Received: May 23, 2021. Revised: June 30, 2021.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Sparse Representation of Social Media, Internet Query, and Surveillance Data to Forecast Dengue Case Number using Hybrid Decomposition-Bidirectional LSTM

Wiwik Anggraeni[1,2] Eko Mulyanto Yuniarno[3] Reza Fuad Rachmadi[3] Pujiadi[4] Mauridhi Hery Purnomo[1,3,5]*

[1] *Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[2] *Department of Information Systems, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[3] *Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[4] *Dengue Fever Eradication, Malang Regency Public Health Office, Malang, Indonesia*
[5] *University Center of Excellence on Artificial Intelligence for Healthcare and Society (UCE AIHeS), Indonesia*
* Corresponding author's Email: hery@ee.its.ac.id

Abstract: Dengue fever is an endemic disease that occurs throughout the year. Forecasting cases of dengue fever based on actual data is needed for monitoring and taking action. Recently, developing countries have faced problems related to the dengue fever surveillance system caused by the data delay factor. On the other hand, availability and access to health-related information on the internet have changed people's behaviors and habits. However, the effect of internet data usage has not been widely studied, especially in areas with different levels of internet penetration. This study examines the impact of dengue fever case reported data, Google Trends, Twitter, and climate data in areas with many cases and varying levels of internet penetration to forecast dengue fever cases. Split time-series cross-validation (STSCV) and blocked time-series cross-validation (BTSCV) are used to obtain various training and testing results. The hybrid Decomposition-Bidirectional Long Short-Term Memory(D-BiLSTM) method is proposed. D-BiLSTM applied to eight different scenarios across multiple level areas. According to the results of the experiments, the D-BiLSTM model with STSCV outperforms the BTSCV. In the high internet penetration area, the average error is 9,517, while in the low internet penetration area, it is 5,188. In areas with high internet penetration, adding the variables Google Trends and Twitter does not significantly reduce the error forecasting. However, in the low penetration area, the inclusion of Google Trends and Twitter significantly decreases errors. In general, the D-BiLSTM model performed well. Then, when compared with other approaches, the D-BiLSTM model as a whole can reduce the average RMSE and the average MAE of the comparison model by 94,120 and 45,132, respectively, in areas of high internet penetration with the best SMAPE model of 0.310. In the low internet penetration area, the average decline in RMSE and MAE was 54,390 and 19,362, with the best SMAPE model performance of 0.183.

Keywords: Dengue fever, Dynamic forecasting, Social media, Internet query, Decomposition, Bidirectional long short-term memory.

1. Introduction

Dengue Fever (DF) is a contagious disease caused by the dengue virus carried by the Aedes Aegypti mosquito [1]. These mosquitoes are found in tropical and subtropical areas, including Indonesia. DF is an endemic disease that occurs throughout the year, especially during the rainy season. This disease infects a large number of people in a short period [2]. In addition, DF is also the fastest spreading viral disease [3]. There are 50-100 million cases reported worldwide in 100 countries each year which cause 24,000 deaths. Approximately 2.5 billion people live in an endemic state of DF. Based on data compiled by WHO from 1968-2009, Indonesia was the country with the highest cases of DF in Southeast Asia [4]. DF has become one of Indonesia's major public

health problems over the past 47 years as it causes rapid and multiple deaths [2]. DF cases in Indonesia increased significantly in 2019 [5], and many of them died. Delays in handling triggered this condition. One of the causes of delays in handling can be related to the data reporting system.

In several developing countries, including Indonesia, the provision of data quickly becomes a problem [6]. A robust, accurate, and reliable disease surveillance system for DF disease is currently unavailable [7, 8]. So far, the government has relied on hospital-based reporting, which is often left behind and sometimes causes delays in data availability [8], therefore requires revision [7]. The delay in reporting can create gaps in disease surveillance [6], leading to delays in disease detection and management [9]. It shows the need for alternative data sources that can describe cases of dengue fever in near real-time.

On the other hand, the current availability and accessibility of health-related information on the internet have altered how people use the internet [10, 11]. Although not everyone who searches for health-related terms is ill, there is a close relationship between the number of people who search for news about a particular illness and the number of people who have the symptoms. However, it can be an indication of the disease's spread [12]. Internet data, including social media has been widely used to monitor diseases such as influenza [13, 14], DF [6-8, 15-18], Zika [19], Chikungunya [9], Malaria [20], Lyme [21], and hand, foot, and mouth disease [22].

Although research has been carried out related to internet data or social media associated with DF, to the best of our knowledge, it has not yet been examined in detail about how different usage affects areas with different internet penetration rates. Most of the previous studies involved internet search data at the national level, whether used in predicting the number of DF cases [23] and DF outbreaks [24]. Previous studies only focused on the association between social media data and the number of reported cases and had not yet reached the predicted level [6, 15]. In addition, previous studies also involved data from related agencies and social media data only [6, 15, 18]. Otherwise, the number of DF cases is influenced by climatic factors such as rainfall [23], temperature [25], and humidity [26]. The use of climatic factors as a component in the forecasting process can also improve performance [27].

Various conditions and the results of the discussion of previous studies regarding the need to see the effect of usage in areas with different internet penetration rates and involving other factors in predicting DF cases are a challenge for the following study. Therefore, this study attempts to solve this challenge. This study aims to see how Google search data and social media influence forecasting DF cases number. This impact is investigated for areas with varying internet penetration rates by factoring in the real-time climate factor in each area. The climate factors involved include temperature, humidity, rainfall, and wind speed. In addition, this study also proposes a combination of dynamic time series and deep learning models to forecast the number of DF cases in several months ahead. This combination is called Decomposition-Bidirectional LSTM (D-BiLSTM). This model is equipped with a cross-validation process so that the model is more robust and has improved forecasting performance. Furthermore, this model is designed dynamic in terms of adding new input data.

Ultimately, the google query and Twitter results are expected to complement the surveillance data in providing information about future DF cases so that the health office can respond quickly in reducing morbidity and mortality. In addition, the optimal model obtained makes it possible to predict the number of DF events in real-time. This model will reduce DF's social costs and economic losses by eliminating the gap in reporting time in traditional surveillance systems. The remainder of this paper is organized as follows. Section 2 describes the previous related studies. In Section 3, the research areas, data sources, and methods are used. Section 4 discusses the results and discussion. Section 5 presented conclusions, as well as directions for future work.

2. Related works

2.1 Involvement of internet query factors, social media, and climate in forecasting the DF cases number

In recent years, digital footprints have become a potential source of data for health-related purposes. The digital tread is usable to explore disease patterns and health dynamics in a population. Wider internet penetration, increased use of mobile phones, and artificial intelligence in the field of digital epidemiology are promising approaches to assist disease surveillance systems caused by delays in data reporting [28]. This approach has the potential to address gaps in conventional surveillance systems, which often experience delays in reporting, under-reporting, and a lack of supporting budgets because the costs involved are very high.

So far, there have been several studies involving internet search data. Most of these studies use Google

Search data which is better known as Google Trends. These studies prove that Google Trends data correlates well with data on cases reported in the Health Agency, be it cases of DF [6, 8, 15, 16], Malaria [20], Zika [19], and hand, foot, and mouth [22], pertussis [29], and influenza [30]. Besides, Google Trends data is also stated to increase the accuracy of disease prediction, be it the prediction of the number of events [13, 17, 19], population health behavior [18], and outbreaks [7]. Google Trends data can also indicate national epidemiological trends in annual and seasonal variations between years [16]. These studies declare the potential use of Google Trends data. It can be obtained more efficiently, faster, and at a lower cost than traditional reporting systems. Although Google Trends have been widely used in previous DF case research, according to our knowledge, still very rare analyzed how Google Trends influenced areas with different internet penetration rates. This study is relevant to what was suggested by [8, 9], which stated that the following research should analyze areas with different internet penetration rates. In addition, previous studies conducted analyzes at the national level. Dengue information at the national level is not ideal for making decisions locally [7]. The spatiotemporal policy is more suitable, especially for areas with high cases [9]. It is because information at the national level is aggregated from a heterogeneous spatial environment.

On the other hand, there are still few studies that report the potential use of Twitter data. This condition poses a challenge because the public interest in using Twitter is getting bigger. Besides, nowadays, Twitter has become one of the popular social media in Indonesia, besides Instagram and YouTube [31]. As of May 2020, Twitter users in Indonesia increased by 24% compared to 2019 [32]. Indonesia has become the fifth largest country for Twitter users after the UK and other significant countries [33]. In addition, it is also rare to combine the influence of climate variables and internet search data or social media. Most of the previous research used search query data or social media alone. Several previous studies stated that climatic factors primarily influenced cases of dengue fever. Human mobility, mosquito control, and temperature have different effects on the [23, 25, 34]. The trend in DF has a strong correlation with temperature and humidity. Besides that, it is also following the discussion presented by [18] that climate should detect DF cases. Likewise, climatic factors such as temperature, rainfall, humidity, and wind speed affect DF in Indonesia [35].

2.2 The approach used in forecasting the DF cases number

In a previous study, Google Trends data and Twitter data on DF were correlated using the Spearman correlation [15, 18] and Pearson correlation [6, 8, 27]. Some of them also involve a lag factor or delay effect [6, 18, 27]. Meanwhile, prior studies generally utilized the Autoregressive Likelihood Ratio [13], ARIMAX [17, 20], ARIMA [16, 20], SARIMA [28, 29], and SARIMA Rule-Based [35] to estimate the number of cases and disease outbreaks. Furthermore, the Time Series Decomposition method [29] and the Autoregressive Model with Google Search [7] are also used. Besides, previous studies that employed social media data for case prediction and DF outbreaks often employed the Autoregressive Likelihood Ratio approach [13], ARIMAX [17], Autoregressive Model with Google Search [7]. Approaches in [7, 13, 17] are based on the classic time series approach.

This classic time series approach has some drawbacks and advantages. ARIMA and SARIMA are the most widely used forecast models. They perform well when the data is linear [36] and short-term forecasting [22]. But that performance declines when used for medium- and long-term forecasting. In addition, it needs stationary tests in mean and variance over time, which takes a lot of time. ARIMA and SARIMA methods are forecasting methods that do not involve the influence between variables. In addition, ARIMA and SARIMA require us to find different parameters or degrees for other datasets, namely p, d, q [22, 29]. A combination of incorrect parameters will get bad results. Thus, the model's performance so a strong dependency on parameter setting (p d q). ARIMA and SARIMA have been developed by involving several other variables known as ARIMAX models and SARIMAX methods. Although it can affect other variables, the model is still difficult to overcome the non-linearity problem [36].

The classic Time series decomposition approach has the advantage of recognizing the trend, seasonal, cyclical, or random components in the data. However, this model cannot involve other variables such as ARIMA. Besides, Likelihood ratio autoregression is a model approach that can include more than one variable. It uses the likelihood ratio approach to the testing problems in threshold autoregression. This condition makes it difficult to choose the criteria bound. Therefore, it takes a strategy to address this threshold problem. The usual technique used is an estimation.

The relationship between DF and climatic factors is complex, so the classical time series model is not easy to get a fit prediction result [36]. The deep learning approach offers more advantages in the health sector than traditional statistical models [37, 38] and is more frequently applied in predicting disease prevalence [39]. LSTM is often used for time series prediction. It has been successfully used to predict influenza trends and hand, foot, and mouth disease epidemics [40, 41]. LSTM was more widely used for data with large values in the previous study, including many disease events. However, LSTM is declared challenging to predict in areas with less incidence of disease [36] and is less accurate [40]. Research with location differentiation based on penetration in this study shows that the existing data is small and even zero, not stationary, and seasonal. Based on these facts, in this study, LSTM will be combined with the decomposition approach. It extracts time-series data that are often nonlinear and nonstationary without leaving the time domain [29]. So far, several decomposition-based hybrid models such as ANN-decomposition and SVM-decomposition have been developed to investigate time series in other cases. The findings show that the hybrid decomposition model will increase the original data's regularity and get more reliable forecasting results than the conventional model. Based on the advantages, disadvantages, and conditions of existing data that have been mentioned earlier, in this study, the Decomposition will be combined with a particular type of LSTM, namely Bidirectional LSTM, to produce forecasts with good performance. Then, to witness how robust the combination model of D-BiLSTM is, special cross-validation will be carried out for time series data in this study. Several types of cross-validation are not all suitable for time series data [41].

3. Materials and method

3.1 Study area

This study uses two areas, each representing an urban area with a high internet penetration rate and an area where most areas can be categorized as rural with a lower internet level. These areas are the city of Surabaya and the Malang Regency.

Surabaya is the capital of the East Java province and Indonesia's second-largest city after Jakarta. Surabaya is also one of Southeast Asia's oldest port cities, with 160 sub-districts and a population density of 8,268 people per square kilometre. Surabaya City is located at 7° 9‘-7°21' South Latitude and 112°36'-112°57'. Surabaya is mostly lowland, with an elevation of 3-6 meters above sea level [42].

Surabaya's Communication and Information Office announced that the city had won the Indonesia Digital Society Award for having the most digital community in Indonesia. Then, according to a survey conducted by the Indonesian Internet Service Providers Association from the second quarter of 2020 to the second quarter of 2021, Surabaya is a provincial capital with a high internet penetration rate (83.0%), exceeding provincial and even national penetration rates of 73.7% [43].

Then, Malang regency is a plateau surrounded by several mountains and lowlands or valley areas at an altitude of 250-500 masl. Malang Regency is the second-largest district in East Java after Banyuwangi Regency and is the district with the largest population in East Java. Malang Regency has coordinates 112^{o} 17’ to 112^{o}57’ East Longitude and 7^{o}44’ to 8^{o}26’ South Latitude. Malang Regency is also the third-largest district in Java Island. The total area is 3,530.65 km2 consisting of 378 villages with a population density of 831.33 / km^2 [44]. The level of internet penetration in the Malang Regency is 18.3%.

3.2 Dataset

The data used in this study consisted of DF surveillance data, climate data including temperature, rainfall, humidity, and wind speed, Google Trends, and Twitter. The data period used is 2009-2019. The surveillance data were obtained from the Health Office in Malang and Surabaya Regencies. In contrast, the climate data were obtained from the Meteorology, Climatology, and Geophysics Agency of Karangploso and Juanda stations. Data from Google and Twitter is obtained by crawling data using several top keywords related to the term dengue fever commonly used by Indonesians, such as demam (in English fever), demam berdarah (in English dengue fever), dengue, dengue virus. dengue fever.

The data patterns and changes over time of each variable involved in the study are displayed in Fig. 1. The RCN variable represents the reported DF case number, TEMP represents temperature, and RH is relative humidity. Moreover, RF is rainfall, WDSP represents windspeed, TW is the number of tweets related to DF, and GT represents the number of Google Trends.

3.3 Method

This study analyses the correlation between data on Google Trends, Twitter, and climate affect the number of reported DF cases. Besides, this study also analyzes their influence on forecasting results.

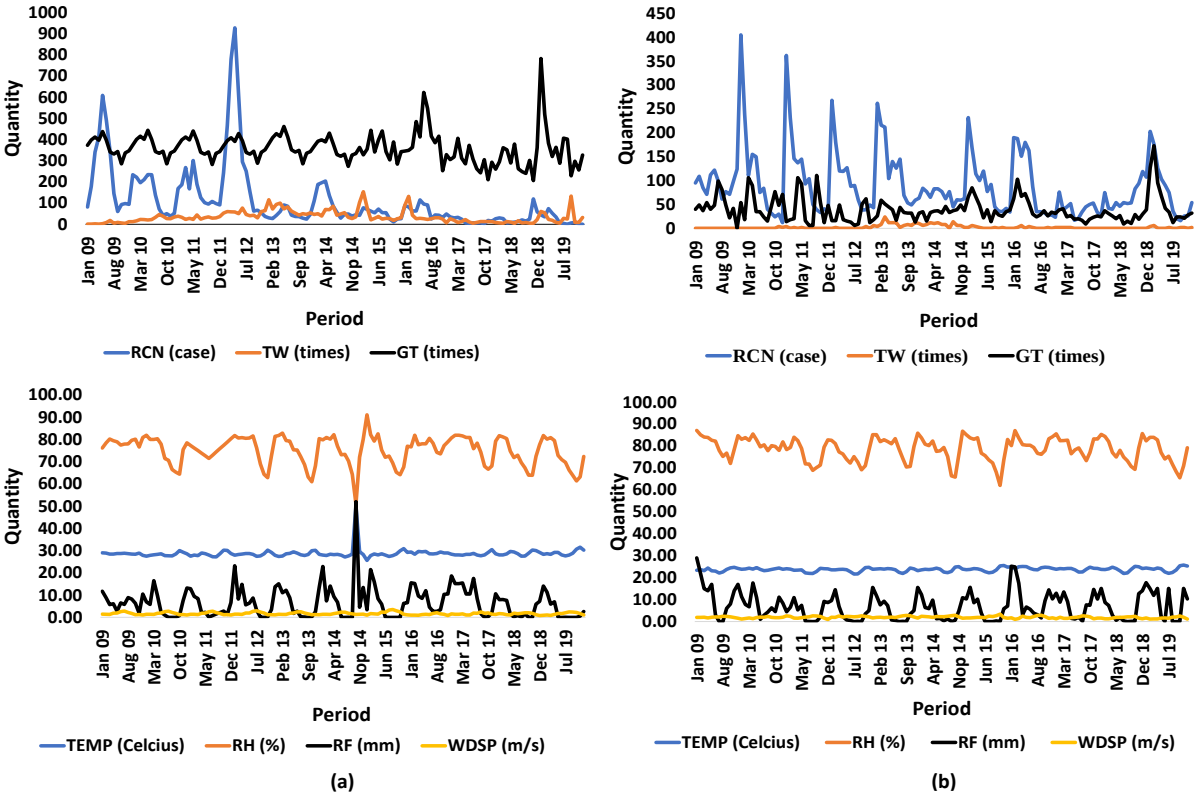

Figure. 1 Data plots of variables involved in each area: (a) Surabaya city, (b) Malang district.

This objective is obtained by building a DF surveillance forecasting model using climate data, including temperature, rainfall, humidity, wind speed, Google Trends, and Twitter. Modeling is carried out using a hybrid model, namely D-BiLSTM. The performance of this model has compared to other candidate models, such as ARIMAX [17], LASSO [34], SVM [37], and Neural Network (NN) [37]. The framework of this study is shown in Fig. 2.

3.3.1. Correlation analysis

How strong the relationship between the variables involved can be found using the correlation coefficient. The specific purpose of the correlation analysis in this study is to see how big the relationship between data on Google Trends, Twitter, and climate on the number of DF cases reported in each area. This study carried out the correlation with the Pearson product-moment[6, 8]. The Pearson product-moment equation is presented in Eq. (1). The $r_{A,B}$ show the correlation coefficient for variables *A* and *B*. Addition, *n* the amount of data, a_t and b_t are the values of attributes *A* and *B* in the data *t*. While $\bar{A}, \bar{B}$ are the average of attribute *A* and attribute *B values*, as well as σ_A, σ_B show the standard deviation of attribute *A* and attribute *B*.

$$r_{A,B} = \frac{\sum_{t=1}^{n}(a_t-\bar{A})(b_t-\bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i-b_i)-n\bar{A}\bar{B}}{n\sigma_A\sigma_B} \quad (1)$$

3.3.2. Decomposing the involved variables

All data next are decomposed using Additive Decomposition. This decomposition produces three values for each variable, namely, trend, seasonality, and random components. In this study, time series decomposition uses the moving average (MA) method to examine trend cycles and seasonal behavior. The trend component (T_t) is obtained using a 2 x m-MA, while the de-trend is obtained by subtracting the DF (Y_t) surveillance data trend. The seasonal component of each season is estimated by averaging the de-trend value in that season. The value of this seasonal component is then adjusted to ensure that it is close to zero. The seasonal variable can be calculated by stringing together these monthly values and then replicating the series for each year. As a result, the residue component (S_t) is found by subtracting the seasonal components and the trend cycle. R_t denotes the residue.

Mathematically, the Decomposition can be modeled like Eq. (2). Meanwhile, the additive model to capture seasonal variations in successive periods is presented in Eq. (3). The de-trend (D_t) processes and the complete decomposition are given in Eq. (4) and Eq. (5) [29].

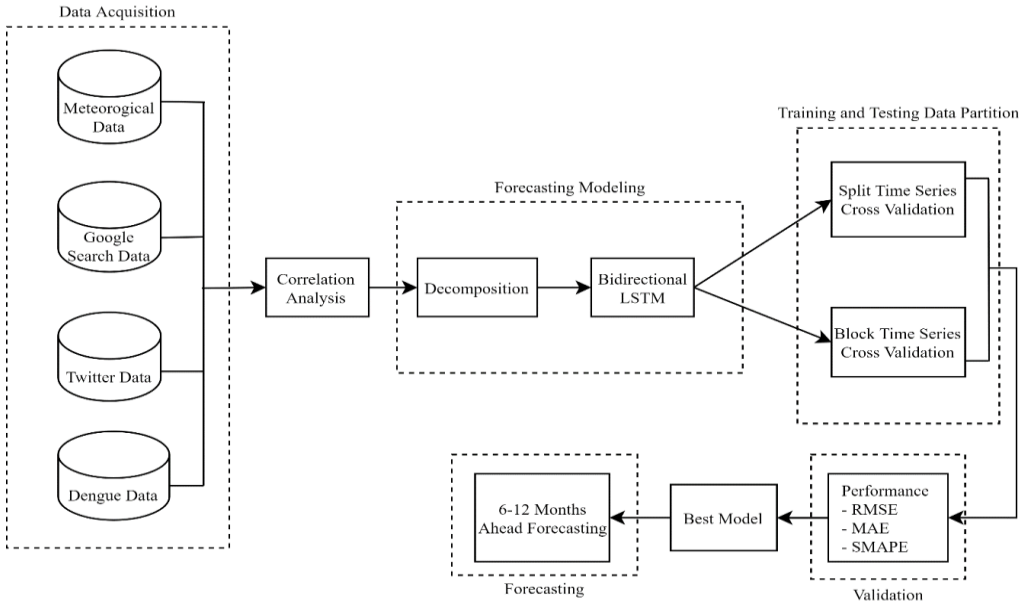

Figure. 2 Summarized framework for the construction of the D-BiLSTM forecasting model for DF cases number

$$Y_t = f(T, S, R) \quad (2)$$

$$Y_t = T_t + S_t + R_t \quad (3)$$

$$D_t = Y_t - T_t \quad (4)$$

$$R_t = Y_t - T_t - S_t \quad (5)$$

3.3.3. BiLSTM modeling

The LSTM model is a Recurrent Neural Network model that can be used to handle forecasting time-series data. This LSTM covers the drawbacks of gradient problems from the input applied to the hidden layer, increasing or decreasing significantly during a circular connection. Bidirectional LSTM (BiLSTM) improves LSTM performance by studying both forward and backward input sequences, combining and embedding both meanings in the hidden states. The bidirectional LSTM calculates the hidden forward sequence $\overrightarrow{h_t}$ then $\overleftarrow{h_t}$ retains information from the future in the backward run, potentially adding a necessary background to the prediction process.

In this study, we use a bidirectional LSTM model with architecture, as displayed in Fig. 3. This architecture consists of two BiLSTM blocks, each with one forward layer (FL) and one backward layer (BL). The $\overrightarrow{s_{l,i}}$ and $\overleftarrow{s_{l,i}}$ represents forward hidden state and backward hidden state vectors at layer l ∈ {1, 2} and frequency index i ∈ {1, 2, . . ., 21}. These output hidden states are concatenated ([$\overrightarrow{s_{l,i}}$; $\overleftarrow{s_{l,i}}$]) at each frequency index i and layer l before further processing. Also, wi's (i ∈ {1, 2, ..., 21}) are the scalar weights. At each frequency index, we built a two-layer BiLSTM architecture with a residual relation between the outputs of the first and second layers.

In BiLSTM block 1, the input is the result of the decomposition process of each variable denoted by x_i with i = 1,2, .., 21. The forward layer (FL_1) provides hidden state $\overleftarrow{s_{1,i}}$. Each hidden state can be denoted as x and y, where $\overrightarrow{\varphi_1}$ and $\overleftarrow{\varphi_1}$ shown parameters of FL_1 and BL_1.

$$\overrightarrow{s_{1,i}} = f(\overrightarrow{s_{1,i-1}}, x_i, \overrightarrow{\varphi_1}) \quad (6)$$

$$\overleftarrow{s_{1,i}} = f(\overleftarrow{s_{1,i+1}}, y_i, \overleftarrow{\varphi_1}) \quad (7)$$

The forward and backward hidden states in LSTM cells are now concatenated at each frequency index (i), and the resulting vector is denoted as $s_{i,k}$, where $s_{1,k} = [\overrightarrow{s_{1,k}}; \overleftarrow{s_{1,k}}]$. Meanwhile, in the BiLSTM block 2, we provide $s_{1,I}$ (i = 1,2,…,21) as input, corresponding forward as ($\overrightarrow{s_{2,i}}$), and backward ($\overleftarrow{s_{1,k}}$) hidden state vector, which is shown in Eq. (8) and Eq. (9).

$$\overrightarrow{s_{2,i}} = f(\overrightarrow{s_{2,i-1}}, s_{1,i}, \overrightarrow{\varphi_2}) \quad (8)$$

$$\overleftarrow{s_{2,i}} = f(\overleftarrow{s_{2,i+1}}, s_{1,i}, \overleftarrow{\varphi_2}) \quad (9)$$

The parameters for the FL_2 and BL_2 are $\overrightarrow{\varphi_2}$ and $\overleftarrow{\varphi_2}$, respectively. Furthermore, the concatenated output hidden states of $BiLSTM_2$ are $s'_{2,i} = [\overrightarrow{s'_{2,i}}, \overleftarrow{s'_{2,i}}]$, where $s'_{2,i} = s_{1,i} \oplus s_{2,i}$. We now obtained

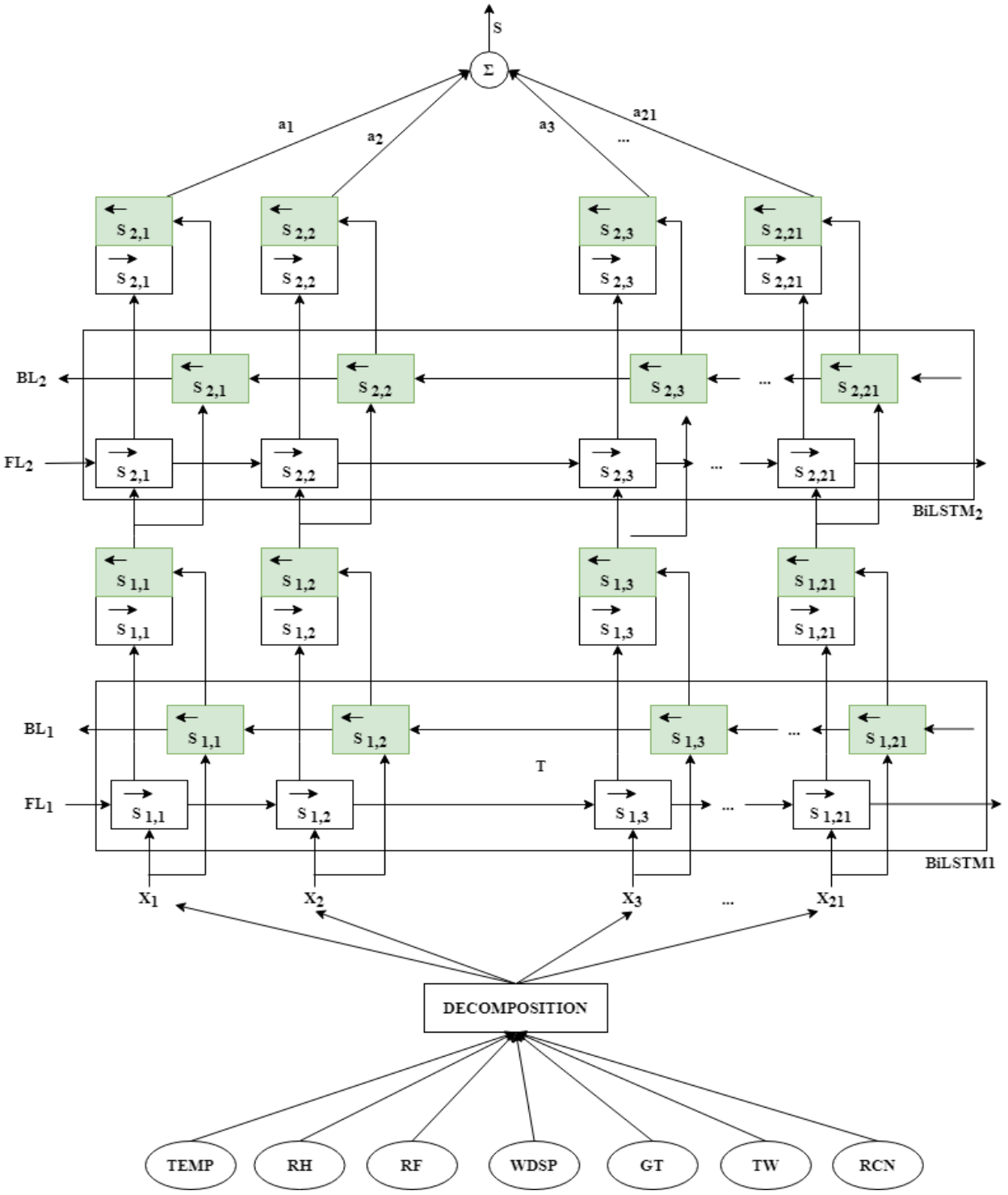

Figure. 3 Decomposition-BiLSTM architecture

single vector (s) by a weighted averaging mechanism as defined in Eq. (10).

$$s = \sum_{i=1}^{21} a_i s_i \quad (10)$$

where $a_i = \frac{exp(s_i^T w + b_w)}{\sum_{i=1}^{21} exp(s_i^T w + b_w)} \in (0,1)$, w_i, b_i is the trained parameter with other LSTM parameters. Finally, we get the expected output value y, shown in Eq. (11). In Eq. (11), v and b_v is another set of trainable weight parameters, and σ is the sigmoid function.

$$y = \sigma(v^T s + b_v) \quad (11)$$

3.3.4. Cross-validation for model robustness testing

Cross-validation techniques are used here to maintain the robustness of the D-BiLSTM model against various types and patterns of data in producing forecasts. In addition, this cross-validation technique permits the model to continue to improve its predicting capacity by capturing changes in the relationship between internet browsing data behavior, climate, and the number of reported cases. Two types of cross-validation are used here: five-fold split time-series cross-validation (STSCV) and blocked time-series cross-validation (BTSCV). The selection of cross-validation is based on the different patterns of series, types of forecasts, and evaluation objectives. STSCV divides the training set into two parts at each iteration because the training set is always earlier than the test set. Unlike ordinary cross-validation, which allows the training set to be located after the test set. Meanwhile, BTSCV divides the data by increasing the margin in two places. The first is between the training and testing folds to prevent the model from observing the lag values used as regressors and responses. The second is between the parts used in each iteration to avoid the model from remembering the pattern from iteration to the next iteration [41].

4. Result and discussion

This study has two main objectives. The first aim is to find the best combination of D-BiLSTM models to forecast the number of reported cases involving the Google Trends, Twitter, climate, and the number of cases reported variables in the previous period.

At the same time, the second objective is to find out how Google Trends and Twitter data involves in the DF Case forecasting results. There are eight models in each area. Each of these models was tested on different cross-validation scenarios to get a more robust model. The performance of each model in each area is compared with others models [16,34,37]. Then, a forecast for the next 6 and 12 periods is made using the best model for each area so that the Health Office and other stakeholders can plan preventive actions based on the results.

4.1 Time-series correlation analysis

The relationship between climate, Google Trends, and Twitter with the DF cases number are shown in Table 1. The correlation coefficient (r) between variables in Table 1 showed correlations between dengue case numbers and other variables. Temperature and rainfall are climate variables that have a small trend correlation among other climate variables. However, even though they are small, they can affect the number of DHF cases indirectly. It is relevant to previous studies presented by [23, 27, 34]. Temperature affects the growth of mosquitoes as vectors of DF disease [34]. The little correlation value here can be impacted by the fluctuating cases number in certain months where the value can drop significantly. This condition is consistent with [23, 27] that the continuity of relations should be analyzed annually, likewise with rainfall. The maximum rainfall of only 29 mm in the Malang Regency is too low, so that it is not sufficient to inhibit mosquito growth. On the other hand, rainfall can significantly inhibit mosquito reproduction to a minimum limit of 52 mm [34]. This condition happened in the Surabaya area, where it once reached a value of 52 mm so that the correlation value appears to be greater than the Malang district area.

Tabel 1. The Pearson correlation between the variable of DF cases number with other variables in each area. TEMP: temperature, RH: relative humidity, RF: rainfall, WDSP: wind speed, TW: Twitter; GT: Google Trends, RCN: reported case number

Variable	Reported Cases in Surabaya	Reported Cases in Malang
TEMP	-0.077	0.059
RH	0.293	0.373
RF	0.076	0.137
WDSP	0.128	-0.21
GT	0.275	0.253
TW	0.035	0.135
RCN Lag 1	0.868	0.545

The positive correlation value between Google Trends and reported cases shows that the greater the number of searches on google, the greater the number of reported cases. This positive relationship is relevant to the research conducted by [6-8, 15, 17]. However, in relation to correlation coefficients, this finding is somewhat in contrast to previous studies. Studies conducted by [6, 7] show relatively high correlation values (> 0.7) between google trends and DF reported at the national level. It could be due to different search query keywords. The level of internet penetration and the keywords used can influence the search frequency [45]. So it is natural that the search correlation value at the level differentiated by the level of penetration tends to be different from the local level without distinction (national level). The level of internet penetration between regions is very different, and local keywords can also be different from others [15]. The decreased correlation results when keywords were separated were also shown by [8, 17]. However, even though the value is only 0.275

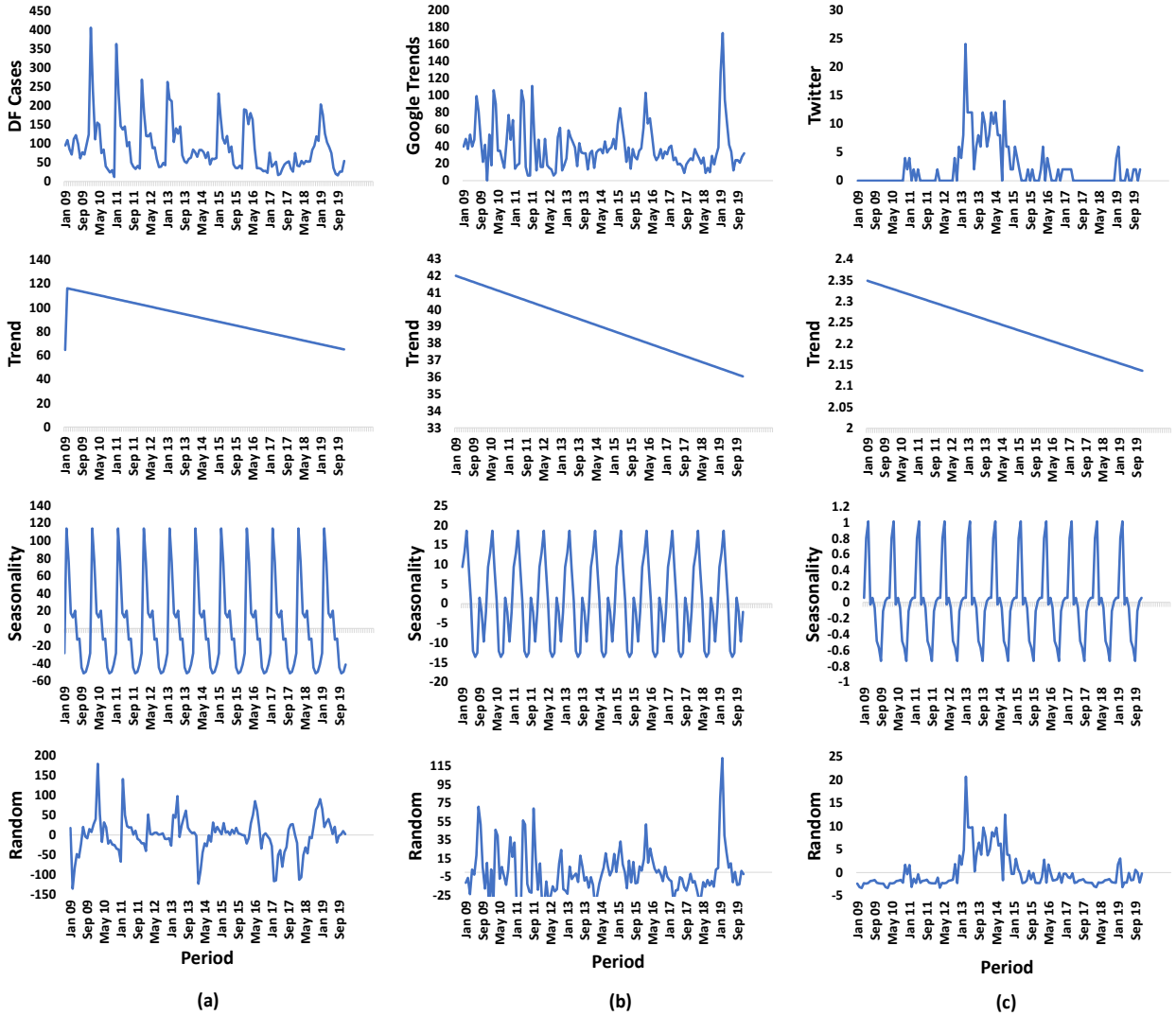

Figure. 4 Time series decomposition of the Malang area data broken down into trends, seasonality, and random components: (a) reported DF cases, (b) Google Trends, and (c) and Twitter

Table 2. Variable combination scenario implemented in each area. Output: RCN

Scenario	Best model	Input
1	M1	TEMP, RH, RF, WDSP
2	M2	TEMP, RH, RF, WDSP, GT
3	M3	TEMP, RH, RF, WDSP, TW
4	M4	TEMP, RH, RF, WDSP,GT, TW
5	M5	TEMP, RH, RF, WDSP, RCN Lag 1
6	M6	TEMP, RH, RF, WDSP, GT, RCN Lag 1
7	M7	TEMP, RH, RF, WDSP, TW, RCN Lag 1
8	M8	TEMP, RH, RF, WDSP,GT, TW, RCN Lag 1

for the Surabaya area and 0.253 in Malang Regency, this value is still said to affect the 0.01 level significantly.

Although Twitter's influence is not as strong as Google Trends, the positive correlation still shows that an increase also followed the rise in the number of Twitter in the number of DF cases reported. Twitter in Indonesia in the early study period was still rare and had only increased in the last three years [32].

Table 3. The result of D-BiLSTM models. Model 1-4 excludes the DF cases number at lag 1 in the independent variable combination. Model 5-8 includes the DF cases number at lag 1 in the independent variable combination.

Model	Surabaya-STSCV			Malang-STSCV			Surabaya-BTSCV			Malang-BTSCV		
	RMSE	MAE	SMAPE	RMSE	MAE	SMAPE	RMSE	MAE	SMAPE	RMSE	MAE	SMAPE
M1	100.304	72.683	49.174	86.422	65.887	53.056	109.387	73.294	48.427	89.201	68.270	49.053
M2	100.238	72.636	49.204	83.986	62.827	48.097	109.623	73.530	48.745	88.965	68.001	48.730
M3	100.288	72.710	49.208	83.870	62.719	47.949	109.485	73.381	48.774	89.515	68.633	49.584
M4	100.034	72.463	48.751	84.006	62.907	48.218	109.583	73.500	48.838	89.085	68.153	48.964
M5	100.109	72.525	49.131	84.187	63.084	48.597	109.824	73.782	49.419	89.579	68.674	49.741
M6	100.060	72.560	48.782	83.705	62.508	47.781	109.887	73.854	49.491	89.394	68.459	49.430
M7	100.142	72.513	49.197	83.712	62.546	47.705	109.728	73.652	49.003	89.316	68.383	49.026
M8	100.184	72.685	49.101	83.173	61.955	46.816	109.975	73.966	49.635	89.508	68.608	49.586

Table 4. The result of a 12-months forecast using D-BiLSTM models. Model 1-4 excludes the reported DF cases at lag 1. Model 5-8 includes the DF cases number at lag 1. Mark-bold models denote the best model

Area	Error	M1	M2	M3	M4	M5	M6	M7	M8
Surabaya	RMSE	53.785	44.491	68.861	68.712	**30.111**	32.694	68.883	68.763
	MAE	51.541	42.390	65.887	65.754	19.110	25.986	65.907	65.798
	SMAPE	0.548	0.502	0.592	0.592	0.310	0.389	0.592	0.592
Malang	RMSE	32.723	30.749	31.509	33.001	33.885	29.892	**28.651**	30.157
	MAE	29.117	27.253	27.901	29.377	30.125	26.467	25.727	26.700
	SMAPE	0.202	0.191	0.195	0.203	0.207	0.187	0.183	0.189

Table 5. The t-value result in the hypothesis test by comparing models that exclude and include the internet query variables. The threshold used is 1.645 with a significant level of 0.05. One-tailed test, H_0: $M_i - M_j \geq 0, i = 1,5;\ j = 2,3,4,6,7,8$

Paired Models	Surabaya		Malang	
	t-value	Conclusion	t-value	Conclusion
M1 - M2	1.885	M2 is better than M1	6.049	M2 is better than M1
M1 - M3	-6.831	M3 is not better than M1	11.407	M3 is better than M1
M1 - M4	-6.729	M4 is not better than M1	-19.932	M4 is not better than M1
M5 - M6	-50.510	M6 is not better than M5	18.622	M6 is better than M5
M5 - M7	-10.674	M7 is not better than M5	19.989	M7 is better than M5
M5 - M8	-50.787	M8 is not better than M5	17.874	M8 is better than M5

4.2 Decomposition-bidirectional LSTM (D-BiLSTM) model

The decomposition of each independent variable produces trend, seasonality, and random components. The results of the vector decomposition of the number of registered DF events, Google Trends, and Twitter are displayed in Fig. 4. The trend component indicates that the number of reported DF cases has a declining trend and Twitter and Google Trends variables. The many jagged random components, especially in the DF case number, indicate that there are events that cannot be explained by the periodicity of this data, which other factors can cause. Then, the seasonal component seems to repeat itself, and this occurs every close to 12 months.

This finding is relevant to previous research conducted in Indonesia [6]. This period is identical to the time of the rainy season. This seasonal period also looks similar to Google Trends and Twitter. Note that the seasonal and random components have negative values due to the decomposition and relative to the moving average month length. These negative scores on the random and seasonal components are for comparison and do not imply negative case numbers or search numbers.

DF cases data from Malang and Surabaya, climate data, local search queries, and social media are used to construct the BiLSTM model. Table 2 presents eight scenarios of variable combinations that produce the eight best models in each scenario. These scenarios are carried out in each area. Models from

Scenario 1-4 exclude dengue data reported case number at lag 1 [30], while models from Scenario 5-8 include reported DF case number at lag 1 [7].

In each scenario, tests were carried out using two types of 5-fold cross-validation, namely STSCV and BTSCV. This cross-validation is carried out in D-BiLSTM using 500 epochs, the mean square error loss function, and the sigmoid activation function. The other parameters were: 4 hidden layers, 32 units per hidden layer, 32 batch size, 0.5 drop out, 0.005 learning rate. Meanwhile, the optimizer used is the best optimizer selected from the optimization algorithm that produces the smallest root mean square error. The optimization algorithms were Adam, Adadelta, Stochastic Gradient Descent (SGD), Adagrad, Adamax, Nadam, RMSprop, and Ftrl. Testing was carried out using Python 3.5.3 includes the Keras module.

The best performing model in each area using two types of cross-validation with the best optimizer isshown in Table 3. The performance metrics used are root mean square error (RMSE), mean absolute error (MAE), and root mean absolute percentage. error (SMAPE) [13].

The performance of each model in Table 3 is the best average of each fold on STSCV and BTSCV. This condition is because they have experimented with the testing data in each area. The metric used is RMSE, and the best model has the smallest RMSE [40]. The use of two kinds of cross-validation aims to make the model learn with more data composition. In this study, the average performance of the D-BiLSTM model using STSCV has better performance than BTSCV. This performance is proper for both models. This condition can be caused because the conditions at the time of STSCV could get a more significant proportion of training data than BTSCV. So, the model can learn better with more data. It is relevant to previous studies, stating that training using more data will increase the model's performance.

Next, the retrain model uses cross-validation to get the results of forecasting data validation. The results of the retrain process are predicting the number of DF cases for data validation in the next period again. The results of the best model performance in each area are shown in Table 4.

Table 4 presents that the performance of the D-BiLSTM model involving DF case lag 1 produces better performance than the model without involving these variables. These findings are consistent with those presented in the experimental data testing shown in Table 3. This condition can occur because the DF case lag 1 has a very high correlation with the current DF case. Highly correlated variables can improve the performance of a model [6]. Table 4 also shows that adding Google Trends and Twitter variables can increase or decrease forecasting performance. In the Surabaya area, which represents high internet penetration rates, models without Google Trends and Twitter data have the best performance. Whereas in the Malang Regency area, a model that involves all data simultaneously is the best model.

The RMSE value in Table 4 also shows a different effect in areas with varying levels of internet penetration. In areas with high penetration, namely Surabaya, data on google trends and twitter is not proven to reduce error significantly. The RMSE value is getting bigger by involving these two variables, except M2, which requires Google Trends without DF cases in the previous period. M2 decreased by 9,295. For Malang Regency as the representative of areas with low internet penetration, RMSE in M2 and M3 looks lower than M1. It does not apply to M4. However, M6, M7, and M8 have a smaller RMSE value than M5. M2 has a decrease in the average RMSE of 1,099. A comparison test of two samples was carried out to determine whether the RMSE difference is significant enough to increase or decrease performance. Table 5 shows the t-student values obtained by paired samples test carried out on sample pairs M1 with M2, M3, M4, and M5 with M6, M7, M8.

The t-value in Table 5 shows that the addition of Twitter as a predictor variable is not proven to improve forecasting performance. However, for Google Trends, it is still quite significant to improve performance. However, if Google Trends include together with Twitter, it is still not enough to increase performance. It is somewhat different from the Malang Regency area, where the internet penetration rate is lower. The addition of the variables Google Trends and Twitter proved to significantly increase performance except when the two data were involved together and without involving the DF case in the previous period.

These results cannot be compared with previous studies because, to the best of the authors' knowledge of the DF case, most previous studies did not pay attention to the differences between areas with different internet penetration rates and local searches. However, if the previous research used national-level data, the results of this study are partly relevant. Previous research conducted by [8, 17] stated that using data at the national aggregate level, Google Trends data can improve forecasting quality. This finding is relevant for models in low internet penetration but not appropriate for models in high penetration areas with DF case lag 1. Several

Table 6. Comparison of model performance in areas with higher internet penetration rates in Surabaya. The mark-bold numbers represent the two highest-performing models.

Method	Error	M1	M2	M3	M4	M5	M6	M7	M8
LASSO [34]	RMSE	139.596	**141.882**	**140.185**	**138.189**	**74.519**	**75.159**	**75.89**	**77.764**
	MAE	92.529	91.665	92.256	90.235	44.263	46.433	46.604	46.730
	SMAPE	0.370	0.384	0.355	0.381	0.332	0.335	0.318	0.334
ARIMAX [18]	RMSE	**114.903**	168.958	228.106	216.911	128.291	125.818	141.761	128.291
	MAE	111.948	159.291	222.477	210.725	86.542	82.423	99.836	86.542
	SMAPE	0.621	0.704	0.747	0.740	0.323	0.313	0.341	0.323
NN [36]	RMSE	177.188	175.520	176.496	174.682	172.270	170.071	171.489	171.795
	MAE	100.390	98.878	99.413	97.735	99.054	96.885	97.672	97.555
	SMAPE	0.173	0.181	0.177	0.185	0.168	0.181	0.179	0.177
SVM [36]	RMSE	155.971	156.673	156.281	156.553	156.264	156.334	156.383	156.846
	MAE	81.837	82.336	81.628	82.207	81.635	81.918	82.234	82.260
	SMAPE	0.331	0.331	0.332	0.328	0.333	0.332	0.330	0.330
D-BiLSTM	RMSE	**53.785**	**44.491**	**68.861**	**68.712**	**30.111**	**32.694**	**68.883**	**68.763**
	MAE	51.541	42.390	65.887	65.754	19.110	25.986	65.907	65.798
	SMAPE	0.548	0.502	0.592	0.592	0.310	0.389	0.592	0.592

Table 7. Comparison of model performance in areas with lower internet penetration rates in Malang. The mark-bold numbers represent the two highest-performing models.

Method	Error	M1	M2	M3	M4	M5	M6	M7	M8
LASSO [34]	RMSE	**65.273**	**63.718**	**66.851**	**63.502**	**56.782**	**57.753**	**57.102**	**58.267**
	MAE	47.053	45.088	48.153	44.488	32.646	35.058	33.353	34.653
	SMAPE	0.320	0.314	0.311	30.675	33.924	32.057	32.872	32.405
ARIMAX [18]	RMSE	330.946	126.736	183.013	172.123	237.541	122.822	207.708	165.942
	MAE	78.057	112.082	174.602	162.850	112.599	65.229	98.943	82.594
	SMAPE	41.997	36.165	39.720	38.580	29.403	24.776	28.209	25.441
NN [36]	RMSE	104.763	102.854	103.429	101.395	102.262	100.385	100.805	101.760
	MAE	80.740	79.316	79.595	77.647	79.339	77.733	77.783	78.662
	SMAPE	0.138	0.139	0.144	15.751	13.643	14.905	15.168	14.571
SVM [36]	RMSE	90.376	88.456	87.828	88.715	69.874	68.790	68.927	68.979
	MAE	50.535	50.141	49.669	51.043	48.172	47.121	47.334	47.232
	SMAPE	0.301	0.303	0.301	31.152	29.280	29.357	29.237	29.301
D-BiLSTM	RMSE	**32.723**	**30.749**	**31.509**	**33.001**	**33.885**	**29.892**	**28.651**	**30.157**
	MAE	29.117	27.253	27.901	29.377	30.125	26.467	25.727	26.700
	SMAPE	0.202	0.191	0.195	0.203	0.207	0.187	0.183	0.189

explanations can be given regarding this contradiction. First, a prior research used a forecasting model that was analyzed per keyword [17], whereas, in this study, we used an aggregate of several keywords with the highest number of uses. This selection certainly causes different types of keywords to be used. Furthermore, varying types of keywords can cause accuracy differences [15]. In addition, [8] does not involve climate variables.

The only variables involved are reported data and Google Trends. It is relevant to what was stated by [17] that the differences in the variables involved could affect the model's performance. In addition, the random values of Google Trends and the very random Twitter, as shown in Fig. 2, can also influence this finding. Although the correlation is proven to be significant, it cannot improve the prediction result [21], likewise with Twitter data. However, let's look at the cases of other diseases. The findings of this study for the Surabaya area are relevant to those expressed by [21], who stated that online search data does not significantly improve forecasting performance.

To see how the performance position of the D-BILSTM model is compared to others, the D-BiLSTM model is compared with other models that have been used in previous studies. The results of the proposed performance model and its comparison in areas with a higher internet penetration rate -

Surabaya- are shown in Table 6. In contrast, areas with lower internet penetration are shown in Table 7. Table 6 and Table 7 show that the D-BiLSTM model has better performance than its rival model. The Lasso model is the second-best model, where even though the average RMSE is still above 100, it is still lower than the others. The outcomes of this research indicate that the D-BiLSTM model succeeded in reducing the average error. D-BiLSTM reduces RMSE by 93,129 and 94,054 for M1 in Surabaya and Malang. Thereafter 93,129 and 50,950, for M2, 106,406 and 41,419 for M3, 102,872 and 37,721 for M4, 102,725 and 86,504 for M5. Subsequently, 99,152 and 54,743 for the M6, 67,498 and 39,753 for the M7, 64,911 and 29,974 for the M8. The M5 model is the best model for D-BiLSTM in the Surabaya region, having a SMAPE of 0.310 and an MAE of 19,110. In the Malang area, the best model, M7, has a SMAPE value of 0.183 and MAE 25,727. These values are still below the 10% range of the Surabaya and Malang data intervals so that the forecasting results can still be said to be excellent.

The reported DF cases and forecast with the D-BiLSTM model and Lasso comparison with the best performance are shown in Fig. 5. Forecasting involves variables following the best models of D-BiLSTM, namely M5 in Surabaya and M7 in Malang. Fig. 5 shows that the forecast results of the D-BiLSTM model are closer to the actual data than the LASSO model. The trend and seasonal factors can influence this phenomenon, and random factors of dengue fever are associated with climate variables, Google Trends, and Twitter. The D-BiLSTM model studies data based on these three components, as shown in Fig. 2, so that the results can be closer to the actual data than studying one factor as a single value.

If observed in Fig. 2, the trend between reported DF and Twitter in the Malang area is the same.

Similarly, a seasonal pattern where the period is 12 months experienced the peak of events in February 2019. Then for the random component between Twitter rose also in February 2019 where previously tended to low. However, the random pattern for more varied DF cases up and down is drastic and up in December 2018. It is a similarity that makes the model easier to learn. This condition is the opposite in Surabaya. Here, models with Twitter and Google

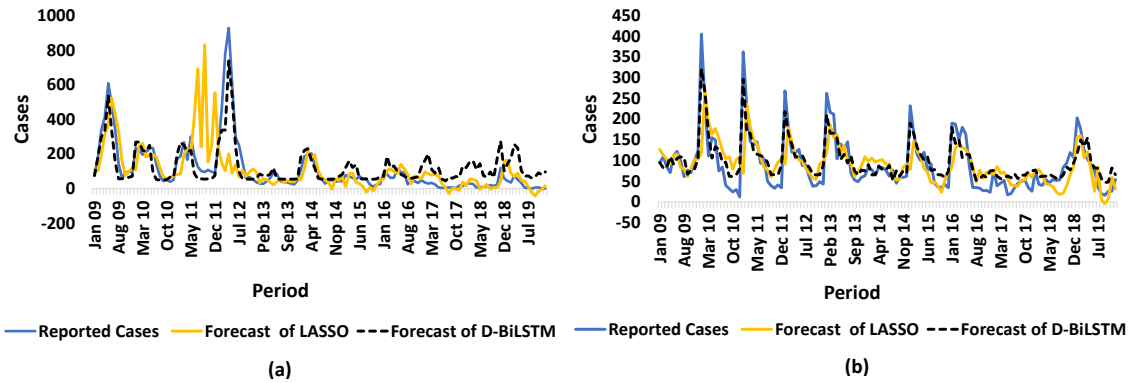

Figure. 5 Comparison of reported DF case data and forecast results with the D-BiLSTM and Lasso models in each area: (a) Surabaya (b) Malang.

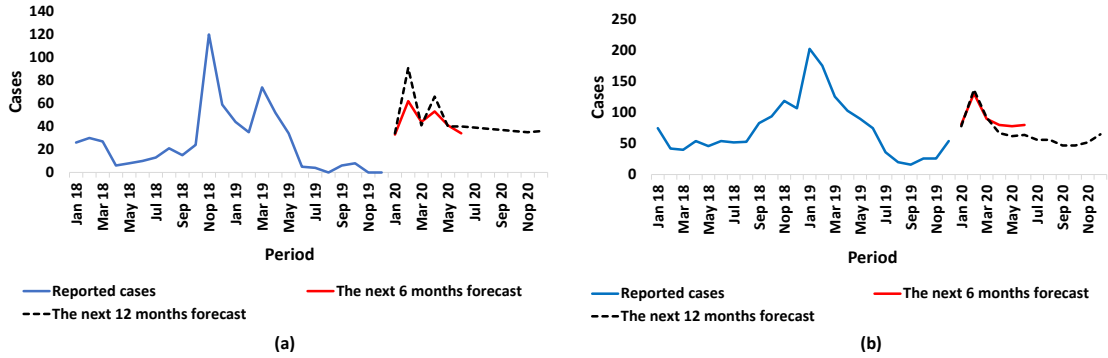

Figure. 6 Forecast results for the next 6 and 12 months using the D-BiLSTM model in each area: (a) Surabaya (b) Malang.

Trends are not the best models. If observed, the pattern of random DF cases reported more sloping in recent years, but this contrasts to Twitter and Google Trends that tend to increase. Similarly, the value of reported DF cases tends to decrease but rises for Twitter and Google Trends. This condition makes the D-BiLSTM model more responsive.

4.3 The next 6-12 periods forecasting

Action planning to avoid an increase in the number of cases can be done appropriately by forecasting the number of cases for several months ahead. The forward forecast period used is 6 and 12 months. This choice is based on the need for budget planning and actions taken in the study area [43]. Comparison of D-BiLSTM forecast skills for nowcast, six months, and 12 months ahead of the forecast using Model 5 is presented in Fig. 6 (a) and Fig. 6 (b).

Fig. 6 depicts the forecast patterns for the next 6 and 12 months in all areas have a similar pattern. However, relative to the previous 12 months, the forecasting results for the next six months tend to be identical to the value in the same month last year. This situation was relevant to the study presented by [13]. The more extended the forecast period in the future, there is a tendency for more extensive errors, which means that the pattern will be more different than the short term.

5. Conclusion and future works

To the best of our experience, this is the first study evaluating forecasting performance using internet query and social media data for dengue fever which separates searches based on differences in internet penetration rates between regions. Previous studies mostly used tracing data collected nationally, which was linked to the number of national cases, so that only policies that could be taken were at the national level. Whereas in reality, regional or local parties also need to prepare policies that are per the conditions of their respective regions where these conditions are not necessarily the same when viewed nationally.

In this research, the influence of Google Trends and Twitter variables is presented in 8 scenarios that produce 1440 models based on the combination of variables, the type of cross-validation, and the optimizer used. There are eight best models in each area. Twitter data is more influential on DF cases reported in areas with higher penetration rates. Google Trends is more significant than Twitter in areas with low and high internet penetration. Even though it is influential, forecasting involving Google Trends and Twitter at the local level does not necessarily increase forecasting performance, especially for areas with high internet penetration. Dengue fever forecasting is reported to have the best performance when it involves climate variables without Google Trends and Twitter variables in areas with high internet penetration rates. Meanwhile, the best model for low internet penetration rates involves climate, Twitter, and DF in the previous lag.

Based on correlation and forecasting, Google Trends and Twitter are not used to replace the reported data. Still, based on the similarity of the reported dengue fever case, Google Trends, and Twitter decomposition patterns, they may help describe the public's response to disease behavior. However, Twitter and Google Trends can be valuable sources of information. Nevertheless, this study has limitations, namely the limited number of areas.

For this reason, in the following research, the combination model obtained will be applied to different and more local areas. In addition, the analysis is focused on separating the number of low and high cases so that it is more apparent how the contribution of social media data to the development of DF cases.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Wiwik Anggraeni: conceptualization, methodology, formal analysis, writing—original draft preparation and editing. Eko Mulyanto Yuniarno: conceptualization, validation, formal analysis, writing—review. Reza Fuad Rachmadi: data curation, validation, writing—review. Mauridhi Hery Purnomo: supervision, conceptualization, formal analysis, writing—review. All authors read and approved the final manuscript.

Acknowledgments

We would like to express our gratitude to the Ministry of Research, Technology, and Higher Education of the Republic of Indonesia for providing research funding through the Doctoral Dissertation Research grant scheme, University Center of Excellence on Artificial Intelligence for Healthcare and Society (UCE AIHeS), and the Malang Regency Public Health Services for their assistance.

References

[1] W. Int, “Dengue and severe dengue”, 2018. https://www.who.int/news-room/fact-

sheets/detail/dengue-and-severe-dengue (accessed Mar. 27, 2019).

[2] WHO, "WHO | Dengue guidelines for diagnosis, treatment, prevention and control: new edition", *WHO*, 2017. https://www.who.int/rpc/guidelines/9789241547871/en/ (accessed Mar. 28, 2019).

[3] "WHO | 29 July 2016, vol. 91, 30 (pp. 349–364)", *WHO*. http://www.who.int/wer/2016/wer9130/en/ (accessed Nov. 25, 2019).

[4] F. Y. Nejad and K. D. Varathan, "Identification of significant climatic risk factors and machine learning models in dengue outbreak prediction", *BMC Medical Informatics and Decision Making*, Vol. 21, No. 1, p. 141, 2021.

[5] B. H. D. P2P, "Kesiapsiagaan Menghadapi Peningkatan Kejadian Demam Berdarah Dengue Tahun 2019 [Preparedness for Facing the Increased Incidence of Dengue Hemorrhagic Fever in 2019]| Direktorat Jendral P2P." http://p2p.kemkes.go.id/kesiapsiagaan-menghadapi-peningkatan-kejadian-demam-berdarah-dengue-tahun-2019/ (accessed Mar. 27, 2020).

[6] A. Husnayain, A. Fuad, and L. Lazuardi, "Correlation between Google Trends on dengue fever and national surveillance report in Indonesia," *Global Health Action*, Vol. 12, No. 1, p. 1552652, 2019.

[7] S. Yang, S. C. Kou, F. Lu, J. S. Brownstein, N. Brooke, and M. Santillana, "Advances in using Internet searches to track dengue", *PLoS Computational Biology*, Vol. 13, No. 7, 2017.

[8] R. A. Strauss, J. S. Castro, R. Reintjes, and J. R. Torres, "Google dengue trends: An indicator of epidemic behavior. The Venezuelan Case", *International Journal of Medical Informatics*, Vol. 104, pp. 26–30, 2017.

[9] R. Strauss, E. Lorenz, K. Kristensen, D. Eibach, J. Torres, J. May, and J. Castro, "Investigating the utility of Google trends for Zika and Chikungunya surveillance in Venezuela", *BMC Public Health*, Vol. 20, 2020.

[10] "World Telecommunication/ICT Indicators Database." https://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx (accessed Feb. 26, 2021).

[11] E. Hagg, V. S. Dahinten, and L. M. Currie, "The emerging use of social media for health-related purposes in low and middle-income countries: A scoping review," *International Journal of Medical Informatics*, Vol. 115, pp. 92–105, 2018.

[12] E. H. Chan, V. Sahai, C. Conrad, and J. S. Brownstein, "Using Web Search Query Data to Monitor Dengue Epidemics: A New Model for Neglected Tropical Disease Surveillance", *PLoS Neglected Tropical Diseases*, Vol. 5, No. 5, p. e1206, 2011.

[13] P. Rangarajan, S. K. Mody, and M. Marathe, "Forecasting dengue and influenza incidences using a sparse representation of Google trends, electronic health records, and time series data", *PLoS Computational Biology*, Vol. 15, No. 11, 2019.

[14] Y. Zhang, H. Bambrick, K. Mengersen, S. Tong, and W. Hu, "Using Google Trends and ambient temperature to predict seasonal influenza outbreaks", *Environment International*, Vol. 117, pp. 284–291, 2018.

[15] P. Guo, L. Wang, Y. Zhang, G. Luo, Y. Zhang, C. Deng, Q. Zhang, and Q. Zhang, "Can internet search queries be used for dengue fever surveillance in China?", *International Journal of Infectious Diseases*, Vol. 63, pp. 74–76, 2017.

[16] A. Wilder-Smith, E. Cohn, D. C. Lloyd, Y. Tozan, and J. S. Brownstein, "Internet-based media coverage on dengue in Sri Lanka between 2007 and 2015", *Global Health Action*, Vol. 9, No. 1, p. 31620, 2016.

[17] W. Anggraeni and L. Aristiani, "Using Google Trend data in forecasting number of dengue fever cases with ARIMAX method case study: Surabaya, Indonesia", In: *Proc. of 2016 International Conference on Information Communication Technology and Systems (ICTS)*, Surabaya, Indonesia, pp. 114–118, 2016.

[18] H. T. Ho, T. M. Carvajal, J. R. Bautista, J. D. R. Capistrano, K. M. Viacrusis, L. F. T. Hernandez, and K. Watanabe, "Using Google Trends to Examine the Spatio-Temporal Incidence and Behavioral Patterns of Dengue Disease: A Case Study in Metropolitan Manila, Philippines", *Tropical Medicine and Infectious Disease*, Vol. 3, No. 4, 2018.

[19] Y. Teng, D. Bi, G. Xie, Y. Jin, Y. Huang, B. Lin, X. An, D. Feng, and Y. Tong, "Dynamic Forecasting of Zika Epidemics Using Google Trends", *PLoS ONE*, Vol. 12, No. 1, 2017.

[20] M. Verma, K. Kishore, M. Kumar, A. R. Sondh, G. Aggarwal, and S. Kathirvel, "Google Search Trends Predicting Disease Outbreaks: An Analysis from India", *Healthcare Informatics Research*, Vol. 24, No. 4, p. 300, 2018.

[21] M. Kapitány-Fövény, T. Ferenci, Z. Sulyok, J. Kegele, H. Richter, I. V. Nagy, and M. Sulyok, "Can Google Trends data improve forecasting of

Lyme disease incidence?", *Zoonoses Public Health*, Vol. 66, No. 1, pp. 101–107, 2019.

[22] Z. Du, L. Xu, W. Zhang, D. Zhang, S. Yu, and Y. Hao, "Predicting the hand, foot, and mouth disease incidence using search engine query data and climate variables: an ecological study in Guangdong, China", *BMJ Open*, Vol. 7, No. 10, 2017.

[23] G. Zhu, T. Liu, J. Xiao, B. Zhang, T. Song, Y. Zhang, L. Lin, Z. Peng, A. Deng, W. Ma, and Y. Hao, "Effects of human mobility, temperature and mosquito control on the spatiotemporal transmission of dengue", *Science of the Total Environment*, Vol. 651, pp. 969–978, 2019.

[24] A. A. E. Metwally, "Google Search Trend of Dengue fever in developing Countries in 2013-2014: An Internet-Based Analysis", *Journal of Health Informatics in Developing Countries*, Vol. 9, No. 1, 2015.

[25] R. Jain, S. Sontisirikit, S. Iamsirithaworn, and H. Prendinger, "Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data", *BMC Infectious Diseases*, Vol. 19, No. 1, p. 272, 2019.

[26] Z. Husnina, A. C. A. Clements, and K. Wangdi, "Forest cover and climate as potential drivers for dengue fever in Sumatra and Kalimantan 2006–2016: a spatiotemporal analysis", *Tropical Medicine and International Health*, Vol. 24, No. 7, pp. 888–898, 2019.

[27] A. Appice, Y. R. Gel, I. Iliev, V. Lyubchich, and D. Malerba, "A Multi-Stage Machine Learning Approach to Predict Dengue Incidence: A Case Study in Mexico", *IEEE Access*, Vol. 8, pp. 52713–52725, 2020.

[28] M. Salathé, "Digital epidemiology: what is it, and where is it going?", *Life Sciences, Society and Policy*, Vol. 14, No. 1, p. 1, 2018.

[29] Y. Zhang, H. Bambrick, K. Mengersen, S. Tong, L. Feng, L. Zhang, G. Liu, A. Xu, and W. Hu, "Using big data to predict pertussis infections in Jinan city, China: a time series analysis", *International Journal of Biometeorology*, Vol. 64, No. 1, pp. 95–104, 2020.

[30] Y. Zhang, L. Yakob, M. B. Bonsall, and W. Hu, "Predicting seasonal influenza epidemics using cross-hemisphere influenza surveillance data and local internet query data", *Scientific Reports*, Vol. 9, 2019.

[31] "BULETINAPJIIEDISI74November2020.pdf." Accessed: Mar. 31, 2021. [Online]. Available: https://apjii.or.id/downfile/file/BULETINAPJIIEDISI74November2020.pdf

[32] A. S. Jati, "Jumlah Pengguna Twitter Meningkat, Tapi... [The number of Twitter users is increasing, but ...]", *detikinet*. https://inet.detik.com/cyberlife/d-5001786/jumlah-pengguna-twitter-meningkat-tapi (accessed Feb. 27, 2021).

[33] P. KOMINFO, "Indonesia Peringkat Lima Pengguna Twitter [Indonesia Ranks Five Twitter Users]", *Website Resmi Kementerian Komunikasi dan Informatika RI*. http:///content/detail/2366/%20indonesia-peringkat-lima-penggunatwitter/0/sorotan_media (accessed Feb. 27, 2021).

[34] S. Mala and M. K. Jat, "Implications of meteorological and physiographical parameters on dengue fever occurrences in Delhi", *Science of the Total Environment*, Vol. 650, pp. 2267–2283, 2019.

[35] A. Q. Munir, S. Hartati, and A. Musdholifah, "Early Identification Model for Dengue Haemorrhagic Fever (DHF) Outbreak Areas Using Rule-Based Stratification Approach", *International Journal of Intelligent Engineering & Systems*, p. 15, 2018.

[36] S. Jiang, R. Xiao, L. Wang, X. Luo, C. Huang, J. Wang, K. Chin, and X. Nie, "Combining Deep Neural Networks and Classical Time Series Regression Models for Forecasting Patient Flows in Hong Kong", *IEEE Access*, Vol. 7, pp. 118965–118974, 2019.

[37] J. Xu, K. Xu, Z. Li, F. Meng, T. Tu, L. Xu, and Q. Liu, "Forecast of Dengue Cases in 20 Chinese Cities Based on the Deep Learning Method", *International Journal of Environmental Research and Public Health*, Vol. 17, No. 2, 2020.

[38] K. Lan, D. Wang, S. Fong, L. Liu, K. K. L. Wong, and N. Dey, "A Survey of Data Mining and Deep Learning in Bioinformatics", *Journal of Medical Systems*, Vol. 42, No. 8, p. 139, 2018.

[39] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects", *Science*, Vol. 349, No. 6245, pp. 255–260, 2015.

[40] S. Chae, S. Kwon, and D. Lee, "Predicting Infectious Disease Using Deep Learning and Big Data", *International Journal of Environmental Research and Public Health*, Vol. 15, No. 8, p. 1596, 2018.

[41] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation", *Information Sciences*, Vol. 191, pp. 192–213, 2012.

[42] "Pemerintah Kota Surabaya." https://www.surabaya.go.id/id/page/0/8227/geografi (accessed Mar. 31, 2021).

[43]"Dinas Komunikasi dan Informatika Kota Surabaya [Surabaya City Government]." https://dinkominfo.surabaya.go.id/index.php?pages=detail_berita&id_berita=190 (accessed Mar. 31, 2021).
[44]"malangkab-Kondisi Geografis.pdf [Geographical Condition of Malang District]." Accessed: Mar. 05, 2021. [Online]. Available: http://malangkab.go.id/uploads/dokumen/malangkab-Kondisi%20Geografis.pdf
[45]G. Cervellin, I. Comelli, and G. Lippi, "Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings", *Journal of Epidemiology and Global Health*, Vol. 7, No. 3, pp. 185–189, 2017.