International Journal on
Advanced Science
Engineering
Information Technology

Vol.8 (2018) No. 1
ISSN: 2088-5334

Comparative Study of K-means and Fuzzy C-means Algorithms on The Breast Cancer Data

Ashutosh Kumar Dubey[#], Umesh Gupta[#], Sonal Jain[#]

[#] *Institute of Engineering and Technology, JK Lakshmipat University,*
Near Mahindra SEZ, Ajmer Road, Jaipur - 302 026, (Rajasthan) India
E-mail: ashutoshdubey123@gmail.com

Abstract— **Breast cancer is one of the most common forms of cancer having a worldwide prevalence. Continuous research is going on for detecting breast cancer in its early stage as the possibility of cure is very high in the early stage. The two main objectives of this work were: firstly, to compare the performance of k-means and fuzzy c-means (FCM) clustering algorithms; and secondly, to make an attempt to carefully consider and examine, from multiple points of view, the combination of different computational measures for k-means and FCM algorithms for a potential to achieve better clustering accuracy. K-means and FCM algorithms have been considered to understand the impact of clustering on the breast cancer data. The execution of k-means algorithm is based on centroid, distance, split method, threshold, epoch, attributes, and number of iterations; while FCM is executed on the basis of fuzziness value and termination condition. The breast cancer Wisconsin (BCW) dataset was used for the experimentation and the comparison. The combination of variance and same centroid offers better outcome in terms of k-means algorithm. The highest and lowest clustering accuracies are (94.7%, 77.1%) and (94.4%, 88.5%) for foggy and random centroid, respectively. The overall average positive prediction accuracy obtained by this approach is approximately 92%. In case of FCM, the highest and lowest clustering accuracies are (97.2%, 91.1%), (97.2%, 90.9%), (97.8%, 90.4%), and (97.1%, 90.2%) for different combination of fuzziness and termination criteria. The average highest and lowest clustering accuracies are (95.7%, 94.7%), (95.9%, 93.6%), (95.3%, 94.2%), and (95.6%, 93.7%) for the same combination in the case of FCM. K-means algorithm was more prominent and consistent in terms of computation time as FCM required more time to carry out several fuzzy calculations and iterations. The findings of this work provide an incisive and extensive understanding of the computational parameters used with k-means and FCM algorithms. The computational results indicate that FCM algorithm was found to be prominent and consistent than k-means algorithm when executed with different iterations, fuzziness values, and termination criteria. It is more potentially capable in clustering BCW dataset as the clustering accuracy is more important than time.**

Keywords— **Breast cancer; breast cancer Wisconsin dataset; k-means; fuzzy c-means.**

I. INTRODUCTION

Breast cancer is the most frequently diagnosed cancer in women [1]. There are 458,000 deaths per year from breast cancer worldwide, making it the most common cause of female cancer with a high mortality in both the developed and developing countries [1, 2].

An early diagnosis of the breast cancer can be helpful as the chances of a complete cure are high [3, 4]. Its symptoms may vary according to the conditions, so the features identification is important. In this regard, the pattern detection is very important so that even hidden patterns can be identified correctly. Data mining techniques are capable in identifying the hidden patterns [4]; and can efficiently be used in classification, estimation, prediction, association rules, clustering, and visualization [5].

Prediction, classification, and estimation are included in the supervised learning category. In these techniques, clustering is important for data grouping as it is capable to cluster the data based on the property or symptom of the disease. K-means or hard c-means and fuzzy c-means are the mostly used clustering algorithms. The main benefit of k-means algorithm is that if the k (number of cluster) is small then the achieved computational speed is high even for large variables. The use of k-means algorithm is increasing day by day in the field of medical research because of its better clustering capabilities. It is basically a partitioning method applied to analyze data and to treat the observations of the data as objects based on locations and distance between the various input data points [6]. Mary et al. [7] also used k-means algorithm for cluster point refinement and used ant colony optimization (ACO) for cluster quality improvement. Wang et al. [8] formulated a clustering method named "molecular regularized consensus patient stratification (MRCPS)", which was based on optimization process. The main benefit of this method is its capability to cluster both the numerical and categorical data. Dubey et al. [9] presented extensive experimentation with k-means clustering on the BCW dataset and found that the k-means algorithm is capable in the classification of this dataset. Bhardwaj et al. [10] proposed improved k-means for increasing the cluster quality of complex datasets. Lu et al. [11] presented a methodology for health data analytics for modelling cancer patient records and suggested data mining techniques for large and heterogeneous clinical datasets. Fuzzy c-means (FCM) clustering is an extension of hard c-means clustering method. FCM is a clustering algorithm, which is applied in the areas of feature analysis and clustering [6]. Rahideh et al. [12] presented a classification approach, which is based on k-means and fuzzy c-means algorithm. It offers better accuracy, sensitivity, and specificity than that of the non-clustering method. Festa et al. [13] proposed a biased random-key genetic algorithm for data clustering, which was found to be useful in comparison to the other related methods. Chen et al. [14] proposed a hybrid intelligent model, which was used to analyze the clinical breast cancer data. It was found to be efficient in feature selection. Wei et al. [15] proposed a novel clustering algorithm; this method showed a greater efficiency in DNA sequence classification and their relationship. Vanisri et al. [16] used k-means algorithm for clustering breast cancer data and fuzzy c-means for optimizing the system. The performance in terms of memory, process time, and fitness point was better in case of a simultaneous clustering scheme. Ahmad et al. [17] designed a new algorithm based on k-means clustering algorithm, which was suitable to work with mixed numerical and categorical features. It was more efficient in comparison to other clustering algorithms. Yin et al. [18] suggested two clustering algorithms: FCM and k-means algorithms for arterial input function (AIF). They compared the performance of these two clustering methods using both simulated and clinical data. Zainuddin et al. [19] proposed the variant of FCM algorithm for finding the relevant clusters. Zheng et al. [20] proposed k-means algorithm and support vector machine (K-SVM) algorithms to extract useful information and diagnose tumors. They used k-means algorithm to recognize the hidden patterns of the benign and malignant tumors separately and employed a support vector machine (SVM) to obtain the new classifier to differentiate the emerging tumors. Rachman et al. [21] applied Fisher's ratio on the selected informative features for creating new data and then applied fuzzy c-means for classification and achieved better results. Deepthi et al. [22] was evaluated integrated feature selection techniques with semi-supervised fuzzy c-means algorithm using gene expression datasets and found to be useful in the accurate prediction of disease subtypes. The main objective of this research paper was to evaluate the performance accuracy of k-means clustering algorithm and FCM on the BCW dataset.

II. Material and Methods

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

K-means and FCM clustering algorithms have been used in this study for breast cancer data analysis. BCW includes clinical cases from the University of Wisconsin Hospital [23]. This dataset consist of total 699 records. Attribute information is shown in Table 1.

TABLE I
Attributes Information [23]

S.no	Attribute	Domain
1	Sample code number (SCN)	Id number
2	Clump thickness (CT)	1 - 10
3	Uniformity of cell size (UCS)	1 – 10
4	Uniformity of cell shape (UCSh)	1 – 10
5	Marginal adhesion (MA)	1 – 10
6	Single epithelial cell size (SECS)	1 – 10
7	Bare nuclei (BN)	1 – 10
8	Bland chromatin (BC)	1 – 10
9	Normal nucleoli (NN)	1 – 10
10	Mitoses	1 - 10
11	Class	Benign(2), Malignant(4)

Initially only 20 rows are considered for the experimentation and explanation. K-means algorithm is applied first for the BCW data clustering according to the k-mean algorithm presented in [9].

The data obtained from the BCW dataset (rows 1-20) are shown in Table 2. The attributes values shown in Table 2 are in the scale of 1-10. The value one shows the normal state and the value ten shows the highest abnormal state. It means from 1-10 the abnormality is increases. First 20 records are analyzed for final calculations of foggy k-means algorithm. Random data points are selected as the initial centroid. As in our case k=2, so two clusters are taken. Initially, there was a need of two centroids. In the first step the application size was 2 as the dataset is partitioned in 10 divisions. In the first iteration the first two SCN are evaluated. Therefore, the first two records (1000025 and 1002945) were not considered for centroid calculation. Thus, the counting was started from SCN=1015425, and finally the total records are 18. First generated random numbers are 0.28 and 0.09. These numbers are generated based on the Java random class. The rows are selected for the centroid (C) calculation based on these random numbers according to the following equations:

$$Ci = Pi \text{ X (number of row-1) } +1 \qquad (1)$$

Where C_i denotes the i^{th} centroid and P_i is the random number for the i^{th} iteration.

Rows 6 (S. no. 8) and 2 (S. no. 4) were found based on the above calculation. For this experimentation, we considered the attribute number two, i.e., second column UCS. On the basis of this selection, the first centroid obtained was 1, 8.

The first centroid obtained is (1, 8), based on the initial centroid the next centroids are calculated in the k-means clustering. So if 1 is considered then “Before Centroid” value is 1 and the concerning attribute of respective SCN will decide the after centroid. It is clear from Table 3 that the “Before Centroid” value is 1 and the concerning attribute value that is second in our case is 8. The sum of “Before Centroid” and the attribute value produce the “After Centroid” value. For the next case it becomes the “Before Centroid” value. The terminology used in Table 3 is the sample code number (SCN), which denotes the code number rows from the Table 1, the “Before Centroid” denotes the previous centroid obtained and attribute shows the value of UCS attribute selected in this process. The “After Centroid” is the addition of the “Before Centroid” and the “Second Attribute”. Based on the “After Centroid”, initial centroid1 (CS1) from first epoch is calculated. Initial centroid2 (CS2) is calculated by the same process from the second cluster. CS1 and CS2 are shown in Table 4. Final centroid1 (CL1) and final centroid2 (CL2) are calculated at the last epoch by the same process as shown in Table 4. The process is continued till epoch-1 and the positive predictive value (PPV) is calculated (Table 4).

Epoch is the stopping condition, for example if the epoch is 8 then the PPV for the first cluster is calculated after 7th iteration and for the second cluster, the same iteration is followed for the calculation. Here PPV determines the positive prediction of clustering. It is calculated based on the true positives and false positives obtained from first cluster and second cluster. The same process is applied with the random k-means algorithm. In this process the centroid initialization is done though k random points. The value of k, application size, iteration and other computational attributes are kept same as above. In random approach, random numbers are not required since initial centroid is (0, 0), but the row selection is random. Then the next centroid is calculated according to the cluster row values as shown in Table 5. The same process is applied for the random centroid except the centroid initialization. CS1 is calculated based on the “After Centroid” value. The process is continued till epoch-1 and the PPV is calculated (Table 6).

TABLE II

SELECTED VALUES FROM THE WHOLE DATASET (1-20)

S. no	SCN	CT	UCS	UCSh	MA	SECS	BN	BC	NN	Mitoses	Class
1	1000025	5	1	1	1	2	1	3	1	1	2
2	1002945	5	4	4	5	7	10	3	2	1	2
3	1015425	3	1	1	1	2	2	3	1	1	2
4	1016277	6	8	8	1	3	4	3	7	1	2
5	1017023	4	1	1	3	2	1	3	1	1	2
6	1017122	8	10	10	8	7	10	9	7	1	4
7	1018099	1	1	1	1	2	10	3	1	1	2
8	1018561	2	1	2	1	2	1	3	1	1	2
9	1033078	2	1	1	1	2	1	1	1	5	2
10	1033078	4	2	1	1	2	1	2	1	1	2
11	1035283	1	1	1	1	1	1	3	1	1	2

TABLE III

INITIAL CENTROID CALCULATION BASED ON THE FIRST EPOCH (FOGGY CENTROID CALCULATION MECHANISM)

S. no	SCN	Before centroid	Attribute	After centroid
1	1016277,6,8,8,1,3,4,3,7,1,2	1	8	9
2	1017122,8,10,10,8,7,10,9,7,1,4	9	10	19
3	1044572,8,7,5,10,7,9,5,5,4,4	19	7	26
4	1050670,10,7,7,6,4,10,4,1,2,4	26	7	33

$$CS_1 = (1/C_i) \sum_{j=1}^{c_i} X_i \qquad (2)$$

TABLE IV

PPV RESULT BASED ON FOGGY CENTROID

S. no		Y1	P1	Y2	P2	K1	K2	CS1	CS2	CL1	CL2	PPV
1		5	0.29	1	0.09	1.00	8.00	8.25	2.00	8.50	2.03	1.00
2		10	0.59	2	0.15	3.00	1.00	1.27	5.42	2.04	8.54	1.00
3		1	0.07	2	0.12	4.00	1.00	1.41	5.66	1.71	6.33	1.00
4		2	0.15	0	0.04	1.00	1.00	3.55	2.77	8.38	2.15	1.00
5		16	0.95	15	0.90	7.00	1.00	1.75	6.83	2.17	8.51	1.00
6		15	0.90	4	0.26	1.00	1.00	3.44	2.88	8.38	2.15	1.00
7		5	0.31	15	0.93	10.00	1.00	2.06	8.66	2.00	8.99	1.00
8		5	0.34	9	0.55	10.00	2.00	2.35	8.50	2.25	8.56	1.00
9		16	0.95	14	0.83	1.00	7.00	6.80	1.76	9.01	2.19	1.00
10		5	0.29	1	0.09	1.00	8.00	8.25	2.00	8.50	2.03	1.00

Total number of records=20

Parameters: Euclidean distance, split-simple, epochs=4, attribute number=2, number of iterations=10

Y1 and Y2: row positions, P1 and P2: initial random number1 and 2, CS1 and CS2: initial centroid 1 and 2, CL1 and CL2: final centroid 1 and 2

TABLE V
INITIAL CENTROID CALCULATION BASED ON THE FIRST EPOCH (RANDOM CENTROID CALCULATION MECHANISM)

S. no	Before centroid	Attribute	After centroid
1	0	1	1
2	1	2	3
3	3	1	4
4	4	1	5
5	5	3	8
6	8	1	9
7	9	7	16
8	16	4	20
9	20	1	21
10	21	1	22
10	22	7	29
11	29	1	30

TABLE VI
PPV RESULT BASED ON RANDOM CENTROID

S. no	K1	K2	CS1	CS2	CL1	CL2	PPV
1	2.50	3.66	2.50	3.66	8.49	2.02	1.00
2	2.50	3.00	2.50	3.00	8.57	2.04	1.00
3	2.50	2.66	2.50	2.66	6.33	1.71	1.00
4	2.50	4.16	2.50	4.16	8.53	2.24	1.00
5	2.41	4.16	2.41	4.16	8.51	2.17	1.00
6	2.50	4.16	2.50	4.16	8.53	2.24	1.00
7	2.33	4.16	2.33	4.16	8.50	2.09	0.50
8	1.75	4.16	1.75	4.16	8.95	1.98	1.00
9	2.50	4.16	2.50	4.16	8.53	2.24	1.00
10	2.00	4.16	2.00	4.16	9.02	2.19	1.00

Total number of records=20
Parameters: Euclidean, split-simple, epochs=4, attribute number=2, number of iterations=10

Then FCM algorithm is applied on the same 20 items of the dataset. The algorithm is shown below for the data computation.

Notations used in algorithm:
C: Clusters
f: Fuzziness
D: Data points
M: Data dimensions
CC: Cluster center
ED: Euclidean distance

Algorithm: Fuzzy C-Means

Step 1: Let there be a data point D= {D1, D_2…….., D_n}. The dimension of the data is M which is to be clustered.
Step 2: Assume the number of clusters to be made is C, which is 2 in our case.
Step 3: Then the cluster fuzziness, f, is chosen, which should be greater than 1.
Step 4: Let the initial membership matrix, U, to be M X C X D. It should follow the following equation for each i and fixed value of m.

$$\sum_{j=1}^{C} U_{ijm} = 1.0 \qquad (3)$$

Step 5: Then degree of membership is calculated.

$$CC_{jm} = \frac{\sum_{i=1}^{D} U_{ijm}^{f} x_{id}}{\sum_{i=1}^{D} M_{ijd}^{f}} \qquad (4)$$

Step 6: Calculate the distance between the data point and the cluster center

$$ED_{ijm} = x_{im}-CC_{jm} \qquad (5)$$

Step 7: Update degree of membership

$$U_{ijm} = \frac{1}{\sum_{c=1}^{C} \left(\frac{ED_{ijm}}{ED_{icm}}\right)^{\frac{2}{f-1}}} \qquad (6)$$

Step 8: Finally the iteration is terminated based on the epsilon value and the final cluster centers are obtained.
Step 9: Final clustering results.

There are some terminologies that need to be explained for an understanding of this algorithm. First is the membership value, it shows the degree of an object corresponding to the fuzziness and it is between 0 and 1. Second important parameter is fuzziness which shows the degree of truth, it should be greater than 1, and in our case, it is 2. Third important parameter is the termination criterion or epsilon value, which should be between 0 and 1. In this approach the data points have their membership values with the cluster centers, which are updated iteratively until found lower than the epsilon value.

Table 2 shows the data from the BCW dataset (rows 1-20) considered for the analysis. According to the mentioned algorithm the degree of membership (DOM) is calculated first, and shown in Table 7. There are two sequences in each column of the first row. The second value of DOM is calculated based on the random number generated by Java random class and the first value is obtained by the subtraction from the second value by one. The power is calculated based on $DOM^{fuzziness}$. In this experimentation, the fuzziness value is considered as two. The data point considered here is the CT value from Table 2.

The cluster center is calculated for all the nine columns. It is evaluated first on the basis of the array sequences T[0,0],T[1,0],T[2,0]……………..T[19,0] and then secondly on the array sequence T[0,1], T[1,1], T[2,1]……………..T[19,1]. This means that for the first column, total 20 values are contributed in the calculation of the numerator and denominator. Based on the numerator and denominator cluster-center can be calculated. So for the first pass it is evaluated nine times and each time 20 values are contributed for finding the cluster-center. In the same way, in the second pass, it is evaluated nine times and each time 20 values are contributed for finding the cluster-center. The result of the two pass is shown in Table 8.

Based on the numerator and denominator the cluster center is calculated.

$$\text{Cluster-center} = \frac{\text{Numerator}}{\text{Denominator}} \qquad (7)$$

Subsequently for updating the degree of membership, the closeness of the data point to the center of the vector was calculated. It depends on the number of dimension. In our case the number of dimension is nine. The main purpose of this step is to stop the iteration when the data points are close enough to the center vector.

The formula is the same but the phases are different so the values retrieved are different for each iteration. If the difference value is greater than the maximum difference, the difference value is assigned to the maximum difference and if it is smaller than the termination criterion epsilon, the iteration is stopped, otherwise the same process is repeated unless a smaller value is obtained. This is shown in Table 9. In our case the max-difference obtained is 0.5401 and the termination criterion is 0.00005. As the termination criterion is small than the maximum difference, so the next iteration starts for finding the new degree of membership. Then the data are arranged by comparing the membership value of the position of T[0][0]-T[0][1], T[1][0]-T[1][1]…….. T[19][0]-T[19][1]. If the value of degree of membership of T[0,0] is higher than T[0,1] then the data point is added in first cluster otherwise in the second cluster. The same procedure is applied to all the values, and by this, PPV value is calculated as shown in Table 10.

TABLE VII
DEGREE OF MEMBERSHIP

S. no	Array sequence	DOM	Power	Data point
1	T[0][0] T[0][1]	0.99 0.01	0.99 1.63E-6	5
2	T[1][0] T[1][1]	0.99 0.00	0.98 3.44E-5	5
3	T[2][0] T[2][1]	0.99 0.03	0.99 1.55E-5	3
4	T[3][0] T[3][1]	0.99 0.00	0.98 3.69E-5	6
5	T[4][0] T[4][1]	0.99 0.00	0.98 8.64E-5	4
6	T[5][0] T[5][1]	0.99 0.00	0.98 8.40E-5	8
7	T[6][0] T[6][1]	0.99 0.07	0.98 6.34E-5	1
8	T[7][0] T[7][1]	0.99 0.01	0.99 7.75E-7	2
9	T[8][0] T[8][1]	0.99 0.008	0.993 3.386E-6	2
10	T[9][0] T[9][1]	0.99 0.00	0.98 3.63E-5	4
11	T[10][0] T[10][1]	0.99 0.00	0.99 3.49E-6	1
12	T[11][0] T[11][1]	0.99 0.02	0.98 6.08E-5	2
13	T[12][0] T[12][1]	0.99 0.00	0.98 6.85E-5	5
14	T[13][0] T[13][1]	0.99 0.00	0.98 9.86E-5	1
15	T[14][0] T[14][1]	0.99 0.00	0.98 4.49E-5	8
16	T[15][0] T[15][1]	0.99 0.00	0.98 5.68E-5	7
17	T[16][0] T[16][1]	0.99 0.00	0.98 8.96E-5	4
18	T[17][0] T[17][1]	0.99 0.00	0.98 4.81E-5	4
19	T[18][0] T[18][1]	0.99 0.00	0.98 2.90E-5	10
20	T[19][0] T[19][1]	0.99 0.00	0.99 1.15E-5	6

TABLE VIII
CLUSTER CENTERS

S. no	Data point	Numerator	Denominator	Cluster center
1	CT column wise (Table 2)	86.90	19.76	4.39
2		56.24	19.76	2.84
3		56.24	19.76	2.84
4		51.29	19.76	2.59
5		60.21	19.76	3.04
6		71.02	19.76	3.59
7		65.15	19.76	3.29
8		41.43	19.76	2.09
9		27.69	19.76	1.40
10		0.009	8.74E-4	4.48
11		0.002	8.74E-4	3.17
12		0.002	8.74E-4	3.16
13		0.002	8.74E-4	3.00
14		0.002	8.74E-4	3.29
15		0.003	8.74E-4	4.10
16		0.003	8.74E-4	3.63
17		0.002	8.74E-4	2.43
18		0.001	8.74E-4	1.20

TABLE IX
UPDATED DEGREE OF MEMBERSHIP

S. no	Array sequence	Data point	T1	T2	Sum	DOM	New_DOM
1	T[0][0]	CT row wise (Table 2)	4.3595	5.2546	1.6883	0.9987	0.5923
	T[0][1]		5.2546	5.2546	2.4527	0.0012	0.4076
2	T[1][0]		8.1069	7.3925	2.2026	0.9941	0.4540
	T[1][1]		7.3925	7.3925	1.8315	0.0058	0.5459
3	T[2][0]		4.0507	4.9311	1.6748	0.9960	0.5970
	T[2][1]		4.9311	4.9311	2.4818	0.0039	0.4029
4	T[3][0]		9.0936	8.6192	2.1131	0.9939	0.4732
	T[3][1]		8.6192	8.6192	1.8983	0.0060	0.5267
5	T[4][0]		4.0519	4.8516	1.6975	0.9907	0.5891
	T[4][1]		4.8516	4.8516	2.4336	0.0092	0.4108
6	T[5][0]		16.0597	15.1182	2.1284	0.9908	0.4698
	T[5][1]		15.1182	15.1182	1.8861	0.0091	0.5301
7	T[6][0]		8.0306	8.0312	1.9998	0.9920	0.5000
	T[6][1]		8.0312	8.0312	2.0001	0.0079	0.4999
8	T[7][0]		4.6583	5.4920	1.7194	0.9991	0.5815
	T[7][1]		5.4920	5.4920	2.3899	8.8084E-4	0.4184
9	T[8][0]		6.5092	7.3762	1.7787	0.9981	0.5621
	T[8][1]		7.3762	7.3762	2.2841	0.0018	0.4378
10	T[9][0]		4.2076	5.1461	1.6685	0.9939	0.5993
	T[9][1]		5.1461	5.1461	2.4958	0.0060	0.4006
11	T[10][0]		5.7691	6.5614	1.7730	0.9981	0.5639
	T[10][1]		6.5614	6.5614	2.2935	0.0018	0.4360
12	T[11][0]		5.0978	5.9802	1.7266	0.9921	0.5791
	T[11][1]		5.9802	5.9802	2.3761	0.0078	0.4208
13	T[12][0]		2.5100	2.4197	2.0760	0.9917	0.4816
	T[12][1]		2.4197	2.4197	1.9293	0.0082	0.5183
14	T[13][0]		4.8797	5.5696	1.7676	0.9900	0.5657
	T[13][1]		5.5696	5.5696	2.3027	0.0099	0.4342
15	T[14][0]		12.3538	11.5367	2.1466	0.9932	0.4658
	T[14][1]		11.5367	11.5367	1.8720	0.0067	0.5341
16	T[15][0]		6.0778	5.7865	2.1032	0.9924	0.4754
	T[15][1]		5.7865	5.7865	1.9064	0.0075	0.5245
17	T[16][0]		4.5162	5.4627	1.6834	0.9905	0.5940
	T[16][1]		5.4627	5.4627	2.4630	0.0094	0.4059
18	T[17][0]		4.3360	5.2509	1.6818	0.9930	0.5945
	T[17][1]		5.2509	5.2509	2.4664	0.0069	0.0069
19	T[18][0]		11.0215	10.3322	2.1378	0.9946	0.4677
	T[18][1]		10.3322	10.3322	1.8788	0.0053	0.5322
20	T[19][0]		4.6054	5.4452	1.7153	0.9966	0.5829
	T[19][1]		5.4452	5.4452	2.3979	0.0033	0.54012

TABLE X
PPV VALUES BASED ON FCM

S. no	Fuzziness value	Termination Criteria	PPV
1	2	2.0E-5	0.90
2	2	3.0E-5	0.75
3	2	4.0E-5	0.80
4	2	5.0E-5	0.90
5	2	6.0E-5	0.85

III. RESULTS AND DISCUSSION

In this section the results based on k-means and FCM algorithms are presented and discussed with different computational parameters. At first, we considered k-means algorithm for the result analysis. The results were calculated based on the total 699 records, and were analyzed on the basis of the centroid, distance, split method, threshold, epoch, BCW attribute, and number of iterations. The first parameter centroid is meant for foggy and random centroid. In case of foggy centroid, the first centroid is calculated based on the random values. In case of the random centroid, the initial centroid is considered as (0, 0). The second parameter is distance, the distance between cluster centers and the data points are compared by Euclidean distance algorithm. The third parameter is split method. This was considered because the main problem in cluster selection is to split it in divisive clustering. Fourth parameter is threshold, which shows the stopping condition of the loop.

Constant Epoch and same centroid are considered in the case of threshold. In constant epoch, the epoch values are decided variably. The epoch variations can be 4, 5, 6, 7, and 9. In case of same centroid, the process is stopped if the means do not change anymore. The fifth parameter is epoch, which is a part of the threshold for stopping the iteration. The sixth parameter is the data attribute, which decides the selection of any attribute from the nine attributes of calculation. The last parameter is number of iteration which determines the filtration of cluster center in each pass.

The results are compared with the help of positive predictive value (PPV). It is calculated as:

$$\text{PPV} = \frac{\text{number of true positive}}{\text{number of true positive} + \text{number of false positive}} \quad (8)$$

It is used for finding the accuracy of positive predictive value from the obtained clusters.

Where a "true positive" implies that the test results are making a positive prediction, and the related case has a positive result, and a "false positive" implies that the test results are making a positive prediction, and the related case has a negative result. The numbers of true positives are also denoted by sensitivity × prevalence and the numbers of false positives are denoted by (1 − specificity) × (1−prevalence). False discovery rate (FDR) was considered for the comparison of unclassified data between k-means and FCM process.

$$\text{FDR} = 1 - \text{PPV} \quad (9)$$

The computation measures used in k-means algorithm with their attribute values are shown in Table 11.

The PPV results based on the centroid variations (foggy/random), epoch variations, attribute selection, and iterations of the four cases are shown in Fig. 1. For the first case centroid variations, the only difference is in the selection of initial centroid. Random initialization is considered for foggy centroid, and zero coordinates are fixed in this case. Due to the initialization the variations is observed in case of foggy approach but not in the case of random approach. For the second case epoch variations, only stopping condition is determined, so no variation is observed in the case of epoch variations. In the third case attribute selection, it shows significant variations, as it participates differently in the breast cancer cause and so affects the PPV.

For the fourth case iterations, no variation is observed. As iterations only determine the partitions, so if the number of iterations is more, the partitions are also more. The variations are not due to the changes in the iteration but the iteration determines the partitions. The effects of k-means algorithm considering highest variance and same centroid with foggy and random centroid are shown in Fig. 2. In the case of the highest variance and the same centroid, the main benefit is that the process is stopped if the means do not change anymore. This provides an opportunity of finding nearer data points, as the process is not restricted to the epoch and the possibility of better cluster selection is improved.

The combination of the highest variance and the same centroid provides better results in comparison to any combinations for k-means algorithm. The comparison of PPV for all the cases discussed above is shown in Fig. 1 and Fig. 2. The highest and lowest clustering accuracies are (94.7%, 77.1%) and (94.4%, 88.5%) for foggy and random centroid, respectively. The highest PPVs are considered for each case. It shows average highest accuracy of 92% obtained overall by k-means algorithm. Computation time comparisons with different iterations using k-means algorithm are shown in Table 12. The partitions increase if the number of iterations is more, but in each partition the attributes are less so the computation time is reduced in the case of iterations with higher number.

TABLE XI
COMPUTATION MEASURES USED IN K-MEANS

S. no	Attributes	Measures / Values
1	Centroid	Foggy(F), Random(R)
2	Distance	Euclidean
3	Split	Simple(S) and variance(V)
4	Threshold	Constant epoch(CE) and same centroid (SC)
5	Epochs(E)	3,4,5,6,7,9 (3-10)
6	BCW attributes (A)	2,3,4,5,6,7,8,9 (1-9)
7	Number of iterations (I)	4,5,6,7,8,9,10(4-10)

Secondly FCM algorithm is considered for the analysis of results with the same dataset. The results are analyzed for the fuzziness value (FV) and termination criteria (TC). In this approach, first degree of membership is calculated on the basis of random numbers; then the FV is considered. It controls the sharing of the data among fuzzy clusters and may affect the results; therefore the selection of FV is important [24]. Several research works have been published in the direction of the selection of an appropriate FV. In an earlier study [24], FV is suggested to be between 1.1 and 5. A heuristic rule is also prescribed elsewhere for the selection of a FV that falls in the range of 1.5-2.5 [25]. They also suggested that the most optimal selection of FV is 2 [26]. Another study [27] suggested the range of FV as 2-3.5 and the optimal interval as 2.5-3 by using cluster validity index. In our approach, the optimal range considered is 2-5 based on the previous published literature. Then cluster centers are calculated. The degree of membership is updated for the data point in the clusters. A termination condition is applied as the iteration should be stopped at some point where the data points are nearer to the center vectors. The TCs are set between 0 and 1, mainly to control the number of iterations. To obtain a better refinement the range considered in this paper is 2.0E–5 to 6.0E–5. The comparison of PPV values considering different fuzziness values (FV1=2, FV2=3, FV3=4 and FV4=5) and termination criteria (TC1=2.0E-5, TC2=3.0E-5, TC3=4.0E-5, TC4=5.0E-5 and TC5=6.0E-5).

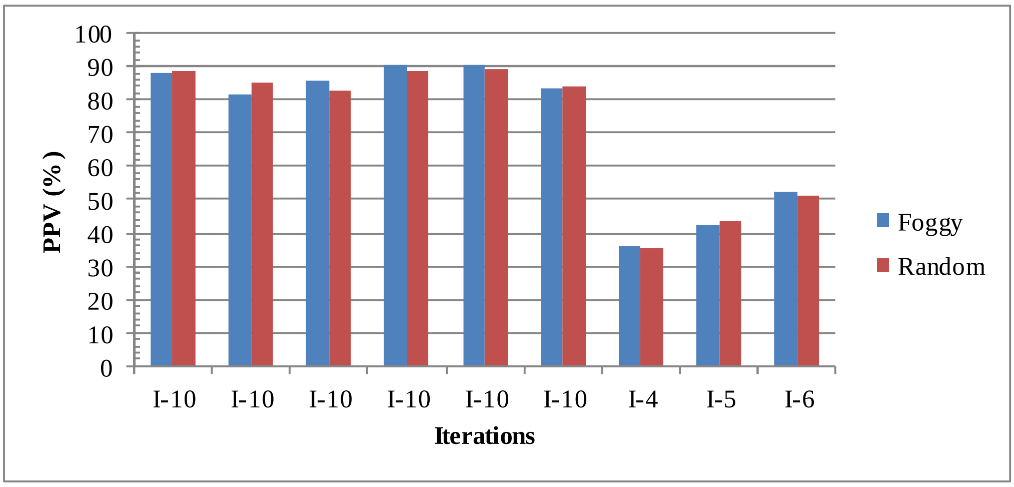

Fig. 1 Effect of k-means considering variable attributes and iterations with foggy and random centroid

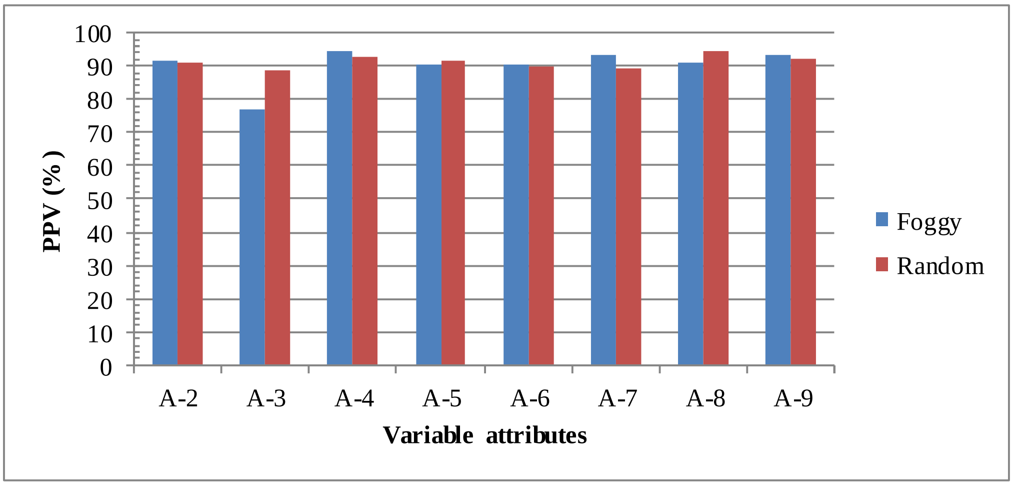

Fig. 2 Effects of k-means considering variance and same centroid with foggy and random approach

TABLE XII
TIME COMPARISON WITH DIFFERENT ITERATIONS USING K-MEANS

S. no	Iterations	Time (MS)	Time (MS)	Time (MS)	Time (MS)	Time (MS)	Time (MS)	Time (MS)	Time (MS)	Time (MS)	Time (MS)
1	I-4	1084	1721	5642	1986						
2	I-5	443	483	379	778	354					
3	I-6	332	246	189	253	449	379				
4	I-7	198	170	168	174	224	233	226			
5	I-8	282	148	171	250	207	171	148	175		
6	I-9	208	158	157	222	211	180	169	140	138	
7	I-10	231	147	136	131	239	142	138	143	155	225

The comparison of PPV values considering different FVs and TCs using FCM are shown in Fig. 3 to Fig. 6. The results are based on the consideration of different iterations and changes of degree of membership due to the random number initialization. So the difference in the Fig. 3 to Fig. 6 is the random initialization for the calculation of degree of membership. Fig. 3 to Fig. 6 shows the PPV results based on variable FVs and TCs. The parameters for all the figures are same but the cluster centres are different. The results clearly indicate that the clustering accuracy for each figures are different but the margin is very less. So it is proved by Fig. 3 to Fig. 6 that with different cluster centres results are vary but the margin is very low. So the accuracies obtained are not biased. The highest and lowest clustering accuracies are (97.2%, 91.1%), (97.2%, 90.9%), (97.8%, 90.4%), and (97.1%, 90.2%) for Fig. 3, 4, 5, and 6, respectively. The average highest and lowest clustering accuracies are (95.7%, 94.7%), (95.9%, 93.6%), (95.3%, 94.2%), and (95.6%, 93.7%) for Fig. 3, 4, 5, and 6, respectively. The FDR values are the false positive numbers. If the PPV value is 90% then the FDR value is 10%. The results based on FDR indicate that the FCM algorithm false discovery rates are low in comparison to the k-means. It is clearly shown in Fig. 7.

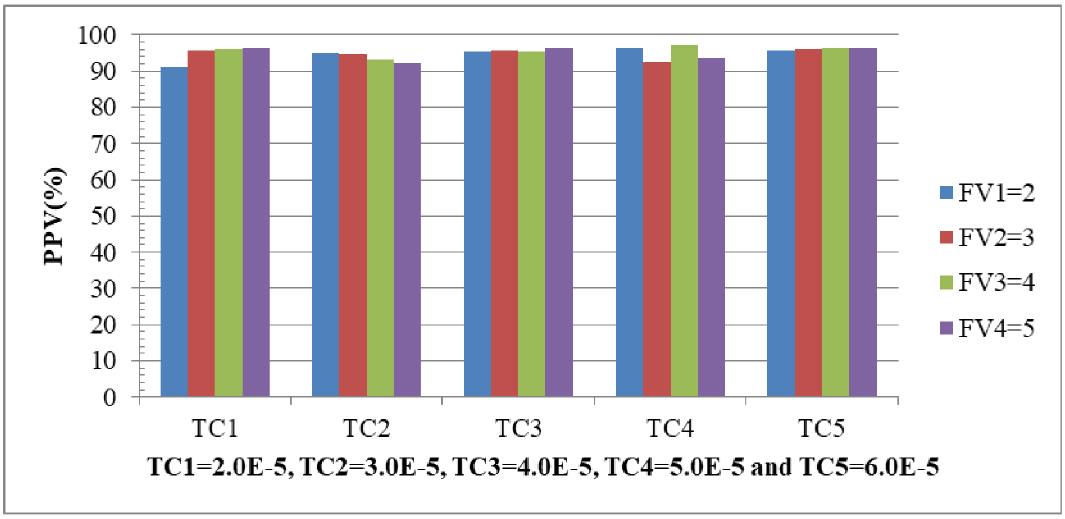

Fig. 3 PPV result based on degree of membership (random initialization) phase-1

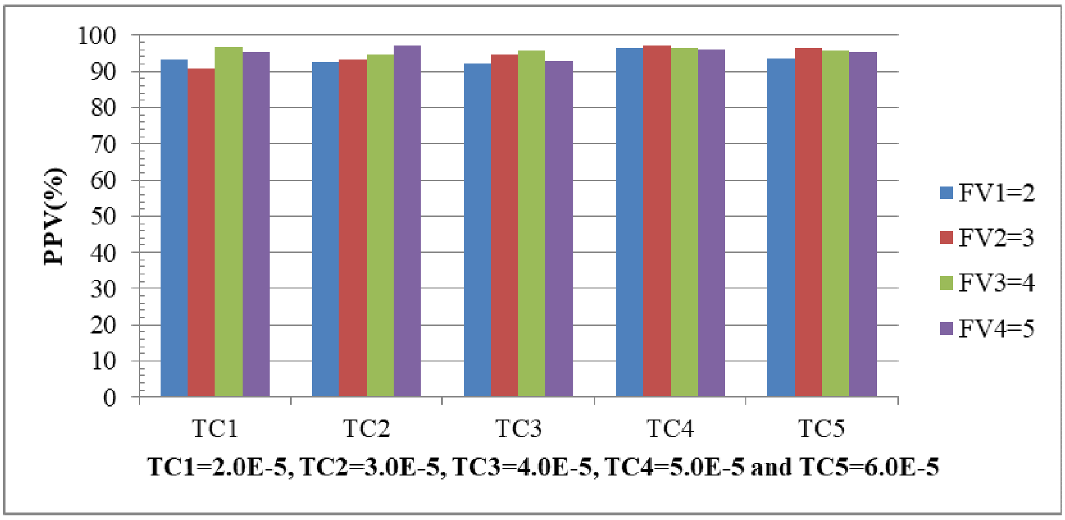

Fig. 4 PPV result based on degree of membership (random initialization) phase-2

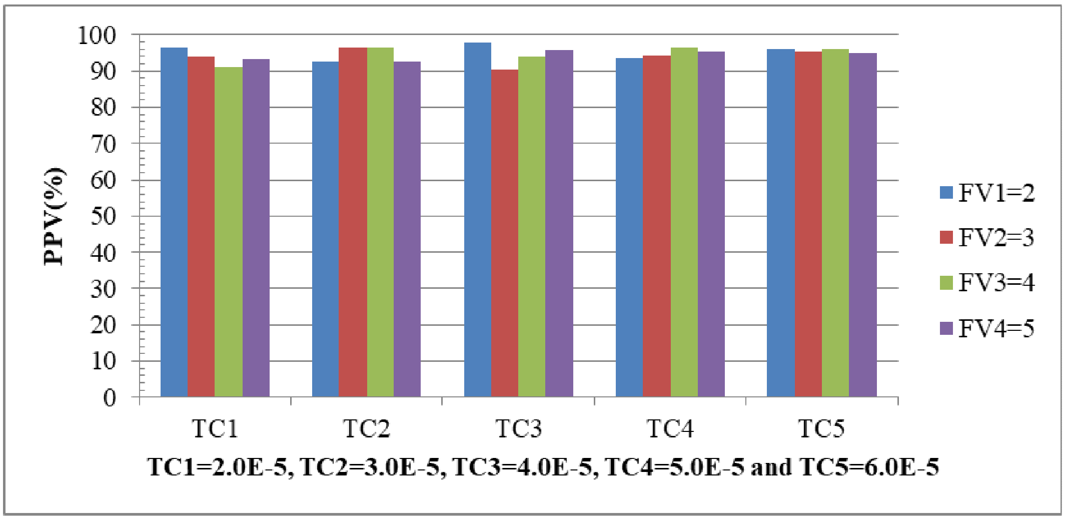

Fig. 5 PPV result based on degree of membership (random initialization) phase-3

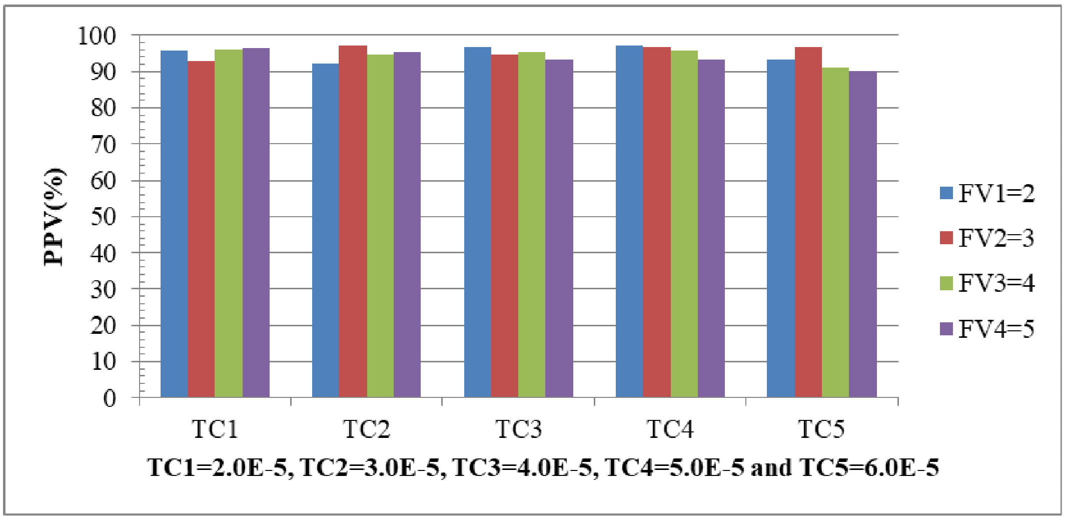

Fig. 6 PPV result based on degree of membership (random initialization) phase-4

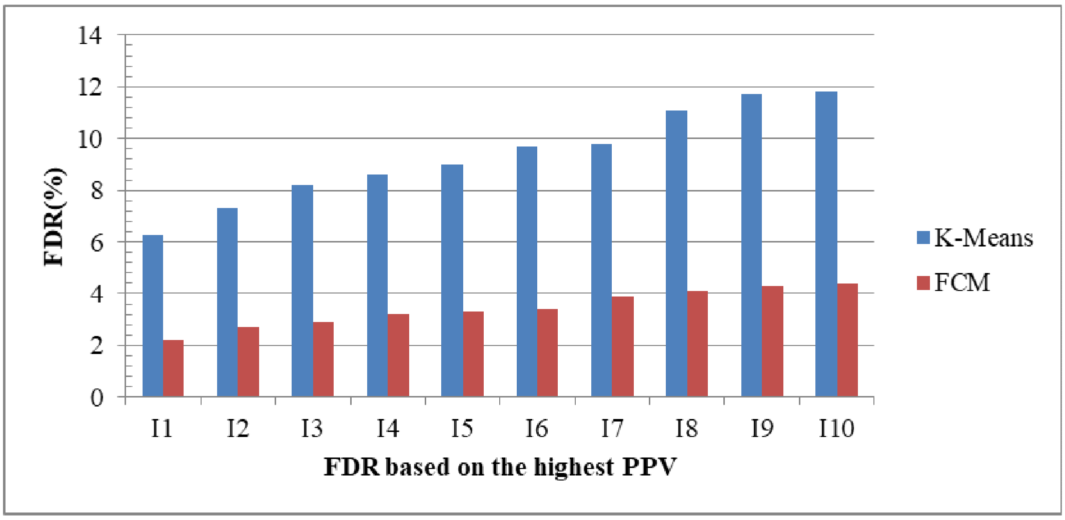

Fig. 7 FDR results based on the highest PPV achieved in case of k-means and FCM

So in all the cases, different parameters and populations were considered and it was found that FCM outperforms in comparison to k-means algorithm for the breast cancer data clustering.

Computation time comparisons with different FVs and TCs using FCM are shown in Table 13. The computation time, in case of FCM, is high due to the several iterations and fuzzy measures' calculations (Table 13). The calculative parameters, degree of membership evaluation and updating and the steps involve in comparing the termination criteria is extensive. So the time taken in the FCM process is more comparison to the k-means algorithm.

TABLE XIII
TIME VALUES CONSIDERING VARIOUS FUZZINESS VALUE AND TERMINATION CRITERIA USING FCM

S.no	Termination Criteria	Time (MS)	Time (MS)	Time (MS)	Time (MS)
		FV=2	FV=3	FV=4	FV=5
1	2.0E-5	928	1035	1538	1545
2	3.0E-5	1032	1506	1145	1521
3	4.0E-5	1548	1190	1464	889
4	5.0E-5	1339	1513	939	1461
5	6.0E-5	961	1232	1343	1473

In this study, the performance of k-means and FCM algorithms for BCW dataset was compared. BCW dataset is considered to delineate the impact of clustering on the basis of different parameters. The key findings are as follows:

1. The results obtained by k-means algorithm show that the variation in the total PPV is due to the random initialization in case of foggy centroid attribute values, as it is participating in determining the mode for the mean and variance in the centroid.
2. In the case of same centroid, the process is stopped if the means do not change anymore. Therefore, the process is not restricted to the epoch and the possibility of better cluster selection is improved. The combination of highest variance and the same centroid provides good results in case of k-means.
3. FCM algorithm produces better results in comparison to the k-means algorithm. The highest accuracy obtained is 97% and 92% respectively for FCM and k-means algorithms. FCM provides an iterative analysis so it gives better results for all selections.
4. The results are approximately same in the case of BCW dataset for a fuzziness value of 2-5. So for BCW dataset a fuzziness value 2-5 can be considered.
5. In case of computation time, k-means algorithm is far better than FCM algorithm. As the computation time is high in FCM algorithm. In case of k-means algorithm, if the number of iterations is more, the number of partitions is also more, which contain less attributes, so the computation time is low in case of high number of iterations compared to a low number of iterations. But in case of FCM algorithm, the computation time is high due to the various calculations, fuzzy measures, and a large number of iterations.
6. In k-means algorithm the data point should belong to one cluster but in case of FCM, it may belong to more than one cluster as membership is assigned to each data point.
7. As the computations are checked several times and the results obtained are uniform so FCM clustering is relatively efficient.

IV. CONCLUSIONS

K-means and FCM clustering algorithms were used in this study for the clustering of the BCW dataset. K-means algorithm is a simple and easy way to classify datasets through assuming k clusters with fixed apriori. FCM algorithm provides an iterative process with the update of cluster centers by updating and assigning membership values. In this work, a computational formulation is presented for integrative clustering with multi variant parameters including BCW data for obtaining good clustering accuracy. K-means algorithm is presented with foggy and random centroids considering the centroid, distance, split method, threshold, epoch, BCW attribute and number of iterations. This work is elaborated in several ways and makes certain important observations. The results of k-means algorithm indicated good accuracy in case of highest variance and same centroid. The consistency and uniformity in case of FCM algorithm is more prominent than k-means algorithm as the results of several repetitions suggest. The highest accuracy obtained is 97% and 92% for FCM and k-means algorithms, respectively. But the computation time is higher in FCM compared to k-means algorithm, thus k-means algorithm is efficient in terms of computation time. This implies that the FCM algorithm produces better results in comparison to the k-means algorithm but with the higher computation time in case of BCW dataset. In future the work is extended in the direction of clustering the non-clustered data obtained from k-means and FCM algorithms.

REFERENCES

[1] S.A. Eccles, E.O. Aboagye, S. Ali, A.S. Anderson, J. Armes, F. Berditchevski, J.P. Blaydes, K. Brennan, N.J. Brown, H.E. Bryant and N.J. Bundred, "Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer," *Breast Cancer Res*, vol. 15, pp. 1-37, 2013.

[2] J. Ferlay, H.R. Shin, F. Bray, D. Forman, C. Mathers and D.M. Parkin, "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008," *Int J Cancer, vol.* 127, pp. 2893-2917, 2010.

[3] A.K. Dubey, U. Gupta and S. Jain, "Breast cancer statistics and prediction methodology: a systematic review and analysis," *Asian Pac J Cancer P*, vol. 16, pp. 4237-4245, 2015.

[4] A.K. Dubey, U. Gupta and S. Jain, "A survey on breast cancer scenario and prediction strategy," *In proceedings of the 3rd international conference on frontiers of intelligent computing: theory and applications (FICTA),* (pp. 367-375), Springer International Publishing, 2015.

[5] R. Jain, *Introduction to Data Mining Techniques*. Available: http://www.iasri.res.in/ebook/expertsystem/datamining.pdf.

[6] S. Ghosh and S.K. Dubey, "Comparative analysis of k-means and fuzzy c-means algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 4, pp.35-38, 2013.

[7] C. Mary and S.K. Raja, "Refinement of clusters from k-means with ant colony optimization," *J Theor Appl Inf Technol*, vol. 6, pp. 28-32, 2009.

[8] C. Wang, R. Machiraju and K. Huang, "Breast cancer patient stratification using a molecular regularized consensus clustering method," *Methods*, vol. 67, pp. 304-312, 2014.

[9] A.K. Dubey, U. Gupta and S. Jain, "Analysis of k-means clustering approach on the breast cancer Wisconsin dataset," *Int J Comput Assist Radiol Surg, vol.* 11, pp. 2033-2047, 2016.

[10] S. Bhardwaj and V. Verma," Improved k-means clustering algorithm using back propagation method," *Int* J *Control Theor AP,* vol. 9, pp. 5169-5180, 2016.

[11] J. Lu, A. Hales and D. Rew, "Modelling of cancer patient records: a structured approach to data mining and visual analytics," *In International Conference on Information Technology in Bio-and Medical Informatics,* (pp. 30-51), Springer, 2017.

[12] A. Rahideh and M.H. Shaheed, "Cancer classification using clustering based gene selection and artificial neural networks," *In international conference on control, instrumentation and automation,* (pp. 1175-1180), IEEE, 2011.

[13] P. Festa, "A biased random-key genetic algorithm for data clustering," *Math Biosci*, vol. 245, pp. 76-85, 2013.

[14] C.H. Chen, "A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection," *Appl Soft Comput,* vol. 20, pp. 4-14, 2014.

[15] D. Wei, Q. Jiang, Y. Wei and S. Wang, "A novel hierarchical clustering algorithm for gene sequences," *BMC bioinformatics*, vol. 13, pp. 174, 2012.

[16] D. Vanisri and C. Loganathan, Fuzzy pattern cluster scheme for breast cancer datasets," *In international conference on communication and computational intelligence,* (pp. 410-4), IEEE, 2010.

[17] F.K. Ahmad and N. Yusoff, "Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier," *In international conference on intelligent systems design and applications,* (pp. 121-5, IEEE, 2013.

[18] J. Yin, H. Sun, J. Yang and Q. Guo, "Comparison of k-means and fuzzy c-means algorithm performance for automated determination of the arterial input function", *PloS one*, vol. 9(2), pp. e85884, 2014.

[19] Z. Zainuddin and O. Pauline, "An effective fuzzy c-means algorithm based on symmetry similarity approach", *Appl Soft Comput*, vol. 35, pp. 433-448, 2015.

[20] B. Zheng, S.W. Yoon and S.S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms," *Expert Syst Appl*, vol. 41, pp. 1476-1482, 2014.

[21] A.A. Rachman and Z. Rustam, "Cancer classification using fuzzy c-means with feature selection," I*n international conference on mathematics, statistics, and their applications,* (pp. 31-34), IEEE, 2016.

[22] P.S. Deepthi and S.M. Thampi, "Predicting cancer subtypes from microarray data using semi-supervised fuzzy C-means algorithm," *J Intell Fuzzy Sys,* vol. 32, pp. 2797-805, 2017.

[23] K. Bache and M. Lichman, *UCI machine learning repository,* University of California, School of Information and Computer Science, Available: http://archive. ics. uci. edu/ml.

[24] C.B. James, *Pattern recognition with fuzzy objective function algorithms*, New York: Plenum Press, 1981.

[25] N.R. Pal and J.C. Bezdek, "On cluster validity for the fuzzy c-mean model," *IEEE Trans Fuzzy Syst,* vol. 3, pp. 370–379, 1995.

[26] J.C. Bezdek, "A physical interpretation of fuzzy ISODATA," *IEEE Trans Syst Man Cy B*, vol. 6, pp. 387–389, 1976.

[27] K. Zhou, C. Fu and S. Yang, "Fuzziness parameter selection in fuzzy c-means: The perspective of cluster validation," *Science China Information Sciences,* vol. 57, pp. 1-8, 2014.