



Modelling Dengue Spread as Dynamic Networks of Time and Location Changes

Arfinda Setiyoutami¹ Diana Purwitasari^{2,3} Wiwik Anggraeni^{1,3,4}
 Eko Mulyanto Yuniarno^{1,5} Mauridhi Hery Purnomo^{1,3,5*}

¹Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

²Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

³University Center of Excellence on Artificial Intelligence for Healthcare and Society (UCE AIHeS), Indonesia

⁴Department of Information System, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

⁵Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

* Corresponding author's Email: hery@ee.its.ac.id

Abstract: Since local human movements can influence dengue spread, a network-based prediction model considers the dynamic relation between dengue case incidences and their location over time. Some approaches often generated the networks in a certain period with a single spanning time until several months or in one year, called static networks. However, one annual-based dengue spread model could have different simulations depending on the selected months to show different dynamicity. Other approaches do not involve any networks for generating the spread models but employ them for validating the models with simulations. Considering the evolution of dengue circumstances that could quickly change between periods, we proposed a Dengue Spread Dynamic Network (DSDN) model with some timespan and location boundaries variants. DSDN includes five network models with nodes representing localities and links showing dengue spread which varied every day depending on infections presence and environmental conditions in a certain period. With our proposed method, daily dengue spread from one location to another can be predicted based on the location-based incidence historical data as outbreaks prevention initiative. We also analyzed how dengue spreads differently in outbreak and non-outbreak periods using Dynamic Network Link Prediction (DNLP) method. From our experiments result, Neighbor Network which modeled that dengue only spreads between neighboring localities produced an accuracy of 92.54% for the entire period. When applied only to outbreak data, there was a performance increase of 3.39 points, which suggested that link prediction performs better when dengue is rapidly spreading. In addition to that, our experiments concluded that dengue potentially spreads to a location with no current infections if local incidence often occurred in the past.

Keywords: Dengue spread dynamic network, Link prediction, Disease spread model, Long short-term memory.

1. Introduction

Dengue is a disease caused by an arthropod-borne dengue virus (DENV), which is transmitted between humans through the bit of female *Aedes* mosquitoes [1]. The disease is distinguished by its severity, including the classic dengue fever (DF), severe dengue with plasma leakage/ dengue hemorrhagic fever (DHF), and dengue with systemic shock/ dengue shock syndrome (DSS) [2]. During the past 60 years, dengue has spread geographically, mostly in tropical and sub-tropical countries. It was responsible for 1.14 million disability-adjusted life-

years in 2013, with an estimation of 100 million cases per year in over 125 countries [3, 4]. Due to the burden posed by dengue fever infection, it is necessary to model its spread and analyze potential endemic areas characteristics, so that preventive and control measures can be carried out.

Authors in [5] proposed dengue endemic areas stratification method as an early outbreak identification in a certain year. It provided area risk mapping to support the Ministry of Health in initiating prevention measures before the outbreak. However, the work did not address how dengue

spreads between areas, and how the spread pattern differed during outbreak and non-outbreak periods.

Since the movements of the mosquito vector are very restricted, human mobilities play a key role in confirming infection risk and modelling the patterns of virus spread [6]. The spread between entities, which can either be humans or locations, creates a relationship that forms a network. Previous works on disease spread network modelling includes simulating the implementation of standard compartmental models, such as susceptible-infectious-recovered (SIR) [7], improved SIR (ISIR) [8], as well as susceptible-exposed-infectious-recovered (SEIR) model [9] on existing contact network data. These approaches could estimate how a disease progresses in a network based on a predefined formula, and examine the appropriate immunization strategies. By understanding disease transmission through a network structure, it helps decision makers to determine infection distribution and disease control properly [10].

Those simulations were used as the foundation for developing epidemic control strategies. However, it requires knowledge of epidemiology to calculate the model formula, which health agencies do not always have, especially in developing locations. Moreover, in areas where dengue rapidly spreads, it is necessary to have an approach that can quickly identify the spread pattern only from the historical data.

Researches [11, 12] proposed a dengue spread network model, where a neighborhood was represented as a node, with links representing people who moves from their residences to their place of daily activities. These works analyzed which movements impact the dynamics of dengue, and which nodes that become the most important outbreak drivers. Meanwhile, authors in [13] used two-mode network for modelling dengue epidemic behaviour from the perspective of complex network. In the projected one-mode network, two locations were connected if both share the same week of incidents.

These studies were able to identify which nodes or locations that have higher infection rates, thus may help health agencies to establish a disease control management when an outbreak occurs. However, the works did not discuss how the network model evolved, and how nodes interacted with each other. Considering that dengue circumstances can change from time to time, it can have an impact on the relationship between locations. Therefore, it is necessary to analyze dengue progression in the form of dynamic model, so that it can help the decision

makers to infer disease characteristics and predict the spread.

Our research proposes Dengue Spread Dynamic Network (DSDN) model with nodes representing locations, and links representing virus spread. In determining the links, we compiled five different scenarios to generate networks with different timespan and nodes grouping or clustering. Then, we predict how dengue infection will develop, both during the outbreak and non-outbreak periods using dynamic network link prediction (DNLP) method. From the link prediction results, we analyzed how dengue spreads between localities in a predetermined period.

The remainder of this article is organized as follows. Section 2 reviews the related studies in network representation and dynamic network link prediction. Section 3 describes the methodology of this research, including how to construct the dataset, infer dynamic network from the dataset, and predict dengue spread using link prediction approach. Section 4 explains the experimental results and discussion. Section 5 describes the conclusion of this research, as well as direction for future works.

2. Related works

2.1 Network representation for modelling the spread of infectious diseases

There are various network structures that have been used in modelling the spread of infectious diseases. Previous studies mostly used standard compartmental model to simulate disease spread in a general network data. Authors in [7] applied SIR on human contact networks in several environments, such as conference, hospital, school, and gallery. Meanwhile, authors in [8] applied the Improved SIR (ISIR) model on an artificial social contact network generated by BA generator, and authors in [9] applied SEIR model on high-resolution human contact network between conference attendees. Other than that, authors in [14] proposed SIR-network model with city's neighborhoods being the nodes, and fractions of people moving between neighborhoods as the directed edges.

Other studies used historical disease incidences data to quickly analyze the spread, without formulating the epidemic model beforehand. Authors in [13] generated a location-based network structure, where the spread of dengue is modelled by establishing weekly dengue cases in different locations as a complex two-mode network. In this type of two-mode network, nodes are separated into primary and secondary sets, where links are only

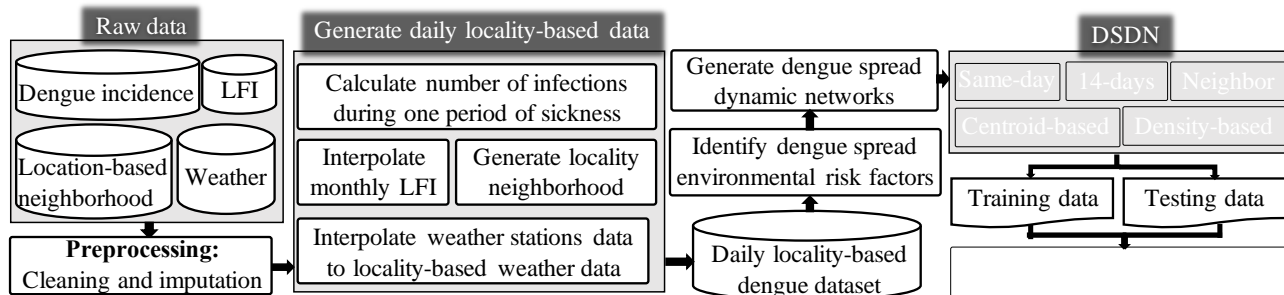


Figure. 1 Generating DSDN for predicting dengue spread using E-LSTM-D

specified between nodes in different sets. The two-mode network was then projected into one-mode network, where locations were connected through edges which represented co-occurring dengue incidences in one week. By modelling the spread of dengue from a location perspective, it was possible to identify which localities were only slightly affected by the virus. It could help investigators to identify the precautions which were taken to suppress the virus spread.

The network used in the previous study is a static network that describes the spread in only one period. The work did not discuss how the network evolved, and how nodes interacted with each other. In order to understand how dengue progresses, it is necessary to know how it spreads over time, so that its characteristics can be inferred. By understanding the dynamics of dengue, it is also possible to make predictions on how dengue circumstances will develop. The proposed method models dengue spread in the form of dynamic networks which consists of graph snapshots, each represents dengue relationship between locations in one day.

2.2 Link Prediction in dynamic networks

Dengue epidemic is a dynamic phenomenon, which can be represented in the form of dynamic networks. Link prediction of a dynamic network tries to predict how its structure evolves, thus explaining the relationships between topologies [15]. In dynamic networks, temporal information needs to be considered and included in the analysis, which is often overlooked by models designed for static networks. Therefore, constructing dengue spread network model requires method that is designed for dynamic networks. By predicting the links that are going to appear or disappear, it is possible to understand how one location gets infected by dengue virus, and how it recovers.

Several studies on DNLP include the use of Random walk which was able to predict future links efficiently in temporal uncertain social networks [16]. Authors in [17] previously applied Random walk on

static networks, while authors in [18] applied it for learning dynamic/ time-dependent network without loss of information. Other than that, authors in [19] proposed link prediction approach based on the attraction force between nodes (DLPA). Using this method, it was possible to detect the missing links and predict potential links in the upcoming period. In addition, with the development of deep learning, authors in [20] proposed Deep Dynamic Network Embedding (DDNE), which made embeddings for new links using deep architecture and measured the similarity of nodes to address the neighbor's influence.

These methods were able to model a network evolution and predict whether there would be missing links or potential links. However, the methods tend to ignore historical information contained in the earlier period, as it only made use of a few historical snapshots. Authors in [15] proposed E-LSTM-D, an end-to-end deep learning framework, which incorporated encoder-decoder architecture to learn network representations, and a stacked long short-term memory (LSTM) to learn the temporal features. This method allowed historical information to be fully used when generating the model, by learning the time dependencies between network snapshots. While most existing methods only focused on predicting links that were going to appear, E-LSTM-D was also able to predict links that were going to disappear, which suited the dynamics of a disease spread network. On top of that, by fine tuning the network structure, this method also allowed prediction for networks on various scales, such as the dengue spread networks that were generated in this research.

3. Research methodology

3.1 Dataset

Malang Regency is one of the dengue endemic areas located in East Java province with the second highest number of dengue incidences among other regencies and municipalities in Indonesia. It consists

of 390 villages which located in 33 sub-districts. This research focused on the dataset that was obtained from Malang Regency's Health Office and Meteorology Climatology and Geophysics Council (BMKG). As displayed in Fig. 1, the raw data consists of dengue fever incidence, larvae free index (LFI), location-based neighborhood, and weather.

Dengue fever incidence data during 2017 to mid-2019 were obtained from community health centers on sub-district level, hereinafter referred to as locality. The data included patient registry which consists of an individual's demographic (age, sex), location (sub-district, village, reporting locality/hospital), and periods indicating the disease progression (dates of symptoms onset, as well as hospital's admission and discharge). The incidences trend showed a spike in January which continued to peak in February 2019. The increase in cases had been occurred from September 2018, which indicated an outbreak period.

Larvae free index (LFI) is a measurement used to identify the presence of mosquito vector in an area. The value was obtained from larvae inspection activity in residential houses, which is calculated as the percentage of the number of houses with zero larvae compared to the total number of houses inspected [21]. LFI raw data contained monthly index in each locality.

Location-based neighborhood data contained neighboring sub-districts in northern, eastern, southern, and western boundaries. The incorporation of location-based neighborhood data aimed to analyze how dengue incidence in an area would be affected by its environmental conditions [22].

Daily weather data consisted of maximum, minimum, and average temperature, rainfall, average humidity, sunshine duration, as well as maximum and average wind speed. The data were recorded in two weather stations in Malang Regency and one neighboring station in Pasuruan Regency as supplementary data. The importance of including weather in the dataset was based on the analysis that temperature, rainfall, humidity, and wind speed were significant weather factors associated with dengue cases [23].

3.2 Create Daily Locality-Based dataset

As illustrated in Fig. 1, preprocessing was the initial step carried out after raw data collection. It consisted of correcting inaccurate and duplicate entries. The inaccuracy included discrepancies in dengue incidence data due to the manually recorded patient registry, incorrect sub-district neighborhood mapping, and undefined values in weather data. In

addition to that, imputation was performed on unrecorded weather values in a certain period.

Prior to creating the dengue spread model from the dataset, it is necessary to standardize its unit. As dengue incidence data can change rapidly every day in each locality, the unit was standardized into daily-based and locality-based. Generating daily locality-based dataset included calculating number of infected people during period of sickness, interpolating monthly LFI and stations-based weather data, as well as breaking down sub-districts neighborhood data into locality-based.

Period of sickness started from three days before the patient's sick date, while ended on 11 days after. This calculation was based on the minimum days in dengue Intrinsic Incubation Period (IIP), which lasts for 3 to 8 days. The number of days in the period of sickness is 14, which was the upper limit of dengue sickness duration [24]. Hence, dengue infection on day t was calculated by accumulating co-occurring incidences during day $t-3$ to $t+11$.

Interpolation was performed on LFI and weather data. Obtaining daily LFI data used Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) method. PCHIP was proven to be able to fill in missing or incorrect values [25], as well as replacing negative values for fitting rainfall data [26]. Whereas generating locality-based weather data used Kriging interpolation algorithm, as it was able to estimate precipitation, temperature, wind speed, humidity, cloudiness, sunshine duration, as well as rainfall data through interpolation [27].

3.3 Identify Risk Factor of Dengue Spread

Feature ranking method was used to determine which factors were the most influential in correlation to the number of dengue incidences [28]. It included

Table 1. Feature score and significance analysis of environmental variables to dengue incidence

Environmental Factors	Feature #	Feature Score
Infected people at neighboring localities	F1	0.275
Minimum temperature	F2	0.044
Maximum temperature	F3	0.040
Average temperature	F4	0.053
Average humidity	F5	0.054
Rainfall	F6	0.035
Sunshine duration	F7	0.057
Maximum wind speed	F8	0.034
Average wind speed	F9	0.022
Larvae Free Index	F10	0.387

calculating the feature score using Random Forest as shown in Table 1, then incorporating the features into a dengue incidence prediction model. Experiments were carried out using a combination of features with the best scores. Based on the prediction accuracy result, features that were considered as environmental risk factors are five features with the highest scores including the number of infected people in neighboring locality (F1), average temperature (F4), average humidity (F5), sunshine duration (F7), and LFI (F10).

3.4 Generate dengue spread dynamic network

Fig. 2 displays undirected network with nodes representing localities. It was visualized using Bokeh visualization library, which was displayed on top of Malang Regency map view. The nodes were plotted based on the longitude and latitude coordinates of each locality, while edges were generated from neighboring localities matrix. Locality i is linked to locality j when j is the immediate neighbor of i . In the figure, the nodes are visually separated by circular boundaries representing neighborhood groups, with three localities in the first group, five localities in the second group, and the rest of 31 localities in the other group.

We modelled DSDN using five different scenarios based on our hypothesis about how dengue spreads spatially in a network. The network links indicated dengue spread between localities, which defined by co-occurring infections on each day. The dynamic network consists of sequential graphs, where $G_k = (V, E_k)$ is the k th snapshot of the network, V is the set of nodes representing locality, and E_k describes temporal links within timespan $[t_{k-1}, t_k]$. For each G_k , adjacency matrix A_k represents the links between nodes. Each element a in $A_k \in \{0, 1\}$ shows whether an infection exists in locality i, j ($a_{k;i,j} = 1$) or does not exist ($a_{k;i,j} = 0$).

Table 2 shows the characteristics of each generated network model. In our models, we assume that when dengue spread occurred in two localities, each locality influenced each other equally. Thus, each network model was in the form of unweighted symmetrical directed network (the values of $a_{d;i,j}$ and $a_{d;j,i}$ were always equal). Same-day Network was modelled based on dengue infections at different localities in one day, which was determined when the number of co-occurring infections in locality i and j on day d is greater than 0 ($infect_{i,d} > 0 \ \& \ infect_{j,d} > 0$). Meanwhile, other four models covered a longer timespan by taking into account all

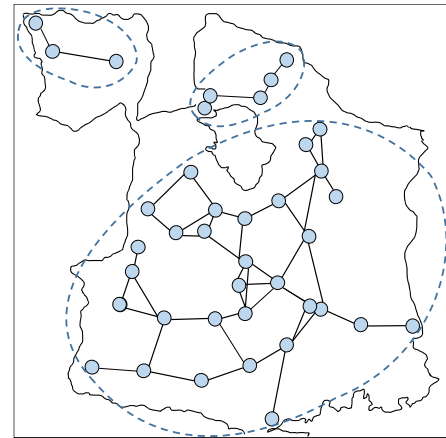


Figure. 2 Visualization of locality neighborhood groups in the form of undirected network

Table 2. The characteristics of the generated dengue spread network models

Network	Conditions for links $i \leftrightarrow j$	
	Group/cluster	Infection presence
Same-day	-	$infect_{i,d} > 0 \ \& \ infect_{j,d} > 0$
14-days	-	$\sum_{d=t-3}^{t+11} infect_{i,d} > 0$ & $\sum_{d=t-3}^{t+11} infect_{j,d} > 0$
Neighbor	$neigh_i = neigh_j$	
Centroid-based	$clust_{i,d} = clust_{j,d}$	
Density-based	$clust_{i,d} = clust_{j,d}$	

period of sickness $d - 3$ to $d + 11$ ($\sum_{d=t-3}^{t+11} infect_{i,d} > 0 \ \& \ \sum_{d=t-3}^{t+11} infect_{j,d} > 0$). In addition to that, group-based and cluster-based models also included grouping/clustering constraints, where the links between nodes only existed when localities i and j on day d were in the same group or cluster ($neigh_i = neigh_j / clust_{i,d} = clust_{j,d}$).

Fig. 3 displays the visualization of all five dynamic network models within period t to $t + 2$ using the same dataset. It was generated based on randomly selected samples in the dataset, which included incidences on day-816 (t), 817 ($t + 1$), and 818 ($t + 2$) in 7 localities from different neighborhood groups. Localities L1 and L2 represented group 1, localities L3 and L4 were in group 2, while localities L5, L6, and L7 were in group 3. The infections and recoveries which occurred during period $t - 2$ to $t + 2$ were as follows:

- day $t - 2$: infections existed in all localities L1, L2, L3, L4, L5, L6, and L7.
- day $t - 1$: infections still existed in localities L2, L3, L4, L5, and L7, while infections in localities L1 and L6 were recovered, and no new infections existed.

- day t to day $t + 2$: infections still existed in localities L2, L3, L4, L5, and L7, and no new infections existed.

It can be seen from the visualization that the dynamics of each network model is different depending on its characteristics and how it was generated.

For dengue spread prediction purposes, we transformed each network model into an adjacency matrix with the dimension of $T \times \sum locality \times \sum locality$, where T was the number of days in one

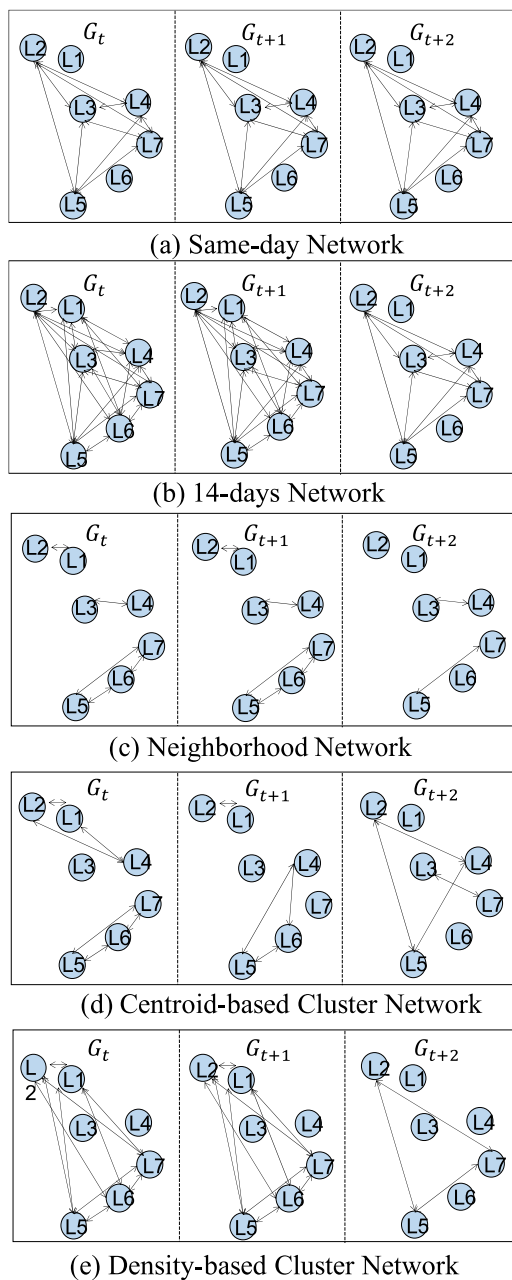


Figure. 3 Visualization of dengue spread network models of localities P1 to P7 showing different dynamic network evolution within the same period

network, and $\sum locality$ was the number of localities. The matrix element was defined as 1 if locality i and locality j nodes were connected through link, otherwise it was stated as 0. Two connected localities were considered to be related in terms of affecting dengue spread in day t based on the conditions determined in each network.

3.4.1. DSDN based on infections in the same day

Same-day Network was modelled based on our hypothesis that two locations influence each other in the spread of dengue if there are co-occurring infections in the same day. Fig. 3 (a) shows the dengue spread dynamics, where infections only existed in localities L2, L3, L4, L5, and L7 on day t . It can be seen in graph G_t that those five nodes were linked and considered influential on the spread. On day $t + 1$ and day $t + 2$, there were no new infections and recoveries occurred, thus the visualizations of graphs G_{t+1} and G_{t+2} were the same as G_t .

3.4.2. DSDN based on co-occurring infections during period of sickness

This 14-days Network covered a longer timespan compared to Same-Day Network, as it was generated based on the period of sickness within $t - 3$ to $t + 11$. We hypothesized that two locations influence each other in the spread of dengue if there were co-occurring infections within the 14 days period of sickness. As shown in Fig. 3 (b), there were infections in all localities. This was due to the infections on previous days which were within one period of sickness. In this case, there were infections in all localities on day $t - 2$. As displayed in graph G_t and G_{t+1} , all localities were linked. Meanwhile, on day $t + 2$ infections in localities L1 and L6 were recovered, and there were no new infections. Thus, graph G_{t+2} displays no links from localities L1 and L6 to others. The model concluded that localities with no infections within 14 days did not affect the dengue spread.

3.4.3. DSDN based on co-occurring infections during period of sickness within a neighborhood group

Neighbor Network was generated based on the same method as 14-days Network, which included all co-occurring infections within the period of sickness. This model also included neighborhood group characteristics, as defined in Fig. 2. We hypothesized that two locations in one neighborhood group influence each other in the spread of dengue if there

are co-occurring infections within the period of sickness. The links in Neighbor Network as shown in Fig. 3 (c) were basically the same with the links in 14-days Network. However, in this model links only existed between localities in the same neighborhood group. Graphs G_t and G_{t+1} display the same links, as there were infections in all localities within one neighborhood group. On the other hand, graph G_{t+2} shows no links from/to localities L1 and L6, as all infections in both localities were recovered.

3.4.4. DSDN dengue spread network based on co-occurring infections during period of sickness within a centroid-based attributed cluster

Centroid-based Network was created using the same period of sickness as 14-days Network and Neighbor Network, with different localities grouping. It was generated based on our hypothesis that dengue spreads between localities with similar environmental conditions. We used K-Means clustering algorithm to distribute the nodes into several clusters on each day. Based on feature scoring results as displayed in Table 1, features F1, F4, F5, F7, and F10 were specified as the clustering attributes.

Simulations were carried out with number of centroids ranging from 2 to 39, which equals to the number of localities. It resulted in nodes clustering which consisted of 2 to 18 clusters per day, with an average Silhouette score of 0.607. The clustering process was applied to the dataset on each day. Hence, the cluster of each node, as well as the network structure, might change every day according to the attributes value.

Fig. 3 (d) shows that two localities with co-occurring infections were not always linked, as the link only appeared when the two localities were in the same cluster. In graph G_t , localities L5, L6, and L7 were connected as its values for features F4, F5, and F7 were exactly the same. Meanwhile, localities L1, L2, and L4 were connected as its values for features F4, F5, F7, and F10 were similar. In addition to that, the values for feature F1 were all under 10, which showed that there was only a small number of dengue incidences occurred around the localities. On the other hand, locality L3 fell into a separate cluster as it had the greatest value difference compared to other localities, which indicated a different environmental condition.

In graph G_{t+1} , localities L1 and L2 were connected as its values for features F1, F4, F5, and F7 were similar. Meanwhile, localities L4, L5, and L6 were also connected due to the features value similarity. However, localities L3 and L7 fell into two separate clusters, as its values for the five

features were not similar. The values for feature F1 were much higher than other localities, which also indicated a different environmental condition in the two localities.

It can be concluded from the Centroid-based Network model that the most significant value influencing the cluster formation was feature F1, as it had a wider range compared to the other features. Thus, localities with a high number of dengue incidences in its neighborhood were put into separate clusters. This also applied to the links in graph G_{t+2} , where although the values of features F4, F5, F7, and F10 for all localities were similar, localities L3 and L7 had much higher F1 values compared to localities L2, L4, and L5.

3.4.5. DSDN based on co-occurring infections during period of sickness within a density-based attributed cluster

Since the greater number of localities in one area indicates a denser population, it causes the distance between the localities to be closer. This Density-based Network incorporated the coordinates of each locality by applying DBSCAN clustering algorithm [29]. All nodes were put into clusters based on the distance between localities and the value of its attributes on each day. Simulations were carried out to find the most suitable epsilon value based on the best Silhouette score. It resulted in nodes clustering with epsilon value ranging from 1.5 to 2.8, number of clusters between 2 to 9, with the average Silhouette score of 0.615. The minimum sample / point value was set to 1 so that all data points could be assigned to the clusters, and nothing was classified as noise [30]. The cluster composition might also vary each day, which depends on attributes value of the particular localities.

Fig. 3 (e) shows that in graphs G_t, G_{t+1}, G_{t+2} , localities L3 and L4 were never linked to other localities. Based on the data, localities L3 and L4 had different values for features F4, F5, and F7, while localities L1, L2, L5, L6, and L7 had exactly the same values for the three features. Compared to Centroid-based Network model, where feature F1 significantly affected cluster formation, in Density-based Network, the values of all five features affected the cluster formation. This was concluded based on the data, where localities were included into one cluster if the values for features F4, F5, and F7 were exactly the same, and features F1 and F10 were similar. When there was a big difference in the values for the two features, the respective localities were put into separate clusters.

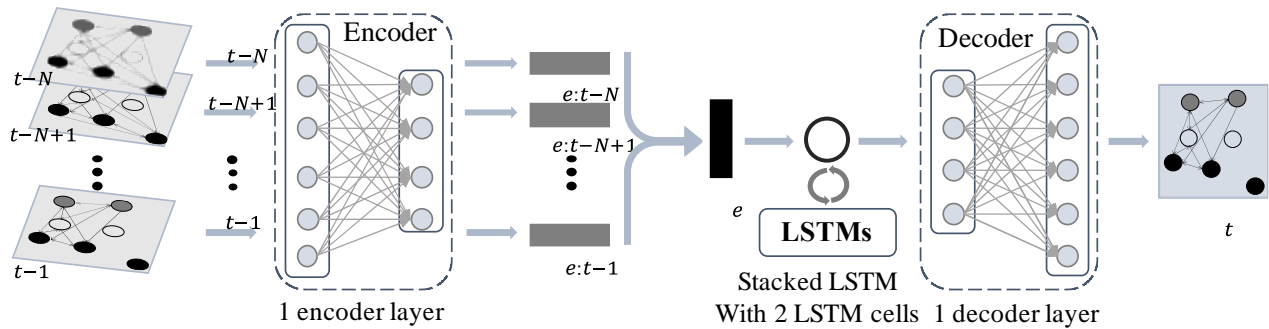


Figure. 4 E-LSTM-D framework used for predicting dengue spread

3.5 Predicting dengue spread

Fig. 4 shows the architecture of E-LSTM-D framework used for predicting dengue spread in this research. E-LSTM-D framework consists of encoder-decoder architecture and stacked LSTM. Encoder layer was placed at the entrance of the model to learn the network structure, and represented the network as high-dimensional data into a lower dimensional vector space. Whereas decoder layer acted as a graph reconstructor at the end of the model to transform the latent features back into a matrix form. Between the encoder and decoder layers, the stacked LSTM layer, which consisted of multiple LSTM cells was placed to learn the pattern of the network evolution.

Using E-LSTM-D, the evolution of a sequence of graphs $\{G_1, \dots, G_T\}$ was learned to predict the future links that may appear or disappear. A sequence of length N was used to add more information hence resulting in a more precise inference. As the input, S was a sequence of graphs with sequence length N which consists of graphs $\{G_{t-N}, \dots, G_{t-1}\}$. It was first received by the encoder layer that was placed at the entrance of the model. The encoder layer processed each term in an input sequence separately. Then, by element-wise adding, it concatenated all the activations using $ReLU(x) = \max(0, x)$ as the activation function, to generate output Y_e .

Output Y_e from encoder layer was then fed into the stacked LSTM layer, which consisted of two LSTM cells. This layer generated H as the output representing the features of target snapshot. As for the decoder layer, it had the mirrored structure of the encoder. However, unlike the encoder, the output layer of the decoder used activation function $sigmoid(x) = [1/(1 + e^{-x})]$. Because in this research the number of layer in the decoder was 1, the activation function used in the decoder layer is only $sigmoid$. The decoder layer received feature H to be processed and reconstructed into a form of predicted graph G_t . To be able to produce a predicted graph

with a structure that fit the input graph, the decoder output layer had the same number of units as the number of nodes.

As for the parameters of E-LSTM-D, we set the encoder layer as 1 layer with 128 units, and the number of LSTM cells as 2, each with 256 units. Meanwhile, the number of units in the decoder layer was 39, which equals to the number of nodes in the network. We evaluated the performance of each model using different length of historical snapshots, which are 3, 11, and 14. This was to represent the number of days in period of sickness from $t - 3$ to $t + 11$, which lasted for 14 days. For example, if 3 was used as the number of historical snapshot, three graphs on the previous periods $\{G_{t-3}, \dots, G_{t-1}\}$ were used as input to predict graph G_t . By distinguishing the length, the ability of E-LSTM-D in learning the model was analyzed, whether this method was able to study a lot of information from longer historical snapshot length, or had better performance at model with a shorter length.

4. Results and Discussion

There are three experiments conducted in this study. In the first experiment, we applied link prediction methods to the generated networks. We evaluated the performance of E-LSTM-D against node2Vec [16] and CTDNE [17] using two different evaluation metrics. The first one was area under Receiver Operation Characteristics curve (AUC), which was the mostly adopted metric to calculate the link prediction accuracy [31]. AUC shows the plotting result of True Positive Rate (TPR) against False Positive Rate (FPR), which indicated a near perfect prediction if the score approached 1.0. Other metric was error rate, which compared the number of links that were falsely predicted, to the total number of existing links. This was an addition to AUC, to make the performance evaluation became more comprehensive [15].

Table 3. Performances of E-LSTM-D link prediction against node2Vec and CTDNE applied to five generated dengue spread dynamic networks

Network	AUC		
	Node2Vec	CTDNE	E-LSTM-D
Same-day	.692	.691	.838
14-days	.688	.689	.859
Neighbor	.691	.679	.926
Centroid-based	.676	.680	.662
Density-based	.700	.679	.837

Table 4. Performances of E-LSTM-D link prediction on five generated dengue spread dynamic networks with 3, 11, and 14 days as the length of historical snapshots

Network	AUC			Error Rate		
	3	11	14	3	11	14
Same-day	.838	.751	.738	.478	.627	.725
14-days	.859	.821	.819	.271	.356	.369
Neighbor	.926	.909	.904	.267	.322	.363
Centroid-based	.662	.635	.625	1.35	1.20	1.23
Density-based	.837	.794	.809	.377	.428	.435

Table 5. Performances of E-LSTM-D link prediction on non-outbreak and outbreak period

Network	AUC						Error Rate					
	3		11		14		3		11		14	
	No-Out	Out	No-Out	Out	No-Out	Out	No-Out	Out	No-Out	Out	No-Out	Out
Same-day	.825	.864	.753	.766	.711	.802	.869	.214	1.15	.290	1.19	.290
14-days	.866	.958	.812	.920	.776	.910	.507	.074	.710	.110	.777	.138
Neighbor	.885	.959	.849	.944	.844	.940	.513	.064	.707	.100	.717	.117
Centroid-based	.838	.825	.806	.820	.788	.795	.704	.807	.775	.683	.829	.754
Density-based	.847	.919	.805	.868	.782	.835	.610	.136	.756	.230	.818	.273

The hyperparameters used for node2Vec and CTDNE were the followings: embedding size = 128, number of walks per node = 10, walk length = 80, and context window = 2. Prior to applying both methods, the network models were transformed into undirected network. All the graphs within the network period length (838 days for Same-day Network and 824 for the others) were used as prediction input, with the first 80% of the data being the training set, while the remainder defined as the test set.

As for E-LSTM-D, historical snapshots length was required to be determined. For the first experiment, we used 3 as the historical snapshot length. We only used 810 snapshots of the dataset, which was subtracted by 14 from a total of 824 samples. This was to accommodate 14 as the historical snapshots length required in the second experiment, so that all link prediction models had the same training and test set. We divided the first 648 snapshots as the training set, and the rest 162 samples (20% of the dataset) as the test set.

Table 3 shows the performance of all link prediction methods measured using AUC, where E-LSTM-D outperformed other two methods for all network models except Centroid-based Network. Compared to the other two methods, E-LSTM-D made use of historical snapshots length to predict the links in the next immediate period. This indicated that sequential data from previous periods were able to

supervise the method to achieve higher prediction accuracy. This also suggested that it was necessary to learn the historical relationship between nodes to produce better predictions. Meanwhile, Centroid-based Network had widely varied links between nodes in each period, which were more difficult to be predicted using methods that relied heavily on historical data patterns such as E-LSTM-D.

For the second experiment, we compared the performance of E-LSTM-D using 3, 11, and 14 as the historical snapshot length as shown in Table 4, where the link prediction applied to Neighbor Network model achieved the highest scores among all models for all evaluation metrics and historical snapshot lengths. Neighbor Network was a model in which the links between nodes were more fixed compared to the other models, because the links only existed between nodes in one neighborhood group. For example, as localities 13, 19, and 26 were in one neighborhood groups, locality 13 was never connected to localities other than 19 and 26. E-LSTM-D was able to learn this kind of characteristic in the link prediction. However, this was not the case for the other models, because the links in other models did not depend on fixed group as in the Neighbor Network, so the other networks were more sparse. This was shown from the prediction result for Centroid-based Network, where it had low AUC score and high Error Rate scores, due to its varying daily cluster for each locality.

Our third experiment was to apply E-LSTM-D link prediction to the network models in non-outbreak and outbreak periods with results as shown Table 5. Based on the dengue fever incidence data trend, we divided the dataset into two periods, non-outbreak (January 2017 – August 2018) and outbreak (from September 2018 onwards). For non-outbreak period dataset, we divided the first 466 snapshots as the training set, and the rest 116 samples as the test set. Meanwhile, for outbreak period dataset, the first 160 snapshots were included in the training set, and the rest 40 samples were in the test set. In this experiment, the link prediction applied to Neighbor Network also had the best performance for all metrics in non-outbreak/outbreak periods.

Compared to the use of all data in the link prediction model, using the data only in outbreak period could improve the prediction performance.

For the evaluation of historical snapshots length variation, the previous study concluded that longer historical snapshots were able to increase the model’s performance [15]. In contrast to that, the performance of the link prediction models in this research tend to decrease as the number of snapshots increased. This was due to how the network was generated based on calculating the presence of cumulative dengue infections in one period of sickness (from $t - 3$ to $t + 11$), so that the network dynamics tend not to change much within a small period of time. This meant that the snapshots from a closer period actually had a bigger influence on the current snapshot than from a further period, thus resulting in a better link prediction performance.

To support the above evaluation, Fig. 5 shows the visualization of the link prediction using 3 as historical snapshot length represented by graphs G_{t-3} ,

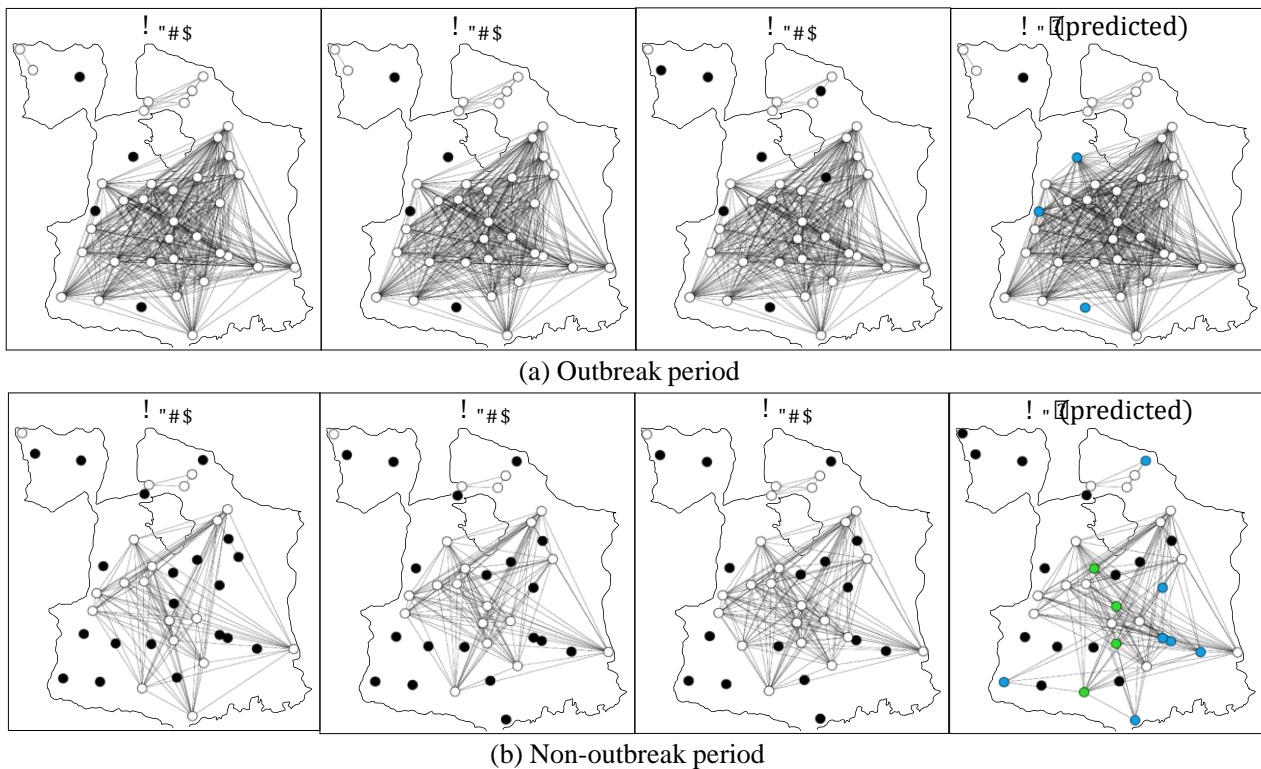


Figure. 5 Visualization of the prediction results using Neighbor Network model with 3 as the number of historical snapshots (G_{t-3} to G_{t-1} as input)

This applied to all models, for all evaluation metrics and historical snapshot lengths. On the other hand, using only non-outbreak period data generally resulted in a lower prediction performance. However, this did not happen to link prediction with Centroid-based Network, where the performance using only non-outbreak data was better than using all data, with higher AUC and lower Error Rate scores. This was mainly influenced by how the training set and the test set were determined.

G_{t-2} , and G_{t-1} to predict G_t . We used data on day-813, day-814, day-815 as the input for the outbreak period, and data on day-585, day-586, day-587 for the non-outbreak period. Nodes in black represent localities with no dengue infections, while nodes in white describe localities where infections existed. The link between white nodes is a two-way directed link which illustrates that these localities influenced each other in the spread of dengue. It can be seen that the relationship between nodes in the outbreak period

is denser, which was also indicated by the presence of more white nodes compared to graphs in the non-outbreak period. This shows that dengue was rapidly spreading during the outbreak period.

In Fig. 5(a), the network structure of the first (G_{t-3}) and second (G_{t-2}) day are the same, and changes only occur on the third day (G_{t-1}). In G_t , it was predicted that the link that existed in G_{t-3} and G_{t-2} would appear on G_t , even though that particular link was missing on G_{t-1} . Thus, the number of white nodes in G_{t-3} and G_{t-2} was the same as the number in G_t . This also occurred to the interconnected nodes in the three input graphs, which were predicted to have the same links in G_t .

Meanwhile, some of the black nodes in the input graphs were predicted to have only one-way link, which were visualized as blue nodes. This means that localities denoted by blue node had the potential to be infected with dengue coming from other localities. The three blue nodes represent localities L3, L16, and L36. Of the three nodes, localities L16 and L36 had links from all white nodes in the neighborhood group. It was predicted that all localities that had dengue infection would influence the spread to localities L16 and L36. Another blue node representing locality L3 was predicted to only have one link from locality L1. Based on dengue fever incidence data, locality L3 was the locality with the lowest incidence frequency (108 days) compared to other locality L16 (120 days) and L36 (154 days). L1 also had the least number of concurring incidences with L3 (52 days), but with the highest number of existing links in G_t (3 times) when there was no relationship between G_{t-3} to G_{t-1} .

In Fig. 5 (b), changes in the network structure occurred in each period, where there were several links that were added and removed in G_{t-2} and G_{t-1} . In G_t , locality L13 that was visualized as white node in all three input graphs was predicted to have recovered from the infections, thus illustrated as black node. Based on the dengue fever incidence data, the frequency of dengue incidence in locality L13 was much lower (104 days) compared to the frequency of non-incidences (373 days).

It was also predicted that there were blue nodes represent localities L5, L7, L17, L24, L28, L33, and L37. From those seven, L24 was the only locality which had dengue incidences in G_{t-2} , and L28 had dengue infections in G_{t-1} . Meanwhile, the other five nodes had no incidences in graphs G_{t-3} , G_{t-2} , and G_{t-1} . However, compared to other black nodes within the same neighborhood group, there were more historical dengue incidences occurred in localities L5 (307 days), L7 (330 days), L17 (374 days), L33 (287 days), and L37 (336 days).

In addition, there were also four nodes that were visualized in green, which were localities L3, L15, L21, and L23. In the input graphs, those four nodes were visualized as white nodes with two-way links from and to other nodes. However, it was predicted that those nodes only had one-way links toward others. Based on historical data, there were less dengue incidences occurred in localities L3 (62 days), L15 (104 days), L21 (202 days), and L23 (111 days).

5. Conclusion and Future Works

In this research, dengue spread was modelled into 5 types of network based on the number of co-occurring infections in localities, as well as neighborhood group and cluster boundaries. The prediction of dengue fever spread using DNLP approach with E-LSTM-D resulted in the best accuracy when applied to the Neighbor Network which modelled that dengue only spread between localities within the same neighborhood group. The accuracy increased when the prediction was applied only to the outbreak period data, where the AUC score increased by 0.0339 while the Error Rate decreased by 0.202. This suggested that E-LSTM-D performance was improved when applied to network with more inter-node links, which indicated rapidly spreading infections.

The prediction of dengue spread also included the result that localities did not always have a two-way relationship with each other. There were localities that did not have current infections, but could potentially be affected by the spread from others, when historically there had been many dengue incidences in that localities. On the other hand, there were also localities that had incidences, but were not affected by the spread from others, when there were less frequent dengue incidences in that localities. For future works, other method could be applied to incorporate factors influencing dengue spread as the attributes of the network, so that the spread model could be generated based on the defined parameter of each factor over time.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Arfinda Setiyoutami: conceptualization, methodology, formal analysis, writing—original draft preparation and editing. Diana Purwitasari: validation, formal analysis, writing—review and editing. Wiwik Anggraeni: data curation, validation, formal analysis, writing—review. Eko Mulyanto

Yuniarno: supervision, conceptualization, formal analysis, writing—review. Mauridhi Hery Purnomo: supervision, conceptualization, formal analysis, writing—review. All authors read and approved the final manuscript.

Acknowledgments

This work was supported by Indonesia Endowment Fund for Education (LPDP) from Ministry of Finance under Indonesian Education Scholarship for Master Program 2018 with registry number 201812110113706.

References

- [1] Guzman and R. E. Istúriz, “Update on the global spread of dengue”, *International Journal of Antimicrobial Agents*, Vol. 36, pp. S40–S42, 2009.
- [2] A. T. Bäck, and Å. Lundkvist, “Dengue viruses – an overview,” *Infection Ecology & Epidemiology*, Vol. 3, No. 1, pp. 19839, 2013.
- [3] J. P. Messina, O. J. Brady, N. Golding, M. U. G. Kraemer, G. R. W. Wint, S. E. Ray, D. M. Pigott, F. M. Shearer, K. Johnson, L. Earl, L. B. Marczak, S. Shirude, N. D. Weaver, M. Gilbert, R. Velayudhan, P. Jones, T. Jaenisch, T. W. Scott, R. C. Reiner Jr, and S.I. Hay, “The current and future global distribution and population at risk of dengue”, *Nature Microbiology*, Vol. 4, No. 9, pp. 1508–1515, 2019.
- [4] J. D. Stanaway, D. S. Shepard, E. A. Undurraga, Y. A. Halasa, L. E. Coffeng, O. J. Brady, S. I. Hay, N. Bedi, I. M. Bensenor, C. A. Castañeda-Orjuela, T. Chuang, K. B. Gibney, Z. A. Memish, A. Rafay, K. N. Ukwaja, N. Yonemoto, and C. J. L. Murray, “The global burden of dengue: an analysis from the Global Burden of Disease Study 2013”, *The Lancet Infectious Diseases*, Vol. 16, No. 6, pp. 712–723, 2016.
- [5] 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 Systems*, Vol. 12, No. 2, pp. 246–259, 2019.
- [6] S. T. Stoddard, B. M. Forshey, A. C. Morrison, V. A. Paz-Soldan, G. M. Vazquez-Prokopec, H. Astete, R. C. Reiner Jr, S. Vilcarromero, J. P. Elder, E. S. Halsey, T. J. Kochel, U. Kitron, and T. W. Scott, “House-to-house human movement drives dengue virus transmission”, *Proceedings of the National Academy of Sciences U. S. A.*, Vol. 110, No. 3, pp. 994–999, 2013.
- [7] P. Holme, “Temporal network structures controlling disease spreading”, *Physical Review E*, Vol. 94, No. 2, p. 022305, 2016.
- [8] Z. Zhang, H. Wang, C. Wang, and H. Fang, “Modeling Epidemics Spreading on Social Contact Networks”, *IEEE Transactions on Emerging Topics in Computing*, Vol. 3, No. 3, pp. 410–419, 2015.
- [9] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Régis, J. Pinton, N. Khanafer, W. Van den Broeck, and P. Vanhems, “Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees”, *BMC Medicine*, Vol. 9, No. 1, pp. 87, 2011.
- [10] L. Danon, A.P. Ford, T. House, C. P. Jewell, M. J. Keeling, G. O. Roberts, J. V. Ross, and M. C. Vernon, “Networks and the epidemiology of infectious disease”, *Interdisciplinary Perspectives on Infectious Diseases*, Vol. 2011, pp. 284909, 2011.
- [11] M. L. V. Araújo, J. G. V. Miranda, R. Sampaio, M. A. Moret, R. S. Rosário, and H. Saba, “Nonlocal dispersal of dengue in the state of Bahia”, *Science of the Total Environment*, Vol. 631–632, pp. 40–46, 2018.
- [12] R. M. Lana, M. F. da C. Gomes, T. F. M. de Lima, N. A. Honório, and C. T. Codeço, “The introduction of dengue follows transportation infrastructure changes in the state of Acre, Brazil: A network-based analysis”, *PLoS Neglected Tropical Diseases*, Vol. 11, No. 11, 2017.
- [13] H. A. M. Malik, A. W. Mahesar, F. Abid, A. Waqas, and M. R. Wahiddin, “Two-mode network modeling and analysis of dengue epidemic behavior in Gombak, Malaysia”, *Applied Mathematical Modelling*, Vol. 43, pp. 207–220, 2017.
- [14] L. M. Stolerman, D. Coombs, and S. Boatto, “Sir-network model and its application to dengue fever”, *SIAM Journal on Applied Mathematics*, Vol. 75, No. 6, pp. 2581–2609, 2015.
- [15] J. Chen, J. Zhang, X. Xu, C. Fu, D. Zhang, and Q. Zhang, “E-LSTM-D: A Deep Learning Framework for Dynamic Network Link Prediction”, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–14, 2019.
- [16] N. M. Ahmed, and L. Chen, “An efficient algorithm for link prediction in temporal uncertain social networks”, *Information Sciences*, Vol. 331, pp. 120–136, 2016.
- [17] A. Grover and J. Leskovec, “node2vec: Scalable Feature Learning for Networks”, In: *Proc. of the*

- 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 2016.
- [18] G. H. Nguyen, J. B. Lee, R. A. Rossi, N. K. Ahmed, E. Koh, and S. Kim, “Continuous-Time Dynamic Network Embeddings”, In: *Proc. of The Web Conference 2018*, Lyon, France, pp. 969–976, 2018.
- [19] K. Chi, G. Yin, Y. Dong, and H. Dong, “Link prediction in dynamic networks based on the attraction force between nodes”, *Knowledge-Based Systems*, Vol. 181, pp. 104792, 2019.
- [20] T. Li, J. Zhang, S. Y. Philip, Y. Zhang, and Y. Yan, “Deep dynamic network embedding for link prediction”, *IEEE Access*, Vol. 6, pp. 29219–29230, 2018.
- [21] F. D. A. Suryanegara, S. Suparmi, and N. Setyaningrum, “The Description of Larva Free Index as COMBI (Communication for Behavioral Impact) Dengue Hemorrhagic Fever Prevention Indicator,” *Jurnal Fakultas Kesehatan Masyarakat (The Indonesian Journal of Public Health)*, Vol. 13, No. 3, pp. 338–344, 2018.
- [22] W. Anggraeni, G. Pramudita, E. Riksakomara, R. P. Wibowo, F. Samopa, Pujiadi, and R. S. Dewi, “Artificial Neural Network for Health Data Forecasting, Case Study: Number of Dengue Hemorrhagic Fever Cases in Malang Regency, Indonesia”, In: *Proc. of International Conference on Electrical Engineering and Computer Science (ICECOS)*, Pangkal Pinang, Indonesia, pp. 207–212, 2018.
- [23] D. Liu et al., “A dengue fever predicting model based on Baidu search index data and climate data in South China”, *PLoS One*, Vol. 14, No. 12, pp. e0226841, 2019.
- [24] J. M. Heilman, J. De Wolff, G. M. Beards, and B. J. Basden, “Dengue fever: a Wikipedia clinical review”, *Open Medicine*, Vol. 8, No. 4, pp. e105-15, 2014.
- [25] V. H. Quej, J. Almorox, J. A. Arnaldo, and L. Saito, “ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment”, *Journal of Atmospheric and Solar-Terrestrial Physics*, Vol. 155, pp. 62–70, 2017.
- [26] I. Azizan, S. A. B. A. Karim, and S. Suresh Kumar Raju, “Fitting Rainfall Data by Using Cubic Spline Interpolation”, *MATEC Web of Conferences*, Vol. 225, 2018.
- [27] S. K. Adhikary, N. Muttil, and A. G. Yilmaz, “Genetic programming-based ordinary kriging for spatial interpolation of rainfall”, *Journal of Hydrologic Engineering*, Vol. 21, No. 2, pp. 4015062, 2016.
- [28] A. Setiyoutami, W. Anggraeni, D. Purwitasari, E. M. Yuniarno, and M. H. Purnomo, “Extracting Temporal based Spatial Features in Imbalanced Data for Predicting Dengue Virus Transmission”, In: *Proc. of Advances in Computer Communication and Computational Sciences: IC4S 2019*, Bangkok, Thailand, 2019.
- [29] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, In: *Proc. of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, 1996.
- [30] G. Boeing, “Clustering to reduce spatial data set size”, arXiv preprint arXiv:1803.08101, 2018.
- [31] B. Chen, Y. Hua, Y. Yuan, and Y. Jin, “Link Prediction on Directed Networks Based on AUC Optimization”, *IEEE Access*, Vol. 6, pp. 28122–28136, 2018.