Microarray Gene Expression Dataset Feature Selection and Classification with Swarm Optimization to Diagnosis Diseases

Peddarapu Rama Krishna[1], Dr. Pothuraju Rajarajeswari[2]
Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India[1]
Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India[2]

***Abstract*—Bioinformatic data concentrated on the accumulation of data pace in the undesired information. Bioinformatics data has vast data-intensive biological information through the computation of data. However, bioinformatics data utilizes statistical methods with gene expression for cancer diagnosis and prognosis. Microarray data provides rough approximations for gene expression analysis. Microarray dataset evaluates the massive gene features presence of sample size and characteristics of microarray data. Hence, it is necessary to evaluate the features in the microarray dataset to exhibit effective outcomes through patterns of gene expression. This paper presented a re-sampling of random probability Swarm Optimization (RRP_SW). With RRP_SW model uses the random re-sampling model estimation of features. The features are evaluated through the computation of a multi-objective optimization model. In the microarray, dataset re-sampling estimated the features in the datasets. The features are samples through the computation of probability values in the datasets for classification. With the RRP_SW model, extreme learning is utilized for the classification of features in the microarray dataset with the benchmark datasets.**

Keywords*—Feature Selection; classification; gene expression data; Microarray; RRP_SW; hybrid feature selection

I. INTRODUCTION

Many scientists have been drawn to the study of gene expression levels using microarrays because it is one of the gold standard instruments for doing so. According to the medical community, cancer ranks among the deadliest conditions imaginable. Medical science can manage and cure the disease, but only if caught early. Microarray samples typically have a high feature count, low sample size, and high levels of noise [1]. The features of the microarray dataset have not changed significantly over the past few decades. The microarray dataset is notoriously difficult to analyze due to its high dimensionality, which is defined as an excessive number of features for various examples with an unbalanced number of classes [2].

One uses the feature selection approach to identify disease-associated genes. Classification precision is commonly used as a metric by which to judge the quality of feature selection [3]. Accordingly, categorizations play a significant role in recognizing genes. Disease categorization in gene expression data is often referred to simply as classification. The generalization capability of a classification model can be flawed [4] due to the curse of a limited sample size and large dimensions. Considering these characteristics of microarray datasets, the reduction of dimensions is very much essential before the classification. Dimensionality reduction in gene expression data is commonly thought to be achieved through feature selection [5]. Therefore, accurate disease detection or gene identification using these datasets requires efficient feature selection and appropriate classifiers. On the other hand, bad class imbalance can contribute to erroneous classification results. Thus, it is essential to employ a reliable resampling method in order to address this issue [6].

Many scholars have focused on feature selection strategies in recent years [7]. Many techniques have been proposed for feature selection in order to identify genes that are by effected disease [8]. A hybrid method involving mRMR and SVM-RFE was suggested for selecting relevant genes [9]. An improved version of SVM-RFE, which also relies on a form of mutual information, was suggested [10]. A further disadvantage of SVM-RFE is how long it takes to complete a single analysis. Faster feature selection using a two-stage support vector machine - random forest ensemble [11]. An adapted form of RFE was suggested, in which the target number of features to be dropped varies with each iteration. However, while this method guarantees faster results, the grade of the features selected may suffer. In order to RFE on the fly [12]. These techniques have improved efficiency while decreasing time spent on it. The primary goals of this thesis are to increase processing speed and solve the problem of feature selection so that the quality of feature selections can be enhanced. Class imbalance [13] describes a situation in which there is a large disparity between sample sizes that come from various social strata. Inequitable distribution of resources among groups can lead to unpredictable categorization outcomes. If there are two classes represented in the test group, for instance, and sample X is twice as large as sample Y, then the distribution is skewed. This test dataset's samples were all correctly identified as X, yielding an accuracy of 66.67%, which is higher than 50%. Therefore, it is reasonable to infer that class imbalance will undermine the reliability of the categorization scheme. Therefore, researchers suggested resampling strategies to deal with these issues [14]. Over-sampling and under-sampling are the two common resampling techniques used historically. It's possible that overfitting or data

loss occurred because the samples were randomly chosen from the minority or eliminated from the major and then replicated [15].

Cancer classification using microarray data aims to identify the relevant hidden gene patterns for an accurate diagnosis [16]. The microarray data classification aims to find the significance of the identified genes and their correlation at the genome level. The features are selected based on the identification of a number of gene classes and select the features for the reduction of genes for classification samples. Support Vector Machines (SVM), Decision Trees (DT), Artificial Neural Networks (ANN), K-Nearest Neighbors (kNN), Extreme Learning, and Regularized Extreme Learning [17] are frequently used categorization techniques.

The challenges in today's microarray data are the availability of large numbers of genes and relatively few samples. The number of samples available is limited due to the difficulty in collecting microarray samples [18]. Microarray gene expression data are used to identify a subset of genes that are either co-expressed or expressed differently. The differential genes used to classify the samples based on the expression pattern identified. Co-expressed genes recognize groups with similar patterns of expression as a functional enhancement for the analysis of biological pathways [19]. On the other hand, as biomarkers, deferentially expressed genes are used to define tumors and various tumor sub-types. The attempt to find molecular invariant or differential behavior relevant to a given biological problem has been applied to the gene expression analysis problem [20]. By reducing the number of features and thus increasing the co-relationship between gene expression levels, classification accuracy is improved. Microarray and gene expression analysis has acquired a position in biology and medicine in recent studies. It still requires a much more efficient classification technique to analyze the enormous amount of data [21]. Also, an effective way to determine the relevance of the gene and thus create an excellent diagnostic prediction algorithm is necessary. It is hard to determine gene dependency. Methods of gene selection are therefore required that evaluate each gene separately based on its characteristics [22]. This information must be extracted from microarray data and is an essential issue to address. Extracting interesting gene patterns based on the information obtained is a desirable goal. In order to address all these issues, a more optimized and cohesive framework is needed.

A. Contribution and Organization of Paper

This paper proposed an RRP_SW model for the feature selection and classification in the microarray dataset for disease diagnosis. The specific contribution of the research is presented as follows:

1) To evaluate the gene expression evaluates the re-sampling-based model for the computation of features. Microarray datasets are pre-processed and evaluated based on the sampling process for the evaluation of features in the datasets.

2) Through the re-sampling the features are evaluated, and the probability features are computed based on the estimation of optimal values. The particle swarm optimization features are computed based on the estimated variables.

3) The particle swarm optimization model evaluates the feature estimation variables through probability estimation. The simulation analysis expressed that the proposed RRP_SW model exhibited higher accuracy for the classification of the benchmark datasets.

It is structured as follows in the paper: The relevant works for the microarray datasets are given in Section II. In Section III, we detail our approach to researching the RRP_SW model, and in Section IV, we share our findings from running simulations of the model. In Section V, we show the overall conclusion reached using the proposed RRP_SW model.

II. Related Works

In study [23] proposed an algorithm for the two phases such as the wrapping and filtering process. Initially, the developed model evaluates the steps to minimize the prediction number for the variation in the target based on relevance value. The proposed heuristics model ratio was evaluated based on the compromised rule between relevance and complementary values. With the wrapping phase, the graph-based model is employed for the relevance feature for the complementary values between each other through discriminative features. Through the graph-based feature selection algorithm model, the complementary features are estimated based on the relevance values. The experimental analysis uses the 13-microarray gene dataset with 8 binary and five multi-class microarray datasets. With the 10-fold validation model Support Vector Machine (SVM), Naïve Bayes (NB) and Artificial Neural Network (ANN) are employed. The experimental results demonstrated that hybrid model exhibited the improved performance compared with the conventional classifier model.

As a means of selecting features from a high-dimensional microarray dataset, the study in [24] introduced the Altruistic Whale Optimization Algorithm (AltWOA). AltWOA uses the conventional Whale Optimization algorithm for the efficient propagation of the efficient features optimum in the iteration process. The AltWOA model comprises of the eight high dimensionality dataset exhibits the improved performance compared with the classical technique for the analysis in terms of accuracy and feature selection.

In study [25] adopted ensemble-based feature selection model based on consideration of genetic algorithm and t-test for the computation with the consideration of the optimal feature subsets based on consideration of different datasets. Using an analysis of the Nested Genetic algorithm's performance on a variety of DNA Methylation data sets. When applied to the colon cancer dataset, the Nested-GA dataset created using the Incremental Feature Selection (IFS) strategy for the best subset of genes shows superior performance after 5-fold validation. The experimental validation of the independent dataset provides a classification accuracy of 99.9% based on the consideration of the biological features for the validation of the resulting analysis. For the DNA methylation model the Nested -GA model exhibit the effective feature selection for Gene Expression. The experimental analysis expressed that developed Nested-GA model exhibits the higher classification optimal feature subset

compared with the other algorithm. Through the DNA-Methylation data, the model exhibited an accuracy value of 98.4%.

In study [26] evaluated the CCFS features for the random dataset with the utilization for the cooperation filter criteria. The optimization model uses the fitness function with the estimation of optimal solution space through a gravitational search algorithm. With the CCFS model, several microarray high dimensional datasets are evaluated and compared with the feature selection with Interact (INT) and Maximum Relevancy Minimum Redundancy (MRMR). The experimental analysis expressed that non-parametric statistical analysis is performed for the non-parametric features based on selected features with improved accuracy, sensitivity, and specificity.

In study [27] comparatively evaluated the different feature sets based on the wrapper and fuzzy rough set for the feature selection. The evaluation is based on the consideration of execution time, classification accuracy, and selected feature numbers. The experimental analysis results expressed that feature selection is evaluated based on cancer microarray gene datasets. The results expressed that KNN model exhibited higher accuracy compared with the conventional classifier model. The fuzzy rough set model feature selection model exhibits the computational with the higher and minimal number of genes to estimate the filter correlation features.

In study [28] developed a distributed feature selection model for the fuzzy set model features. The datasets are classified based on the different subsets based on the fuzzy shuffling and set theory. Every subset is individually evaluated HCPF (Hesitant fuzzy set-based feature selection algorithm using Correlation coefficients for Partitioning Features). With the merging procedure, the feature subset is updated and improves the classification accuracy. For the high-dimensional microarray datasets, the technique is tested using a centralized algorithm and 22 sets of distributed features. The experimental analysis demonstrated that the developed model achieves significant results compared with the other non-parametric features approach.

III. Feature Selection with the Re-Sampling Probability Estimation

Feature selection is a crucial step in the analysis of high-dimensional gene expression datasets, such as those obtained from Marray experiments. One effective method for feature selection is the Re-sampling Probability Estimation (RPE) technique. Let $X = [xij]$ be the gene expression matrix where xij denotes the expression level of the j-th gene in the iii-th sample, and $y = [yi]$ be the vector of class labels for the samples. First, each gene is ranked based on a statistical measure. Suppose we use the t-statistic for ranking genes defined in Eq. (1).

$$t_j = \frac{\bar{x}_{i1} - \bar{x}_{2j}}{\sqrt{\frac{x_{1j}^2}{n_1} + \frac{x_{2j}^2}{n_2}}} \quad (1)$$

where $\bar{x}_{i1}$ and $\bar{x}_{2j}$ the mean expression levels of gene j in the two classes, $\frac{x_{1j}^2}{n_1}$ and $\frac{x_{2j}^2}{n_2}$ are the standard deviations, and n_1 and n_2 are the number of samples in each class. To assess the stability of the rankings, we employ bootstrapping. In each iteration k, a bootstrap sample x_k is generated by sampling with replacement from X. The t-statistic is then computed for each gene in the bootstrap sample stated in Eq. (2).

$$t_j^k = \frac{\bar{x}_{i1}{}^k - \bar{x}_{2j}{}^k}{\sqrt{\frac{x_{1j}^{(k)2}}{n_1^k} + \frac{x_{2j}^{(k)2}}{n_2^k}}} \quad (2)$$

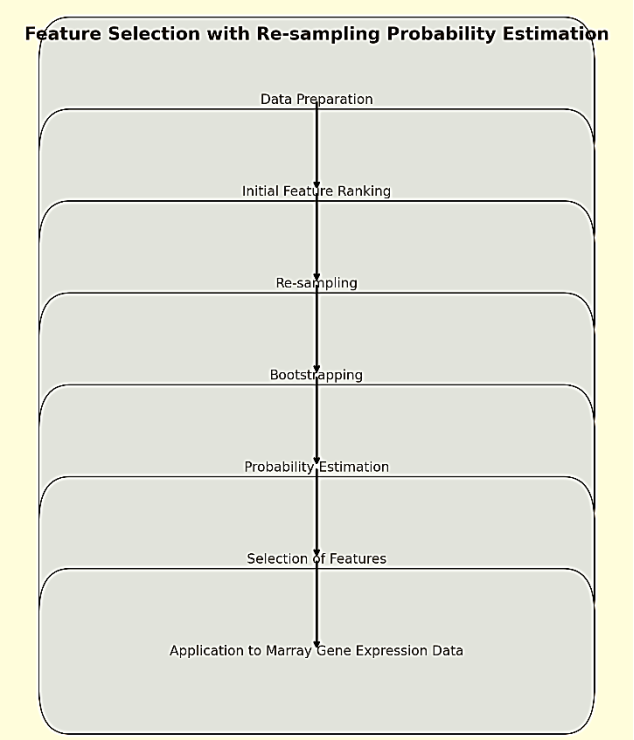


Fig. 1. Flow Chart of Re-Sampling in M-array

After BBB bootstrap iterations, each gene j will have a distribution of t-statistics t_j^k. To estimate the probability that gene j is consistently ranked among the top features, we calculate the frequency p_j with which gene j appears in the top M rankings defined in Eq. (3).

$$p_j = \frac{1}{B}\sum_{k=1}^{B} t_j^{(k)} \quad (3)$$

A threshold τ is set to select the genes with a high re-sampling probability. The selected set of genes $\{S\}$ is given in Eq. (4).

$$S = \{j \mid pj \geq \tau\} \quad (4)$$

This threshold can be chosen based on domain knowledge or statistical criteria such as the false discovery rate (FDR). In Marray gene expression datasets, which typically involve thousands of genes across multiple samples, applying the RPE method helps in identifying a subset of genes that are most relevant to the biological question at hand. For instance, in distinguishing between different disease states, the selected genes are those that consistently show significant differential

expression across multiple bootstrap samples, thus providing a robust and reliable feature set for further analysis.

Algorithm 1: Feature Selection with Re-sampling Probability Estimation

Input: Gene expression matrix X (samples x genes), class labels y, number of bootstrap iterations B, threshold τ

Output: The selected set of informative genes S

1. Data Preparation
   Normalize the data matrix X
   Handle missing values if any
2. Initial Feature Ranking
   for each gene j in X do
      Compute t-statistic t_j based on class labels y
   end for
3. Re-sampling
   Initialize an empty list to store bootstrap rankings
   for k = 1 to B do
      Generate a bootstrap sample x^K by sampling with replacement from X
      Compute t-statistics t_j^k for all genes in the bootstrap sample x^K
      Rank the genes based on t_j^k
      Store the rankings in the list
   end for
4. Probability Estimation
   Initialize a dictionary to count top-M appearances for each gene
   for each gene j do
      Set count[j] = 0
      for each bootstrap iteration k do
         if gene j is in the top-M rankings in bootstrap sample k then
            Increment count[j] by 1
         end if
      end for
      Compute probability $p_j = count[j]/B$
   end for
5. Selection of Features
   Initialize an empty set S
   for each gene j do
      if $p_j >= \tau$ then
         Add gene j to set S
      end if
   end for
6. Return the set of selected genes S

A hybrid approach RRP_SW method is proposed as a hybrid feature selection that combines the advantages of minimizing redundancy and maximizing relevancy (mRMR) and adaptive genetic algorithm (AGA). The architecture of the proposed model is illustrated in Fig. 2.

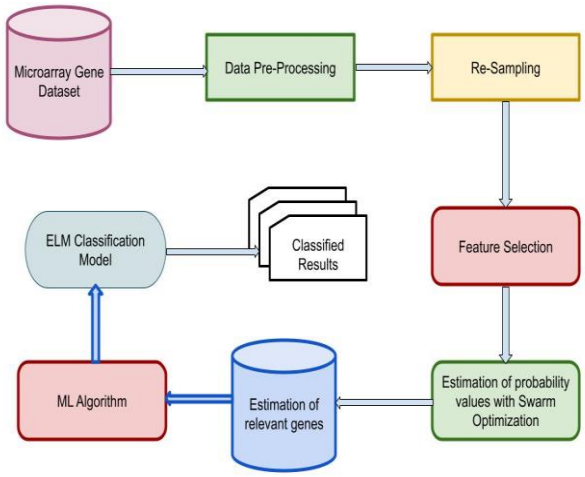


Fig. 2. Architecture of RRP_SW.

A. Classification Problem

The problem of determining the categories of new observation based on the previously analyzed similarity of data is called as classification in machine learning. Classification can be formally defined as:

Definition: X = {x1, · · · xn} are set of given data points. , each of them belongs to a finite set of classes Y = {y1, · · · , ym}, the classification task is to generate a function f : X → Y wherein, elements of X maps to elements of Y xi is known as an instance (or sample), which has a definite set of features $F = \{f1,\cdots,fl\}$ that may be numerical or categorical either. These features are often termed as variables or attributes, which are used interchangeably in this thesis. Every data point xi has an association with a label yi , that shows its class from set Y. The main aim of classification is to design a model, that can determine their label yi for the given data point xi.

Fig. 1 illustrates the system architecture of the proposed techniques. The raw data are gene expression microarray datasets. Due to the potential for the data to be inconsistent and chaotic, it must first undergo pre-processing. The suggested resampling method is then used to generate the balanced datasets. The proposed feature selection technique is then applied at step 48 to choose relevant characteristics.(genes). Finally, various classifiers are utilized to evaluate the efficiency and efficacy of the procedure. The process flow of Proposed RRP_SWM is presented in Fig. 3.

B. Proposed method Re-sampling Based Swarm Optimization

The imbalance of class problem has been addressed using this method. The data on gene expression is biologically specific and, therefore, should not be changed arbitrarily. Therefore, the suggested method intends to deal with the imbalance of class for microarray data without resorting to over-fitting 49 of model and hence losing information, while maintaining the integrity of the inspired biological value. It is believed in this strategy that samples with the same label undergo the same distribution. Under this hypothesis, a data matrix was built that includes a small group. Then, the new sample value for that location was determined by selecting one value at random from each column. To ensure that each class had an equal number of representatives, the current sample was saved, and the procedure

was repeated k times. When all was said and done, k examples were collected that mirrored the original dataset's feature distribution in the microarray data. The RRP_SW method is illustrated below as Algorithm 2:

Algorithm 2: RRP_SW for the Feature Extraction

Input: X - Given minority sample matrix for the data, k- new sample count
Output: X - New data matrix
while (k >= 1) : do
for j = 1, 2, ..., n (n column size of X): do
Random value V chosen from Xj (X column in j^{th} features)
Save V to the respective position of a new sample.
End
Update the new X sample;
k = k − 1;
end
Return X;

In this case, the rows represent samples and the columns indicate genes (features) that will be used to evaluate the given microarray class data (represented by the matrix X).

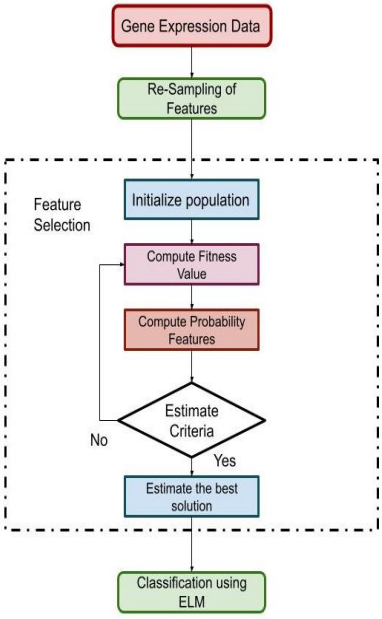


Fig. 3. Flow Chart of RRP_SW.

C. Large Scale Swarm Optimization

In place of SVM, large-scale swarm optimization (LSSO) was used to expedite the weights allocation procedure. LSSOs were specifically created for the purpose of classifying massive amounts of data, such as text. Text data has very big dimensions, and so do the microarray datasets. This means that microarray databases will also work well with LSSO. The goal function of the large-scale liner SVM is given by the Eq. (5):

$$min_w f(w) \equiv \|w\|_1 + C\sum_{i\in I(w)} b_i(w)^2 \quad (5)$$

Were,

$$b_i(w) \equiv 1 - y_i w^T x_i$$

$$I(w) \equiv \{(i|b_i(w) > 0)\}$$

For the ith sample, the feature vector is denoted by xi, while the sampling procedure for feature vector yi is denoted by w. Consequently, in large-scale swarm optimization, the loss function is a square pivoted L1 regularized function. The degree to which the weight vector is sparse is determined by the punishment factor C > 0.(w). As C grows, the weight vector(w) becomes sparser, which penalizes genes with lower significance and thus higher weights to 0. The ultimate decision function looks like Eq. (6) for all swarm optimizations:

$$f(x^*) = sign(w.x^*) \quad (6)$$

The unknown sample feature vector is denoted by x $*$. One variable is updated by cyclic coordinate descent method to generate $w^{k.j} \in R^n, j = 1, \ldots., n+1$ from the current solution w k . Where, j and k refer as feature(variable) and iteration respectively. Thus, $w^{k,1} = w^k, w^{k,n+1} = w^{k+1}$, and hence it is mentioned in Eq. (7).

$$w^{k,j} = \left[w_1^{k+1}, \ldots., w_{j-1}^{k+1}, w_j^k, \ldots., w_n^k\right] \text{ for } j = 2, \ldots., n \quad (7)$$

The one-variable optimization problem shown below was solved, for updating $w^{k,j}$ to $w^{k,j+1}$ as in Eq. (8).

$$min_s gj(z) = \left|w_j + z\right| + L_j'(0,w)z + \frac{1}{2}L_j''(0{:}w)z^2 + constant \quad (8)$$

Where,

$$e_j = [0, \ldots.,0,1,0, \ldots.,0]^T \in R^n$$

$$L_j(z;w) \equiv C \sum_{i\in I(w+zej)} b_i(w + zej)^2$$

And it can be stated as in Eq. (9) and Eq. (10).

$$L_j'(0,w) = -2C\sum_{i\in I(w+zej)} y_i x_i b_i(w) \quad (9)$$

$$L_j''(0{:}w) = max\left(2C\sum_{i\in I(w)} x_{ij}^2, 10^{-12}\right) \quad (10)$$

Since Lj (z;w) is not a double differentiable, so Eq. (10) is an approximate expression.

The variables are evaluated based on the consideration of variable j and z as in Eq. (11).

$$w_j^{kj+1} = w_j^{wj} + z^* \quad (11)$$

D. Mutual Information Relevance and Redundancy

The mutual information about the microarray dataset is evaluated with the re-sampling random probability Swarm Optimization (RRP_SW). The RRP_SW dataset are estimated for the entropy features as presented in Eq. (12).

$$H(X) = \sum_{x=1}^{N_x} P_x(X)\log(P_x(X)) \quad (12)$$

In the above Eq. (8), the probability class for the features are denoted as $(P_x|x = 1,2, \ldots., N_x)$. The average conditional probability of the variables is computed based on the feature vector as in Eq. (13).

$$H(S|X) = \sum_{s=1}^{N_s} P(s)\left(\sum_{x=1}^{N_s} P_x(x|s)log\left(P_s(x|s)\right)\right) \quad (13)$$

With the Eq. (9) the feature vector is represented as N_s for the samples in dataset and the class x conditional probability is

denoted as the $P_x(x|s)$. The entropy values for the conditional probability are evaluated based on the consideration of initial probability features. The class features are independent based on conditional entropy values based on mutual information. The microarray dataset mutual information is represented as $I(X;S)$ with consideration of variables x and s represented as in Eq. (14).

$$I(X;S) = H(x) - H(X|S) \quad (14)$$

The above Eq. (9) is redefined as in Eq. (15).

$$I(X;S) = I(S;X) = \sum P(x.s) log\left(\frac{P(x,s)}{P(x)P(s)}\right) \quad (15)$$

The RRP_SW mutual information property is computed based on the symmetricity property with variables S and X as $I(X;S) = I(S;X)$ .With RRP_SW the attribute mutual information I is estimated based on discrete variables S and X as in Eq. (15). The redundancy is reduced in the feature through the consideration of mutual information with the computation of maximal dissimilarity between the genes in the microarray datasets. Here, the gene subsets are evaluated based on the consideration of minimal redundancy average value as in Eq. (16).

$$minimum(W) = \frac{1}{|s|^2}\sum_{i,j\in s} I(i,j) \quad (16)$$

Where, $I(i,j)$ represented the mutual information in the ith and jth genes in the microarray dataset with the gene denoted as $|s|$. Through mutual information genes are expressed with the mutual information $I(h,i)$ as in Eq. (13). The relevance of the mutual information between gene target classes is defined as $h_1, h_2, \ldots\ldots\ldots, h_k$. The maximal relevance between the gene variables subset s as defined in Eq. (17).

$$maximum(V) = \frac{1}{|s|^2}\sum_{i\in s} I(h,i) \quad (17)$$

The swarm optimization model for the proposed RRP_SW model is presented in Fig. 4.

E. Experimental Evaluation

The experimental analysis of the RRP_SW performs the verification based on the consideration of different experimental datasets. The dataset for analysis comprises a set of heterogeneous classes of more than two. The dataset for the analysis is presented as follows:

1) Datasets: The Microarray gene expression data set has been utilized for a vast range of experimental analyses of the datasets. The dataset for the analysis is comprised of information about diseases such as breast, small round blue cell tumor (SRBCT), Lymphoma, Lung, and other cancer datasets. The characteristics and features of the different diseases' datasets are presented in Table I.

The final process in proposed RRP_SW comprises the Back-Propagation mechanism with the assigned neural network value with fine-tuning of the error through iteration process. The rate of error in the network are fine-tuned with the assigned weights to perform reliable model design through generalization and improvisation as shown in Fig. 5.

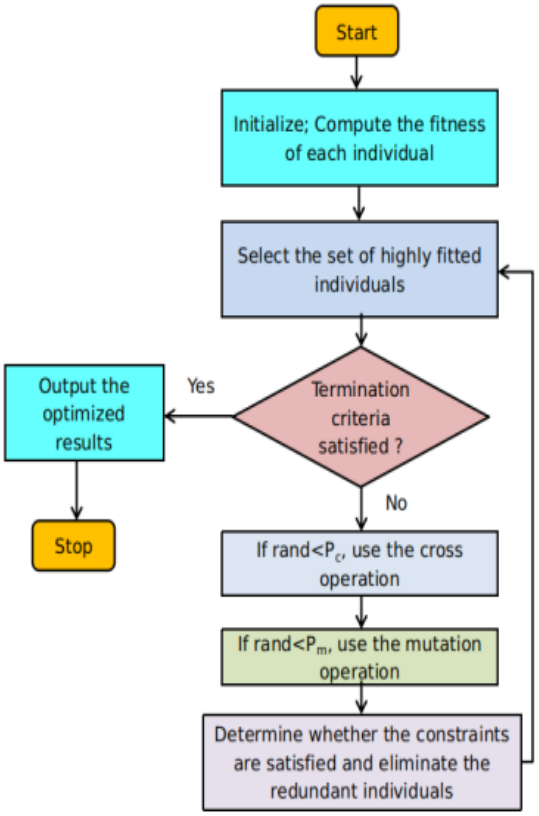


Fig. 4. Flow chart of swarm optimization.

TABLE I. CHARACTERISTICS OF DATASETS

Datasets	Number of Features	Number of Samples	Number of Classes	Class Description
Breast	24481	97	2 (46- 51)	46 Normal 51 Cancer
Lung	12600	203	5 (139-17-6-21-20)	139 AD 17 NL 6 SMCL 21 SQ 20 COID
Lymphoma	4026	62	3 (42-9-11)	42 DLBCL 9 FL 11 CLL
SRBCT	2308	*63*	4 (23-8-12-20)	23 EWS 8 NHL 12 NB 20 RMS

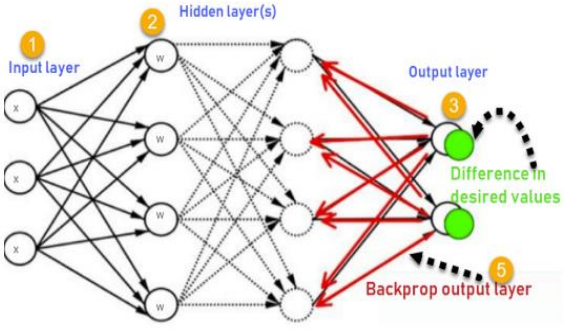


Fig. 5. Architecture of Back Propagation Neural Network (source:guru 99.com).

The Neural Network outputs are stated as L with the set of the training set N with the sample set of (x,t) as stated in Eq. (18).

$$\sum_{j=1}^{L} \beta_j \emptyset \left(\omega_j x_i + b_j\right), \quad i \in [1, N] \tag{18}$$

The neural network model input, output, and target layers are presented in Eq. (19).

$$y_i = \sum_{j=1}^{L} \beta_j \emptyset \left(\omega_j x_i + b_j\right) = t_i + \epsilon_i, \quad i \in [1, N] \tag{19}$$

The RRP_SW comprises of two stages with the conversion of hidden neurons in to represented input data. The input layer comprises of the biases and weights for the estimation of the data presented in hidden layer with the non-linear activation function. The evaluation process uses the extreme learning process as shown in Fig. 6. The matrix computation in RRP_SW model extreme learning process is presented as $\beta = (\beta_1^T \dots \dots \dots \beta_L^T)^T$, $T = (y_1^T \dots \dots \dots y_L^T)^T$

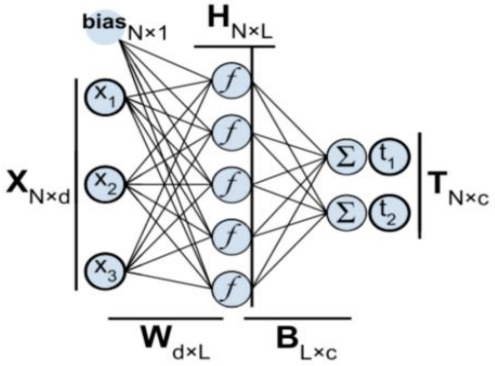


Fig. 6. ELM in RRP_SW.

The microarray data set comprises of the correlated or irrelevant information for the hidden layer model with L1 regularization. With extreme learning process L1-regulated with the pruning of neurons for the robust performance of network. ELM model comprises of the building model to derive the relevance of output.

IV. Simulation Analysis

The proposed RRP_SW model comprises of the 10 times features with the selected microarray datasets under different genes target number. The targeted gene microarray dataset comprises of the different number of genes. The analysis of microarray dataset gene for the selection are presented in Table II.

TABLE II. Number of Genes Selected on Microarray Datasets With RRP_SW Selection

Dataset	Gene in Steps									
	1	2	3	4	5	6	7	8	9	10
Breast	22	47	69	94	135	128	163	165	185	215
Lung	29	67	91	108	119	137	161	165	184	214
Lymphoma	14	39	66	81	121	127	156	170	186	208
SRBCT	24	49	74	93	134	136	163	180	195	210

TABLE III. A Breast Dataset

Gene	Count in Top 10	Probability p_j
22	90	0.90
47	85	0.85
69	92	0.92
94	88	0.88
135	80	0.80
128	83	0.83
163	89	0.89
165	91	0.91
185	87	0.87
215	84	0.84

TABLE III B. Lung Dataset

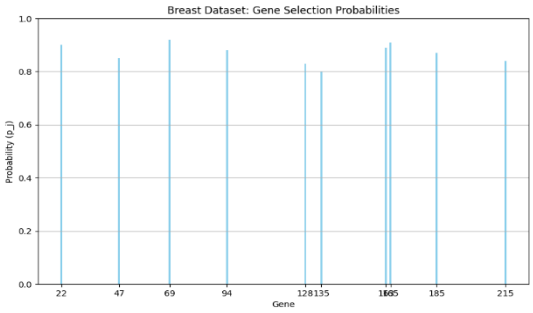
Gene	Count in Top 10	Probability p_j
29	86	0.86
67	89	0.89
91	91	0.91
108	85	0.85
119	83	0.83
137	84	0.84
161	90	0.90
165	87	0.87
184	82	0.82
214	88	0.88

TABLE III C. Lymphoma Dataset

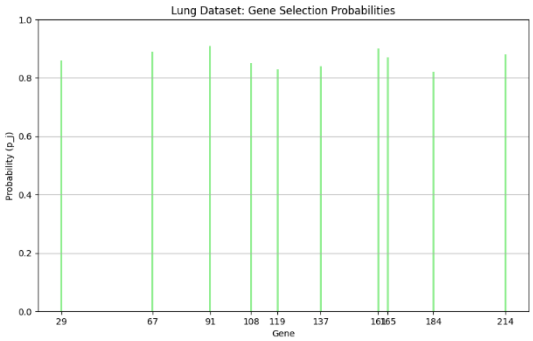
Gene	Count in Top 10	Probability p_j
14	95	0.95
39	87	0.87
66	89	0.89
81	82	0.82
121	85	0.85
127	88	0.88
156	86	0.86
170	91	0.91
186	84	0.84
208	90	0.90

TABLE III D. SRBCT Dataset

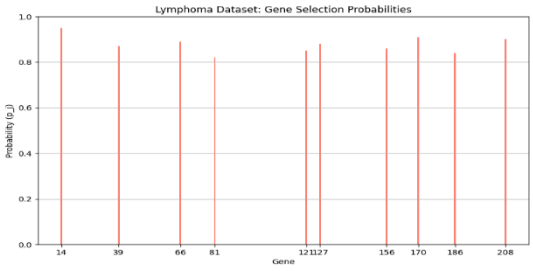
Gene	Count in Top 10	Probability p_j
24	92	0.92
49	85	0.85
74	87	0.87
93	89	0.89
134	83	0.83
136	88	0.88
163	90	0.90
180	86	0.86
195	84	0.84
210	91	0.91



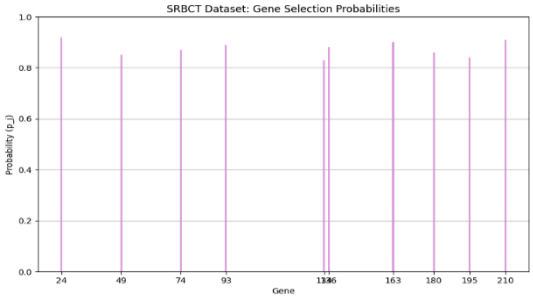
(a)



(b)



(c)



(d)

Fig. 7. Feature Selection with (a) Breast (b) Lung (c) Lymphoma (d) SRBCT.

The provided Tables III (A, B, C, and D) and Fig. 7(a) – Fig. 7(c) present datasets related to different types of cancer: breast, lung, lymphoma, and SRBCT (small round blue cell tumors). Each table lists gene counts and their corresponding probabilities of being in the top 10 genes associated with each cancer type. The gene counts represent how frequently each gene appears in the top 10 list, while the probabilities (pjp_jpj) indicate the likelihood of each gene being among the top 10 based on the dataset.

In each table:

- The gene counts range from 22 to 215 in Table IIIA (Breast), 29 to 214 in Table IIIB (Lung), 14 to 208 in Table IIIC (Lymphoma), and 24 to 210 in Table IIID (SRBCT).
- The probabilities (pjp_jpj) vary between 0.80 and 0.92 in Table IIIA, 0.82 and 0.91 in Table IIIB, 0.82 and 0.95 in Table IIIC, and 0.83 and 0.92 in Table IIID.

These tables likely serve as data points for statistical analysis or machine learning models aiming to identify genes most relevant to each cancer type based on their frequency and probability of occurrence in top-ranking lists. The variability in gene counts and probabilities across different cancer types reflects the diversity and specificity of genetic factors associated with each disease, crucial for advancing targeted diagnostic and therapeutic strategies in oncology research. The rate of classification accuracy for the microarray dataset are evaluated with RRP_SW model. The classification accuracy is estimated with the average results obtained with the 30 times of the classification process presented in Table IV.

TABLE IV. CLASSIFICATION ACCURACY OF RRP_SW SELECTION AND ELM

Dataset	Classification Accuracy Rate %									
	1	2	3	4	5	6	7	8	9	10
Breast	98.67	98.45	99.04	98.67	95.62	99.04	98.34	95.42	99.73	96.43
Lung	98.56	98.35	98.73	97.34	96.72	95.34	99.04	98.75	98.27	97.56
Lymphoma	99.03	98.42	99.43	99.04	95.24	97.31	99.45	99.46	97.63	98.36
SRBCT	99.28	98.15	98.52	98.74	94.92	98.35	99.34	97.61	97.94	97.84

When compared to the other feature selection method, the RRP_SW model's swarm optimization-based classification accuracy is significantly higher. The RRP_SW model is comparatively examined with the ReliefF, SFS and mRMR algorithms. The classification accuracy of the RRP_SW model is evaluated for the different microarray dataset such as SRBCT, Lung, breast and Lymphoma dataset shown in Fig. 6.

The RRP_SW model perform the feature selection with the balanced dataset for the experimental analysis of the raw datasets for the gene selection. The feature selection is performed with the SVM based model for the defined objective set function. The SVM classifier uses the class value C = 1 for the dataset 128 genes.

The Fig. 6 to Fig. 8 provides the comparative performance analysis of the raw dataset balance with consideration of measures such as accuracy, MCC and AUC. Experimental analysis of boldface microarray dataset variables is presented in Table V. With the comparative examination of the microarray dataset Leukemia exhibits the effective and significant performance for the different variables. The comparative analysis expressed that MCC, ACC and AUC model exhibits the

significant performance for the balances dataset such as Colon and Breast cancer. Additionally, it is observed that for ovarian dataset the performance of RRP_SW exhibits the minimal performance compared with the other datasets. Additionally, it is expressed that raw datasets exhibits the significant performance for the increase in genes to resolve the imbalance class in the microarray datasets.

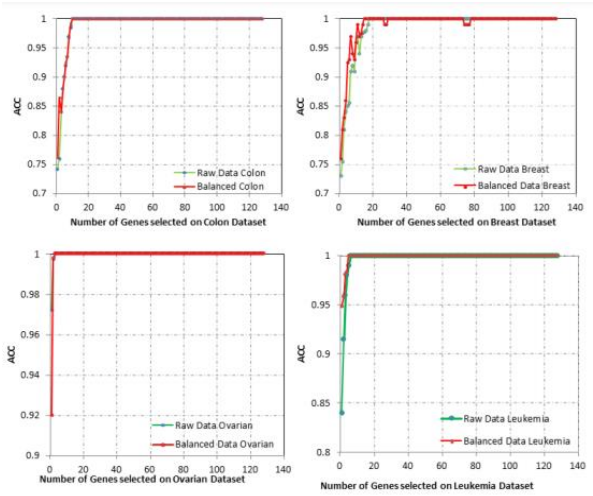


Fig. 8. Accuracy analysis for the different datasets.

TABLE V. COMPUTATION OF FEATURE CLASSIFICATION

	Raw Datasets			Balanced Datasets		
	ACC	AUC	MCC	ACC	AUC	MCC
Breast	98.64	99.05	99.85	99.87	99.35	99.87
Lung	99.85	99.34	99.73	99.46	1.0	99.93
Lymphoma	99.04	1.0	99	99.97	99.85	99.58
SRBCT	98.96	99.97	98.86	99.86	99.96	99.49

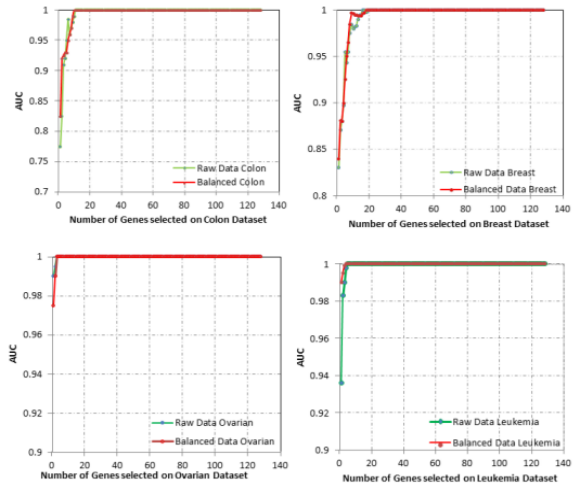


Fig. 9. Comparison of AUC for the different datasets.

With the proposed RRP_SW model the feature selection is performed with the consideration of different feature selectors. The experimental analysis is performed with the consideration of the different methods such as mRMR and SVM-RFEVSS. Through swarm optimization model the classifier model exhibits the balanced dataset for the selected 1 to 128 genes as in Table VI. Fig. 9 shows comparison of AUC for the different datasets.

TABLE VI. CLASSIFICATION OF RRP_SW

	SVM-RFEVSS	RRP_SW
Breast	4671.82	995.06
Lung	98.87	99.03
Lymphoma	2346.14	687.48
SRBCT	786.60	99.84

In Table VI-time consumption is measured for the LSSO-RFEVSS and SVM-RFEVSS which reveals that proposed RRP_SW model exhibits the minimal time for the reduced time consumption for the feature selection for the high dimensional dataset with different classification method are shown in Table VII.

TABLE VII. COMPARISON OF CLASSIFICATION

Datasets	kNN			SVM			ELM – RRP_SW		
	ACC	AUC	MCC	ACC	AUC	MCC	ACC	AUC	MCC
Breast	77.455	81.67	78.96	83.79	84.73	93.46	99.84	99.73	91.39
Lung	80.44	83.46	81.34	87.50	81.39	91.46	98.45	98.94	90.77
Lymphoma	79.04	79.87	79.45	88.92	80.35	96.35	99.49	99.03	93.73
SRBCT	82.35	80.57	84.68	90.74	85.70	93.72	99.78	99.38	99.83

Furthermore, it is observed that classifier SVM model exhibits the significant performance for the RRP_SW model. Through the classifier model the features in the microarray dataset is evaluated based on the sample size with the microarray data analysis. Fig. 10 shows classification for different datasets.

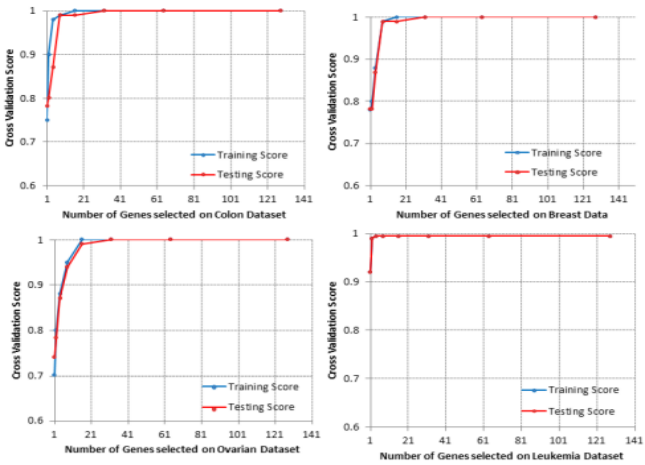


Fig. 10. Classification for different datasets.

Microarray data is characterized by a high degree of dimensionality, a relatively small sample number, and the presence of class imbalance. The Class imbalance issue is rarely

addressed in this field of study, among them. In this chapter, a simple but effective method known as RRP_SW was proposed for pre-processing the datasets, and the intern solved this problem. The balanced datasets were obtained by using these methods. For example, many scholars in this area rely on the tried-and-true SVM-RFE method. To lessen the time needed for SVM-RFE's processing, a newer version of RFE, RFEVSS, was suggested. A bigger initial step size helped reduce recursion time; further reduction of the step size when the features are to be eliminated further decreased recursion time, guaranteeing high-quality, meaningful gene selection. There is a vast pool of genes at play in the human body, but only a small subset is actually involved in illness development. Therefore, effective feature picking must be implemented. Even though a practical version of swarm optimization, called LSSO, was developed. For microarray datasets, Large Linear Support Vector Machine (LSSO) is a pure linear classifier based on a support vector, which acquires the benefits of SVM while decreasing the expense of computational effort (large scale linearly separable data). The results section demonstrates that the resulting method, which is referred to as LSSO-RFEVSS, is an effective and efficient feature selector in comparison to other current feature selectors. Finally, experiments were run to determine the effect of various classifiers on the findings, and it was found that Logistic Regression was superior in the vast majority of instances.

A. Limitations and Future Scope

The limitations of the study on Feature Selection with Re-Sampling Probability Estimation (RPE):

1) Computational complexity: The RPE method involves bootstrapping and recalculating t-statistics for multiple iterations, which can be computationally intensive, especially with large datasets. This may limit its applicability to high-dimensional gene expression data where computational resources are constrained.

2) Dependency on bootstrap size: The performance of the RPE technique is dependent on the number of bootstrap iterations. A higher number of iterations may improve stability but also increase computational time and resource usage. Conversely, too few iterations might lead to unreliable feature rankings.

3) Threshold selection: The choice of threshold τ for selecting significant genes can be subjective and may impact the results. The threshold is typically chosen based on domain knowledge or statistical criteria, which might not always capture the most relevant features accurately.

4) Feature redundancy: While RPE helps in identifying relevant genes, it might not fully address the issue of feature redundancy. Some genes might be highly ranked but redundant in terms of the information they provide, potentially leading to overfitting in subsequent models.

5) Limited generalizability: The RPE method was evaluated on specific cancer datasets (e.g., breast, lung, lymphoma, SRBCT). Its effectiveness on other types of gene expression datasets or in different biological contexts is not fully explored.

6) Assumption of consistent gene distribution: The RRP_SW method assumes that samples with the same label have similar distributions, which may not always hold true in real-world datasets. This could lead to inaccurate balancing and potential loss of biological significance.

7) Potential for overfitting: The feature selection process, particularly when combined with a high number of genes and complex models, may lead to overfitting, where the model performs well on training data but poorly on unseen test data.

Future research on the RRP_SW method for gene expression data could focus on integrating it with advanced machine learning techniques like deep learning to enhance classification accuracy. Exploring hybrid feature selection methods, improving scalability, and handling missing, or noisy data are key areas. Extending the method to other omics data, developing enhanced evaluation metrics, and ensuring model interpretability will also be valuable. Additionally, applying RRP_SW in real-world biomedical research, benchmarking against other methods, and creating user-friendly tools can further its impact and usability.

V. Conclusion

Microarray dataset comprises of the expression of genes characterized by the higher number of gene features in the samples. To evaluate the feature selection-based approach is proposed re-sampling random probability Swarm Optimization (RRP_SW). The RRP_SW model effectively minimizes the dimensionality of the data in specified time through minimal redundancy features in the datasets. The re-sampling-based model effectively increases the classification accuracy for the 20000 gene dataset. The RRP_SW perform the gene selection with the reduced gene minimal than 300 for the accuracy of classification. The RRP_SW model exhibits improved feature selection compared with the conventional feature selection model for the different benchmark datasets. The experimental analysis stated that proposed RRP_SW model exhibist the significant performance for the feature selection and classification of microarray datasets.

References

[1] H. S.Basavegowda, and G. Dagnew, "Deep learning approach for microarray cancer data classification," CAAI Transactions on Intelligence Technology, vol. 5, no.1, pp.22-33, 2020.

[2] T. Bhaskar, M.N. Narsaiah and M. Ravikanth, "Central Medical Centre Healthcare Data Security with Lightweight Blockchain Model in IoT Sensor Environment," *Journal of Sensors, IoT & Health Sciences*, vol.01, no.01, pp.15-26,2023.

[3] E. A.Alhenawi, R. Al-Sayyed, A. Hudaib, and S. Mirjalili, "Feature selection methods on gene expression microarray data for cancer classification: A systematic review," Computers in Biology and Medicine, vol.140, pp.105051, 2022.

[4] A. Jahwar, and N.Ahmed, "Swarm intelligence algorithms in gene selection profile based on classification of microarray data: a review," Journal of Applied Science and Technology Trends, vol.2, no.01, pp.01-09, 2021.

[5] M.Abd-Elnaby, M. Alfonse, and M. Roushdy, "Classification of breast cancer using microarray gene expression data: a survey," Journal of Biomedical Informatics, vol.117, pp.103764, 2021.

[6] S.H. Shah, M.J. Iqbal, I. Ahmad, S. Khan, and J.J. Rodrigues, "Optimized gene selection and classification of cancer from microarray gene expression data using deep learning," Neural Computing and Applications, pp.1-12, 2020.

[7] P. Brundavani, D. Vishnu Vardhan and B. Abdul Raheem, "Ffsgc-Based Classification of Environmental Factors in IOT Sports Education Data during the Covid-19 Pandemic," *Journal of Sensors, IoT & Health Sciences*, vol.02, no.01, pp.28-54,2024.

[8] G.Zhang, J.Hou, J. Wang, C.Yan, and J. Luo, "Feature selection for microarray data classification using hybrid information gain and a modified binary krill herd algorithm," Interdisciplinary Sciences: Computational Life Sciences, vol.12, no.3, pp.288-301, 2020.

[9] O.Alomari, S.N.Makhadmeh, M.A. Al-Betar, Z.A.A. Alyasseri, I.A. Doush, et al., "Gene selection for microarray data classification based on Gray Wolf Optimizer enhanced with TRIZ-inspired operators," Knowledge-Based Systems, vol.223, pp.107034, 2021.

[10] S. O.Abdulsalam, A. A.Mohammed, J. F.Ajao, R. S.Babatunde, R. O. Ogundokun, C. T. Nnodim, and M. O. Arowolo, "Performance evaluation of ANOVA and RFE algorithms for classifying microarray dataset using SVM," In European, Mediterranean, and Middle Eastern Conference on Information Systems , pp. 480-492, 2020.

[11] E.H.Houssein, D.S. Abdelminaam, H.N. Hassan, M.M. Al-Sayed, and E. Nabil, "A hybrid barnacles mating optimizer algorithm with support vector machines for gene selection of microarray cancer classification," IEEE Access, vol.9, pp.64895-64905, 2021.

[12] B.Haznedar, M.T.Arslan, and A. Kalinli, "Optimizing ANFIS using simulated annealing algorithm for classification of microarray gene expression cancer data," Medical & Biological Engineering & Computing, vol.59, no.3, pp.497-509, 2021.

[13] K.Rezaee, G.Jeon, M.R.Khosravi, H.H. Attar, and A. Sabzevari, "Deep learning-based microarray cancer classification and ensemble gene selection approach," IET Systems Biology, vol. 16, no.3-4, pp.120-131, 2022.

[14] A.S.M. Shafi, M.M. Molla, J.J. Jui, and M.M. Rahman, "Detection of colon cancer based on microarray dataset using machine learning as a feature selection and classification techniques," SN Applied Sciences, vol.2, no.7, pp.1-8, 2020.

[15] R.Tabares-Soto, S.Orozco-Arias, V.Romero-Cano, V.S. Bucheli, J.L. Rodríguez-Sotelo, and C.F. Jiménez-Varón, "A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data," PeerJ Computer Science, vol.6, pp.e270, 2020.

[16] S. Venkatramulu,Md. Sharfuddin Waseem ,Arshiya Taneem ,Sri Yashaswini Thoutam,Snigdha Apuri and Nachiketh, "Research on SQL Injection Attacks using Word Embedding Techniques and Machine Learning," *Journal of Sensors, IoT & Health Sciences*, vol.02, no.01, pp.55-64,2024.

[17] A.Dabba, A. Tari, S.Meftali, and R. Mokhtari, "Gene selection and classification of microarray data method based on mutual information and moth flame algorithm," Expert Systems with Applications, vol.166, pp.114012, 2021.

[18] M.Rostami, S.Forouzandeh, K.Berahmand, M.Soltani, M. Shahsavari, and M. Oussalah, "Gene selection for microarray data classification via multi-objective graph theoretic-based method," Artificial Intelligence in Medicine, vol.123, pp.102228, 2022.

[19] V.Nosrati, and M. Rahmani, "An ensemble framework for microarray data classification based on feature subspace partitioning," Computers in Biology and Medicine, vol.148, pp.105820, 2022.

[20] G.Dagnew, and B.H.Shekar, "Ensemble learning-based classification of microarray cancer data on tree-based features," Cognitive Computation and Systems, vol.3, no.1, pp.48-60, 2021.

[21] Bejjam Komuraiah, "IoT Health Science Data Analytics Model for the Prevalence of Anxiety and Depression in Working Professionals," *Journal of Sensors, IoT & Health Sciences*, vol.02, no.02, pp.30-40,2024.

[22] K.Cahyaningrum, and W.Astuti, "Microarray gene expression classification for cancer detection using artificial neural networks and genetic algorithm hybrid intelligence," In 2020 international conference on data science and its applications, pp. 1-7, 2020.

[23] H.Chamlal, T.Ouaderhman, and F.E. Rebbah, "A hybrid feature selection approach for Microarray datasets using graph theoretic-based method," Information Sciences, 2020.

[24] R.Kundu, S.Chattopadhyay, E.Cuevas, and R. Sarkar, "AltWOA: Altruistic Whale Optimization Algorithm for feature selection on microarray datasets," Computers in Biology and Medicine, vol.144, pp.105349, 2022.

[25] S.Sayed, M.Nassef, A.Badr, and I. Farag, "A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets," Expert Systems with Applications, vol.121, pp.233-243, 2019.

[26] M.K. Ebrahimpour, H.Nezamabadi-Pour, and M. Eftekhari, "CCFS: A cooperating coevolution technique for large scale feature selection on microarray datasets," Computational biology and chemistry, vol.73, pp.171-178, 2018.

[27] C.A.Kumar, M.P. Sooraj, and S. Ramakrishnan, "A comparative performance evaluation of supervised feature selection algorithms on microarray datasets," Procedia computer science, vol.115, pp.209-217, 2017.

[28] M. K.Ebrahimpour, and M. Eftekhari, "Distributed feature selection: A hesitant fuzzy correlation concept for microarray high-dimensional datasets," Chemometrics and Intelligent Laboratory Systems, vol.173, pp.51-64, 2018.