Received: January 14, 2022. Revised: February 24, 2022.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Fuzzy C-Means and Social Network Analysis Combination for Better Understanding the Patient-based Spread of Dengue Fever with Climate and Geographic Factors

Wiwik Anggraeni[1,2,5] 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: Climate and geography factors significantly influence the spread of dengue fever. It's critical to figure out the specifics of climatic and geographic conditions and the relationships between current patients. For further preventive measures, it is also necessary to identify the transmission source patients. This study aims to improve the understanding of dengue fever patients spreading under climate and geographic locations. Patients are clustered based on climatic and geographical variables, and influential patients are found in the established network using Fuzzy C-Means and Social Network Analysis. The scenario of cluster numbers' alteration and degree of fuzziness with Fuzzy C-Means made the three groups of patient clusters. A total of 52% of the patients are included in the lowland cluster, based on climate conditions and altitude. The patients grouped better with this approach than with the other compared methods. The Calinski Harabasz score has an average difference of 1644.105. The following relationship in the network is constructed once the cluster has been formed. It is given a fuzzy rule representing the distance of residence and the period of illness between patients. According to the Social Network Analysis approach applied to the region and month scenario, the three areas sensitive to the spreading center are Dau, Kepanjen, and Karangploso. The most significant patient data distribution occurred during the rainy season, peaking in January-February. The centrality algorithm shows that patients with male characteristics and the age range of children and adults could be the source of the disease's spread every month. Kepanjen area is the site of residence with the most significant impact with a proportion of 62.5 % in a year, and the initial illness is the fever related to dengue. Analysis results of this study can use by the Public Health Office to plan and manage resources to prevent extensive spread.

Keywords: Fuzzy C-means, Social network analysis, Dengue fever spreading, Patient, Climate, Geographic.

1. Introduction

Dengue Fever (DF) is a highly contagious mosquito-borne disease [1]. Every year, between 50 and 100 million DF cases are reported across 100 countries, resulting in 24,000 deaths. According to WHO data, Indonesia has the most significant DF cases number in Asia [2]. For the past 47 years, DF has been a major public health issue in Indonesia. This condition is because DF has resulted in a high rate of mortality [3] and a significant increase in the cases number, particularly in 2019 [4].

Many factors, including climatic and geographical conditions, contribute to the proliferation of DF cases. The role of climate in disease propagation is more closely linked to diseases spread by animals, such as DF disease. In the case of DF, a combination of climatic variables was used to see the spread of cases [5, 6], the development of Aedes mosquitoes [7, 8], DF transmission [9], modeling the number of cases [10, 11], effect analysis

[12], predictions of outbreaks [13-16], and dengue fever incidence rates [17], as well as predictions of spread associated with social media data [18]. To gain more detailed information, most studies combine various variations of climate data with other variables.

Other factors, such as geographical conditions, are thought to influence the spread of DF (in addition to climate). There is relatively little research that investigates the impact of altitude. Several prior studies [7, 8, 12, 19, 20] suggested that regional conditions could be a significant factor. According to studies conducted by [21], different terrain conditions exhibited varied patterns of case numbers.

Prior investigations have determined the level of influence of the variables involved. However, until the conditions under which the spread of these cases can occur, the data acquired are not specified. Climate variables, sets of conditions, and the climate at the time of the DF incident, for example, have not been identified in detail. Furthermore, the previous study has not covered the relationship between patients in each group and the criteria for which individuals are susceptible to spreading the disease.

A clustering strategy may be used to categorize groups of meteorological and geographical factors at the time of illness occurrence. K-means [22], Hierarchal Clustering [23], and Fuzzy C-Means (FCM) are some of the most widely recommended algorithms in the health industry. FCM, on the other hand, is still infrequently employed in the context of clustering, much less in relation to DF. FCM is used mostly for classification and segmentation [24, 25].

Social network analysis (SNA) technique can be used to detect relationships between objects in a cluster and identify the effects of objects. SNA can study the structure of social ties inside a group using the graph-based work idea [26], revealing informal relationships between individuals. SNA is commonly used in the health industry to define public health communication [27], human-to-human infection analysis [28], disaster-prone area investigation [29], and disease distribution analysis [30].

The majority of these studies have little to do with DF illness. Furthermore, they have not thoroughly examined the impact of climatic and geographical variables on the distribution of DF cases, particularly the distribution of patients. Previous research hasn't specified which patients are likely to be susceptible to DF, impacting the disease's spread. Identifying the patient as the epicenter of the disease's dissemination and the patient's characteristics is critical. The detection of these patients is required to end the outbreak by breaking the chain of transmission, reducing the number of dengue fever sufferers and the region impacted to a minimum [31], as well as the patient's characteristics. The finding of this study has the potential to hasten the government's national healthcare system goals [32]. Furthermore, determining the meteorological and geographical conditions of the patient cluster is critical. Both conditions are important to investigate because rising disease epidemics are linked to climatic change include temperature, humidity, rainfall, and wind speed [33].

This research contributes to a framework that will aid in analyzing DF patient distribution based on climatic conditions and altitude. The novel framework proposed in this research is a different approach from previously. This methodology combines fuzzy C-means, which clusters patients depending on climatic and geographical variables, with social network analysis, which finds qualities and centrality of the formed relationships. The results of this attribute measure and centrality are then utilized to dig deeper into the locations and times where a substantial amount of spread is possible. In addition, knowing how patients interact and identify patients who may be super-spreaders in a cluster. Our approach will provide an accurate and rapid understanding of dengue fever in a particular region, which can be the size of one or more other regions. The approach can find areas that require special attention for medical intervention. Our approach provides a range of more details about the distribution of dengue infection, than previous studies.

In detail, this research contributes in a novel method to provide several parts, including:

- DF patients clustering, which based on climatic conditions, including temperature, rainfall, humidity, wind speed, and geographical conditions represented by the area's altitude. These results display the climatic and geographical conditions where the patients are more spread out.
- Relationships between patients by considering the period of illness and their location as the weight of the relationship. Fuzzy rules are used to determine the weight.
- Identification of the vulnerable areas to being the center of the spread of DF.
- Identification of the months that are vulnerable to a high DF spread rate.
- Identifying susceptible patients is central to the spread of DF and their characteristics.
- Visualization of the relationship between patients in the form of graphs.

The rest of this paper is organized as follows. Section 2 describes the previous related studies. In section 3, the research areas, data sources, and methods used is described extensively. Section 4 presents the results and discussion. Section 5 presented conclusions, and directions for future work.

2. Related works

2.1 Approaches to describe objects' relation

Social network analysis (SNA) is an approach for studying human relationships that employs graph theory. SNA is frequently used to examine the relationships between things in various fields. In comparison to other disciplines, SNA is still not commonly used in the health sector. Furthermore, SNA is still exceptionally rarely used for disease transmission.

However, SNA has been used in the medical field to diagnose infectious diseases. A literature survey revealed a link between disease dissemination and SNA in a study undertaken by [27]. The aspects of SNA that make it suited for usage in public health and epidemiological applications are examined in this article. The findings reveal three primary characteristics or link patterns significant to public health applications. The three essential characteristics are degree centrality, eigenvector centrality, and betweenness centrality.

The study [28] used a social media data categorization approach to acquire information from users when analyzing human-to-human infection. The naive bayesian classification is used to analyze the disease's spread. Then, according to [30], SNA can depict the phenomenon of illness spreading into a tissue. There are two network models in the network. The first network is analyzed from both a global and local perspective, with proximity, betweenness, and short path-length measurements. According to this study [30], several locations have been contaminated with the disease in recent years.

SNA has also been used to describe the relationship between people. This study uses degree centrality, eigenvector centrality, and betweenness centrality. This centrality algorithm is better than other algorithms [34]. In addition, SNA has also been used for disaster data classification using Twitter data. The classification results show information on the most frequent disasters and their areas [28].

These studies show the advantages of SNA, but they have not explored SNA in detail until identifying the actors that can be the source of the spread of the disease and their characteristics. In addition, no previous studies link in detail the climatic conditions and altitude of the area with the distribution of actors.

2.2 Approaches in object grouping

Clustering is one of the most used ways of grouping items. K-means, fuzzy C-means (FCM), and Hierarchal clustering algorithms are extensively employed in the health industry [35]. In classification, K-means is more often used [22]. In a network system, K-means is sometimes integrated with other methods for classification, such as random forest [36]. Meanwhile, market clusters are frequently utilized with the Hierarchical model [23]. FCM is rarely utilized for clustering, although it is frequently used for classification [24] and segmentation [25]. FCM is more commonly utilized for segmenting fraud detection [24] and brain tumors [35].

Because of its capacity to deal with overlapping cluster boundaries, FCM is utilized [24, 37]. FCM, on the other hand, is still infrequently employed in the context of disease transmission, particularly in the case of dengue fever. In some places of India, SOM has been used to categorize dengue fever precisely [38]. Areas with low, medium, and high DF status are the results of this study. On the other hand, the clustering uses limited variables. Furthermore, there has been no investigation into the area's climatic characteristics and altitude. Similarly, no patient-related information has been made public.

2.3 Climate and geographical as influencing factors of DF spread

There has been a lot of research done on the effect of climate on the spread of disease, particularly DF. Previous research looked at the spread of dengue and chikungunya cases using a combination of minimum-maximum temperature and precipitation, as well as mosquito density [5]. The same variable was also used to look at dengue disease cases and mosquito development [10, 14]. Rainfall factors, IOD, and Nano 34 were introduced [11]. Another study modeled DF incidence in numerous regions using minimum-maximum temperature and rainfall [15], [16]. Furthermore, rainfall combined with an average temperature and humidity has been utilized to detect DF patients [6]. Another study looked at the effect of climate on the DF incidence by omitting the minimum-maximum temperature and then adding the wind factor [7]. DF outbreaks [8, 9], and DF incidence in numerous places have all been linked to average temperature, humidity, and rainfall.

These findings suggest that mosquitoes can spread to tropical and subtropical regions as a marker of the impact of temperature and precipitation on mosquito presence and development [5]. The DF

virus is then transmitted at low temperatures in the winter [7] and minimum temperatures in the preceding four months [9]. Furthermore, the climate has both a direct and indirect effect on the incidence of dengue fever, which is mediated by mosquito density [8].

Mosquito density and climatic conditions have a significant relationship. The rainfall circumstances of the previous month and two months ago, on the other hand, had no substantial impact on mosquito development and spread. Furthermore, temperature, rainfall, and sunlight significantly impact DF transmission [11]. Mosquito control is particularly required during high temperatures, as mosquito transmission occurs [11].

The spread of dengue fever in Brazil is predicted to be heavily influenced by rainfall during the past two months and the minimum temperature. At an average minimum temperature of 21 °C, the incidence of dengue fever increases [3]. In China [12], where the highest temperature is 21.6–32.9 °C, and the minimum temperature is 11.2–23.7 °C, the temperature significantly impacts dengue fever. This temperature is not the same as the temperature in Brazil or Taiwan.

Several climate conditions influence DF, according to these studies. Unfortunately, no extensive analysis of the temperature, rainfall, or humidity at the time of the increase in dengue fever cases was conducted. Furthermore, the research described above did not consider wind speed or geographical factors, particularly the area's altitude.

3. Materials and method

3.1 Study area

This study uses Malang regency as the case study location. The Malang regency was chosen since it is East Java's second-largest district [39]. Furthermore, Malang regency is the third region in East Java with the largest number of DF cases in 2016, and so on [4].

Many beaches flank the Malang regency with warmer temperatures and several mountains with cooler temperatures. Malang regency is divided into 33 sub-districts, 12 sub-districts, and 378 settlements, covering 3,530.65 km^2 and a population density of 831.33/km^2 [39]. Different geographical characteristics, a hot climate, and a dense population render this area vulnerable to DF.

3.2 Dataset

The data used in this investigation includes information about DF patients and weather information such as air temperature, humidity, rainfall, and wind speed. Furthermore, data on geographic circumstances in the form of an area's elevation is involved. Dengue patient's data, a daily report from January 2018 to December 2019, is obtained from the Malang district health office. The altitude data of the region is the height above sea level for each sub-district. This altitude data is collected from the central bureau of statistics. As for data on air temperature, humidity, rainfall, and wind speed obtained by the meteorology, climatology, and geophysics agency of Karangploso. All data used have the same period from January 2018 to December 2019.

3.3 Method

This study employs a hybrid approach that combines fuzzy C-means (FCM) with social network analysis (SNA). The procedure is generally divided into three stages: data preprocessing, clustering with FCM, and identifying distinct climatic and patient distribution circumstances using SNA. Clustering's performance is compared to other methods such as K-means, hierarchical, mean-shift, and DB-SCAN. Various centrality techniques, such as the proximity centrality algorithm, the betweenness centrality algorithm, and others, depict the magnitude of the characteristics and the value of centrality in the SNA. The study's framework is illustrated in Fig. 1.

3.3.1. Data preprocessing

Data preprocessing was done to guarantee that it is ready to be processed at the following level. The following are the preprocessing steps employed in this study:

- Eliminate the column not used in the clustering analysis procedure from the patient data. This procedure creates a new data frame with four climatic variables, one altitude variable, and one unique identifier (id).
- The column 'No id' has been removed from the next dataframe, resulting in a data frame containing only climate variables used as input for the clustering training procedure.
- We used the min-max normalization method from Eq. (1) to avoid variables dominating the data from the preceding procedure.

Eq. (1) uses several variables of $\grave{v}_i$ (data value after normalization), v_i (data value before normalization), min_A (minimum value in the data

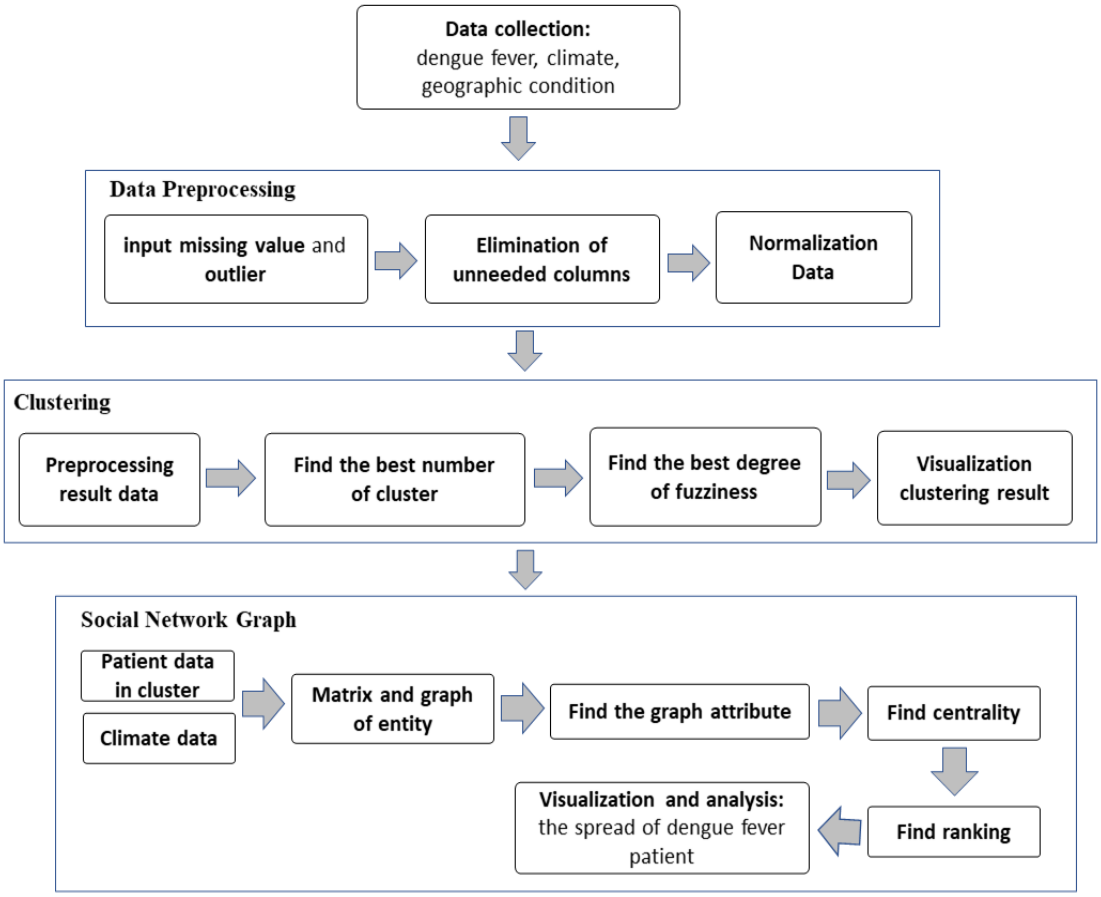

Figure. 1 The framework of the proposed approach

before normalization), max_A (maximum value in the data before normalization), new_min_A (minimum value in the data after normalization), and new_max_A (maximum value in the data after normalization). Furthermore, the results of the data distribution for each climate variable can be seen in Fig. 2.

$$\grave{v_i} = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \quad (1)$$

3.3.2. Patient data clustering

The goal of the clustering approach is to group patients based on altitude and climatic variables. Fuzzy C-means (FCM) are used in this procedure. FCM employs a fuzzy grouping paradigm, which allows data to belong to any classes or clusters produced with varying degrees or membership levels ranging from 0 to 1 [24]. The degree of membership in a cluster determines the level of data presence [37]. Suppose input from the fuzzy system is defined as $U = (u_1, u_2, \ldots, u_n)$ and membership degree from a data point k in cluster i is symbolized as $\mu_{ik}(\mu_k) \in [0,1]$ with $(1 \le i \le n;\ 1 \le k \le c)$, then the partition matrix in FCM is defined as Eq. (2).

$$\mu_{ik} = \begin{bmatrix} \mu_{11}[\mu_1] & \ldots & \mu_{1c}[\mu_1] \\ \vdots & \ldots & \vdots \\ \mu_{ni}[\mu_n] & \ldots & \mu_{nc}[\mu_n] \end{bmatrix} \text{ with } \sum_{k=1}^{c} \mu_{ik} = 1 \quad (2)$$

$$F_w(U,V) = \sum_{i=1}^{n}\sum_{k=1}^{c}((\mu_{ik})^w\, {d_{ik}}^2) \quad (3)$$

Objective function used in FCM is shown in Eq. (3), where $F_w(U,V)$ is the objective function towards U and V, c is the cluster number in X, n is the processed data number, w is weight with $w \in [1,\infty]$, and X is the matrix of processed data with $n \times m$ (n shows samples number and m shows data criteria). U is the initial partition matrix (membership function), V is the center cluster matrix, and $d_{ik} = d(x_i - v_k) = \left[\sum_{j=1}^{m}(x_{ij} - v_{kj})^2\right]$ is the distance of every data in the cluster. To update every value from partition matrix, we used Eq. (4). The center of the k^{th} cluster is denoted by V_{kj}, which is shown in Eq. (5).

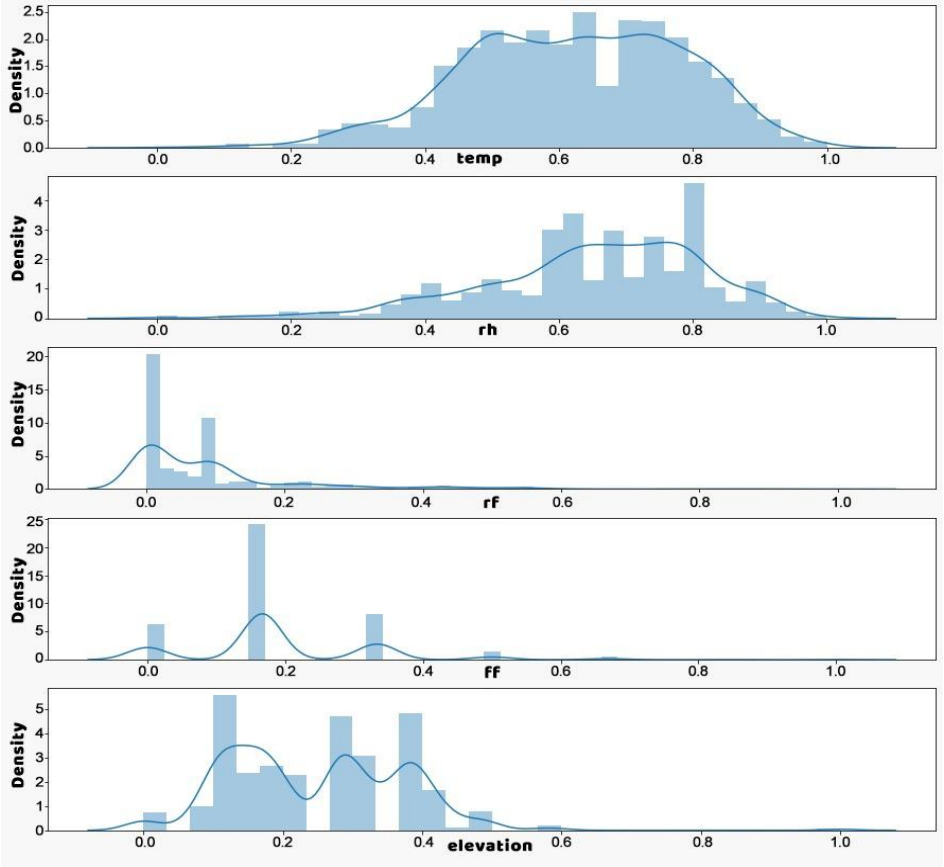

Figure. 2 Distribution of pre-processed climate data

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}(X_{ij}-X_{kj})^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}(X_{ij}-X_{kj})^2\right]^{\frac{-1}{w-1}}} \quad (4)$$

$$V_{kj} = \frac{\sum_{i-1}^{n}((\mu_{ik})^{w} * X_{ij})}{\sum_{i-1}^{n}(\mu_{ik})^{w}} \quad (5)$$

Clustering was done in patient data. Climate data include temperature, humidity, rainfall, wind direction and location's altitude are used to cluster the patient's data. The number of clusters *(c)* and the degree of fuzziness *(d)* are varied in this scenario *(m)*. The degree of fuzziness refers to how difficult it is to tell if an object belongs in cluster A or B. The degree of fuzziness is 0.1-100, and the number of clusters is changed starting at 2. However, the number of clusters in this experiment was 2, 3, and 4, and the degree of fuzziness was 1.5–4.0 with a 0.5 interval. This decision was based on an experiment in which values outside the interval did not affect the center clustering findings. Patients in each cluster, where patients in the same cluster are in nearly identical climatic and altitude settings, are the results of this stage.

3.3.3. Implementation of social network analysis (SNA) approach

There are several processes carried out at the SNA implementation stage, including:

a. Making Matrix and Graph of Relationships between Entities

After the patient, elevation, and climate data are combined, the next step is to determine the relationship between the entities. In this study, nodes represent patients, and edges represent relationships between patients. Edge represents a fuzzy value that represents the location of residence between patients and the time span of illness between one patient and others. The range of illness is ten days based on the incubation period of DF [40].

The membership function of the illness time is using the *S* curve. The illness range membership function is written in Eq. (6). In addition to the duration of illness, this study also involves a membership function that represents the patient's location. The membership function is written in Eq. (7).

$$S(x) = \begin{cases} 0, & if\ x = 0 \\ 2(x/10)^2, & if\ 0 \le x \le 5 \\ 1 - 2((10-x)/10)^2, & if\ 5 \le x \le 10 \\ 1, & if\ x \ge 10 \end{cases} \quad (6)$$

$$W(x) = \begin{cases} 0 & ,if\ it's\ in\ the\ same\ area \\ 0.5 & ,if\ it's\ in\ a\ different\ area\ but\ is\ immediately\ adjacent \\ 1 & ,if\ it's\ in\ a\ different\ area\ and\ not\ directly\ adjacent \end{cases} \quad (7)$$

$$\mu_{S \cap W} = min(S(x), W(x)) \quad (8)$$

Therefore, the illness range and location membership functions are integrated into a single weight using a fuzzy relation, as shown in Eq (8).

b. SNA Attribute Calculation

The graph that has been created is used to calculate the attributes. The graph properties computed include:

- *The number of nodes.* Based on the clustering results, the number of these nodes will change dynamically.
- *The total number of edges.* It is determined by whether or not the relationship between patients is constructed using the fuzzy rules on Eq (8).
- *Average Degree*, which is the average number of node-to-node links. The greater the average degree value, the more likely a patient is to infect a large number of others. Eq. (9) and Eq. (10) are used to calculate average degree, where *k* is the average degree value, *N* is the number of nodes in a network, and *E* is the number of edges.

$$k \equiv \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2E}{N} \quad (9)$$

$$k^{in} = k^{out} = \frac{E}{N} \quad (10)$$

- *Average Path Length.* The path is interpreted as the distance between one node and another. The equation of the average path length is shown in Eq. (11), where *i* is the i^{th} node, *j* is the j^{th} node, *I(i,j)* is the shortest path between *i* and *j* nodes, and *n* is the number of nodes.

$$average\ path\ length = \frac{\sum_{i>j} I(i,j)}{\frac{n(n-1)}{2}} \quad (11)$$

- *Density*, is the density of a graph. Density values range from 0 and 1. If the density value is closer to 1, the more nodes are connected. For a directed network, the maximum number of possible bonds between actors is $k \times (k-1)$. The equation for measuring density is displayed in Eq. (12), where D is the density, L is the number of bonds observed in the network, and k is the number of bonds in the network.

$$D = \frac{L}{k \times (k-1)} \quad (12)$$

- *Network Diameter*, which is the closest path between nodes. The model for measuring network diameter is shown in Eqs. (13) and (14), where d is the diameter and $d(u,v)$ is the distance of *u* and *v*.

$$d = max_{u,v \in V}\ d(u,v) \quad (13)$$

$$N(d_{max}) \approx N \quad (14)$$

- *Modularity*, which is an indication of the formation of groups on the graph. Modularity in a graph is defined as in Eq. (15), where Q is modularity; m is the number of edges; A_{ij} is an element of the adjacency matrix A in rows i and j; k_i and k_j are degrees on i and j; c_i and c_j are the sum of the parts of i or that pass through all vertices of i or j; δ is worth 1 if x is equal to y, and vice versa.

$$Q = \frac{1}{2}m \times \sum\left(\left(\frac{A_{ij} - k_i \times k_j}{2m}\right)\delta(c_i, c_j), i, j\right) \quad (15)$$

- *Number of Community*, namely the many groups formed. The purpose of finding a community is to identify a model based on the topology.

c. Implement Centrality Algorithm to Calculate the Centrality Value

The graph that was created earlier is used to determine centrality. Degree centrality, betweenness centrality, proximity centrality, and eigenvector centrality were utilized as measures of centrality in this study [27, 34]. Each one employs a unique centrality algorithm.

One approach to assess centrality in a social network is betweenness centrality. Eq. (16) shows the value betweenness centrality of each node in the network, where σ_{st} is the shortest distance from *s* to *t*, and $\sigma_{st}(v_i)$ is the shortest distance from *s* to *t* passing through v_i bonds.

$$C_B(v_i) = \sum_{v_i \neq v_s \neq v_t \in V, s<t} \frac{\sigma_{st(v_i)}}{\sigma_{st}} \quad (16)$$

The average distance from the initial node to all other nodes in the network is shown by closeness

```
1.  Graph A = /* read input graph */
2.  Worklist a1 = {v:v ε A.nodes}
3.  foreach n: Node ε a1 {
4.    forall v: Node ε A://Hitung shortest path DAG G
5.      compute σ nv
6.      compute pred(n,v)
7.    forall v: Node ε G://Hitung lintasan mundur DAG
      G
8.      compute δn(v)
9.      BC(v) += δn(v)
10.   forall (u,v): Edge ε G://Reset atribut graf
11.   reset (u,v)
```

Figure. 3 Pseudocode for centrality calculation

Table 1. Attribute ranking criteria

No	Attribute Name	Remark
1	Average degree	The higher, the better
2	Network diameter	The smaller, the better
3	Average path length	The smaller, the better
4	Density	The higher, the better
5	Modularity	The smaller, the better
6	Number of communities	The smaller, the better

centrality. This method can determine how quickly a node can communicate with other nodes. Closeness centrality is a measurement of a patient's proximity to all other patients. Eq. (17) yields the value of each node's proximity centrality, where $g(v_i, v_j)$ is the distance between nodes v_i and v_j, and n is the number of nodes in the network.

$$C_c(v_i) = \frac{n-1}{\sum_{j\neq i}^{n} g(v_i,v_j)} \quad (17)$$

There was also the degree centrality, which reflects how many interactions a node has. Eq. (18) is used to find the degree centrality value of node n_i, where $d(n_i)$ is the number of interactions that node n_i has with other nodes in the network [26].

$$CD(n_i) = d(n_i) \quad (18)$$

Eigenvector centrality is a type of measurement that gives more weight to nodes related to other nodes and has high centrality values. Eqs. (19) and (20) give the eigenvector centrality value of a node, where α is the normalization constant (vector scale), and β represents how much a node has a centrality weight in a node with a high centrality value. A is an adjacency matrix, I is the identity matrix, and l is a matrix. The radius power of a node is β. Furthermore, the pseudocode for the centrality calculation is shown in Fig. 3.

$$C_I(\beta) = \sum(\alpha + \beta_{CJ}) A_{JI} \quad (19)$$

$$C(\beta) = \alpha(I - \beta A) - 1Al \quad (20)$$

d. Rankings

There are two types of rankings: attribute ranking and centrality ranking. The higher the centrality value, the higher the ranking for the centrality ranking criteria. However, this condition is not the same as attribute ranking. Rank qualities according to the criteria are listed in Table 1 [27].

4. Result and discussion

The main goals of this study are finding clusters of DF patients based on climatic and geographical conditions, identifying areas with a high distribution rate, identifying months that are susceptible to high spread, and identifying patient characteristics that have the potential to spread, among other patients.

4.1 The cluster of patient spreading

This study compares performance in clustering using a variety of test scenarios to determine the ideal number of clusters. Several criteria were used in the experiment, including the Silhouette score (SH), Calinski Harabasz score (CH), and Dunn index (DB). Table 2 shows the results of comparing the performance of each scenario for each parameter. The ideal number of clusters was found to be 3 with a degree of fuzziness *(m)* of 3.5 out of the three metrics displayed in Table 2. The largest SH, the largest CH, and the smallest DB show the same conclusion.

Table 3 shows the results of fuzzy C-means clustering on DF patient data with 3 clusters and 3.5 m values. "Temp mean" and "Temp median" represent average and median temperature, "RH mean" and "RH median" represent average and median humidity, "RF mean" and "RF median" represent average and median rainfall, "FF mean" represents "FF median" average and median wind speed, and "Elevation mean" and "Elevation median" represent mean and median elevation, and in the column "Number of Patients," there is also extra information in the form of the number of cluster members.

Cluster 0 has moderate temperatures, moderate wind speeds, high humidity and rainfall, and a

Table 2. Clustering scenario performance based on Silhouette score, Calinski Harabasz score, and Dunn Index

Degree of Fuzziness	Silhouette Score			Calinski Harabasz Score			Dunn Index		
	Number of Cluster			Number of Cluster			Number of Cluster		
	3	4	5	3	4	5	3	4	5
m = 1.5	0.195	0.135	0.010	941	774	736	3.875	3.325	6.824
m = 2.0	0.185	0.067	-0.013	888	1030	669	4.182	3.861	6.227
m = 2.5	0.174	0.057	0.001	863	1114	753	3.113	4.175	5.324
m = 3.0	0.268	0.072	0.026	1413	1180	753	2.012	4.683	13.211
m = 3.5	**0.279**	0.191	0.159	**1947**	1293	1310	**1.740**	2.214	3.259
m = 4.0	0.269	0.219	0.184	1893	1271	954	1.891	3.993	4.931

Table 3. Characteristics of each formed cluster

Cluster	Temp_ mean	RH_ mean	RF_ mean	FF_ mean	Elevation_ mean	Temp_ median	RH_ median	RF_ median	FF_ median	Elevation_ median	# Patient
0	24.675	85.567	23.417	0.928	446.456	24.650	87.500	10.500	1.000	427.000	180
1	26.069	83.861	9.249	0.859	304.725	26.100	84.000	6.600	1.000	391.000	1282
2	24.041	80.886	8.368	1.619	575.913	24.100	82.000	1.600	1.600	583.000	1004

medium altitude average, as seen in Table 3. Cluster 0 is tilted toward the moderate lands' characteristics. Cluster 1 has the highest temperature, lowest wind speed, moderate humidity, low rainfall and humidity values, and the lowest average altitude among the other clusters. Cluster 1 has a lowland bent to its characteristics. Cluster 2 has the lowest temperature characteristics and the highest average altitude among the other clusters, with a high value on wind speed, a low value on rainfall and humidity factors. Highland characteristics describe this cluster.

Table 3 also shows that Cluster 1 had the highest proportion of DF patients, accounting for 52.0 % of all patients. Cluster 2 accounts for 40.7 % of patients, while Cluster 0 accounts for 7.3 %. The majority of the patients live in lowland areas surrounded by forests, beaches, and numerous rivers. It has a significant impact on the growth of DF-causing mosquitos and the disease's dissemination. This condition is consistent with [19], who said about variances in geographic conditions with many instances, and what [10] said about watersheds having a higher number of cases than others.

Similarly, those who claim that areas near woods with plenty of rain can boost mosquito populations [6]. Furthermore, because lowland areas are densely populated, the dispersion rate in this area is high [6]. The fact that Cluster 1 has a higher mean temperature can decrease the time between larval development and the extrinsic incubation period backs up [43]. Mosquitoes do not transmit far with decreased wind speeds, allowing them to thrive in the vicinity.

The distribution of patients is visualized based on each variable to further examine clustering findings with FCM. The scatter plot in Fig. 4 depicts the results of the 2-dimensional distribution visualization, where Temp denotes temperature, RH denotes humidity, RF denotes rainfall, FF denotes wind speed, and Elevation denotes altitude. FCM did a good job separating patients into their different groups, as shown in Fig. 4. There are no patients who fall into two or more clusters, as can be shown. This is represented by the number of members in each cluster, which equals the number of patient records combined. Furthermore, each cluster serves as a separator.

When Table 3 and Fig. 4 are compared, it can be seen that the majority of the patients were found in Cluster 1. Their conditions are an altitude of less than 500 masl (with mean of 384.73 masl) and a temperature of 23-30 °C (with mean 26.07 °C). According to a study conducted by [44, 45], the optimal temperature range for mosquitoes to survive is between 20-30 °C, and [12], the best temperature range for mosquitoes to survive is between 21.6-32.9 °C. In terms of humidity, most patients were distributed at an average of 83.86 %. This finding differs from research in [46], which claims that the ideal humidity for growth is 60-80 %.

The difference of findings can be caused by the slightly different altitude and rainfall conditions from the area where this case study is located. Meanwhile, the most patient distribution occurred when the average rainfall was 9.25 mm, and the average wind speed was 0.86 m/s. For rainfall and wind speed can not be compared with other studies because this is still rarely mentioned. In a study conducted in Indonesia [14, 47], it was not stated how much rainfall values support DF distribution. In fact, the

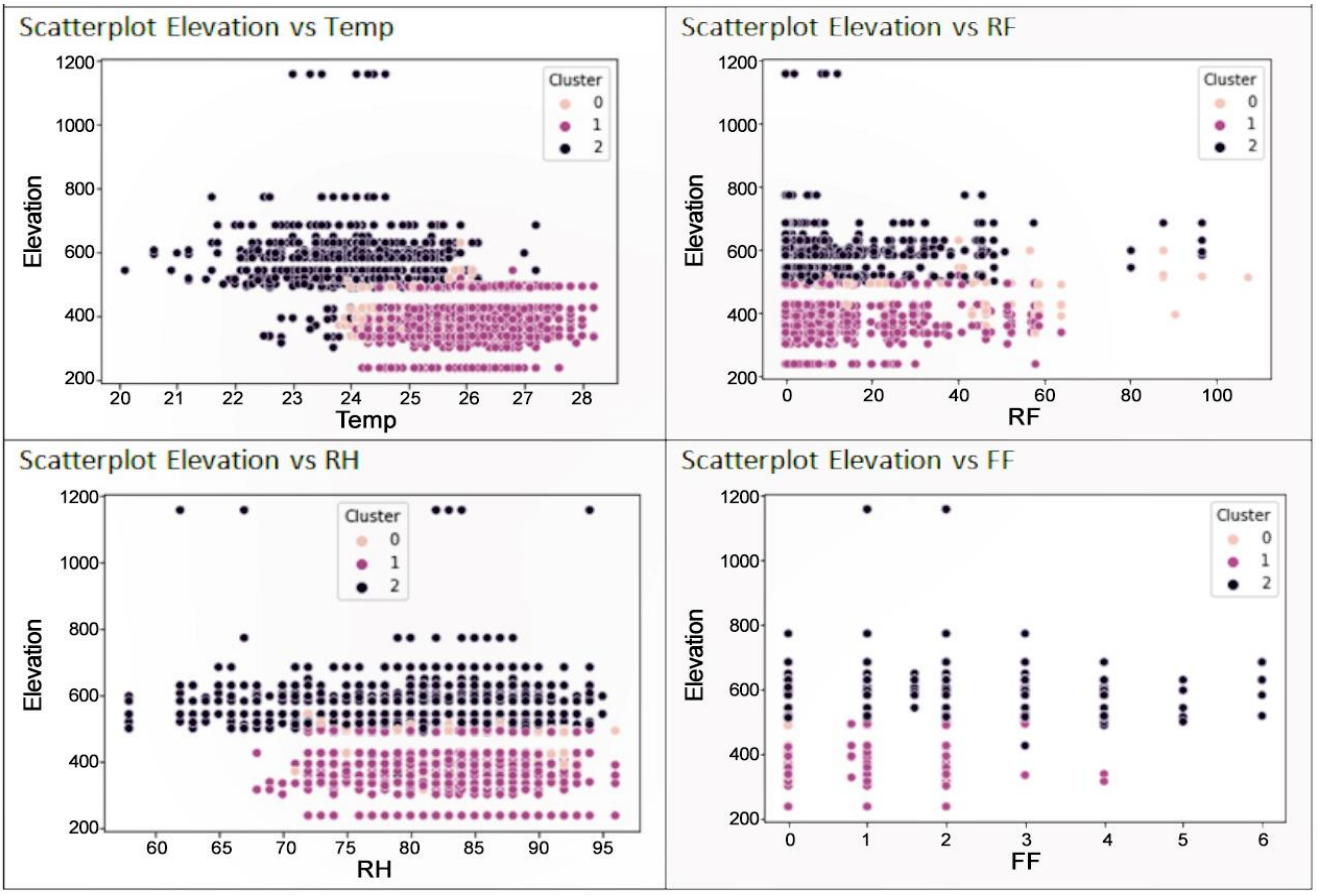

Figure. 4 Scatter plot of the patients' distribution based on climate and elevation variables

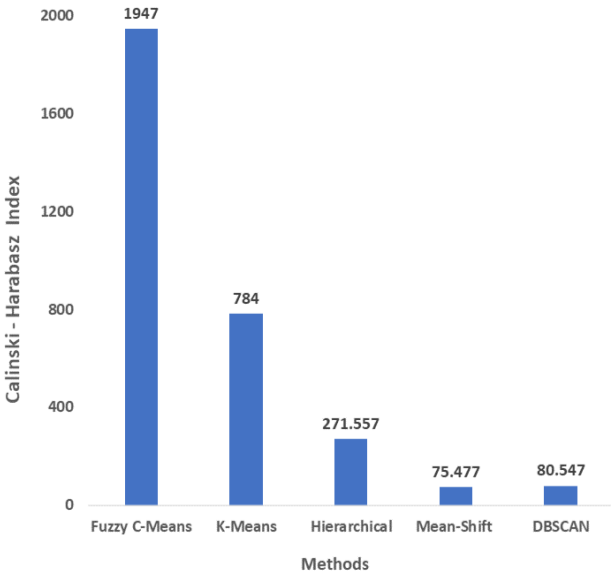

Figure. 5 Calinski-Harabasz Index FCM comparison results with other methods

wind speed variable was not included in the study [14, 15].

Then, to ensure that the clustering results from the FCM are suitable for use, the next validation process is carried out. Validation is done by comparing the Fuzzy C-Means algorithm cluster results with other methods. The methods used in this comparison are K-Means, Complete Hierarchical Clustering, Mean-Shift Clustering, and DBSCAN Clustering. The comparison results are shown in Fig. 5. Fig. 5 displays the comparison of FCM with other methods by using the Calinski-Harabasz Index. The better cluster results than the others will have the highest value on the index. It can be seen in Fig. 5 that the fuzzy C-means algorithm has the highest value compared to other algorithms. The difference in the 3D distribution of patients for FCM and the comparison method is shown in Fig. 6.

Fig. 6 illustrates that, of the five methods examined, fuzzy C-means clustering has the best cluster separation, evidenced by each cluster being formed together and not separated. This condition backs up the statement [24] that FCM has been successfully applied in various fields. This is because FCM considers the potential of cluster overlap, which is subsequently resolved using the idea of [24, 37].

4.2 Identification of patient distribution by region and month

After collecting information on the spread of patients concerning the altitude and climate, the next step was to determine the spread of patients in each

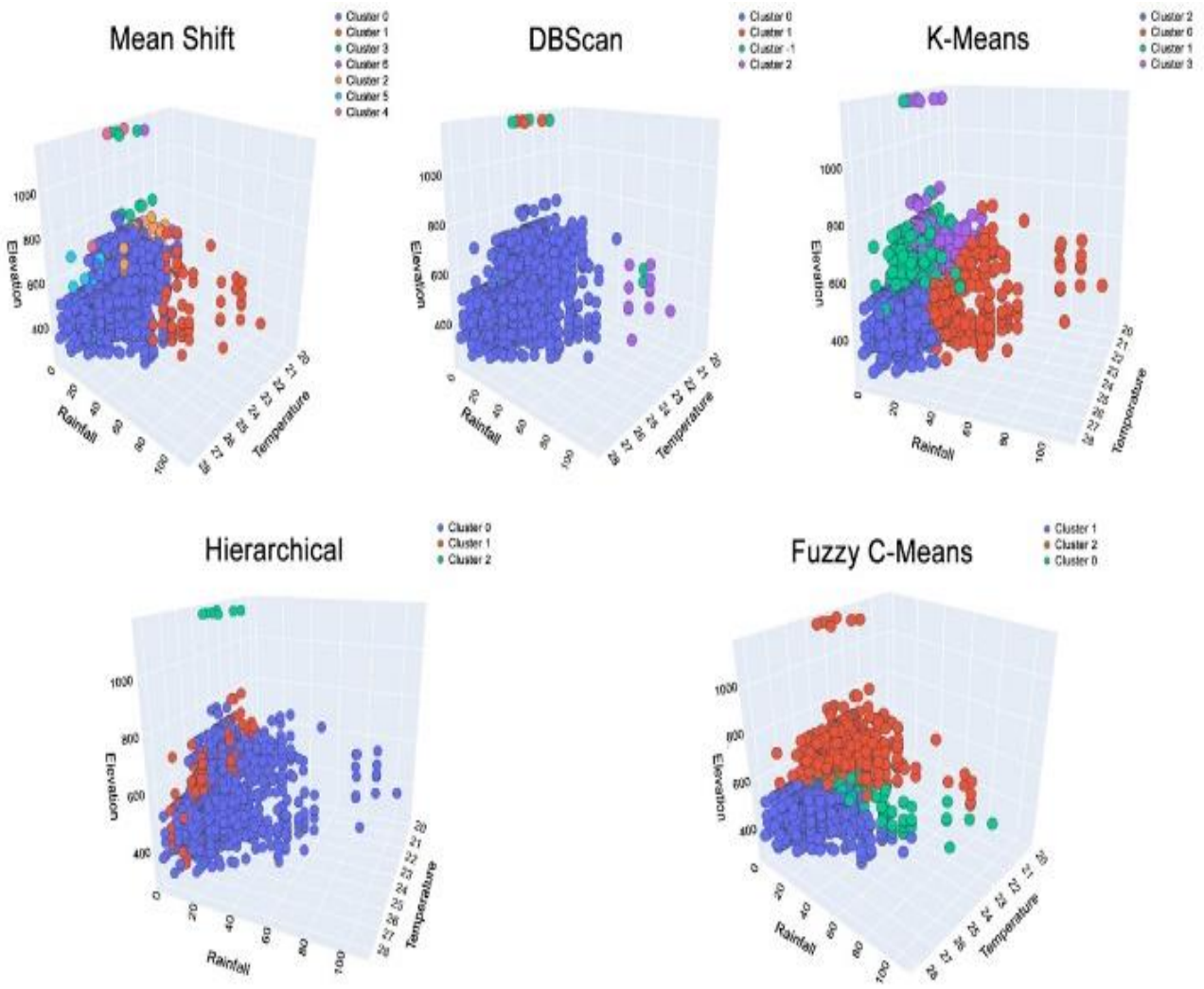

Figure. 6 A 3D plot comparison of the proposed method and others

location and the spread level each month. Because the disease is easily disseminated, knowing which locations to watch out for is essential—in addition, learning when to take preventative measures to avoid an increase in DF occurrences. In addition, the analysis includes information about patients who could be a source of disease transmission. Table 4 shows the regions with the ten most significant distribution levels and related attribute sizes.

According to Table 4, Dau has the highest DF cases. This is demonstrated by the 224 existent nodes or patients and the 8446 edges generated by the relationships. There was just one patient who was not related to the others, based on the number of existing patients. This means that a patient becomes ill after all other patients have recovered and they are not in the same area or on the same side of the border as other patients.

Based on the 75,411 average degree value in the Dau area, it is estimated that someone suffering from DF has a relationship with 75 other sufferers. An average patient can infect/transmit the disease to 75 other people in the same region or directly adjacent to the patient's location. The average path length, which has a value of 3,774, represents the distance of pain and area between one patient and another, suggesting that many patients are suffering from the same ailment and are in close proximity. The network diameter value, which is less than the others, also supports this. This smaller diameter indicates that the maximum distance between the area and the patient's duration of illness in the Dau area is shorter than the others.

Based on the average degree attribute value, the six regions with the highest-ranking are Dau, Gondang Legi, Sumbermanjing Wetan, Pakisaji, Kepanjen, and Turen. This suggests that the rate of disease transmission in these six areas is higher than in other areas. This could be information for the Health Office about the highest-priority needs for preventive, therapeutic, or eradication efforts. The six

Table 4. Ten regions with the highest number of DF cases and the attributes of the graph formed

Area	Attribute Value								
	# Node	# Node w/o relation	# Edge	Avg. Degree	Network Diameter	Avg. Path Length	Graph Density	Modularity	Number of Communities
Dau	224	1	8446	75.411	11	3.744	0.338	0.67	10
Gondang Legi	173	7	2292	27614	11	3.565	0.167	0.751	12
Pakisaji	168	3	2106	25.527	10	3.208	0.156	0.742	15
Kepanjen	149	1	1826	24.676	12	3.744	0.168	0.704	14
Singosari	140	4	1358	19.971	24	7.729	0.148	0.73	16
Sumbermanjing Wetan	123	6	1504	25.709	15	5.033	0.222	0.736	13
Turen	122	5	1296	22.154	12	4.052	0.191	0.695	11
Dampit	121	1	978	16.3	10	3.338	0.137	0.749	16
Wagir	105	4	746	14.627	18	5.524	0.145	0.778	15
Karangploso	95	4	992	21.802	11	3.517	0.242	0.691	9

regions that require priority attention, from first to sixth priority, are Pakisaji, Dampit, Karangploso, Gondang Legi, Dau, and Kepanjen, as seen by the diameter network constructed. Table 4 shows that the network diameter is shrinking, implying that the transmission rate increases in a shorter time and a smaller area. Pakisaji, Dampit, Karangploso, Gondang Legi, Dau, and Kepanjen have the highest priority for handling based on the average path length attribute. This regional group is the same as the network diameter ranking regional group.

Upon looking at the density graph, the places that need to be treated quickly are Dau, Karangploso, Sumbermanjing Wetan, Turen, Kepanjen, and Gondang Legi. With a graph density of 0.191, we enter the turbulent region as a new priority. Dau, Karangploso, Turen, Kepanjen, Singosari, and Sumbermanjing Wetan are the six highest priority regions that demand rapid processing based on the modularity attribute. Based on all of the criteria employed, Dau and Kepanjen are the two places that consistently emerge as priority locations based on the attributes. Fig. 7 depicts the graph visualization for the Dau region. The graph is overly dense because there are so many nodes and edges in the Dau region of Fig. 7. There are nodes with many relationships till the color turns dark. Some nodes are separated but still establish other relationships and produce new communities. Then some nodes don't have any relationships at all. This indicates that the patient represented by this node is not in the same stage of their illness and is located in a different area not directly adjacent to theirs.

In Fig. 8, a sample of the graph for additional places where the number of nodes and edges formed is not too significant makes it easier to evaluate every patient. In Kepanjen, a sample of the patient graph (Fig. 8) reveals 17 patients with 64 interactions created. In Kepanjen, the network divides into three communities: one with three patients (patient IDs 441, 444, and 449), another with four patients (patient IDs 2072, 2115, 2117, and 2186), and a third with ten patients (patients with IDs 863, 871, 883, 894, 907, 917, 955, 972, 985, and 1012). There were no unrelated patients in this network. This fact suggests that all patients in their particular communities suffer from the same illness and are located in the same or adjacent places. The weight of the relationship between two nodes determines the thickness of the edge between them. The greater the weight, the thicker the edge. This indicates that the two patients are nearby.

The next scenario involves a distribution study based on the disease's spread through time. In this example, the time period specified in months. This choice is based on the Public Health Office's requirements for monitoring and planning the budget and activities to be undertaken [4]. Months are required rather than days. Table 5 shows the distribution of graph attribute values from January to December in the six regions with the most instances.

Judging by the number of existing patients, represented by the total number of nodes, cases with a higher number occur throughout the rainy season months of January, February, March, and even April. This finding is consistent with [6], who found that the average number of high cases occurred during the

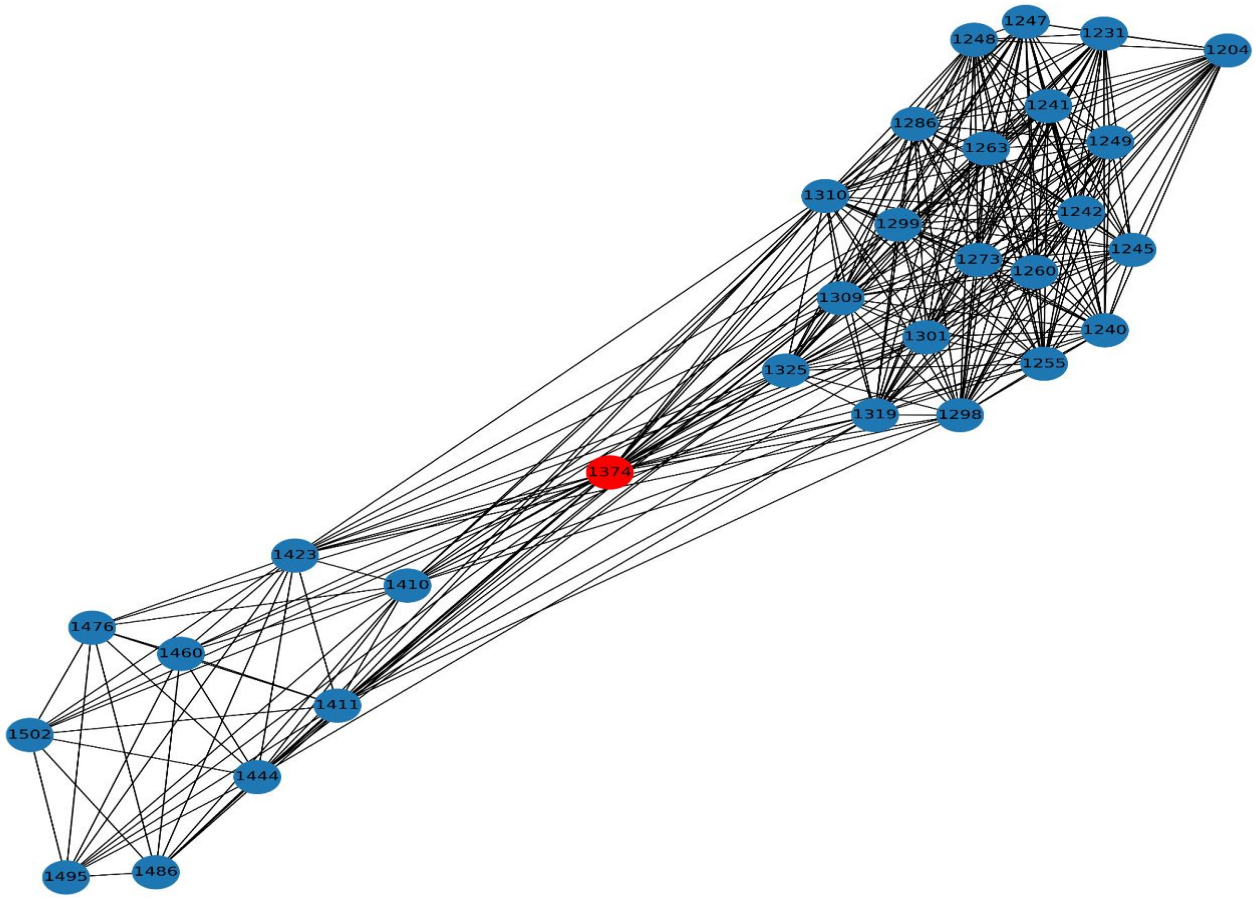

Figure. 7 Visualization of patients' relation

rainy season. However, this finding contrasts with what [7] claimed in his case study area, where they stated that a significant proportion of instances happened during the summer. It can happen because of the difference in the number of seasons and the temperatures between seasons 2 and 4. In addition, the number of nodes without a relationship is zero from January to May and November. This condition suggests that the patients in that month are connected, whether they become ill within the same ten days, live in the same area, or are directly adjacent to one another. This occurrence indicates that a patient can infect others or that a prior patient can infect another patient.

Table 5 further demonstrates that each month has a distinct amount of nodes generated. It is calculated based on the number of dengue fever patients in each location—similarly, the number of created edges. Edges have a higher number of edges in months with a significant number of nodes. During the rainy season, the average relationship between dengue fever patients is higher than during the dry season. This discrepancy in average could be attributed to the season's climatic circumstances. The rainy season's climatic circumstances are more suitable for mosquito reproduction, resulting in a higher spread rate.

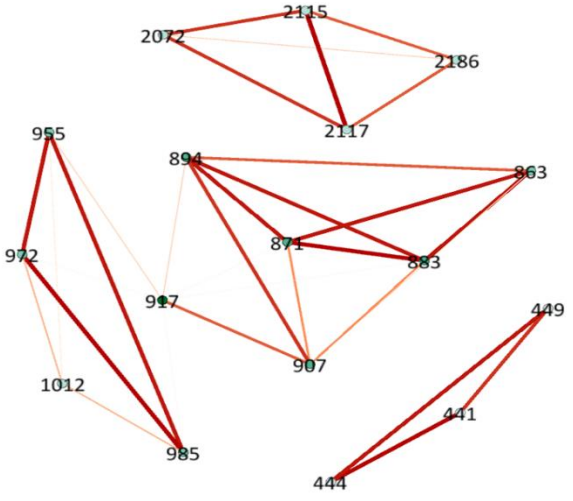

Figure. 8 A Sample of the patients' distribution graph

This statement corroborates the findings of [6, 11]. This assertion is backed up by the average distance between nodes, which is depicted by the average network diameter in the dry season, which is lower than in the rainy season. The average number of patients in a network is also higher during the rainy season. As a result, the possibility of dispersal is larger during the rainy season than during the dry season. This is in line with what was said by [7, 43]. However, this is in contrast to what [11] claimed, which stated that rain caused DF cases to drop. Heavy rains do not support mosquito density because most mosquito eggs and larvae will drift away from the breeding areas.

The average path length represents the average value of the number of patients in a network. During the rainy season, a patient's influence is greater than during the dry season. The height of the rainy season, January, has an average degree value of 46,044, suggesting that a DF patient can infect 46 other persons in the same area or directly next to the patient's location for ten days. It turns out that the 137 sick persons are all related. Three thousand one hundred fifty-four relationships have been developed due to these 137 individuals. This is unlike August-September when the dry season is typically at its peak. In August, just 64 relationships for 23 individuals were found in the six regions with the most cases.

The average path length in January is 1.756, which is 0.082 different from June, indicating that many people are suffering from the same condition and are in close proximity to one another. This is further confirmed by the fact that the network diameter value is lower than in June. This means the greatest distance between the area and the duration of illness is shorter than the others. The density graph depicts the degree of node connectedness. In comparison to other months, September, which is normally the end of the dry season, has the highest level of connectivity between nodes. In fact, as compared to the rainy season's height in January or February, the number of patients isn't quite as high. Compared to other months, this shows that the rate of illness spread in this month is the highest. In September, the result was the same with modularity.

4.3 Identification of patients with potential source of spread and their characteristics

This identification is accomplished by centrality ranking. This centrality rating can reveal information about patients who have a significant impact on establishing a network in certain settings. Two scenarios, dependent on geography and time (month), are still used in this research. Table 6 shows the results of the centrality ranking for the top ten patients in the regional scenario.

Table 6 shows the centrality algorithm results to calculate proximity centrality for ten patients with the highest value. According to the findings of calculations utilizing closeness centrality, the patients with ID 1876 from the Ampelgading area had the highest closeness centrality value. The centrality

Table 5. The network attribute values formed in the six regions with the most DF cases

Area	Attribute Value								
	# Node	# Node w/o relati-on	# Edge	Avg. Degree	Network Diameter	Avg. Path Length	Graph Density	Modula-rity	Number of Commu-nities
January	137	0	3154	46.044	4	1.756	0.339	0.649	8
February	133	0	3038	45.684	4	1.664	0.346	0.625	8
March	125	0	2880	46.080	4	1.521	0.372	0.59	7
April	79	0	840	21.266	4	1.731	0.273	0.735	9
May	86	0	914	21.256	4	1.69	0.25	0.791	9
June	63	1	520	16.744	5	1.842	0.275	0.788	8
July	26	2	90	7.500	2	1.262	0.326	0.744	6
August	23	3	64	6.400	3	1.318	0.377	0.774	5
September	17	3	62	8.857	2	1.34	0.681	0.399	4
October	15	2	36	5.538	3	1.292	0.462	0.615	3
November	49	0	444	18.122	4	1.556	0.378	0.633	6
December	40	2	284	14.947	4	1.362	0.404	0.63	5

Table 6. Centrality value for the ten highest-ranking patients

Ranking	Centrality											
	Closeness			Betweenness			Degree			Eigenvector		
	ID	Value	Area	ID	Value	Area	ID	Value	Area	ID	Value	Area
1	1876	0.459	Ampelgading	1567	0.235	Donomulyo	1875	0.468	Ampelgading	1080	0.707	Ngantang
2	1644	0.365	Tumpang	1659	0.213	Bululawang	1644	0.359	Tumpang	320	0.707	Pujon
3	1948	0.323	Bululawang	1926	0.199	Singosari	443	0.286	Wonosari	443	0.524	Wonosari
4	1840	0.309	Dau	1522	0.189	Tajinan	1872	0.274	Poncokusumo	1861	0.5	Kromengan
5	443	0.298	Wonosari	1136	0.186	Sumbermanjing Wetan	2034	0.271	Bululawang	705	0.461	Bantur
6	1422	0.284	Ngajum	813	0.158	Wagir	1494	0.269	Kalipare	2354	0.461	Pagak
7	1494	9.283	Kalipare	1711	0.124	Poncokusumo	1936	0.25	Tirtoyudo	1113	0.448	Gedangan
8	1936	0.266	Tirtoyudo	1628	0.11	Turen	1422	0.247	Ngajum	1775	0.434	Wajak
9	1867	0.261	Karangploso	1469	0.097	Ampelgading	1640	0.246	Dau	1494	0.426	Kalipare
10	1872	0.238	Poncokusumo	1374	0.097	Dau	1354	0.217	Tajinan	602	0.4	Lawang

of proximity is equivalent to 0.459. This patient ID node has quick access to other nodes in the patient database. Furthermore, it has the quickest path to other patient nodes and, of course, excellent visibility to see what is going on in the network. Patients with ID 1644 from the Tumpang area with a value of 0.365 and patients with ID 1948 from the Bululawang area with a value of 0.323 were ranked 2 and 3, respectively.

In the meantime, Table 6 shows that patients with ID 1567 in the Donomulyo area have the greatest betweenness centrality value, which is 0.235. This means that patients with ID 1567 are sandwiched between two major groups in the resulting network. Patients with ID 1567 and the greatest betweenness centrality rating are almost probably the most influential or powerful people in the network.

Because this node in the network connects numerous smaller communities to build one larger community, this patient significantly impacts what happens in the network. Of course, if this node isn't in the network, it will break the link between multiple smaller communities. When a patient node is removed from the network, the network is split into two smaller communities.

Fig. 8 serves as an illustration. Node ID 917 is the node in the same place as node ID 1567 in Table 6. As shown in Fig. 8, node 917 can connect community 1 (nodes with patient IDs 955, 972, 985, and 1012) and community 2 (nodes with patient IDs 863, 871, 883, 894, and 907) into a larger community network (nodes with patient IDs 863, 871, 883, 894, and 907) into a larger community network (nodes with patient IDs 863, 871, 883, 894, 907, 917, 955, 972, 985, and 1012).

If there is no node 917, the ten-node community will be split into two smaller and distinct communities, the community 1 network and the community 2 network. Patients with ID 1659 from Bululawang had a betweenness value of 0.213, and patient ID 1926 from Singosari had a value of 0.199, respectively, with a betweenness ranking of 2 and 3. These patients had the same function as the rank 1 node but lacked the potency of node ID 1567.

Meanwhile, patients with ID 1875 from Ampelgading, ID 1644 from Tumpang, and ID 443 from Wonosari got the top three rankings for centrality value. The degree centrality score represents the number of interactions a patient has. Patients with ID 1080 from Ngantang, ID 320 from Pujon, and ID 443 from Wonosari are then considered eigenvector centrality.

The centrality value is used to determine which patients can influence the spread of dengue disease in specific months in the monthly identification scenario. Only patients were taken in this scenario from January 1 to December 31, 2019, with the highest cases in six regions. As a result, we generated a new patient ID. Table 7 shows the results of monthly centrality calculations in 2019 for the six regions with the most cases, based on the preceding regional scenario's results. The patients with the highest scores among all the other patients for that month are displayed below. The centrality algorithm is used to rank the items. Patients with ID 516 in January had the highest closeness centrality value compared to other patients, according to calculations using the closeness centrality algorithm. This patient hails from the Kepanjen area. The person in question is a female, 34 years old, with DF as her initial diagnosis. The first checkup is at the Kepanjen public health center, followed by treatment at the Wava Husada Kepanjen Hospital.

Meanwhile, based on betweenness centrality, Table 7 demonstrates that patients with ID 552 from Kepanjen have the highest value betweenness

centrality. This demonstrates how patients with the ID 552 become intermediary participants in the network that forms, connecting two enormous groups. Patients with ID 516 and ID 511 had the greatest scores for degree of centrality and eigenvector centrality. They were both from Kepanjen and had DF status when first diagnosed.

Patients who have received the top ratings in previous months are similarly rated. Each one can be identified and used as preliminary analysis material by the Public Health Service to determine the next steps in preventing the spread of DF disease. Descriptive information of patient characteristics with the highest ranking in region and month scenarios is displayed in Fig. 9, Fig. 10, Fig. 11, and Fig. 12.

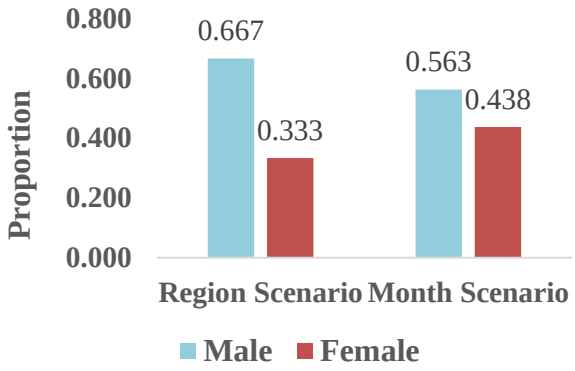

Figure. 9 Characteristics of patients with the highest centrality value based on gender

Fig. 9 depicts the characteristics of patients with a high centrality value as measured in terms of area and month. In terms of area, males are more dominant than females in patients with high centrality. The proportional difference is 33.34 percent. This disparity can be described as significant. However, it backs up claims from [43], but the reason for this has yet to be addressed. This patient was predominantly in the 20–60-year age range, which accounted for 56.67 percent of the total. However, In the monthly analysis, patients with a large influence are more male, as seen in Fig. 9 month scenario, despite the difference being just 12.50 %. This backs up [43]'s claim that males make up a larger proportion of the population, despite being impossible to say that men are more dominant. There is no significant proportion between male and female DF patients since their characteristics are generally male and female.

Following that comes the teenage age group in the Fig. 10 region scenario, which ranges from 11 to 19 years old, with a 36.67 percent difference in proportion. It differs from [6], who claims that persons aged 60 to 69 had the highest risk of spreading the disease. However, he later noticed that the average age of patients is down in several nations, while the proportion of minors is increasing [6]. Then, in this investigation, exceptional patients in the older category were discovered. They are at the top of the list because they have significant ties with other patients and many of them. This condition suggests that people in their eighties and nineties are still at risk of contracting the disease. It can also infect

Table 7. Highest-ranking patients on each centrality measure in each month

Month	Centrality											
	Closenenss			Betwenness			Degree			Eigenvector		
	No_ID	Value	Area	No_ID	Value	Area	No_ID	Value	Area	No_ID	Value	Area
January	516	0.396	Kepanjen	522	0.023	Kepanjen	516	0.346	Kepanjen	511	0.172	Kepanjen
February	479	0.412	Kepanjen	469	0.010	Kepanjen	479	0.386	Kepanjen	479	0.176	Kepanjen
March	200	0.362	Gondanglegi	440	0.011	Kepanjen	451	0.371	Kepanjen	245	0.167	Gondanglegi
April	435	0.250	Kepanjen	435	0.022	Kepanjen	96	0.231	Dau	646	0.281	Sumbermanjing Wetan
May	425	0.243	Kepanjen	424	0.039	Kepanjen	427	0.235	Kepanjen	427	0.290	Kepanjen
June	703	0.243	Sumbermanjing Wetan	417	0.034	Kepanjen	703	0.242	Sumbermanjing Wetan	60	0.331	Dau
July	45	0.320	Dau	45	0.040	Dau	45	0.320	Dau	354	0.408	Pakisaji
August	415	0.234	Kepanjen	415	0.013	Kepanjen	415	0.227	Kepanjen	414	0.477	Kepanjen
September	365	0.563	Pakisaji	365	0.074	Pakisaji	365	0.563	Pakisaji	365	0.433	Pakisaji
October	413	0.298	Kepanjen	413	0.044	Kepanjen	413	0.286	Kepanjen	374	0.498	Pakisaji
November	410	0.399	Kepanjen	685	0.032	Sumbermanjing Wetan	409	0.396	Kepanjen	409	0.255	Kepanjen
December	528	0.333	Kepanjen	11	0.027	Dau	400	0.333	Kepanjen	500	0.325	Kepanjen

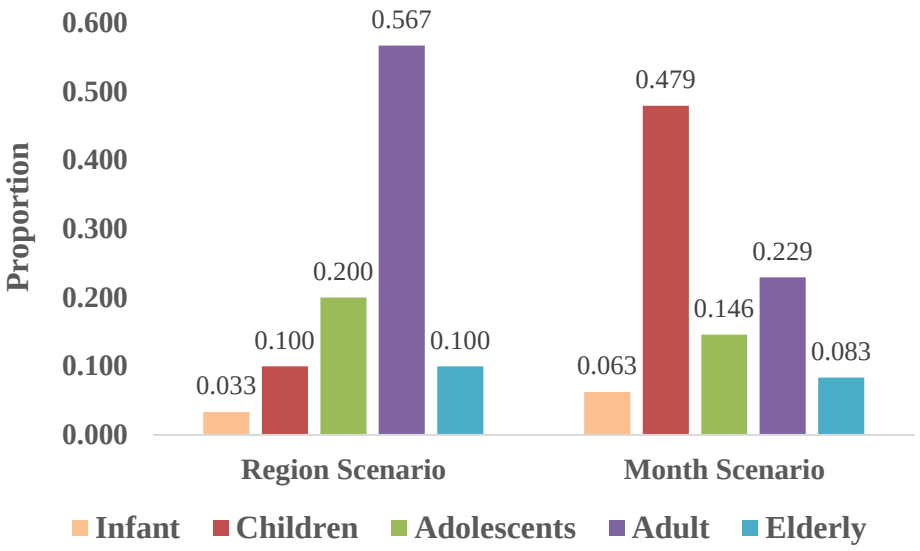

Figure. 10 Characteristics of patients with the highest centrality value based on age group

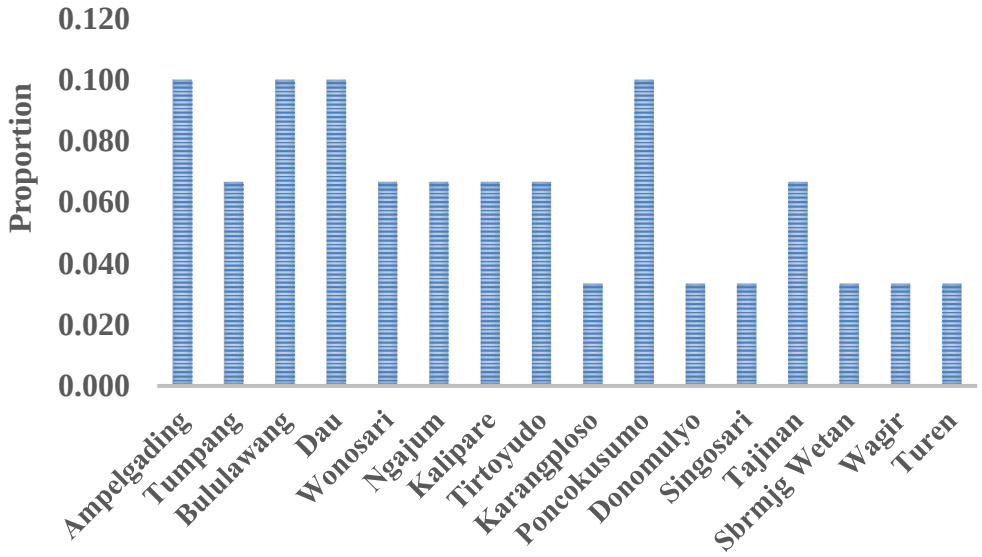

Figure. 11 Characteristics of patients with the highest centrality value based on region scenario

anyone in proximity. Besides, when considering age groupings in the month scenario, it becomes clear that patients with a significant monthly influence are usually children aged 2 to 10 years old. With the second proportion, the adult group aged 20-60 years, the disparity is enormous. This difference is approximately 25 %. It demonstrates that children aged 2 to 10 years can be contagious and infect other patients. The data in this age group support [6]'s statement that the average age of patients is declining while the proportion of children is growing in several countries. The age groups are discovered differently than in the regional scenario in the monthly scenario because the monthly scenario only uses case data from 2019. Furthermore, this data on the features of different sexes and age groups can provide further basic knowledge for future early warning systems development.

The distribution of patients in Fig.11 showed that patients with the highest centrality scores were spread across several areas. The four areas with the highest scores are Ampel Gading, Bululawang, Dau, and Poncokusumo. It is distributed globally within a few years as a scenario region. In comparison, the spread of patients specifically for a certain period, such as the month scenario, is shown in Fig. 12. Based on Fig. 12, patients with the highest-ranking require more monthly monitoring is usually found in the Kepanjen area. Following that are the areas of Dau, Pakisaji, Sumbermanjing Wetan, and Gondanglegi. There are DF cases in the Kepanjen region except for September, although it is not the rainy season.

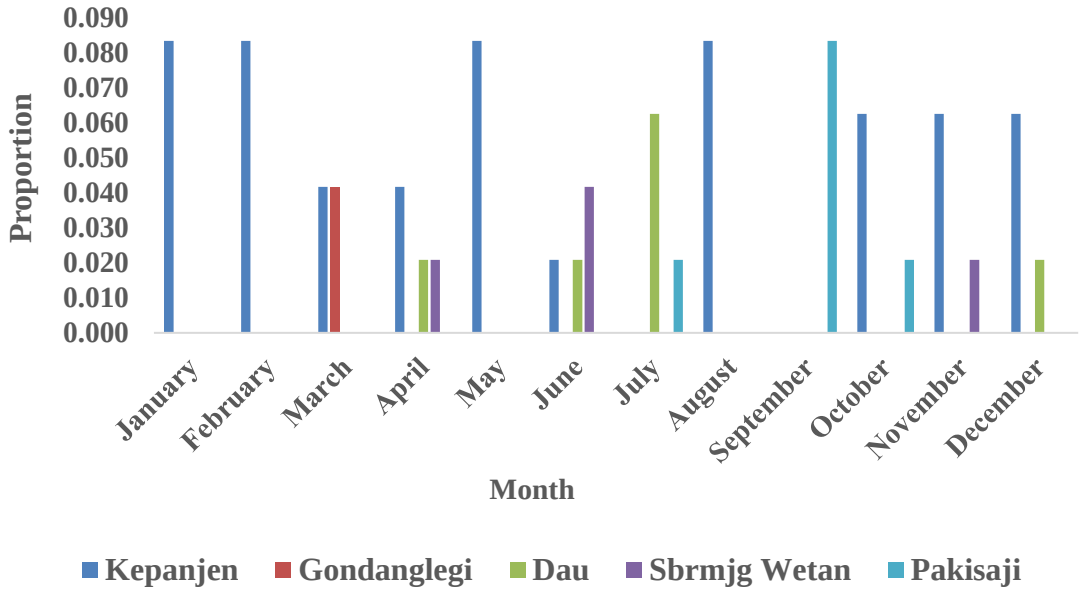

Figure. 12 Characteristics of patients with the highest centrality value in each month based on month scenario

5. Conclusion and future works

In order to understand the distribution of patients and instances of dengue fever, this study combines information from the domains of health, computer science, and social network science. The discussion includes climatic and geographical characteristics of the location to make the analysis more realistic. Temperature, humidity, rainfall, and wind speed are all climatic elements. The elevation is one of the geographical elements. To visualize the distribution of patients, a combination of fuzzy C-means and social network analysis is recommended. Patients were clustered into three clusters depending on climate and elevation, with more patients spread out in clusters identical to the characteristics of lowland areas. Patients clustered with fuzzy C-means performed better than those clustered with K-means, complete hierarchical clustering, mean-shift clustering, and DBSCAN clustering. The relationship between patients is depicted in a graph utilizing the idea of social network analysis, which can exhibit characteristics that can subsequently be converted into information about patient distribution. Various centrality algorithms were used to identify patients who could be the center of disease spread in regional and month settings. According to the detection results, male patients have the largest impact on the network. There is a difference in the proportion difference between each sex by location and month in the analysis scenario. Because the data period used is different, this finding occurs. The same can be said for the findings of age group data. This disparity in information on the features of sex and age groups, on the other hand, can be used to supplement basic knowledge in order to establish an early warning system in the future. A more comprehensive information system could make it easier for health agencies to adjust their dengue preventive tactics as conditions change.

This research can be expanded in the future by including other characteristics as precondition criteria involved in weighing patient interactions. Demographic and environmental variables are among them. This is conducted so that the information obtained is more detailed and accurate to the requirements. This data can then be utilized to plan and manage resources and form the basis of recommendations for preventing the spread of DF.

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 Office for their assistance.

References

[1] WHO, *Weekly Epidemiology Report for 29 July*. 2016. [Online]. Available: http://www.who.int/wer/2016/wer9130/en/

[2] 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.

[3] 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, 2021).

[4] 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, 2021).

[5] L. P. Campbell, C. Luther, D. M. Llanes, J. M. Ramsey, R. D. Lozano, and A. T. Peterson, "Climate Change Influences on Global Distributions of Dengue and Chikungunya Virus Vectors", *Philosophical Transactions of the Royal Society of London. Series B, Biological of Sciences*, Vol. 370, No. 1665, p. 20140135, 2015.

[6] J. L. Duarte, F. A. D. Quijano, A. C. Batista, and L. L. Giatti, "Climatic Variables Associated With Dengue Incidence in a City of the Western Brazilian Amazon Region", *Revista da Sociedade Brasileira de Medicina Tropical*, Vol. 52, 2019.

[7] M. K. Butterworth, C. W. Morin, and A. C. Comrie, "An Analysis of the Potential Impact of Climate Change on Dengue Transmission in the Southeastern United States", *Environmental Health Perspectives*, Vol. 125, No. 4, pp. 579–585, 2017.

[8] L. Xu, L. C. Stige, K, S, Chan, J. Zhou, J. Yang, S. Sang, M. Wang, Z. Yang, Z. Yan, T. Jiang, L. Lu, Y. Yue, X. Liu, H. Lin, J. Xu, Q. Liu, and N. C. Stenseth, "Climate Variation Drives Dengue Dynamics", In: *Proc. of National Academy of Sciences*, Vol. 114, No. 1, pp. 113–118, 2017.

[9] S. Atique, S. S. Abdul, C. Y. Hsu, and T. W. Chuang, "Meteorological Influences on Dengue Transmission in Pakistan", *Asian Pacific Journal of Tropical Medicine*, Vol. 9, No. 10, pp. 954–961, 2016.

[10] R. Li, L. Xu, O. N. Bjørnstad, K. Liu, T. Song, A. Chen, B. Xu, Q. Liu, and N. C. Stenseth, "Climate-Driven Variation in Mosquito Density Predicts the Spatiotemporal Dynamics of Dengue", In: *Proc. of National Academy of Sciences*, Vol. 116, No. 9, pp. 3624–3629, 2019.

[11] Y. H. Lai, "The Climatic Factors Affecting Dengue Fever Outbreaks in Southern Taiwan: an Application of Symbolic Data Analysis", *Biomedical Engineering OnLine*, Vol. 17, No. 2, p. 148, 2018.

[12] J. Xiang, A. Hansen, Q. Liu, X. Liu, M. X. Tong, Y. Sun, S. Cameron, S. H. Easey, G. Soo, H. C. Williams, P. Weinstein, and P. Bi, "Association between Dengue Fever Incidence and Meteorological Factors in Guangzhou, China, 2005–2014", *Environmental Research*, Vol. 153, pp. 17–26, 2017.

[13] A. Adde, P. Roucou, M. Mangeas, V. Ardilon, J. Claude, D. D. Rousset, R. Girod, S. Briolant, P. Quenel, and C. Flamand, "Predicting Dengue Fever Outbreaks in French Guiana Using Climate Indicators", *PLoS Neglected Tropical Disease*, Vol. 10, No. 4, p. e0004681, 2016.

[14] A. L. Ramadona, L. Lazuardi, Y. L. Hii, Å. Holmner, H. Kusnanto, and J. Rocklöv, "Prediction of Dengue Outbreaks Based on Disease Surveillance and Meteorological Data", *PLOS ONE*, Vol. 11, No. 3, p. e0152688, 2016.

[15] 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 and Systems*, Vol. 12, No. 2, pp. 246–260, 2019, doi: 10.22266/ijies2019.0430.24.

[16] W. Anggraeni, H. N. Pradani, S. Sumpeno, E. M. Yuniarno, R. F. Rachmadi, Pujiadi, and M. H. Purnomo, "Prediction of Dengue Fever Outbreak Based on Climate and Demographic Variables Using Extreme Gradient Boosting and Rule-Based Classification", In: *Proc. of 2021 IEEE 9th International Conference on Serious Games and Applications for Health (SeGAH)*, pp. 1–8, 2021.

[17] S. Naish, P. Dale, J. S. Mackenzie, J. McBride, K. Mengersen, and S. Tong, "Climate change and Dengue: a Critical and Systematic Review

of Quantitative Modelling Approaches", *BMC Infectious Disease*, Vol. 14, No. 1, p. 167, 2014.

[18] W. Anggraeni, E. M. Yuniarno, R. F. Rachmadi, Pujiadi, and M. H. Purnomo, "A Sparse Representation of Social Media, Internet Query, and Surveillance Data to Forecast Dengue Case Number using Hybrid Decomposition Bidirectional LSTM", *International Journal of Intelligent Engineering and Systems*, Vol. 14, No. 5, pp. 209–225, 2021, doi: 10.22266/ijies2021.1031.20.

[19] C. C. Tam, M. O. Driscoll, A. F. Taurel, J. Nealon, and S. R. Hadinegoro, "Geographic Variation In Dengue Seroprevalence and Force of Infection in the Urban Paediatric Population of Indonesia," *PLoS Neglected Tropical Disease*, Vol. 12, No. 11, p. e0006932, 2018.

[20] "Forecasting and Modeling Techniques to Study Climate's Impact on Public Health." https://sinews.siam.org/Details-Page/forecasting-and-modeling-techniques-to-study-climates-impact-on-public-health (accessed Sep. 08, 2019).

[21] W. Anggraeni, A. Abdillah, Pujiadi, L. T. Trikoratno, R. P. Wibowo, M. H. Purnomo, and Y. Sudiarti, "Modelling and Forecasting the Dengue Hemorrhagic Fever Cases Number Using Hybrid Fuzzy-ARIMA", In: *Proc. of 2019 IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH)*, pp. 1–8, 2019.

[22] S. A. Rizvi, M. Umair, and M. A. Cheema, "Clustering of Countries for COVID-19 Cases Based on Disease Prevalence, Health Systems And Environmental Indicators", *Chaos Solitons Fractals*, Vol. 151, p. 111240, 2021.

[23] B. Bhattacharjee, M. Shafi, and A. Acharjee, "Network Mining Based Elucidation of The Dynamics of Cross-Market Clustering and Connectedness in Asian Region: an MST and Hierarchical Clustering Approach", *Journal of King Saud University - Computer and Information Sciences*, Vol. 31, No. 2, pp. 218–228, 2019.

[24] S. Subudhi and S. Panigrahi, "Use of Optimized Fuzzy C-Means Clustering and Supervised Classifiers for Automobile Insurance Fraud Detection", *Journal of King Saud University - Computer and Information Sciences*, Vol. 32, No. 5, pp. 568–575, 2020.

[25] A. Bal, M. Banerjee, A. Chakrabarti, and P. Sharma, "MRI Brain Tumor Segmentation and Analysis using Rough-Fuzzy C-Means and Shape Based Properties", *Journal of King Saud University - Computer and Information Sciences*, pp. 1–19, 2018.

[26] N. Meghanathan, "Assortativity Analysis of Real-World Network Graphs based on Centrality Metrics", *Computer and Information Science*, Vol. 9, No. 3, p. 7, 2016.

[27] S. Stephens and J. D. Applen, "Rhetorical Dimensions of Social Network Analysis Visualization for Public Health", In: *Proc. of 2016 IEEE International Professional Communication Conference (IPCC)*, pp. 1–4, 2016.

[28] K. Zhang, X. Liang, J. Ni, K. Yang, and X. Shen, "Exploiting Social Network to Enhance Human-to-Human Infection Analysis without Privacy Leakage", *IEEE Transactions on Dependable and Secure Computing*, Vol. 15, No. 4, pp. 607–620, 2018.

[29] K. Aziz, D. Zaidouni, and M. Bellafkih, "Social Network Analytics: Natural Disaster Analysis Through Twitter", In: *Proc. of 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, 2019.

[30] 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.

[31] "wabah_penyakit_menular.pdf. [infectious disease outbreak]" Accessed: Feb. 29, 2020. [Online]. Available: https://www.bphn.go.id/data/documents/wabah_penyakit_menular.pdf

[32] JKKI, "Memahami Sistem Kesehatan [Understanding the Health System]", 2019. https://www.kebijakankesehatanindonesia.net/20-sistem-kesehatan/79-Memahami-Sistem-Kesehatan (accessed Feb. 29, 2020).

[33] "Knowledge Centre Perubahan Iklim - Beranda [Climate Change Knowledge Center - Home.]", http://ditjenppi.menlhk.go.id/kcpi/ (accessed Feb. 29, 2020).

[34] O. Serrat, "Social Network Analysis", in Knowledge Solutions: Tools, Methods, and Approaches to Drive Organizational Performance, O. Serrat, Ed. Singapore: Springer, pp. 39–43, 2017.

[35] C. J. J. Sheela and G. Suganthi, "Automatic Brain Tumor Segmentation from MRI using Greedy Snake Model and Fuzzy C-Means Optimization", *Journal of King Saud University - Computer and Information Sciences*, pp. 1-10, 2019.

[36] S. S. Khah, P. F. Marteau, and N. Béchet, "Intrusion Detection in Network Systems Through Hybrid Supervised and Unsupervised Machine Learning Process: A Case Study on the ISCX Dataset", In: *Proc. of 2018 1st International Conference on Data Intelligence and Security (ICDIS)*, pp. 219–226, 2018.

[37] J. Aparajeeta, P. K. Nanda, and N. Das, "Modified Possibilistic Fuzzy C-means Algorithms for Segmentation of Magnetic Resonance Image", *Applied Soft Computing*, Vol. 41, pp. 104–119, 2016.

[38] S. R. Mutheneni, R. Mopuri, S. Naish, D. Gunti, and S. M. Upadhyayula, "Spatial Distribution and Cluster Analysis of Dengue Using Self Organizing Maps in Andhra Pradesh, India, 2011–2013", *Parasite Epidemiology and Control*, Vol. 3, No. 1, pp. 52–61, 2018.

[39] "malangkab-Kondisi Geografis.pdf.", Accessed: Mar. 05, 2021. [Online]. Available: http://malangkab.go.id/uploads/dokumen/malangkab-Kondisi%20Geografis.pdf

[40] A. Setiyoutami, D. Purwitasari, W. Anggraeni, E. M. Yuniarno, and M. H. Purnomo, "Modelling Dengue Spread as Dynamic Networks of Time and Location Changes", *International Journal of Intelligent Engineering and Systems*, Vol. 14, No. 3, pp. 346–358, 2021, doi: 10.22266/ijies2021.0630.29.

[41] J. L. Duarte, F. A. D. Quijano, A. C. Batista, and L. L. Giatti, "Climatic Variables Associated with Dengue Incidence in a City of the Western Brazilian Amazon Region", *Revista da Sociedade Brasileira de Medicina Tropical*, Vol. 52, 2019.

[42] O. Serrat, "Social Network Analysis", *Knowledge Solutions*, Springer Singapore, pp. 39–43, 2017.

[43] S. Atique, S. S. Abdul, C. Y. Hsu, and T. W. Chuang, "Meteorological Influences on Dengue Transmission in Pakistan", *Asian Pacific Journal of Tropical Medicine*, Vol. 9, No. 10, pp. 954–961, 2016.

[44] C. W. Tsai, T. G. Yeh, and Y. R. Hsiao, "Evaluation Of Hydrologic and Meteorological Impacts on Dengue Fever Incidences in Southern Taiwan using Time-Frequency Analysis Methods", *Ecological Informatics*, Vol. 46, pp. 166–178, 2018.

[45] K. L. Ebi and J. Nealon, "Dengue in a changing climate", *Environmental Research*, Vol. 151, pp. 115–123, 2016.

[46] P. Siriyasatien, A. Phumee, P. Ongruk, K. Jampachaisri, and K. Kesorn, "Analysis of Significant Factors for Dengue Fever Incidence Prediction", *BMC Bioinformatics*, Vol. 17, No. 1, p. 166, 2016.

[47] 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 and Systems*, pp. 246-260, 2018, doi: 10.22266/ijies2019.0430.24.