Research on Biomedical Engineering (2023) 39:573–586
https://doi.org/10.1007/s42600-023-00289-5

ORIGINAL ARTICLE

Development of novel methodology for gene identification-based classification of leukaemia disorder

J. Briso Becky Bell[1] · Ananth Rajkumar[2] · S. Maria Celestin Vigila[3] · M. Gerald Arul Selvan[2] · J. S. Binoj[4]

Received: 4 December 2022 / Accepted: 28 May 2023 / Published online: 19 June 2023
© The Author(s), under exclusive licence to The Brazilian Society of Biomedical Engineering 2023

Abstract

Purpose Many classifier approaches and algorithms are developed in recent days to handle large dimensionality problems in biomedical data mining and machine learning. The research focus to identify the optimum search methodology to predict the gene samples for leukaemia.

Methods Different search classifier such *T*-test, Principal Component Analysis (PCA) and Genetic Algorithm (GA) are considered in this study and the search results of each classifiers are analysed. The classifiers are used to filter the top important and sort the mutually exclusive illness samples. The classifiers *T*-test and PCA are blended with Linear Discriminant Analysis (LDA), Self-Organizing Map (SOM) and Random Optimized Search (ROS) to predict the performance of the coupled classifiers.

Results The confusion matrix is employed to calculate accuracy and compare the considered classifiers' performance and accuracy. GA classifiers show a better performance than the other classifier-based feature selection algorithms with substantially unique gene characteristics. The mean, best and average generations of GA are considered to determine the accuracy of the generations.

Conclusion The ROS-based LDA classifier improves the classification results and GA enhances the gene retrieval. The performance analyses of the different generations of GA are examined using the confusion matrix and the most optimal classifier is identified as GA-Avg-120G.

Keywords Bio-informatics · Classifiers · Optimized search · Genomic pattern recognition · Evolutionary learning

✉ J. S. Binoj
binojlaxman@gmail.com

J. Briso Becky Bell
brisobell@gmail.com

Ananth Rajkumar
ananth@sxcce.edu.in

S. Maria Celestin Vigila
celesleon@yahoo.com

M. Gerald Arul Selvan
gerald@sxcce.edu.in

[1] Department of Information Technology, Kings Engineering College, Irungattukottai 602117, India

[2] Department of Mechanical Engineering, St. Xaviers Catholic College of Engineering, Chunkankadai 629003, India

[3] Department of Information Technology, Noorul Islam Centre for Higher Education, Kumaracoil 629180, India

[4] Institute of Mechanical Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai 602105, Tamil Nadu, India

Introduction

In the last decades, the analysis of genomic data is widely carried out to acquire more information about genic variants. Genomic pattern recognition can be used to distinguish samples from patients having two mutually exclusive diseases with common genomic features. Gene expression data are generated using gene microarray mass spectrometry technology (Goswami et al. 2009) that provides more advancement in clinical diagnostics tests for every gene-based disorder. The goal of feature selection is to select a set of genes that can be used to differentiate between Acute Lymphoblast Leukaemia (ALL) and Acute Myeloid Leukaemia (AML) patients (Furong et al. 2020).

Recently, due to technological development more advancement is made in bio-informatics. Researchers use data banks to extract gene features from multi-dimensional gene and protein database. The most popular algorithms used for efficient retrieval of important gene are based on

sequential feature selection approach, because cross-validation is used for evaluating the performance of the selected features. In order to access datasets of huge gene count with small sample quantity, it is important to reduce either gene count or dimensionality of features using machine learning capability. As higher gene count is not necessary for obtaining a desired learning result and much lower gene count can lead to misclassification in the learning algorithm. The reduction of genes is important because it increase learnability and also reduces the computational time (Sun et al. 2009). The two main processes involved in the reduction of gene features are feature selection method and feature transformation algorithm. The feature selection methods as described by Tsanas et al. (2010) are used to reduce and select a set of features from the original data, and the feature transformation algorithms given by Bakshi (1998) transform the high-dimensional feature space into reduced dimensional space. The gene selection approach reduces the gene set by extracting a small set of vital genes which can perform at best efficiency in classification between binary category small sample datasets. Feature selection methods are roughly classified into two groups: filter approach and optimal subset search approach. Filter approach evaluates and selects the gene subsets by finding general characteristics of the data without the involvement of the chosen learning algorithm. The optimal subset search can be either a wrapper-based or an embedded-based optimal search approach. A wrapper approach uses the chosen learning methods for evaluating each of the candidate feature subset by assessing the performance of learner. Embedded approach search for features better fit for the chosen learning method, as the learning method takes a long time to run.

In applications of disorder versus normal observation, mutually linked diseases, classification plays a vital role for judging the predictive samples to the high-confidence class by learning the major distinct gene features. If the misclassification error cost and class distribution are known for samples, one can easily compute the correct threshold. However, the misclassification error costs are difficult to assess even by the human experts in the field, as there are high-dimensional redundant gene features. So there is a need to minimize the dimensionality of features using feature selection approach to classify the diseases and for effective identification of clinical diagnosis.

Many research studies are done lately mainly in the feature selection of leukaemia, Dynamic Weight LogitBoost (DWLB), an ensemble-based approach developed by Subash Chandra Bose et al. (2021), for selecting features in leukaemia data tumour detection. Whereas the feature selected are highly redundant. Cucoo Search (CS) by Sampathkumar et al. (2020) is a hybrid methodology using crossover-based search for detecting significant features in leukaemia data, as due to the usage of multiple search algorithms it has one class learning problem at classification. A Particle Swarm Optimization (PSO)-based feature optimization done by Srisukkham et al. (2017) has used a PSO methodology to select the significant features in leukaemia data, but it has long learning time and search time. The large margin hybrid algorithm is a feature selection method developed by Zhang et al. (2019); the approach uses *K*-Nearest Neighbour (K-NN), Partial Least Square Discriminant Analysis (PLS-DA) and Least Square Support Vector Machines (LS-SVM) as classifier wrappers to perform high-degree feature section in various datasets, as wrappers are usually prone to high testing delay.

Most of the current feature selection algorithms can be assessed by the performance of classifier model specified by Mundra and Rajapakse (2009) and Kumar and Choudhary (2012). The accuracy of the training samples data is not an actual estimate for a model's performance of independent dataset, as re-substituting the training samples may be typically over-optimistic. In order to obtain the performance of a selected model, the accuracy should be tested with another independent dataset which has not been used to build the actual model. The sample set data are partitioned into equal halves of training samples and another equal half of test samples; each of the test set and the training samples must roughly have the same categorical proportions in the partitions (Ab Hamid et al. 2021). The present feature selection algorithms that select features by ranking (*T*-statistic, Pearson Correlation Coefficient (PCC), Signal to Noise Ratio and the *F*-statistic) given by Hancer et al. (2018) show individual features from the total list of features; thereby, it creates redundant information, as it will not consider interaction between features and certainly not all the features are needed for effective classification tasks (Chang and Moura 2010). These types of feature selection methods are generally used as a pre-processing step as they are quick to retrieve features and most of the advanced feature selection methods (*K*-means, Principal Component Analysis (PCA), Single Value Decomposition (SVD)) (Japkowicz 2001; Mitchell and Mitchell 1997) are used to improve the classifier's performance (Bishop and Nasrabadi 2006). The sequential selection techniques select a set of features by removing or adding features until it reaches certain stopping criteria.

Random Optimized Search (ROS) uses forward sequential feature selection embedded within a linear classifier in order to select the important features (Alpaydin 2020). The ideal aim of classification is to minimize the mislabelling error; the feature selection method does a sequential search using the misclassification error of the learning methods. The learning methods act on every candidate feature subset, as the accuracy for each subset is observed; the training set is used for selecting the features thus to fit the classifier model and test set is applied on learned classifier to evaluate the performance. The highest performance-based

candidate feature subset is selected as the finally selected features. In each feature selection method, in order to estimate and compare the accuracy of each candidate feature subset, a confusion matrix is set to calculate misclassification error. By using misclassification error of the testing, training can be improved with other feature subset, so the prediction accuracy of the testing set can be gradually learned.

However, ranking the features in binary classes can result in skewed learning, which reduces classification accuracy. As feature selection selects highly redundant genes with significant genes, which leads to one class learning problem thus reducing the classification performance of classifier. With this feature selection technique more feature subsets are gained for to be trained based on optimal classification rate, thus obtaining high classification rates by reducing the learning of redundant features. In the evolutionary learning approaches given by Hernandez Hernandez et al. (2007) and Gunavathi and Premalatha (2014), Genetic Algorithm (GA) is used to perform classification at various generations for various features and perfect features are identified only at optimized iterative classifications at each consecutive generation.

In this paper, novel methodologies are developed to predict the most accurate methodology to identify features in a high-dimensional gene data expression. Then by using the identified features based on reduced gene data expression, learning is done by the classifier, so one can classify data samples into either of the two categories such as ALL sample or AML sample. By analysing the performance of classifier, it shows that the developed methodology has the efficiency of classification for subgroups in leukaemia disorder (Bell and Vigila 2018).

The paper begins with the introduction of the feature selection approach to classification problems and various research studies in this area. The "Methods" section portrays the developed methodologies used to select the features and classify the samples. The results obtained for the developed methodologies and the list of experimental data is discussed in the "Results" section. The conclusion is provided at the "Conclusions" section at the end of the paper based on the observation of the obtained results.

Methods

In this research study three optimal subset search approaches are developed for gene identification-based sample classification of leukaemia disorder; the schematic representation of the developed methodologies are shown in Fig. 1. The first approach is the feature ranking-based approach; in this approach, *T*-test is used to select the subset of features in a high-dimensional gene data expression, so that the reduced gene expression data are used for classification of disease samples, and three types of classifiers are used for classification. The classifier Linear Discriminant Analysis (LDA) is used as a linear classifier to assess the performance of reduced gene expression. In the ROS-based LDA classifier *k*-fold cross-validation is used for classifying the reduced gene expression. And the Self-Organizing Map (SOM) classifier classifies sample by forming a neural network-based

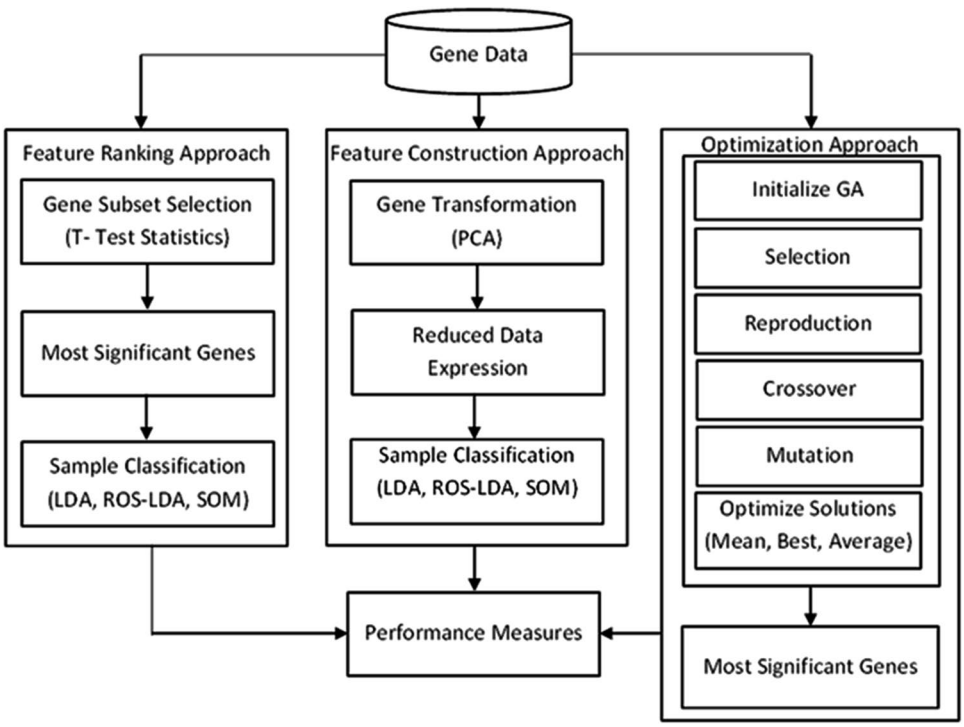

Fig. 1 Gene selection with optimal subset search-based learning

training and testing method in order to classify the reduced data (Tuv et al. 2009). The second approach is the feature transformation-based approach, in which the PCA is used to transform features for reducing the gene expression data and for classification; three classifiers similar to the previous approach are used to find the classifier accuracy of each learner. The last approach is an optimization approach; here the high-dimensional gene data are reduced using optimization-based selection. In this approach, GA is used to retrieve the most significant genes and the operations such as genetic selection, crossover, reproduction and mutation are performed at each generation. At first, GA performs classification of categorical samples, by obtaining the performances of each classification for various disjoint feature subsets and accuracy is verified for gaining higher classification rate in each generation, thus selecting the feature subset producing the higher classification rate as the top significant genes. The approach uses hybrid classifier for learning to give high significant features without redundancy. Three different levels of learning average are employed in this approach, to learn the high-dimensional data by predicting the mean and best at different generations. After various classifications the classification rates for each of the classifiers are obtained. Finally, performance metrics such as accuracy and error rate are employed to evaluate the classifier training, validation and test performances; the mentioned three optimal subset search approaches have nine different flow methodologies to obtain highly significant features and high-performance accuracy. The proposed three search methodologies are studied and compared to predict the most efficient method with high performance that can be more suitable for classification of leukaemia disorder. The main functional components used in the proposed methodologies are discussed below.

T-test

T-test is a sequential feature selection method used to select the subsets of features. The gene expression leukaemia data values (ALL, AML) are compared using two-sample *T*-test (Guyon and Elisseeff 2003), to evaluate differentially expressed genes in binary categorical dataset. *T*-test is performed in each gene for identifying significant changes in expression values between the two categories of samples. By conducting a single-category *T*-test on every gene, the statistical significant intensity for each gene is identified; the gene having the highest statistical significant intensity is selected as the significantly expressed gene. The formula to determine the test significance is shown in Eq. 1.

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \tag{1}$$

where t is the test significance value, $\bar{x}$ and $\bar{y}$ are the mean of samples and s_x and s_y are the respective standard deviations of samples for two categories of data, in which n and m are the sizes of samples on each category.

Principal component analysis

PCA is used to transform the features as clusters, so that the clusters can be accessed by linear classifier algorithms. In high-dimensional datasets with multiple features, similar features are grouped as a cluster. The primary principle of PCA is to identify features with similar characteristic and group them as a cluster; the generated clusters are called as the principal component (PCs). By taking redundancy of features as an advantage, one can make the problem smooth by transforming a set of feature variables with a single constructed feature (Ilin and Raiko 2010).

In this feature transformation-based aspect, the complete set of PCs is found. But in common whilst adding first few PC variances it may exceed 80% of the total variance of the actual data.

Thus, the PC scores are the representation of X in the PC space, and similarly, the PC variances are usually the eigenvalues of the covariance matrix of X.

$$wcoeff_1^T = \frac{1}{N}\sum_{n=1}^{N}\left\{(u_1^T X_n - u_1^T \sum_{i=1}^{n}\frac{X_i}{N}\right\}^2 \tag{2}$$

where X is the data matrix, u is the variable weight which is a vector of length N with each element only positive and the inverse of sample variance is the variable weight. The *wcoeff* is the observed weight coefficient found using Eq. 2. In the first output, *wcoeff* has the coefficients of first PC. The PC data can be processed via a classifier learning algorithm for finding its performance. One of the main advantages of PCA was dimensionality reduction, where the primarily formed new principal features consist 80% of the total variance, of the actual data. So, the classifier can perform faster and the performance of the classifier could be improved effectively.

Linear discriminant analysis

LDA is a linear classifier which assesses the training data with various classes on the basis of diverse Gaussian distributions. In order to train a learner, the fitting function calculates the Gaussian distribution in every class using certain parameters, for predicting the classes of new test data by classification scheme. For creating a classifier or learner using LDA (Hu et al. 2010), one has to consider a gene expression data having each class y and expression data x; learning can be done by finding mean and standard deviation as given by Eqs. 3 and 4, respectively.

$$\hat{\mu}_k = \frac{\sum_{n=1}^{N} M_{nk} x_n}{\sum_{n=1}^{N} M_{nk}} \tag{3}$$

$$\hat{\Sigma} = \frac{\sum_{n=1}^{N} \sum_{k=1}^{K} M_{nk} \left(x_n - \hat{\mu}_k\right) \left(x_n - \hat{\mu}_k\right)^T}{N - K} \tag{4}$$

where $\hat{\mu}$ is the calculated mean of data expression x_n, $\hat{\Sigma}$ is the standard deviation of data expression x_n, M is an n row k column class membership matrix and if sample n is not from class k, the value of M_{nk} is 0; else, the value of M_{nk} is 1.

Random optimized search by LDA

Feature selection can be followed sequentially by classification and this kind of feature extraction is usually known as a wrapper-based selection approach (John and Kohavi 1997). It generates random subsets of gene features and evaluates the quality independently by learning on a classifier. So, it uses a pool of most optimal gene features in order to analyse the strength of classifier. It allows searching of a subset of features having high accuracy-based performance using LDA classifier over-randomized subsets of gene features. ROS incorporates additionally stratified k-fold cross-validation in order to validate testing performances. This partitions the observations as a training set and a test dataset. Then it classifies so as to minimize the expected misclassification cost using Eq. 5.

$$\hat{y} = \arg \min_{y=1,\ldots K} \sum_{k=1}^{K} \hat{\hat{P}}(k|x) C(y|k) \tag{5}$$

where k is a scalar. When $0 < k < 1$, it randomly selects nearly $k \times n$ observations for the test set. The k is set to value of 1/10 by default, $\hat{y}$ is the predicted class of test samples and K is the categorical class count. The $\hat{\hat{P}}(k|x)$ is the posterior probability of class k for observation x and $C(y|k)$ is the cost of mislabelling a sample as y when its real class is k. Thus, it randomly partitions samples that are set onto a training set and a test set with stratification, by using the class category detail in various group, and it has roughly the same class proportions for both training and test sets. Here, the misclassification cost is found for various feature subsets; the feature subset which has the minimal misclassification cost is taken as most significant feature. Whilst using LDA iterative classifier by random optimal search principle, the maximum performance accuracy can be obtained.

SOM

SOM is a statistical learning tool (Ron and George 1997), which uses neural network-based algorithm and using a SOM to cluster the data. The condition for training a network is easy for static network that is for network without feedback. The standard SOM uses back propagation algorithm for training the network on the basis of minimizing of an energy function by finding the instantaneous error as in Eq. 6.

$$E(m) = \frac{1}{2} \sum_{q=1}^{q} \left[d_q - y_q\right]^2 \tag{6}$$

where E is the error function, y_q is the target output of the neural network and d_q represents the predicted output for the qth input pattern. Each weight is changed as per the gradient descent update rule.

$$\nabla W_{ij} = -K \frac{dE}{dW_{ij}} \tag{7}$$

where K is a constant of proportionality and W_{ij} are the weights of the connection between neuron i and neuron j; the weights of the connection are calculated using Eq. 7. The weight adjustment procedure is continuously repeated until the variation between the actual output and predicted output is closer to some acceptable tolerance level. The Feed Forward Neural Network (FFNN) is used to train and classify the samples as per the categorical classes. The two-layer FFNN is shown in Fig. 2, with a sigmoid transfer function in the hidden layer, a transfer function in the output layer. The hidden layer uses the input data and finds the weighted input w. It spreads the initial weight w across the input space (Horn et al. 2009); firstly, there are some distance b from the training vectors to input vectors, where hidden neurons are set to 10, so that a hidden layer competitive network is formed. The number of output neurons is set to 2, which is equal to the number of categories. By using the constructed network, the test data is evaluated into positive or negative class. The classified samples are evaluated for accuracy performance measures.

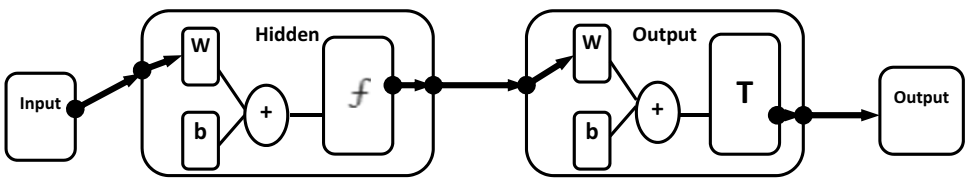

Fig. 2 SOM classifier using Feed Forward Neural Network

Genetic algorithm

GA is an optimization-based embedded classifier with linear minimization fitness function. It finds the local unconstrained minimum for gene expression data by using the objective function with n number of features as the designated dimension of fitness function, as it creates the initial population for GA. Here each row of the population matrix is a random sample of row indices of the gene features (Gunavathi and Premalatha 2014). A population option is used to set the population size, so that the GA uses population size to count the number of individuals there in every generation. When the population size is excessively high, the GA searches for solution space more exhaustively, so that the chance to find the local minimum by the GA is reduced. The fitness function accepts length of the number of features and it returns a scalar value at each run. GA accepts a population by number of features as expression vector.

$$\min_x f(x) = 100(x_1^2 - x_2)^2 + (1 - x_1)^2 \tag{8}$$

where x is the gene data expression vector and the objective function $f(x)$ is a minimization function holding a linear classifier as shown in Eq. 8. Each row of input population vector produces an objective function value, and the error rate of current classification is found and compared with the actual categorical values at each generation. The objective function values are operated by GA optional operation such as mutation, reproduction, selection and crossover with another gene vector. A selection option specifies how parent population vector is chosen by GA for the next generation. In mutation option the input population vector values are changed to be considered for next-generation parent population. In reproduction option, two parent populations are concatenated to form a reproduced child population for next generation. In crossover option, part data vector values of two parent populations are taken to form a new population for next generation. Whilst these population vectors taken as input for fitness functions until convergence of the resulting classification reaches minimal error, thereby iterating every generation till the end of generation.

Performance measures

The performance measures are used to calculate the accuracy for each of the feature-based classifier methods. In order to measure the performance of the classifier using the classification data the confusion matrix is constructed. The confusion matrix given in Fig. 3 shows various output cases, such as the True Positives (*TP*) that indicate the total predicted category 1 class samples to the total learned test samples at classification. The True Negatives (*TN*) predict the total predicted category 2 class samples to the total learned test samples at classification. The False Positives (*FP*) are the total wrongly predicted category 2 class samples as category 1 class samples at classification and the False Negatives (*FN*) are the total wrongly predicted category 1 class samples as category 2 class samples at classification. It also shows the performance of *T*-test features on SOM classifier for leukaemia dataset-based classification. In the confusion matrix measure the error rate (*ERR*) and the accuracy (*ACC*) are calculated using the formulas given by Eqs. 11 and 12. The other formulas used are given in Table 1, where the True Positive Rate (*TPR*) is the number of predicted True Positives to the total number of True Positive and False Negatives. The True Negative Rate (*TNR*) is the number of predicted True Negative to the total number of True Positives and False Negatives. The Positive Predictive Value (*PPV*) is the number of predicted True Positives to the total number of True Positives and False Positives. The False Positive Rate (*FPR*) is the number of predicted False Positive to the total number of False Positives and True Negative.

Error rate or misclassification cost is calculated for all the classifiers using Eq. 13. Another measure to test the

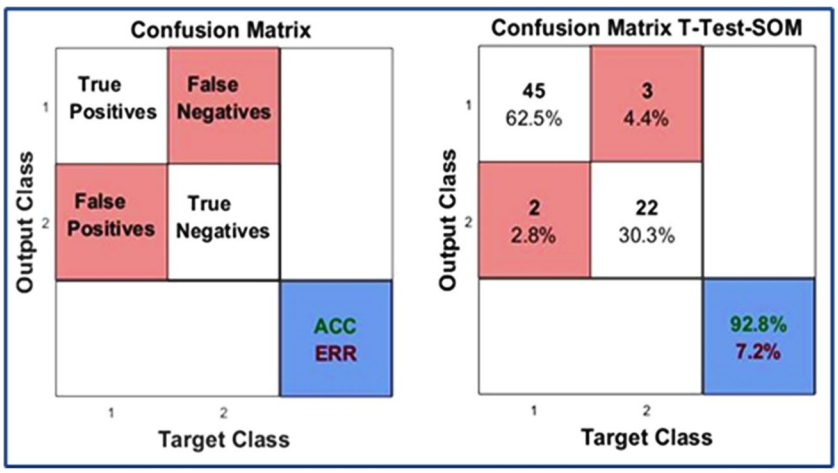

Fig. 3 Confusion matrix

Table 1 Performance measures

Measures	Equations
Sensitivity/Recall/True Positive Rate/$(1-\beta)$ / $(1-\gamma)$	$TPR = TP/(TP + FN)$
Specificity/True Negative Rate	$TNR = TN/(FP + TN)$
Precision/Positive Predictive Value / $(1-\Omega)$	$PPV = TP/(TP + FP)$
False Positive Rate/Fallout/$(1-\alpha)$	$FPR = FP/(FP + TN)$
Prevalence	$PRE = (\sqrt{(TPR(-TNR + 1)) + TNR - 1)/((TPR + TNR - 1))}$
Receiver Operating Characteristics	$ROC = \sum_i \{\left[\left(TPR_i\right).\left(\alpha_i - \alpha_{i-1}\right)\right] + 1/2[TPR_i - \left(1 - \beta_{i-1}\right).\left(\alpha_i - \alpha_{i-1}\right)]\}$(9)
Precision Recall Characteristics	$PRC = \sum_i \{\left[\left(PPV_i\right).\left(\gamma_i - \gamma_{i-1}\right)\right] + 1/2[PPV_i - \left(1 - \Omega_{i-1}\right).\left(\gamma_i - \gamma_{i-1}\right)]\}$ (10)

classifiers for fitting data is the Receiver Operating Characteristic (*ROC*) (Fawcett 2006); *ROC* plot is plotted using Eq. 9, the *TPR*s and *FPR*s based on thresholding of outputs varying from 0 to 1. In this plot, only fewer *FPR* is accepted, for taking the farther left and up to the line of reach, thereby getting a high *TPR*. For a good classification, the line goes from the bottom left corner to the top left corner, and then to the top right corner or close to that. In Precision Recall Characteristic (*PRC*) the positive predictive values and sensitivity values are taken to plot the curve using Eq. 10. The top left corner to top right corner and to bottom right corner shows best predictive region. The area under *ROC* and *PRC* are used to visualize the performance of classifiers.

$$CRR = \frac{(TP + TN)}{(P + N)} \tag{11}$$

$$ACC = [\{(TPR)(PRE)\} + \{(TNR)(1 - PRE)\}] \tag{12}$$

$$ERR = 1 - \frac{(TP + TN)}{(P + N)} \tag{13}$$

Results

The dataset is retrieved from the NCBI disease database from which high-resolution leukaemia datasets are generated using the microarray (Sayers et al. 2012). Each sample set includes 25 AML and 47 ALL leukaemia disease data. The pre-processed dataset of leukaemia data requires certain pre-processing steps for improving the accuracy of selection. The various approaches used in this research study are discussed below.

The feature ranking approach is a filtering approach for feature selection (Cheng et al. 2013). *T*-test finds the index of most significant gene feature; by this method 7129 genes are ranked by considering the absolute value of the test statistic *t*-score. The various ALL group average denoted by red line circle head and AML group average denoted by blue colour with triangle head for each gene is displayed. The *T*-test score for each gene features is shaded with bright orange colour and significant gene is marked in green asterisk symbol as shown in Fig. 4. The top 10 ranked features, extracted using highest *t*-score, are identified as the significant features. The identified top 10 genes with its corresponding DB ID, gene description, *T*-score, and gene ID are shown in Table 2.

The expression and class categorical samples are divided into 36 training sets, 18 validation sets, and 18 test sets. The sample size for classification is shown in Table 3. During the learning *T*-test-based features in LDA classifier, original ALL and AML samples are given as green and yellow colours, respectively; also, magenta and sky-blue colours denote classified ALL and AML samples. The red and blue colours indicate trained ALL and AML samples, respectively, as shown in Fig. 5. The random search function (Zhang et al 2019) searches for a subset of features using LDA classifier. The over-randomized subsets of gene features are used for evaluating the performance of the classifier with the actual set; from the performed results, it is observed that the first-item feature set have occurred most frequently in the subsets, so it conveys a good classification (Waltz

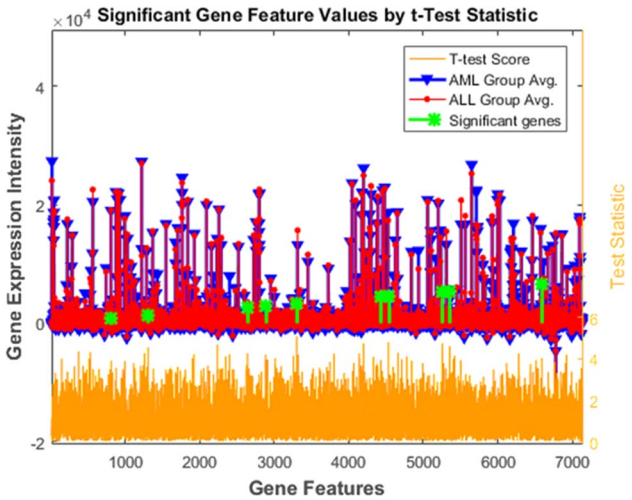

Fig. 4 Significant genes by *T*-test statistics

Table 2 *T*-test-based feature selection

DB ID	Gene description	*T*-score	Gene ID
3301	Canalicular multi-specific organic anion transporter	5.2310	U49248_at
4535	Retinoblastoma Binding Protein P48	4.9564	X74262_at
5254	MCM3 = mini-chromosome maintenance deficient 3	4.9153	D38073_at
4196	PRG1 = proteoglycan 1, secretory granule	4.9026	X17042_at
2242	Peptidyl-prolyl cistrans isomerase, mitochondrial precursor	4.8450	M80254_at
5352	GB DEF = non-muscle myosin heavy chain-B mRNA	4.7531	M69181_at
1306	CRYZ = crystallin zeta (quinone reductase)	4.7059	L13278_at
379	AARS = alanyl-tRNA synthetase	4.4304	D32050_at
532	HMG1 = high-mobility group protein 1	4.3983	D63874_at
4661	Biphenyl hydrolase-related protein	4.3953	X81372_at

Table 3 Leukaemia dataset sample size for sample classification

Samples	Training samples	Validation samples	Test samples
72 (total)	36	18	18
47 (ALL)	23	12	12
25 (AML)	13	6	6

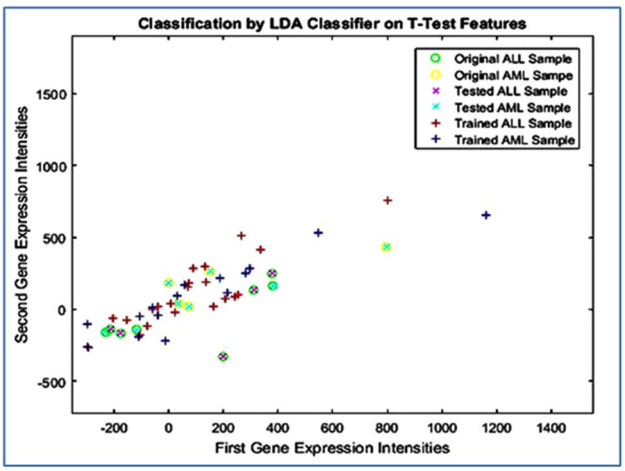

Fig. 5 Classification by LDA on *T*-test features

et al. 2006). LDA can be used to classify the ALL and AML samples. As gene expression data has only a small sample size, cross-validation using 10 hold-outs is taken for obtaining better efficiency during the classification performance (Mesko et al. 2010). The training and test sets for system evaluation are prepared based on *K*-fold principle and the selected gene feature subset is used only for training, and the validation is performed with only test subset. In completion of each loop the accuracy and the error rate are recorded for each classification.

From the graph plot shown in Fig. 6, the best correct classification is given by blue line curve and the mean correct classification rate is given by red line curve.

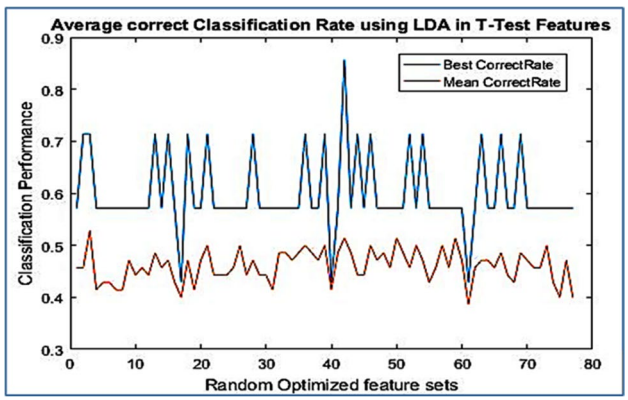

Fig. 6 Classification rate in *T*-test features by LDA-ROS

It is observed that on considering about 10 features in *T*-test-based selection a possible better classification is obtained. Likewise, the mean correct rate occurs at the maximum for a small number of gene features and it decreases gradually. As discussed by Han and Kamber (2006), high performance is achieved using the most frequently selected feature for classification of the gene test data.

In SOM, during the classification of samples (Raj et al. 2016), initially some significant features or clusters are identified to learn the data in order to classify the ALL from AML samples. Around 7129 genes are there before selection, and after classification using *T*-test, the size is reduced to 10.

A one-hidden layered SOM with 10 hidden neurons is constructed and trained. As the NN is designed with random initial weights, the output slightly varies every time after training the network. The training is done until the network continues to improve the validation set, and after the completion of the training the testing continues. The test dataset provides an independent test performance of network accuracy. The input feature size of network is 20 and output size

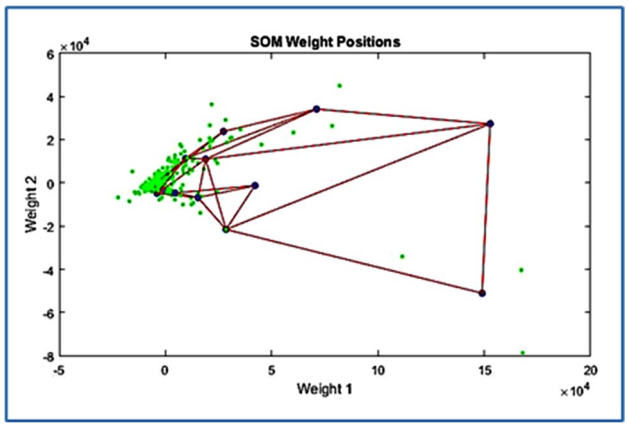

Fig. 7 Weight in SOM classifier trained *T*-test features

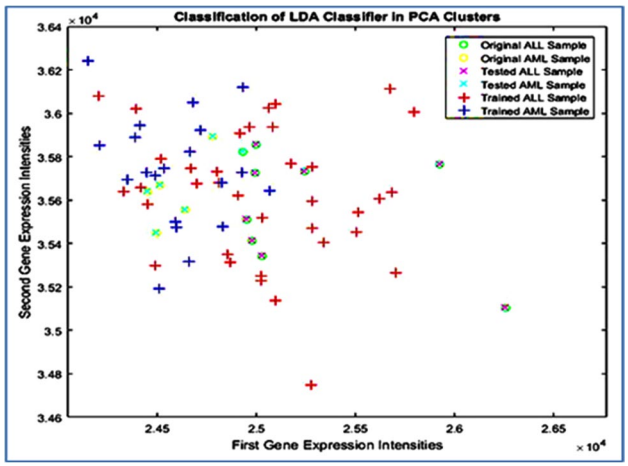

Fig. 9 Classification by LDA on PCA clusters

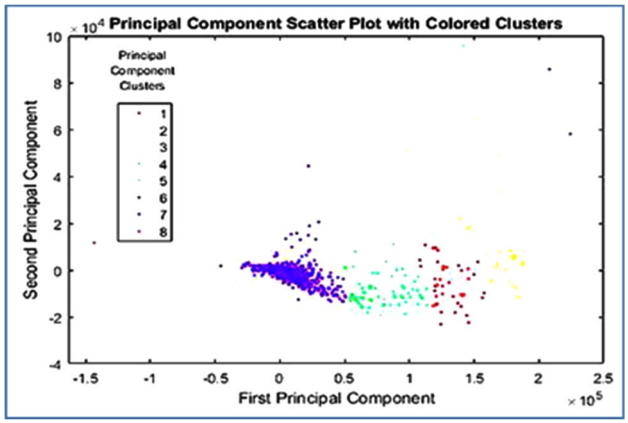

Fig. 8 Significant clusters in PCA-based transformation

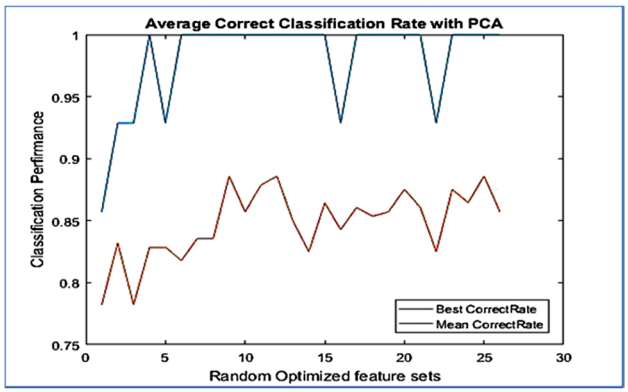

Fig. 10 Classification rate in PCA clusters in LDA-ROS

is 2 because the network has ALL and an AML sample, after the first training of *T*-test feature-based samples.

In the training process of the gene expression samples, the training vectors are represented as green spots and the hidden nodes are constructed for each of the 10 features with calculated weights. These nodes form the SOM network given in blue colour and are connected with lines are shown in Fig. 7. In the testing stage, the SOM network is formed in order to correctly classify the class of test datasets, and thereby, the accuracy of specific test set can be obtained by using the transfer function in the output layer.

In feature construction-based approach, PCA is used to reduce the genes as clusters. As the genes are transformed to PCs and the eight clusters based on first two PCs are indicated using purple and bright blue colours as shown in Fig. 8.

The PCA-based clusters are learned using LDA classifier; in this process, 20 PCs are used for training the samples. The actual ALL and AML samples are denoted by green and yellow points as shown in Fig. 9. Moreover, in the same figure, the magenta and sky-blue colours are used to represent the classified ALL and AML samples; similarly, the red and blue colours indicate trained ALL and AML categorical samples, respectively. In this process, when an actual ALL class of sample is given in green circle and if the predicted ALL sample is pink cross, it shows correct prediction. But when the sample is indicated as sky blue cross, it means wrong prediction. Also, actual AML class of sample is given in yellow circle and if the predicted AML sample is blue cross, it shows correct prediction, and the sample as pink cross is wrong prediction.

On using LDA-ROS classifier to operate the PCA clusters a high classification performance is obtained. The best and mean correct classification performance range deviation for the 10 optimized feature sets are shown in Fig. 10.

In PCA-based SOM learning the clusters are used to construct the network, similar to that of *T*-test feature-based classification. In this learning process, 20 PCs with calculated weights are used instead of features, and SOM network is formed with red nodes and blue connector lines as shown

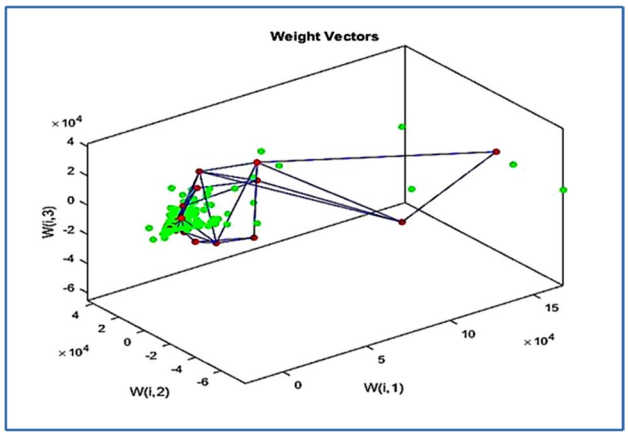

Fig. 11 Weight in SOM classifier trained PCA clusters

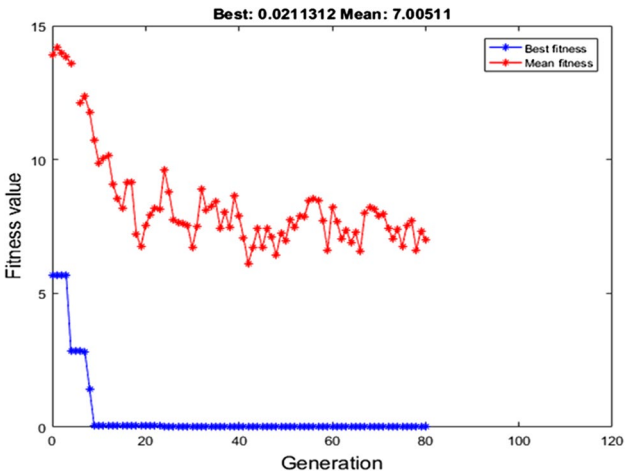

Fig. 12 Mean, best fitness value

in Fig. 11, where W(l,1), W(l,2) and W(l,3) are the weight vectors from each of the three PCs. The hidden layer is modelled and the gene expression validation sets and test set samples are classified into ALL and AML samples.

The another approach used in this study is Genetic Algothrim (GA), in which the considered number of variables and the population size are 10 and 72, respectively. The generation of GA is generally categorized into three groups: 50 (mean), 80 (best) and 120 (average). The selection function uses stochastic uniform function, and the crossover function uses 2-point crossover function as various GA options to improve accuracy. The mean and best fitness value for various generations can be viewed in Fig. 12. The GA optimizes at exact 10 generation, given by black dotted curve to obtain the best fitness, and the GA optimizes at 10 generation and has attained mean fitness at each future generations denoted by blue dotted curve. The best fitness value is 0.0211 at 10 generation and the mean fitness value is 7.005 at 40 generation. High-performance optimization is achieved through GA.

The top features are selected by the GA, and thus, retrieved data are tabulated in Table 4 along with the database ID, gene description and unique gene ID. Figure 13 displays the Gene Expression Intensity for various gene features, in which the average of various ALL group is denoted by green line circle marker and AML group average is denoted by blue colour triangle marker. In this figure, the highly optimized test subset genes are identicated with dual marked red asterisk sign.

Discussion

The results obtained from the various classifiers are studied and the performance measure for the classifiers is critically analysed by three methods: first by accuracy measure, secondly an *ROC* based measure is used and lastly by *PRC* measure. In performing the calculation using accuracy measure, the confusion matrix is constructed for *T*-test-based classification approach, PCA-based classification approach and GA-based classification approach using the computed *TP*, *TN*, *FP* and *FN* values. The accuracy is calculated for each of the feature selection-based classification methods

Table 4 Top features selected by GA

DB ID	Gene description	Gene ID
6515	DHPS = deoxyhypusine synthase	U26266_s_at
3189	HOXA9 = homeo box A9	U41813_at
6218	ELA2 = elastatse 2, neutrophil	M27783_s_at
1007	Collagen, type vi, alpha 2, N-terminal domain	HG4480-HT4833_at
3126	IAP homolog B (MIHB) mRNA	U37547_at
2834	GB DEF = orphan receptor ROR gamma mRNA	U16997_at
6638	GB DEF = CMKBR5 gene, non-functional mutant	X99393_s_at
1202	GB DEF = carbonic anhydrase-related protein (CARP) mRNA	L04656_at
4330	ITPKB = inositol 1,4,5-trisphosphate 3-kinase B	X57206_at
2238	GATA2 = GATA-binding protein 2	M77810_at

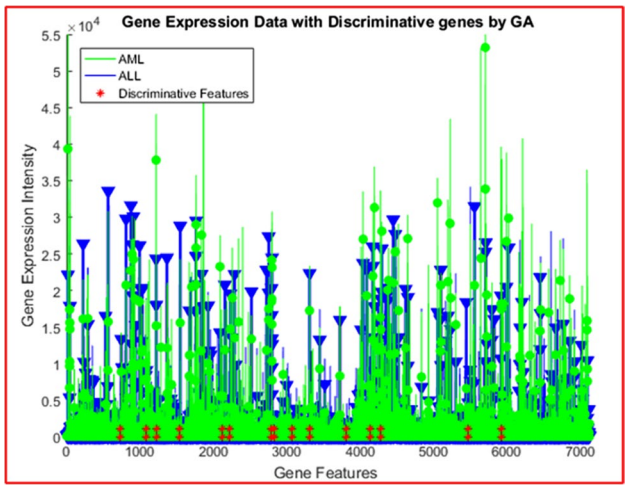

Fig. 13 Significant genes on GA

from the confusion matrix. The confusion matrix of PCA-SOM and GA at 120 generations is given in Fig. 14. The performance charts with the maximum accuracy and minimum error rate for all trained and tested classifiers are tabulated in Table 5. On comparing the data of accuracy during classification, the results indicate that GA-based classification proves to be better approach when compared to the other approaches.

The sensitivity, specificity, *PPV* and fallout for all feature selection-based classifier approaches are computed, and are listed in Table 6. Based on those above data the *ROC* and *PRC* performance measures are calculated using the equations given in Table 1. The chart shows the area under the *ROC* at Fig. 15, for the various feature selection-based classifier approaches given by their *TPR* and *FPR* ratio. GA gives the highest non-parametric performance as the *ROC* curve converges with upper left quarter, thus showing perfect classification. The area under the *PRC* is shown in Fig. 16 for the various feature selection-based classifier

Table 5 Accuracy for various feature selection-based classifiers

Approach	Method	Accuracy	*ERR*
Feature ranking	*T*-test-LDA	0.5929	0.4070
	T-test-ROS-LDA	0.8300	0.1699
	T-test-SOM	0.9285	0.0714
Feature construction	PCA-LDA	0.7428	0.2571
	PCA-ROS-LDA	0.8485	0.1514
	PCA-SOM	0.9444	0.0555
Optimization	GA-mean-50G	0.9305	0.0694
	GA-best-80G	0.9861	0.0138
	GA-Avg-120G	1.0000	0.0000

methods using precision and recall ratio of classifier prediction. GA gives the highest non-parametric performance as the *PRC* curve converges with upper right quarter, thus showing perfect classification. Similarly, PCA-based SOM classification perform better than *T*-test-based SOM classification in both *ROC*- and *PRC*-based metrics; also, *T*-test and PCA both classify better at ROS-based LDA method rather than ordinary LDA method in both *ROC*- and *PRC*-based metrics.

From this analysis, from the three data classifier approaches, in feature ranking approach *T*-test based on SOM classifier performed well, in feature construction-based approach PCA based on SOM classifier has given higher accuracy rate and on optimization approach the GA classification for average 120 generations shows higher accuracy other than best 80 generation and mean 50 generation optimizations. Thus, the GA shows higher performance, whilst comparing feature ranking-based classification, feature construction-based classification and optimization-based classification approaches.

Thus, the comparison analysis shows that the developed methods prove to perform better than DWLB (Subash Chandra Bose et al. 2021), CS (Sampathkumar et al. 2020) and

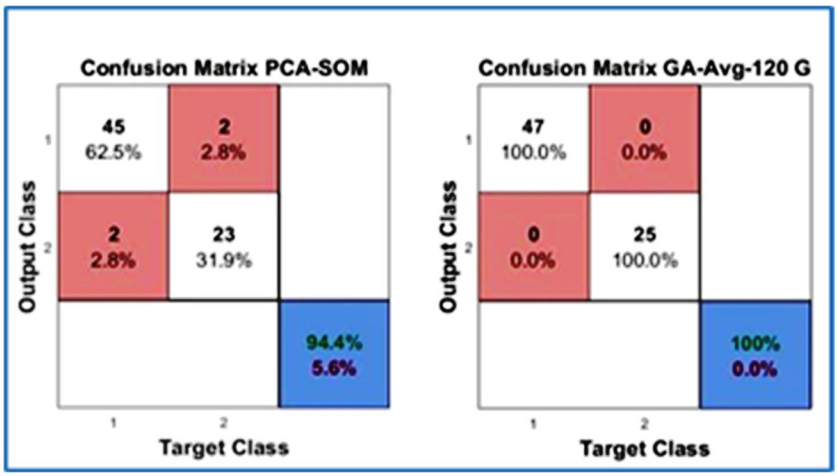

Fig. 14 Confusion matrix for PCA-SOM and GA-Avg-120G

Table 6 Sensitivity, specificity, *PPV* and fallout for various selection-based classifiers

Approaches	Method	Sensitivity	Specificity	*PPV*	Fallout
Feature ranking	*T*-test-LDA	0.5778	0.6200	0.7324	0.4493
	T-test-ROS-LDA	0.8797	0.7405	0.8592	0.7738
	T-test-SOM	1.0000	0.8000	0.9000	1.0000
Feature construction	PCA-LDA	0.6956	0.8333	0.8888	0.5882
	PCA-ROS-LDA	0.8911	0.7717	0.8754	0.7975
	PCA-SOM	0.9787	0.8800	0.9388	0.9565
Optimization	GA-mean-50G	0.9565	0.8846	0.9361	0.9200
	GA-best-80G	1.0000	0.9615	0.9787	1.0000
	GA-Avg-120G	1.0000	1.0000	1.0000	1.0000

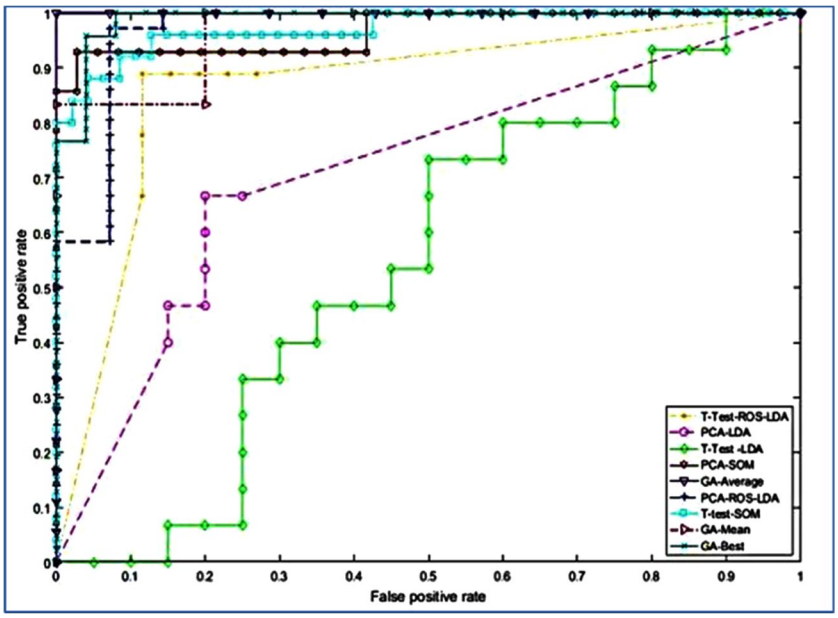

Fig. 15 Area under the *ROC* for various methods

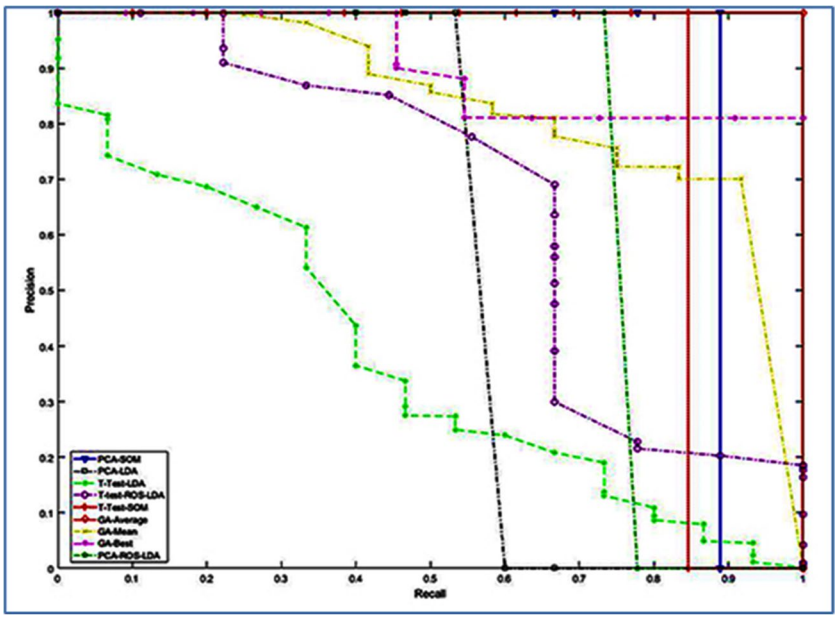

Fig. 16 Area under *PRC* for various methods

Table 7 Comparison of accuracy on classification

Approaches	Accuracy (%)
Wrapper-based GA (Liu et al. 2018)	91.87
Dynamic Weight LogitBoost (Subash Chandra Bose et al. 2021)	92
T-test-SOM	92.85
PCA-SOM	94.44
Cucoo Search (Sampathkumar et al. 2020)	96.90
GA-Avg-120G	100

wrapper-based GA (Liu et al. 2018) based on accuracy rankings provided in Table 7. As GA at 120 generation is the best in classifying leukaemia data with higher accuracy.

Conclusions

A novel search-based feature selection approach was developed in this research study to increase the classification performance. The performance of classifier such as the feature ranking-based classification (*T*-test), feature transformation-based classification (PCA) and embedded-based methods (GA) were analysed and compared by classifying leukaemia samples. The feature ranking- and feature transformation-based classification approaches are coupled with generic LDA, LDA-ROS and SOM classifiers to analyse the classifier's accuracy. In embedded approach (GA) the classification was performed under three different generation (mean, best and average)-based iteration. The performance of *T*-test-based model blended with SOM classifier show high accuracy in feature ranking-based classification approach. Similarly, in feature transformation-based approach, PCA-SOM classifier shows higher accuracy rate, and in embedded classifier approach, GA classifier yields best classification rate at average generation optimization. Amongst the considered classifiers the overall performance and accuracy was high for Genetic Algorithm-based classifiers. As a result, the identified classifier GA-Avg-120G shows a high accuracy in selecting leukaemia data compared to all other selection classifiers considered.

Declarations

Conflict of interest The authors declare no competing interests.

References

Ab Hamid TM, Sallehuddin R, Yunos ZM, Ali A. Ensemble based filter feature selection with harmonize particle swarm optimization and support vector machine for optimal cancer classification. Mach Learn Appl. 2021;5:100054.

Alpaydin E. Introduction to machine learning. MIT press; 2020.

Bakshi BR. Multiscale PCA with application to multivariate statistical process monitoring. AIChE J. 1998;44(7):1596–610.

Bell JB, Vigila SM. Gene selection approaches for classifying disease relevant data sample. Int J Eng Technol. 2018;7(3.27):62–9.

Bishop CM, Nasrabadi NM. Pattern recognition and machine learning. New York: Springer; 2006.

Chang HH, Moura JM. Biomedical signal processing. Biomed Eng Des Handb. 2010;2:559–79.

Cheng WY, Yang TH, Anastassiou D. Biomolecular events in cancer revealed by attractor metagenes. PLoS Comput Biol. 2013;9(2):e1002920.

Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006;27(8):861–74.

Furong H, Peiwen G, Fucui L, Xuewen L, Weimin Z, Wendong H. AML, ALL, and CML classification and diagnosis based on bone marrow cell morphology combined with convolutional neural network. Medicine. 2020;99(45):e23154.

Goswami RS, Sukhai MA, Thomas M, Reis PP, Kamel-Reid S. Applications of microarray technology to Acute Myelogenous Leukemia. Cancer Inform. 2009;7:13–28.

Gunavathi C, Premalatha K. Performance analysis of genetic algorithm with kNN and SVM for feature selection in tumor classification. Int J Comput Inform Eng. 2014;8(8):1490–7.

Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–82.

Han J, Kamber M. Data mining concepts. Model and Techniques. 2006.

Hancer E, Xue B, Zhang M. Differential evolution for filter feature selection based on information theory and feature ranking. Knowl-Based Syst. 2018;140:103–19.

Hernandez Hernandez JC, Duval B, Hao JK. A genetic embedded approach for gene selection and classification of microarray data. Eur Conf Evol Comput Mach Learn Data Min Bioinforma. 2007; 90–101.

Horn J, De Jesús O, Hagan MT. Spurious valleys in the error surface of recurrent networks—analysis and avoidance. IEEE Trans Neural Netw. 2009;20(4):686–700.

Hu Q, Che X, Zhang L, Yu D. Feature evaluation and selection based on neighborhood soft margin. Neurocomputing. 2010;73(10–12):2114–24.

Ilin A, Raiko T. Practical approaches to principal component analysis in the presence of missing values. J Mach Learn Res. 2010;11:1957–2000.

Japkowicz N. Supervised versus unsupervised binary-learning by feedforward neural networks. Mach Learn. 2001;42(1):97–122.

John G, Kohavi R. Wrappers for feature subset selection. Artif Intell. 1997;97(1):272–324.

Kumar C, Choudhary A. A top-down approach to classify enzyme functional classes and sub-classes using random forest. EURASIP J Bioinf Syst Biol. 2012;1:1–4.

Liu XY, Liang Y, Wang S, Yang ZY, Ye HS. A hybrid genetic algorithm with wrapper-embedded approaches for feature selection. IEEE Access. 2018;6:22863–74.

Mesko B, Poliskal S, Szegedi A, Szekanecz Z, Palatka K, Papp M, Nagy L. Peripheral blood gene expression patterns discriminate among chronic inflammatory diseases and healthy controls and identify novel targets. BMC Med Genomics. 2010;3(1):1–3.

Mitchell TM, Mitchell TM. Machine learning. New York: McGraw-hill; 1997.

Mundra PA, Rajapakse JC. SVM-RFE with MRMR filter for gene selection. IEEE Trans Nanobiosci. 2009;9(1):31–7.

Raj S, Ray KC, Shankar O. Cardiac arrhythmia beat classification using DOST and PSO tuned SVM. Comput Methods Programs Biomed. 2016;1(136):163–77.

Ron K, George HJ. Wrappers for feature subset selection. Artif Intell. 1997;97(1–2):273–324.

Sampathkumar A, Rastogi R, Arukonda S, Shankar A, Kautish S, Sivaram M. An efficient hybrid methodology for detection of cancer-causing gene using CSC for micro array data. J Ambient Intell Humaniz Comput. 2020;11:4743–51.

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, Feolo M. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2012;40:13–25.

Srisukkham W, Zhang L, Neoh SC, Todryk S, Lim CP. Intelligent leukaemia diagnosis with bare-bones PSO based feature optimization. Appl Soft Comput. 2017;56:405–19.

Subash Chandra Bose S, Sivanandam N, Praveen Sundar PV. Design of ensemble classifier using Statistical Gradient and Dynamic Weight LogitBoost for malicious tumor detection. J Ambient Intell Human Comput. 2021;12(6):6713–23.

Sun Y, Todorovic S, Goodison S. Local-learning-based feature selection for high-dimensional data analysis. IEEE Trans Pattern Anal Mach Intell. 2009;32(9):1610–26.

Tsanas A, Little MA, McSharry PE. A simple filter benchmark for feature selection. J Mach Learn R. 2010;1(1–24).

Tuv E, Borisov A, Runger G, Torkkola K. Feature selection with ensembles, artificial variables, and redundancy elimination. J Mach Learn Res. 2009;10:1341–66.

Waltz RA, Morales JL, Nocedal J, Orban D. An interior algorithm for nonlinear optimization that combines line search and trust region steps. Math Program. 2006;107(3):391–408.

Zhang J, Xiong Y, Min S. A new hybrid filter/wrapper algorithm for feature selection in classification. Anal Chim Acta. 2019;1080:43–54.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.