Efficient Evolutionary based Clustering Approaches for Healthcare Data

Thesis submitted in partial fulfilment of the requirements for the award of degree of

Master of Engineering

in

Computer Science and Engineering

Submitted By

Meghna Dhalaria

(Roll No. 801632025)

Under the supervision of:

Dr. Maninder Kaur

Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY

PATIALA – 147004

June 2018

Certificate

I hereby certify that the work which is being presented in the thesis entitled, "*Efficient Evolutionary based Clustering Approaches for Healthcare Data* ", in partial fulfilment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar Institute of Engineering and Technology, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Maninder Kaur* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Meghna
(Meghna Dhalaria)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Maninder Kaur
(Dr. Maninder Kaur)

Assistant Professor,
CSED

Acknowledgement

First and foremost, I would like to thank my supervisor, Assistant Professor **Dr. Maninder Kaur** of Thapar Institute of Engineering and Technology for giving me an opportunity to work under her on the challenging and burning topic and providing me ample guidance and support through the course of the research. I am also thankful to them for their time, patience, discussions. I am grateful for their guidance, encouragement and support throughout my M.E research work.

I am also thankful to **Dr. Ashutosh Mishra**, P.G. coordinator, for the motivation and inspiration that triggered me for the thesis work.

I would like to thank **Dr. S.S. Bhatia**, Senior Professor and Dean of Academic Affairs, Thapar Institute of Engineering and Technology, Patiala, for giving provisions of the entire required infrastructure such as computer labs, library facilities, immensely useful for learners to equip themselves with the latest in the field.

I am equally grateful to **Dr. Maninder Singh,** Associate Professor and Head, Computer Science and Engineering Department, and also thankful to all my respected teachers in the Department for their direct or indirect help, inspiration and motivation.

I also would like to express my greatest gratitude to my parents, my brother and my sister for their endless love, constant support, encouragement and patients throughout my research without which I would not have reached the present position.

Last but not least, I would like to thank almighty for everything.

(Meghna Dhalaria)

Abstract

Good health care is one of the most significant factors which can make a contribution to the individual well-being of everyone in the modern world. The detection of diseases is a crucial and difficult task in healthcare. The recognition of diseases from numerous features or signs is a prime issue which is not free from false presumptions frequently followed with the aid of unpredictable effects. The healthcare enterprise gathers large amounts of disease data that unfortunately, are not mined to decide concealed facts for effective diagnosing. As the quantity of stored data increases, clustering play a vital role in extracting knowledge and finding patterns to provide better care and effective diagnostic capabilities. Clustering aims to arrange a set of data objects into clusters; such that objects inside a cluster are "similar" to each other than they are to objects in the different clusters. There are various numbers of applications for clustering which includes marketing, scientific and engineering, ecommerce, image segmentation business etc. The current work in the thesis focuses on the two diseases namely Wisconsin Breast Cancer and Epileptic seizure. The work relies on the finding the optimal solution based on clustering techniques. The proposed clustering techniques based evolutionary algorithms namely GA-clustering, PSO-clustering and DE-clustering are applied on breast cancer wisconsin dataset and their effectiveness is evaluated on the basis of DB index and classification parameters. In another work, a novel partitioning based clustering using DE approach is proposed that is applied on epileptic seizure recognition dataset and its results are compared with DE-clustering approach on the basis of cluster validity measures namely DB index, Dunn index and computational time. So, clustering techniques are of vital importance that it organizes the data, thereby generating patterns that can be further utilized for better analysis of diseases.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

DBI	Davies Bouldin Index
DE	Differential Evolution
DEC	Differential Evolution Clustering
DI	Dunn Index
EEG	Electroencephalogram
GA	Genetic Algorithm
ML	Machine Learning
MLP	Multi Layer Perceptrons
PCDE	Partitioning Clustering using Differential Evolution
PSO	Particle Swarm Optimization
SVM	Support Vector Machine
WCGA	Wisconsin Clustering Genetic Algorithm
WCDE	Wisconsin Clustering Differential Evolution
WCPSO	Wisconsin Clustering Particle Swarm Optimization

Chapter 1
Introduction

Good health care is one of the most significant factors which can make a contribution to the individual well-being of everyone in the modern world. The significance of health care in modern day society may be gauged from the reality that proper health is recognized as one of the essential rights of an individual. Health care systems throughout the globe should turn out to be more productive at treating patients and decreasing expenses. Occurrence of chronic diseases is expanding throughout the world and supervision of them represents one of the greatest healthcare challenges. Because of this reason, healthcare systems are attempting to discover better solutions that enhance quality, effectiveness and reduce costs. In addition, the use wearable technologies for helping individuals with chronic situations have been related with significant enhancements in personal satisfaction. Mobile computing and sensing technologies can possibly enhance healthcare quality and effectiveness. As compared with standard clinical practice for checking patients, estimation depends on observation data gathered in research centre settings or in person. Subsequently, exact and continuous monitoring has turned out to be progressively imperative in healthcare, where utilizing technology for patient monitoring can assist to examine the impact of mental illness on patient's every day performance to expand adequacy in treating mental problems and to examine perceived stress at workplaces. Because of its embedded sensors, advanced mobile phones have turned out to be fit for observing various measurements of human conduct, including social, mental, and physical communication measurements. Nowadays, healthcare institutions are more dependent on advances technology, and the utilization of Machine Learning (ML) techniques can give useful support to help physicians in many ways. In the most recent decade, ML got consideration from numerous areas, including healthcare with a point of enhancing administration quality and care. To-date, advanced ML methods that are utilized as a part of healthcare intended at solving prognostic problems, consisting human behaviour and mental-health fields.

The current research focuses on aspects of two diseases namely Wisconsin Breast Cancer and Epileptic Seizures as given below.

1.1. Breast cancer

Breast cancer, is a kind of tumour that often influences women. It is caused by two aspects. Modifiable aspects are the aspects that can be managed like behaviour and environmental problems whereas Non-modifiable aspects are those that can't be managed like hereditary of breast cancer [1]. Normally tumours can be classified as non-cancerous (benign) and cancerous (malignant). After lung cancer it is the second most widespread disease across the world and the maximum incessant cause of death. The possibilities of survival are higher, if it is diagnosed in early phases. Due to the fact that its symptoms differ from person-to- person, it's far essential to describe specific features of dissimilar patients and design a patient-particular remedy. According to a survey, in India 3.57 % of women is inclined to cancer as the early detection procedures at the occurrence of breast most cancers are still non-existence in the accurate disease prediction [2]. Furthermore, the dearth of consciousness, treatment facilities and proactive measures expand the survival risk. Early analysis of the syndromes may additionally direct to conquer the breast cancer throughout proper remedy. 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 [3]. For the reason that early year of cancer studies, researchers have used the conventional microscopic method to evaluate tumour of breast cancer patients. For the detection and treatment of disease, exact prediction of tumours is significantly important. The machine learning methods are dynamically used by existence researchers to acquire appropriate cancer facts from the databases. It's approximated that up to 30% of all breast malignancy tumours, even those got early will metastasize to different organs in the body, consisting of lungs, livers, bones or brains For the identification of breast malignancy, different methods are utilized as a part of mammography is the most encouraging procedure and utilized by radiologist often. Mannogram picas are commonly of low contrast and noisy. In breast mammography, bright regions constitute cancer. There are numerous features in mammography that assist physicians to locate abnormalities in early degree and those functions can be directly extracted by means of picture processing techniques. A portion of the hazard factors, such as age, hereditary hazard and family history improve the probability of a lady creating bosom tumour. With early detection, 97% of ladies can live for 5 years. In any case, large portions of tumour are analyzed in the last phase of the disease. So, the techniques for accurate and early detection of breast malignancy are necessary.

1.2. Epileptic Seizures

The disease in which patients suffer seizures caused by a brain functionality disorder is called epilepsy [4]. While more than fifty million people around the area are recognized with epilepsy, in the United States, approximately three million patients have been affected by epilepsy [5]. Epilepsy is the third most common brain disorder. Meanwhile, there are several possible causes of epilepsy, one of which is a molecular mutation, which results in irregular neuronal behaviour or migration of neurons. Although the main cause of epilepsy remains unknown, early diagnosis can be useful for treating epilepsy. Epilepsy patients can be treated with drugs or surgical procedures [6]. However, these methods are not fully effective. Unfortunately, seizures that cannot be completely treated medically limit the active life of the patient. In these cases, patients cannot independently work and do some activity. This leads to social isolation of individuals and economic difficulties. Early prediction of epileptic seizures ensures enough time before it actually occurs; it is very useful because the attack can be avoided by the drug. Epileptic seizures have four different states: the preictal state, which is a state that appears before the seizure begins, the ictal state that begins with the onset of the seizure and ends with an attack, the postictal state that starts after ictal state, and interictal state that starts after the postictal state of 1st seizure and ends before the begin of preictal state of consecutive seizure. Similarly, seizures can be predicted by detecting the beginning of the preictal state.

Observing brain activity via the EEG (electroencephalogram) has turned into an imperative instrument in the detection of epilepsy. The electroencephalogram recordings sufferers affected by epilepsy demonstrate two classes of atypical activity 1. ictal, the activity recorded for the duration of an epileptic seizure; and inter-ictal, abnormal alerts recorded all through epileptic seizures; Electroencephalogram signature of ictal period of an epileptic seizure is consists of a non-stop discharge of polymorphic waveforms of variable frequency and amplitude, rhythmic hyper synchrony, electro cerebral inactivity or sharp and spike wave complexes determined over a length longer than the common length of those abnormalities throughout the inter-ictal intervals. The Electroencephalogram signature of an inter-ictal activity is occasional temporary waveforms, as remote spikes; sharp waves spike or spike-wave complexes.

1.3. Machine Learning in Healthcare

With the rise of Machine Learning (ML), it plays a major role in the field of health care. Powerful ML usage empowers healthcare professionals in better decision-making, distinguishing patterns and innovations, and enhancing the productivity of research and medical trials. Machine learning enables building models to rapidly examine data and convey results, leveraging the ancient and real-time data.

1.3.1. Machine learning

Machine Learning is an umbrella term (make computer to act intelligently) for techniques that can figure out how to make forecast, estimations or recognize certain pattern from the information or data. Rather than being explicitly programmed to carry out a task, the computer utilises algorithms that can draw conclusions beyond experiences and enhance its accuracy throughout the system.

A machine learning (ML) algorithm is essentially a process or sets of procedures that helps a model adapt to the data given an objective. An ML algorithm normally specifies the way the data is transformed from input to output and how the model learns the appropriate mapping from input to output. Thus

ML algorithm = model + learning algorithm.

The model specifies the mapping function and holds the parameters while the learning algorithm updates the parameters in an effort to help the model satisfy the objective. Machine learning is split into three categories: Unsupervised, Supervised and reinforcement learning.

- **Supervised Learning:** It is a learning in which there is knowledge about the data or prior knowledge about class. It is further categorized into classification and regression.

- **Classification:** Classification problem is the point at which class variable is a group or a category, such as “black” or “white” or “Yes” and “No”. “Black” or “white”.

- **Regression:** Regression problem is the factor at which class variable is a real value, such as “height” or “Rupees.”

- **Unsupervised Learning:** Unsupervised learning is the learning in which there is no knowledge about the data or doesn't have prior knowledge about class. It is further categorized into two parts namely association and clustering.

- **Association:** Association rule is a market based analysis problem. Framing the association rule (X->Y) between set of items. Suppose if someone purchases item X then what is the probability that Y also go with it.

- **Clustering:**

  It is unsupervised learning technique. Clustering is a set of data items which are similar to one another in the same group and dissimilar to the items within different groups. Clustering helps divide data into a number of subsets. Each of these subsets contains data that is similar to one another. These subsets are known as clusters. Clustering only uses the input data, to determine patterns, anomalies or similarities in its input data. There are various clustering algorithm has been designed to perform clustering namely partitioning, hierarchical, grid-based and model based algorithms. Good clustering algorithm aims to create clusters whose:

- intra-cluster similarity is high (The data that is present inside the cluster is similar to one another)

- inter-cluster similarity is less (Each cluster holds data that isn't similar to the other)

- **Reinforcement Learning:** In Reinforcement learning a computer program will cooperate with a dynamic situation in which it must play out a specific purpose, (for example, playing a diversion with an opponent or driving a bike). The program is provided feedback in phrases of rewards and punishments because it navigates its problem space. Using this, machine is trained to make particular decisions.

There are various applications in which machine learning is used such as financial services, health care, oil and gas, marketing and sales, government and transportation. The current research focuses on the application of health care.

1.3.2. Applications of Machine Learning in Healthcare

As healthcare produces substantial information, the challenge is to acquire this data and efficiently use it for analysis, prediction and treatment. Powerful machine learning usage

empowers healthcare professionals in better decision-making, distinguishing patterns and innovations, and enhancing the productivity of research and medical trials. Machine learning enables building models to rapidly examine data and convey results, leveraging the ancient and real-time data. With machine learning, healthcare carrier companies could make better decisions on patient's diagnoses and treatment alternatives, which lead to usual development of healthcare services. Formerly, it turned into challenging for healthcare professionals to gather and examine the huge extend of data for efficient predictions and treatments since there were no advances or devices accessible. Now with machine learning, it's been comparatively easy. Machine learning algorithms can likewise be useful in giving crucial measurements, ongoing information and progressed examination as far as the patient's disease, family history, clinical trial information, and so on to specialists.

There are numerous regions where machine learning can be applied to trade the future of healthcare.

1.3.2.1. Disease Prediction

The contemporary approach to healthcare is to prevent the sickness with early intrusion rather than go for a treatment after analysis. Historically, doctors or physicians use a hazard calculator to evaluate the possibility of disease expansion. These calculators use essential statistics which includes medical conditions, demographics, life workouts and more to calculate the probability of growing a certain disease. Such calculations are achieved by using equation-primarily based mathematical methods and equipments. The challenge here is the low accuracy rate with a similar equation-based approach. But with latest development in technology consisting of huge facts and machine learning, it is feasible to get more exact results for disease prediction. Doctors are collaborating with analysts and computer researchers to develop better tools to predict the diseases. Specialists within the area are working on the methodologies to recognize, create and fine-tune machine learning algorithms and models which can convey accurate predictions. To build up a strong and more precise machine learning model, we can use information collected from research performed, medical health records, patient demographics and other sources. The distinction among traditional and machine learning approach for prediction of disease is the range of dependent variables to consider. In a conventional method, consider small number of variables that you can count on your figure including age, gender, height, weight and more (due to computational limitation).

Whereas machine learning is processed on computing devices can consider a massive number of variables, which ends up in a better accuracy of healthcare data.

1.3.2.2. Drug Discovery

Drug discovery and progress is very expensive and time-consuming work. Typically, a new drug improvement takes more than 10 years to get into a marketplace and prices approximately around 2.6 billion dollars. A drug discovery initiative is aimed at locating a compound that reacts with the centred molecules of the frame, inflicting ailment to therapy. Yet, there is an incredible possibility that the centre or supporting medication compound responds to non-focused molecules within the body adversely, that may probably cause threatening and hazardous aspects outcomes. As pharmaceutical organizations can't forecast a potential medication compound impact on focused and non-focused molecules using conventional computational technologies, the possibilities of drug failure are higher in medical trials. This situation makes drug discovery very expensive and time-consuming process. In this case, machine learning is used for better prediction and can save a lot of assets. Machine learning primarily based technique (considering the massive quantity of clinical data for authorised and failed drugs) to recognize a toxic compound that may cause symptoms can save numerous assets by going into clinical trials. Around ninety percent of drugs can't make it through the trial procedure. By using automating the compound molecules response tactics using machine learning, pharmaceuticals can enhance the drug discovery and improvement technique and reduce the time-to-marketplace, automating the drug discovery procedures can reduce costs by about 70%.

1.3.2.3. Electronic Health Records

It includes entire clinical and health information in a single device to make certain data availability and accessibility. Machine Learning based EHR Model Transfer approach allows to observer predictive models across different EHR structures. These models can be trained using datasets from one EHR and can be used to predict results for another framework. These frameworks have heterogeneous information sources, with records that are available in many forms-unstructured and structured, such as snap shots, text, scientific imaging and etc. Storing these data is not a challenge, but it's far hard to deploy this information for evaluation and predictions due to inconsistent formats. These are the few possible regions where Machine Learning can assist the healthcare industry out of numerous scenarios. With

machine learning applications, healthcare and medicine phase can proceed into a new realm and completely remodel the healthcare operations.

1.4. Optimization using Evolutionary Approaches

Optimization is a technique during which the most ideal estimations of selection variables are acquired under the specified set of constraints and according to a particular optimization cost function. The most ordinary optimization technique applies to a design as a way to maximize the possible reliability or minimize the cost or another specific goal. Fields of decision-making, industry, technology and engineering are all rich in issues that involve the usage of optimization method. Because, maximum real world optimization issues appears to be both on a very basic level and essentially hard, research into better algorithms stays precious and continues, so that, it is straightforward to assure to discover the best solution using an proficient optimization algorithm. These days, there exist a lot of optimization algorithms that work using heuristic-based and gradient-based search strategies in stochastic and deterministic contexts. Which will extend the applicability of the optimization method to various problem domains; physical and natural principles are mimicked to expand robust optimization algorithms. Ant Colony Optimization (ACO), Differential Evolution algorithm (DE), Particle Swarm Optimization (PSO) and Genetic algorithm (GA) are few optimization algorithms. Over the most recent decade, evolutionary algorithms have been extensively utilized in various problem domains and effectively finding the best optimal solutions.

To enhance the performance of the algorithm various machine learning technique have been used in evolutionary algorithms. Machine learning techniques consist of interpolation and regression, principle component analysis, statistical methods, artificial neural networks, reinforcement learning, support vector machine etc.

The work focuses on mainly three evolutionary algorithms namely Particle Swarm Optimization Algorithm (PSO), Differential Evolution Algorithm (DE) and Genetic Algorithm (GA).

1.4.1. Genetic Algorithm (GA)

This technique was introduced in 1970 by John Holland, which is inspired of natural selection that belongs to the bigger class of evolutionary algorithms. It is typically used to produce high-quality solutions to optimization. The first step is to generate a populace of random solutions. The subsequent stage is to choose two parents to perform recombination to

generate an offspring. The selection function ought to vary from being a random selection procedure to a fitness proportional scheme. There are different crossover strategies that can be used. A few common crossover techniques are uniform crossover, *k*-point crossover and binomial crossover. Once one or more offspring are generated, they are mutated the use one of the various mutation operations. Subsequently, a mutated offspring is comparison with its parent in the populace. The offspring can replace its parent if it has a higher fitness and be a part of the populace for the next generations. This entire method is repeated till the stopping criterion is met. The best individual of the population is returned. This individual represents the best solution. The main operators of the genetic are as per the following:

- **Selection:** In the selection stage of GA, offspring's are selected from the mating pool that is further involved for crossover operation. Selection of fittest parent is crucial for generating fittest next generation after crossover. Fittest parents are crossed (multiple way: one way, two-way etc.) new offspring's are generated. The probability of crossover is empirically set based on the problem that is being tried to be solved.
  Parent Selection: There are various algorithms for Parent Selection:

  - **Roulette Wheel:** The parent with largest proportion in the sample has the higher probability of being selected as the fittest parent.
  - **Stochastic Universal Sampling:** Same as roulette wheel but fewer complexes. Only one runs (iteration) to select all the possible fittest parents.
  - **Tournament Selection:** For some fixed iteration select K best individuals. For instance : 24 → 12–12 groups → best 4- best 4 groups → best 2-best 2 groups → 1–1 (winners)
  - **Rank Selection:** Similar to roulette but suitable when all the parents in the current population have same fitness score. Ranking them would help find the fittest.

- **Crossover:** It is a probabilistic methodology that swaps the component of the solution with some other in solution representation or chromosome. The principle role is to give blending of convergence and solutions in a subspace.
  Crossover: There are various ways for performing Crossover:

- **One way crossover:** Parents are crossed at one point. Half of one parent is appended of half of another parent. Vice-versa.

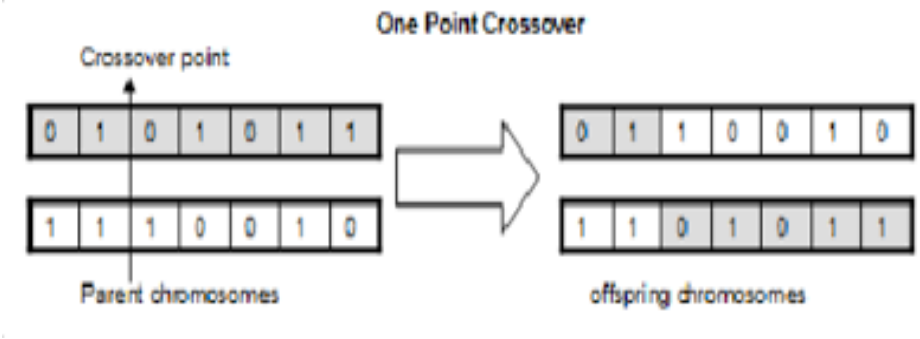

Figure 1.1(a): Representation of single point crossover

- **Multi-way crossover:** Generalization of one point crossover. Generating offspring by swapping alternating sections of 2 parents.

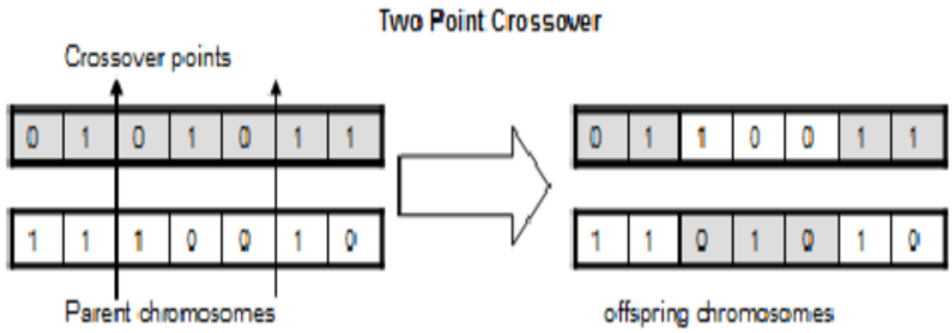

Figure 1.1(b): Representation of two point crossover

- **Universal Crossover:** In this method, every gene of either parent tested with a probability of either being swapped or not. This method brings in high diversity in the new offspring's.

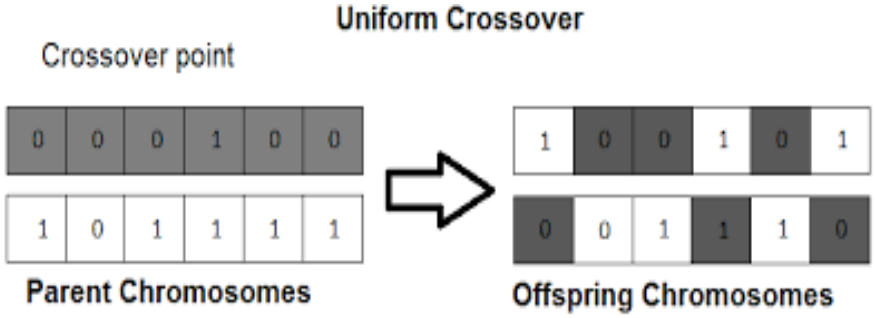

Figure 1.1(c): Representation of uniform point crossover

- **Mutation:** It is achieved by flipping the part one solution randomly, which increases the variety of the populace and provides a mechanism to getting away from a local optimum.

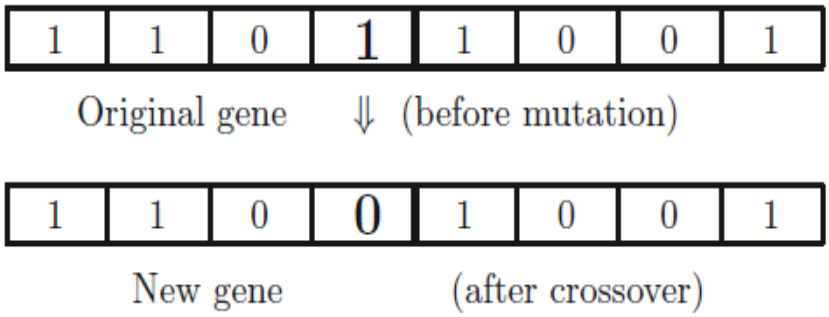

Figure 1.1(d): Representation of single site mutation

Choice of parameters in GA

Choice of parameters is one of the most important issues in genetic algorithm. The probability of crossover *(pc)* is high, usually in the range of 0.7 to 1.0. Whereas the probability of mutation *(pm)* is small, usually in 0.001 to 0.05. Crossover occurs sparsely, if *pc* is too small which is not efficient for evolution. If the *pm* is too high, the solutions could still jump around even if the optimal solution is approaching. And also right population size also plays a significant role. There is not sufficient evolution going on if the size of population is small and there is possibility that whole population wiped out.

1.4.2. Differential Evolution algorithm (DE)

This technique was developed in 1996 by R. Storn and K. Price. It is robust and fast optimization technique that uses a stochastic, populace based search methodology. DE is used for the real value problems. DE includes three main operator namely mutation, crossover, and selection. Firstly create a mutant vector (donor vector) using mutation by randomly selected three vectors; secondly, crossover is done between donor vector and parent vector to create trail vector. While the trial vector is formed, selection is performed to maintain only one of the vectors. The simplest way is to preserve the best fitness vector. In different terms, if trial vector has less fitness then only the parent vector will take place. If not, the trial vector exchanges the parent vector at once and will become suitable for selection inside the creation of the subsequent donor vector. The essential components of the differential evolutionary algorithm are as per the following:

- **Mutation:** For each vector x_i at time t, first randomly choose three different vectors x_a, x_b and x_c at t .The mutation is done to generate donor vector v_j. Donor vector are generated by including a weighted difference of two populace vector to a third vector.
- **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 $pCR \in [0, 1]$. There are two methods to perform crossover: Binomial and Exponential crossover.
    - **Binominal Crossover:** It 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.
    - **Exponential Crossover:** Within the exponential scheme, a section of the donor vector is chosen, and this section starts with the random integer and length that can encompass many additives.
- **Selection:** It is same as that used in genetic algorithms. Fittest value is chosen and, for the minimization issues, the minimal objective cost.

Choice of parameters in DE

The choice of parameters is essential. Both parametric studies and empirical observations propose that parameter values need to be best-tuned. The scale component *F* is the most important one. Even though $F \in [0, 2]$ is acceptable in hypothesis, $F \in (0, 1)$ is more proficient in practice. In fact, *F* perfect choice is usually in range of 0.7 to 0.9. The probability of crossover (*pCR*) is in the range of 0.1 to 0.8, and the best choice *pCR* is 0.5. The population size should depend upon the dimensionality *d* of the problem.

1.4.3. Particle Swarm Optimization (PSO)

This technique was introduced in 1995 by Dr. Kennedy and Dr. Eberhart, which is motivated from the social behaviour of flocking of birds or fish schooling. The first step is to initialize population of particles. Every particle is initialized with a random position and velocity. For every particle, fitness cost is evaluated. Every particle is pulled closer to the position of the current ***gbest*** (global best) and its own ***pbest*** (personal best), while in the mean time it has an

inclination to move randomly. When location attains by the particle is higher than before found locations, it updates that place of particle with the new current best. During iteration, there is a current best for all particles at time *t.* The purpose is to discover the ***gbes*t** among all the current best solutions until the target no longer enhance or after a specific range of iterations.

The two essential steps in PSO are updated the position and the velocity. The velocity is updated on the premise of momentum or inertia term, cognitive term, and the social term. Thus in each iteration, velocity value of particle is updated towards the attainment of local best and global best. The essential components of the particle swarm optimization algorithm are as per the following:

- **Velocity Update:** Let v_i and x_i be the velocity and position for particle *i* respectively. The velocity is update by using equation (1):

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

Inertia Component — Cognitive Component — Social Component

Here v_i^t and x_i^t is the velocity and position for particle *i* at *t* time, *w* is the inertia weight, $\alpha_1\alpha_2$ are the acceleration constant, $\beta_1\beta_2$ are the random values, 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.$\alpha_1\beta_1[pbest - x_i^t],\ is$ called as the cognitive component, shows the finest solution of individual particle and $\alpha_2\beta_2[gbest - x_i^t]$ called the social component, shows the best solution of all the particle(swarm).

- **Particle Update:** As soon as the velocity for each particle is calculated, the particle position of updated by using equation (2):

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

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

Choice of parameters in PSO

Inertia weight (w) usually in the range of [0 to 1.2], acceleration constant ($\alpha_1\alpha_2$) [0 to 2], and random values ($\beta_1\beta_2$) are in the range of [0 to 1]. The population size should depend upon the dimensionality d of the problem.

1.5. Road Map of Thesis

Chapter 1(Introduction) discusses the brief introduction to the healthcare, two disease namely breast cancer and epileptic seizure, machine learning, role of machine learning in healthcare, different applications of machine learning in healthcare, optimization using evolutionary approaches.

Chapter 2(Literature Review) summarizes the work done by the authors in the field of two diseases namely breast cancer and epileptic seizure. The survey is soughed year wise, brief introduction about the technique, nature of the objective function and stopping criteria is discussed. In the end, the chapter identifies the gaps among the techniques so that they can be addressed in the proposed method and the summarization of various approaches is described in the form of the table.

Chapter 3 (Problem Definition) includes the basic definition of the clustering algorithm. Along with the problem statement description of the problem is also included in the chapter. The clustering validation metric are discussed that the used to validate the proposed algorithm and verify the quality of clusters formed by the algorithm.

Chapter 4 (Objectives) discusses the objectives of the work.

Chapter 5 (Proposed Methodology) discusses the proposed algorithm along with flow chart. The pseudo code of the proposed technique is also included at the end of the chapter.

Chapter 6 (Simulation and Results) gives the brief description of the data sets that are used to test the proposed algorithm. In the second section, results obtained are shown in the tabular form.

Chapter 7 (Conclusion and future scope) discusses the conclusions drawn from the results and analysis chapter. In the future scope, the directions are given in which the work may be extended.

Chapter 2

Literature Review

Numerous contributions have been explored with the aid of the researchers in the field of health care, with each methodology being categorized according to its algorithm type. This chapter categorizes the Literature survey in the fields of two diseases namely Breast cancer and Epileptic Seizures.

2.1. Survey on Breast Cancer Disease

Breast cancer, is a kind of tumour that often influences women. After lung cancer it is the second most widespread disease across the world and the maximum incessant cause of death. Some of the former techniques which were used for the detection of Breast cancer are discussed:

Majali et al. [7] introduced an analytic system using association approach and classification technique using Data Mining. Frequent pattern algorithm in association rule is used to discover the pattern of malignant and benign. They used Decision tree (DT) algorithm for the prediction of the possibility of cancer in term of age. They implement FP growth algorithm to improve the performance of algorithm by generating the frequent item set without candidate generation. On Wisconsin dataset the classification accuracy of proposed method is 90%.

Sentruk et al. [8] examined the overall performance of seven different classification prediction models namely logistic regression, Discriminant Analysis (DA), K- Nearest Neighbor (KNN), Decision Tree (DT), Naive Bayes (NB), SVM and Artificial Neural Networks (ANN) for diagnosis breast cancer through Rapid Miner Tool. They concluded that Support Vector Machine (SVM) algorithm perform better than the other six algorithm. The accuracy achieved by SVM algorithm is 96%.

Sivakami [9] proposed an ailment reputation prediction model by making use of a hybrid technique of SVM and DT. To alert the seriousness of the disease the procedure of the framework comprises of two fundamental parts to be specific data treatment and choice extraction, and DT –SVM. They measure the performance of the proposed version with IBL-

SMO (instance based learning sequential minimal optimization) and naive bayes, and show that proposed algorithms performs well as compared to the other algorithm with 91% accuracy.

Venkatesan et al. [10] examined the breast cancer records using various classification algorithms such as Alternating Decision Tree (ADT), J48, Best First Tree (BF Tree), CART (classification and regression tree) with the help of tool (WEKA). The classifier has implemented for cross validation which makes use of 10 folds in which 9 folds for training purpose and rest for testing purpose and percent cut up uses 2/3 of the data for training and 1/3 for testing. The outcomes revealed that the J48 classifier performs well in terms of accuracy.

Shah et al. [11] conducted the comparison between three classification algorithms namely K-NN, Naive Bayes and Decision tree. The experiment is conducted through WEKA tool. They concluded that Naive Bayes (NB) is better in comparison to other algorithm because it take least time i.e. 0.02 second and providing highest accuracy.

Soumadip et al. [12] implemented different classification techniques which include SVM and MLP using back propagation neural network on Wisconsin dataset that are collected from UCI repository for detection of breast cancer. They finalized that SVM classifier has the ability to improve the conventional classification algorithm.

Elouedi et al. [13] suggested a hybrid detection method of breast cancer primarily based on Clustering and Decision tree. They have done feature extraction of malignant instance and then k-mean clustering is applied to the malignant instance. After that to every cluster they applied C4.5 (Decision tree) and compare the accuracies. They combined the results of both malignant and benign are applied to C4.5 algorithm to discover out the better consequences of classification based totally on the accuracies and confusion matrix. They acquire better outcomes than those acquired with the original dataset.

Kathija et al. [14] applied confusion matrix accuracy and 10 cross validation on Wisconsin diagnosis breast cancer dataset to locate the smallest subset characteristics and can make certain exceptionally accurate ensemble type of breast cancer either malignant or benign. For the breast cancer data the best classifier is support vector machine and naive bayes classifier.

Dumitru et al. [15] applied Naive Baye classifier on Wisconsin prognostic breast cancer dataset containing 198 numbers of patients and a binary decision class: recurrent event having 47 numbers of instance and 151 instances are non recurrent events. 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. [16] applied the supervised fuzzy clustering technique on breast cancer dataset. The experimental result shows the accuracy obtained by supervised fuzzy clustering method is 95.57%.

Chang et al. [17] made a study for predicting cancer by means of various algorithm particularly logistic regressions, Neural Network, Decision tree and genetic algorithm. Creating breast cancer classification model they are concerned with 10 attribute. The experimental consequences showed that logistic regression model has highest accuracy for predicting breast cancer than other applied techniques whereas decision tree has lowest accuracy. Moreover, genetic algorithm has highest accuracy in creating adequate classification rules and classification of breast cancer.

Alshammari et al. [18] implemented four classifications techniques namely two types of Artificial Neural Network i.e. Radial Basic Function (RBF), Multilayer preceptron, Naive Bayes and C4.5 for predicting the survivability of breast cancer. The experimental outcomes revealed that the accuracy (0.893), sensitivity (0.891) and specificity (0.985) of C4.5 (Decision tree) were higher than other for predicting the survivability of cancer.

Pour et al. [19] build a contrast between neural network classification methods with model based mining methods for detection of cancer. The experimental outcomes revealed that by adding the ensemble technique can enhance the consequences of neural network method and model based data mining method. Moreover, neural network with ensemble technique had maximum accuracy rate as compared to Model based data mining methods.

Rajesh et al. [20] suggested an algorithm for analysis of breast cancer for both malignant and benign. They used C4.5 (Decision tree) for predicting breast cancer. The outcomes demonstrate that C4.5 (Decision tree) had 93% accuracy for detecting breast cancer.

Hota [21] applied various techniques particularly DT (Decision tree) and ANN (Artificial Neural Network), statistical, unsupervised ANN techniques used for breast cancer classification dataset. On this work distinctive models are blended to perform ensembling. The results show that accuracy of ensembling model is higher than the individual model.

Pritom et al. [22] predicts whether the cancer is malignant and benign. The dataset was collected from of the UCI repository having 35 attributes. They implemented various algorithms such as SVM, Naive Bayes and C4.5 classification algorithm. The outcomes of these algorithms, Naive Bayes, SVM and C4.5 has 73.73%, (75.75%) and (67.17%) accuracy respectively. Support vector machine provide better performance.

Fung et al. [23] applied decision tree algorithm on breast cancer datasets that was collected from Leiden University. The informational collections have 574 patients who've surgery at that hospital. By initial analysis of three years they create the repetition of breast cancer by using a DT (Decision tree) algorithm. The accuracy obtained by the Decision tree is 70%.

Joshi et al. [24] implemented various clustering and classification technique to create a pattern of breast cancer patients. Various classifiers are used to finding the healthy patients. The 47 classification algorithms were used for identifying healthy individuals from sick patients. Their consequences revealed that, 13 techniques within those 47 techniques were same (76% healthy people and 24% sick people). The following techniques are: GD, J48, AdaboostM1, Multilayer perceptron, Classification via Regression, simple Logistic, Bayes Net, Multi Class Classifier, SMO and Attribute selected technique.

Padmavathi et al. [25] used different technique namely RBF (Radial Basis Function), MLP (Multilayer perceptron) and Logistic regression and 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.

Salama et al. [26] measured the confusion matrix and accuracy based totally on 10 fold cross validation method of various classification approaches such as Decision tree (J48), MLP, IBK and SMO for detecting cancer in three different datasets of breast cancers Their outcomes revealed that, the aggregate of IBK, J48, MLP, and SMO gives the maximum accuracy as

compared to single individual technique in all three datasets for detecting of breast cancer is either malignant or benign.

Saleema et al. [27] presented a model for determining important response variable which include stage of cancer, patient survival and patient age at diagnosis by concerning the standard classifiers. The model had three stages namely: pre-processing, problem specific processing and classifier namely Naive Bayes, K-Nearest Neighbour and Decision tree for predicting purposes. Their outcome revealed that, decision tree algorithm is better than the other classifiers.

Mittal et al. [28] presented an effectively hybridized classifier which is build by combining supervised classifier referred to as stochastic gradient descent (SGD) with an unsupervised classifier self –organizing maps (SOM) for detecting breast cancer. They evaluate their results with three supervised ML strategies specifically DT (Decision Tree), RF (Random Forest) and SVM (Support Vector Machine). The hybrid model builds up by integrating SOM with stochastic gradient descent gave 83.52% accuracy over training set and 83.68% accuracy over the testing.

Dubey et al. [29] 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.

Table 2.1: shows the summarization of various approaches

Author	Year	Methodology	Results
Abdulhamit subasi et al. [30]	2017	Rotation forest model	Rotation forest classification model obtained the accuracy 99.48%.
Jimin Guo et al. [31]	2017	Decision tree	The Decision tree classifier is used to predict whether or not a patient developed early disorder recurrence and approximated accuracy is 70%.
			Accuracy obtained by these three

Kate R.J. et al. [32]	2017	Artificial-neural network, Decision tree, Logistic regression	classifier are: Artificial neural network is 91.2% ,Decision tree is 93.6% Logistic regression is 89.2%
M.R. Mohebian et al. [33]	2017	Multi layer perception Decision tree, Support vector machine	Experimental results show that, the SVM perform better than the other approaches. The accuracy obtained by SVM is 78%.
A.I Pritom [34]	2016	Naive bayes algorithm, Support vector machine, Decision tree	After attribute selection support vector machine provide better performance.
Walaa Gad et al. [35]	2016	SVM- Kmeans	The accuracy obtained by SVM-Kmean is 99.8%.
Rashmi G.D et al. [36]	2015	Applied Naive bayes Prediction and Naive bayes Classification algorithms	The Naive bayes classifier algorithm has obtained maximum accuracy approximately 89-95%.
Jahanvi Joshi et al. [37]	2014	k-mean clustering, Expectation maximization, Farthest first clustering algorithm, HCM algorithm	K-mean algorithm gives more optimum outcome. The results showed that 83% of peoples are healthy and 17% of patients are sick.
A.M. Cvetkovic et al. [38]	2015	Support vector machine, Logistic regression, Decision tree, Naive bayes and Artificial neural network.	According to the following parameter such as accuracy (0.9657), sensitivity (0.991) and specificity (0.889) the decision tree classification give better results.
Ahmad LG et al. [39]	2013	Decision tree, Support vector machine, and Artificial neural network	Support vector machine have better accuracy i.e. 90.7% as in comparison to different algorithm.
C. Shah et al. [40]	2013	Naive bayes, K-Nearest Neighbour, Decision tree, Algorithm	Naive bayes algorithm provides better performance i.e. 92.99%.

Q. Fan et al. [41]	2010	Decision Tree, Artificial Neural Network, CHAID and QUEST	Their experimental results showed that decision tree C5.0 gives 71% more accurate values than the others.

2.2. Survey on Epileptic Seizures

Epilepsy is a typical interminable neurological issue, which is set apart by seizures. It is usually caused by excessive discharge of cortical cells from the brain. It is undetectable and may lead to paroxysms. Epileptic seizures cannot be cured but it can be controlled through proper medicinal drugs. Some of the former techniques which were used for the recognition of epileptic seizure in Electroencephalography signals are discussed:

Rasekhi et al. [42] have proposed an algorithm for seizure prediction with the help of univariate linear features. In [42], the authors used only six EEG channels in their proposed model and extracted 22 univariate linear properties. Thus, a 132-dimensional feature space is created. It was assumed that preictal time starts 10 to 40 minutes before the ictal state with a difference of 10 minutes. Prediction of epileptic seizures is considered by classifying a binary class that classifies test data into either preictal state or ictal state. On average, the prediction sensitivity after applying this algorithm is 73.90%.

Teixeira et al. [43] have proposed a model for prediction of epileptic seizures by choosing only six channels of EEG signals and have extracted 22 linear univariate features for every channel. The overall feature space expands to 132 dimensions. In [43], the authors have used only six electrodes for EEG data acquisition. The main purpose behind this minimum electrode selection is to set free the patient from wearing a large number of electrodes, as patients are often unwilling to wear so many electrodes on their scalp due to discomfort. Therefore, in order to give comfort to the mind, only six channels have been acquired and used for prediction purpose. The authors have selected these electrodes by using three different approaches.

Bandarabadi et al. [44] have proposed an algorithm to predict epilepsy seizures that can extend the life of epilepsy-affected patients. They have extracted spectral power features, and after suitable selection of features, features are passed into Support Vector Machines for classification. They have observed sensitivity of 75.8%; it means that their classifier has

predicted 66 seizures out of total 87. They have concluded that, by applying these methods, after reducing proposed features subset can improve seizure prediction performance. In [44], the authors have used wavelet method for prediction of seizures. They have extracted features including wavelet energy and wavelet entropy. Two or three channels have been selected for testing purposes on a dataset of six patients. Sensitivity has been reported as 88% with average anticipation time of 22 minutes.

Zandi et al. [45] have also proposed a model for predicting seizures using scalp EEG signals on the basis of zero crossings. The authors in [45] have computed the histogram of all intervals in a moving average window and have selected values from particular bins for observations. Once the whole process is completed, last 5 seconds of observations are compared with different reference sets of points, containing interictal and preictal states. They have measured a similarity index on the basis of variation Bayesian Gaussian mixture model of EEG data.

Yusof et al. [46] proposed a brand new mutation operation for rapid feature selection with the aid of GA depending on elitism of the allele. The fittest chromosomes were preserved by normal elitism in GA, which were then evaluated by utilizing the fitness function. The highest fit allele was 14 conserved and the evaluation of fitness of the allele executed primarily based on the frequency of the occurrences. The chromosome that underwent this mutation approach was found to have the highest fitness as it was created based on the fittest alleles. The proposed approach increased the search for the fittest chromosomes, and minimized the time taken for optimal convergence.

Guo et al. [47] implemented Genetic Programming to perform automatic feature extraction from the original database in the epileptic EEG classification process. The discriminatory performance of k-Nearest Neighbors (k-NN) classifier was enhanced and the input feature dimensionality was reduced. However, the main drawback of the method is that the Genetic Programming based feature extraction system is costly.

Wang et al. [48] proposed a hierarchical Electroencephalogram classification system that can be used for the epileptic seizures detection. The three steps of the proposed system were,

- Illustration of original EEG alerts using wavelet packet coefficients and extraction of features with the aid of the best basis-based totally on wavelet packet entropy.

- Use of Cross-validation approach and k-Nearest Neighbor classifier within the training degree of Hierarchical Knowledge Base creation.
- Computation of accuracy of classification and rates of rejection using the top-ranked discriminative rules from Hierarchical Knowledge Base.

Janjarasjitt [49] used wavelet to extract the characteristics of the epileptic EEG alerts during seizures and seizure free intervals. The log variance of wavelet coefficients of epileptic EEG data was used in classifying the epileptic EEG data. The computational results revealed significant difference between the log variance of wavelet coefficients of EEG data of seizures and that of non-seizure periods. Based on the results the authors have concluded that the region of the brain where epileptic EEG data were recorded has an influence on the analysis and diagnosis of epilepsy. The wavelet based scale variance is a log variance for wavelet coefficient of EEG signals. The classification could be done by utilizing this variance as a feature 33 in epileptic EEG data. In classification process, an unsupervised classification algorithm "k-Means" algorithm uses a squared error criterion was used. K-Means algorithm initiates with a random initial partition. Clustering was done until converge condition is obtained. The computational complexity of K-Means is linear to the wide variety of patterns i.e., O (n).

The following were the steps in the k-means clustering algorithm are as follows:

1. Initialize k cluster centers by randomly choosing k patterns.
2. Allocate each object to the closest cluster center.
3. Calculate the centers of cluster using the current cluster.
4. If the criteria are not satisfied, go to step 2 otherwise generate the cluster.

The performance of the classification was evaluated by two parameters, Sensitivity and Specificity.

Acharya et al. [50] examined the performance of five different classification models such as Fuzzy Neural Network, k-Nearest Neighbour, Decision tree and SVM for prediction of epileptic's seizures. They concluded that Support Vector Machine performs better as comparison to other model.

2.3. Research Gaps

- The major work in literature has been focused on classification and k-mean clustering for both the datasets (Breast cancer dataset and Epileptics seizures). No attention has been paid on the evolutionary approaches in these fields.

- Various classification techniques in literature have been applied on small datasets for prediction of Breast cancer dataset and Epileptics seizures. No work has been attempted for large input data.

Chapter 3

Problem Definition

The detection of diseases is a crucial and difficult task in healthcare. The recognition of breast cancer and epileptic seizures from numerous features or signs is a prime issue which is not free from false presumptions frequently followed with the aid of unpredictable effects. The healthcare enterprise gathers large amounts of breast cancer and epileptic seizures disease data that unfortunately, are not mined to decide concealed facts for effective diagnosing. As the quantity of stored data will increases, clustering play a vital role in extracting knowledge and finding patterns to provide better care and effective diagnostic capabilities.

3.1. Problem statement

In the current era, Medical Databases are so huge and complicated, clustering algorithms are of vital importance that it organizes the data, thereby generating patterns that can be further utilized for better analysis of diseases. As large data sets have become more common in biological and data mining applications, clustering is a significant challenge.

The problem relies on analysis of breast cancer and epileptic seizures diseases datasets utilizing efficient clustering algorithms by grouping data into related groups such that

- Intra cluster distance is minimum.
- Inter cluster distance is maximum.
- Free from the problem of local minima.

3.2. Validity measures 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. The validity measures used to quantify our proposed algorithm are DBI.

3.2.1. Davies-Bouldin (DB) Index

Donald W. Bouldin and David L. Davies introduced DBI in 1979. It is a used for evaluating clustering algorithms [51]. DBI validates the clustering algorithm using the 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 as shown in equation (3). Therefore lower the value of DBI better is the clustering algorithm.

$$DB = \frac{1}{c}\sum_{i=1}^{c} Max_{i \neq j}\left\{\frac{d(x_i)+d(x_j)}{d(c_i c_j)}\right\} \quad (3)$$

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 $d(c_i c_j)$ distance between the centroids. The way it is designed confirms the idea that no cluster should be similar to another one in the data set, and therefore best clustering scheme would minimizes the DBI value. The value of k (number of cluster) for which value of DBI is least is termed as the optimal value of k.

3.2.2. Dunn Index

J.C. Dunn introduced DI in 1974. It is a metric used for evaluating clustering algorithms whose aim is to recognize the set of clusters that are closely packed, having small variance value between the data points of a cluster, are well- separated, far apart from the dissimilar data points, i.e. means of clusters values have sufficient distance among them. For the given data set, higher the value of DI better is the clustering approach [52]. As shown in equation (4).

$$Dunn = min_{1 \leq i \leq c}\left\{\min\left\{\frac{d(c_i c_j)}{max_{1 \leq k \leq c}(d(X_k))}\right\}\right\} \quad (4)$$

d ($c_i c_j$) represents the inter-cluster distance, c is the number of cluster and $d(X_k)$ denotes the intra-cluster distance.

3.2.3. Cluster validity indices in terms of classification evaluation parameters

3.2.3.1. Accuracy

Accuracy is termed as correctness of classified instances to the aggregate number of instances. It is calculated as given in equation (5).

$$Accuracy = \frac{(tn+tp)}{(tn+tp+fn+fp)} \quad (5)$$

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-Breast cancer patient suffering from breast cancer.

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

3.2.3.2. Sensitivity

Sensitivity is the ratio of actual positive cases which are correctly identified. It is also known as true positive. It is calculated as given in equation (6).

$$Sensitivity = \frac{tp}{(fn+tp)} \quad (6)$$

3.2.3.3. Specificity

Specificity is the ratio of true negative cases which are correctly identified. It is also known as true negative. It is calculated as given in equation (7).

$$Specificity = \frac{tn}{(fp+tn)} \quad (7)$$

3.2.3.4. Precision

Precision is the ratio of actually true predicted instances out of the total number of true instances. The Precision is computed as in equation (8).

$$Precision = \frac{tp}{(tp+fp)} \quad (8)$$

3.2.3.5. Recall

Recall is the ratio of actual true instances out of all the items which are true. The Recall is computed as in equation (9).

$$Recall = \frac{tp}{(fn+tp)} \quad (9)$$

3.2.3.6. 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. It is calculated as given in equation (10). If the f-score value is 1 then it means perfect precision and recall and for 0 it means worst.

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

Chapter 4
Objectives

The objectives of the current work are as under:

1. To study the existing approaches related to the prediction and analysis of the diseases under consideration.
2. Application of clustering algorithms based on different evolutionary techniques on Wisconsin Breast Cancer dataset.
3. Comparative analysis of various evolutionary clustering approaches based on different evaluation parameters namely Davies Bouldin Index (DBI), Accuracy, Sensitivity, Specificity, Precision, Recall and F-measures.
4. Development of a novel partition based evolutionary clustering algorithm for Epileptic Seizures dataset.
5. Comparison of the proposed technique with the base evolutionary clustering approach in terms of DBI, Dun index and computational time.

Chapter 5

Proposed Methodology

In the process of extraction of useful information from the data set, clustering plays a significant role in finding the patterns. The approach is applied on the datasets namely Breast Cancer Wisconsin (Diagnostic) and Epileptic Seizure Recognition datasets collected from UCI Repository. The work is validated by using the cluster validation metrics- DBI and Dunn Index.

5.1. Genetic Algorithm based clustering

The work proposes a genetic algorithm based clustering named **W**isconsin Clustering Genetic Algorithm (WCGA) for solving breast cancer disease. Genetic algorithm (GA) is inspired of natural selection that belongs to the bigger class of evolutionary algorithms. It is typically used to produce high-quality solutions to optimization. The GA-clustering is based on the genetic algorithm steps which are discussed below:

5.1.1. 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. Every string is a series of real numbers demonstrating the *k* cluster centres.

Let *X*=5 where *X* be the number of attribute in a dataset and *k*=2 where *k* be the number of cluster. Then the chromosome

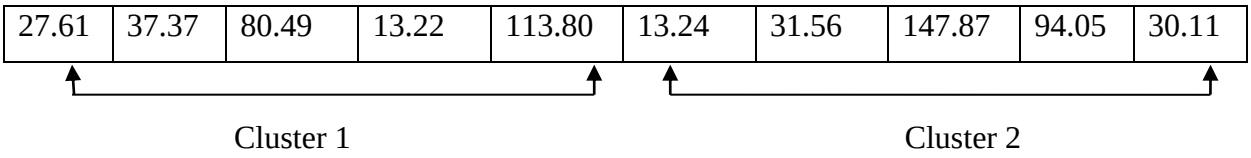

Figure 5.1: Solution Representation of WCGA

represents 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)].

5.1.2. Population Initialization

The *K* cluster centres encoded in each chromosome are initialized to *K* randomly chosen points from the data set. This process is repeated for each of the *P* chromosomes in the population, where *P* is the size of the population.

5.1.3. Fitness Function

The importance of the clustering evaluation is to partition a given data into clusters with the subsequent properties: similarities within the clusters and dissimilarities between clusters. DB is used for measuring the validity of cluster. As the lower DB index, better the quality of cluster by making more compact and separated clusters.

5.1.4. Selection

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

5.1.5. Crossover

It is a probabilistic methodology that swaps the component of the solution with some other in solution representation or chromosome. The principle role is to give blending of convergence and solutions in a subspace. In this two-point crossover is used with fixed crossover probability p_c. The parts of the chromosomes lying to the right of the crossover point are exchanged to create two offspring.

5.1.6. Mutation

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

5.1.7. Termination Criteria

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

5.2. PSO based clustering

The work proposes a PSO based clustering named **W**isconsin Clustering Particle Swarm Optimization (WPSO) for solving breast cancer disease. PSO 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 PSO-clustering is based on the particle swarm optimization algorithm steps which are discussed below:

5.2.1. 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 nX k size linear array representing the k cluster centroids in sequence. Every string is a series of real numbers demonstrating the *k* cluster centres.

Let X=5 where X be the number of attribute in a dataset and k=2 where k be the number of cluster.

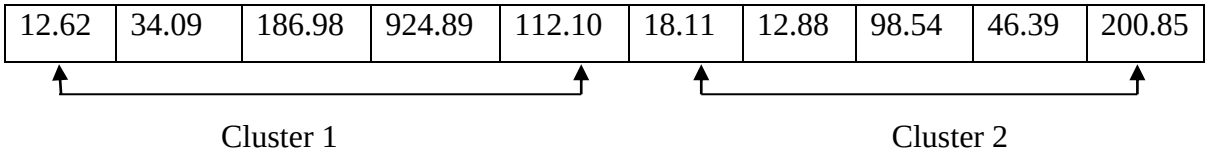

Figure 5.2: Solution Representation of WCPSO

represents the two cluster centers i.e. [(12.62, 34.09, 186.98, 924.89, 112.10) and (18.11, 12.88, 98.54, 46.39, 200.85)].

5.2.2. Population Initialization

Initialize positions vector ***X*** and associated velocity *V* of all particles in the population randomly. Here the position ***X*** of the particle is the center position of each cluster.

5.2.3. Fitness Function

The importance of the clustering evaluation is to partition a given data into clusters with the subsequent properties: similarities within the clusters and dissimilarities between clusters. DB is used for measuring the validity of cluster. As the lower DB index, better the quality of cluster by making more compact and separated clusters.

Evaluate particle's fitness evaluation with *pbest* solutionP_i. If current value is higher than P_i , then set P_i cost same to the current value and the P_i position equal to the current position. Examine fitness evaluation with the populace's overall previous best. If current value is higher than the P_g (particle global best), then reset P_g to the current particle's value and position.

5.2.4. Velocity Update

Let v_i and x_i be the velocity and position for particle *i* respectively. The velocity is update by the using equation (11):

$$v_i^{t+1} = wv_i^t + \alpha_1\beta_1[pbest - x_i^t] + \alpha_2\beta_2[gbest - x_i^t] \quad (11)$$

Inertia Component — Cognitive Component — Social Component

Here v_i^t and x_i^t is the velocity and position for particle *i* at *t* time, *w* the inertia weight, α_1, α_2 are the coefficients, $\beta_1\beta_2$ are the random values, *pbest* is the particle best solution and *gbest* is the global best solution. wv_i^t is the inertia component, $\alpha_1\beta_1[pbest - x_i^t],\ is$ called as the cognitive component and $\alpha_2\beta_2[gbest - x_i^t]$ called the social component.

5.2.5. Particle Update

As soon as the velocity for each particle is calculated, the particle position of updated by using equation (12):

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (12)$$

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

5.2.6. Termination Criteria

The complete procedure will continue (fitness function, velocity and particle updating) till the maximum number of iterations.

5.3. DE based clustering

The work proposes a Differential Evolution algorithm based clustering named **W**isconsin Clustering Differential Evolution (WCDE) for solving breast cancer disease. 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 DE-clustering is based on the differential evolution algorithm steps which are discussed below:

5.3.1. 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 nX k size linear array representing the k cluster centroids in sequence. Every string is a series of real numbers demonstrating the *k* cluster centres.

Let X=5 where X be the number of attribute in a dataset and k=2 where k be the number of cluster.

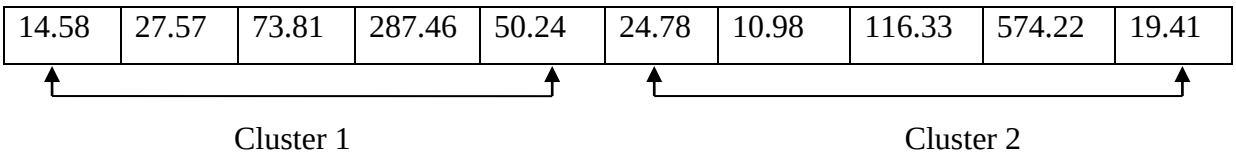

Figure 5.3: Solution Representation of WCDE

represents the two cluster i.e. (14.58, 27.57, 73.81, 287.46, 50.24) and (24.78, 10.98, 116.33, 574.22,19.41).

5.3.2. Population Initialization

The *K* cluster centres encoded in each chromosome are initialized to *K* randomly chosen points from the data set. This process is repeated for each of the *P* chromosomes in the population, where *P* is the size of the population.

5.3.3. Fitness Function

The importance of the clustering evaluation is to partition a given data into clusters with the subsequent properties: similarities within the clusters and dissimilarities between clusters. DB is used for measuring the validity of cluster. As the lower DB index, better the quality of cluster by making more compact and separated clusters.

5.3.4. Mutation

For each vector x_i at time t, first randomly choose three different vectors x_a, x_b and x_c at t .The mutation is done to generate donor vector v_j. Donor vector are generated by including a weighted difference of two populace vector to a third vector. As shown in equation (13):

$$v_i^{t+1} = x_a^t + F(x_b^t - x_c^t) \tag{13}$$

Here *F* is a parameter, represents the differential weight where $F \in [0, 2]$, v_j is donor vector generated by mutation.

5.3.5. 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 $pCR \in [0, 1]$. In this binomial crossover is used to perform crossover on each d variables or component by generating a consistently appropriated random number $r \in [0, 1]$, 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 shown below in equation (14):

$$u_{j,i}^{t+1} = \begin{Bmatrix} v_{j,i} & if\ r_i \le pCR, \\ x_{j,i}^t & otherwise, \end{Bmatrix} j = 1,2,3 \ldots . d. \tag{14}$$

5.3.6. Selection

Selection is same as that used in genetic algorithms. Trail vector is compare with the parent vector and choose the fittest value as shown in equation (15).

$$x_i^{t+1} = \begin{Bmatrix} u_i^{t+1} & if\ f(u_i^{t+1}) \le f(x_i^t) \\ x_i^t & otherwise. \end{Bmatrix} \tag{15}$$

5.3.7. Termination Criteria

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

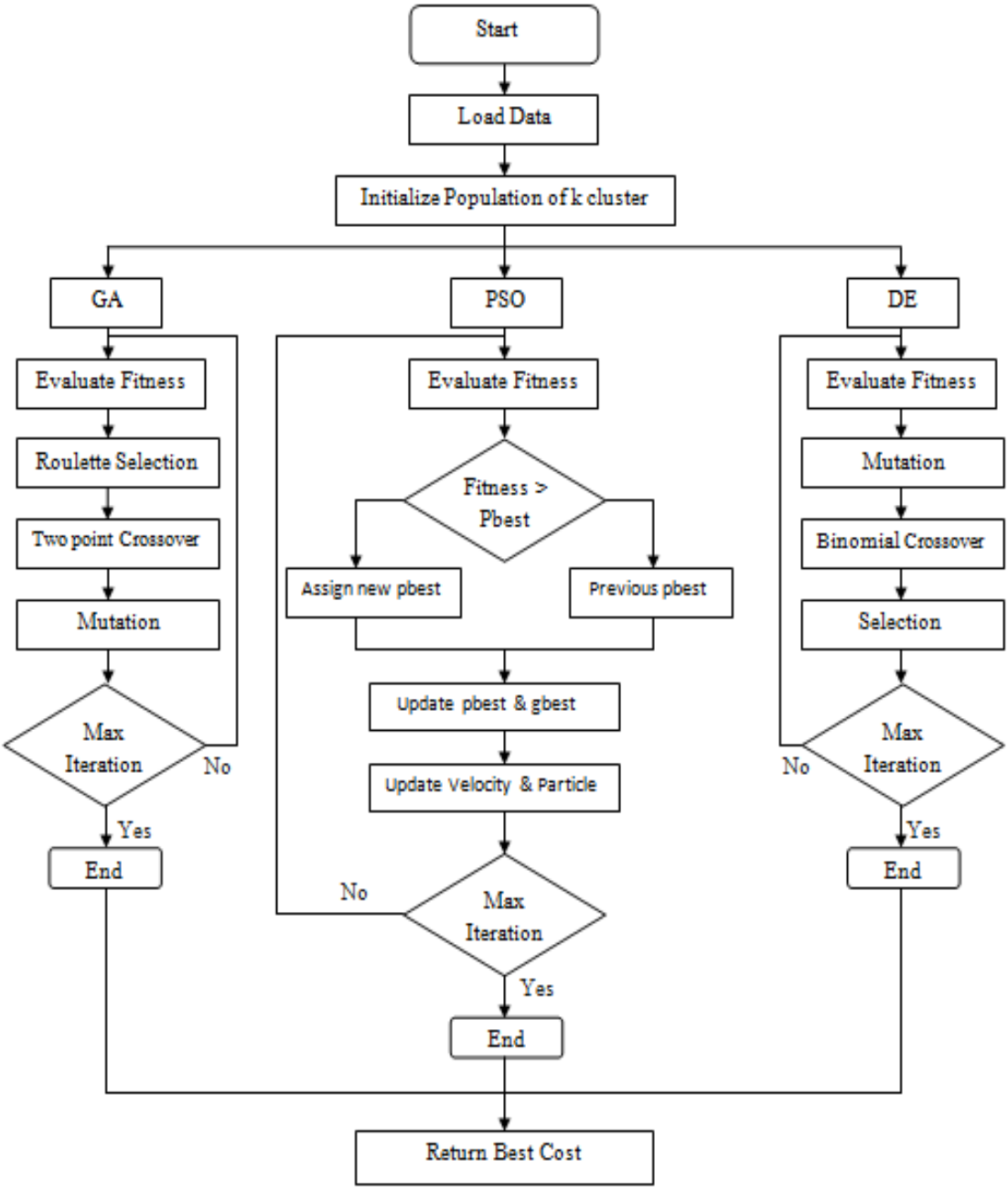

Figure 5.4: Workflow of Proposed Methodology

5.4. Partitioning based clustering

Partitioning method tends to be more efficient and easily adapted for huge datasets. These methods primarily based on greedy heuristics approach that are used iteratively to obtain a local optimum solution. In partitional technique, the data set is divided into the clusters using some technique. For the given database of n objects, a partitional algorithm partitions the data into 'k' (number of clusters, which are pre-decided), so that optimization of an objective function can be done. It uses an iterative relocation method that tries to improve the

partitioning by moving objects from one group to another. Each cluster has at least one object and each object belongs to only one cluster. The main objective of partition clustering is to split the data points into K partitions. Each partition will reflect one cluster. The technique of partition depends upon certain objective functions. Partitioning clustering algorithms decide a grouping solution by way of maximise the similarities among objects inside the same groups while minimise the dissimilarities between distinct groups. Partitioning clustering algorithms based on Evolutionary algorithm such as Differential evolution have been proposed to handle the issue of finding the optimal partition of a data.

The work proposes a novel partitioning based clustering using differential evolution approach named **P**artitioning Clustering using Differential Evolution (PCDE) and Differential Evolution Clustering is applied to the Epileptics Seizures Dataset and comparing the performance.

Partitioning Method

The novel partitioning technique divides the existing samples into distinct partitions within the feature space. Let A = {a1, a2, a3,...., an} be the pool of samples. The mean of A say, $\bar{x}_{\mathrm{m}}$ is computed as given in equation (16)

$$\bar{x}_{\mathrm{m}} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad \forall\, j \in [1: tuple_{size}] \tag{16}$$

where x_{ij} represents the j^{th} entry for the tuple of $a_i{}^{th}$ sample. The distance D = {d_1, d_2,..., d_n} is calculated by computing the distance of each sample $x_i \in A$ from the mean $\bar{x}_{\mathrm{m}}$. The minimum and maximum in D say (d_{min} and d_{max}) respectively is used to partition the samples in A. To this end, first the factors in D are sorted and then divided into two subsets D1 and D2 such that D1 contain the minimum value (d_{min}) and D2 contain the maximum value (d_{max}). For illustration, consider the randomly generated pool A of samples with two-dimensional feature space as shown below in Figure 5.5:

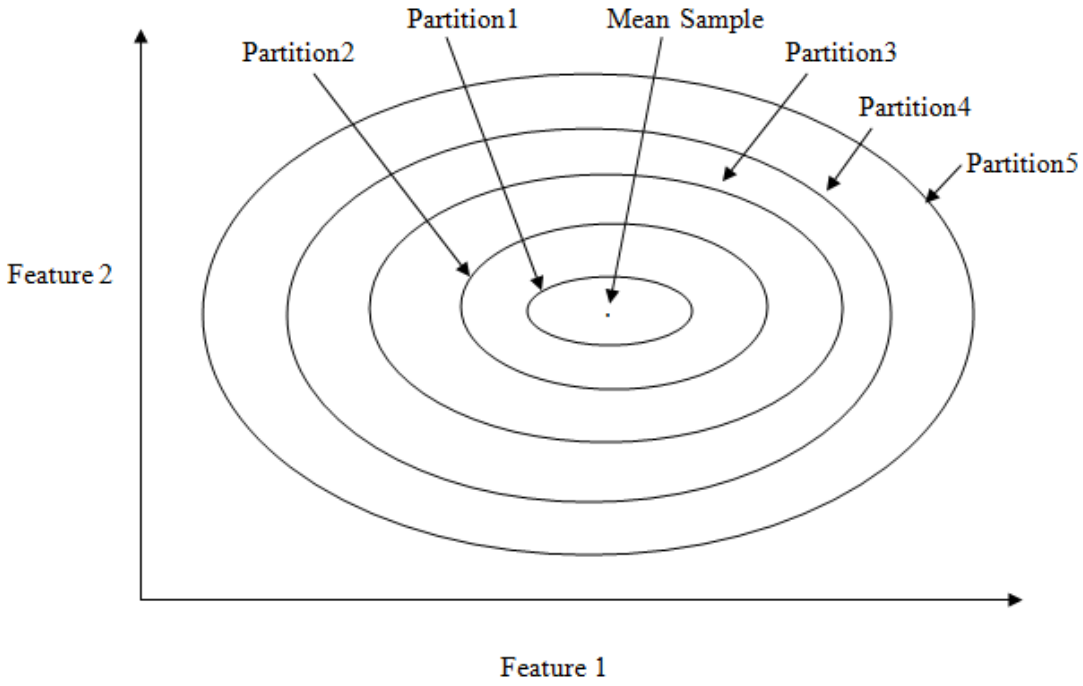

Figure 5.5: Illustration of the Proposed Methodology

Algorithm 1 Partitioning Clustering using Differential Evolution approach (PCDE)

Input : InputFile, K

Begin

X.Data=Read(InputFile)

X.Mean=mean(Data)

X.D= Euclidean_distance(X.Data, X.Mean)

X.D= Sort(X.D, 'ascending')

min= X.D(1)

max = X.D(Data_Size)

[BestSol, BestCost]= DE(X.Data, min, max,K)

End

DE(X.Data,min, max, K)

Input:- Pop_Size- Population Size, beta_max-upper bound of scaling factor, Max_Iteration-Maximum Iterations, pCR-Crossover Probability, beta_min-Lower bound of scaling factor.

Output:-BestSolution, BestCost along with the partition range of cluster.

Begin

```
Initialize(nVar, VarSize, Max_Iteration)
  Set VarMin= min_tuple(X)
  Set VarMax= max_tuple(X)
        For(i=1 to Pop_Size)
           Pop.solution=Initialize_Population(min,imax,K-2)
           Pop.cost= clusteringcost(pop.solution,min,max,X)
              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+□.*(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]=clusteringcost (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
```

For Partitioning based Clustering using DE approach, mean of the data *X.Mean* is calculated and measures the distances using Euclidean Distance *(X.D= X.Data, X.Mean),* sort the *X.D* in ascending order and select the min and max value. After that Differential evolution approach is applied to calculate the best cost and best solution.

DE, is the idea of a populace of individuals, where in every individual consists of a candidate solution and quality is indicated by fitness.

P_1	P_2	P_3	P_4	P_6	...	P_{n-2}

Figure 5.6: Solution representation for PCDE

For partitioning method, if there is n clusters (number of clusters) then n-2 partitions take place. In this work the value of n=5 so there is n-1 partitions it means it is divide into 3 partitions.

Chapter 6
Simulation Results

The chapter gives the results of various evolutionary based clustering approaches for different diseases along with the comparative analysis of these approaches based on various evaluation parameters.

6.1. Data Sets Used

The proposed techniques are applied on the following two disease datasets.

6.1.1. Breast Cancer Wisconsin (Diagnostic) Data Set

The data set consists of total of 569 instances with 32 being the number of attributes of the instance [53]. The data set contains no missing value. Each record is marked as one of the two diseases namely:-malignant and benign. The last column of the data set consists of the information of the belongingness of the instance to the particular disease, hence that column have being deleted while giving input to the proposed algorithm namely named **W**isconsin Clustering Genetic Algorithm (WCGA), **W**isconsin Clustering Particle Swarm Optimization (WCPSO) and named **W**isconsin Clustering Differential Evolution (WCDE).

6.1.2. Epileptic Seizure Recognition Data Set

The data set consists of total of 11500 instances with 179 being the number of attributes of the instance [54]. The data set contains no missing value. Each record is marked as 1, 2, 3, 4, 5 where 1 stands for Seizure activity recording, 2 stands for recorder the EEG from the location wherein tumour was located, 3 stands for the place of the tumour became in the brain and recording the EEG activity from the healthy brain area, 4 stands for the closed eyes, which means that once they recording EEG signal the patient had their eyes closed and 5 stands for the eyes opened, this means that once they recording EEG signal the patient had their eyes opened. The last column of the data set consists of the information of the belongingness of the instance to the particular disease, hence that column have being deleted while giving input to the proposed algorithm namely **P**artitioning Clustering using Differential Evolution (PCDE) and **D**ifferential Evolution Clustering (DEC).

Table 6.1: Description of Data Set (where, # stands for number of)

Data Set	#Instances	#Attributes	#Missing Value
Breast Cancer Wisconsin Diagnostic	569	32	No
Epileptic Seizure Recognition	11500	179	No

6.2. Results

The proposed Evolutionary approach is implemented using MATLAB (R2016a) on Intel core i5 processor, 4 GB RAM run on 64-bit O.S (Operating System). Each data set is made to run for 10 times, where each run consists of 100 iterations of the proposed evolutionary algorithm.

Table 6.2: The control parameters 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	c1	2	beta_max	0.8
mu	0.02	c1	2	Pcr	0.2

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, c1-Cognitive Coefficient, c2-Social coefficient, beta_min-Lower bound, beta_max-Upper bound, pCR-Crossover probability (parameters values tabulated in table 6.2)

6.2.1. Results of Wisconsin Breast Cancer Dataset

Table 6.3: The results of WCGA, WCPSO and WCDE evaluated on the basis of DBI and Computation time.

	WCGA		WCPSO		WCDE	
Runs	**DBI**	**CPU time**	**DBI**	**CPU time**	**DBI**	**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

The table 6.3 depicts that the results obtained from WCGA, WCPSO, WCDE based on validity measures namely DBI 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 0.93021 respectively. It is clear from the table that WCDE outperforms in term of the DBI that is 0.88622 and it also has lower CPU time that is 231.1172 in comparison to the WCGA and WCPSO.

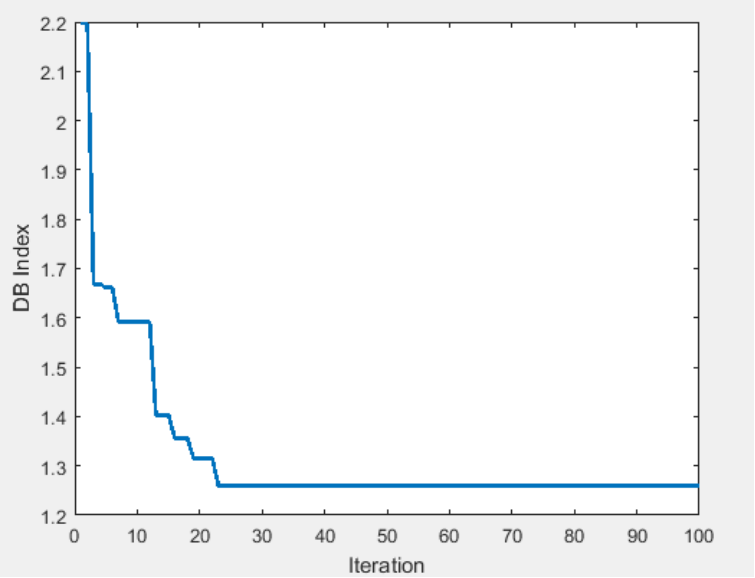

Figure 6.1(a): depicts the db-index vs number of iterations for WCGA approach

Figure 6.1(a) (WCGA) shows that graph plotted between number of iteration and DB Index values. In this graph at 23 iterations the results of WCGA goes stagnant and the DB index value is near about 1.2.

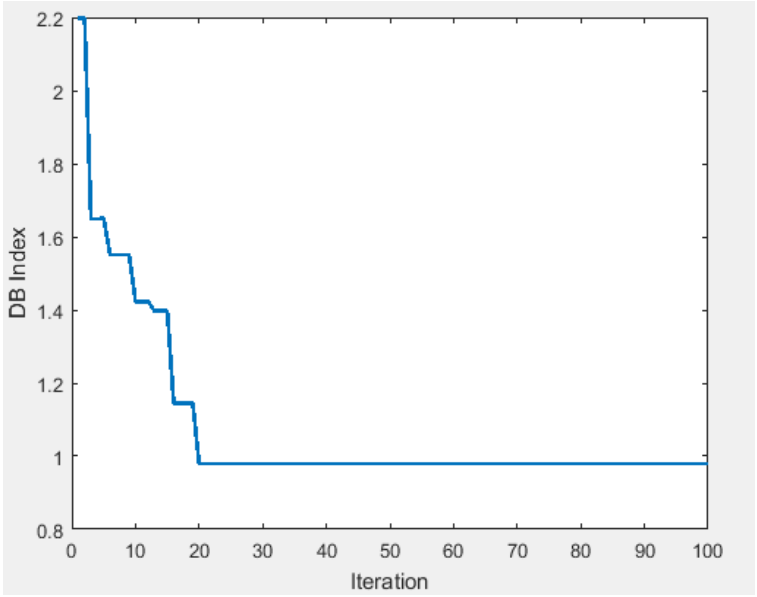

Figure 6.1(b): depicts the db-index vs number of iterations for WCPSO approach

Figure 6.1(b) (WCPSO) shows that graph plotted between number of iteration and DB Index values. In this graph at 20 iterations the results of WCPSO goes stagnant and the DB index value is near about 0.9.

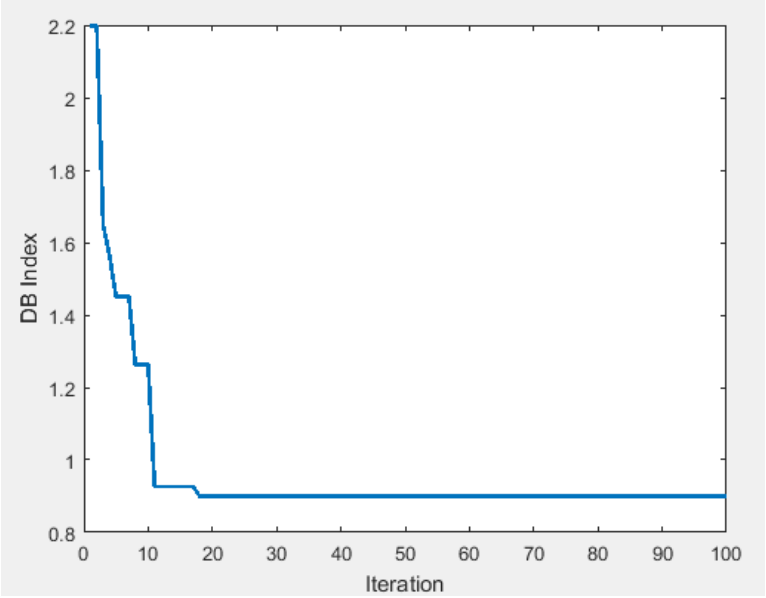

Figure 6.1(c): depicts the db-index vs number of iterations for WCDE approach

Figure 6.1(c) (WCDE) shows that graph plotted between number of iteration and DB Index values. In this graph at 17 iterations the results of WCDE goes stagnant and the DB index value is near about 0.8.

According to Figure 6.1 (a), (b) and (c) Shows that WCDE gives better results in term of DB index as compared to the WCPSO and WCDE. As DB index gives the optimal results in the lower values.

Table 6.4: The results of WCGA cluster validity (in terms of classification parameters)

WCGA						
No. of Runs	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**F-measure**
Run_1	0.6796	0.7302	0.7120	0.6570	0.7302	0.6916
Run_2	0.6077	0.7070	0.7560	0.6832	0.7070	0.6948
Run_3	0.7206	0.6858	0.5630	0.5726	0.6858	0.6241
Run_4	0.7118	0.6906	0.5462	0.5845	0.6906	0.6331
Run_5	0.5990	0.6547	0.6903	0.6172	0.6547	0.6353
Run_6	0.7678	0.6585	0.6515	0.6691	0.6585	0.6637
Run_7	0.7192	0.6604	0.5485	0.5953	0.6604	0.6261
Run_8	0.7258	0.6557	0.7675	0.6261	0.6557	0.6405
Run_9	0.6425	0.6331	0.6885	0.6770	0.6331	0.6543
Run_10	0.6011	0.6434	0.7384	0.6414	0.6434	0.6423
Average	0.67751	0.67194	0.66619	0.63234	0.67194	0.65058

Table 6.4 depicts that the results obtained from WCGA based on parameters such as accuracy, sensitivity, specificity, precision, recall and f-measures. The values obtained accuracy-0.67751, sensitivity- 0.67194, specificity-0.66619, precision-0.63234, recall-0.67194, f-measure-0.65064.

Table 6.5: The results of WCPSO cluster validity (in terms of classification parameters)

WCPSO						
No. of Runs	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**F-measure**
Run_1	0.6763	0.6717	0.6947	0.6742	0.6717	0.6729
Run_2	0.6907	0.7215	0.6514	0.7526	0.7215	0.7367
Run_3	0.6116	0.7356	0.6210	0.7896	0.7356	0.6146
Run_4	0.7663	0.6464	0.6415	0.6179	0.6464	0.6318
Run_5	0.8067	0.6323	0.6291	0.7277	0.6323	0.6766
Run_6	0.7216	0.7104	0.7322	0.7257	0.7104	0.7179
Run_7	0.6149	0.6534	0.7007	0.6340	0.6534	0.6435
Run_8	0.7434	0.7402	0.6695	0.6093	0.7402	0.6684
Run_9	0.6935	0.7151	0.7194	0.6153	0.7151	0.6527
Run_10	0.7601	0.6509	0.6950	0.7314	0.6509	0.6888
Average	0.70851	0.68775	0.67545	0.68777	0.68775	0.67039

Table 6.5 depicts that the results obtained from WCPSO based on parameters such as accuracy, sensitivity, specificity, precision, recall and f-measures. The values obtained accuracy-0.70851, sensitivity- 0.68775, specificity-0.67545, precision-0.68777, recall-0.68775, f-measure-0.67039.

Table 6.6: The results of WCDE cluster validity (in terms of classification parameters)

WCDE						
No. of Runs	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**F-measure**
Run_1	0.7608	0.8002	0.7031	0.7248	0.8002	0.7606
Run_2	0.8917	0.6670	0.5322	0.7174	0.6670	0.6912
Run_3	0.7661	0.6858	0.6585	0.6957	0.6858	0.6907
Run_4	0.8906	0.7610	0.6782	0.6503	0.7610	0.7013
Run_5	0.7509	0.8047	0.6331	0.7647	0.8047	0.7841
Run_6	0.8674	0.7585	0.7199	0.7241	0.7585	0.7409
Run_7	0.7889	0.6604	0.7686	0.6510	0.6604	0.6556
Run_8	0.7957	0.6924	0.8459	0.6259	0.6924	0.6574
Run_9	0.8478	0.7830	0.7899	0.6665	0.7830	0.7200
Run_10	0.7785	0.7430	0.7062	0.6231	0.7430	0.6777
Average	**0.81384**	0.73560	0.70356	0.68435	0.73560	0.70800

Table 6.6 depicts that the results obtained from WCDE based on parameters such as accuracy, sensitivity, specificity, precision, recall and f-measures. The values obtained accuracy-0.81384, sensitivity- 0.70356, specificity-0.70356, precision-0.68435, recall-0.70356, f-measure-0.70800.

It is clear that the results obtained from WCGA, WCPSO and WCDE cluster validity in terms of classification parameters WCDE performs better as compared to the WCGA and WCPSO in terms of accuracy.

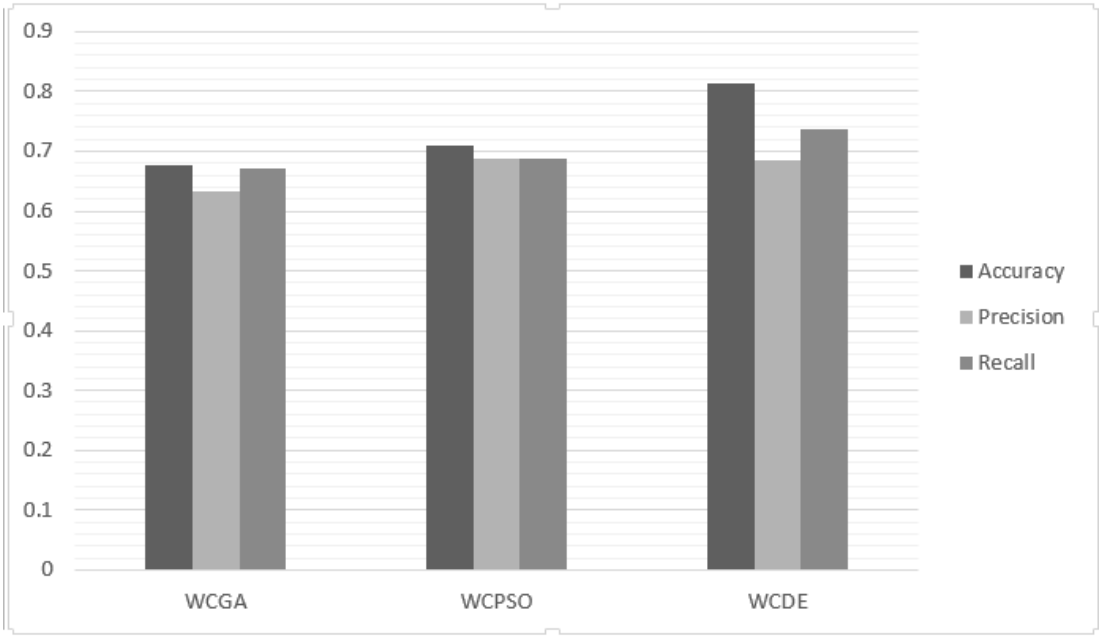

Figure 6.2: Comparative analysis of WCGA, WCPSO and WCDE based on the accuracy, precision and recall

Figure 6.2 shows the comparative analysis of WCGA, WCPSO and WCDE on the basis of accuracy, precision and recall. From this figure it is concluded that WCDE outperforms in comparison with WCGA and WCPSO.

6.2.2. Results of Epileptic Seizure Recognition Dataset

Table 6.7: The results of DEC and PCDE evaluated on the basis of DBI, Dunn and Computation time.

	DEC			PCDE		
No. of Runs	**DB Index**	**Dunn**	**CT**	**DB Index**	**Dunn**	**CT**
Run 1	0.8987	0.8670	575.45	0.8674	0.9441	452.37
Run 2	0.8796	0.8756	468.22	0.8770	0.9234	538.96
Run 3	0.9952	0.9342	585.27	0.8515	0.9548	485.32
Run 4	0.8584	0.7944	483.54	0.8242	0.9704	441.21
Run 5	0.8417	0.9575	562.85	0.8686	0.8125	389.59
Run 6	0.9266	0.8935	489.74	0.8516	0.8737	552.37
Run 7	0.8802	0.8541	583.67	0.8905	0.9537	396.67
Run 8	0.86 17	0.8834	467.62	0.8472	0.9615	497.50
Run 9	0.9584	0.9032	598.68	0.9103	0.8737	573.63
Run 10	0.8643	0.8467	576.84	0.8381	0.9704	573.16
Average	0.9003	0.8809	539.18	**0.8626**	0.9238	490.07

From the above tabulated table 6.7 it is concluded that the proposed algorithm, PCDE is perform better as compared to DEC. The values obtained are DB Index- 0.8626 Dunn-0.9238 computational times- 490.07

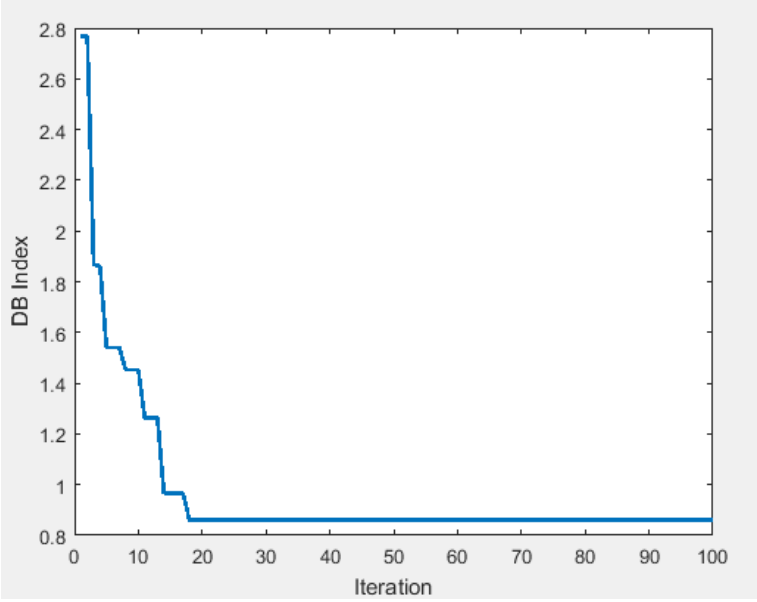

Figure 6.3 (a): depicts the db-index vs number of iterations for PCDE approach

Figure 6.3 (a) (PCDE) shows that graph plotted between number of iteration and DB Index values. In this graph at 17 iterations the results of PCDE goes stagnant and the DB index value is near about 0.8.

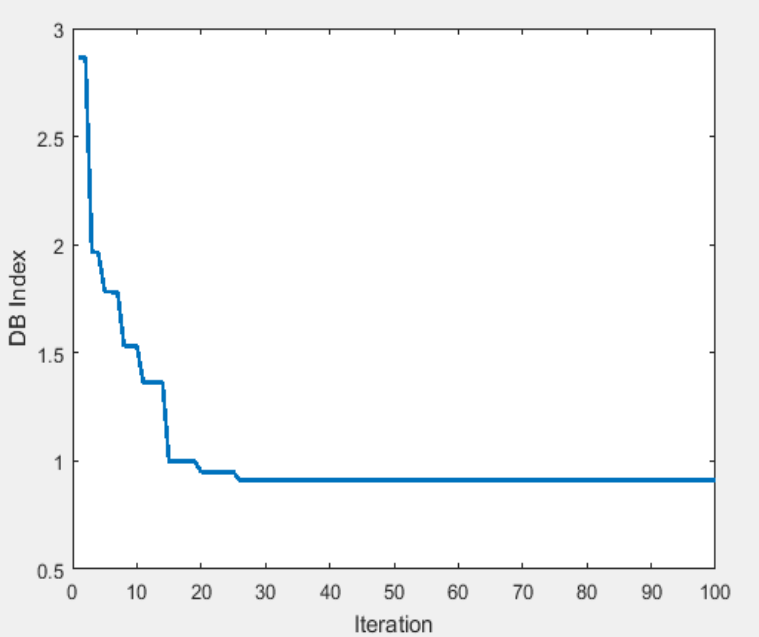

Figure 6.3 (b): depicts the db-index vs number of iterations for DEC approach

Figure 6.3 (b) (DEC) shows that graph plotted between number of iteration and DB Index values. In this graph at 25 iterations the results of DEC goes stagnant and the DB index value is near about 0.9.

From above two figures 6.3(a) (b) it is concluded that PCDE performs better as compared to DEC.

Chapter 7

Conclusion and Future scope

7.1. Conclusion

In current era, medical databases are so huge and complicated, clustering algorithms are of vital importance that it organizes the data, thereby generating patterns that can be further utilized for better analysis of diseases. In this work, various clustering techniques are applied based on evolutionary approaches to find the optimal results. The performance is examined on the basis of clustering metric namely DBI and DI. The experimental results show that on breast cancer wisconsin dataset the Wisconsin Clustering Differential Evolution performs better than other two techniques. The novel proposed technique, partitioning based clustering using DE is applied on epileptic seizure dataset and the results are compared with the Differential Evolution Clustering. The results conclude that Partitioning based Clustering using Differential Evolution gives better results as compared to Differential Evolution Clustering.

7.2. Future scope

- The work can be tried by considering other evolutionary techniques.
- The proposed Partitioning Clustering based on Differential Evolution can be applied on large scale datasets using Hadoop Framework.

References

[1] Williams, Kehinde, Peter Adebayo Idowu, Jeremiah AdemolaBalogun, and Adeniran Ishola Oluwaranti, "Breast cancer risk prediction using data mining classification techniques," Transactions on Networks and Communications 3(2), 01-11, 2015.

[2] Jaimini Majali, Rishikesh Niranjan, Vinamra Phatak, OmkarTadakhe, "Data Mining Techniques For Diagnosis And Prognosis Of Cancer," International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 3, March 2015, pp.613-616.

[3] Vikas Chaurasia, Saurabh Pal, "A Novel Approach for Breast Cancer Detection using Data Mining Techniques," International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), Vol. 2, Issue 1, January 2014, 2456-2465.

[4] U. R. Acharya, S. V. Sree, G. Swapna, R. J. Martis, and J. S. Suri, "Automated EEG analysis of epilepsy: a review," Knowledge-Based Systems, vol. 45, pp. 147–165, 2013.

[5] R. S. Fisher, W. Van Emde Boas, W. Blume et al., "Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE)," Epilepsia, vol. 46, no. 4, pp. 470–472, 2005.

[6] M. Guenot, "Surgical treatment of epilepsy: Outcome of various surgical procedures in adults and children," Revue Neurologique, vol. 160, no. 5, pp. S241–S250, 2004.

[7] Majali, Jaimini, et al. "Data Mining Techniques for Diagnosis and Prognosis of Cancer." International Journal of Advanced Research in Computer and Communication Engineering 4(3), 613-616, 2015.

[8] Senturk, ZehraKarapinar, and Resul Kara, "Breast Cancer Diagnosis via Data Mining: Performance Analysis of Seven different algorithms," Computer Science & Engineering 4(1), 35-46, 2014.

[9] Sivakami, K, "Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model," International Journal of Scientific Engineering and Applied Science (IJSEAS)–1(5), 418-429, 2015.

[10] Venkatesan, E., and T. Velmurugan, “Performance analysis of decision tree algorithms for breast cancer classification,” Indian Journal of Science and Technology 8(29), 1-8, 2015.

[11] Shah Chintan, 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.

[12] Ghosh Soumadip, 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.

[13] Elouedi Hind, et al., "A hybrid approach based on decision trees and clustering for breast cancer classification," Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of. IEEE, 2014.

[14] Kathija, Shajun Nisha., ”Breast Cancer Data Classification Using SVM and Naïve Bayes Techniques,” International Journal of Innovative Research in Computer and Communication Engineering- Vol. 4, Issue 12, December 2016.

[15] 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.

[16] Nauck, D and Kruse, R. (1999), “Obtaining interpretable fuzzy classification rules from medical data.” Artificial intelligence in medicine, 16(2), 149-169.

[17] Chang PW, Liou MD, editors, “Comparison of three Data Mining techniques with Genetic Algorithm in analysis of Breast Cancer data,” Asian Journal of Computer Science and Information Technology 3:5 (2013) 81 - 87.

[18] Alshammari , Sultanah M, Shah , Tawfiq M, Huang , Yan , editors, “Data Mining Techniques for Predicting Breast Cancer Survivability Among Women in the United States,” UNT Digital Library. 2015.

[19] Ghassem Pour S, Mc Leod P, Verma B, Maeder A, editors, “Comparing Data Mining with Ensemble Classification of Breast Cancer,” Masses in Digital Mammograms. 2012.

[20] Hota HS, “Diagnosis of Breast Cancer Using Intelligent Techniques,” International Journal of Emerging Science and Engineering (IJESE) 2013; 1:45–53.

[21] Rajesh K, Anand S, “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.

[22] A. I. Pritom, “Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique,” pp. 310–314, 2016.

[23] Jimin Guo Benjamin 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.

[24] Joshi J, Doshi R, Patel J, “Diagnosis and Prognosis Breast Cancer Using Classification Rules,” International Journal of Engineering Research and General Science. 2014; 2:315–323.

[25] Padmavati J, “A Comparative study on Breast Cancer Prediction Using RBF and MLP,” International Journal of Scientific & Engineering Research. 2011; 2:1–5.

[26] Salama GI, Abdelhalim MB, Zeid MA, “Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers,” International Journal of Computer and Information Technology. 2012; 1:36–43.

[27] Saleema JS, Deepa Shenoy P, Venugopal KR, Patnaik LM, “Cancer Prognosis Prediction Model using Data Mining Techniques,” Data Mining and Knowledge Engineering. 2014; 6:25–47.

[28] Mittal Dishant, Dev Gaurav, and Sanjiban Sekhar Roy, "An effective hybridized classifier for breast cancer diagnosis," 2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM). IEEE, 2015.

[29] 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.

[30] E. Aličković and A. Subasi, “Breast cancer diagnosis using GA feature selection and Rotation Forest,” Neural Comput. Appl., vol.28, no. 4, pp. 753–763, 2017.

[31] J. Guo et al., “Revealing determinant factors for early breast cancer recurrence by decision tree,” Inf. Syst. Front., 2017.

[32] R. J. Kate and R. Nadig, “Stage-specific predictive models for breast cancer survivability,” Int. J. Med. Inform., vol. 97, pp. 304–311, 2017.

[33] M. R. Mohebian, H. R. Marateb, M. Mansourian, M. A. Mañanas, and F. Mokarian, “A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer

Recurrence (HPBCR) Using Optimized Ensemble Learning," Comput. Struct. Biotechnol. J., vol. 15, pp. 75–85, 2017.

[34] A. I. Pritom, "Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique," pp. 310–314, 2016.

[35] Walaa Gad , "SVM-Kmeans: Support Vector Machine based on Kmeans Clustering for Breast Cancer Diagnosis," International Journal of Computer and Information Technology (ISSN: 2279 – 0764) Volume 05 – Issue 02, March 2016,252-257.

[36] G. D. Rashmi, A. Lekha, and N. Bawane, "Analysis of efficiency of classification and prediction algorithms (Naïve Bayes) for Breast Cancer dataset," 2015 Int. Conf. Emerg. Res. Electron. Comput. Sci.Technol., pp. 108–113, 2015.

[37] 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.

[38] B. R. A. Cirkovic, A. M. Cvetkovic, S. M. Ninkovic, and D. Nenad, "Prediction Models for Estimation of Survival Rate and Relapse for Breast Cancer Patients," IEEE International Conference on Bioinformatics and Bioengineering, 2015.

[39] A. LG and E. AT, "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence," J. Heal. Med. Informatics,vol. 4, no. 2, pp. 2–4, 2013.

[40] C. Shah and A. G. Jivani, "Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction," 2013 Fourth Int. Conf. Comput. Commun. Netw. Technol., pp. 1–4, 2013.

[41] Q. Fan, C. Zhu, and L. Yin, "Predicting Breast Cancer Recurrence Using Data Mining Techniques," pp. 310–311, 2010.

[42] J. Rasekhi, M. R. K. Mollaei, M. Bandarabadi, C. A. Teixeira, and A. Dourado, "Preprocessing effects of 22 linear univariate features on the performance of seizure prediction methods," Journal of Neuroscience Methods, vol. 217, no. 1-2, pp. 9–16, 2013.

[43] C. A. Teixeira, B. Direito, M. Bandarabadi et al., "Epileptic seizure predictors based on computational intelligence techniques: A comparative study with 278 patients," Computer Methods and Programs in Biomedicine, vol. 114, no. 3, pp. 324–336, 2014.

[44] M. Bandarabadi, C. A. Teixeira, J. Rasekhi, and A. Dourado, “Epileptic seizure prediction using relative spectral power features,” Clinical Neurophysiology, vol. 126, no. 2, pp. 237–248, 2015.

[45] A. S. Zandi, R. Tafreshi, M. Javidan, and G. A. Dumont, “Predicting epileptic seizures in scalp EEG based on a variational bayesian gaussian mixture model of zero-crossing intervals,” IEEE Transactions on Biomedical Engineering, vol. 60, no. 5, pp. 1401–1413, 2013.Yusof, T. Kreuz, C. Rieke et al., “On the predictability of epileptic seizures,” Clinical Neurophysiology, vol. 116, no. 3, pp. 569–587, 2012.

[46] Yusof, T. Kreuz, C. Rieke et al., “On the predictability of epileptic seizures,” Clinical Neurophysiology, vol. 116, no. 3, pp. 569–587, 2012.

[47] L. Guo, D. Rivero, J. Dorado, J.R. Rabunal., and A. Pazos, “Automatic epileptic seizure detection in EEGs based on line length feature and artificial neural networks,” J Neurosci Methods, Vol 191, No.1, 2010b, pp. 101-109.

[48] Wang, J. Lina, and J. Gotman, “Discriminating preictal and interictal states in patients with temporal lobe epilepsy using wavelet analysis of intracerebral EEG,” Clinical Neurophysiology, vol. 123, no. 10, pp. 1906–1916, 2012.

[49] Janjarasjitt, M. Hekim, and M. Ozer, “EEG signals classification using the K-means clustering and a multilayer perceptron neural network model,” Expert Systems with Applications, Vol. 38, 2011, pp. 13475– 13481.

[50] Acharya, J. Lina, and J. Gotman, “Comparison Machine Learning Algorithms for Recognition of Epileptic Seizures in EEG,” Computational and Mathematical Methods in Medicine, 2012.

[51] Davies, David L., and Donald W. Bouldin. "A cluster separation measure."Pattern Analysis and Machine Intelligence, IEEE Transactions on 2 (1979): 224-227.

[52] Dunn, J. C. (1973). "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". Journal of Cybernetics 3 (3): 32–57.

[53] Michalski,R.S. Learning.“UCI repository of machine learning databases,” 1987.[Online].Available:https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+Diagnostic).

[54] Michalski,R.S. Learning.“UCI repository of machine learning databases,” 1987.[Online].Available:https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition.

List of Publications

[1] Meghna Dhalaria, Maninder Kaur. “Prediction of Epileptic Seizures using Machine Learning Approaches: A Survey” (Submitted).

[2] Meghna Dhalaria, Maninder Kaur. “Analysis of Evolutionary Approaches on Wisconsin Breast cancer” (Submitted).

[3] Meghna Dhalaria, Maninder Kaur. “Novel Partitioning based Clustering using DE Approach on Epileptic Seizures” (Submitted).

Plagiarism Report

8%	5%	6%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	research.ijcaonline.org Internet Source	1%
2	Samuel Giftson Durai, S. Hari Ganesh, A. Joy Christy. "Novel Linear Regressive Classifier for the Diagnosis of Breast Cancer", 2017 World Congress on Computing and Communication Technologies (WCCCT), 2017 Publication	<1%
3	Paterlini, S.. "Differential evolution and particle swarm optimisation in partitional clustering", Computational Statistics and Data Analysis, 20060301 Publication	<1%
4	Ching-Yi Chen, Fun Ye. "Particle swarm optimization algorithm and its application to clustering analysis", IEEE International	<1%