Journal of Cancer Research and Clinical Oncology (2023) 149:15171–15184
https://doi.org/10.1007/s00432-023-05308-7

RESEARCH

Classification of cancer cells and gene selection based on microarray data using MOPSO algorithm

Mohammad Reza Rahimi[1] · Dorna Makarem[2] · Sliva Sarspy[3] · Sobhan Akhavan Mahdavi[4] · Mustafa Fahem Albaghdadi[5] · Seyed Mostafa Armaghan[6]

Received: 4 August 2023 / Accepted: 16 August 2023 / Published online: 27 August 2023

Abstract

Purpose Microarray information is crucial for the identification and categorisation of malignant tissues. The very limited sample size in the microarray has always been a challenge for classification design in cancer research. As a result, by pre-processing gene selection approaches and genes lacking their information, the microarray data are deleted prior to categorisation. In essence, an appropriate gene selection technique can significantly increase the accuracy of illness (cancer) classification.

Methods For the classification of high-dimensional microarray data, a novel approach based on the hybrid model of multi-objective particle swarm optimisation (MOPSO) is proposed in this research. First, a binary vector representing each particle's position is presented at random. A gene is represented by each bit. Bit 0 denotes the absence of selection of the characteristic (gene) corresponding to it, while bit 1 denotes the selection of the gene. Therefore, the position of each particle represents a set of genes, and the linear Bayesian discriminant analysis classification algorithm calculates each particle's degree of fitness to assess the quality of the gene set that particle has chosen. The suggested methodology is applied to four different cancer database sets, and the results are contrasted with those of other approaches currently in use.

Results The proposed algorithm has been applied on four sets of cancer database and its results have been compared with other existing methods. The results of the implementation show that the improvement of classification accuracy in the proposed algorithm compared to other methods for four sets of databases is 25.84% on average. So that it has improved by 18.63% in the blood cancer database, 24.25% in the lung cancer database, 27.73% in the breast cancer database, and 32.80% in the prostate cancer database. Therefore, the proposed algorithm is able to identify a small set of genes containing information in a way choose to increase the classification accuracy.

Conclusion Our proposed solution is used for data classification, which also improves classification accuracy. This is possible because the MOPSO model removes redundancy and reduces the number of redundant and redundant genes by considering how genes are correlated with each other.

Keywords Classification of cancer cells · Gene selection · Multi-target particle swarm optimisation · Microarray

✉ Mohammad Reza Rahimi
m_rahimi17@yahoo.com

✉ Mustafa Fahem Albaghdadi
mustafafahem20@gmail.com

Dorna Makarem
dorna.makarem@alumnos.upm.es

Sliva Sarspy
sliva.sarspyy@gmail.com

Seyed Mostafa Armaghan
amer.armaghan@aut.ac.ir

[1] Software Engineering, Qeshm Institute of Higher Education, Qeshm, Iran

[2] Escuela Tecnica Superior de Ingenieros de Telecomunicacion Politecnica de Madrid, Madrid, Spain

[3] Department of Computer Science, College of Science, Cihan University-Erbil, Erbil, Iraq

[4] Sadjad University Bachelor of Engineering–BE, Computer Engineering, Mashhad, Iran

[5] Information Technology Unit, Al-Mustaqbal University College, Babylon 51001, Iraq

[6] Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran

Introduction

Since its inception in 1996, microarray technology has also been referred to as (deoxyribonucleic acid) DNA arrays, gene chips, DNA chips, and biochips (Sarhan 2009; Venkatesan et al. 2022). One of the most recent advancements in molecular biology is the use of microarray technology, which enables the simultaneous monitoring of the expression of thousands of genes in a single hybridisation experiment. Aside from its scientific promise in the fundamental study of gene expression, this technique also has significant applications in the control and interactions of genes in pharmacological and clinical research. Microarrays, for instance, can be used to identify diseased genes for therapeutic medications or to assess their effect by comparing gene expression in healthy and unhealthy cells (Daoud and Mayo 2019; Lai et al. 2020).

Each of the microarray's thousands of spots carries a unique known DNA sequence, or marker (Lu and Han 2003; Trik et al. 2023). A robotic arrayer prints these dots on a glass slide. The most often used microarrays are two of them: oligonucleotide arrays, also known as oligo for short, and complementary DNA arrays (arrays based on complementary DNA (Guillen and Ebalunode 2016; Samiei et al. 2023). In general, there are two key distinctions between (Complementary DNA) cDNA microarray and oligonucleotide microarray: first, the length of the DNA fragment in cDNA microarray is longer than the length of DNA fragment in oligonucleotide microarray; and second, in cDNA microarray experiments, two (ribonucleic acid) RNA samples the control sample and the experimental sample are labelled with two different fluorescents (Cy3 and Cy5) (Wang et al. 2020; Ayyad et al. 2019).

A cDNA microarray is used to compare the levels of gene expression in two distinct samples. The procedure for gathering six-step microarray data in an experiment is shown in Fig. 1. These steps can be used to obtain a DNA microarray, as shown in the figure: selecting samples from various samples (such as malignant and healthy samples); reverse transcription, cDNA production, and mRNA isolation from samples; marking cDNA samples with the fluorescent dyes Cy3 and Cy5 in green and red, respectively; spooning the samples onto a slide that has the appropriate gene sequences already printed on it; Scrub the slide's surface; scanning with a hybrid array. The labelled samples are combined in the cDNA microarray and then hybridised with DNA molecules on the slide's surface. The hydrogen bond is created more strongly the more base pairs that are complementary to one another (Mokhlesi et al. 2020; Chen et al. 2022). The slide is washed following the hybridisation of the labelled samples with the DNA molecules on the slide's surface. Weaker bonds are lost after washing whereas stronger ones are preserved. The number of samples that have a strong connection with the surface sequences determines the intensity and quality of the final signal.

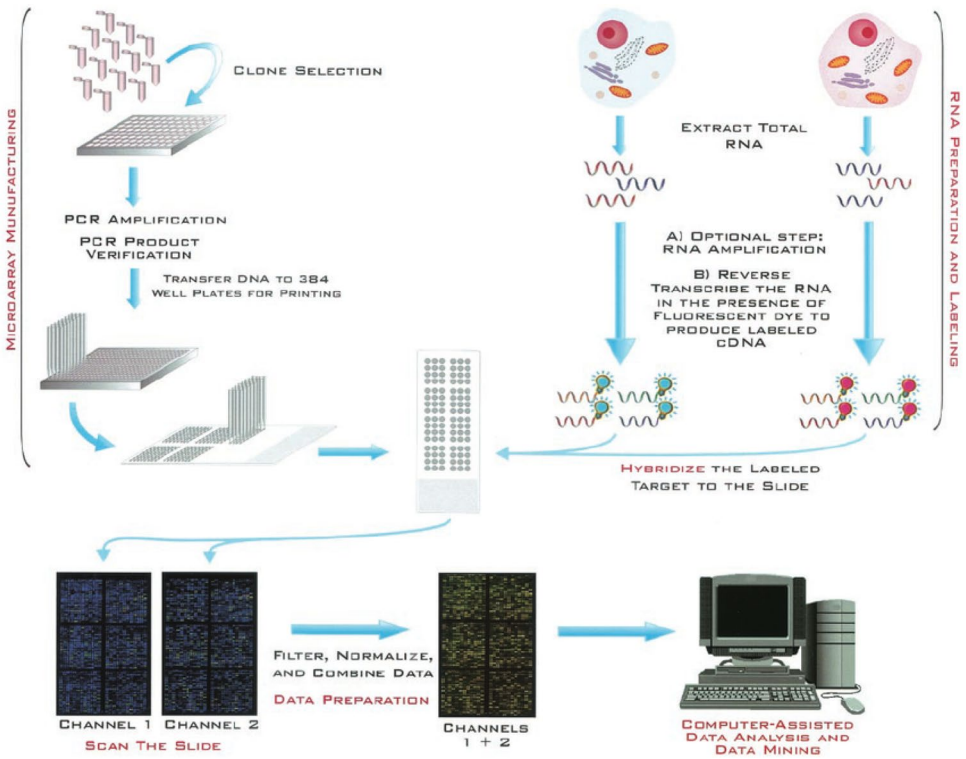

Fig. 1 Several stages of microarray data collecting. The microarray experiment consists of six phases

The microarray data are presented as a matrix with thousands of columns and a few hundred rows, where each column corresponds to a gene and each row to a sample. Microarray data analysis has had issues because of the high dimension of the characteristics and the relatively small number of samples. The issues are as follows: (a) increasing computational expense and classification difficulty, (b) decreasing the validity of classifiers in forecasting fresh samples and their capacity for generalisation, (c) when identifying genes with varying expression and creating predictive models, there is a considerable likelihood that unrelated genes will manifest themselves due to the high number of features compared to samples, and (d) it is challenging to interpret disease-causing genes since, biologically speaking, only a small subset of genes are connected to the illness. As a result, the data pertaining to the vast majority of genes behave as background noise, which can cancel out the impact of that small fraction. Thus, a more accurate interpretation of the function of information-containing genes results from concentrating on a smaller range of gene expression data. Therefore, reducing the number of genes or, to put it another way, choosing discriminating genes for categorisation, is the first crucial step in the analysis of microarray data.

This paper presents an ideal technique for this purpose that is based on a hybrid model of multi-objective particle swarm optimisation and linear Bayes discriminant analysis. Multi-objective particle swarm is a powerful and complex computational algorithm that extracts and selects features related to cancer cells in a simultaneous and optimal manner. This algorithm has the ability to reduce data dimensions and extract important information by considering several different objectives, which is very necessary in the analysis of complex biological data. The main motivation in this study is the importance of microarray data in cancer diagnosis and improving the accuracy and reliability in cancer diagnosis, which can help to improve efficiency in the classification of diseases, especially cancer, by choosing the right genes. In general, the main contribution of the authors in this paper is the classification of cancer cells and the selection of genes based on microarray data using the MOPSO algorithm and the evaluation of the results of this method in comparison with other methods, which has led to improved classification accuracy for four cancer databases.

When this approach is used, the noise in the microarray data is reduced, and the classification accuracy of the data also increases. The used microarray databases are then introduced in the next section. The overall method for feature extraction and classification of microarray data is proposed in the third section, and its various processes are described. In the fourth section, the suggested hybrid model is introduced and its many processes are thoroughly detailed. This model is based on a hybrid approach that combines multi-objective particle swarm optimisation and linear Bayes discriminant analysis. The fifth section presents the outcomes of the application on four databases, and the last section concludes with summaries.

Databases

We employ four microarray databases in this article, which we shall describe below. It is important to note that high-density oligonucleotide arrays were used to measure all the samples (Gupta et al. 2022; Khezri et al. 2022). The information in this article was taken from the source (Lei et al. 2022).

Leukaemia: This database includes 72 samples from microarray studies with 7129 different levels of gene expression. The separation of acute lymphoblastic leukaemia (ALL) (Alharbi and Vakanski. 2023) and acute myeloid leukaemia (AML) (Haoyan et al 2022) data, two categories of blood cancer, is the key issue. The information is split into two groups: The testing method employed 34 control samples (20 related to ALL and 14 related to AML), whereas the training process used 38 cancer samples (27 related to ALL and 11 related to AML) are split. Table 1 contains a brief description of the data set used in the proposed framework. 97 breast cancer samples from microarray studies with 24,481 different levels of gene expression are included in this database. The information is split into 2 groups: 78 samples (34 related to relapsed samples) and 19 control samples (12 related to relapsed samples (Haoyu et al 2022; Yanwei et al. 2023b) and 7 related to non-recurring samples (Khezri and Zeinali 2021) were used in the testing process. and 44 examples including non-relapse samples) that were used during training are separated.

Lung cancer: There are 181 samples from microarray studies with 12,533 different levels of gene expression in this collection. The information is split into 2 groups: 32 samples (16 linked to MPM and 16 items connected to ADCA samples) are used in the training process, while 149 control samples (15 related to MPM samples and 134 related to ADCA samples) are used in the test process (Shayan et al. 2021; Shekar and Dagnew 2019).

Table 1 Datasets

Datasets	Genes	Class	Data		
			Samples	Train	Test
Leukaemia	7129	ALL	72	38	34
Breast	24,481	AML	97	72	25
Lung	12,533	AML	181	128	53
Prostate	12,600	ALL	136	102	34

136 prostate cancer samples from microarray studies with 12,600 different levels of gene expression are included in this database. The information is broken down into two groups: 102 samples (52 related to tumour samples and 50 related to non-tumour samples) and 34 control samples (25 related to tumour samples and 9 related to non-tumour samples) used in the testing process. There are various normals employed in the training process.

The general process of feature extraction and microarray data classification

Gene selection, a procedure in which a few genes are chosen before classification, is one of the crucial difficulties in microarray data processing. Microarray data processing has had issues because of high dimensions, a small sample size, and intrinsic variability in biological and laboratory processes. Therefore, reducing the number of genes or, to put it another way, choosing distinguishing genes for categorisation, is the first crucial step in the analysis of microarray data. Gene selection is the process that follows. The general steps of feature extraction and microarray data categorisation are depicted in Fig. 2. The broad order of these processes is as follows, which is covered in great depth as follows:

- pre-processing of gene expression data;
- deciding on a group of information-carrying genes;
- Data classification

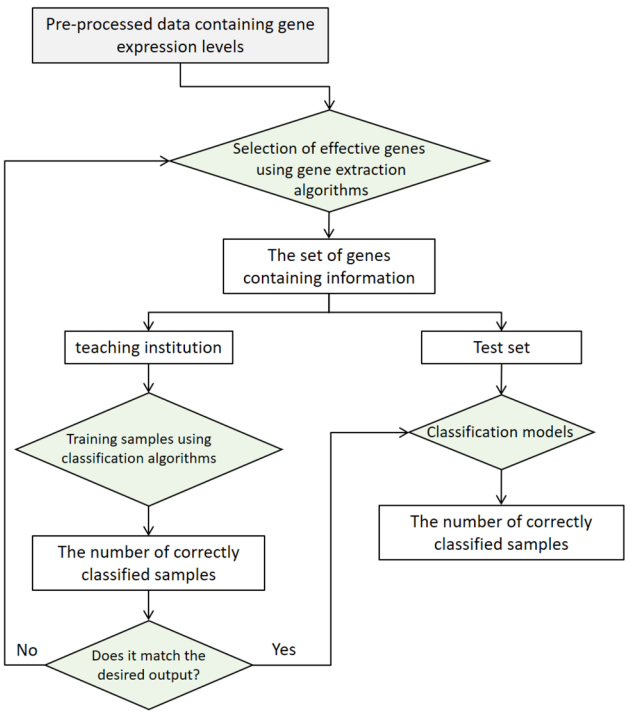

Fig. 2 Block diagram of different stages of microarray data analysis

- Analysing and verifying the results that were obtained.

Gene expression data pre-processing

Before applying feature extraction and classification algorithms, it is necessary to pre-process the used data. The data used in this article are called using Weka software, which have the attribute-relation file format (ARFF) (Guo et al. 2018; Sun et al. 2022). Due to the large changes of data, discretisation of its values is necessary to achieve proper classification accuracy. Discretisation is the process of converting continuous characteristics and variables into discrete values or characteristics (Rezaei et al. 2023; Khezri et al. 2023). Usually, during this process, the data are divided into k sections with the same length (same intervals) or $k\%$ of the total data (same frequencies).

Selecting the set of genes containing information for microarray data classification

Research has demonstrated that categorising microarray data can be used to accurately diagnose cancer. Despite this, the primary issue in analysing microarray data is its high dimension, which is brought about by the comparison of a very large number of variables (genes) with a very small number of samples. Even though there are many genes in the microarray data, very few of them significantly affect how accurately the genes are classified. Numerous genes serve the same purpose in both healthy and diseased (cancer) states. In addition, some genes show up in the data as noise. While the presence of noisy genes does not contribute to the development of cancer, their impact on classification accuracy is detrimental. As a result, by pre-processing gene selection approaches and genes lacking their information, the microarray data are deleted prior to categorisation. By doing this, the classifier's effectiveness will be improved, and the computational complexity will be decreased (Sun et al. 2019; Almugren and Alshamlan 2019).

Generally speaking, there are three models for feature (gene) selection (Basavegowda and Dagnew 2020; Trik et al. 2022). The filter model, which performs feature selection and classification in two distinct processes, is the first model. This model chooses genes with a strong capacity for discrimination as effective genes. This model is straightforward and quick to compute and is independent of classification or learning algorithms. The wrapper model, the second model, combines feature selection and classification into a single operation. In order to find efficient genes, this model employs categorisation. In other words, the wrapper model tests the chosen gene subset using a learning process. The filter model is less precise than the wrapper model. The articles have discussed a number of techniques for choosing appropriate subsets based on the wrapper model. This is accomplished in Mostavi et al. (2020),

by combining evolutionary methods with the K closest neighbour classifier. Using adaptive operators, parallel genetic algorithms have been created in (Yanwei et al. 2023a). In addition, in study Liu et al. (2015), a collection of genes were chosen and classified using a hybrid genetic algorithm and support vector machine model. The issue of gene selection and classification is given as a multi-stage optimisation problem in study Sharma and Rani (2019), where the goal is to simultaneously reduce the number of features (genes) and the number of samples that are incorrectly categorised.

Finally, in hybrid models, the process of selecting a set of effective genes during the training process is performed by a special classifier. An example of this method is the use of a support vector machine with the elimination of recursive features. The idea of this method is to delete genes one by one and check the effect of their deletion on the expected error (Ghosh et al. 2019; Zhang et al. 2020). Recursive feature removal algorithm is a backward feature ranking method. In other words, the genes that are removed in the last step obtain the best classification results, while these genes may not correlate well with the classes alone. Hybrid models can be considered the generalised state of the envelope model. Two other examples of the hybrid model are mentioned in the research of Sridevi et al. (2022).

Evaluation and validation of results

Evaluation of the outcomes of applying the classification algorithms is the final step in the analysis of microarray data. The k-fold validation method is used in this article to analyse the outcomes, where k is the number of repetitions and its value is taken to be 10. For this reason, the ten-fold algorithm is validated by first splitting the samples into 10 equal parts. Then, each time the algorithm is run, 0.1 of the total data is utilised as test samples, and the other data is thought of as teaching examples. The experiment is repeated ten times using separate training and test data, and the result is calculated by averaging the results.

The proposed model based on the combined algorithm of MOPSO/LBDA

The suggested model is based on linear Bayesian discriminant analysis, multi-objective particle swarm optimisation, and Pearson correlation analysis. The proposed algorithm's block diagram is shown in Fig. 3. The following are the main steps of this algorithm, which are described in detail as follows:

- Using Pearson correlation analysis to pre-process data and choose a group of genes;
- Choosing the group of genes that contain information and classifying them using a combination MOPSO/LBDA algorithm.

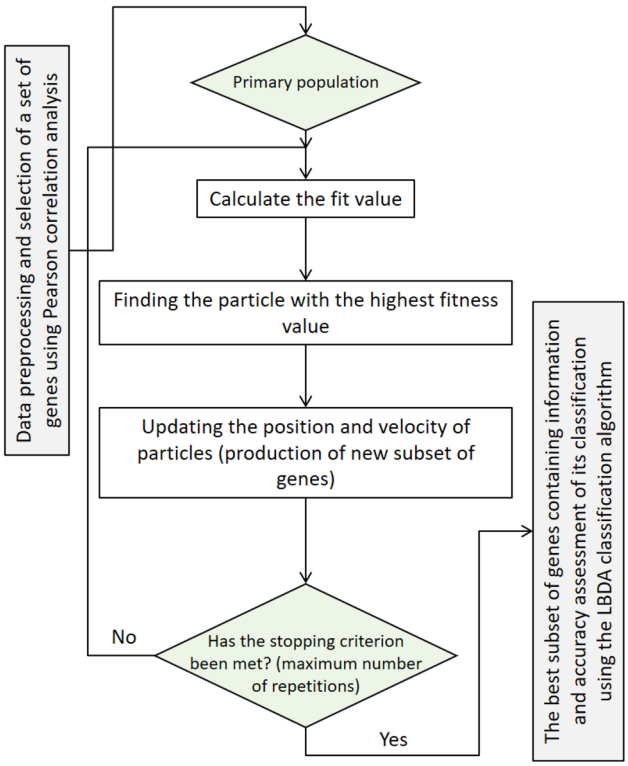

Fig. 3 Block diagram of the proposed algorithm

Data pre-processing and selection of a set of genes using Pearson correlation analysis

Two ideal binary feature markers are defined according to Eq. (1) to score each gene. The first pointer in class A has a value of one, while the second pointer in class A has a value of zero, while the third pointer in class A has a value of one:

$$\begin{aligned} \text{Ideal1} &= (1,1,1,1,0,0,0,0,0,0) \\ \text{Ideal2} &= (0,0,0,0,1,1,1,1,1,1) \end{aligned} \tag{1}$$

Genes are chosen as genes providing information for categorisation if they are similar to the markers, or if the distance between the genes and the markers is modest. In this article, Pearson's correlation analysis criterion which is described as follows (Taheri et al. 2023) is used to determine how far apart each gene is from the markers:

$$PC = \frac{\sum_{i=1}^{n} \left(\text{ideal}_i - \mu_{\text{ideal}}\right)\left(g_i - \mu_g\right)}{\sqrt{\sum_{i=1}^{n} \left(\text{ideal}_i - \mu_{\text{ideal}}\right)^2}\sqrt{\sum_{i=1}^{n} \left(g_i - \mu_g\right)^2}} \tag{2}$$

When g_i is the value of the gene vector and iideal is the binary value of the ideal marker vector, n is the number of training samples, μ_g is the mean of the gene, μ_{ideal} is the mean of the ideal$_i$ marker, and ig is the value of the gene vector. The Person correlation (PC) criterion is determined for each gene for the initial pre-processing procedure, and the k genes with

the smallest PC are chosen as the genes carrying the primary information.

Multi-objective particle swarm optimisation algorithm

Kennedy and Eberhart first put forth the concept of particle swarm optimisation in 1995 (Singh et al. 2016). PSO is an iterative, evolutionary computing method that draws inspiration from nature. The social behaviour of animals, such as the group movement of birds and fish, served as inspiration for this algorithm. Each solution to a problem is represented in this method as a particle. Each particle uses both its own experience and the population's experiences to determine the best solution to the issue. There are a specific number of particles in the particle swarm optimisation (PSO) algorithm. Each particle has two defined positions and speeds, which are represented by a position vector and a velocity vector, respectively. The best location of each particle in the past is kept in one memory, while the best position of all particles is kept in another memory. The particles choose their course of action for the following turn based on the experience gained through these memories. Each moving particle modifies its position by altering its speed in accordance with both its own optimal position and the optimal position for all moving particles. Therefore, the movement of each particle is determined by three variables: the particle's current position, its best position to date (Pbest), and the best position to date for the entire collection of particles (Gbest) (Abd-Elnaby et al. 2021). PSO has been utilised with success in a variety of domains, including function optimisation, the training of neural networks, fuzzy control systems, etc. The binary type of PSO method, which is applied in cases of binary discrete variables, was also introduced by Kennedy and Aberhat in 1997. In MOPSO, the speed of each particle is determined by the amount of bits changed during each iteration, and the position of each particle is described as a binary vector of zero and one (Debata and Mohapatra 2022).

Linear Bayesian discriminant analysis method

An adjustable technique called the LBDA is utilised to avoid overfitting (Zeinali et al. 2023) in high-dimensional data. With the use of training data and this method, the amount of correction may be swiftly and automatically predicted. This classifier can identify data with noise and features that are difficult to categorise accurately (Petrini et al. 2022). Regression is carried out using the Bayes framework, which forms the foundation of this classification (Fabin et al. 2023). In this method, the relationship between the targets and the feature vector will be linear. This connection looks like as follows:

$$t = w^T x + n \tag{3}$$

where w is the weight vector ($w \in R^D$), n is white noise, and t and x are the target and feature vectors, respectively. In this instance, the similarity function for the regression's w-weights is written as follows:

$$p(D|\beta, w) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp\left(-\frac{\beta}{2}\left\|X^T w - t\right\|^2\right) \tag{4}$$

where D stands for two parameters that are $\beta, \{X, t\}$ the inverse of variance, N stands for the number of samples in the training set, and ($X \in R^{D\times N}$) is a row matrix comprising feature vectors. The prior distribution for the weight vectors must be identified in order to characterise a Bayes set. This distribution, which is described as follows, offers fundamental details about the weight vector:

$$p(w|a) = \left(\frac{a_i}{2\pi}\right)^{\frac{D}{2}} \exp\left(-\frac{1}{2} w^T I'(a) w\right) \tag{5}$$

where is a_i square diagonal matrix of dimensions $D+1$, where D is the number of features, and it reflects the inverse of the variance of the initial distribution for the weight vectors. Applying Bayes' rule in the manner described below will yield the posterior distribution given a prior distribution and a similarity function:

$$p(w|\beta, aD) = \frac{p(D|\beta, W)p(w|a)}{\int p(D|\beta, w)p(w|a)dw} \tag{6}$$

The posterior distribution will be Gaussian because the prior distribution and similarity function are both Gaussian. As a result, the distribution's mean and covariance are as follows:

$$\begin{aligned} m &= \beta(\beta XX^T + I'(a))^{-1} xt \\ C &= \beta(\beta XX^T + I'(a))^{-1} \end{aligned} \tag{7}$$

As a result, the probability distribution of the regression objectives for the new feature vector can be computed using the posterior distribution. The predicted distribution in this instance is displayed as follows:

$$\begin{aligned} p\left(\hat{t}|\beta, a, \hat{x}, D\right) &= \int p\left(\hat{t}|\beta, \hat{x}, w\right) p(w|\beta, a, D)dw \\ \hat{t} &= w^T \hat{x} \end{aligned} \tag{8}$$

A Gaussian distribution with mean and variance is used to represent the distribution in Eq. (8), and it determines several values for the new input vector. Regression objectives for class 1 samples $\frac{N}{N_1}$ in the LBDA algorithm are set in and for class 2 samples in where N is the total number of training samples, the number of class 1 samples, and the number of class 2 samples $-\frac{N}{N_2}$ [21]. Thus, the follow-

ing formula is used to determine the probability of class 1 samples:

$$p(\hat{y}=1|\beta,a,\hat{x},D)=\frac{P\left(\hat{t}=\frac{N}{N_1}|\beta,a,\hat{x},D\right)}{P\left(\hat{t}=\frac{N}{N_1}|\beta,a,\hat{x},D\right)+P\left(\hat{t}=-\frac{N}{N_2}|\beta,a,\hat{x},D\right)} \tag{9}$$

Feature selection and classification using the combined MOPSO/LBDA model

In this paper, the gene selection process is carried out using the MOPSO algorithm, and the classification produced by MOPSO is evaluated using the LBDA method. The MOPSO algorithm's working principle is that it comprises of 30 particles, each of which has 70 binary bits. These 70 binary bits specify the positions of each particle. Each bit represents a gene, therefore bit zero denotes the absence of selection for the trait (gene) corresponding to it and bit one denotes the presence of selection for the characteristic (gene) corresponding to it. A particle swarm is depicted in Fig. 4. After each update, particles are assessed. The LBDA classifier determines each particle's appropriateness to assess the quality of the gene set that it has chosen because each particle's position represents a set of genes. The best fit for each particle is known as pbest, while the best fit for the entire group is known as gbest. Until the stopping requirement is satisfied, this process will be repeated. The maximum repetitions or maximum fit rate are two definitions of the halting criterion (Daoud and Mayo 2019). The following relation updates the particle speed in the MOPSO algorithm:

$$\begin{aligned} v_{pd}^{\text{new}} &= w\times v_{pd}^{\text{old}}+c_1\times \text{rand}_1\left(p\text{best}_{pd}-x_{pd}^{\text{old}}\right)\\ &\quad +c_2\times \text{rand}_2\left(g\text{best}_d-x_{pd}^{\text{old}}\right)\end{aligned} \tag{10}$$

The velocity of the particles in the binary MoPSO method is calculated as follows if v_{pd}^{new} is not within the range $[v_{min}, v_{max}]$:

$$v_{pd}^{\text{new}}=\max\left(\min\left(v_{\max},v_{pd}^{\text{new}}\right),v_{\min}\right) \tag{11}$$

We employ the sigmoid transformation function to change the velocity vector into the probability vector as follows:

$$S(v_{pd}^{\text{new}})=\frac{1}{1+\exp(-v_{pd}^{\text{new}})} \tag{12}$$

As a result, using relation (12), the particle positions are likewise changed as follows:

$$X_{pd}^{new}=\begin{cases}1;\ \text{rand}<S(v_{pd}^{\text{new}})\\ 0;\ \text{otherwise}\end{cases} \tag{13}$$

w is the inertia weight, d is the problem's dimension, C_1 and C_2 are acceleration factors, and rund, rund_1 and rund_2 are random values in the range [1 and 0]. In addition, X_{pd}^{new} is the particle's former position, v_{pd}^{new} is the particle's speed in the

MOPSO algorithm

Beginning

The subsequent stages are continued after randomly generating primary particles up until the stopping requirement is satisfied:

1. LBDA classifier evaluation of particle suitability for p=1:N (N number of particles)

2. If (fit rate x_p) > (fit rate $pbest_p$) then

$$pbest_p = x_p$$

3. The end

4. If (fitness level of one of the x_ps) > (fitness level gbest) then

The position of that particle = gbest

6. The end

7. for D=1:d (number of particle dimensions)

$$v_{pd}^{new}=w\times v_{pd}^{old}+c_1\times rand_1\left(pbest_{pd}-x_{pd}^{old}\right)+c_2\times rand_2(gbest_d-x_{pd}^{old}$$

*if v_{pd}^{new} is not within the range $[v_{min}, v_{max}]$: **then***

$$v_{pd}^{new}=max\ (min\left(v_{max},v_{pd}^{new}\right),v_{min})$$

Changing the speed vector into a probability vector using the sigmoid transformation function

$$S\left(v_{pd}^{new}\right)=\frac{1}{1+exp\ (-v_{pd}^{new})}$$

$$if\ (rand<S\left(v_{pd}^{new}\right))\ then\ x_{pd}^{new}=1.$$

Else if $x_{pd}^{new}=0$.

Up until the stop requirement is satisfied, repeat the previous steps.

end

Fig. 4 Representation of multi-objective particle swarm in MOPSO algorithm

Table 2 Values of parameters used in MoPSO algorithm

Parameter values used in MoPSO algorithm	
m: number of particles	30
C_1, C_2: acceleration factors	2
V_{max}: maximum particle velocity	6
V_{min}: minimum particle velocity	− 6
R_1, R_2: random numbers	[0, 1]
N: length of each particle (number of bits)	70
Max iter: maximum iteration	50
w: inertial weight	$w = w_{max} - (((w_{max} - w_{min}) \times \text{iter})/\text{Maxiter})$
W_{max}: maximum inertia weight	0.995
W_{min}: minimum inertia weight	0.5

previous stage, and v_{pd}^{new} is its speed in the new stage. The following succinctly describes the MOPSO algorithm:

Implementation results

On a machine with a 3.4 GHz processor and 1 GHz RAM memory, all phases of implementing the suggested method were completed. Table 2 displays the values of the parameters utilised in the MOPSO algorithm.

Table 3 Quantitative accuracy values obtained from applying the proposed algorithm in 10 times of running the algorithm on four databases

Number of algorithm executions	Databases							
	Leukaemia		Lung cancer		Breast cancer		Prostate cancer	
	ACC (%)	Avg	ACC (%)	Avg	ACC (%)	Avg	ACC (%)	Avg
1	91.54	43	99.71	41	95.12	44	95.42	55
2	93.12	49	99.53	41	94.28	25	96.25	53
3	90.59	49	99.37	51	95.8	42	96.22	46
4	86.35	68	99.65	40	93.19	35	99.3	51
5	90.32	48	97.28	39	95.2	44	97.05	51
6	87.35	46	99.54	41	94.54	30	97.02	41
7	88.34	40	99.54	41	95.1	49	95.18	41
8	88.41	38	99.64	54	95.1	43	95.31	41
9	90.02	66	99.42	41	92.13	35	97.34	58
10	89.19	63	99.87	40	91.21	39	97.29	46
Average	89.26	51	99.51	43	94.36	39	96.71	48

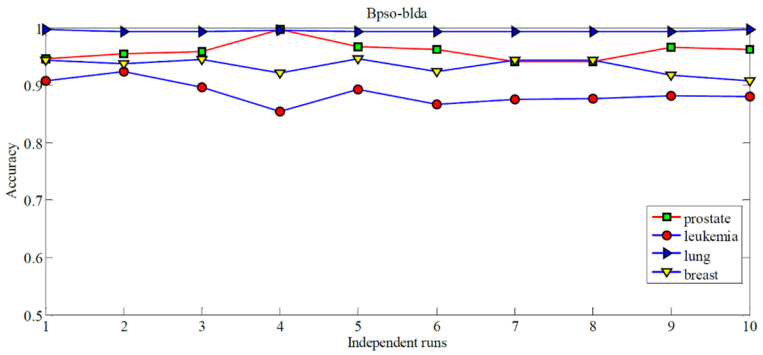

Fig. 5 Drawing the curve of the average classification accuracy of the proposed algorithm in 10 times of its execution on four databases

Table 4 Comparison of the quantitative results of classification accuracy in the proposed algorithm and other methods in the blood cancer database

Gene selection algorithm	Classification	Classification accuracy (%)
MOPSO	52 M	88.6
PSO	ANN	87.2
–	Nero-Fuzzy	87.6
–	KNN	73.4
–	Bayesian	92.1

Table 5 Comparison of the quantitative results of classification accuracy in the proposed algorithm and other methods in the lung cancer database

Gene selection algorithm	Classification	Classification accuracy (%)
MOPSO	LBDA	99.6
PSO	SVM	99.2
IPSO	KNN	97.4
PSO	Ensemble Neural Network	99.98
PSO	ANN	98.6
–	Bayesian	90.04

Table 6 Comparison of the quantitative results of classification accuracy in the proposed algorithm and other methods in the breast cancer database

Gene selection algorithm	Classification	Classification accuracy (%)
MOPSO	LBDA	94.02
PSO	SVM	85.09
GA	SVM	96.15
PSO	Bayesian	75.01
PSO	ANN	95.02
–	Ensemble neural network	92.04

Table 7 Comparison of the quantitative results of classification accuracy in the proposed algorithm and other methods in the prostate cancer database

Gene selection algorithm	Classification	Classification accuracy (%)
MOPSO	LBDA	96.18
IPSO	KNN	93.1
Hybrid PSO/GA	SVM	94.52
EA	KNN	89.02
PSO	ANN	90.31
–	Ensemble neural network	87.24

The accuracy of the suggested methodology as determined by the ten-fold validation method and the average accuracy across four databases are both shown in Table 3. The columns Acc (%) and Avg (N) in this table, respectively, denote the accuracy after 10 iterations of the algorithm and the average number of genes chosen for each iteration. As can be shown, the blood cancer database's maximum classification accuracy is 92.39 when an average of 48 genes are chosen. With a classification accuracy of 47.85, the lowest in this database, 67 genes have been chosen. Similar to this, the average number of genes chosen in the lung, breast, and prostate databases is 39, 43, and 50, respectively, and the greatest classification accuracy is 99.66, 94.63, and 96.68. The average number of selected genes in this situation is 38, 38, and 40, respectively, with the lowest values of classification accuracy in these databases being 96.26, 90.78, and 94.05, respectively. Figure 5 graphically displays the proposed algorithm's average classification accuracy after 10 iterations across four databases.

Table 4 compares the proposed algorithm's classification accuracy in the leukaemia database to those of other approaches found in the references. As can be shown, the suggested approach improves classification accuracy by 2.8, 1.2, and 21.9 percent when compared to PSO + ANN (Lai et al. 2020), Nero-Fuzzy (Alharbi and Vakanski 2023), and KNN (Mostavi et al. 2020) methods. Only the Bayesian classification algorithm is more accurate than the proposed algorithm (Petrini et al. 2022). Similarly, Table 5 compares the performance of the suggested algorithm with those of other techniques in the lung cancer database. According to this table, the suggested algorithm's classification accuracy is 0.4, 2.9, respectively, when compared to PSO + SVM (Petrini et al. 2022), IPSO + KNN (Abd-Elnaby et al. 2021), PSO + ANN, and Bayesian techniques. It gets better by 1.1 and 11.7%. The suggested algorithm's classification accuracy in this database is just somewhat worse than that of the PSO + Ensemble NN combination technique (Debata and Mohapatra. 2022). In the breast cancer database, Table 6 compares the performance of the proposed approach with PSO + SVM, GA + SVM (Yan et al. 2023), and Bayesian methods. This table demonstrates the suggested algorithm's superiority to the PSO + SVM and Bayesian approaches, with the proposed algorithm's rate of improvement being 9.4 and 26.1 percent, respectively. Finally, Table 7 compares the proposed approach to IPSO + KNN, PSO/GA + SVM combined algorithm, and KNN + EA (Sree Devi et al. 2022) when used on a database of prostate cancer cases. This table

illustrates how the proposed algorithm is superior to all other suggested techniques, with the health improvement rates being equal to 4.1, 2.7, and 7.9 percent, respectively.

Two perspectives feature selection and classification method can be used to analyse the reasons why the proposed algorithm is preferable. MOPSO and evolutionary computational methods like the genetic algorithm are quite similar. The foundation of MOPSO is social interaction in biological populations. By exchanging information among their participants and using deterministic and probabilistic principles, MOPSO and evolutionary algorithms are examples of crowd-based search techniques that enhance the search process. Genetic operators such as crossover and mutation operators are not available in MOPSO, nevertheless. Of course, one may compare the intersection operator to the social model of particle interaction. For instance, the mutation parameter in the genetic algorithm is comparable to the *rand1* and *rand2* (Relation 10) parameters that impact the speed of particles. In actuality, the only distinction between them is that, whereas in MOPSO, new particles must be processed in each iteration without any probabilities, crossover and mutation operators in genetic algorithms are probabilistic. The information sharing technique of MOPSO differs greatly from the genetic algorithm. Evolution in the genetic algorithm is accomplished by the use of crossover and mutation operators. Chromosomes exchange information with one another, and the entire population travels collectively in the direction of the target location. This model simulates a single area search in the issue space. Since each particle is uniformly dispersed over the problem space in MOPSO, this model's shortcoming is that it is susceptible to becoming stuck in local optima and only gbest offers information for other particles. It is a one-way sharing system, and only the best solution can evolve. The performance of the MOPSO is affected by the variables w and the acceleration factors c1 and c2. These parameters (Table 2) can be appropriately configured in order to quickly get the required outcomes. The movement of the particles will be extremely slow and time-consuming if the values of these parameters are chosen too tiny, and the algorithm will be compromised and the desired collection of relevant characteristics will not be produced if the values of these parameters are chosen too big. As a result, the MOPSO algorithm can choose the crucial features with ease when the parameters are set properly.

Regarding the classification selection approach, it can be explained that while SVM generally performs well at classifying tumours, it has a significant computational complexity when determining the optimal set of parameters (Sree Devi et al. 2022). However, the LBDA classifier can be quickly and simply built. Tables 3, 4, 5, 6 show that using the MOPSO gene selection approach along with the LBDA classification generally produces superior results.

Figure 6a–d shows how effectively the proposed method can divide gene expression levels into two classes. Examples are the 27 samples connected to the ALL class and the 11 samples related to the AML class in the blood cancer data Fig. 6a, which clearly illustrates this split. Similarly, 44 samples are associated to the samples where the cancer did not recur, whereas 34 samples are related to the samples where the cancer recurred. Figure 6b similarly depicts this division. Similar circumstances can be seen in the data on lung and prostate cancer. In the case of lung cancer, Fig. 6c shows that 16 samples are associated to malignant pleural mesothelioma (MPM) samples and 16 samples are related to Adenocarcinoma (ADCA). In addition, Fig. 6d does a good job of separating the 52 samples linked to tumour samples from the 50 samples connected to non-tumour samples in prostate cancer data. Tables 8, 9, 10 and 11 also display how many functional genes were discovered after using the suggested approach.

Conclusion

In this work, a novel and effective algorithm for selecting and classifying genes was proposed. The programme's performance was assessed using data from four microarray databases and is based on a hybrid model of multi-objective particle swarm optimisation and linear Bayesian discriminant analysis. Using the ten-fold validation method, all the samples were initially split into training samples and test samples. Then, a set of genes were chosen from training samples using Pearson's correlation analysis. The following phase involved selecting a gene set for each particle using the MOPSO method, and then using the LBDA classifier to determine whether or not that gene set was suitable for that particle. The particle with the best match (classification accuracy) after 50 repetitions is regarded as the gene set carrying information. The results of the implementation

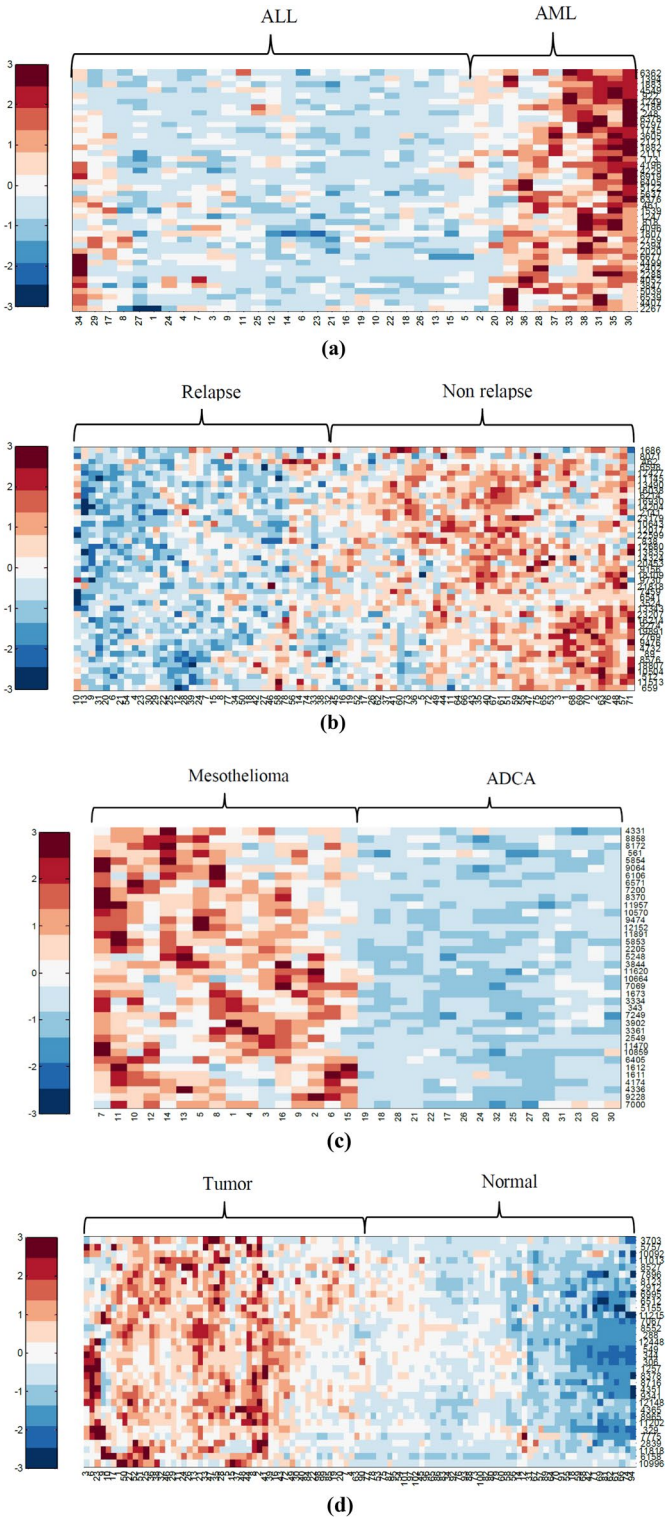

Fig. 6 The set of genes containing selected information by applying the proposed algorithm in the databases: **a** leukaemia; **b** breast cancer; **c** lung cancer and **d** prostate cancer

Table 8 The number of genes effective in the incidence of leukaemia obtained by applying the proposed algorithm

Row	Genes effective in the occurrence of leukaemia								
1	2267	2288	2394	1247	5122	173	1745	1249	6362
2	4407	2402	2759	1539	6405	2111	6797	922	5039
3	6539	4499	1807	461	6919	1882	6378	4549	
4	3847	6677	4096	6376	4229	2121	248	1883	
5	4052	2020	818	5637	4196	3605	2186	1394	

Table 9 The number of genes effective in the incidence of breast cancer obtained by applying the proposed algorithm

Row	Genes effective in the occurrence of breast cancer								
1	659	8576	19,891	3524	18,109	12,680	23,179	16,038	462
2	11,513	89	9274	6541	9156	838	2141	13,490	8071
3	512	4732	16,214	7459	20,453	22,599	14,204	11,145	1686
4	10,504	9476	23,207	21,818	14,324	12,017	16,930	12,427	
5	18,807	2769	13,343	9730	13,835	10,643	6214	6598	

Table 10 The number of genes effective in the occurrence of lung cancer obtained by applying the proposed algorithm

Row	Genes effective in the occurrence of breast cancer							
1	7000	1612	3361	1673	5248	9474	6571	8172
2	9228	6405	3902	7069	2205	10,570	6106	8858
3	4336	10,859	7249	10,664	5853	11,957	9064	4331
4	4174	11,470	343	11,620	11,891	8370	5854	
5	1611	2549	3334	3844	12,152	7200	561	

Table 11 The number of genes effective in the occurrence of prostate cancer obtained by applying the proposed algorithm

Row	Genes effective in the occurrence of breast cancer						
1	10,996	329	9341	306	8552	6995	11,013
2	6158	11,202	4351	344	7067	2912	10,092
3	11,818	8965	8716	549	11,215	8123	5757
4	2839	4365	8378	12,448	5155	7896	3703
5	7775	12,148	1257	288	6512	8527	

demonstrated that the suggested technique causes the dimension of the microarray data to be reduced. The MOPSO model is used to classify data, which also improves classification accuracy. This is possible because the MOPSO model eliminates redundancy and reduces the number of redundant and excess genes by taking into account how genes are correlated with one another. The suggested algorithm, meanwhile, occasionally performs worse than alternative algorithms. For instance, the classification accuracy for lung cancer is 100% for the PSO-Ensemble Neural Network approach and 99.5% for the suggested technique. This is because using a set of classifiers yields better results than using a single classifier. Future research will focus on fusing the suggested algorithm with more sophisticated ones in an effort to boost classification rates.

Author contributions All the authors contributed to the study conception and design. Data collection, simulation and analysis were performed by MRR, DM, MFA and SS. The first draft of the manuscript was written by SAM and SMA, and all the authors commented on previous versions of the manuscript.

Funding This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of data and materials It is not possible to share the data of this research publicly because the people who participated in this research for evaluation and test results do not consent to publication. Therefore, the data will be shared only if the respected editor of the journal or the reviewers request the data.

Declarations

Conflict of interest We certify that there is no actual or potential conflict of interest in relation to this manuscript.

References

Abd-Elnaby M, Alfonse M, Roushdy M (2021) Classification of breast cancer using microarray gene expression data: a survey. J Biomed Inform 117:103764

Alharbi F, Vakanski A (2023) Machine learning methods for cancer classification using gene expression data: a review. Bioengineering 10(2):173

Almugren N, Alshamlan H (2019) A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. IEEE Access 7:78533–78548

Ayyad SM, Saleh AI, Labib LM (2019) Gene expression cancer classification using modified K-Nearest Neighbors technique. Biosystems 176:41–51

Basavegowda HS, Dagnew G (2020) Deep learning approach for microarray cancer data classification. CAAI Trans Intell Technol 5(1):22–33

Chen C, Jianhua W, Devin K, Zilong Z, Feifei C, Da Z, Mulin JL, Quan Z (2022) webTWAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study. Nucleic Acids Res 50(D1):D1123–D1130

Daoud M, Mayo M (2019) A survey of neural network-based cancer prediction models from microarray data. Artif Intell Med 97:204–214

Debata PP, Mohapatra P (2022) Identification of significant biomarkers from high-dimensional cancerous data employing a modified multi-objective meta-heuristic algorithm. J King Saud Univ-Comput Inf Sci 34(8):4743–4755

Fabin C, Liang H, Niu B, Zhao N, Zhao X (2023s) Adaptive neural self-triggered bipartite secure control for nonlinear MASs subject to DoS attacks. Inf Sci 631:256–270

Ghosh M, Adhikary S, Ghosh KK, Sardar A, Begum S, Sarkar R (2019) Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods. Med Biol Eng Comput 57:159–176

Guillen P, Ebalunode J (2016) Cancer classification based on microarray gene expression data using deep learning. In: 2016 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE, pp 1403–1405

Guo Y, Liu S, Li Z, Shang X (2018) BCDForest: a boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data. BMC Bioinform 19(5):1–13

Gupta S, Gupta MK, Shabaz M, Sharma A (2022) Deep learning techniques for cancer classification using microarray gene expression data. Front Physiol 13:952709

Haoyan Z, Xudong Z, Liang Z, Ben N, Guangdeng Z, Ning X (2022) Observer-based adaptive fuzzy hierarchical sliding mode control of uncertain under-actuated switched nonlinear systems with input quantization. Int J Robust Nonlinear Control 32(14):8163–8185

Haoyu Z, Quan Z, Ying J, Chenggang S, Dong C (2022) Distance-based support vector machine to predict DNA N6-methyladine modification. Curr Bioinform. 17(5):473–482

Khezri E, Zeinali E (2021) A review on highway routing protocols in vehicular ad hoc networks. SN Comput Sci 2:1–22

Khezri E, Zeinali E, Sargolzaey H (2022) A novel highway routing protocol in vehicular ad hoc networks using VMaSC-LTE and DBA-MAC protocols. Wirel Commun Mob Comput. https://doi.org/10.1155/2022/1680507

Khezri E, Zeinali E, Sargolzaey H (2023) SGHRP: secure greedy highway routing protocol with authentication and increased privacy in vehicular ad hoc networks. PLoS ONE 18(4):e0282031

Lai YH, Chen WN, Hsu TC, Lin C, Tsao Y, Wu S (2020) Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. Sci Rep 10(1):4679

Lei XP, Li Z, Zhong YH, Li SP, Chen JC, Ke YY, Lv A, Huang LJ, Pan QR, Zhao LX, Yang XY, Chen ZS, Deng QD, Yu XY (2022) Gli 1 promotes epithelial-mesenchymal transition and metastasis of non-small cell lung carcinoma by regulating snail transcriptional activity and stability. Acta Pharm Sin B 12(10):3877–3890

Liu B, Tian M, Zhang C, Li X (2015) Discrete biogeography based optimization for feature selection in molecular signatures. Mol Inf 34(4):197–215

Mokhlesi DG, Khorami E, Boukani B, Trik M (2020) Improve replica placement in content distribution networks with hybrid technique. J Adv Comput Res 11(1):87–99

Mostavi M, Chiu YC, Huang Y, Chen Y (2020) Convolutional neural network models for cancer type prediction based on gene expression. BMC Med Genom 13:1–13

Petrini I, Cecchini RL, Mascaró M, Ponzoni I, Carballido JA (2022) Statistical learning analysis of thyroid cancer microarray data. International work-conference on bioinformatics and biomedical engineering. Springer International Publishing, Cham, pp 90–102

Rezaei M, Rahmani E, Khouzani SJ, Rahmannia M, Ghadirzadeh E, Bashghareh P, Taheri F (2023) Role of artificial intelligence in the diagnosis and treatment of diseases. Kindle 3(1):1–160

Samiei M, Hassani A, Sarspy S, Komari IE, Trik M, Hassanpour F (2023) Classification of skin cancer stages using a AHP fuzzy technique within the context of big data healthcare. J Cancer Res Clin Oncol. https://doi.org/10.1007/s00432-023-04815-x

Sarhan AM (2009) Cancer classification based on microarray gene expression data using DCT and ANN. J Theor Appl Inf Technol 6(2):208–216

Sharma A, Rani R (2019) C-HMOSHSSA: gene selection for cancer classification using multi-objective meta-heuristic and machine learning methods. Comput Methods Progr Biomed 178:219–235

Shayan S, Jamaran S, Askandar RH, Rahimi A, Elahi A, Farshadfar C, Ardalan N (2021) The SARS-Cov-2 Proliferation blocked by a novel and potent main protease inhibitor via computer-aided drug design. Iran J Pharm Res 20(3):399

Shekar BH, Dagnew G (2019) Grid search-based hyperparameter tuning and classification of microarray cancer data. In: 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), IEEE, pp 1–8

Singh BK, Verma K, Thoke AS (2016) Fuzzy cluster based neural network classifier for classifying breast tumors in ultrasound images. Expert Syst Appl 66:114–123

Sree Devi KD, Karthikeyan P, Moorthy U, Deeba K, Maheshwari V, Allayear SM (2022) Tumor detection on microarray data using grey wolf optimization with gain information. Math Probl Eng. https://doi.org/10.1155/2022/4092404

Sun L, Zhang X, Qian Y, Xu J, Zhang S (2019) Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification. Inf Sci 502:18–41

Sun J, Zhang Y, Trik M (2022) PBPHS: a profile-based predictive handover strategy for 5G networks. Cybern Syst 53(6):1–22

Taheri MS, Ghomi Z, Mirshahi R, Moradpour M, Niroomand M, Yarmohamadi P, Zeidabadi H (2023) Usefulness of subtraction images for accurate diagnosis of pituitary microadenomas in dynamic contrast-enhanced magnetic resonance imaging. Acta Radiol 64(3):1148–1154

Trik M, Molk AMNG, Ghasemi F, Pouryeganeh P (2022) A hybrid selection strategy based on traffic analysis for improving performance in networks on chip. J Sens. https://doi.org/10.1155/2022/3112170

Trik M, Akhavan H, Bidgoli AM, Molk AMNG, Vashani H, Mozaffari SP (2023) A new adaptive selection strategy for reducing latency in networks on chip. Integration 89:9–24

Venkatesan C, Balamurugan D, Thamaraimanalan T, Ramkumar M (2022) Efficient machine learning technique for tumor

classification based on gene expression data. Int Conf Adv Comput Commun Syst (ICACCS) 1:1982–1986

Wang J, Jiang X, Zhao L, Zuo S, Chen X, Zhang L, Lin Z, Zhao X, Qin Y, Zhou X, Yu XY (2020) Lineage reprogramming of fibroblasts into induced cardiac progenitor cells by CRISPR/Cas9-based transcriptional activators. Acta Pharm Sin B 10:313–326. https://doi.org/10.1016/j.apsb.2019.09.003

Yan C, Ben N, Xudong Z, Guangdeng Z, Ahmad A (2023) Event-triggered adaptive decentralized control of interconnected nonlinear systems with Bouc-Wen hysteresis input. Int J Syst Sci. https://doi.org/10.1080/00207721.2023.2169845

Yanwei Z, Ben N, Guangdeng Z, Ning X, Ahmad AM (2023a) Event-triggered optimal decentralized control for stochastic interconnected nonlinear systems via adaptive dynamic programming. Neurocomputing. https://doi.org/10.1016/j.neucom.2023.03.024

Yanwei Z, Ben N, Guangdeng Z, Xudong Z, Khalid HA (2023b) Neural network-based adaptive optimal containment control for non-affine nonlinear multi-agent systems within an identifier-actor-critic framework. J Franklin Inst 360(12):8118–8143

Zeinali-Rafsanjani B, Alavi A, Lotfi M, Haseli S, Saeedi-Moghadam M, Moradpour M (2023) Is it necessary to define new diagnostic reference levels during pandemics like the Covid19-? Radiat Phys Chem 205:110739

Zhang L, Deng S, Zhang Y, Peng Q, Li H, Wang P, Fu X, Lei X, Qin A, Yu XY (2020) Homotypic targeting delivery of siRNA with artificial cancer cells. Adv Healthc Mater 9(9):e1900772