Research Article

An Efficient Framework for Performing Discriminative Classification Technique Using WALIF & PSO-GSA Algorithms for Cancer Disease Prediction based on Gene Expression Data

E. Monica Sushil Cynthia[1*] & Dr. S. Kannan[2]

[1*] Research Scholar, Department of Computer Applications, Madurai Kamaraj University, Tamil Nadu, India.

[2] Associate Professor, Department of Computer Applications, Madurai Kamaraj University, Madurai, Tamil Nadu, India.

Email: monicasushil777@gmail.com[1*] & skannanmku@gmail.com[2]

Article History: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

Abstract

Cancer is one of the deadly diseases that affect many people globally. This disease spreads to different parts of the body. So it becomes essential to predict the abnormal growth and the extent of their spread. The goal of this study is to look into the many strategies for cancer disease identification using gene expression data as well as the obstacles that come with them. For gene selection, the model comprises preprocessing of the micro array data of gene expression. Analyzing the characteristics of gene, provide a deep understanding about cancer disease classification. The use of machine learning approaches and statistical methods are used to identify abnormal genes or mutated genes that could be modeled efficiently. We propose a design for predictive gene selection using WALIFS algorithm and PSO-GSA algorithm is used for cancer disease identification. A comparative study with various classification algorithms is made to determine the most appropriate algorithm to classify the gene expression data for cancer. Hence this research work delivers uniqueness and predicts cancer disease using gene expression data with high accuracy.

Keywords: Gene expression data, Feature selection, WALIFS, PSOGSA, Cancer disease Classification algorithm.

1. Introduction

Bioinformatics is a new field that combines computer science, mathematics, and information technology to make decisions and evaluate genetic data. Bioinformatics takes use of the synergies that exist between computational and biological disciplines. While the science of bioinformatics was founded with the goal of extracting data from the 3 billion bases of human DNA, it has now evolved to include the study of data content and data flow in biological systems and processes in general. Rather than cell and molecular biomedicine, computational biologists are primarily interested in developmental, population, and theoretical biology. It is unavoidable that molecular biology plays a key role in computational biology; nonetheless, this is not the focus of computational biology. It appears that computational biologists favour statistical models for biological events over physico-chemical models in these areas of computational biology.

1.1 Motivation

Current microarray research focuses on identifying gene sets that are differentially expressed (DE) or differentially compound (DC) in distinct biological states (e.g. diseased versus non-diseased). It is discovered that a few genes exhibit significant increases or decreases in expression variability in a variety of human disorders (variance). Such differential variability (DV) patterns are also scientifically intriguing since these observed differences in expression variability might be mediated by changes in primary expression dynamics.Gene expression tests are a rapid and easy technique to identify illness indicators that are important in clinical treatment. This investigation looks at the challenge of recognising differentially expressed genes using microarray data. Genes having fundamentally distinct expression in two user-defined sets of microarray trials are known as differentially expressed genes, or discriminator genes. This involves systematically evaluating the performance of each methodology using simulated and biological data at various noise levels.

2. Related work

Natural language processing (NLP) techniques [1], rule-based techniques [2], ontology-based methods, and machine learning (ML) methods [3] are only a few examples of previous work on automatic information extraction from the literature. Because each sort of methodology may capture a distinct component of the extraction challenge, these strat3egies are frequently utilized in combination. Kim et al. [4] presented a machine learning-based protein annotation using data from the biological literature. The authors provide a method for learning rules from a collection of relevant and irrelevant texts that the system is given at the start. Rules that distinguish between relevant and irrelevant utterances are developed using the previously known instances (relevant and irrelevant phrases). The user then chooses a subset of rules from the total retrieved rules that are relevant to him. The chosen rules are put together to establish relations, and the rules associated with each relation are then used to annotate proteins on test data.

The problems of reconstructing gene-regulatory networks from time series of gene expression data were outlined by Stark J et al. [5]. While several computational methods have been studied for reconstructing gene networks from empirical gene expression data, Bayesian network (BN) based techniques have shown considerable promise in inferring causal links between genes and are gaining popularity. The goal of one of the first key studies advocating this method was to use Bayesian networks to infer gene regulatory networks in Saccharomyces cerevisiae from gene expression profiles [6]. Stuart Kauffman [7] employed the simplest dynamic models — synchronous Boolean network models – as a model for gene regulatory networks in the 1960s. The concept behind Boolean networks is that binary on/off switches that operate in discrete time steps may represent fundamental features of gene regulation.

3. Proposed Methodology

The suggested work for the gene expression identification and recognition model is discussed in this section. The following processing steps for gene identification in illness prediction are included in the suggested model of this research work.

3.1 Dataset Description

The datasets were obtained from the NCBI's Gene Expression Omnibus (GEO) repository. The datasets are publicly available in soft file format on the internet. The Table.1 provides a more thorough description of Gene Expression data for Cancer disease.

Table 1. Sample Cancer Disease Gene Expression Dataset

Name of the cancer	Reference series	Organism	No. of Genes collected	No. of samples collected
Prostate Cancer	GDS5072	Homo sapiens	30331	11
Breast Cancer	GDS5076	Homo sapiens	13291	4
Cervical Cancer	GDS5040	Homo sapiens	33297	6
Tongue Cancer	GDS4562	Homo sapiens	45235	96

3.2 Preprocessing of Gene Expression Data

This section depicts the preparation phases of gene expression data for cancer, with a focus on the filtering and normalization procedures, as the decisions made here have a significant impact on the probe set used in subsequent studies. (1) Transformation, (2) Missing Value Removal, (3) Filtering Data, and (4) Proposed Filtering Algorithm are among the preprocessing functions. The roles of pre-processing are defined in the following sections.

(a) Transformation of Gene Expression Data

After we've entered the raw picture data into the gene expression matrix, the next step is to study it and see if we can extract any knowledge about important biological processes. There are two simple methods for examining the gene expression matrix:

1. Associating expression profiles of genes by relating rows in the expression matrix;
2. Associating expression profiles of samples by relating columns in the matrix.

(b) Missing Value Removal of Gene Expression Data

Missing value imputation involves using information about the data to assess the entries that have been lost. There are two types of information that are often available. The correlation structure between elements in the data matrix is the first type of information. Because genes participating in parallel biological processes have similar expression patterns, there is correlation between rows in a gene expression data matrix.

It's possible that a new methodology will allow for the selection of close-to-ideal threshold values for sample means and variances for gene filtering. By analyzing a large number of publicly available and simulated datasets. This well-established adaptive methodology improves the sensitivity of discovering differentially expressed genes as compared to previous microarray data filtering strategies that used preset threshold values.

(c) Filtering Data

The methodology for gene filtering entails removing genes from components, which we presume are often non-informative genes. It is well understood that genes with low mean expression or low variation of expression are more likely to be non-informative. Gaussian components must have the same characteristic. When the sample means or variances are decomposed into Gaussian components, the components can be ordered according to their location parameter (mean of the Gaussian component). Then, on the left hand side of the signal scale, delete genes that correspond to components with the lowest values of this parameter. The algorithm anticipates that including them into the subsequent assessment would result in more false disclosures than real DEG detection.

(d) Proposed Pre-Processing Method for Cancer identification

Two techniques are omitted in this case. The first is based on the "top three" rule (indicated by the acronym "top3" in the following text). More specifically, we predict three components with maximum values for the location parameter, referred to as high-level expressed genes, medium-level expressed genes, and low-level expressed genes, to be useful, and we gather genes that fall into these categories. Other genes aren't connected.

```
Top-3 Gene Preprocessor (D Dataset)
Input:    S-Size of the gene
X-Column Matrix
L-Length of the gene
Output: P-Preprocessed Data
{
ReadGeneSet ($D^{'}$) from Dataset D
S=Size ($D^{'}$) //read the size of the time domain for sample
//Represent x as column-vector
X= X ($D^{'}$)
// compute length of the Data $D^{'}$
L = Gene Length ($D^{'}$)
//Applying Filtering
For each i=1 to n from N // no. of genes in the dataset $D^{'}$
{
For each j=1 to l from L
{
If (!isNull($n_1$))
{
Read Gene($n_1$)
// Check for missing values
If (missing>=$n_1$){
            If (Min ($D^{'}$) <=$n_1$ | Max ($D^{'}$) >=$n_1$) // checks for over expression genes
            {
            T=Select Gene (Top($n_1$)) // select genes to be top N of the highest maximum values
            } }}
P= Filtered Data (T)
}
//represent the filtered gene in the form of the original one
Return P}
```

Figure 1. Algorithm for Top-3 Gene Pre-Processor

3.3 Feature Selection Method on Gene Expression

In order to decrease the dimension and redundancy of gene expression data, feature selection is a key stage in the classification process. An effective and robust feature selection strategy accelerates the learning process of classifiers while also stabilizing classification accuracy.Choosing a subset of comparable characteristics for use in model creation is known as feature selection. When using a feature selection strategy, the essential assumption is that the data contains a lot of redundant or unnecessary information. Irrelevant features provide no benefit in any context, whereas redundant features provide no more information than the currently selected features. Feature selection approaches are commonly used in domains with a large number of characteristics and a small number of samples (or data points).

Weighting Attribute with Locality Interaction based Feature Selection (WALIFS)

In machine learning, examining high-dimensional data is a significant problem. Numerous effective and efficient feature-selection techniques have recently been developed to combat the curse of dimensionality. Most feature-selection algorithms, on the other hand, presume feature independence; they identify significant characteristics based on their distinguishing strong correlation with the target notion. When the assumption of feature independence is accurate, these algorithms can perform well. However, they may struggle in domains where there are feature interactions. Because of feature interactions, a single feature having only a smidgeon of a correlation with the target idea might become extremely connected when seen in conjunction with other features. The removal of these characteristics can severely impair the classification model's performance.

We outline the Weighting Attribute with Locality Interaction based Feature Selection (WALIFS) on the neighborhood interaction gain in this segment, and then go on to our proposed feature subset selection methodology.

In addition to that the feature Fj has an impact on the relationship between feature Fi and class C. Because of the positive neighborhood interaction gain, we can't depict their relationship without considering both of them at the same time, and adding another characteristic would increase the degree of reliance. That is, the addition of feature Fj improves the accuracy of predicting the class variable C. We must consistently increase the weight of feature Fj. Because the new feature has a negative neighborhood interaction benefit, it will reduce the level of reliance. That is, the addition of feature Fj has a detrimental impact on the prediction of the class variable C. We should consistently reduce the weight of feature Fj. As a result, we may use the neighborhood interaction gain to characterize the neighborhood cooperation weight factor. It's possible to look at feature connections and use that information to drive feature selection and design.

The WALIFS of feature Fi with respect to feature Fj is defined as

$$NIW_{\delta}(F_i, F_j) = 1 + \frac{NIG_{\delta}(F_i, F_j)}{NH_{\delta}(F_i) + NH_{\delta}(F_j)}$$

Feature selection algorithms in the past seldom considered redundancy and cooperation at the same time. As a result, certain highly valued characteristics are lost throughout the feature selection process. To address this problem, we first compute the neighborhood common information between a feature and the class, then alter it using the interaction weight factor, which can disclose if a feature is redundant or interactive. The attuned relevance metric will be used to rank the candidate characteristics. Algorithm displays the consistent descriptive pseudo code.

Different search techniques can be used to choose features. In this part, we use the sequential forward search approach for efficiency's sake. The procedure is terminated by a predetermined threshold K, where K is a locality parameter. We rank the features in descending order corresponding to the modified relevance measure and then pick the first K features, where K has been indicated in advance, for a dataset D with the original set F = {F1, F2,...,Fn} and the class C. WALIFS is a feature-based algorithm for ranking features. We use the neighborhood common information LMI(Fi; C) as a measure of relevance after initializing parameters that include the selected feature subset and the weight for each feature.

```
WALIFS Algorithm
{
Input:   $D$ - Dataset
         $F = \{F_1, F_2,...,\}$ - Feature Set
         $C$- Target Class
         $K$- Number of selected feature
         $\delta$ - Locality Size
Output: $S$ Selected Feature sub-set
Set $S$, $k$ sub-set to null
Initial the weighted parameter $w$ to each feature
For each feature $F_i$ from $F$
{
Calculate Locality Mutual Information for each feature $LMI_\delta$ $(F_i; C)$
}
While ($k$< K)
{
For each feature $F_i$from$F$
Calculating relevance measure $R(F_i, C)=w (F_i) x LMI_\delta(F_i, C)$
}
Choose the feature $F_j$from relevance measure
Measure the sub-set $S=S U \{F_j\}$ and eliminate the feature from feature set $F =F - \{F_j\}$
For each feature $F_i$ from F
{
Compute the weighted attribute locality interaction $LIW(F_i, F_j)$
Evaluate the weighted parameter w with respect to the $LIW_\delta$
}
Select the next feature k from K (k=k+1)} }
```

Figure 2. Proposed Feature Selection (WALIFS) for Gene Expression

We might use a specific classifier to choose the subset of characteristics that produces the best accuracy to establish the threshold K. Otherwise, the procedure may be stopped until $|LMI(F; C) - LMI\delta(S; C)| \leq \varepsilon$ is fulfilled, where ε is a very small positive number. Now we'll look at the algorithm's difficulty. Assume n is the total number of candidate features. Figure 2 shows the algorithm. Table 2 shows a vector for feature selection obtained by proposed feature selection algorithm.

Table 2. Feature Selection Vector for Proposed Methodology

WALIFS	989(201461_s_at[MAPKAPK2]),10290(210830_s_at[PON2]),797(201269_s_at[NUDCD3]),5112(205585_at[ETV6]),254(200726_at[PPP1CC]),511(200983_x_at[CD59]),3501(203974_at[HDHD1]),22052(49111_at[ARRB1])

3.4 Proposed PSO-GSA Based Classification

Agents are discussed as objects in the suggested process, and their performance is judged by their masses. The gravitational pull attracts all of these items, causing a global movement of all items toward the ones with heavier weights. As a result, masses collaborate through gravitational pull, which is a direct kind of communication. The heavy masses, which correspond to good solutions, move more slowly than the lighter masses, ensuring the algorithm's exploitation stage.

Position, inertial mass, active gravitational mass, and passive gravitational mass are the four requirements for each mass (agent) in GSA. The mass's location is related to the problem's solution, and its gravitational and inertial masses are calculated using a fitness function. To put it another way, each mass presents a solution, and the algorithm is traversed by changing the gravitational and inertia masses properly.

We presume that the heaviest mass is bothered by the masses as time passes. In the search space, this mass will indicate an optimal solution. The GSA might be viewed as a mass-segregation scheme. It's like a little artificial universe with masses obeying Newton's principles of gravity and motion. More precisely, the laws that govern the masses are as follows: The gravitational force between two particles is directly proportional to the product of their masses and inversely proportional to the distance between them, according to the law of gravity: each particle attracts every other particle, and the gravitational force between two particles is directly proportional to the product of their masses and inversely proportional to the distance between them, R.

In PSO, an agent's orientation is determined by only two best positions, pbest and gbest. In GSA, however, the agent's direction is determined by the total force obtained by all other agents. In PSO, updating is done without regard for the quality of the solutions, and fitness values are unimportant in the process, but in GSA, force is proportional to fitness value, and agents observe the search area around them through the lens of force. For updating the velocity, PSO uses a form of memory (due to pbest and gbest). GSA, on the other hand, has no memory and updates based only on the present position of the agents.In PSO, the distance between solutions is ignored during updating, but in GSA, the force is inversely proportional to the distance between solutions.

PSOGSA is a hybrid meta-heuristic optimization technique based on population. The primary idea behind combining the particle swarm algorithm and the gravitational search algorithm is to overcome both systems' flaws. In later versions, the gravity search algorithm has a sluggish searching pace.

Any meta-heuristic population-based algorithm has two essential qualities that are continually debated. The first is the algorithm's capacity to search over the whole solution space, known as exploration, and the second is the algorithm's ability to reach the optimal solution, known as exploitation. Particle swarm optimization has already demonstrated its effectiveness in exploitation, and GSA is an excellent examiner. Figure 3 depicts the suggested architecture for PSO-GSA.

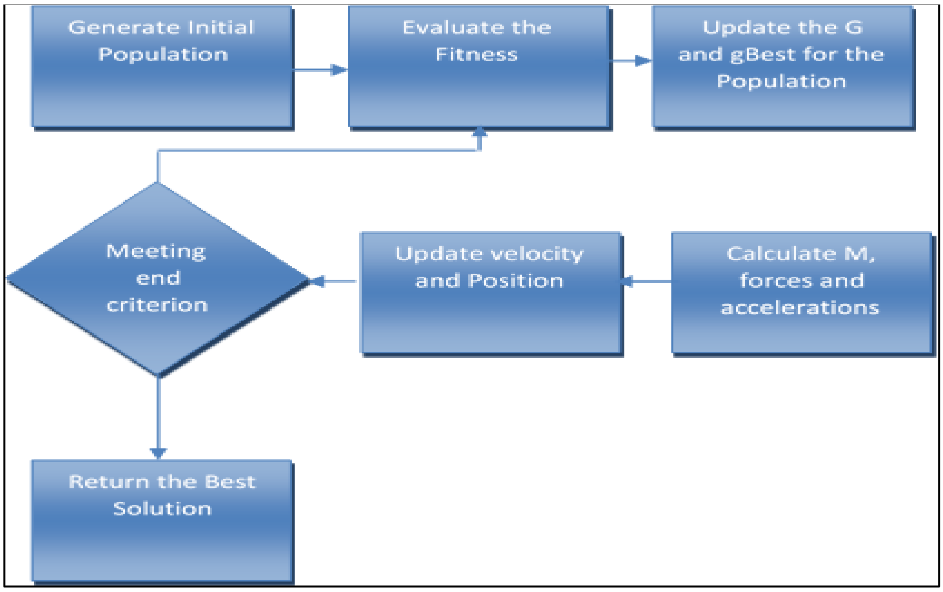
Generate Initial Population
Evaluate the Fitness
Update the G and gBest for the Population
Meeting end criterion
Update velocity and Position
Calculate M, forces and accelerations
Return the Best Solution

Figure 3. Proposed Feature selection Optimization architecture using PSO-GSA

The steps of PSOGSA are depicted in the flowchart image. PSOGSA is far more powerful than PSO or GSA since it replaces the drawbacks of both algorithms with their strengths. It has previously addressed a variety of optimization problems in several disciplines.

In the discipline of software engineering, search-based optimization is used in two stages. After selecting an acceptable demonstration for the programme under test, a fitness function that meets the test sufficiency conditions is well-defined. This fitness function is used to determine whether the candidate solution is capable of achieving the desired result. We chose weighted path based coverage as our fitness function since path coverage is our test adequacy requirement.

It is desirable to recognize the routes of the programme under test here, according to the test adequacy requirements. Using cyclomatic complexity, the upper bound of path numbers is first identified from the control flow graph. Weights are assigned to each edge crossed to track the travelled pathways. Weight 100 is allocated to sequential statements in the control flow graph, whereas weight 20 is allocated to false edges and 80 is allocated to true edges whenever conditional and loop statements are encountered. Finally, the total weight of each path from start to finish is calculated by combining the edge weights. The fitness value for each path is calculated as the sum of these edge weights.

$$f(x) = \sum_{i=1}^{n} \quad P(i)$$

Where, P (i) = fittest path.

$$P(i) = \sum_{i=1}^{m} \quad W(i)$$

Where, W (i) = weight assigned to the fittest path. This fitness function is assessed, to spontaneously create targeted input test information covering all the paths of the program under test. PSO, GSA and PSOGSA have been utilized for the automatic generation of targeted test information.

4. Performance Evaluation

The results of accuracy ratio for the classification methodologies for various feature selection strategies are given in Table 3.

(A) Accuracy

The evaluation of accuracy ratio for the classification methodologies are illustrated in Figure 4. In this figure, the accuracy ratio for the proposed Feature Selection methodology WALIFS provides high accuracy for all classification methodologies employed as shown in Figure 4.

Table 3.Accuracy Results for the Classification Methodologies for the Feature Selection Strategies

Classification / Features	WALIFS	RELIEF	BAT	MRMR
SVM	63.4821	63.3194	64.2756	64.0000
Neural Network	85.2344	85.0000	85.3711	85. 1781
SVM + Neural Network	69.5792	33.5639	33.5630	33.5639

Neural Network + PSO	99.9800	62.5000	75.0000	88.1002
Neural Network + GSA	75.0000	75.0000	50.0000	66.6777
Neural Network + PSO + GSA	99.9800	75.0000	87.5000	95.0542

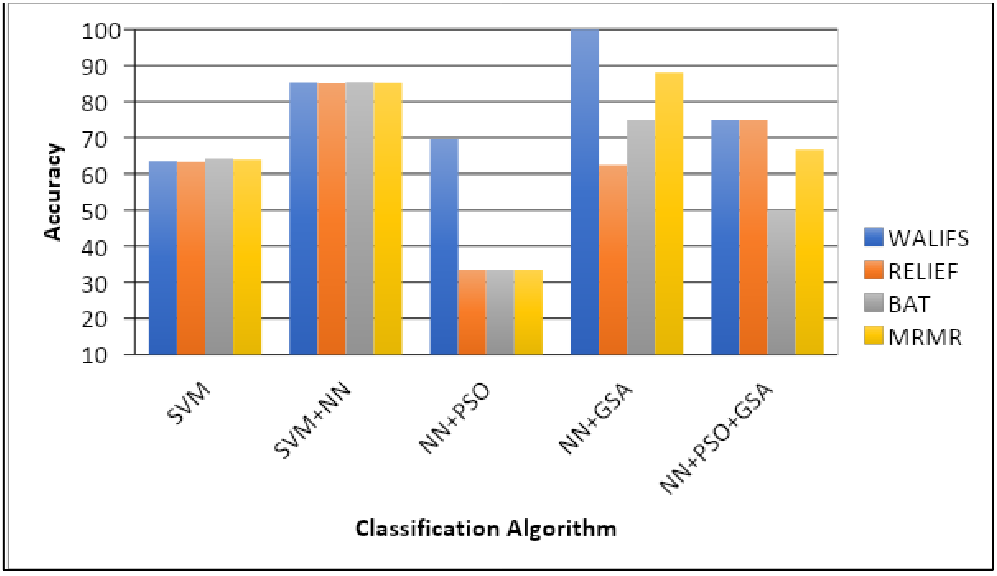

Figure 4. Accuracy Evaluation for Cancer Classification Methodologies

(B) Sensitivity

Sensitivity also recognized as the True Positive rate or Recall is figured as,

$$Specificity = \frac{No. of\ True\ Postivies}{(No. of\ True\ Positives + No. of\ False\ Negatives)}$$

$$Specificity = \frac{TN}{(TP + FN)}$$

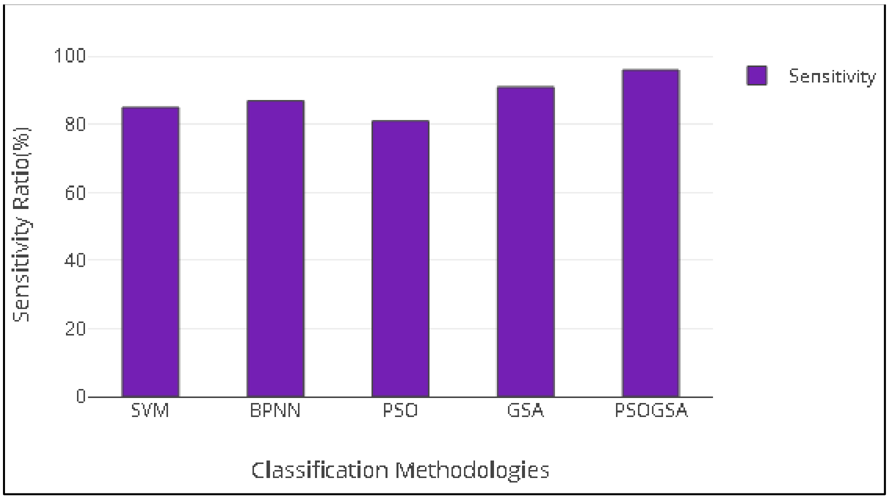

Figure 5. Sensitivity Evaluation for Cancer Disease Classification Methodologies

Since the formula doesn't comprise FP and TN, Sensitivity might provide you a biased result, particularly for imbalanced classes. Sensitivity ratio for various classification algorithms is compared. Proposed methodology outperforms than existing one as shown in Figure 5.

(C) Specificity

Specificity, also recognized as True Negative Rate is computed as,

$$Specificity = \frac{No.of\ True\ Negatives}{(No.of\ True\ Negatives + No.of\ False\ Positives)}$$

$$Specificity = \frac{TN}{(TN + FP)}$$

Specificity ratio for various classification algorithms is compared. i.e The existing classifiers such as SVM,BPNN,PSO, GSA is compared with proposed method PSO-GSA. Proposed methodology outperforms than existing one as shown in Figure 6.

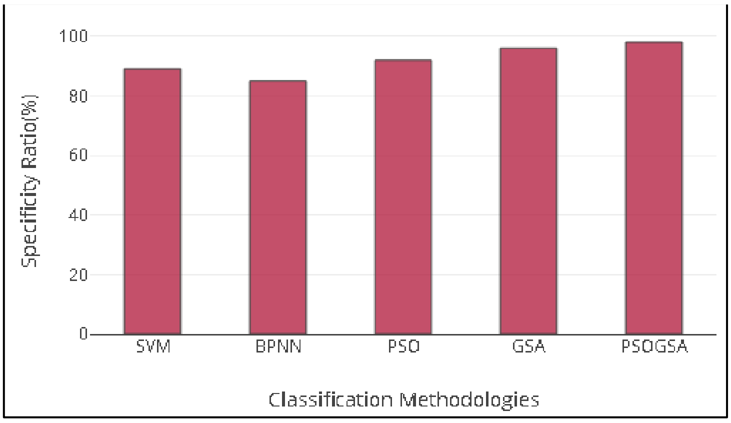

Figure 6. Specificity Evaluation for Classification Methodologies

5. Conclusions

This paper includes the discussion of Micro Array Data Analysis methodologies, as well as a processing component for cancer disease detection. This study went into the construction of gene expression data for cancer recognition algorithms in great depth. In this work, a new approach of analysis on gene selection for cancer classification based on gene expression profile was presented. The approach lead to a fast and adequate classification system and it outperforms earlier published work. cancer. . This work helps the clinicians and researchers for efficient diagnosis, arriving at better treatment protocols for drug discovery and personalized medications.

Future Enhancement: In future more different classifiers can be used as box members. Moreover this system can be analyzed to find out the patients who will be more prone to disease analyzed in future based on the gene biomarkers identification. This system can be analyzed to other benchmark especially multiclass datasets. The proposed methodology outperforms the other existing cancer identification algorithms with high accuracy ratio and efficient in classifying Cancer disease.

References

[1] Song and Yongling, "Text mining biomedical literature for constructing gene regulatory networks," Ph.D. dissertation, University of Florida, 2007, aAI3425541.

[2] Narayanaswamy.M, K. E. Ravikumar, and K. Vijay-Shanker, "Beyond the clause: extraction of phosphorylation information from medline abstracts," Bioinformatics, vol. 21, no. suppl 1, pp. i319–i327.

[3] Tang.Y-T., S.-J. Li, H.-Y. Kao, S.-J. Tsai, and H.-C. Wang, "Using unsupervised patterns to extract gene regulation relationships for network construction," PLoS ONE, vol. 6, no. 5, p. e19633, 05 2011.

[4] Kim.J,H., A. Mitchell, T. K. Attwood, and M. Hilario, "Learning to extract relations for protein annotation," Bioinformatics, vol. 23, no. 13, pp. i256–i263, 2007.

[5] Stark J, Brewer D, Barenco M, Tomescu D, Callard R, et al. (2003) Reconstructing gene networks: what are the limits?.BiochemSoc Trans 31: 1519-25.

[6] Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. J ComputBiol 7: 601-20.

[7] Kauffman S (1969) Homeostasis and differentiation in random genetic control networks. Nature 224: 177-8.

[8] Alshamlan, Hala M., Ghada H. Badr, and Yousef A. Alohali. "Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification." Computational biology and chemistry 56 (2015): 49-60.

[9] Sara Tarek, RedaAbdElwahab and Mahmoud Shoman, "Gene expression based cancer classification," Egyptian Informatics Journal, December 2016.

[10] Hanaa Salem, GamalAttiya and Nawal El-Fishawy, "Classification of Human Cancer Diseases by Gene Expression Profiles," Applied Soft Computing, Vol. 50, pp.124-134, January 2017.