 2017. Prediksi Kerawanan Wilayah Terhadap Tindak Pencurian Sepeda Motor Menggunakan Metode (S)ARIMA Dan CART. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)* 11 (2), 119-130
[12] Chang 2012 “Seasonal Autoregressive Integrated Moving Average Model for Precipitation Time Series,” *J. Math. Stat.* 8 (4) 500–505
[13] Hikichi S E, Salgado E G and Beijo L A 2017 “Forecasting number of ISO 14001 certifications in the Americas using ARIMA models,” *J. Clean. Prod.* 147 242–253
[14] S.W. Astuti and Jamaludin, 2018. Forecasting Surabaya – Jakarta Train Passengers with SARIMA model. IOP Publishing IOP Conf. Series: Materials Science and Engineering 407 (2018) 012105 doi:10.1088/1757-899X/407/1/012105

[15] WU, Chih-Hung, et al. A New Forecasting Framework for Bitcoin Price with LSTM. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2018. p. 168-175.

[16] Olah, C. (2015). Understanding lstm networks, 2015. URL http://colah.github.io/posts/2015-08-Understanding-LSTMs.