

PAPER • OPEN ACCESS

Classification of Breast Cancer using Fast Fuzzy Clustering based on Kernel

To cite this article: Zuherman Rustam and Sri Hartini 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **546** 052067

View the [article online](#) for updates and enhancements.



IOP | ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the [collection](#) - download the first chapter of every title for free.

Classification of Breast Cancer using Fast Fuzzy Clustering based on Kernel

Zuherman Rustam^{1*}, Sri Hartini¹

¹Department of Mathematics, University of Indonesia, Depok 16424, Indonesia.

*Corresponding author: rustam@ui.ac.id

Abstract. Breast cancer is the second leading cause of death in women in the world. The classification is the initial process of executing patient treatment, which is important as it increases life expectancy as well as quality. In this paper, a new method is proposed based on kernel, which is modified from KC-Means: it combines K-Means, Fuzzy C-Means algorithm, and kernel function. The C-Means algorithm is applied on the centers of a fixed number of groups founded by K-Means, and the kernel function is expected to improve the accuracy of classification with its ability to separate data which cannot be separated linearly. We applied the proposed method on a dataset of 201 breast cancer and 85 non-breast cancer samples from the UC Irvine Machine Learning Repository. Results concluded that fast fuzzy clustering has an accuracy of 85.26%, but fast fuzzy clustering based on kernel is 89.74%, with a better running time on average than 90.95% with the same method.

Keywords: Breast cancer; KC-Means; kernel.

1. Introduction

One in four of all new cancer cases diagnosed in women worldwide is breast, contributing to 15% of the total female cancer deaths [1]. In 2019, an estimated 268.600 new cases of invasive breast cancer have already been diagnosed in the U.S [2]. Furthermore, one in eight women in the UK are diagnosed [3]. There are significant increases in breast cancer in several Asian countries in recent years, including China, Singapore, and Thailand which have incidence rates increasing by 3% to 4% every year [4].

Early diagnosis of breast cancer coupled with advanced treatment has a significant impact in improving the rates of survival [4]. After being diagnosed, patients will be tested to find out what stage the cancer has reached, to determine how serious it is and how best to treat it [5]. Correct and early treatment will have a big impact compared to a longer treatment delay time [6].

Because conventional diagnosis based on past experience and knowledge can lead to the wrong diagnosis, many researchers have started producing algorithms and models for getting a precise diagnosis [7]. There are many algorithms that can be used for breast cancer classification, including the clustering algorithms. This is a method of separating data into several clusters where every data in a cluster has a similar characteristic. K-means and fuzzy c-means (FCM) are the mostly used algorithms [8].

A lot of studies and experiments using k-means and fuzzy c-means algorithms have been done. Wisconsin data made a comparison of the two and found that FCM produces a better result in comparison to the k-means algorithm, although it has a higher computation time [8,9]. To minimize



these limitations, experiments were carried out to combine the two methods. Combination of k-means and FCM algorithm in raw mammograms imagery have successfully detected breast cancer as well as identified which stage it was at [10]. FCM Crow search algorithm combines FCM with crow search algorithm has also been successfully implemented in several breast cancer datasets. This method works efficiently and the number of parameters for crow search is very minimal, leading to a reduced complexity [11]. The hybrid of k-means and the extreme Learning Machine can classify the mammograms image of breast cancer in a shorter period of time and with greater accuracy [12].

Kernel function is used to separate data which cannot be done linearly. It has been used with fuzzy c-means clustering to sort datasets of lung cancer, IRIS, Wine, Checkerboard, Time Series, Yeast, and breast cancer [13]. Furthermore, fuzzy kernel c-means has been used to solve Intrusion Detection Systems (IDS) with feature selection using kernel matrix concept [14]. Weighted KM-SVM also uses a kernel function to diminish error rates in classification data of breast and kidney cancer [15]. Meanwhile, fuzzy k-medoids has been used kernel function to sort datasets of breast cancer Wisconsin, diabetes, German, image segmentation, sonar, thyroid, IRIS, wine, and soybean [16]. In this study, the method proposed incorporates KC-means based on kernel (fast fuzzy clustering) using K-means, Fuzzy C-means algorithm, and kernel function.

2. Materials and Methods

2.1. Materials

The dataset was taken from UC Irvine Machine Learning Repository from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia, and was provided by M. Zwitter and M. Soklic [17]. This dataset consists of 201 breast cancer and 85 non-breast cancer samples. The instances have categorical characteristic and are described by 9 attributes, both linear and nominal. The dataset attributes are shown in Table 1 [18].

2.2. Methods

2.2.1. K-means algorithm. K-means algorithm is one of the simpler and important forms of clustering [19]. It was introduced by Lloyd in 1982 as a probabilistic technique to find clusters in a set of data points. Given a data set of n observations $\{x_1, x_2, \dots, x_n\}$ and c centers of the clusters $\{v_1, v_2, \dots, v_c\}$. Suppose $r_{ik} \in \{0,1\}$ where $k = 1, 2, \dots, c$ describing which of the c clusters the data point x_i is assigned to. In K-means algorithm, the goal is to partition the data set into some number c of the cluster by finding values of the $\{r_{ij}\}$ and the $\{v_j\}$ such that the objective function given by equation (1) is minimum [20].

$$J = \sum_{i=1}^n \sum_{j=1}^c r_{ij} \|x_i - v_j\|^2 \quad (1)$$

where

$$r_{ik} = \begin{cases} 1 & , \text{if } k = \arg \min \|x_i - v_j\|^2 \\ 0 & , \text{otherwise} \end{cases} \quad (2)$$

The algorithm of K-means is given as Fig.1 [21].

1. Let T be the maximum number of iterations allowed, $0 < \varepsilon < 1$. Let $v_j^{(t)}$, $t = 0$ be the initial centers, $j = 1, 2, \dots, c$, and $V^{(t)}$ be the set of centers in the iteration t ;
2. Compute the value of $\|x_i - v_j\|$;
3. Assign the elements x_i to the clusters according to
$$r_{ik} = \begin{cases} 1 & , \text{if } k = \arg \min \|x_i - v_j\|^2 \\ 0 & , \text{otherwise} \end{cases}$$
4. Update the cluster centers by take $t = t + 1$ and
$$v_j^{(t)} = \frac{\sum_{i=1}^n r_{ik} x_i}{\sum_{i=1}^n r_{ik}}$$
5. If $\|V^{(t-1)} - V^{(t)}\| < \varepsilon$ or $T = t$, then the iteration stops. Otherwise, go back to step (2); End.

Figure 1. K-means algorithm.**Table 1.** Breast Cancer Dataset Attribute

Attribute Name	Description
Age	Patient's age in years
Menopause	The period in a woman's life when menstruation ceases
Tumor-size	Patient's tumor-size on the breast
Inv-nodes	Node-size in main portion of the breast
Node-caps	Node is present or not in cap of the breast
Deg-malig	Stage of breast cancer
Breast	Left, right, or both breasts.
Breast-quad	Portion of the breast, i.e left up, left low, right-up, right-low, central
Irradiate	Present or not (YES/NO)
Class	No-recurrence-events, recurrence-events (reduce the risk of breast cancer)

2.2.2. Fuzzy c-means algorithm. Fuzzy c-means is a clustering technique where data belongs to one or more cluster depending on the degree in each [22]. It is different with K-means where every data belongs to one cluster only [20]. Fuzzy c-means was originally founded by Dunn and then modified by Besdek. The goal is to partition the data set into some number c of the cluster. Given a data set of n elements $\{x_1, x_2, \dots, x_n\}$ and c centers of the clusters $\{v_1, v_2, \dots, v_c\}$. Suppose $u_{ij} \in [0, 1]$ is a membership or degree of the data point x_i in j^{th} -cluster which satisfies equation (3) and (4) [23].

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

and

$$0 < \sum_{i=1}^n u_{ij} < n, \quad j = 1, 2, \dots, c \quad (4)$$

where $c < n$ is a positive integer. The objective function of fuzzy c-means algorithm is defined in equation (5) [23].

$$J_m = \sum_{i=1}^n \sum_{j=1}^c (u_{ij})^m (d(x_i, v_j))^2 \quad (5)$$

This function must attain the minimum value so that fuzzy c-partition will be found [23], as shown in Figure 2. [21].

1. Let T be the maximum number of iterations allowed, $0 < \varepsilon < 1$. Let $v_j^{(t)}$, $t = 0$ be the initial centers, $j = 1, 2, \dots, c$, and $V^{(t)}$ be the set of centers in the iteration t ;
2. Compute the value of $\|x_i - v_j\|$;
3. Compute the membership and assign the elements x_i to the clusters according to

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{\|x_i - v_j\|}{\|x_i - v_k\|} \right)^{\frac{2}{m-1}}} \quad \text{where } m \neq 1$$
4. Update the cluster centers according to

$$v_j = \frac{\sum_{i=1}^n (u_{ij})^m x_i}{\sum_{i=1}^n (u_{ij})^m}, \quad j = 1, 2, \dots, c$$
5. If $\|V^{(t-1)} - V^{(t)}\| < \varepsilon$ or $T = t$, then the iteration stops. Otherwise, go back to step (2);
- End.

Figure 2. Fuzzy c-means algorithm.

2.2.3. *Fast Fuzzy Clustering.* KC-means clustering is a combination of the k-means and fuzzy c-means method. This combined method is also called fast fuzzy clustering because it makes fuzzy c-means work faster with the accuracy close to only its method. The KC means is given in Figure 3 [21].

1. Apply k-means algorithm on the dataset;
2. Let V be the set of final centers of clusters which was obtained from step 1;
3. Suppose V as a new data set;
4. Apply fuzzy c-means algorithm on V ;
5. Regain corresponding data set clusters based on fuzzy c-means clustering output;
6. End.

Figure 3. KC-means algorithm.

2.2.4. *Kernel function.* In general, real-world issues are far more complex. It cannot always be resolved using linear functions. Kernel offers an alternative solution by projecting the data into a high dimensional feature space so that it can solve non-linear problems. With that ability, this function can separate data that cannot be separated linearly [25]. The kernel function for every $x \in \mathbb{R}^n$ is given by [20]

$$K(x, y) = \langle \phi(x), \phi(y) \rangle \quad (6)$$

where $\phi(x)$ is the feature space mapped by input data x . The distance between $\phi(x)$ and $\phi(y)$ in feature space can be defined as [20]

$$(d(x, y))^2 = K(x, x) - 2K(x, y) + K(y, y) \quad (7)$$

There are several kernel functions [24]

1. Linear kernel function:

$$K(x, y) = x \cdot y \quad (8)$$

2. Polynomial kernel function:

$$K(x, y) = (x \cdot y + 1)^h \quad (9)$$

3. Gaussian radial basis function kernel function:

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \quad (10)$$

2.2.5. *Fast fuzzy clustering based on kernel.* This paper proposes fast fuzzy clustering based on kernel. This method is a modified KC-means method that has been explained in section 2.2.3. All three kernel functions mentioned in section 2.2.4 is used to obtain the representation of structure of data as best as possible. The method is given below (Figure 4.)

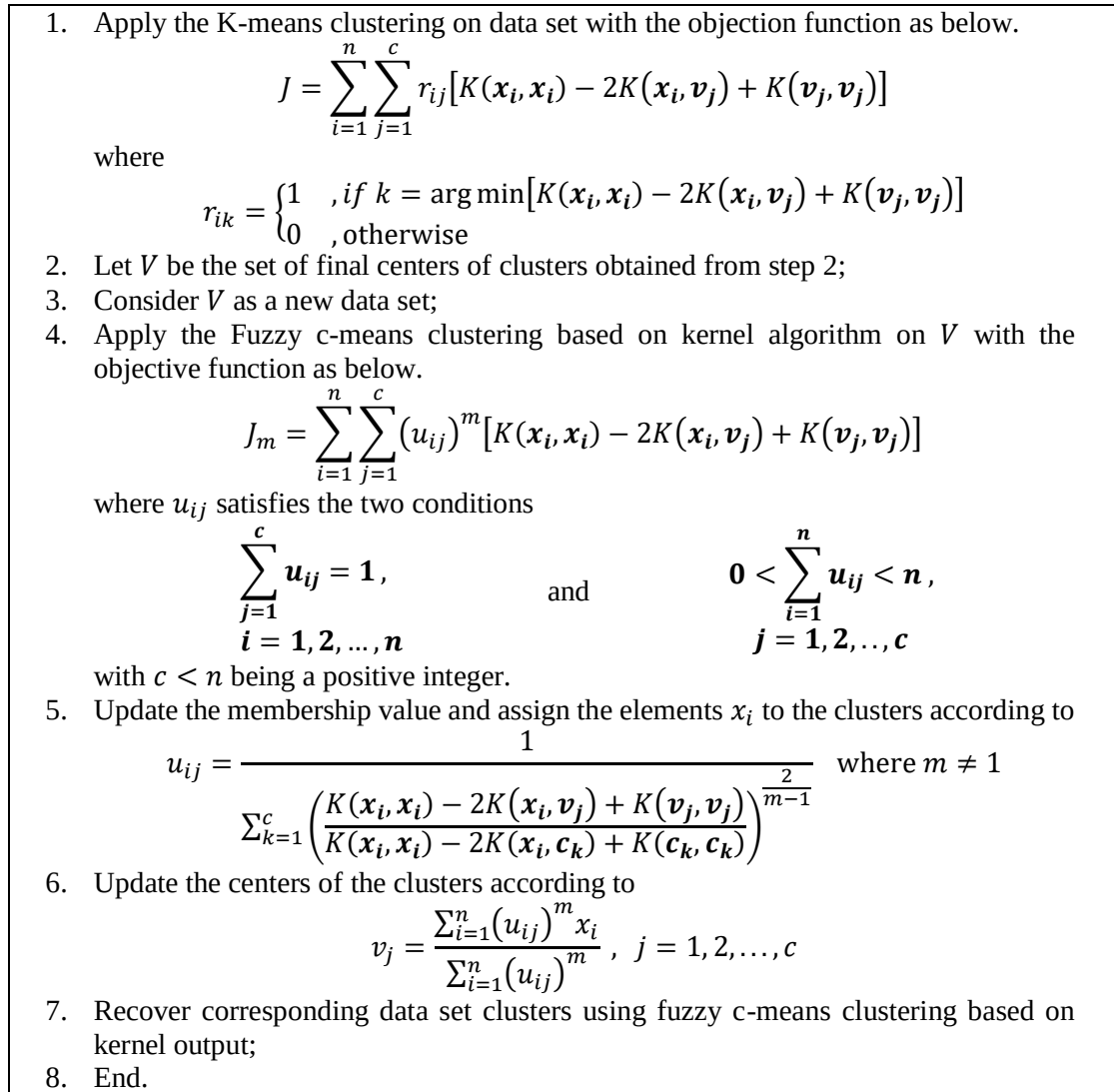


Figure 4. KC-means based on kernel algorithm.

2.2.6. *Performance Measures.* The confusion matrix is used to estimate the performance of a model and shows the classified and misclassified rate of the system. The performance of the methods in this paper is compared by the accuracy as computed below.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (11)$$

where

TP: the number of breast cancer samples which are correctly diagnosed;

FP: the number of non-breast cancer samples which are incorrectly diagnosed;

TN: the number of non-breast cancer samples which are correctly diagnosed;

FN: the number of breast cancer samples which are incorrectly diagnosed [26].

3. Result and Discussions

The performance of fast fuzzy clustering with or without kernel function was examined. These methods are conducted with the initial data set as well as training data which start from 10 to 90 percent. The performance of fast fuzzy clustering is shown in Table 2.

Table 2. Performance Measures of Fast Fuzzy Clustering using Euclidean Distance.

Percentage of Training Data	Accuracy	Running Time
10	78.83	0.09
20	80.04	0.14
30	79.87	0.09
40	85.57	0.20
50	87.68	0.61
60	91.18	0.47
70	91.67	0.30
80	87.41	0.27
90	85.07	0.27

The average accuracy and running time of the fast-fuzzy clustering method is 85.26 percent and 0.27 seconds, respectively. Meanwhile, for that which is based on kernel, different parameters are examined using the same training data ranging from 10 to 90 percent. The result after we calculate the average of its accuracy and running time is shown in Table 3.

Table 3. Performance Measures of Fast Fuzzy Clustering based on Kernel

Kernel Parameter	Average of Accuracy	Average of Running Time
0.0001	65.33	7.77
0.001	78.51	7.76
0.05	89.74	7.74
0.1	90.17	7.76
1	90.55	7.94
5	90.25	7.75
10	90.42	7.76
50	90.38	7.83
100	90.71	7.78
1000	90.95	7.81

We can see from Table 3 that the average value of accuracy and the running time have variations according to kernel parameters used, shown in Figure 5 and 6 below.

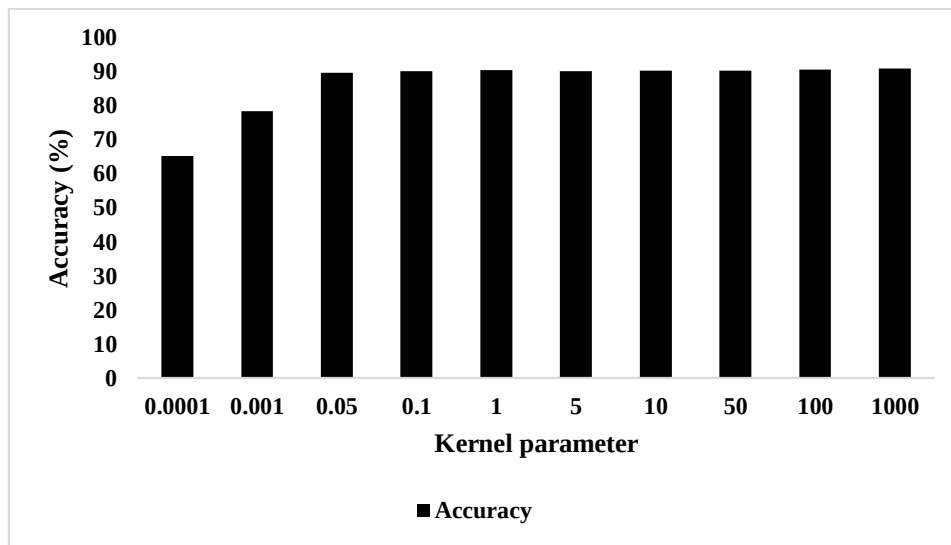


Figure 5. The average of running time from fast fuzzy clustering based on kernel with different kernel parameter

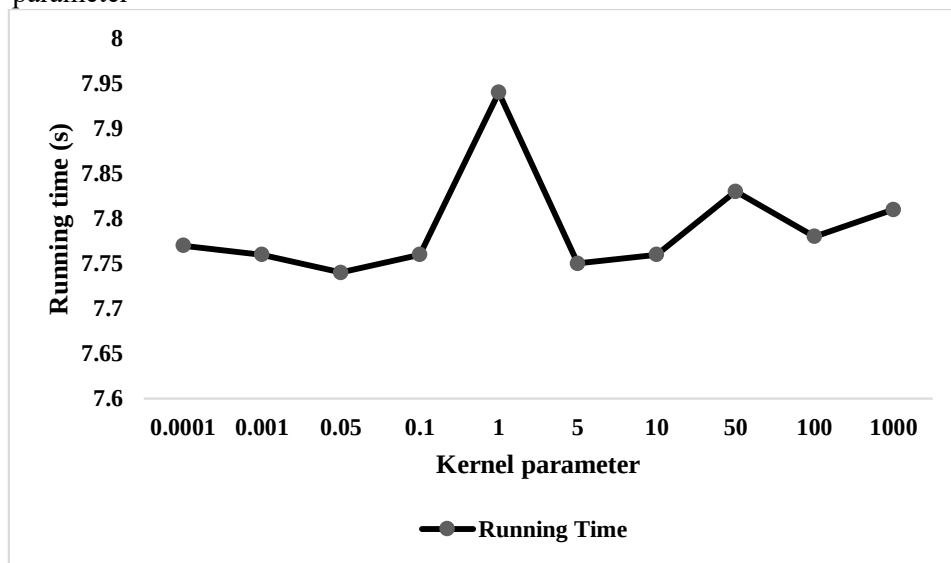


Figure 6. The average of running time from fast fuzzy clustering based on kernel with different kernel parameter

The highest average value for accuracy (90.95 percent) is obtained when kernel parameter $\sigma = 1000$ which resulted after an average of 7.81 seconds of running time. This performance is shown in Table 4. With this kernel parameter, accuracy can reach 95.52 percent in 2.23 seconds if 90% from training data is taken.

Table 4. Performance Measures of Fast Fuzzy Clustering using Kernel Parameter $\sigma = 1000$

Percentage of Training Data	Accuracy	Running Time
10	88.27	10.91
20	89.74	10.75
30	89.31	10.30

40	88.02	9.56
50	91.50	8.77
60	91.91	7.52
70	93.14	6.00
80	91.11	4.25
90	95.52	2.23

Meanwhile, the best average running time (7.74 seconds) is obtained when kernel parameter $\sigma = 0.05$ where the accuracy is 89.74%. These two results are better in accuracy compared to fast fuzzy clustering without using the kernel function, but still need improvement in the running time performance.

4. Conclusion

Classification of breast cancer is important in providing a more accurate diagnosis for the patient. Many methods were developed to achieve great accuracy and fast running time: this study introduced a new approach to breast cancer diagnosis using fast fuzzy clustering based on kernel, modified from the KC-means. This method was applied to the data set from UC Irvine machine learning repository, containing 201 breast cancer and 85 non-breast cancer samples.

The comparison in fast fuzzy clustering with or without kernel found that KC-means provides approximately 85.26% accuracy. Meanwhile, KC-means based on kernel provide 89.74%, with a better running time than 90.95% accuracy using the same method.

For future research, it is possible to develop a new model or method to increase the running time as well as the accuracy so that this classification can be used for better diagnosis. Others kernel parameters can also be used to get a higher level of accuracy for small data, although the running time is slower than without using kernel.

References

- [1] International Agency for Research on Cancer. (2018). Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018 [Press release]. 12 September. Available at: https://www.iarc.fr/wp-content/uploads/2018/09/pr263_E.pdf (Accessed: 21 February 2019)
- [2] Breast Cancer Statistics. Accessed: 21 February 2019. <https://www.breastcancer.org/>
- [3] Breast Cancer Care. Accessed: 21 February 2019. <https://www.breastcancercare.org.uk/>
- [4] Bhatia, H. (2016). 'Breast cancer in Asia', *GenRe*. p. 2,3. Available at: www.genre.com/knowledge/publications/ri16-4-en.html
- [5] American Cancer Society. Accessed: 21 February 2019 <https://www.cancer.org/>
- [6] Smith, E.C., Ziogas, A., & Anton-Culver, H. (2013). Delay in surgical treatment and survival after breast cancer diagnosis in young women by race/ethnicity. *JAMA Surgery*, 148(6), 516-523. doi: 10.1001/jamasurg.2013.1680.
- [7] Muhic, I. (2013). Fuzzy analysis of breast cancer disease using fuzzy c-means and pattern recognition. *Southeast Europe Journal of Soft Computing*. doi: 10.21533/scjournal.v2i1.45.
- [8] Dubey, A.K., Gupta, U., & Jain, S. (2018). Comparative study of k-means and fuzzy c-means algorithms on the breast cancer data. *International Journal on Advanced Science Engineering Information Technology*, 8, 18-29. doi: 10.18517/ijaseit.8.1.3490.
- [9] Sathish, A., Sundaram, J.M. (2014). A comparative study on k-means and fuzzy c-means algorithm for breast cancer analysis. *International Journal of Computational Intelligence and Informatics*, 4(1), 54-58.
- [10] Sable, S.D., Deshmukh, A. (2015). Implementation and analysis of k-means and fuzzy c-means clustering techniques. *International Journal of Industrial Electronics and Electrical Engineering*, 3(6), 173-176.
- [11] Parvathavarthini, S., Visalakshi, N.K., Shanthi, S., & Mohan, J.M. (2018). Crow search optimization based fuzzy c-means clustering for optimal centroid initialization. *TAGA Journal of*

- Graphic Technology*, 14, 3034-3045.
- [12] Chidambaranathan, S. (2016). Breast cancer diagnosis based on feature extraction by hybrid of k-means and extreme learning machine algorithms. *ARPJ Journal of Engineering and Applied Sciences*, 11(7), 4581-4586.
 - [13] Kannan, S.R., Ramthilagam, S., Devi, R., Hong, T.P. (2014). Fuzzy c-means in finding subtypes of cancers in cancer database. *Journal of Innovative Optical Health Sciences*, 7(1), 1-20. doi: 10.1142/S1793545814500187.
 - [14] Rustam, Z., Talita, A.S. (2015). Fuzzy kernel c-means algorithm for intrusion detection systems. *Journal of Theoretical and Applied Information Technology*, 81(1), 161-165.
 - [15] Kim, S. (2016). Weighted k-means support vector machine for cancer prediction. *SpringerPlus*, 5(1), 1162. doi: 10.1186/s40064-016-2677-4
 - [16] Rustam, Z., Talita, A.S. (2015). Fuzzy kernel k-medoids algorithm for multiclass multidimensional data classification. *Journal of Theoretical and Applied Information Technology*, 80(1), 147-151.
 - [17] UCI Machine Learning Repository. Accessed: 10 February 2019. <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer>
 - [18] Joshi, J., Doshi, R., & Patel, J. (2014). Diagnosis of breast cancer using clustering data mining approach. *International Journal of Computer Applications*, 101(10), 13-17. doi: 10.5120/17722-7611
 - [19] Dubey, A.K., Gupta, U., & Jain, S. (2016). Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. *International Journal of Computer Assisted Radiology and Surgery*, 11(11), 2033-2047. doi: 10.1007/s11548-016-1437-9
 - [20] Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.
 - [21] Atiyah, I.A., Mohammadpour, A., Taheri, S.M. (2018). KC-means: a fast fuzzy clustering. *Hindawi: Advances in Fuzzy Systems*, 2018, 1-8. doi: 10.1155/2018/2634861
 - [22] Saxena, N., Kumar, M., (2018). A comprehensive study on data clustering for breast cancer prognosis and risk exposure. *International Journal of Pure and Applied Mathematics*, 118(24), 1-17.
 - [23] Bezdek, J.C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press.
 - [24] Liu L., Shen B., Wang X. (2014) *Research on Kernel Function of Support Vector Machine*. In: Huang YM., Chao HC., Deng DJ., Park J. (eds) *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*. Lecture Notes in Electrical Engineering, 260, 827-834. Springer, Dordrecht. doi: 10.1007/978-94-007-7262-5_93
 - [25] Cristianini, N., Shawe-Taylor, J. (2014). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press.
 - [26] Rahideh, A., Shaheed, M.H. (2011). "Cancer classification using clustering based gene selection and artificial neural networks." in *Proceedings of the 2nd International Conference on Control, Instrumentation, and Automation* (pp.1175-1180), Shiraz, New Jersey: IEEE. doi: 10.1109/ICCIAutom.2011.6356828