**I J C T A**, 9(3), 2016, pp. 1733-1744


# An Analysis of Particle Swarm Optimization Technique for Breast Cancer Dataset

**Vijaylakshmi S.* and Priyadarshini J.****

***Abstract:*** This paper gives the current overview of use of Particle Swarm Optimization techniques on breast cancer data. We analyze the breast Cancer data available from the WBC, WDBC from UCI machine learning with the aim of developing accurate prediction models for breast cancer using Particle Swarm Optimization (PSO) Technique. PSO is a population-based stochastic search algorithm that mimics the capability of swarm (cognitive and social behavior). Breast Cancer Diagnosis and Prognosis are two medical applications pose a great challenge to the researchers. The use of machine learning and data mining techniques has revolutionized the whole process of breast cancer Diagnosis and Prognosis. Breast Cancer Diagnosis distinguishes benign from malignant breast lumps and Breast Cancer Prognosis predicts when Breast Cancer is likely to recur in patients that have had their cancers excised. Thus, these two problems are mainly in the scope of the classification problems. This study paper summarizes various review and technical articles on breast cancer diagnosis. In this paper we present an overview of the current research being carried out using the data mining techniques to enhance the breast cancer diagnosis and prognosis.

***Keywords:*** PSO, GA, optimization, Breast Cancer, Diagnosis, Prognosis

## 1. INTRODUCTION

As breast cancer recurrence is high, good diagnosis is important. Many studies have been conducted to analyze Breast Cancer Data. Breast Cancer Diagnosis is distinguishing of benign from malignant breast lumps and Breast Cancer Prognosis predicts when Breast Cancer is to recur in patients that have had their cancers excised. Breast cancer poses a serious threat and is the second leading cause of death in women today and most common cancer in developed countries. The lack of awareness initiatives, structured viewing, and affordable treatment facilities continue to result in poor survival. Cancer of the breast is the second most common human neoplasm, accounting for around one quarter of all cancers in females after cervical carcinoma. We present the overview of research in use of particle swarm optimization techniques in breast cancer. Particle Swarm Optimization has become a popular technology in current research and for medical domain applications. Breast cancer has become the leading cause of death in women in developed countries.

Swarm intelligence is based on a population of individuals [1]. In swarm intelligence, an algorithm maintains and successively improves a collection of potential solutions until some stopping condition is met. The solutions are initialized randomly in the search space. The search information is propagated through the interaction among solutions. Based on the solutions convergence and divergence, solutions are guided toward the better and better areas. In swarm intelligence algorithms, there are several solutions which exist at the same time. The premature convergence may happen due to the solution getting clustered together too fast. The population diversity is a measure of exploration and exploitation. Based on the population diversity changing measurement, the state of exploration and exploitation can be obtained. The population diversity definition is the first step to give an accurate observation of the search state. Many studies of population diversity in evolutionary computation algorithms and swarm intelligence have been proposed in [2]–[8].

* Research Scholar, SCSE, Vellore Institute of Technology Chennai Campus, Chennai, India, *Email: vijayalakshhmi.s2014@vit.ac.in*

** Associate Professor, SCSE Vellore Institute of Technology Chennai Campus, Chennai, India, *Email: priyadarshini.j@vit.ac.in*

Particle swarm optimization (PSO) is a population-based stochastic algorithm modeled on social behaviors observed in flocking birds [9], [10]. A particle flies through the search space with a velocity that is dynamically adjusted according to its own and its companion's historical behaviors. Each particle's position represents a solution to the problem. Particles tend to fly toward better and better search areas over the course of the search process [11], [12]. Different topology structure can be utilized in PSO, which will have different strategy to share search information for every particle. PSO algorithms have recently been shown to produce good results in a wide variety of real-world data.

The rest of the paper is organized in two parts. First part concern with review of research in the application of PSO techniques on breast cancer and second part consist experimental work. First part is organized as follows: The next section discusses About Knowledge discovery and database is subsequently discussed and its tasks that are related to PSO techniques. Next section concerns with previous work of research review of application of data mining in breast cancer data. Second part consist dataset description and experiment and discusses the results and future work.

## 2. RELATED WORK

In the literature, there are many studies done on cancer detection and/or data mining. [13] useddata mining for the diagnosis of ovarian cancer. For the analysis, serum proteomics that distinguish the serum ovarian cancer cases from non-cancer ones are used. An SVM (Support Vector Machine) based method is applied and statistical testing and GA (Genetic Algorithms) based methods are used for feature selection. [14] aimed to propose a new 3-D microwave approach based on SVM classifier whose output is transformed to a posteriori probability of tumor presence. Gene expression data sets for ovarian, prostate and lung cancers are analyzed in another paper[15]. An integrated gene search algorithm (preprocessing: GA and correlation based heuristics, making predictions/ data mining: decision tree and SVM algorithms) for genetic expression data analysis is proposed. In [16] the clinical and imaging diagnostic rules of peripheral lung cancer by data mining techniques that are Association Rules (AR) of knowledge discovery process and Rough Set (RS) reduction algorithm and Genetic Algorithm (GA) of generic data analysis tool (ROSETTA) are extracted. [17] deals with complementary learning fuzzy neural network (CLFNN) for the diagnosis of ovarian cancer. CLFNN-micro-array, CLFNN-blood test, CLFNN-proteomics demonstrates good sensitivity and specificity. So, it is shown that CLFNN outperforms most of the conventional methods in ovarian cancer diagnosis. [18] applies the classification technology to construct an optimum cerebrevascular disease predictive model. Classification algorithms used are decision tree, Bayesian classifier, and back propagation neural network.

The objective of [19] is to develop an original method to extract sets of relevant molecular biomarkers (gene sequences) that can be used for class prediction and as a prognostic and predictive tool. With the help of the analysis of DNA microarrays, molecular biomarkers are generated and this analysis is based on a specific data mining technique: Sequential Pattern Discovery. The performance of data classification by integrating artificial neural networks with multivariate adaptive regression splines (MARS) approach is explored for mining breast cancer pattern[20].

This approach is based on firstly to use MARS in modeling the classification problem, then obtained significant variables are used as input variables of designed neural networks model. A comparison of three data mining techniques artificial neural networks, decision trees, and logistic regression is realized in a study to predict the survivability of breast cancer [21]. Accuracy rates are found as 93.6%, 91.2%, and 89.2% respectively. Many aspects of possible relationships among DNA viruses and breast tumors are considered [22]. Feasible clusters in DNA virus combinations that depend on the observed probability of breast cancer, fibro adenoma and normal mammary tissue are created in this study and viral prerequisites for breast carcinogenesis and the protective are determined. Obtaining bioinformatics about breast tumor and DNA viruses, and building an accurate diagnosis model for breast cancer and fibro adenoma are aimed [23].

A hybrid SVM-based strategy with feature selection to render a diagnosis between the breast cancer and fibro adenoma and to find important risk factor for breast cancer is constructed. DNA viruses, HSV-1, EBV, CMV, HPV and HHV-8 are evaluated. There is also another study related to breast cancer. Bre[as]t cancer pattern is mined using discrete particle swarm optimization and statistical method [24]. Besides, to detect breast cancer, association rules (AR) and neural network (NN) are used this time [25]. AR is used to reduce the dimension of the database and NN is used for intelligent classification. In Menendeza et al. (2010), a Self-Organizing Map (SOM) based clustering algorithm for preprocessing of samples from a breast cancer screening program is introduced. Prediction of the recurrence of breast cancer is investigated [13]. The accuracy of Cox Regression and SVM algorithms are compared and it is shown that a parallelism of adequate treatment and follow-up by recurrence prediction prevent the recurrence of breast cancer. In this study, different from the studies stated above, breast cancer is tried to be predicted whether as a benign or malignant case through seven different algorithms which have not been tried for breast cancer yet in the literature and a performance analysis is aimed to be performed.

In this section some of the related prior work on data mining methods for breast cancer diagnosis is discussed. Song et al. [26] presented a new approach for automatic breast cancer diagnosis based on artificial intelligence technology. They focused to obtain a hybrid system for diagnosing new breast cancer cases in collaboration between Genetic Algorithm (GA) and Fuzzy Neural Network. They also showed that inputs reduction (features selections) can be used for many other problems which have high complexity and strong non-linearity with huge data to be analysed.

Arulampalam and Bouzerdoum [38] proposed a method for diagnosing breast cancer and called Shunting Inhibitory Artificial Neural Networks (SIANNs). SIANN is a neural network stimulated by human biological networks in which the neurons interact among each other's via a nonlinear mechanism called shunting inhibition. The feed forward SIANNs have been applied to several medical diagnosis problems and the results were more favourable than those obtained using Multilayer Perceptions (MLPs). In addition, a reduction in the number of inputs was investigated.

Setiono [27]proposed a method to extract classification rules from trained neural networks and discussed its application to breast cancer diagnosis. He also explained how the pre-processing of data set can improve the accuracy of the neural network and the accuracy of the rules because some rules may be extracted from human experience, and may be erroneous. The data pre-processing involves the selection of significant attributes and the elimination of records with missed attribute values from Wisconsin Breast Cancer Diagnosis dataset. The rules generated by Setiono's method were more brief and accurate than those generated by other methods mentioned in the literature.

Meesad and Yen [28] proposed a hybrid Intelligent System (HIS) which integrates the Incremental Learning Fuzzy Network (ILFN) with the linguistic knowledge representations. The linguistic rules have been determined based on knowledge embedded in the trained ILFN or been extracted from real experts. In addition, the method also utilized Genetic Algorithm (GA) to reduce the number of the linguistic rules that sustain high accuracy and consistency. After the system being completely constructed, it can incrementally learn new information in both numerical and linguistic forms. The proposed method has been evaluated using Wisconsin Breast Cancer Dataset (WBC) data set. The results have shown that the proposed HIS perform better than some well-known methods.

## 3. EXISTING SYSTEM BASED ON CLASSIFICATION

The data mining consists of various methods. Different methods serve different purposes, each method offering its own advantages and disadvantages. However, most data mining methods commonly used for this review are of classification category as the applied prediction techniques assign patients to either a "*benign*" group that is non-cancerous or a "*malignant*" group that is cancerous and generate rules for the

same. Hence, the breast cancer diagnostic problems are basically in the scope of the widely discussed classification problems. In data mining, classification is one of the most important task. It maps the data in to predefined targets. It is a supervised learning as targets are predefined. The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. The following methods are the commonly used methods for data mining classification.

### 3.1. K-Nearest Neighborhood (KNN)

The k-Nearest Neighbor algorithm is based on learning by analogy, that is, by comparing a given test example with training examples that are similar to it. The training examples are described by n attributes. Each example represents a point in an n-dimensional space. In this way, all of the training examples are stored in an n-dimensional pattern space. When given an unknown example, a k-nearest neighbor algorithm searches the pattern space for the k training examples that are closest to the unknown example. These k training examples are the *k* "nearest neighbors" of the unknown example. "Closeness" is defined in terms of a distance metric, such as the Euclidean distance.

The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an example is classified by a majority vote of its neighbors, with the example being assigned to the class most common amongst its k nearest neighbors (*k* is a positive integer, typically small). If $k = 1$, then the example is simply assigned to the class of its nearest neighbor. The same method can be used for regression, by simply assigning the label value for the example to be the average of the values of its k nearest neighbors. It can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones.

The neighbors are taken from a set of examples for which the correct classification (or, in the case of regression, the value of the label) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. The basic k-Nearest Neighbor algorithm is composed of two steps: Find the k training examples that are closest to the unseen example. Take the most commonly occurring classification for these k examples (or, in the case of regression, take the average of these k label values). [42]

### 3.2. Support Vector Machine (SVM)

Support vector machine (SVM) is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. SVMs are well suited to dealing with interactions among features and redundant features. Support Vector Machine a new approach to supervised pattern classification which has been successfully applied to a wide range of pattern recognition problems. Support Vector Machine is a training algorithm for learning classification and regression rules from data. SVM is most suitable for working accurately and efficiently with high dimensionality feature spaces. SVM is based on strong mathematical foundations and results in simple yet very powerful algorithms. The standard SVM algorithm builds a binary classifier. A simple way to build a binary classifier is to construct a hyperplane separating class members from non-members in the input space. SVM also finds a nonlinear decision function in the input space by mapping the data into a higher dimensional feature space and separating it there by means of a maximum margin hyperplane. The system automatically identifies a subset of informative points called support vectors and uses them to represent the separating hyperplane which is sparsely a linear combination of these points. Finally SVM solves a simple convex optimization problem. The machine is presented with a set of training examples, $(x_i, y_i)$ where the xi are the real world data instances and the *yi* are the labels indicating which class the instance belongs to. For the two class pattern recognition problem, $y_i = +1$ or $y_i = -1$. A training example $(x_i, y_i)$ is called positive if $y_i = +1$ and negative otherwise. SVM

construct a hyperplane that separates two classes and tries to achieve maximum separation between the classes. Separating the classes with a large margin minimizes a bound on the expected generalization error.

### 3.3. Decision Trees (DT's)

A decision tree is a tree where each non-terminal node represents a test or decision on the considered data item. Choice of a certain branch depends upon the outcome of the test. To classify a particular data item, we start at the root node and follow the assertions down until we reach a terminal node (or leaf). A decision is made when a terminal node is approached. Decision trees can also be interpreted as a special form of a rule set, characterized by their hierarchical organization of rules. Decision Trees are generated by recursive partitioning. Recursive partitioning means repeatedly splitting on the values of attributes. In every recursion the algorithm follows the following steps:

- An attribute A is selected to split on. Making a good choice of attributes to split on each stage is crucial to generation of a useful tree. The attribute is selected depending upon a selection criterion which can be selected by the criterion parameter.
- Examples in the Example Set are sorted into subsets, one for each value of the attribute A in case of a nominal attribute. In case of numerical attributes, subsets are formed for disjoint ranges of attribute values.
- A tree is returned with one edge or branch for each subset. Each branch has a descendant subtree or a label value produced by applying the same algorithm recursively.

In general, the recursion stops when all the examples or instances have the same label value, i.e. the subset is pure. Or recursion may stop if most of the examples are of the same label value. This is a generalization of the first approach; with some error threshold. However there are other halting conditions such as:

- There are less than a certain number of instances or examples in the current subtree. This can be adjusted by using the minimal size for split parameter.
- No attribute reaches a certain threshold. This can be adjusted by using the minimum gain parameter.
- The maximal depth is reached. This can be adjusted by using the maximal depth parameter.

## 4. PROPOSED SYSTEM

The GA and PSO is widely used Evolutionary algorithm because of its simple process with optimized solution. In this paper, the hybrid PGSO procedure is followed by combining the PSO and GA for optimized feature selection. The GA is embedded within the PSO to improve the PSO by serving as a local optimizer at each iteration as shown in figure 1.

### 4.1. Particle Swarm Optimization (PSO)

In PSO, each particle represent a candidate solutions of a population, simultaneously coexist and evolve based on knowledge sharing with neighbouring particles. Each particle flies on the problem search space and based on the directed velocity vector, it will generate a solution. Each particle changes its velocity to determine the better position by using its own flying knowledge for the best position memory found in the earlier flights and experience of neighbouring particles as the best determined solution of the population. The best position determined by the particles is represented as Pbest and the tendency to move forwards the previous best position of the neighbourhood'sgbest**.** The velocity of the particle is updated using the following equation

$$v^t_{\max count} = w_t v^t_t + c_1 r_1 (p^i_t - x^t_t) + c_2 r_2 (p^2_t - x^t_t)$$

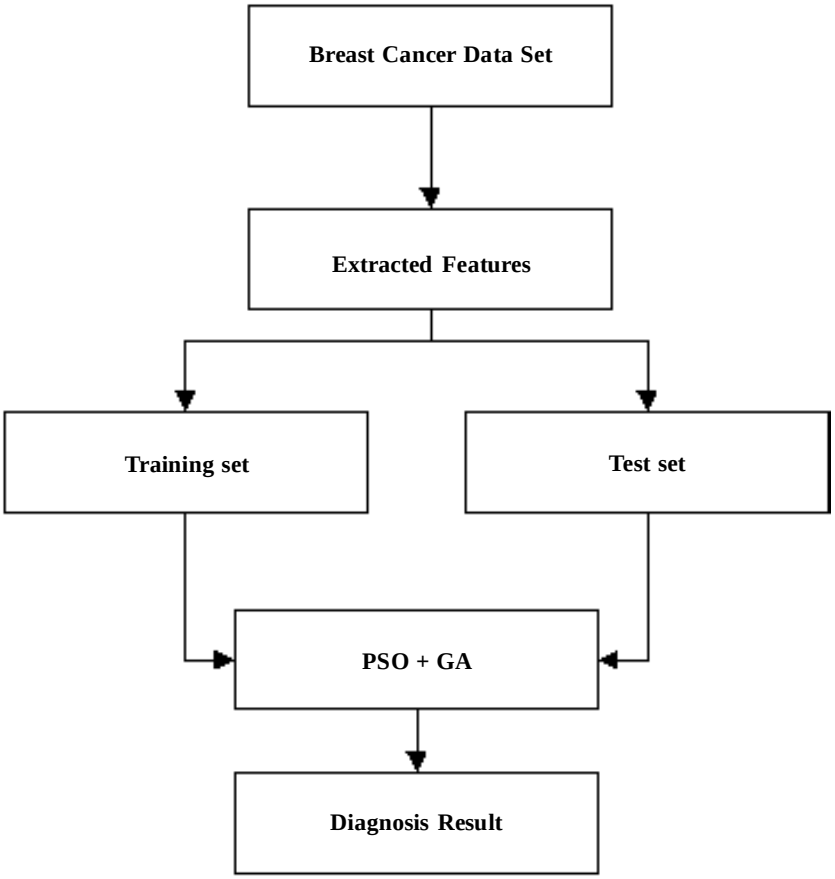


**Figure 1: Block Diagram of the Proposed Method**

Where $x_i^i$ represents the current position of particle $i$, $p_i^i$ is the current best position determined by particle $i$, $p_g^t$ is the global best position determined among all particles in the problem space up to iteration count $i$, $c_1$ and $c_2$ represents the cognitive and social scaling parameters, $r_1$ and are random numbers distributed uniformly in the interval (0,1). $w_t$ is the particle inertia, which minimize the search area dynamically,

$$w_t = (w_{\max} - w_{\min}) \times \frac{\max count - t}{\max count} + w_{\min}$$

Where $w_{\max}$ and $w_{\min}$ indicates the maximum and minimum of *wt* respectively max count represents the maximum iteration and *t* represents the current iteration number. The particle position updated according to the following equation,

$$X^{i}_{\max count} = \begin{cases} 1, rand() < sig\left\{v_i^i + 1\right\} \\ 0, rand() \geq sig\left\{v_i^i + 1\right\} \end{cases}$$

$$sig(x) = \frac{1}{1 + e^{-x}}$$

In this way all particles determines the new positions and update their individual best position $p_i^i$ and global best position $p_g^t$ of the swarm. This process will be continued until maximum iteration reached.

Consider an Information System ($IS$) = ($U$, $A$) and $A = (X \cup Y)$ where $X$ is a non-empty finite set of condition attributes and $Y$ is a non-empty finite set of decision attributes, such that $RED$ ($IS$) $\subseteq X$. The objective function of particle i at position $x$ is determined by the following equation,

$$f\left(x_t^i\right) = a \times y_{x_t^i}(Y) + \beta \times \frac{|x| - |x_t^i|}{|x|}$$

Where $y_{xi}$ ($Y$) is the classification quality of particle condition attribute set $x_t^i$, which contains the RED, and relative to decision table $Y$, and it is shown in the following equation,

$$y_{x_t^i}(Y) = \frac{d_{RED}}{d_i}$$

Where $d_{\mathrm{RED}}$represents a dependency degree of RED on Y and dc represents the dependency degree of $X$ on $Y$. $x_t^i$ is the '1' number of length of selected feature subset for particle $x_t^i$, while population of solutions $P$ (number of particles in the population) is at iteration count $t$. $|X|$ is the total number of condition attributes. The parameter $\alpha$ = [0, 1] and $\beta$ = 1-$\alpha$ represents importance of classification quality and subset length.

### 4.2. Genetic Algorithm

Genetic Algorithm (GA) is one of the computational models used widely because of its evolution. GA optimization technique contains selection, crossover and mutation operations to a population of completing problem solutions. After the three operations are applied a new generation of the populations will be generated and at the same time the GA will generate a set of chromosome randomly at the space. The fitness value will be calculated for the chromosomes and the chromosome with a higher fitness value will be kept and the same operation will be performed until a fixed number of iterations are reached.

To improve the PSO performance, the GA is used as a local optimizer at each iteration. After the initial populations are created, the operation such as selection, crossover and mutation will be applied to the initially created particles. Choose two particles randomly and determine the relative differenced for those two particles by the following equation,

$$d = \frac{f_1 - f_2}{f_1 + f_2}$$

Where $f_1$and $f_2$are the fitness value of particle 1 and particle 2. According to the relative difference value the cross over operation is chosen and it is defined in the following equation,

$$\begin{cases} If\ f_1 < f_2\ and\ R < d, crossover\ operation\ is\ done\ on\ particle\ 2 \\ If\ f_1\ and\ R > d, normal\ crossover\ operation\ is\ choosen \\ If\ f_1 < f_2\ and\ R < d, crossover\ operation\ is\ done\ on\ particle\ 1 \\ If\ f_1 > f\ and\ R > r, normal\ crossover\ operation\ is\ choosen \end{cases}$$

Mutation operation is done on the two selected particles and it is set to probability of 1/$n$ ($n$ = number of particles). The process will be done until the maximum iteration is reached. Then the PSO based optimization is performed for the particles. The PGASO based rough set algorithm is given in the Figure 2.

Input: C, the set of all condition attributes, max count
Output: Reduct (RED)
Step 1: Initializing position, velocity, c1,c2.
Step 2: For i=1 to P
Obtaining the interia weight for particlei by using equation (3)
Calculate the fitness (objective) function for particlei by using the equation (6)
End For
Step 3: Intialize $p_i^t = p_i^{new}$ and Gbest = $f_{best}^t\left(p_t^{best}\right)$
Step 4: For i=1 to Maxcount
Select two random particle and calculate the relative difference using equation 8
Select the cross over operation according to the equation (9)
Mutuate (particlex), Mutuate (particley)
End For
Step 5: while (t< maxcount)
For i=1 to P
Compute the fitness function for particlei using eq (6)
If $pbest_i^t < f\left(x_i^t\right)$

$pbest_i^t < f\left(x_i^t\right)$

$p_i^t = x_i^t$

End if
Find the $f_{best}^t\left(p_i^{best}\right) = \max\left\{p_i^1, p_i^2, \ldots p_i^P\right\}$
$if\ pbest_i^t < f_i^{best}$
$p_i^g = p_i^{best}\ and\ Gbest_i = f_i^{best}\left(p_i^{best}\right)$
End if
Evaluate the velocity for particlei by equation (2)
Update particle position by using equation (4)
End For
End while
Step 6: Redact RED

**Figure 2: Algorithm for optimized feature set.**

## V. EXPERIMENTAL RESULTS AND DISCUSSIONS

Breast cancer is a malignant tumor that develops when cells in the breast tissue divide and grow without the normal controls on cell death and cell division. It is the most common cancer among women. Although scientists do not know the exact causes of most breast cancer, they do know some of the risk factors that increase the likelihood of a woman developing breast cancer. These factors include such attributes as age, genetic risk and family history. Treatments for breast cancer are separated into two main types, local and systematic. Surgery and radiation are examples of local treatments whereas chemotherapy and hormone therapy are examples of systematic therapies. Usually for the best results, the two types of treatment are used together. Although breast cancer is the second leading cause of cancer death in women, the survival rate is high. With early diagnosis, 97% of women survive for 5 years or more.

The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used [29], to distinguish malignant (cancerous) from benign (non-cancerous) samples. A brief description of these datasets is presented in table 1. Each dataset consists of some classification patterns or instances with a set of numerical features or attributes.

**Table 1**
**Description Of The Breast Cancer Datasets**

| *Dataset* | *No. of Attributes* | *No. of Instances* | *No. of Classes* | *Missing Value* |
|---|---|---|---|---|
| Wisconsin Breast Cancer (Original) | 11 | 699 | 2 | Yes |
| Wisconsin Diagnosis Breast Cancer (WDBC) | 32 | 569 | 2 | Yes |
| Wisconsin Prognosis Breast Cancer (WPBC) | 34 | 198 | 2 | Yes |

### 5.1. Wisconsin Breast Cancer Dataset [29]

The data used in this study are provided by the UC Irvine machine learning repository located in breast-cancer-Wisconsin sub-directory, filenames root: breast-cancer-Wisconsin having 699 instances, 2 classes (malignant and benign), and 9 integer-valued attributes. We removed the 16 instances with missing values from the dataset to construct a new dataset with 683 instances. Class distribution: Benign: 458 (65.5%) Malignant: 241 (34.5%). The details of the attributes found in this dataset listed in table 2.

**Table 2**
**Wisconsin Breast Cancer Dataset Attributes**

| | *Attribute* | *Domain* |
|---|---|---|
| 1 | Sample code number | id number |
| 2 | Clump Thickness | 1-10 |
| 3 | Uniformity of Cell Size | 1-10 |
| 4 | Uniformity of Cell Shape | 1-10 |
| 5 | Marginal Adhesion | 1-10 |
| 6 | Single Epithelial Cell Size | 1-10 |
| 7 | Bare Nuclei | 1-10 |
| 8 | Bland Chromatin | 1-10 |
| 9 | Normal Nucleoli | 1-10 |
| 10 | Mitoses | 1-10 |
| 11 | Class | 2 for benign, 4 for malignant |

In this study, the accuracy of three data mining techniques is compared. The goal is to have high accuracy, besides high precision and recall metrics. The result for the various dataset are given in table 3-5 and shown in figure 3 respectively.

**Table 3**
**Comparison of Various Techniques for WBCO Datasets**

| *Technique* | *Accuracy* | *Precision* | *Recall* |
|---|---|---|---|
| KNN | 81.1 | 0.72 | 0.57 |
| SVM | 83.3 | 0.75 | 0.63 |
| Decision Tree | 87.9 | 0.80 | 0.69 |
| PSO | 88.6 | 0.81 | 0.72 |
| PSO+GA | 90.6 | 0.83 | 0.77 |

**Table 4**
**Comparison of Various Techniques for WDBC Datasets**

| *Technique* | *Accuracy* | *Precision* | *Recall* |
|---|---|---|---|
| KNN | 84.5 | 0.77 | 0.59 |
| SVM | 85.5 | 0.80 | 0.72 |
| Decision Tree | 86.7 | 0.83 | 0.81 |
| PSO | 89.4 | 0.86 | 0.89 |
| PSO+GA | 93.6 | 0.87 | 0.92 |

**Table 5**
**Comparison of Various Techniques for WPBC Datasets**

| *Technique* | *Accuracy* | *Precision* | *Recall* |
|---|---|---|---|
| KNN | 86.1 | 0.80 | 0.63 |
| SVM | 88.3 | 0.82 | 0.77 |
| Decision Tree | 89.7 | 0.84 | 0.85 |
| PSO | 94.3 | 0.88 | 0.90 |
| PSO+GA | 95.1 | 0.90 | 0.97 |

The results show that PSO+GA based approach achieved significantly better performance than other classification methods on all three data sets. In addition, the proposed PSO+GA based method is more efficient than the common PSO method.

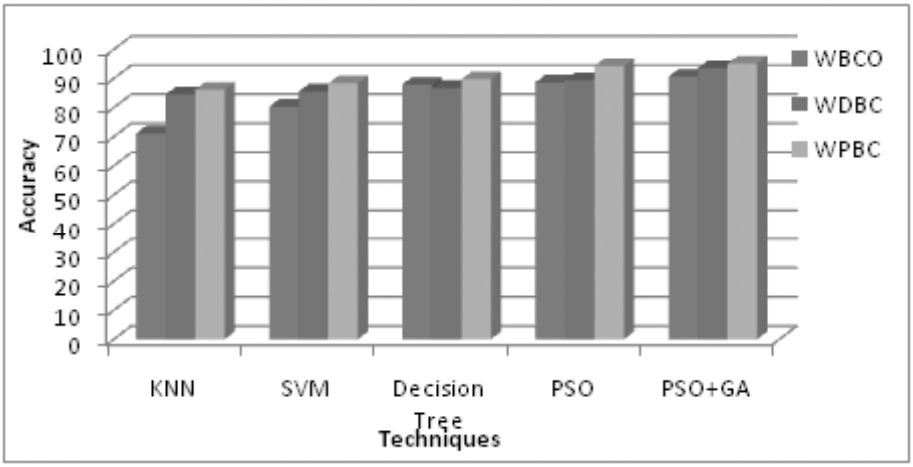


**Figure 3: Comparative analysis of various techniques**

Although data mining methods are capable of extracting patterns and relationships hidden deep into large medical datasets, without the cooperation and feedback from the medical professional, their results are useless. Hence, the diagnosis based on PSO+GA gives better performance and the accuracy of the system is improved.

## 6. CONCLUSION

This paper has outlined, discussed and resolved the issues, algorithms, and techniques for the problem of breast cancer dataset. Unlike classification techniques the proposed approach gives better performance and accuracy. This study clearly shows that the preliminary results are promising for the application of the data mining methods but the evolutionary algorithm based technique. The analysis does not include records withmissing data; future work will include the missing data which also increase the performance of the system.

### *References*

[1] J. Kennedy, R. Eberhart, and Y. Shi, Swarm Intelligence. Morgan Kaufmann Publisher, 2001.

[2] M. L. Mauldin, "Maintaining diversity in genetic search," in Proceedings of the National Conference on Artificial Intelligence (AAAI 1984), August 1984, pp. 247–250.

[3] E. K. Burke, S. Gustafson, and G. Kendall, "A survey and analysis of diversity measures in genetic programming," in Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2002). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, pp. 716–723.

[4] Y. Shi and R. Eberhart, "Population diversity of particle swarms," in Proceedings of the 2008 Congress on Evolutionary Computation (CEC2008), 2008, pp. 1063–1067.

[5] ——, "Monitoring of particle swarm optimization," Frontiers of Computer Science, vol. 3, no. 1, pp. 31–37, March 2009.

[6] S. Cheng and Y. Shi, "Diversity control in particle swarm optimization," in Proceedings of 2011 IEEE Symposium on Swarm Intelligence (SIS 2011), Paris, France, April 2011, pp. 110–118.

[7] S. Cheng, Y. Shi, and Q. Qin, "Population diversity of particle swarm optimizer solving single and multi-objective problems," International Journal of Swarm Intelligence Research (IJSIR), vol. 3, no. 4, pp. 23–60, 2012.

[8] "A study of normalized population diversity in particle swarm optimization," International Journal of Swarm Intelligence Research (IJSIR), vol. 4, no. 1, pp. 1–34, January-March 2013.

[9] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in Proceedings of the Sixth International Symposium on Micro Machine and Human Science, 1995, pp. 39–43.

[10] J. Kennedy and R. Eberhart, "Particle swarm optimization," in Proceedings of IEEE International Conference on Neural Networks (ICNN), 1995, pp. 1942–1948.

[11] R. Eberhart and Y. Shi, "Particle swarm optimization: Developments, applications and resources," in Proceedings of the 2001 Congress on Evolutionary Computation (CEC2001), 2001, pp. 81–86.

[12] S. Cheng, Y. Shi, and Q. Qin, "Population diversity based study on search information propagation in particle swarm optimization," in Proceedings of 2012 IEEE Congress on Evolutionary Computation, (CEC 2012). Brisbane, Australia: IEEE, 2012, pp. 1272–1279.

[13] Kim K.S., W. Kim, K.Y. Na, J.M. Park, J.Y. Kim, K.Y. Lee, J.E. Lee, S.W. Kim, R.W. Park, and Y.S. Jung (2010). "New recurrence prediction model for breast cancer by data mining", p. 136.

[14] Kerheta, A., M. Raffettob, A. Bonia, and A. Massa (2006). "An SVM-based approach to microwave breast cancer detection", Engineering Applications of Artificial Intelligence, No. 19, pp. 807-818.

[15] Shah, S., and A. Kusiak (2007). "Cancer gene search with data mining and genetic algorithms", Computers in Biology and Medicine. No. 37, pp. 251-261.

[16] Qiang, Y., Y. Guo, X. Li, Q. Wanga, H. Chenc, and D. Cuicc (2007). "The diagnostic rules of peripheral lung cancer preliminary study on data mining techniques", Journal of Nanjing Medical University. No. 21(3), pp. 190-195.

[17] Tan, T. Z., C. Queka, G. S. Ng, and K. Razvi (2008). "Ovarian cancer diagnosis with complementary learning fuzzy neural network", Artificial Intelligence in Medicine. No. 43, pp. 207-222,

[18] Yeh, D.Y., C.H. Cheng, and Y.W. Chen (2011). "A predictive model for cerebrovascular disease using data mining", Expert Systems with Applications.

[19] Fabregue, M., S. Bringay, P. Poncelet, M. Teisseire, and B. Orsetti (2011). "Mining microarray data to predict the histological grade of a breast cancer", Journal of Biomedical Informatics. No. 44, pp. 12-16.

[20] Choua, S.M., T.S. Leeb, Y. E. Shaoc, and I.F. Chen (2004). "Mining breast cancer pattern using artificial neural networks and multivariate adaptive regression splines", Expert Systems withApplications, No. 27, pp. 133-142.

[21] Delen, D., G. Walker, and A. Kadam (2005). "Predicting breast cancer survivability: a comparison of three data mining methods", Artificial Intelligence in Medicine, No. 34, pp. 113-127.

[22] Liao H.C., and J.H. Tsai (2007). "Data mining for DNA viruses with breast cancer, fibroadenoma, and normal mammary tissue", Applied Mathematics and Computation, No. 188, pp. 989-1000.

[23] Huang, C., H.C. Liao, and M.C. Chen (2008). "Prediction model building and future selection with support vector machines in breast cancer diagnosis". Expert Systems with Applications, No. 34, pp. 578-587.

[24] Yeh, W.C., W.W. Chang, and Y. Y. Chung (2009). "A new hybrid approach for mining breast cancer pattern using discrete partical swarm optimization and statistical method", Expert Systems with Applications. No. 36, pp. 8204-8211.

[25] Karabatak, M., and M.C. Ince (2009). "An expert system for detection of breast cancer based on association rules and neural network", Expert Systems with Applications, No. 36, pp. 3465-3469.

[26] Song, H., et al., New methodology of computer aided diagnostic system on breast cancer, *in Proceedings of the Second international conference on Advances in Neural Networks*-Volume Part III. 2005, Springer-Verlag: Chongqing, China. pp. 780-789.

[27] Setiono, R., Generating Concise and Accurate Classification Rules for Breast Cancer Diagnosis. *Artificial Intelligence in Medicine*, 2000. 18(3): pp. 205-219.

[28] Meesad, P. and G. Yen, Combined numerical and linguistic knowledge representation and its application to medical diagnosis. *Component and Systems Diagnostics, Prognostics, and Health Management* II, 2003. 4733: pp. 98-109.

[29] [14] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.