International Journal of Engineering & Technology, 7 (4.40) (2018) 162-167

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Spatial Mapping of Toddler Pneumonia Vulnerability in Bojonegoro, Indonesia, Using Hybrid Genetic Algorithm – K-means (GA-Kmeans)

Arna Fariza[1]*, Arif Basofi[1], Dyah Ayu Rosita Devi[1]

[1]Politeknik Elektronika Negeri Surabaya
**E-mail: arna@pens.ac.id*

Abstract

Pneumonia is an acute infection that affects the lung tissue (alveoli) which can be caused by various microorganisms such as viruses, fungi, and bacteria. Pneumonia was the second leading cause of death (13.2%) after diarrhea (17.2%) among under-fives. It shows that pneumonia is a disease that becomes a public health problem that contributes to the high mortality rate of children under five in Indonesia. In recent years, Bojonegoro city gives a large contribution for toddler pneumonia under five years of age in East Java district, Indonesia. In 2012, pneumonia sufferers reached 90.17% of the total number of the toddler and still many the next years. There are five risk factors for pneumonia include the number of children under five, the estimated number of patients, the number of sufferers, environmental factors and nutritional status. A spatial approach is needed to see the spreading of pneumonia vulnerability level in each sub-district in Bojonegoro. This approach can be used by the government as a supporting effort in controlling and preventing pneumonia which is more focused, efficient, and effective. This paper proposes a new approach to generate a vulnerability mapping of toddler pneumonia using hybrid genetic algorithm - K-means (GA-Kmeans) clustering algorithm according to five risk factors. K-means is a clustering algorithm that can produce data groupings based on several attributes well and quickly. However, there is a problem in the initialization stage of the initial random seeds from K-means, which is very difficult to reach an optimum global. The genetic algorithm is used to optimize initial seeds in the K-means algorithm. The vulnerability level of toddler pneumonia is classified into low, medium and high, then it is visualized into spatial mapping. The result of GA-Kmeans test iteration experiments produced best variance cluster 0.99 (almost 1) and determined high levels of vulnerability in 2016 are Kedungadem, Kepohbaru, Baureno, Kanor, Sumberrejo, Balen, Kapas, Bojonegoro, Dander and Ngasem sub-districts.

Keywords: *Pneumonia vulnerability, K-means, genetic algorithms, spatial mapping.*

1. Introduction

Clinical pneumonia (defined as respiratory infections associated with clinical signs of pneumonia, principally pneumonia, and bronchiolitis) in children under five years of age is still the leading cause of childhood mortality in the world [1]. The case management criteria for the classification of acute respiratory infections (ARIs) of the World Health Organization (WHO) are based on infection severity. The clinical spectrum ranges from 'a cough and cold' to life-threatening illnesses such as pneumonia. The WHO estimates that ARIs cause 3 million deaths annually in children below 5 years of age worldwide. Most AR1 deaths occur in developing countries and are mostly due to pneumonia [2]. According to Indonesia basic health research in 2013, pneumonia was the second leading cause of death (13.2%) after diarrhea (17.2%) among under-fives. It shows that pneumonia is a disease that becomes a public health problem that contributes to the high mortality rate of children under five in Indonesia [3].

East Java is a province in Indonesia with high rates of toddler pneumonia. According to the results of recording and reporting in 2012, the number of patients in the district/city of East Java reached 84.392 people [4]. In recent years, Bojonegoro city gives a large contribution for toddler pneumonia under five years of age in East Java district, Indonesia. In 2012, pneumonia sufferers reached 90.17% of the total number of the toddler and still many the next years.

There are several factors that cause pneumonia, they are malnutrition, low birth weight, non-exclusive breastfeeding, lack of measles immunization, parents who smoke, zinc deficiency, mother experiences as caregivers, comorbidities such as diarrhea, heart disease, asthma, maternal education, child care, humidity cold air, vitamin A deficiency, birth order and air pollution outside the house [3][4]. Bojonegoro health service composites the risk factors of toddler pneumonia into 5 factors, includes the number of children under five, the estimated number of patients, the number of sufferers, environmental factors and nutritional status.

A spatial approach is needed to see the spreading of toddler pneumonia vulnerability level in each sub-district in Bojonegoro. This approach can be used by the government as a supporting effort in controlling and preventing pneumonia which is more focused, efficient, and effective. Pneumonia vulnerability level is divided into 3 level for each spatial region of sub-district, namely low, medium and high level. High-risk areas mean they need to handle further by the government, medium risk areas need of supervising and low-risk areas state safe areas.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A clustering algorithm is a method that used to divide unlabeled groups with the multi-attribute. K-means is a clustering algorithm that is widely used because it is simple, fast and efficient. However, there is a problem in the initialization stage of the initial random seeds from K-means, which is very difficult to reach an optimum global [5]. There were several methods developed to improve the initial center of k-means, included subtractive cluster [5], density of objects [6], particle swarm optimization algorithm [7] and genetic algorithm [8]. A genetic algorithm is a supervised optimization algorithm that is widely used because of its speed and optimal global reach capabilities.

This paper proposes a new approach to generate a vulnerability mapping of toddler pneumonia using hybrid genetic algorithm - K-means (GA-Kmeans) clustering algorithm according to five risk factors. The data sources consist of 14 attributes that are decomposed from 5 risk factors. There is three cluster providing the low, medium and high vulnerability level. A genetic algorithm is used to optimize initial seeds in the K-means algorithm. A genetic algorithm capable to identify the right genes through an initial population selection approach. With the fitness function and gene rearrangement operation, it produces high-quality cluster centers. The vulnerability level of toddler pneumonia is classified into low, medium and high, then it is visualized into spatial mapping.

2. Data Collection

Administratively, Bojonegoro district consists of 23 sub-districts as seen in Fig. 1.



Fig. 1: Administration map of Bojonegoro district.

There are 5 factors that affect the spread of toddler pneumonia. The data source is obtained from Bojonegoro Health Service in 2016 included:

1. The Number of toddlers.
2. Number of toddler pneumonia patients
3. Patients found and handled
4. Environmental infection, consist of house eligible, a source of clean water, drinking water source, sanitation, TUPM (Public places and food management).
5. Nutritional status, consist of the number of toddler weighed, more nutrition, good nutrition, less nutrition, bad nutrition, and malnutrition.

It means there are 14 attributes that are processed with the proposed algorithm.

3. Methodology

The methodology of toddler pneumonia vulnerability mapping can be seen in Fig. 2.

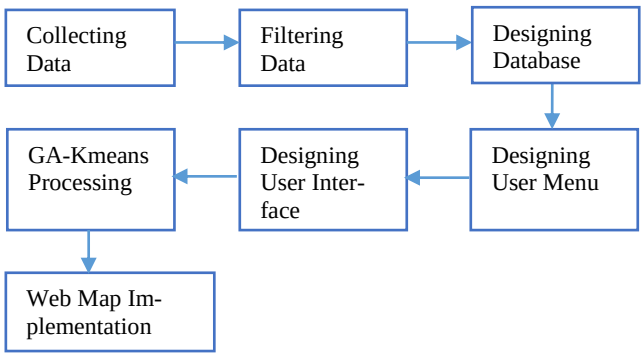

Fig. 2: Methodology of implementation mapping.

1. Collecting Data
   The collection of data is obtained from Bojonegoro Health Service that is consist of the number of infants, the number of patients and the number of estimates of patients who are found and handled, environmental infection factors and nutritional status.
2. Filtering Data
   The data source from Bojonegoro Health Service is necessary to filter according to the correlation of the data. The correlation result is compared with the expert judgment from Bojonegoro Health Service employee. This process produces 14 attributes to the cluster.
3. Designing Database
   It is necessary to create a database to accommodate all attribute data and other information data. We use PostgreSQL to save the spatial and attributes data.
4. Designing User Menu
   User menu is designed because the user is provided with a feature to see how the state and status of a vulnerability in a low, medium or high area
5. Designing User Interface
   The most important thing is to create a user-friendly of the user interface, making it easier for users to operate and obtain information on the web.
6. GA-Kmeans Processing
   GA-Kmeans Processing produces three vulnerability level clusters of each area according to the 14 attributes of risk factors. with some factors that exist by using GA-Kmeans. The cycle of genetic algorithm starts from initializing population randomly, calculating each individual fitness value, selecting the best individual, cross-over, and mutation process to form a new population sequentially. The best individuals obtained from the genetic algorithm optimization process will be the initial center of K-means algorithms. The K-means clustering algorithm is iterated to classify the data into three clusters of vulnerability level.
7. Web Map Implementation
   Spatial vulnerability level mapping of toddler pneumonia in 23 sub-districts is visualized using Google Maps.

3.1. Data Preprocessing

data collected from the Bojonegoro City Health Office was filtered to determine the spread factor of Toddler Pneumonia from 2012 to 2016. The data consists of the number of toddlers, the number of toddler pneumonia patients estimated patients, patient found and handled, environmental infection (consisting of houses eligible, a source of clean water, drinking water source, sanitation, and TUPM), and nutritional status (consisting of the number of toddlers weighed, more nutrition, good nutrition, less nutrition, bad nutrition, and malnutrition). This data pre-processing is processed by calculating the level of correlation to define data linkages and relationships.

Table 1: Correlation values of environmental infection

Year	Correlation Value				
	House eligible	Source of clean water	Drinking water	Sanitation	TUPM
2012	0.81	0.79	079	0.71	0.22
2013	0.88	0.82	0.81	0.64	0.66
2014	0.86	0.61	0.82	0.80	
2015	0.74	0.65	0.61	0.65	
2016	0.76	0.69	0.76	0.76	

Table 1 shows the correlation of environmental infection data of more than 0.5 which states a strong correlation in factors of houses eligible, a source of clean water, drinking water source, sanitation, and TUPM.

Table 2: Correlation values of nutritional status

Year	Correlation Value					
	Toddler weighed	More nutrition	Good nutrition	Less nutrition	Bad nutrition	Mal nutrition
2012	0.84	0.76	0.84	0.70	0.82	0.74
2013	0.89	0.69	0.89	0.69	0.70	0.71
2014	0.79	0.51	0.80	0.61	0.58	0.62
2015	0.62	0.60	0.62	0.51	0.37	0.52
2016	0.77	0.48	0.78	0.65	0.41	0.65

Table 2 shows the correlation of nutritional status data of more than 0.5 excepts more nutrition factor in 2016 (0.48), bad nutrition factor in 2015 (0.37), and in 2016 (0.41), but in average all the factors have a strong correlation. Hence, the number of toddlers weighed, more nutrition, good nutrition, less nutrition, bad nutrition, and malnutrition are the significant factors of this case.

3.2. Pneumonia Factor Tree

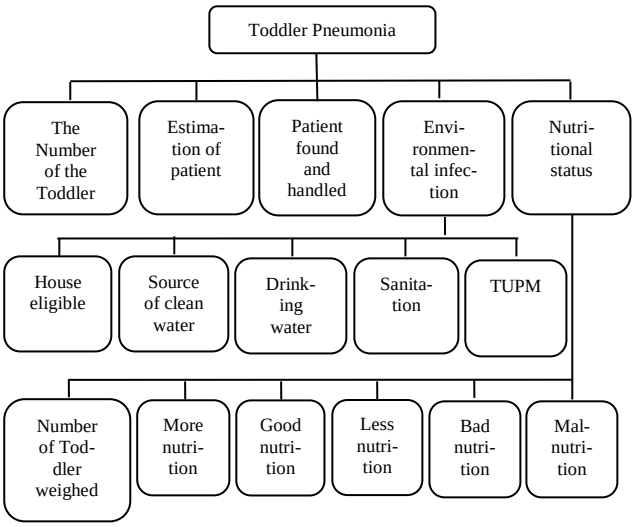

Fig. 3: Toddler pneumonia factor tree

According to the correlation data and expert judgment from Bojonegoro City Health Office, there are 5 factors of pneumonia toddler that influence to toddler pneumonia factor. They are the number of toddlers, estimated patient, patient found and handled, environment infection, and nutritional status. The environment infection is decomposed into house eligible, a source of clean water, drinking water source, sanitation, and TUPM (public places and food management). The nutritional status is decomposed into the number of toddlers weighed, more nutrition, good nutrition, less nutrition, bad nutrition, and malnutrition. All of the 14 attributes can be presented into toddler pneumonia factor tree as seen in Fig 3.

3.3. Hybrid Genetic Algorithm – K-Means (GA-Kmeans) Model

The genetic algorithm model is designed to produce an optimal cluster center according to the toddler pneumonia vulnerability problem. The steps of the genetic algorithm process sequentially consist of initialization of population, evaluating parent of the population according to fitness value, cross-over processing of parent, mutation processing of parent, elitism processing of parent and finding the best individual.

1. Representation of chromosomes
   Chromosomes are the collection of genes that represent the center of the cluster using real number values. Chromosome length is set with multiplication result between the number of attributes and number of clusters (number of attributes × number of clusters). In this study, there are 14 attributes and 3 clusters. It means 3 chromosomes consist of 42 genes or one chromosome consists of 14 genes. The data is normalized in the range 0-1 because each data have different data ranges. The chromosomal representation model used is the float chromosome.

2. Initializing the population
   This process initializes a number of individuals randomly or through certain procedures. Before generating the initial population, we must first determine the number of individuals in the population. for example, the number of individuals is *N*. After that, it only raises the initial population that has random *N* individuals.

3. Calculation of fitness functions
   This process evaluates each population by calculating the fitness value of each chromosome until the stop criteria are met. An individual is evaluated based on a particular function as a measure of its performance. In this case, the fitness function is minimum distance *J* of the minimum distance intra-class between centroid w_r and data x_i among the number of data *N* is defined as (1).

$$J = \sum_{i=1}^{N} min_r d(x_i, w_r) \quad (1)$$

   The flow of the fitness function calculation process includes: calculates the distance of each original data to the centroid represented on the chromosome, take the minimum value from the distance of each data to centroid to *I* and add the shortest distance of data to the centroid that each chromosome has as a fitness value. Then calculate the fitness function using (2)

   Fitness function = $1 / J$ (2)

4. Selection process
   The selection process is used to select parents which individuals are selected for crossover. The selection process here uses a roulette-wheel selection where each chromosome occupies a circle piece on the roulette wheel proportionally according to its fitness value. The process begins by ranking or sorting chromosomes in the population based on fitness and then giving new fitness values based on the sequence. The flow of the selection process includes: the fitness value of each individual is calculated (fi, where i is the 1st to nth individual), calculated the total fitness of all individuals, the probability of each individual is calculated from the probability of random numbers between 1 and 100, random numbers are generated between 1 to 100. From the random number generated, which individual is selected.

5. Crossover Process
   The crossover process is one of the operators in the genetic algorithm that involves two parents to produce new offspring.

Crossover is done by exchanging genes from two parents randomly. Various types of Crossover that are widely used include direct gene exchange and arithmetic gene exchange. The crossover process is carried out on each individual with a specified cross-over probability. An individual who leads to the optimal solution can be obtained through the crossing process, with the note that crossovers can only be done if a random number *r* in the interval [0 1] generated is less than a certain probability *prob*, in other words: $r < prob$. Usually, the *prob* value is set close to 1. The simplest way to do crossovers is a crossing of one intersection. The position of the intersection is random. Determination of individuals who are entitled to carry out crossover operations depends on the predetermined crossover (*Pc*) probability. Steps to perform a crossover operation are as follows:

1) Determine the probability of crossover (*Pc*).
2) Perform the selection process to select individuals who will experience crossover
3) Perform a crossover. Crossover: Exchange directly, with the probability of crossover = 0.9

6. Mutation Process
Mutation is the process of changing the value of one or more genes on a chromosome. Mutations create new individuals by modifying one or more genes in the same individual. Mutations function to replace genes lost from the population during the selection process and provide genes that are not present in the initial population. Each individual experiences a gene mutase with a specified mutation probability. Mutations are carried out by giving inversion values or shifting the value of genes in the selected genes to be mutated. The probability of a good mutation is in the range 0 to 0.15. The probability of a mutation that is too small causes it to be trapped in the local optimum, and the probability of a mutation that is too large causes convergence to be difficult to obtain.

7. Elitism Process
The process of elitism is carried out to preserve the best chromosomes to the new population, so as not to lose the best solution. The elitism used is a ranking elitism system and taken with the best n. This process is done by ranking the fitness value of the parent and child population, then a number of chromosomes are selected as the new population.

4. Result and Discussion

The performance of the GA-Kmeans model of clustering algorithm can be seen in Table 1. The parameters of GA-Kmeans consist of iteration number, crossover and mutation probability. The performance of GA-Kmeans clustering algorithms is evaluated using variance cluster analysis. Variance cluster analysis is used to measure the results of values from the dissemination of clustering data. The variety of clusters result produce from the iteration number, the probability of crossover and the probability. The value of variance cluster *V* is calculated from cluster variance within (V_w) and cluster variance between (V_b) as seen in (3). Cluster variance within means the variance distance of intra-class and cluster variance between means the variance distance of inter-class. The greater the variance value, it reaches the better cluster results, because it has a better similarity level.

$$V = V_w / V_b \tag{3}$$

Cluster variance within (V_w) of *N* data, cluster number *k*, n_i data in the ith cluster and the ith cluster variance V_i^2 is defined as (4)

$$V_w = \frac{1}{N=k}\sum_{i=1}^{k}(n_i - 1) \times V_i^2 \tag{4}$$

A variance of ith cluster variance V_i^2 of n_i data in the ith cluster *i*, and the distance between *j* data and the average of *j* data in the ith cluster is defined as (5)

$$V_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{i}\left(d_j - \overline{d_j}\right)^2 \tag{5}$$

Cluster variance between (V_b) of cluster number *k*, n_i data in the ith cluster and distance between the *i* data and the average of data in the ith cluster is defined as (6)

$$V_b = \frac{1}{k-1}\sum_{i=1}^{k} n_i\left(d_i - \overline{d}\right)^2 \tag{6}$$

The GA-Kmeans model classifies the 23 sub-districts into 3 clusters (low, medium and high) based on 14 attributes of risk factors. The parameter testing of GA-Kmeans model list as

- Iteration number: 1000, 5000 and 10000
- Crossover probability: range of 0.5 – 1.0
- Mutation probability: 0.05, 0.1 and 0.15

Table 3: GA-Kmeans evaluation result

Iteration number	Crossover Probability	Mutation Probability	Variance Cluster
1000	0.5	0.05	0.11
	0.5	0.1	0.16
	0.5	0.15	0.13
	0.6	0.05	0.08
	0.6	0.1	0.13
	0.6	0.15	0.08
	0.7	0.05	0.19
	0.7	0.1	0.24
	0.7	0.15	0.16
	0.8	0.05	0.21
	0.8	0.1	0.71
	0.8	0.15	0.31
	0.9	0.05	0.07
	0.9	0.1	0.18
	0.9	0.15	0.10
	1.0	0.05	0.22
	1.0	0.1	0.16
	1.0	0.15	0.09
5000	0.5	0.05	0.18
	0.5	0.1	0.07
	0.5	0.15	0.11
	0.6	0.05	0.12
	0.6	0.1	0.17
	0.6	0.15	0.12
	0.7	0.05	0.21
	0.7	0.1	0.19
	0.7	0.15	0.17
	0.8	0.05	0.09
	0.8	0.1	0.13
	0.8	0.15	0.12
	0.9	0.05	0.22
	0.9	0.1	0.05
	0.9	0.15	0.13
	1.0	0.05	0.27
	1.0	0.1	0.14
	1.0	0.15	0.22
10000	0.5	0.05	0.05
	0.5	0.1	0.24
	0.5	0.15	0.06
	0.6	0.05	0.41
	0.6	0.1	0.17
	0.6	0.15	0.33
	0.7	0.05	0.45
	0.7	0.1	0.10
	0.7	0.15	0.15
	0.8	0.05	0.05
	0.8	0.1	0.17
	0.8	0.15	0.06
	0.9	0.05	0.09

	0.9	0.1	0.07
	0.9	**0.15**	**0.99**
	1.0	0.05	0.07
	1.0	0.1	0.07
	1.0	0.15	0.06

Table 3 yield the best variance cluster value is 0.99 (almost 1) with the iteration number parameter is 10000, crossover probability is 0.9 and mutation probability is 0.15. From the variance cluster analysis result, we can conclude that the optimum cluster result needs a lot of iteration number, almost all parents are crossed over and enough mutation of parents. From the experimental results that have been done using GA-Kmeans method which has done several experiments of cluster analysis with iteration up to 10000, the probability of crossover is 0.9 and the mutation probability of 0.15 get the best cluster variety and can be used for implementing the spatial map of 23 sub-district in Bojonegoro. The toddler vulnerability level results in 2016 are shown in Tabel 4.

Table 4: Toddler pneumonia vulnerability level in 2016

District	Low	Medium	High
Margomulyo	✓		
Ngraho		✓	
Tambakrejo		✓	
Ngambon	✓		
Bubulan	✓		
Temayang		✓	
Sugihwaras		✓	
Kedungadem			✓
Kephbaru			✓
Baureno			✓
Kanor			✓
Sumberrejo			✓
Balen			✓
Kapas			✓
Bojonegoro			✓
Trucuk		✓	
Dander			✓
Ngasem			✓
Kalitidu		✓	
Malo		✓	
Purwosari	✓		
Padangan		✓	
Kasiman		✓	

As seen in Table 4, from the 23 sub-districts of Bojonegoro in 2016, there are 4 sub-districts in the low level of vulnerability, 9 sub-districts in the medium level of vulnerability and 10 sub-districts in the high-level vulnerability. The mapping of the sub-districts level of vulnerability of toddler pneumonia can be seen in Fig 4. A low area is described with the green color, medium with blue color and high with red color. The visualization of spatial mapping according to GA-Kmeans clustering of toddler pneumonia vulnerability result as shown in Fig 5.

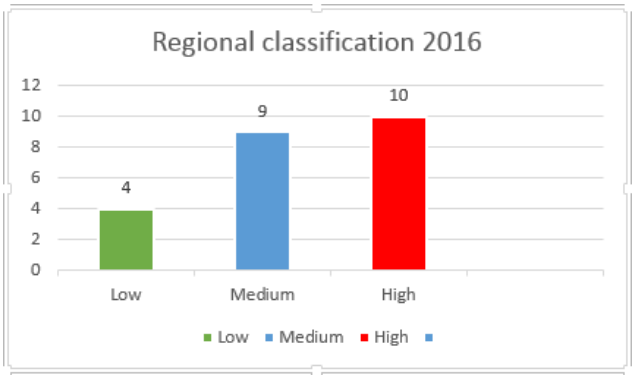

Fig. 4: The recapitulation of toddler pneumonia vulnerability level in 2016



Fig. 5: Spatial mapping of toddler pneumonia vulnerability level in 2016

According to Table 4 and spatial mapping result of toddler pneumonia vulnerability level, there are 10 sub-districts in Bojonegoro that enter the high level of vulnerability, they are Kedungadem, Kepohbaru, Baureno, Kanor, Sumberrejo, Balen, Kapas, Bojonegoro, Dander, and Ngasem. Those sub-districts have to get a handling policy from the government, such as immunization prevention, healthy infrastructure improvements and extraordinary incident of pneumonia planning.

This methodology might also be used to analyze the toddler pneumonia vulnerability every year in Bojonegoro sub-district because there is a form to enter the data with the different year. Therefore, this application can analyze and visualize the pneumonia vulnerability as a support system for the government policy.

In this paper, we also compare the variance cluster value of K-means and GA-Kmeans for the 2016 data source. K-means has a variance cluster value of 0.02 whereas GA-Kmeans has 0.25. Hence, the cluster that is formed by using GA-Kmeans method is more ideal than cluster formed using K-means method.

5. Conclusion

In this paper, GA-Kmeans method can determine the toddler pneumonia vulnerability level into 3 clusters namely low, medium and high level. The variance analysis result of GA-Kmeans gives the best result of 0.99 (almost 1) that it declares the GA-Kmeans method effectively to produce a good cluster. The spatial mapping can be used as the support system for the stakeholder more clearly than text visualization.

K-means has a variance cluster value of 0.02 whereas GA-Kmeans has 0.25. Hence, the cluster that is formed by using GA-Kmeans method is more ideal than cluster formed using K-means method.

Acknowledgement

The author would like to thank Ministry of Research, Technology and Higher Education, the Republic of Indonesia which sponsoring this research and also Bojonegoro health service that supports the data and evaluate the result of the application.

References

[1] Rudan, Igor, Cynthia, Boschi-Pinto, Biloglav, Zrinka, Mulholland, Kim and Campbell, Harry (2008), "Epidemiology and etiology of childhood pneumonia", Bulletin of the World Health Organization, Vol. 86, pp. 408–416.

[2] Gritly, S. M., Elamin, M. O., Rahimtullah, H., Haji, A. Y., Ali, A. D., Mohamed, E. A., & Adetunji, H. A., "Risk Factors of Pneumonia Among Children Under 5 Years at a Pediatric Hospital in Sudan", International Journal Of Medical Research & Health Sciences, Vo.7 No.4 (2018)., pp. 60-68. available online: https://www.ijmrhs.com/medical-research/risk-factors-of-pneumonia-among-children-under-5-years-at-a-pediatric-hospital-in-sudan.pdf, last visit: 30.09.2018

[3] Anwar, A., & Dharmayanti, I, "Pneumonia pada anak balita di Indonesia", Kesmas: National Public Health Journal, Vol.8 No.8 (2014), pp. 359-365. available online:

http://jurnalkesmas.ui.ac.id/kesmas/article/view/405, last visit: 30.09.2018

[4] Susi Susanti, "Pemetaan Penyakit Pneumonia di Provinsi Jawa Timur". *Jurnal Biometrika dan Kependudukan*, Vol.5 No.2 (2017), pp 117-124. available online: https://e-journal.unair.ac.id/JBK/article/view/5843, last visit: 30.09.2018.

[5] Dhanachandra, N., Manglem, K., and Chanu, Y. J., "Image segmentation using K-means clustering algorithm and subtractive clustering algorithm". *Procedia Computer Science*, Vol.54 (2015)., pp. 764-771, available online: https://www.sciencedirect.com/science/article/pii/S1877050915014143, last visit: 30.09.2018

[6] Yuan, Q., Shi, H. and Zhou, X. (2015, June), "An optimized initialization center K-means clustering algorithm based on density". In Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2015 IEEE International Conference on (pp. 790-794).

[7] Gupta, A., Pattanaik, V., and Singh, M. (2017, May), "Enhancing K means by unsupervised learning using PSO algorithm". In Computing, Communication and Automation (ICCCA), 2017 International Conference on (pp. 228-233). IEEE.

[8] Rahman, M. A., and Islam, M. Z., "A hybrid clustering technique combining a novel genetic algorithm with K-Means". *Knowledge-Based Systems*, Vol. 71 (2014), pp. 345-365, available online: https://www.sciencedirect.com/science/article/pii/S0950705114002937, last visit: 30.09.2018