International Journal on Advanced Science Engineering Information Technology

Vol.9 (2019) No. 4
ISSN: 2088-5334

Comparison of Fuzzy C-Means, Fuzzy Kernel C-Means, and Fuzzy Kernel Robust C-Means to Classify Thalassemia Data

Zuherman Rustam[#], Annisa Kamalia[#], Rahmat Hidayat[*], Fajar Subroto[+], Aditya Suryansyah S.[+]

[#]*Department of Mathematics, University of Indonesia, Depok, 16242, Indonesia*
E-mail: rustam@ui.ac.id; annisa.kamalia@sci.ui.ac.id

[*]*Department of Information Technology, Politeknik Negeri Padang, Padang, 25163, Indonesia*
E-mail: rahmat@pnp.ac.id

[+]*Harapan Kita Children and Womens's Hospital, Jakarta, 11420, Indonesia*

***Abstract*— Among the inherited blood disorders in Southeast Asia, thalassemia is the most prevalent. Thalassemias are pathologies that derive from genetic defects of the globin genes. Thalassemia is also considered a health burden among the world's population. Thalassemia cannot be cured, but there is a method to prevent the occurrence of thalassemia by early detection with screening. The aim is to identify the suspected unrecognised diseases in a population that seems healthy and asymptomatic using tests, examinations, or other procedures that can be applied quickly and easily to the target population. Research on thalassemia has been done extensively, such as testing the accuracy of β-thalassemia data in Thailand using the Bayesian Network and Multinomial Logistic Regression. In this study, we will compare the performance of the classification of thalassemia data by Fuzzy C-Means, Fuzzy Kernel C-Means, and Fuzzy Kernel Robust C-Means. The author uses thalassemia data from Indonesia, acquired from Harapan Kita Children and Womens's Hospital, Jakarta, that consists of 82 thalassemia samples from the patients of thalassemia and 68 non-thalassemia samples with 11 features. In total, there are 150 data patients used in this paper. The results show the accuracy of the classification. The accuracy of FCM is 100% when training data is 90%, FRCM is 100% when training data is 90%, and FKRCM, which is the modified Fuzzy, 100% when we use the $\sigma = 0.0001$ and 80% & 90% training data. This result denote that Fuzzy C-Means, Fuzzy Robust C-Means, and Fuzzy Kernel Robust C-Means perfectly classify thalassemia data from Indonesia.**

***Keywords*— thalassemia; fuzzy C-means; fuzzy kernel C-means; fuzzy robust C-means; fuzzy kernel robust C-means.**

I. INTRODUCTION

Thalassemia is an inherited blood disorder that involves the lack of, or a deficiency in the gene that produces haemoglobin, the protein present in red blood cells [1]. The word "Thalassemia" derives from two Greek words, "*thalassa*", which means "sea" and "*haema*", which means "blood". It was named so because of its high prevalence in Mediterranean countries [2]. Earlier, the distribution of thalassemia was predominantly limited to areas in the Mediterranean (Italy, Greece, Turkey, and Cyprus), and across the Middle East through Southern Asia to Southeast Asia (India, Vietnam, and Cambodia) in the so-called 'thalassemia belt.' [3]. Thalassemia is one of the most common chronic diseases in Indonesia.

The *thalassemias* are a group of recessively autosomal inherited conditions characterised by decreased or absence of synthesis of one of the two polypeptide chains (α or β) that form the normal adult human haemoglobin molecule [4]. According to [5], clinically thalassemia is divided into three forms: (a) thalassemia major, indicating patients with severe anaemia and dependent on blood transfusions; (b) thalassemia minor or trait, referring to patients that carry the thalassemia gene (carrier); and (c) thalassemia intermedia, referring to patients with a phenotype ranging in severity from severe anaemia with hepatosplenomegaly and thalassemia-like bone modifications to moderate microcytic hypochromic anaemia. Based on the two polypeptide chains, thalassemia is divided into two: namely beta thalassemia and alpha thalassemia. The majority of thalassemia patients will experience mild anaemia. This anaemia causes a pale face, weakness, decreased appetite, and insomnia. In some cases there is thickening and enlargement of the bones, especially in the head and face bones. Some symptoms and signs of thalassemia include: fatigue, shortness of breath, paleness, yellow skin colour, and swollen stomach.

Prevention of thalassemia is based on prenatal testing, public awareness of the disease, and screening. Screening is the identification of suspected unrecognised diseases in a

population that seems healthy and asymptomatic, by means of tests, examinations or other procedures that can be applied quickly and easily to the target population. In this study, we classify thalassemia using Fuzzy C-Means (FCM), Fuzzy Kernel C-Means (FKCM), and Fuzzy Kernel Robust C-Means (FKRCM), and then we compare the performance of these for thalassemia data in Indonesia.

Several other studies have been conducted related to thalassemia. For example, a comparative analysis of thalassemia screening of KNN, SVM, and Multi-Layer Perceptron [6] as well as classification of *thalassemic* pathologies based on artificial neural network [7]. In Thailand, β-thalassemia data was tested accurately using the Bayesian Network and Multinomial Logistic Regression [8]. In Palestine, β data -thalassemia was identified using balancing techniques, SMOTE and classifiers such as the Decision Tree and Multi-Layer Perceptron [9]. For *thalassemia* screening, *thalassemia* data was classified using a decision tree, K-Nearest Neighbour, and Multi-Layer Perceptron classifier [10]. The classification was conducted using Binomial Logistic Regression Based on Classical and Bayesian Statistics for Screening β-Thalassemia [11].

Fuzzy c-means and fuzzy kernel c-means have been applied in various field, not only for cancer or to classify disease. Application of fuzzy c-means and fuzzy kernel c-means has been used in example for predicting the direction of Indonesian stock price movement [12], and to predict the composite index price [13], and also for forecasting stock market momentum [14], and to solve intrusion data system (IDS) that they claim provides better result [15].

This paper consists of four sections. Section 1 is the introduction. In this paper, the problem to be discussed is thalassemia classification. In section 2, we explain the FCM, FKCM, and FKRCM methods used to classify thalassemia data, and matrix confusion, which is used to calculate the accuracy. In this section we also explain the data used in this study. The experimental results are given in Section 3 along with the discussion. The last section is the conclusion.

II. Material and Method

A. Fuzzy C-Means (FCM) Method

The idea of basic fuzzy clustering called Fuzzy C-Means (FCM) was invented by Bezdek [12]. For a data set $= \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_m}\} \subseteq \mathbb{R}^d$, we define the membership matrix $n \times c$, $U = [u_{ij}]$, where $1 \le i \le n$, $i \le j \le c$, and the cluster centre $\boldsymbol{V} = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \dots, \boldsymbol{v_c}\}$ where each object in V is a d-dimensional vector.

It is based on minimisation of this objective function:

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{C} u_{ij}^m \left\| \boldsymbol{x_i} - \boldsymbol{v_j} \right\|^2 \quad (1)$$

where $m > 1, m \in \mathbb{R}$, u_{ij} is the degree of the membership $\boldsymbol{x_i}$ in the cluster j, x_i is the i-th data on d-dimensional, $\boldsymbol{v_j}$ is the d-dimension of the cluster centre, and $\|*\|$ is any norm representing a similarity between the centre of the data and the data itself. With constraints:

$$\sum_{j=1}^{C} u_{ij} = 1, where\ i = 1,2, \dots, n \quad (2)$$

$$\sum_{j=1}^{C} u_{ij} > 0,\ i = 1,2, \dots, n \quad (3)$$

$$\mathrm{u_{ij}} \in [0.1]\,, \mathrm{j} = 1,2,3, \dots, \mathrm{c}$$

Membership values and cluster centre are updated by using:

$$u_{ij} = \sum_{k=1}^{C} \left(\frac{\|\boldsymbol{x_i} - \boldsymbol{v_j}\|}{\|\boldsymbol{x_i} - \boldsymbol{v_k}\|} \right)^{-\frac{2}{m-1}} \quad (4)$$

$$v_j = \frac{\sum_{i=1}^{n} u_{ij}^m . x_i}{\sum_{i=1}^{n} u_{ij}^m} \quad (5)$$

In this research, we used Fuzzy C-Means (FCM) and the algorithm can be seen in Fig. 1 [17].

Input: $X, c, m_i, m_f, \varepsilon, T$
Output: U *and* V

1. Initial condition : $\boldsymbol{V^0} = [\boldsymbol{v_1}, \boldsymbol{v_2}, \dots, \boldsymbol{v_c}], \boldsymbol{v_j} \in C_j$
2. For $t = 1$ *to* T
3. $m = m_1 + \frac{1}{T}(m_f - m_i)$
4. $b = -\frac{1}{m-1}$

   Calculate membership

5. $u^1 = [u, j], 1 \le i \le n, 1 \le j \le c$ by using $u_{ij} = \frac{\boldsymbol{d^{2b}(x_i, v_j)}}{\sum_{k=1}^{c} \boldsymbol{d^{2b}(x_i, v_k)}}, 1 \le i \le n, 1 \le j \le c$

   Update cluster

6. $\boldsymbol{V^t} = [\boldsymbol{v_1}, \boldsymbol{v_2}, \dots, \boldsymbol{v_c}]$ where $\boldsymbol{v_j} = \frac{\sum_{i=1}^{c} u_{ij}^m x_i}{\sum_{i=1}^{n} u_{ij}^m}, j = 1,2, \dots, c$
7. If $E = \sum_j^c \boldsymbol{d^2}(\boldsymbol{v_{jt}}, \boldsymbol{v_{jt-1}}) \le \varepsilon$,

   STOP ELSE
8. Go to (2)

Fig. 1 Fuzzy C-Means Algorithm

Fuzzy C-Means classification's accuracy is dependent on the types of data. When the data is non-linearly separable, its convergence is slow and inaccurate. To solve this problem, the data set are transformed into another space (feature space) that dimension is much higher than data space [18]. It is expected that the transformed data behaviour can approach the linearly separable data, so that classification accuracy can be improved. We then need a "connector" between data space and feature space, so we can have a better accuracy without directly working at feature space. This concept is called kernel.

B. Fuzzy Kernel C-Means (FKCM) Method

This study will apply the kernel method to the FCM to complete the classification of thalassemia data using Fuzzy Kernel C-Means (FKCM). In Fuzzy Kernel C-Means, we apply the kernel method to the FCM method. With kernel, we can overcome the non-linear problem that are well generalized in combination with linear models.

Suppose $x_1, x_2, \dots, x_n \in \mathbb{R}^n$ is the original dataset in $\mathbb{R}^n$. Then, there is a function ϕ that maps data to a new feature space $\mathbb{F}$ [19]:

$$\phi = \mathbb{R}^n \to \mathbb{F}$$

The kernel function is defined as follows [20]:

$$\boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x_j}) = < \boldsymbol{\phi}(\boldsymbol{x_i}), \boldsymbol{\phi}(\boldsymbol{x_j}) > \quad (7)$$

and the distance function is defined as follows [19]:

$$\boldsymbol{d^2}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{\phi}(\boldsymbol{x_k}) - \boldsymbol{\phi}(\boldsymbol{v_i})\|^2 \quad (8)$$

We use RBF Kernel :

$$\boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x_j}) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (9)$$

,where

$$\boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x_i}) = \boldsymbol{K}(\boldsymbol{v_i}, \boldsymbol{v_i}) = 1 \quad (10)$$

Thus,

$$\boldsymbol{d}^2(\boldsymbol{x}, \boldsymbol{y}) = 2(1 - (\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{y})) \quad (11)$$

The objective functions of Fuzzy Kernel C-Means are as follows:

$J(\boldsymbol{V}, U, \boldsymbol{X}, c, m) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m ||\boldsymbol{\phi}(\boldsymbol{x_k}) - \boldsymbol{\phi}(\boldsymbol{v_i})||^2$ (12)

subject to

$$0 \le u_{ik} \le 1 \quad (13)$$

$$\sum_{i=1}^{c} u_{ik} = 1 \quad (14)$$

$$0 \le \sum_{i=1}^{c} u_{ik} \le n \quad (15)$$

$$i = 1,2,..,c\ ;\ k = 1,2,..,n$$

Where$c \ge 2$ is the number of clusters, n is the number of data, m is the degree of fuzziness with $m > 1$,$\boldsymbol{d}^2(\boldsymbol{x}, \boldsymbol{y})$ is the kernel mapping distance between the x_k data and the cluster centre $\boldsymbol{v_i}$, $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, .., \boldsymbol{x_n}\}$ is the set of data to be clustered, $\boldsymbol{V} = \{\boldsymbol{v_1}, \boldsymbol{v_2}, .., \boldsymbol{v_c}\}$ is the cluster centre set, and $U = [u_{ik}]$ is the matrix of membership function.

The optimum condition of the membership value and cluster centre is as follows:

$$\mathrm{u_{ik}} = \frac{1}{\sum_{\mathrm{j=1}}^{\mathrm{c}}\left(\frac{1-\mathbf{K}(\mathbf{x_k}, \mathbf{v_i})}{1-\mathbf{K}(\mathbf{x_k}, \mathbf{v_j})}\right)^{\frac{1}{\mathrm{m}-1}}} \quad (16)$$

$$\mathrm{v_j} = \frac{\sum_{\mathrm{k=1}}^{\mathrm{N}}(\mathrm{u_{ik}})^{\mathrm{m}}\mathbf{K}(\mathbf{x_k}, \mathbf{v_j})\mathrm{x_k}}{\sum_{\mathrm{k=1}}^{\mathrm{N}}(\mathrm{u_{ik}})^{\mathrm{m}}\mathbf{K}(\mathbf{x_k}, \mathbf{v_i})} \quad (17)$$

In Karayiannis and Bezdek's study of Fuzzy LVQ [21] for different degrees of fuzziness m is used

$$m = m_i + \frac{t}{T}(m_f - m_i)$$

where m_i = initial value of m

m_f = final value of m

Figure 2 shows the algorithm of Fuzzy Kernel C-Means. [22].

Input: $\boldsymbol{X}, c, m_i, m_f, \varepsilon, T$
Output: U *and* $\boldsymbol{V}$

1. Initial condition : $\boldsymbol{V^0} = [\boldsymbol{v_1}, \boldsymbol{v_2}, \dots, \boldsymbol{v_c}], v_j \in C_j$
2. For $t = 1$ *to* T
3. $m = m_i + \frac{t(m_j - m_i)}{T}$
4. $b = -\frac{1}{m-1}$
5. Calculate membership $U^1 = [u_{ij}], 1 \le i \le n, 1 \le j \le c$, by using $u_{ij} = \frac{\boldsymbol{d^b}(\boldsymbol{x_i}, \boldsymbol{v_j})}{\sum_{k=1}^{c}\boldsymbol{d^b}(\boldsymbol{x_i}, \boldsymbol{v_k})}$, $1 \le i \le n, 1 \le j \le c$
6. Update cluster center $\boldsymbol{V^t} = [\boldsymbol{v_1}, \boldsymbol{v_2}, \dots, \boldsymbol{v_c}]$, where $\boldsymbol{v_j} = \frac{\sum_{i=1}^{n} u_{ij}\boldsymbol{x_i}}{\sum_{i=1}^{n} u_{ij}^m}, j = 1,2, \dots, c$
7. If $E = \sum_{j=1}^{c} k^2(\boldsymbol{v_{jt}} - \boldsymbol{v_{jt-1}}) \le \varepsilon$, STOP,else $t = t + 1$

Fig. 2 Fuzzy Kernel C-Means Algorithm

C. Fuzzy Kernel Robust C-Means (FKRCM) Method

In this section, we explain about the FRKCM (Fuzzy Robust Kernel C-Means) method. The reason for using Fuzzy Kernel Robust C-Means is it can show the robustness to outlier, crash, and weighting exponent m. In the FKRCM method, we will map the data from the original data space into a feature space (higher-dimensional space) using the Kernel method.

We first discuss the Fuzzy Robust C-Means. Fuzzy Robust C-Means (FRCM) or Fuzzy Robust Clustering works in almost the same way as Fuzzy C-Means; the differences concern the definition of the membership function, the use of scale in the prototype, and the use of learning rate for each iteration. For the set of data $= \{x_1, \dots, x_m\} \subseteq \mathbb{R}^d$, an objection function in the FRCM algorithm according to [23] is:

$$J(U, \boldsymbol{V}) = \sum_{j=1}^{c}\sum_{i=1}^{n}[u_{ij}^m \boldsymbol{d}^2(\boldsymbol{x_i}, \boldsymbol{v_j}) \quad + \eta_j\left(f(u_{ij})\right)^m] \quad (18)$$

where $u_{ij} \in [0,1]$, C is number of clusters, $U = [u_{ij}]$ is the $n \times c$ membership matrix where $1 \le i \le n,\ 1 \le j \le C$, $1 \le i \le n,\ v_i$ is the d-dimensional cluster centre, $\|*\|$ is any norm shows distance between the data and cluster centre, $m > 1$ is the fuzziness degree , and $f(u_{ij})$ according to [23] :

$$f(u_{ij}) = (1 + u_{ij}\ln u_{ij} - u_{ij}) \quad (19)$$

whereη_j, u_{ij}, v_i according to [23]:

$$\eta_j = \min_k \boldsymbol{d}^2(\boldsymbol{v_i}, \boldsymbol{v_k})\ ,\ for\ i \ne k, \quad (20)$$

$$u_{ij} = exp(-\frac{\boldsymbol{d}^2(\boldsymbol{x_i}, \boldsymbol{v_j})}{\eta_j}) \quad (21)$$

$$\boldsymbol{v_j} = \boldsymbol{v_j} + \alpha_t(\boldsymbol{x_i} - \boldsymbol{v_j})\ exp(-\frac{\boldsymbol{d}^2(\boldsymbol{x_i}, \boldsymbol{v_j})}{\eta_j}) \quad (22)$$

where $\alpha_t = \alpha_0(1 - \frac{t}{T})$ and T is the maximum iteration value, and t shows the t-th iteration.

Next, we will explain about the Fuzzy Kernel Robust C-Means method (FKRCM).

The objective function of Fuzzy Kernel Robust C-Means is:

$$J(U, \boldsymbol{V}) = \sum_{j=1}^{c}\sum_{i=1}^{n}[2u_{ij}^m(1 - \boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{v_j})) + \eta_j\left(f(u_{ij})\right)^m)] \quad (23)$$

Fuzzy Kernel Robust C-Means Algorithm

Input :$\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots., \boldsymbol{x_n}\}$ dataset, $\boldsymbol{V^0} = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \dots, \boldsymbol{v_c}\}$, number of cluster, c, and tolerance, ε

Output : $\boldsymbol{V^0} = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \dots, \boldsymbol{v_c}\}$

For $t = 1{:}T$

1. $\alpha_t = \alpha_0(1 - \frac{t}{T})$
2. Update membership $U^t = [u_{ij}]$

$u_{ij} = exp(-\frac{\boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{vj})}{\eta_j})$, $i = 1.2 \dots n$ *and* $j = 1.2 \dots . c$

$\eta_j = \min_k \| \boldsymbol{v_j} - \boldsymbol{v_k} \|^2\ for\ j \ne k,$

3. Update cluster center : $\boldsymbol{V^t} = \{\boldsymbol{v_1}.\boldsymbol{v_2}, \dots \boldsymbol{v_c}\}$

$\boldsymbol{v_j} = \boldsymbol{v_j} + \alpha_t(\boldsymbol{x_i} - \boldsymbol{v_j})u_{ij}$,$i = 1.2 \dots n$ *and* $j = 1.2 \dots . c$

4. If $\|\boldsymbol{V^t} - \boldsymbol{V^{t-1}}\| \leq \varepsilon$ stop else go to 1

D. Confusion Matrix

Evaluating the performance of a classifier is important. The confusion matrix (Table I) is a useful tool for analysing as much as a good clustering method recognises the tuples from different classes.

TABLE I
CONFUSION MATRIX

		Prediction	
		Thalassemia	Non-thalassemia
Observed	Thalassemia	Correct Thalassemia (True Positive)	Type II Error (False Negative)
	Non-thalassemia	Type I Error (False Positive)	Correct Non-Thalassemia (True Negative)

The definition of true positive, true negative, false positive and false negative is as listed in Table II.

TABLE II
DEFINITION OF TRUE POSITIVE, TRUE NEGATIVE, FALSE POSITIVE, AND FALSE NEGATIVE

	Definition
True Positive (TP)	The number of thalassemia tuples which are correctly classified
True Negative (TN)	The number of non-thalassemia tuples which are correctly classified
False Positive (FP)	The number of non-thalassemia that are classified as thalassemia
False Negative (FN)	The number of non-thalassemia that are classified as thalassemia

A prediction of the accuracy of the formation of classification models can be obtained with the following formula [20]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (24)$$

E. Data

In this study, we used 150 data patients from year 2014 - 2018 taken from Harapan Kita Children and Womens's Hospital with 11 features, namely haemoglobin (g/dL), haematocrit (%), leukocyte count (10^3/µL), basophils (%), eosinophils (%), rod neutrophils (%), segment neutrophils (%), lymphocytes (%), monocytes (%), platelet counts (thousand / µL), and class types of patients. We divided the data into two, thalassemia and normal (non-thalassemia).
The data consisted of 82 thalassemia samples and 68 non-thalassemia samples.

III. RESULTS AND DISCUSSION

A. Experiment and Result

The results of the experiment are given below in tabular form. We use training data diverse from 10% (90% testing data) to 90% (10% training data).

TABLE III
THE PERFORMANCE OF FKCM AND FKRCM, WITH THE RBF KERNEL USING PARAMETER Σ

σ	Fuzzy Kernel C-Means		Fuzzy Kernel Robust C-Means	
	Accuracy (%)	Running Time (s)	Accuracy (%)	Running Time (s)
0.0001	67.80	0.28	98.31	0.50
0.001	96.61	0.19	98.31	0.47
0.05	94.92	0.16	98.31	0.44
0.1	94.92	0.17	98.31	0.42
1	94.92	0.14	98.31	0.44
5	94.92	0.14	98.31	0.44
10	94.92	0.16	98.31	0.42
50	94.92	0.16	98.31	0.44
100	94.92	0.16	98.31	0.44
1000	94.92	0.14	98.31	0.45

Table III shows the performance (accuracy and running time) of FKCM and FKRCM when the parameter of the RBF kernel (σ) is going from 0.0001 to 1000. We use different value of σ to find out the value of σ that produce the best accuracy and running time. The training data we used was 60% training data. In first experiment, the σ = 0.0001 is used. As the result, the accuracy of FKCM is smaller than accuracy obtained using FKRCM, but the running time of FKCM is faster.

We obtained the highest performance in both the FKCM and FKRCM methods when sigma (σ) was 0.001. The highest accuracy for both FKCM and FKRCM is 96.61% and 98.31%, respectively. We can conclude that the changes of parameter σ do not affect the accuracy of FKRCM classifier. Hence, we will use $\sigma = 0.001$ as the parameter of the RBF kernel to compare the accuracy of all methods. Table IV shows a comparison of the accuracy using FCM, FKCM, FRCM, and FKRCM classifier to the thalassemia data with kernel RBF and σ=0.001 as the parameter of the RBF kernel

TABLE IV
THE PERFORMANCE OF FCM, FKCM, FRCM AND FKRCM USING RBF KERNEL WITH Σ=0.001

Training Data (%)	Accuracy			
	Fuzzy C-Means	Fuzzy Kernel C-Means	Fuzzy Robust C-Means	Fuzzy Kernel Robust C-Means
10	73.88	85.82	89.55	93.28
20	78.99	91.60	91.60	94.12
30	66.35	80.77	89.42	95.19
40	91.01	84.27	94.38	97.75
50	90.67	64.00	93.33	98.67
60	91.53	**96.61**	94.92	98.31
70	72.73	72.73	97.73	97.73
80	68.97	75.86	93.10	**100.00**
90	**100.00**	64.29	**100.00**	**100.00**

As we can see, the highest accuracy of FKCM classifier is 96.61% when we use 60% training data, while the highest accuracy of FCM, FRCM, and FKRCM is 100%. Between the four fuzzy-classifier, FKRCM is the only one that can perform perfectly, that is when we use 80% and 90% training data. The graph of the accuracy of fuzzy c-means, fuzzy kernel c-means, fuzzy robust c-means, and fuzzy kernel robust c-means to data training is given below in Figure 3.

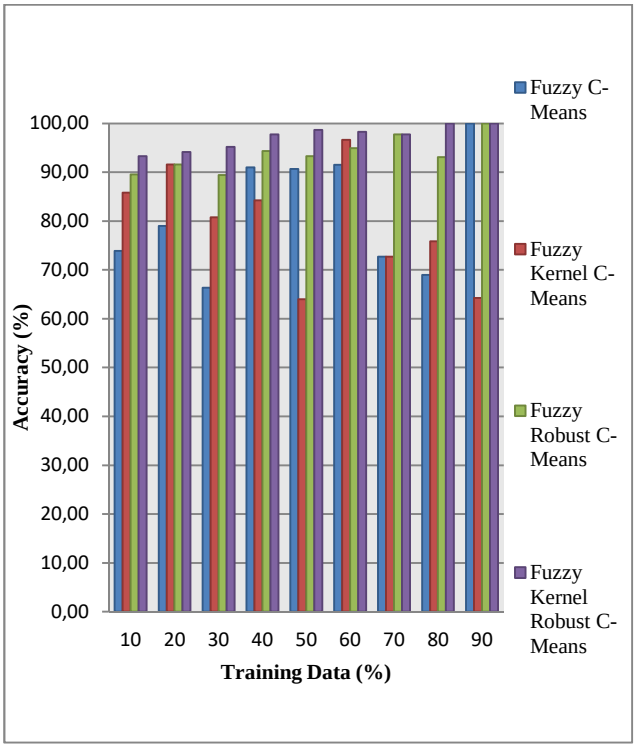

Fig. 3 Comparison of Accuracy Using FCM, FKCM, FRCM, and FKRCM with the RBF Kernel Using Parameter σ=0.001

We also provide the graph of accuracy obtained using all of the method in Table V.Table V shows a comparison of running time using FCM, FKCM, FRCM, and FKRCM classifier to the thalassemia data with kernel RBF and σ=0.001 as the parameter of the RBF kernel. We can see that the increase of data training do not affect the running time of the program.

TABLE V
THE COMPARISON OF RUNNING TIME OF FCM, FKCM, FRCM AND FKRCM USING RBF KERNEL WITH Σ=0.001

Training Data (%)	Running Time (s)			
	Fuzzy C-Means	Fuzzy Kernel C-Means	Fuzzy Robust C-Means	Fuzzy Kernel Robust C-Means
10	0.41	0.16	0.19	0.23
20	0.61	0.16	0.11	0.34
30	0.80	0.08	0.19	0.41
40	1.05	0.09	0.17	0.50
50	1.30	0.11	0.19	0.47
60	1.55	0.14	0.25	0.53
70	1.84	0.13	0.20	0.48
80	2.05	0.14	0.23	0.33
90	2.27	3.56	0.23	0.16

The graph of the time needed for simulating the program for each classifier (fuzzy c-means, fuzzy kernel c-means, fuzzy robust c-means, and fuzzy kernel robust c-means) to data training is given below in Figure 4.

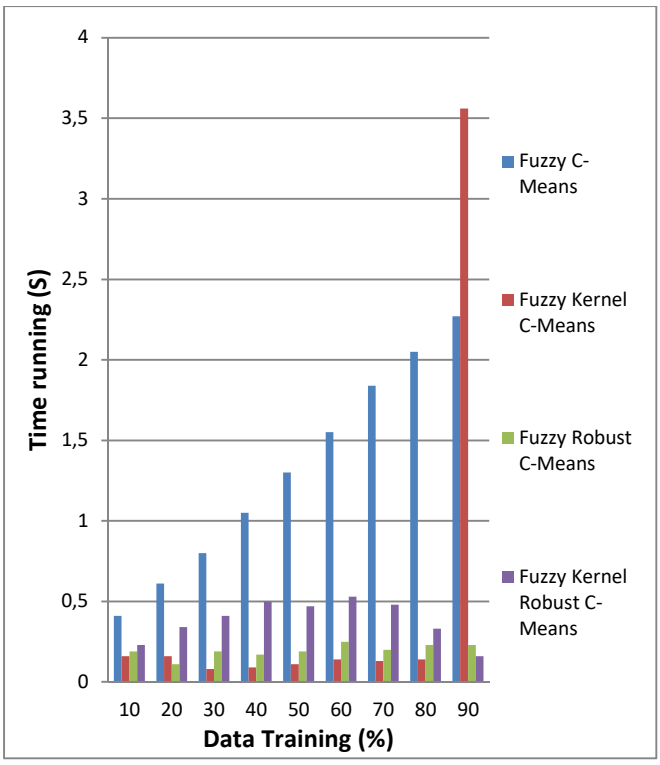

Fig. 4 Comparison of Running Time Using FCM, FKCM, FRCM, and FKRCM with the RBF Kernel Using Parameter σ=0.001

B. *Discussion*

Based on the result of the experiment that has been done, the results of classifying thalassemia data are diverse. The highest accuracy is obtained by fuzzy c-means, fuzzy robust c-means, and fuzzy kernel robust c-means. The performance of the fuzzy kernel robust c-means is the best among the other three classifier. We can say it is because of the robustness of the method and the proper use of sigma. In the FKRCM method, the changes of the sigma do not impact the accuracy, as in the FKCM method; the changes of sigma impact the performance, especially when we apply a small sigma. We hope this study can help with the screening of thalassemia in Indonesia.

IV. CONCLUSION

The purpose of our study is to compare the performance of three classification methods, namely Fuzzy C-Means (FCM), Fuzzy Kernel C-Means (FKCM), and Fuzzy Kernel Robust C-Means (FKRCM) to classify the thalassemia data from Harapan Kita Children and Womens's Hospital, Indonesia. In section 4, we can see the highest accuracy obtained from each method. We first compare the FKCM and FKRCM methods, using 60% of training data with different sigma increasing from 0.0001 to 1000 (table 3). The change of sigma did not change the accuracy of the FKRCM classifier, which was constant at 98.31%. On the other hand, there was a change in the FKCM classifier when the sigma was 0.0001 and 0.001. Here, the highest accuracy of FKCM was 96.61% when the sigma was 0.001. In table 3, it is obvious that fuzzy

kernel robust c-means is outperforming the fuzzy c-means, fuzzy robust c-means, and fuzzy kernel c-means. Its highest accuracy is 100% with 80% and 90% training data. The fuzzy c-means classifier classifies thalassemia perfectly when 90% training data is used, while the highest is 96.61% when 60% training data is used. We can conclude that the best classifier to classify thalassemia data is the fuzzy kernel robust c-means.

ACKNOWLEDGMENT

The University of Indonesia (DRPM-UI) with PITTA 2018 research grant scheme supported this research financially. Moreover, the authors want to thank the head of Harapan Kita Children and Womens's Hospital that allows the author to take the data.

REFERENCES

[1] The World Health Organization (WHO) website. [Online]. Available : https://www.who.int/genomics/public/geneticdiseases/en/index2.html

[2] M. I. Khan, H. N.Khan, and M. Usman, "Beta thalassemia trait; diagnostic importance of haematological indices in detecting beta thalassemia trait patients," *The Professional Medical Journal*, vol. 25, no.4, pp. 545-550, 2018.

[3] P. L. Greenberg, V. Gordeuk, S. Issaragrisil, N. Siritanaratkul, S. Fucharoen, and R. C. Ribeiro, "Major Hematologic Diseases in the Developing World— New Aspects of Diagnosis and Management of Thalassemia, Malarial Anemia, and Acute Leukemia," *American Society of Hematology*, pp. 479-498, 2001.

[4] M. Peters, H. Heijboer, and P. C. Giordano, "Diagnosis and management of thalassaemia", *BMJ*, vol. 7.

[5] X. Gu and Y. Zeng, "A Review of the Molecular Diagnosis of Thalassemia," Hematology, vol. 7, no. 4, pp. 203–209, 2002.

[6] S. R. Amendolia, G. Cossub, M. L. Ganaduc, B. Golosioa, G. L.Masala, and G. M. Mura, "A comparative study of K-Nearest Neighbour, Support Vector Machine and Multi-Layer Perceptron for Thalassemia screening," *Chemometrics and Intelligent Laboratory System*, vol. 69(1-2), pp. 13-20, 2003.

[7] S. R. Amendolia, A. Brunetti, P. Carta, G. Cossu, M. L. Ganadu, B. Golosio, G. M. Mura, M. G. Pirastru, "A Real-Time Classification System of Thalassemic Pathologies Based on Artificial Neural Networks," *Medical Decision Making*, pp. 18-26, 2002.

[8] P. Paokanta, M. Ceccarelli, and S. Srichairatanakool, "The Effeciency of Data Types for Classification Performance of Machine Learning Techniques for Screening β-Thalassemia," IEEE, 2010.

[9] A. S. AlAgha, H. Faris, B. H. Hammo, A. M. AlZoubi, "Identifying β-thalassemia carriers using a data mining approach: The case of the Gaza Strip, Palestine," *Artificial Intelligence in Medicine*, vol. 88, pp. 70-83, 2018.

[10] D. Setsirichok, T. Piroonratana, W. Wongseree, T. Usavanarong, N. Paulkhaolam, C. Kanjanakom, … , N. Chaiyaratana, "Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a Naïve Bayes classifier and a Multilayer Perceptron for Thalassaemia screening," *Biomedical Signal Processing and Control*, *vol. 7, No.* 2, pp. 202-212, 2012.

[11] P. Paokanta, N. Harnpornchai, and N. Chakpitak, "The Classification Performance of Binomial Logistic Regression Based on Classical and Bayesian Statistics for Screening β-Thalassemia", in *The 3rd International Conference on Data Mining and Intelligent Information Technology Applications*, 2011, pp. 236-241.

[12] D.A. Puspitasari, Z. Rustam, "Application of SVM-KNN using SVR as Feature Selection on Stock Analysis for Indonesian Stock Exchange," *Proceeding of 3rd International Symposium on Current Progress in Mathematics and Sciences*, 2017

[13] Z. Rustam, D.F. Vibranti, D. Widya, "Predicting The Direction of Indonesian Stock Price Movement using Support Vector Machines and Fuzzy Kernel C-Means," *Proceeding of 3rd International Symposium on Current Progress in Mathematics and Sciences*, 2017.

[14] Z. Rustam, Fanita, "Predicting The Jakarta Composite Index Price using ANFIS and Classifying Prediction Result Based on Relative Error by Fuzzy Kernel C-Means," *Proceeding of 3rd International Symposium on Current Progress in Mathematics and Sciences*, 2017.

[15] Z. Rustam and A.S. Talita, "Fuzzy Kernel C-Means Algorithm for Intrusion Detection Systems," *Journal of Theoritical and Applied Information Technology*, vol. 81, 2015.

[16] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, NewYork, 1981.

[17] Z. Rustam and D. Zahras, "Comparison between Support Vector Machine and Fuzzy C-Means as Classifier for Intrusion Detection System," in *2nd International Conference on Statistics, Mathematics, Teaching, and Research*, 2018, pp. 1-6.

[18] Z. Rustam and A. S. Talita, "Fuzzy Kernel K-Medoids Algorithm for Multiclass Multidimensional Data Classification", *Journal of Theoretical and Applied Information Technology*, vol. 80, Issue 1, 2015.

[19] A. Wulan, V. M. Jannati, Z. Rustam, and A. F. Ahmad, "Application Kernel Modified Fuzzy C-Means for Gliomatosis Cerebri," *International Conference on Mathematics, Statistics, and Their Applications*, pp. 35–38, 2016.

[20] J. Han, M. Kamber, J. Pei, *Data mining concepts and techniques*. Waltham, Massachusetts: Morgan Kaufmann Publishers, 2012.

[21] N. B. Karayiannis and J. C. Bezdek, "An Integrated Approach to Fuzzy Learning Vector Quantization and Fuzzy C-Means Clustering", *IEEE Trans. Fuzzy Systems*, vol. 5, no. 4, pp. 622-628, 1997.

[22] Z. Rustam and F. Yaurita, "Insolvency Prediction in Insurance Companies Using Support Vector Machines and Fuzzy Kernel C-Means," in *2nd International Conference on Statistics, Mathematics, Teaching, and Research*, 2018, pp. 1-9.

[23] S. R. Kannan, M. Siva, S. Ramathilagam, and R. Devi, "Effective Kernel-Based Fuzzy Clustering Systems in Analyzing Cancer Database," *Data Enabled Discovery and Applications*, pp. 85–92, 2018.