Bio-Inspired Computational Approaches for Breast Cancer Cluster Analysis

Maninder Kaur, Meghna Dhalaria

Abstract*: Breast cancer is the second most widespread disease throughout the world and the maximum incessant cause of death. The possibilities of survival are higher, if it is diagnosed in early phases. Moreover, the lack of awareness, treatment facilities and proactive measures expand the risks of survival. Cluster analysis is a statistical practice that categorizes observations into like sets or groups. The usage of cluster exploration offers a complex challenge as it entails numerous practical selections that define the superiority of a cluster solution. This paper highlights the application of cluster analysis inBreast Cancer dataset with the help of evolutionary approaches .Various evolutionary algorithms like genetic, differential evolution and particle swarm optimization are considered to overcome the problem of local maxima. The work proposes three evolutionary algorithm based techniques named WCGA,WCPSO and WCDE to perform clustering of breast cancer data and evaluate their effectiveness based on clustering validity measure (DBIndex), computation time and in terms of classification parameters. The results show that WCDE outperformed WCGA and WCPSO in terms of DB Index.

***Index Terms*: GA-clustering, DE-clustering, PSO-clustering, Validation index, Breast Cancer.**

I. INTRODUCTION

The diseasebreast cancer is a kind of tumor that often influences women. It is caused by two aspects, one being modifiable aspect,that be managed like behavior and environmental problems and other non-modifiable one that cannot be managed like hereditary of breast cancer. This disease is second most widespread disease across the world after lung cancer and the maximum incessant cause of death. The possibilities of survival are higher, if it is diagnosed in early phases. Because its symptoms differ from person-to-person, it is far essential to describe specific features of dissimilar patients and design a patient-particular remedy. Due to various reasons such as lack of awareness,illiteracy, and financial constraints, women often do not undergo medical care at early stage of prognosis. Early analysis of the syndromes may additionally direct to conquer the breast cancer through proper treatment. In the year 2014, approximate 2,32,714 new breast cancer rates occurred in women, while an aggregate of 2,97,800 female patients died due to cancer in which 16.1% of the aggregate death was in breast cancer happened inside the US. In the early years of cancer studies, researchers have used the conventional microscopic method to evaluate tumor of patients with breast cancer. For the detection and cure of disease, exact forecast of tumor is significantly vital. The researchers are dynamically utilizing various machine-learning methods to acquire appropriate cancer facts from the databases.

Untilnow, both supervised and unsupervised machine learning approaches have been exploited in breast cancer analysis.Amongst the current practices, supervised machine learning method is the prevalent in detection of disease. The authors Joshi et al. [1] implemented various clustering and classification technique to create a pattern of breast cancer patients. The 47 classification algorithms were used for identifying healthy individuals from sick patients. Fung et al. [2] collectedbreast cancer datasets from Leiden University and applied decision tree algorithm on these datasets. By initial analysis of three years, they created the patterns of breast cancer using a decision tree (DT) algorithm. The accuracy obtained by the DT classifier is 70%. Dubey et al. [3] applied k-means to assess the effect of clustering using split method, distance measurement and centroid initialization. They concluded that distance measures such as Manhattan and Euclidean distance gives better results as compare to others. Majali et al. [4] introduced an analytic system using association rule mining and classification technique. The work utilized frequent pattern algorithm in association rule to discover the pattern of malignant and benign. The DT classifier was used for the prediction of the chance of cancer using age criteria. The work was simulated on Wisconsin dataset and the results showed classification accuracy of 90%. Sentruk et al. [5] examined the overall performance of seven different classification forecast models namely Logistic regression,Discriminant Analysis (DA), K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN), Decision Tree (DT), Support Vector Machine (SVM) and Naive Bayes (NB) for diagnosis breast cancer through Rapid Miner Tool. The SVM classifier achieved highest accuracy of 96%. Ghosh et al. [6] implemented different classification technique such as SVM and MLP using Back propagation Neural Network on Wisconsin dataset of breast cancer. They finalized that SVM classifier has the ability to improve the conventional classification algorithm. Shah et al. [7] conducted the comparison between Decision Tree (DT),K-Nearest Neighbor (KNN) and Bayesian Network. The experiment is conducted through WEKA tool. The authors concluded that Naive Bayes (NB) is better in comparison to other algorithm because it take least time i.e. 0.02 second, providing highest accuracy.

Revised Manuscript Received on 22 May 2019.
***** Correspondence Author

Maninder Kaur*, Computer Science and Engineering Department, Thapar institute of Engineering and Technology, Patiala,India.

Meghna Dhalaria, Computer Science and Engineering Department, Thapar institute of Engineering and Technology, Patiala,India.

Hota [8] applied ensemble approach based on Decision tree and Artificial Neural Network, statistical, unsupervised ANN techniques for breast cancer classification dataset. The results showed ensemble model outperformed the single independent model. The authors Rajesh et al. [9] proposed an algorithm for analysis of breast cancer for both malignant and benign.

They used C4.5 algorithm for predicting breast cancer. The outcomes demonstrate that C4.5 (Decision tree) had 93% accuracy for detecting breast cancer. Padmavathi et al. [10] used different technique namely Multilayer perceptron (MLP), logistic regression andRadial Basis Function (RBF) techniques for predicting breast cancer. Their outcomes revealed that, Radial Basis Function (RBF) is better technique for predicting breast cancer as compared to other techniques. Moreover, the time by the RBF for predicting was lesser than the other techniques. Dumitru et al. [11] applied Naive Baye classifier on Wisconsin breast cancer dataof 198 patients and a binary decision class. The performance measures of the Naive Bayes classifier was 74.24 % which is better than the other well-known machine learning technique. Nauck et al. [12] applied the supervised fuzzy clustering technique on breast cancer dataset. The experimental result showed 95.57% accuracy by supervised fuzzy clustering method.Wang et al. [13] used SVM-based ensemble learning algorithm to reduce the diagnosis variance and increase diagnosis. The proposed model reduced the variance by 97.89% and improved accuracy by 33.34%, in comparison to best single SVM model.Liu, N., et al. proposed[14] a novel breast cancer intelligent diagnosis approach that utilized information gain directed simulated annealing genetic algorithm wrapper (IGSAGAW) for feature selection. The proposed feature selection approach improved classification accuracy and minimizes the misclassification cost.

Clustering is a vital technique that empowers combination of information on the basis of the nature or the indication of the disease.The major work in literature has been focused on classification problem for the prediction of Breast cancer with less work on clustering analysis in this direction. Moreover, less attention has been paid on the evolutionary approaches namely genetic, particle swarm optimization and differential algorithm on Breast Cancer Wisconsin (Diagnostic) Dataset.

This paper focuses on the application of various evolutionary approaches for cluster analysis of breast cancer data along with their comparative analysis.

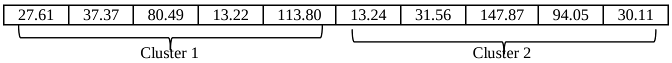
The rest of this paper is organized as follows. Section II presents proposed methodology. The simulations with results discussion is presented in Section III. Section IV concludes this work.

II. PROPOSED METHODOLOGY

In this work, three evolutionary approaches namely **W**iscosin breast cancer **C**lustering using **G**enetic **A**lgorithm,(WCGA), **W**iscosin breast cancer **C**lustering using **P**article **S**warm **O**ptimization(WCPSO) and **W**iscosin breast cancer **C**lustering using **D**ifferential **E**volution(WCDE) are proposed. These approaches are based GA, PSO and DE approaches respectively. For simulation work, breast cancer Wisconsin (Diagnostic) dataset is used to evaluate the efficiency of the evolutionary clustering approaches.

Solution Representation

Given a dataset (mXn size) containing ‘m’ samples with ‘n’ attributes, the solution for clustering the data in k clusters is encoded as nXk size linear array representing the k cluster centroids in sequence. The solution consists of real numbers signifying the k cluster centres as shown in figure 1.



27.61	37.37	80.49	13.22	113.80	13.24	31.56	147.87	94.05	30.11

Figure 1: Solution Representation

Let n=5 where ‘n’ be the number of attribute in a dataset and k=2 where k be the number of cluster. Then the chromosomerepresents the two cluster centers i.e. [(27.61, 37.37, 80.49, 13.22, 113.80) and (13.24, 31.56, 147.87, 94.05, 30.11)].

Fitness Function

The main objective of cluster analysisis to split the given data into a set of clusters that increase the homogeneity within the clusters, and reduce the heterogeneity among clusters. There has been various measures in literature to measure the validity of clusters. The DB index is one of the measure that has been adopted in this work. The lower DB index, better the quality of cluster by making more compact and separated clusters.

Termination Criteria

The complete procedure will continue (fitness function, crossover,selection and mutation) until the maximum number of iterations.

A. Wiscosin breast cancer Clustering using Genetic Algorithm (WCGA)

This section presents genetic algorithm based clustering for breast cancer cluster analysis. Genetic algorithm (GA) is typically used to produce high-quality solutions to optimization. The basic components of the approach are as follows:

Population Initialization

TheK cluster centers of each chromosome in the population are initialized with random set of real numbers. The number of chromosomes initialized is equal to the population sizeP.

Selection

In the selection stage of GA, offspring’s are chosen from the pool that are further involved for crossover operation. Roulette wheel selection technique is applied for the selection procedure.

Crossover

It is a probabilistic methodology that swaps the component of the solution with some other solution representation. In this two-point crossover is used with fixed crossover

Algorithm1 WCGA pseudo code

```
WBC_CL_GA(Max_Iteration, Pop_Size, pm, mu, k)
Input:- Pop_Size-Population Size, Max_Iteration-Maximum Iterations, pc –probability of crossover,pm- probability of mutation, nc- no. of offsprings, mu-Mutation Rate.
Output:-BestSol, BestCost of cluster.
Begin
Initialize a structure pop of size (Pop_SizeX, 1) wherein Pop.Sol represents the solution in terms of k cluster centres, Pop.cost stores the fitness value associated with that solution, Pop_c is a temporary array used for crossover and Pop_m is temporary array used for mutation.
  X=READ('data.csv')
  Set k = 2
  Initialize (nVar, VarSize, Max_Iteration)
CostFunction= clusteringcost(x, m)
  Set VarMin= min_tuple(X)
  Set VarMax= max_tuple(X)
       For(i=1to Pop_Size)
           Pop=Initialize_ population()
           Pop.cost=Evaluate (Pop, CostFunction)
       Endfor
  Pop=Sort_descend(Pop, Pop.cost)
  BestSol=Pop(1)
       For(it= 1 to Max_Iteration)
            For(k=1to nc/2)
                X=RouletteWheelSelection()
                Y=RouletteWheelSelection()
                 Pop_c(1).Sol, Pop_c(2).sol=Crossover(x.Sol,y.Sol,VarMin,VarMax)
                 Pop_c(1).cost=Evaluate (Pop_c(1).Sol, CostFunction)
                 Pop_c(2).cost=Evaluate (Pop_c(2).Sol, CostFunction)
            Endfor
            For(k=1 to nm)
                i=random_index([1 Pop_size])
                Pop_m(k).sol=Mutate(Pop(i).sol,mu,VarMin,VarMax)
                Pop_m(k).cost=Evaluate (Pop_m(k).sol, CostFunction)
            Endfor
                Pop=Merge_Solutions(Pop, Pop_c, Pop_m)
                pop=Sort_descend(Pop,Pop.cost);
                pop=pop(1:Pop_size)
                BestSol=pop(1)
                BestCost(it)=BestSol.Cost
       Endfor
End
```

probability *pc*. The parts of the chromosomes lying to the right of the crossover point are exchanged to create two offspring.

Mutation

It is achieved by flipping the part of one solution randomly, which increases the variety of the populace and provides a mechanism to get away from a local optimum. In this fixed mutation probability pm is used.

B. Wiscosin breast cancer Clustering using Particle Swarm Optimization(WCPSO)

WCPSO approach is based on particle swarm optimization that is motivated from the social behaviour of flocking of birds or fish schooling. It is widely used due to its flexibility and simplicity. In this there is a global communication among the swarm particle and uses real numbers.The basic components of the algorithm are as follows:

Population Initialization

In the initial stage, the positions vector ***X*** and accompanying velocity *V* of all particles in the population is initialized randomly. In this case, the position ***X*** of the particle represent the centroid of cluster position.

Fitness Function Evaluation

If the fitness value of the particle is better than the *pbest* (particle's best) solutionP_i. *Then update the particle fitness with* P_i. Similarly, update the global best(gbest)P_g if there is improvement in the fitness of the current population.

Velocity Update

Let v_i and x_i be the velocity and position for particlerespectively. The velocity is update by the eq(1).

$$v_i^{t+1} = wv_i^t + \alpha_1\beta_1[pbest - x_i^t] + \alpha_2\beta_2[gbest - x_i^t] \qquad (1)$$

Here v_i^t and $x_i^t$is the velocity and position for particle *i* at *t* time , *w* is the inertia weight [0 to1.2],$\alpha_1\alpha_2$

Algorithm2 WCPSO pseudo code

```
WBC_CL_PSO(Max_Iteration,Pop_Size,w,wdamp,c1,c2,k)
Input:- Pop_Size-Population Size, Max_Iteration-Maximum Iterations, w-Inertia Weight, wdamp- Damping Ratio of inertia weight, c1-Cognitive Coefficient, c2-Social Coefficient.
Output:-BestSol, BestCost of cluster.
Begin
Initialize a structure pop of size (Pop_SizeX,1) wherein particle.sol represents the solution in terms of k cluster centres, particle.cost stores the fitness value associated with that solution
  X=READ('data.csv')
  Set k = 2
  Initialize (nVar, VarSize, Max_Iteration)
  CostFunction= clusteringcost(x, m)
  Set VarMin= min_tuple(X)
  Set VarMax= max_tuple(X)
  Set GlobalBest= ∞
 VelMin= VelMax=(VarMax-VarMin)*0.1
       For(i= 1 to Pop_Size)
           Initialize the position and velocity of particle.
           Cost =Evaluate_particle cost()
           Update particle.Best.Cost
                If (GlobalBest.Cost>particle.Best.Cost)
                   Then GlobalBest=particle.Best.Cost
                Endif
       Endfor
       For(i= 1 to Max_Iteration)
              For(pp= 1 to Pop_Size)
                   Update the velocity of particle.
                         If(particle(pp).Velocity<VelMin)
                             particle(pp).Velocity= VelMin
                         Endif
                         If(particle(pp).Velocity<VelMiax)
                             particle(pp).Velocity= VelMax
                         Endif
                     particle(pp).Position= particle(pp).Position + particle(pp).Velocity
                     Check if particle(pp).Position lies
```

```
within the limits of VarMin and VarMax
            Particle(pp).Cost= CostFunction (particle(pp). Position)
            Update PersonalBest and GlobalBest
        Endfor
      BestCost(it)= GlobalBest.Cost
    Endfor
  BestSol= GlobalBest
End
```

are the coefficients [0 to 2], $\beta_1\beta_2$ are the random values lies between [0 to1], pbest is the particle best solution and gbest is the global best solution. wv_i^t is the inertia component, responsible for maintaining the particle moving in the similar way it was initially heading.

Particle Update

As soon as the velocity for each particle is calculated, the particle position is

updated by using the eq(2):

$$x_i^{t+1} = x_i^t + v_i^{t+1} \qquad (2)$$

Here x_i^t is the position of pervious particles and v_i^{t+1} is the new updated velocity.

C. Wiscosin breast cancer Clustering using Differential Evolution(WCDE)

This approach is based on Differential Evolution algorithm for cluster analysis of breast cancer. Differential Evolution (DE) is robust and fast optimization technique that uses a stochastic, populace based search methodology. DE is used for the real value problems. The basic components of the approach are as follows:

Mutation

For each vector α_i at time t, first randomly choose three different vectorsx_a,x_b and x_c at t .The mutation is done to generate donor vector v_j . Donor vector are generated byincluding a weighted difference of two populace vector to a third vector as in eq(3).

$$v_i^{t+1} = x_a^t + F(x_b^t - x_c^t)(3)$$

Here F represents the differential weight where F ϵ [0, 2], v_j is donor vector generated by mutation

Crossover

Donor vector does crossover with the target vector (current generation) to generate the trial vector. pCR (Crossover Probability) parameter is used for controlling the rate pCR ϵ [0, 1].In this binomial crossover is used to perform crossover on each d variables or component by generating a consistently appropriated random number r ϵ [0, 1], if random no is greater than crossover probability then value of target vector become trial vector otherwise value of donor vector become trial vector as in eq (4).

$$u_{j,i}^{t+1} = \begin{cases} v_{j,i} & if\ r_i \le pCR, \\ x_{j,i}^t & otherwise, \end{cases} \quad j = 1,2,3 \ldots .d. \Big\} \qquad (4)$$

Selection

Selection is same as that used in genetic algorithms. Trail vector is

Algorithm3 WCDE pseudo code

```
WBC_CL_DE(Max_Iteration,Pop_Size, beta_min, beta_max, pCR,k)
Input:- Pop_Size-Population Size, Max_Iteration-Maximum Iterations, beta_max-upper bound of scaling factor, pCR-Crossover Probability, beta_min-Lower bound of scaling factor.
Output:-BestSolution, BestCost of cluster.
Initialize a structure pop of size (Pop_SizeX, 1) wherein Pop.position represents the solution in terms of k cluster centres and Pop.cost stores the fitness value associated with that solution.
Begin
  X=READ('data.csv')
  Set k = 2
  Initialize(nVar, VarSize, Max_Iteration)
  CostFunction= clusteringcost(x, m)
  Set VarMin= min_tuple(X)
  Set VarMax= max_tuple(X)
  For(i=1 to Pop_Size)
    Pop=Initialize_Population()
    Pop.cost=Evaluate(Pop, CostFunction)
      If(Pop(i).cost<BestSolution.cost)
        BestSolution=Pop(i)
      Endif
  Endfor
  For(it=1 to to Max_Iteration)
    For(i=1 to Pop_Size)
      x=Pop(i).Position;
      Y=randperm(Pop_Size);
      Y(Y==i)=[]
      p=Y (1)
      q=Y (2)
      r=Y (3)
      beta= random_generation(beta_max ,beta_min ,VarSize)
      a=Pop(p).Position+$\beta$.*(Pop(q).Position-Pop(r).Position)
      a=max(a,VarMin)
      a=min(a,VarMax)
      b=zeros(size(x))
      i0=randi([1 numel(x)])
      For(l=1to numel(x))
        If(l==i0 || rand<=pCR)
          b(l)=a(l)
        Else
          b(l)=x(l)
        Endif
      Endfor
      NewSolution.Position=b
      [NewSolution.Cost, NewSolution.Out]=CostFunction(NewSolution.Position)
        If(NewSolution.Cost<Pop(i).cost)
          Pop(i)=NewSolution;
          If(Pop(i).cost<BestSolution.cost)
            BestSolution=Pop(i)
          Endif
        Endif
    Endfor
    BestCost(it)=BestSolution.Cost;
End
```

compare with the parent vector and choose the fittest value as in eq(5).

$$x_i^{t+1} = \left\{\begin{matrix} u_i^{t+1} & if\ f(u_i^{t+1}) \le f(x_i^t) \\ x_i^t & otherwise. \end{matrix}\right\} \qquad (5)$$

III. Simulation Results

The proposed evolutionary approaches are implemented on Intel core i7 processor, 4 GB RAM, run on 64 bit OS (Operating System)using MATLAB (R2016a). The data is taken from UCI Repository of Breast Cancer Wisconsin (Diagnostic) Dataset[15]. The data set consists 569 instances with 32 attributes. The data set contains no missing value.

Table 1: Attribute information of the Dataset

Attributes	Description
ID number	Unique identificatiom
Diagnosis	Malignant, Benign
Radius	Average of separations from centre to factors on the circumference
Symmetry	distribution of cancer zone
Area	extent of cancerous zone
Perimeter	Distance arund the cancer zone
Compactness	(Perimeter)^2 / area – 1
Smoothness	Difference in lengths of radius
Concavity	Seriousness of concave portion of the form
Concave points	Number of inner parts of the form
Texture	Deviation (Standard) of gray-scale esteems
Fractal Dimension	estimate of coastline – 1

A. Validity measure used

Validity Measures are used to check the quality of the clusters formed by the algorithm. The validity measures quantify the quality of clusters based on the properties that are inherent in the data sets.

Davies-Bouldin (DB) Index

Donald W. Bouldin and David L. Davies introduced DB index in 1979[16]. It is a used for evaluating clustering algorithms. DB Index validates the clustering algorithm using features and quantities inherent in the data set, hence is an internal evaluation scheme. DB Index evaluates the inter cluster differences and intra cluster similarity. Therefore, lower the value of DB index, the better is the clustering algorithm. It is defined as in eq(6):-

$$BD = \frac{1}{c}\sum_{i=1}^{c} Max_{i\neq j}\left\{\frac{d(x_i)+d(x_j)}{d(c_ic_j)}\right\} \qquad (6)$$

C denotes the number of clusters, $d(x_i)$and $d(x_j)$ are clusters sample i and j to their appropriate clusters centroid, i and j are cluster label and $d(c_ic_j)$is the distance between the centroids. The best clustering scheme minimizes the DB index value. The value of k (number of cluster) for which value of DB index is least is termed as the optimal value of k.Various cluster Validity Indices in terms of classification evaluation parameters used are as follows:

Accuracy-: is termed as correctness of classified instances to the aggregate number of instances. It is calculated as in eq(7)

$$Accuracy = \frac{(tp+tn)}{(tp+tn+fp+fn)} \qquad (7)$$

Recall- is the ratio of actual positive cases that are correctly identified. It is also known as true positive.It is calculated as in eq(8).

$$Recall = \frac{tp}{(tp+fn)} \qquad (8)$$

Precision- is the ratio of actually true predicted instances out of the total number of true instances.It is calculated as in eq(9).

$$Precision = \frac{tp}{(tp+fp)} \qquad (9)$$

F-measure-It is defined as the measure of test's accuracy, also known as F-score. It combines both recall and precision of the test to determine the score. If the f-score value is 1 then it means perfect precision and recall and for 0 it means worst.It is calculated as in eq(10)

$$F - measure = \frac{2\times.(Precision\times.Recall)}{(Precision+Recall)} (10)$$

tp (True Positive):represents Breast cancer patient suffering from breast cancer.

tn (True Negative): represents Non-Breast cancer patient are not suffering from breast cancer.

fp (False Positive): represents Non-Breastcancer patient suffering from breast cancer.

fn (False Negative): represents Breast cancer patient are not suffering from breast cancer.

Various control parameter used in evolutionary clustering as well as classification are: PopSize-Population size, Max_It- Maximum iterations, pm-Mutation probability, pc-Crossover probability, mu-Mutation rate, beta-Selection pressure, w-Inertia weight, wdamp-Damping Ratio of inertia weight, c1-Cognitive Coefficient, beta_min-Lower bound, beta_max-Upper bound,c2-Social coefficient, pCR-Crossover probability (parameters values tabulated in table2).

Table 2: The control parameter of WCGA, WCPSO and WCDE

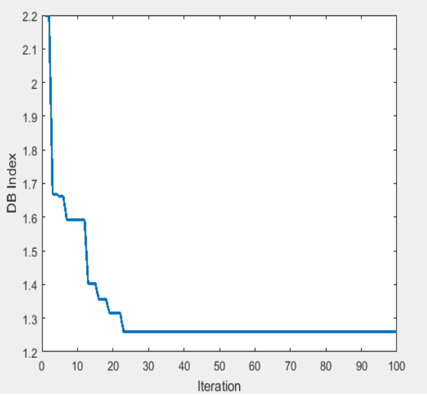
WCGA		**WCPSO**		**WCDE**	
Parameters	**Values**	**Parameters**	**Values**	**Parameters**	**Values**
Max_It	100	Max_It	50	Max_It	50
PopSize	100	PopSize	100	PopSize	100
Pc	0.8	W	1	beta_min	0.2
Pm	0.3	Wdamp	0.99	beta_max	0.8
Mu	0.02	c1	2	Pcr	0.2
		c2	2		

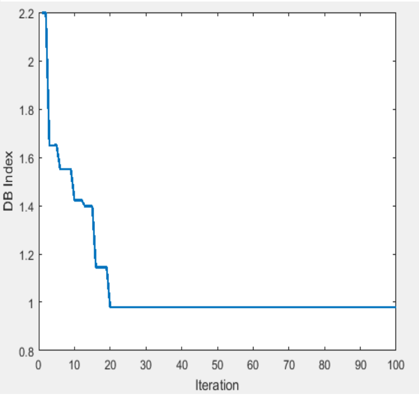
The table 3 shows that the results obtained from WCGA, WCPSO, WCDE based on validity measures namely DB index and computation time. The lower DB index, better the quality of cluster by making more compact and separated clusters.

The db index values for WCGA and WCPSO are 1.04918 and 0.93021respectively. It is clear from the table that WCDE outperforms in term of the DB Index that is 0.88622 and it also has lower cpu time that is 231.1172 in comparison to the WCGA and WCPSO.

Table 3: The results of WCGA, WCPSO and WCDE evaluated on the basis of DB Index and Computation time.

	WCGA		WCPSO		WCDE	
Run	**DBIndex**	**CPU time**	**DBIndex**	**CPU time**	**DBIndex**	**CPU time**
Run1	0.96866	1101.6719	0.94050	255.1875	0.89874	252.5781
Run2	0.96161	2187.0156	0.95705	245.375	0.88748	241.0781
Run3	0.95166	1615.8906	0.88521	252.375	0.90707	229.9844
Run4	0.98966	1557.3281	0.90811	246.3125	0.90707	242.9063
Run5	0.96161	4960.000	0.94050	243.3594	0.85157	233.7188
Run6	0.98966	544.4688	0.92459	249.5938	0.85157	207.5313
Run7	0.95166	505.8125	0.94050	1009.0156	0.88748	183.0781
Run8	1.36881	502.625	0.90811	441.2188	0.86542	237.7344
Run9	1.36881	553.2656	0.95705	538.9688	0.90707	242.1719
Run10	0.97966	445.4688	0.94050	485.3281	0.89874	240.3906
Average	1.04918	950.9547	0.93021	396.6735	**0.88622**	231.1172





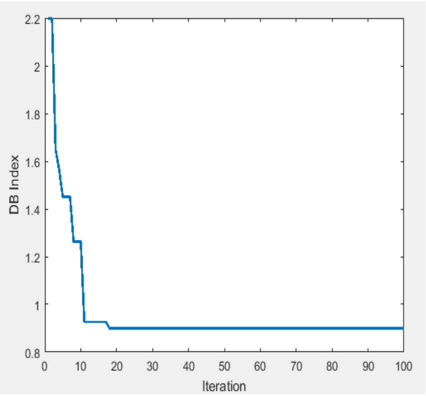


Figure 2: WCGA

Figure 3: WCPSO

Figure 4: WCDE

Table 4: The results of WCGA ,WCPSO,WCDE cluster validity (in terms of classification parameters)

	Evaluation Parameters	**Run_1**	**Run_2**	**Run_3**	**Run_4**	**Run_5**	**Run_6**	**Run_7**	**Run_8**	**Run_9**	**Run_10**	**Average**
WCGA	Accuracy	0.6796	0.6077	0.7206	0.7118	0.5990	0.7678	0.7192	0.7258	0.6425	0.6011	0.67751
	Precision	0.6570	0.6832	0.5726	0.5845	0.6172	0.6691	0.5953	0.6261	0.6770	0.6414	0.63234
	Recall	0.7302	0.7070	0.6858	0.6906	0.6547	0.6585	0.6604	0.6557	0.6331	0.6434	0.67194
	Fmeasure	0.6916	0.6948	0.6241	0.6331	0.6353	0.6637	0.6261	0.6405	0.6543	0.6423	0.65064
WCPSO	Accuracy	0.6763	0.6907	0.6116	0.7663	0.8067	0.7216	0.6149	0.7434	0.6935	0.7601	0.70851
	Precision	0.6742	0.7526	0.7896	0.6179	0.7277	0.7257	0.6340	0.6093	0.6153	0.7314	0.68377
	Recall	0.6717	0.7215	0.7356	0.6464	0.6323	0.7104	0.6534	0.7402	0.7151	0.6509	0.68575
	Fmeasure	0.6729	0.7367	0.6146	0.6318	0.6766	0.7179	0.6435	0.6684	0.6527	0.6888	0.68513
WCDE	Accuracy	0.7608	0.8917	0.7661	0.8906	0.7509	0.8674	0.7889	0.7957	0.8478	0.7785	**0.81384**

	Precision	0.7248	0.7174	0.6957	0.6503	0.7647	0.7241	0.6510	0.6259	0.6665	0.6231	0.68435
	Recall	0.6421	0.6345	0.6973	0.6452	0.7536	0.6373	0.7243	0.6362	0.7623	0.6066	0.67394
	Fmeasure	0.6809	0.6734	0.6964	0.6477	0.7591	0.6779	0.6856	0.6310	0.7111	0.6147	0.67780

Fig 2, 3 and 4 depicts the plot between number of iterations and DB index. Lower the value of DB index, the better the result is. It is clear from the figures that WCDE gives lower values of DB index as compared to the WCGA and WCPSO. The table 4 shows that the results obtained from WCGA ,WCPSO and WCDE based on parameters such as accuracy, sensitivity, specificity, precision, recall and f-measures. The WCDE gives best performance giving accuracy 0.81384, precision-0.68435 and recall-0.67394.It is clear that the results obtained from WCGA, WCPSO and WCDE cluster validity in terms of classification parameters especially in case of accuracy and sensitivity WCDE performs better as compared to the WCGA and WCPSO.

IV. CONCLUSION AND FUTURE WORK

In this paper, three evolutionary approaches such as WCGA, WCPSO, WCDE areapplied for cluster analysis of breast cancer. The acquired outcomes showed that WCDE outperformed as compared to WCGA and WCPSO. The value obtained by WCDE based on cluster validity measure DBIndex is 0.88622 and the approach perform better in case of computational time. In terms of classification parameter, WCDE achieves better results in terms of accuracy and sensitivity. The work can be extended by applying the proposed approaches on large-scale datasets using big data analytics giving clearer picture about the patterns in the data.

REFERENCES

1. Jahanvi Joshi, Rinal Doshi and Jigar Patel, “Diagnosis of Breast Cancer using Clustering Data Mining Approach.” International Journal of Computer Applications (0975 –8887) Volume 101– No.10, September 2014, 13-17.
2. Jimin GuoBenjamin C. M. Fung, Farkhund Iqbal and Peter J. K. Kuppen, “Revealing determinant factors for early breast cancer recurrence by decision tree,” Inf. Syst. Front., 2017.
3. Ashutosh Kumar Dubey, Umesh Gupta and Sonal Jain, “Analysis of k-means clustering approach on the breast cancer Wisconsin dataset.” International Journal of Computer Assisted Radiology and Surgery, Springer, 2016.
4. Jaimini Majali, Rishikesh Niranjan, Omkar Tadakhe and Vinamra Phatak, “Data Mining Techniques for Diagnosis and Prognosis of Cancer.” International Journal of Advanced Research in Computer and Communication Engineering**4(3)**, 613-616,2015.
5. ZehraKarapinar Senturk and Resul Kara, “Breast Cancer Diagnosis via Data Mining: Performance Analysis of Seven different algorithms.” Computer Science & Engineering **4(1)**, 35-46,2014.
6. Soumadip Ghosh, Sujoy Mondal and Bhaskar Ghosh, "A comparative study of breast cancer detection based on SVM and MLP BPN classifier."Automation, Control, Energy and Systems (ACES), 2014 First International Conference on. IEEE, 2014.
7. Chintan Shah and Anjali G. Jivani, "Comparison of data mining classification algorithms for breast cancer prediction." Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, 2013.
8. H.S Hota, “Diagnosis of Breast Cancer Using Intelligent Techniques.” International Journal of Emerging Science and Engineering (IJESE) 2013;1:45–53.
9. K Rajesh and Sheila Anand, “ Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm.” International Journal of Advanced Research in Computer and Communication Engineering. 2012;1:72–77.
10. J Padmavati, “A Comparative study on Breast Cancer Prediction Using RBF and MLP.” International Journal of Scientific & Engineering Research. 2011;2:1–5.
11. Diana Dumitru, "Prediction of recurrent events in breast cancer using the Naive Bayesian classification."Annals of the University of Craiova-Mathematics and Computer Science Series 36.2 (2009): 92-96.
12. Detlef Nauck and RudolfKruse, “Obtaining interpretable fuzzy classification rules from medical data.” Artificial intelligence in medicine,16(2), 149-169,1999.
13. Wang, H., Zheng, B., Yoon, S.W. and Ko, H.S., 2018. A support vector machine-based ensemble algorithm for breast cancer diagnosis. European Journal of Operational Research, 267(2), pp.687-699.
14. Liu, N., Qi, E.S., Xu, M., Gao, B. and Liu, G.Q.,. A novel intelligent classification model for breast cancer diagnosis. InformationProcessing&Management,56(3), 609-623,2019.
15. Michalski,R.S. Learning UCI repository of machine learning databases,” 1987.[Online].Available:https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
16. Davies, David L., and Donald W. Bouldin., A cluster separation measure, Pattern Analysis and Machine Intelligence, IEEE Transactions on 2 (1979): 224-227.

AUTHORS PROFILE

Maninder Kaur is holding an academic position as Assistant Professor in Department of Computer Science and Engineering, Thapar Institute of Engineering & Technology, Patiala. She received her Bachelor’s degree from Sant Longowal Institute of Engineering and Technology and Master’s degree from Punjabi University. She completed her Ph.D. in the field of VLSI physical design automation using evolutionary approach. Her major research experiences and interests include IoT, big data analytics, data mining and swarm intelligence. She has 25 publications including journals, conferences. She has also acted as mentor in various capstone projects in the field of IoT and machine learning. She is currently supervising two Ph.D. and two M.E. students. She has also supervised more than 12 M.E./M.Tech., thesis. She is an associate member of the Institution of Engineers, India. Her current research includes the application of machine learning and swarm intelligence techniques for big data analytics, cyberbullying and sarcasm detection.

Meghna Dhalaria is pursuing Ph.D in computer Science Department, Jaypee University of Information and Technology, Waknaghat. She received her Bachelor’s degree from Baddi University of Emerging Sciences and Technologies. She completed her Master’s degree from Thapar Institute of Engineering and Technology, Patiala. Her current research includes the applications of Machine learning and Deep learning.