Medical & Biological Engineering & Computing
https://doi.org/10.1007/s11517-018-1874-4

ORIGINAL ARTICLE

Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods

Manosij Ghosh[1] · Sukdev Adhikary[1] · Kushal Kanti Ghosh[1] · Aritra Sardar[1] · Shemim Begum[2] · Ram Sarkar[1]

Received: 9 February 2018 / Accepted: 12 July 2018

Abstract

Microarray datasets play a crucial role in cancer detection. But the high dimension of these datasets makes the classification challenging due to the presence of many irrelevant and redundant features. Hence, feature selection becomes irreplaceable in this field because of its ability to remove the unrequired features from the system. As the task of selecting the optimal number of features is an NP-hard problem, hence, some meta-heuristic search technique helps to cope up with this problem. In this paper, we propose a 2-stage model for feature selection in microarray datasets. The ranking of the genes for the different filter methods are quite diverse and effectiveness of rankings is datasets dependent. First, we develop an ensemble of filter methods by considering the union and intersection of the top-*n* features of ReliefF, chi-square, and symmetrical uncertainty. This ensemble allows us to combine all the information of the three rankings together in a subset. In the next stage, we use genetic algorithm (GA) on the union and intersection to get the fine-tuned results, and union performs better than the latter. Our model has been shown to be classifier independent through the use of three classifiers—multi-layer perceptron (MLP), support vector machine (SVM), and K-nearest neighbor (K-NN). We have tested our model on five cancer datasets—colon, lung, leukemia, SRBCT, and prostate. Experimental results illustrate the superiority of our model in comparison to state-of-the-art methods.

Keywords Wrapper method · Filter method · Ensemble · Microarray data · Cancer detection

1 Introduction

DNA microarray provides the expression profiles of many genes which allow insights into the physiological processes and disease etiology meditated by those genes. Regulation of the expression of a gene occurs during the transcription of DNA into messenger ribonucleic acid (mRNA). Even though differential degradation of mRNA in the cytoplasm and others also cause regulation, relative quantity of mRNA species in cells is of great interest. The functions of gene expressions are ascertained from their upregulation and downregulation, which give gene expressions values significant importance. In the normal cell, continuous mutation damages the DNA which may lead to the impairment of cell replication. This is one of the main causes of formation of malignant tumor cells. Microarray gene expression data contain information regarding expression levels of the genes in certain tissue and cell. This data serve as a key source of information in different biological studies and analysis. Microarray data are therefore very useful in the field of tumor and cancerous gene detection.

There are quite a few types of cancers depending on where these cancers form and also depending on the types of cells

✉ Manosij Ghosh
manosij1996@gmail.com

Sukdev Adhikary
sukdev.999@gmail.com

Kushal Kanti Ghosh
kushalkanti1999@gmail.com

Aritra Sardar
sardararitra97@gmail.com

Shemim Begum
shemim_begum@yahoo.com

Ram Sarkar
raamsarkar@gmail.com

[1] Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

[2] Department of Computer Science and Engineering, Government College of Engineering & Textile Technology, Berhampore, West Bengal, India

Published online: 01 August 2018

that form the cancer. So it is an important task to differentiate between a cancer patient and a non-cancerous one and also to distinguish between different types of cancers to ensure appropriate treatment. Here lies the importance of cancer detection. Usually microarray data contain expression profiles of genes of both cancerous (tumor) and non-cancerous (normal) cells. Proper analysis would help the medical practitioner, drug designer to identify the genes responsible for cancers and help them to take the necessary action before the disease goes beyond treatment. Therefore, microarray gene expression data derives its importance from the fact that treatment becomes easier after detection.

Generally, microarray data contain few samples (typically around 100) and large number of features (~ 6000 to ~ 60,000) which lead to the "curse of dimensionality" [1]. In this kind of data most of the features are irrelevant and/or redundant. The most relevant ones are referred to as biomarkers, as the expression values can indicate the occurrence of cancer. Therefore, finding biomarkers is a substantial research problem. The irrelevant data on the other hand, can affect the accuracy of cancer detection system and also increase the computing time for the same. In short, not all the genes are responsible for cancer, a quite small percentage of the total number of genes are responsible for causing cancer. Herein lies the importance of feature selection (FS), which eliminates the irrelevant and/or redundant data from a dataset and makes detection faster and more accurate.

Brute force FS refers to the selection of all types of combination of the features and then their evaluation to find the best subset. Obviously, the cost of this method is too high (as it is an NP-hard problem), thereby making an exhaustive search a non-feasible option. Forward [2] and backward [3] FSs, though improvements on the brute force technique, provide poor accuracy due to the constrained search mechanism. Forward FS adds the features one after another and evaluates the improvement, if there is no improvement, the added feature is removed. Backward FS starts with all the features and then removes a feature from the set if the accuracy of the subset increases after the feature is removed. This process continues iteratively for all features until no more improvement is observed. These methods do not allow a large number of subsets to be evaluated which affect their performance.

FS in contemporary research is mostly done in three different ways—filter methods [4], wrapper methods [5], and hybrid or embedded methods [6]. Filter methods select a subset of features using the intrinsic properties of the data, independent of any learning algorithm. Though this method is fast and computationally less expensive, accuracy is low as the selection is classifier independent. Wrapper methods, on the other hand, utilize the performance of classifiers (learning algorithm) to evaluate the worth of feature subsets. Wrapper method normally outperforms filter method in terms of accuracy but at the cost of higher computational complexity. Hybrid methods try to combine the two approaches and perform FS using filter method along with a wrapper method. Hybrid methods due to their ability to include both intrinsic data properties and a learning algorithm to evaluate feature subsets generally perform better than the other two categories.

Filter methods consist of a variety of techniques including subset selection techniques like minimum redundancy maximum relevancy (mRMR) [7], correlation based feature selection (CFS) [8], and feature ranking techniques like symmetrical uncertainty (SU) [9], ReliefF [10], information gain [11], and chi-square [12]. Feature ranking methods produce an ordering of the features using some information content or distance or statistical (or dependence) criteria. Filter ranking methods are used for their simplicity and good success is reported for practical applications [13–15]. Another approach says that a feature can be regarded as irrelevant if it is conditionally independent of the class labels and vice versa. Therefore, if a feature is to be relevant, it cannot be independent of the class labels, i.e., the feature that has no influence on the class labels can be discarded.

Filter methods can determine subset with optimum features by various measures, like information (or uncertainty), distance, and dependence (or probability) [16]. A brief overview of the three categories is provided hereafter.

1. Information measure: this procedure typically computes the entropy (or uncertainty) of the features with respect to the class labels and hereby calculates the information gain. Obviously a feature X is more preferable than another feature Y if the information gain of X is more than that of Y. Information gain of a feature X is defined by the difference between the prior uncertainty due to the feature X and the expected posterior uncertainty using that feature. SU [9] is an important as well as popular technique following this measure.
2. Distance measure: this is also known as divergence, discrimination, or separability measurement. Normally, this type of measure is used for two-class problem, though they can be extended for multi-class problem as well. If the feature X causes more difference in the conditional probability of the two-class than feature Y, then X is more preferable. Various methods can be used for difference measurement, like Euclidean, Manhattan, Mahalanabish. One of the most well-known techniques of this category is ReliefF [10].
3. Dependency measure: this category of methods, also known as correlation measurement, uses the ability to predict the value of one variable from that of another variable. Using various classical dependence measures, we can find the correlation between a feature and a class. If the correlation value between a feature A and class C is higher than that of feature B and class C, then A is more preferable than B. This type of methods can also be used

to determine the dependence between two features. Chi-square [12] feature ranking, which belongs to this type, is a very useful and well-explored method.

Wrapper methods are used extensively in FS and a few algorithms of this category include genetic algorithm (GA) [17], particle swarm optimization (PSO) [18], ant colony optimization (ACO) [19], gravitational search algorithm (GSA) [20], etc. GA is not only the simplest of all these algorithms but has been extensively used since its inception in the 1970s. GA has been shown to be extremely useful in cases of multi-objective optimizations where it is needed to reduce the feature dimension and increase the classification accuracy. GA is a computationally inexpensive algorithm of this category.

Hybrid methods are generally more efficient than its two counterparts. General approach for hybrid methods in FS is a 2-stage model. The first stage consisting of refining of features by a filter method (generally a ranking technique chooses the top-n features) followed by a wrapper method which finds the most discriminative subset from the top-n.

As stated before, different filter methods decide optimal features differently, i.e., selection process of deterministic factors would vary in different methods. So, different methods, when applied individually, result in different outcomes. If we apply a number of methods individually and combine their results, not only can we get the most important information from all the methods but also there is a high chance that prediction performance of the system would improve. Therefore, aim of combining multiple FS methods is to increase the maximum accuracy achieved by a single method, since the combination is capable of overcoming the errors caused by other methods in different parts of the input space while enhancing the accuracy by providing their own complementary view on importance of features. We can combine the methods by taking the union or the intersection of the results generated. Union is preferable as it combines the best of the three methods to create our new search space. This search space too is quite large at times and may not compose entirely of relevant features. So, further refining through some low-cost wrapper method is justifiable.

2 Related work

A brief overview of some recent works on FS in microarray datasets is presented in this section. We have also highlighted the wide use and applicability of the 2-stage hybrid model. Work reported in [21] has proposed an ensemble approach where genes are ranked by using blogreg, t test, and Fisher, and then the sum of the three rankings is used as the final ranking. Then, 1% of the top genes is selected for the next step. Discriminant independent component analysis (dICA) is subsequently used on the selected features to transform them with PSO as the optimization function and SVM as a classifier. Another method using PSO is [22], where first correlation-based feature selection (CFS) is used to select discriminative features and Taguchi Chaotic Binary PSO (TCBPSO) is applied on the selected features. K-NN classifier is used with leave-one-out-cross-validation (LOOCV) as fitness measure. Taguchi is implemented using signal-to-noise-ratio (SNR) and orthogonal array (OA). They have proposed a method [23] in which initial datasets are pre-processed using a quartile-based technique and after that Hamming distance-based binary PSO (HDBPSO) is applied. Hamming distance is introduced as a proximity measure to update the velocity of particles in PSO.

In distributed FS, proposed in [24], a dataset is divided into several smaller disjoint sets either randomly or using a ranking method and placing features, with similar relevance to the class, together. Filter method is applied to each subset to select

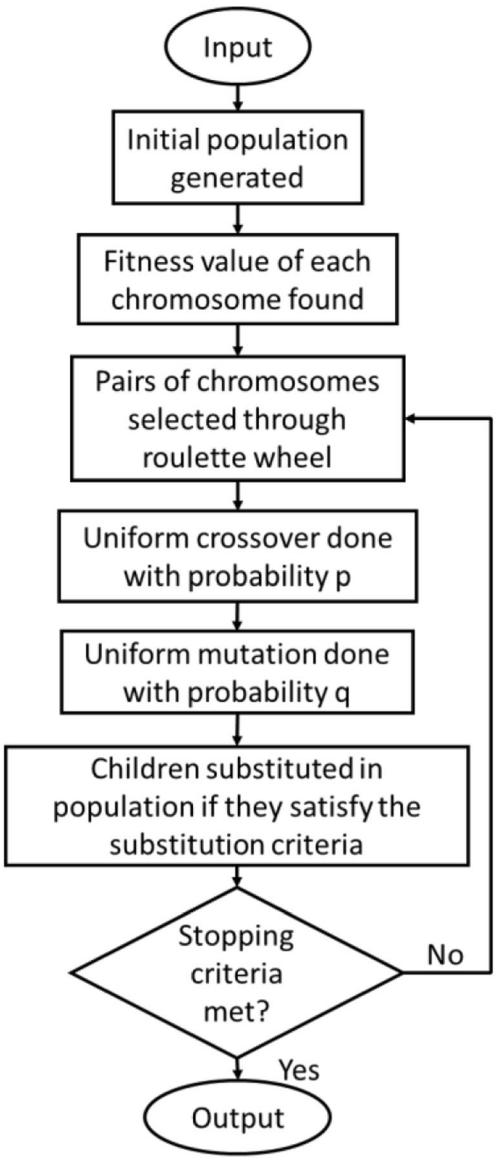


Fig. 1 Flowchart of the GA

Table 1 Description of the datasets used to evaluate the proposed method

Dataset	Total number of samples	Gene dimension	Distribution	
			Class	No. of samples
Colon	36	7464	Colon cancer	18
			Non-cancerous	18
Lung	203	12,601	Adenocarcinoma(AD)	139
			Normal lung(NL)	17
			Small cell lung Cancer(SMCL)	6
			Squamous cell Carcinoma(SQ)	21
			Pulmonary Carcinoid(COID)	20
Leukemia	72	5148	Acute lymphoblastic Leukemia (ALL)	47
			Acute myeloid Leukemia (AML)	25
Prostate	102	12,534	Normal tissue	50
			Prostate tumor	52
SRBCT	83	2309	Ewing's sarcoma (EWS)	29
			Burkitt's lymphoma (BL)	11
			Neuroblastoma (NB)	18
			Rhabdomyosarcoma (RMS)	25

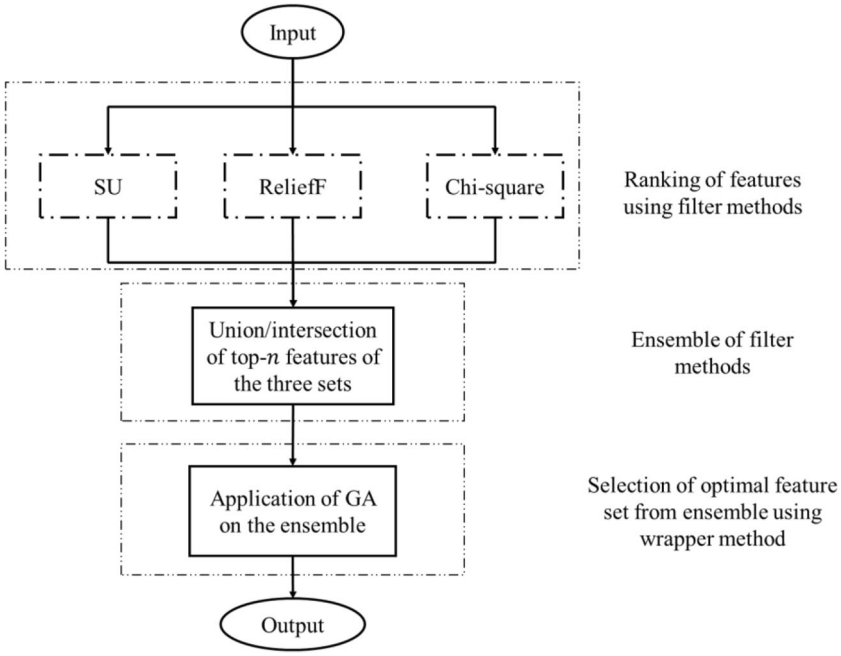


Fig. 2 Flowchart of the proposed 2-stage FS model

a smaller subset of features from each set. The subsets are joined in a method akin to forward FS (except subsets are used instead of features) to form the final selected subset.

A 2-stage approach proposed in [25] consists of feature ranking using information gain (IG) and then applying binary differential evolution (BDE). Another 2-stage approach proposed in [26] consists of five filters—IG, ReliefF, CFS, INERACT, and consistency-based filter. Then, the five sets are fed to a specific classifier whose decisions are combined using simple voting. In [27] also, a 2-stage model is proposed with two filter methods—*F* score and IG in the first stage. The wrapper method—sequential floating search method (SFSM is a combination of sequential forward selection (SFS) and sequential backward selection (SBS)) starts with the

Fig. 3 **a**–**j** Variations of accuracy (**a**, **c**, **e**, **g**, **i**) and number of features (**b**, **d**, **f**, **h**, **j**) with the value of *n* for the five datasets using the MLP classifier are shown. Plots are done for union (E_U), intersection (E_I), ReliefF, chi-square, and SU. ▶


Accuracy vs n for colon
Accuracy (in %)
n
Union ReliefF Chi-Square SU Intersection
a
No. of features vs n for colon
No. of features
n
Union ReliefF Chi-Square SU Intersection
b
Accuracy vs n for lungs
Accuracy (in %)
n
Union ReliefF Chi-Square SU Intersection
c
No. of features vs n for lungs
No. of features
n
Union ReliefF Chi-Square SU Intersection
d
Accuracy vs n for leukeamia
Accuracy (in %)
n
Union ReliefF Chi-Square SU Intersection
e
No. of features vs n leukeamia
No. of features
n
Union ReliefF Chi-Square SU Intersection
f
Accuracy vs n for prostate
Accuracy (in %)
n
Union ReliefF Chi-Square SU Intersection
g
No. of features vs n for prostate
No. of features
n
Union ReliefF Chi-Square SU Intersection
h
Accuracy vs n for SRBCT
Accuracy (in %)
n
Union ReliefF Chi-Square SU Intersection
i
No. of features vs n for SRBCT
No. of features
n
Union ReliefF Chi-Square SU Intersection
j

intersection of two filter methods as the initial stage and limits the search space to the union of the two filter method's results.

In sequential random K-nearest neighbors (SRKNN) [28], authors have taken *k*-base classifiers, each selecting a set of features using forward sequential selection. Thereafter, the feature sets are combined and if the accuracy is better, the new set is chosen. The process is iterated considering the output feature set as the new input set for the next iteration.

In [29], two variations of kernel ridge regression (KRR), namely, wavelet kernel ridge regression (WKRR) and radial basis kernel ridge regression (RKRR) are used to classify the features obtained from modified cat swarm optimization (MCSO) using K-NN classifier.

In [30], authors have proposed a quadratic programming feature selection (QPFS) method based on semidefinite programming model which is relaxed from the quadratic programming model with maximizing feature relevance and minimizing feature redundancy. Lagrange multiplier has been used as proxy measurement to identify the expected features instead of solving a feasible solution for the original max-cut problem.

Variable neighborhood search-based on predominant grouping (PGVNS) [31] consists of variable neighborhood search (VNS), a filter method, and the concept of Markov blankets to group the input space into subsets of features called predominant group. Each predominant group is composed of one predominant feature X along with all redundant features for which X forms a Markov blanket.

3 Brief overview of concepts

We describe three filter methods used in the present work in Sections 3.1–3.3 and GA—the wrapper method in Section 3.4.

3.1 Symmetrical uncertainty (SU)

In information theory founded by Shanon [9], the uncertainty of a variable A is measured by its Entropy $H(A)$. Whereas for two variables A and B, their conditional entropy $H(A|B)$ measures the uncertainty about A when the variable B is known and the mutual information (MI) [32]-MI(A; B) gives the certainty about B that is resolved by A. According to Shanon, entropy $H(A)$ of a variable A is defined by

$H(A) = -\sum_{c \in C} P(a) \log P(a)$, where $P(a)$ represents the probability of the variable A. Conditional entropy $H(A|B)$ is defined by

$$H(A|B) = -\sum_{b\in B} P(b)\left(\sum_{a\in A} P(a/b)\log P(a/b)\right)$$

MI between A and B is given by

$$\mathrm{MI}(A;B) = H(A) - H(A/B) = H(B) - H(B/A) = \mathrm{MI}(B;A)$$
$$\mathrm{MI}(A;A) = H(A)$$
$$\mathrm{MI}(A;B) = \sum_{b\in B}\sum_{a\in A} P(b,a)\log P(b,a)/(P(a)*P(b))$$

Here, the goal is to reduce uncertainty for prediction of the class C for a known feature X, and hence, to increase MI(X; C) as much as possible. In other words, the aim is to select features with higher MI. In principle, MI can be exactly calculated only when the probability distribution function of the data is known. However, data may not be distributed in a fixed pattern and it is required to estimate the MI from the given data.

The fitness of a feature is determined by SU. A feature with high value of SU gets higher priority. SU of two variables A and B is given by the formula

$$\mathrm{SU}(A,B) = \frac{2*\mathrm{MI}(A;B)}{H(A)+H(B)}$$

SU behaves symmetrically towards a couple of variables, i.e., it compensates for the bias of MI towards the feature having large number of different values and normalizes it in [0, 1]. A SU value of 0 represents the independence of the variables, and the value 1 indicates that the variables are strongly connected. The features are ordered in decreasing order of their dependence on the class of the features.

3.2 Chi-square

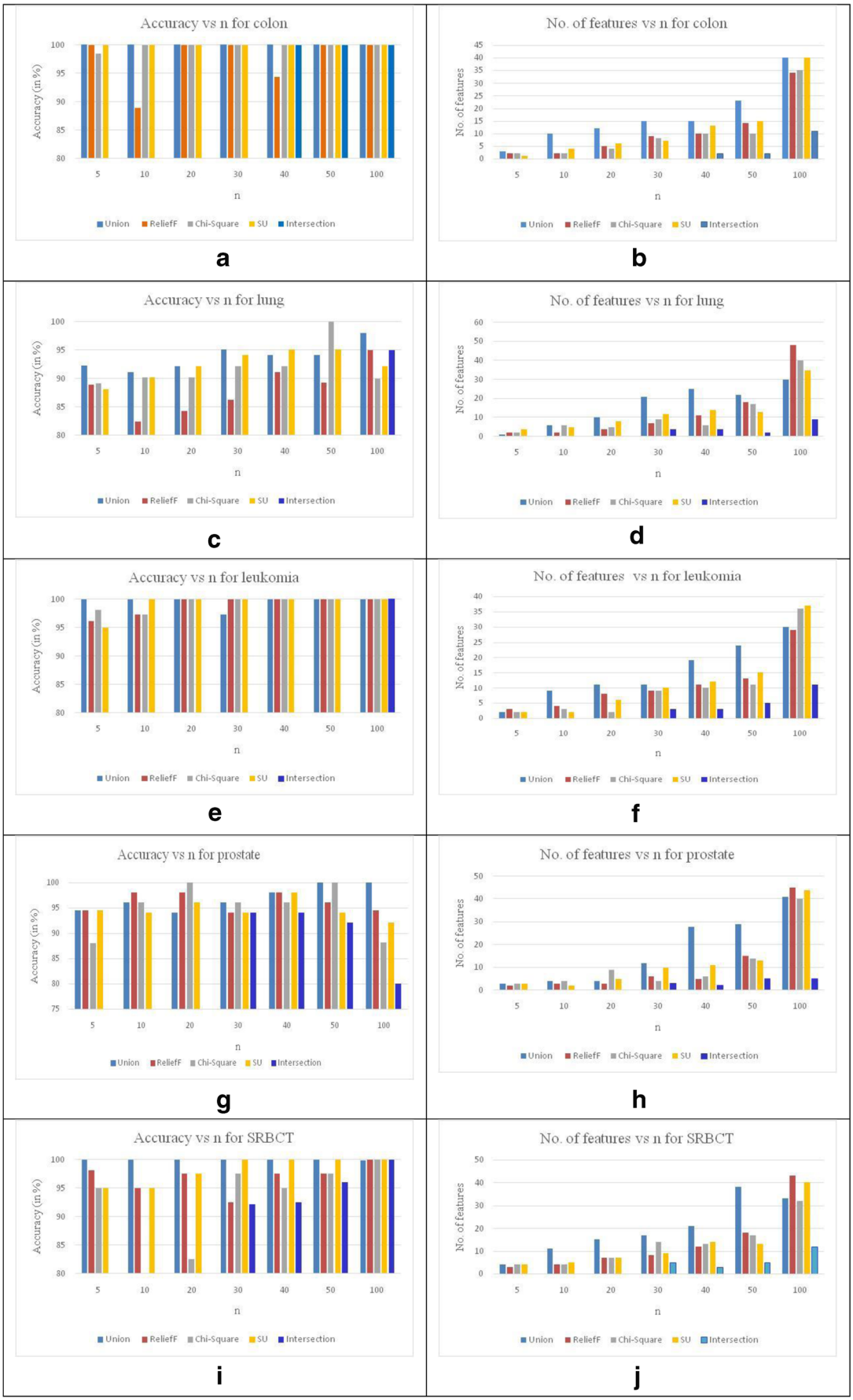
The chi-square statistics measure the lack of independence between the feature f and the class C, and can be compared with chi-square distribution to judge the extremeness. Chi-square is based on comparing obtained values of the frequency of a class due to the split and expected frequency. The range of continuous values is needed to be discretized into intervals.

Let N be the number of examples, N_{ij} be the number of samples of C_i class within jth interval, and M_{ij} be the number of samples in jth interval. Now, the expected frequency of N_{ij} is given by

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{I} \frac{\left(N_{ij}-E_{ij}\right)^2}{E_{ij}}$$

Where I is the number of the intervals and c is the number of classes. If the probability of occurrence of event I is given by p_i, then expected value, $E_i = \mathrm{num} * p_i$, where num is the total number of events. Higher value of chi-square indicates that the feature is more informative.

Fig. 4 **a**–**j** Variations of accuracy (**a**, **c**, **e**, **g**, **i**) and number of features (**b**, **d**, **f**, **h**, **j**) with the value of n for the five datasets using the K-NN classifier are shown. Plots are done for union (E_U), intersection (E_I), ReliefF, chi-square, and SU. ▶


Accuracy vs n for colon
Accuracy (in %)
n
Union ReliefF Chi-Square SU Intersection
a
No. of features vs n for colon
No. of features
n
Union ReliefF Chi-Square SU Intersection
b
Accuracy vs n for lung
Accuracy (in %)
n
Union ReliefF Chi-Square SU Intersection
c
No. of features vs n for lung
No. of features
n
Union ReliefF Chi-Square SU Intersection
d
Accuracy vs n for leukomia
Accuracy (in %)
n
Union ReliefF Chi-Square SU Intersection
e
No. of features vs n for leukomia
No. of features
n
Union ReliefF Chi-Square SU Intersection
f
Accuracy vs n for prostate
Accuracy (in %)
n
Union ReliefF Chi-Square SU Intersection
g
No. of features vs n for prostate
No. of features
n
Union ReliefF Chi-Square SU Intersection
h
Accuracy vs n for SRBCT
Accuracy (in %)
n
Union ReliefF Chi-Square SU Intersection
i
No. of features vs n for SRBCT
No. of features
n
Union ReliefF Chi-Square SU Intersection
j

3.3 ReliefF

This was originally developed by Kira and Rendel as Relief [33]. The key idea of this method is to select features according to how well their values distinguish from nearest instances. So, for a given feature, we search for two nearest neighbors: one from the same class (called nearest hit H) and another from the other class (called nearest miss M). Here, the basic rule is that this value for a good attribute must differ largely from that of another attribute with different class and should have almost similar value as that of an attribute with same class. Traditional Relief method is for two-class problem.

Later it has been modified for multi-class problem. If R_i and H are different for a particular X, then X differentiates them even when they belong to the same class, so corresponding weightage is reduced. If R_i and M are different for a value of X, then as expected, X differentiates between two instances from different class, so the weightage of X is increased.

The difference function diff(X, Y, Z) measures the difference between the values of attribute X for two instances Y and Z. It is defined as

$$\mathrm{diff}(\mathrm{X},\mathrm{Y},\mathrm{Z}) = |(val(X,Y) - val(X,Z))|/max(X) - min(X)$$

The extension of Relief is ReliefF [34], which can work with more noisy data (both for two- and multi-class problems). ReliefF is similar to Relief except that it searches for k number of nearest neighbors in the same class (nearest hits) and in other classes (nearest misses).

3.4 Genetic algorithm (GA)

GA is inspired by the natural process of evolution and selection. Modeling the nature, parents create children and the fitters have the better chances of survival. So, eventually like natural evolution, GA considers only the so called fitter versions of the chromosomes over time. The feature subsets encoded as chromosomes are taken to be individuals and a collection of the chromosomes forms a population. Chromosomes are binary strings where '0' at position i denotes that the ith feature is not selected and '1' denotes that the ith feature is selected. Each chromosome is then given a fitness value denoted by the accuracy of those selected features in predicting the correct classes (gene labels) using a classifier.

We use a multi-objective GA (as described later) which allows us to give importance to both accuracy and the number of selected features, i.e., feature dimension. The flowchart of GA is given in Fig. 1. The chromosomes, generated by genetic operations crossover and mutation, form the new population. However in this work, the parent chromosomes substitute the child if they are superior according to our multi-objective function. The use of this mechanism allows us to implement the elitism concept allowing the parents, better than the children, to pass onto the next generations.

The population is first randomly initialized by creating chromosomes with k random bits as '1' ($k \in [n/2, 3n/4]$, n is the feature dimension) and rest are made '0'. Accuracies of all the chromosomes are then measured. Thereafter, crossover creates new children or offspring from parent chromosomes selected from the population by the use of a roulette wheel. This gives the chromosomes, with higher accuracies, a greater probability of being a parent. Once the children are generated by crossover, we perform mutation on the chromosomes to decrease the possibility of the population getting stuck in a local optima.

3.4.1 Crossover and mutation

Crossover can be done in a number of ways including 1-point crossover, 2-point crossover, and uniform crossover, of which uniform crossover has been shown to give the best results among the three [35]. The two parents selected by the roulette wheel undergo an exchange of their chromosome elements with a probability of $p = 0.5$. We thereafter perform uniform mutation where we flip all the bits of the chromosome with a probability of $q = 0.01$. Then, the children are used to substitute the chromosomes in the population.

3.4.2 Substitution algorithm

Due to the importance of accurately identification of cancerous genes, we allow more importance to accuracy. We set a tolerance limit $\omega = 0.02$, which gives us the limit of accuracy we are ready to compromise to allow a decrease in feature dimension. We take the values of $\mathrm{wt}_1 = 4$ and $\mathrm{wt}_2 = 1$.

c_1, c_2 *are the 2 chromosomes*
if accuracy difference $< \omega$
$\quad$ c_i *with better accuracy is chosen*
else
$\quad$ c_i*with better weighted average is chosen*
Weighted average $= wt_1{}^{*}accuracy + wt_2{}^{*}(1 - ratio\ of\ features\ selected)$

Fig. 5 (**a**–**j**) Variations of accuracy (**a**, **c**, **e**, **g**, **i**) and number of features (**b**, **d**, **f**, **h**, **j**) with the value of n for the five datasets using the SVM classifier are shown. Plots are done for union (E_U), intersection (E_I), ReliefF, chi-square, and SU. ▶

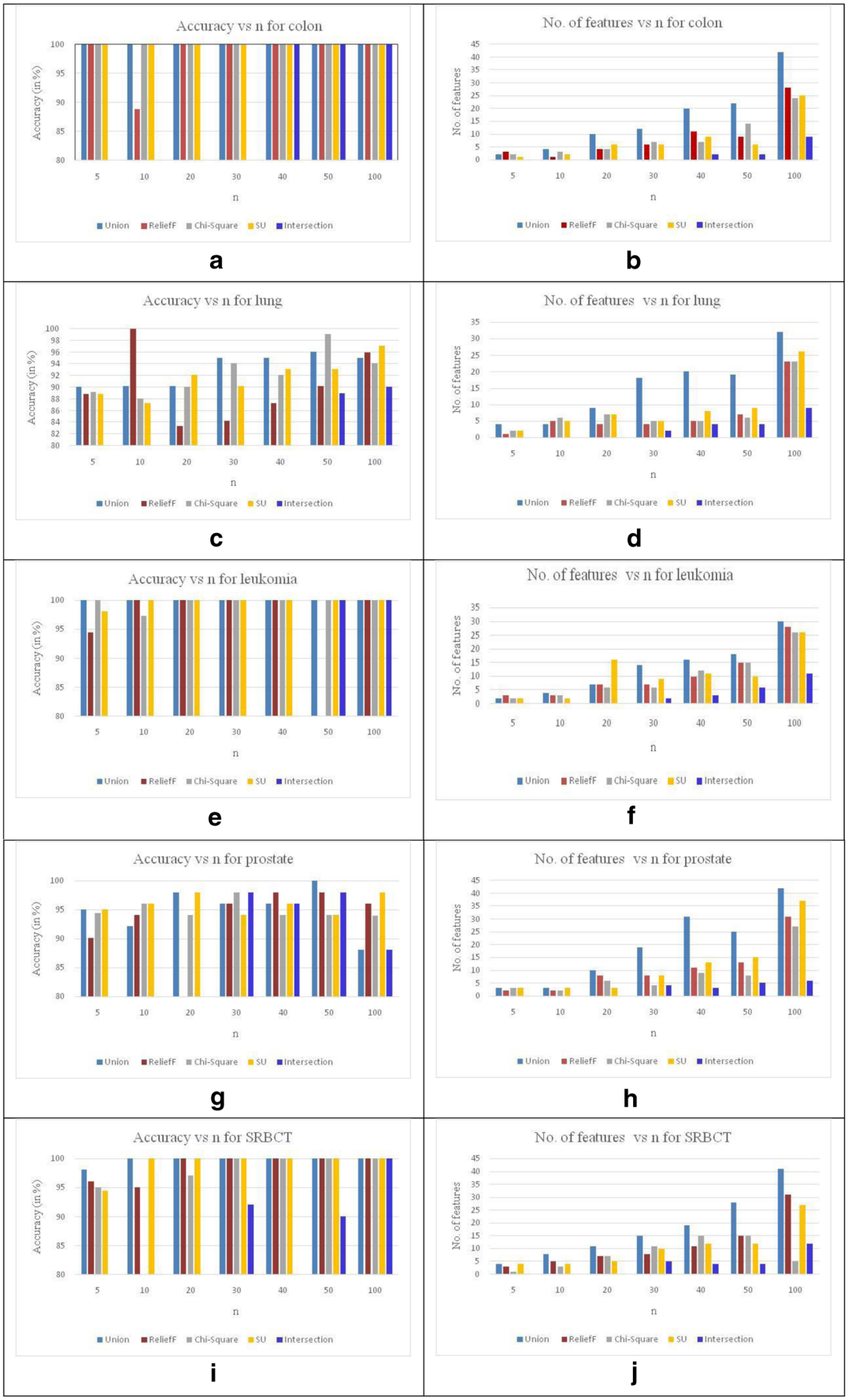

Accuracy vs n for colon
Accuracy (in %)
n
Union ReliefF Chi-Square SU Intersection
a
No. of features vs n for colon
No. of features
n
Union ReliefF Chi-Square SU Intersection
b
Accuracy vs n for lung
Accuracy (in %)
n
Union ReliefF Chi-Square SU Intersection
c
No. of features vs n for lung
No. of features
n
Union ReliefF Chi-Square SU Intersection
d
Accuracy vs n for leukomia
Accuracy (in %)
n
Union ReliefF Chi-Square SU Intersection
e
No. of features vs n for leukomia
No. of features
n
Union ReliefF Chi-Square SU Intersection
f
Accuracy vs n for prostate
Accuracy (in %)
n
Union ReliefF Chi-Square SU Intersection
g
No. of features vs n for prostate
No. of features
n
Union ReliefF Chi-Square SU Intersection
h
Accuracy vs n for SRBCT
Accuracy (in %)
n
Union ReliefF Chi-Square SU Intersection
i
No. of features vs n for SRBCT
No. of features
n
Union ReliefF Chi-Square SU Intersection
j

Table 2 Summarization of results of our 2-stage FS model when GA is applied on E_U

Dataset	Classifiers					
	K-NN		MLP		SVM	
	Accuracy (in %)	No. of features	Accuracy (in %)	No. of features	Accuracy (in %)	No. of features
Colon	100	3	100	2	100	2
Lung	92.24	1	96.07	9	90.19	4
Leukemia	100	2	100	2	100	2
Prostate	96.07	4	98.03	7	100	25
SRBCT	100	4	100	4	100	8

4 Methodology

The justification behind the present work is outlined in Section 4.1, and the proposed model along with how it deals with the shortcomings of existing works has been explained in Section 4.2.

4.1 Justification

From the literature, it can be observed that for FS in microarray data, general research trend is to combine a filter method and a wrapper method following a 2-stage approach in order to find out the most informative genes or biomarkers. The first stage consists of a filter method, which evaluates the genes and the top-n genes are qualified to go to the next stage. Next stage consists of a wrapper method which then tries to pinpoint the most informative gene subset from the n genes. However, the ranking of the genes for the different filter methods are quite diverse, and it has been observed that for some datasets, some methods provide much better results. Therefore, we can conclude that efficiency of general 2-stage model is dependent on the effectiveness of filter method. Due to this, selection of a single filter method in the first stage is not justified. Another crucial factor is the selection of the value of n, a small value leads of elimination of informative features, whereas a large value makes the selection harder. We can infer that this general trend may not work satisfactorily as selection of the filter method and the value of n are extremely crucial to the success of the algorithm.

4.2 Proposed model

The proposed work is an improvement on the generally used 2-stage filter-wrapper method. Different feature ranking methods assign varying ranks to the same feature, and each method has its own merits and demerits. This hypothesis has been used in this work. Instead of just selecting the top-n genes from a particular filter method, we propose a model where we make an ensemble of three different filter methods, one from each category as mentioned earlier—information, distance, and dependency based. The ensemble of filter methods, from three distinct categories, allows us to combine the information we get from the three ranking techniques. If some important features get missed by any method, then there is a high chance of it being chosen by some other method (this chance gets maximized as we use filter rankers of three different categories); thereby, making the combination robust. This helps us to include the efficiency of each of the three ranking techniques. So, we can get rid of the reliance of a particular filter method on a dataset, and more specifically create a model which gives optimal results on all datasets.

A gene can be assumed to be good (useful for classification) if a gene is categorized as good by any of the three filter

Table 3 List of methods with which the performance of the proposed model is compared

Method	Abbreviation
Particle swarm optimization with discriminant independent Component analysis [21]	PSO + dICA
Kernel principal component analysis [37]	Kernel PCA
Probabilistic principal component analysis [38]	Prob PCA
Gaussian process latent variable model [39]	GPLVM
Isomap [40]	ISOMAP
Factor analysis [41]	FA
Principle component analysis [42]	PCA
Independent component analysis [43]	ICA
Discriminant independent component analysis [43]	dICA

Table 4 Comparison of the proposed E_U and E_I with other state-of-the-art methods for gene selection in microarray data

	Proposed models		Other methods											
Dataset	E_U	E_I	ReliefF	Chi-square	SU	PSO + dICA	Kernel PCA	Prob PCA	GPLVM	ISOMAP	FA	PCA	ICA	dICA
Colon	100(2)	100(10)	100(2)	100(2)	100(3)	100(10)	100(5)	94.73(5)	84.21(5)	94.73(5)	89.47(5)	94.73(5)	94.74(10)	94.74(10)
Lung	96.07(9)	91.17(8)	91.17(9)	97.05(9)	98.03(10)	100(10)	98.03(10)	98.03(10)	100(20)	100(10)	98.03(10)	96(10)	96(10)	100(10)
Leukemia	100(2)	100(12)	100(4)	100(6)	100(2)	100(10)	100(10)	100(20)	100(10)	94.44(20)	100(10)	100(10)	100(30)	94.44(20)
Prostate	98.03(7)	100(16)	100(6)	98.07(7)	98.03(8)	96.77(10)	93.54(10)	87(30)	83.87(10)	83.87(30)	87(30)	90.32(10)	93.55(10)	96.77(10)
SRBCT	100(4)	100(12)	100(7)	100(15)	100(7)	100(5)	100(5)	96(5)	100(10)	92(10)	96(5)	100(5)	100(5)	100(5)

methods. We take the top-n genes from each of the three filter rankings and make an ensemble to include the information from all the three rankings. The filter methods we choose from each of the three categories are SU (information based), ReliefF (distance based), and chi-square (dependency based). This new subset of features is then passed to a meta-heuristic wrapper method called GA in order to obtain the best subset of features, which can achieve the best results. In each step of GA, some random procedure is followed. Hence, the removal of irrelevant features through ensemble makes GA perform better. The ensemble of the three rankings can be done by using both *union* and *intersection*. Therefore, we propose two approaches to make the ensemble of three filter methods using union (E_U) and intersection(E_I). As ensemble approach, we prefer union between two because this allows us to include the best features selected by each ranking technique. This leads to minimization of dependence between datasets and ranking techniques. The flowchart of our work is given in Fig. 2.

5 Results and discussion

We have applied our FS model on five popular microarray datasets to prove the effectiveness of the same. The dataset descriptions are summarized in Table 1. The details of the datasets can be found at [36]. It is to be noted that the feature dimensions of the used datasets range from ~2000 to ~13,000. We divide our datasets to form training and test sets such that the ratio of train to test is 1:1.

The proposed model ensembles the top-n (ranked) features from the three ranking techniques described before and then applies GA on them to find the best subset of genes that may be biomarkers. We test our model taking values of n as 5, 10, 20, 30, 40, 50, and 100. To prove the classifier independence nature of our model, we use three classifiers—MLP, K-NN, and SVM. The set E_U consist of all the features in top-n of any of the three feature ranks whereas, set E_I contains features present in top-n of at least two rankings. Figure 3 shows the variations of accuracy and number of features selected verses the value of n for our proposed models—$E_U$and E_I. Figure 3 also contains three 2-stage methods, using the ranking methods (ReliefF, chi-square, SU) and MLP classifier. Each of SU, chi-square, and ReliefF model first chooses top-n features (given by the corresponding ranking algorithm) and then applies GA for further improvement. All these FSs are done first using MLP classifier. The same is repeated for K-NN and SVM to show the classifier independence of our models, and the results are shown in Figs. 4 and 5 respectively.

It is to be noted from the Figs. 3, 4 and 5 that there is no intersection of features if the value of n is taken less than 30. It is worth mentioning that this outcome is noteworthy as this

Table 5 List of all the gene ids and their names when the value of n is set to 5 for E_U

Probe/ uniprotID	Gene name	Citation(s)
Colon		
Z49269	CCL14	[44]
M12272	ADH1C	[45, 46]
U37019	CNN1	[47]
	Lung	
32378_at	PKM	[48]
34319_at	S100P	[49]
34095_f_at	IGHG1	[50, 51]
41238_s_at	CELA3A	[52, 53]
	Leukemia	
Z19554_s_at	VIM	[54–56]
X17042_at	SRGN	[57, 58]
	Prostate	
37639_at	HPN	[59, 60]
32598_at	NELL2	[61–63]
36864_at	PEX3	[64, 65]
32146_s_at	ADD1	[66]
	SRBCT	
796,258	CD99	[67]
52,076	PBX1	[67]
236,282	WAS	[67]
770,394	NF2	[67]
812,105	CDH2	[67]

Table 6 (a–e) TF, KEGG pathway, and GO are listed for the genes obtained by the proposed $E_U$model for all five datasets for $n = 5$

Colon			
Gene name	TF	KEGG pathway	GO
CNN1	CRTC3[0.04]	Endothelin Pathways_Homo sapiens_WP2197[0.001]	GO:CC perinuclear theca (GO:0033011)[0.02] cortical cytoskeleton (GO:0030863)[0.0006] GO: BP regulation of smooth muscle contraction involved in micturition (GO:1904318)[0.0004] regulation of gastro-intestinal system smooth muscle contraction (GO:1904304)[0.0004]
ADH1C	CBFB (human)[0.03]	hsa_00350Tyrosine metabolism_ Homo sapiens [0.001]	
CCL14	MYC[0.04]	has_04062Chemokine signaling pathway_Homo sapiens [0.009] has_04060Cytokine-cytokine receptor interaction_Homo sapiens [0.01]	GO:BP C-C chemokine receptor CCR2 signaling pathway (GO:0038150)[0.002] GO:CC chemokine activity (GO:0008009)[0.007] GO:MF retinol dehydrogenase activity (GO:0004745)[0.0008]
		Lung	
Gene name	TF	KEGG pathway	GO
PKM	MBD2[0.01]	hsa01230_Biosynthesis of amino acids_Homo sapiens [0.003]	GO:CC platelet alpha granule lumen (GO:0031093)[0.009] GO:BP cellular response to insulin stimulus (GO:0032869)[0.003] insulin receptor signaling pathway (GO:0008286) glucose import in response to insulin stimulus (GO:0044381) GO:MF cadherin binding involved in cell-cell adhesion (GO:0098641)[0.001] cadherin binding (GO:0045296)[0.001]
S100P	CRTC1[0.04]		GO:CC interchromatin granule (GO:0035061)[0.01] nuclear dicing body (GO:0010445) GO:BP lymphatic endothelial cell migration (GO:1904977)[0.001] GO:MF adherin binding involved in cell-cell adhesion (GO:0098641)[0.01] cadherin binding (GO:0045296)
IGHG1	IGHM[0.01]	has04974_Protein digestion and absorption_Homo sapiens [0.004] has04972_Pancreaticsecretion_ Homo sapiens [0.004]	GO:BP defense response to Gram-positive bacterium (GO:0050830)[0.005] GO:MF serine-type endopeptidase activity (GO:0004252)[0.01]

Table 6 (continued)

Colon			
CELA3A	CRX[0.04]		GO:BP peptidyl-glycine cholesteryl ester biosynthesis from peptidyl-glycine (GO:0019708)[0.005] GO:MF serine-type endopeptidase activity (GO:0004252)[0.01]
		Leukemia	
Gene name	TF	KEGG pathway	GO
SRGN	ESR1[*p* value: 0.04]	NA	GO:BP negative regulation of interleukin-17 secretion [*p* value0.008500] negative regulation of transforming growth factor-beta secretion [p-value 0.0005500] GO:CC platelet alpha granule lumen[0.009150] mast cell granule[0.0.008500]
VIM	a)APEX1 (human) [0.05] b)FOS (human) [0.05]	has05169_Epstein-Barr virus infection_Homo sapiens has05206_MicroRNAs in cancer_ Homo sapiens	GO: CC actin cytoskeleton (GO:0015629)[0.004] microtubule cytoskeleton (GO:0015630)[0.004] GO: BP muscle filament sliding (GO:0030049)[0.005] GO:MF double-stranded miRNA binding (GO:0098851)[0.002] RNA strand-exchange activity (GO:0034057)[0.003] RNA strand annealing activity (GO:0033592)[0.003]
		Prostate	
Gene name	TF	KEGG pathway	GO
HPN	CEBPE [0.006350]		GO:BP hepatocyte growth factor receptor signaling pathway GO:0048012[0.0005500] positive regulation of gene expression GO:0010628[0.01070] GO:CC nuclear membrane GO:0031965[0.0008350] GO:MF serine-type edopeptidase activityGO:0004252[0.009250]
NELL2	LOC135440[p-value0.01]	has03030_DNA replication_ homosapiens [*p* value 0.001]	GO:BP Cellular response to epidermal growth factor stimulus(GO:0071364)[p-value 0.01] GO:CC Extracellular region (GO: 0005576) [p-value 0.004] GO:MF Calcium ion binding(GO:0005509) [p-value 0.03]
PEX3	ATF1 (human)[0.01]		GO:CC

Table 6 (continued)

Colon			
		hsa04146_Peroxisome_Homo sapiens [0.004]	integral component of peroxisomal membrane (GO:0005779)[0.0006] peroxisomal membrane (GO:0005778)[0.002] extrinsic component of intraperoxisomal membrane (GO:0005780)[0.002] GO:BP peroxisome membrane biogenesis (GO:0016557)[0.0009]
ADD1	SREBF2 (human) [0.04]		GO:CC interchromatin granule (GO:0035061)[0.01] histone locus body (GO:0035363)[0.01] GO:BP ion transmembrane transport (GO:0034220)[0.001] positive regulation of collagen binding (GO:0033343)[0.002] GO:MF protein heterodimerization activity (GO:0046982)[0.02] actin monomer binding (GO:0003785)[0.01]
		SRBCT	
Gene name	TF	KEGG pathway	GO
CD99	PCBP1 [0.04]	hsa04670_Leukocyte transendothelial migration_Homo sapiens hsa04514_Cell adhesion molecules (CAMs)_Homo sapiens	GO:CC plasma membrane proton-transporting V-type ATPase, V0 domain (GO:0000222)[0.05] GO:BP regulation of innate immune response (GO:0045088) regulation of antigen receptor-mediated signaling pathway (GO:0050854)
PBX1	CBEPA [0.04]	hsa05202_Transcriptional misregulation in cancer_Homo sapiens	GO:MF bent DNA binding (GO:0003681) recombination hotspot binding(GO:0010844) recombination hotspot binding(GO:0010844) negative regulation of SREBP signaling pathway by DNA binding (GO:0100060)
WAS	GTF2I[0.01] XBP1[0.03]	hsa05130_Pathogenic *Escherichia coli* infection_Homo sapiens hsa05231_Choline metabolism in cancer_Homo sapiens	GO:BP regulation of actin polymerization or depolymerisation actin filament of cell cortex of cell tip GO:MF protein tyrosine kinase binding 3-phosphoinositide-dependent protein kinase binding
NF2	PLAU (0.04) SREBF2 (0.04)	hsa04390_Hippo signaling pathway_Homo sapiens	GO:BP negative regulation of DNA-dependent DNA replication GO:CC actin cytoskeleton
CDH2	GABPA (0.005) EGR1 (0.03), CRTC3 (0.05)		GO:BP radial glial cell differentiation

Table 6 (continued)

Colon		
	hsa05412_Cell Arrhythmogenic right ventricular cardiomyopathy (ARVC)_Homo sapiens hsa04514_Cell adhesion molecules (CAMs)_Homo sapiens	GO:CC sarcoplasmic reticulum lumen

signifies the different nature of the filter methods (the hypothesis we have assumed at the beginning) and this very reason motivates us not to believe blindly on any single filter method. For the value of n as 40 and 50, E_I gives good results. In case of E_U, as the number of features increases, the results deteriorate (in terms of feature dimension). The best results for union are observed when $n = 10$ and $n = 5$.

The results depicted in Figs. 3, 4 and 5 clearly show the superiority of E_U over the individual ranking model as well as E_I. The summary of the best results of E_U, for all the five datasets and all three classifiers, is listed in Table 2.

The results of E_U and E_I model are then compared with some state-of-the-art methods. Since MLP produces the best results, so the comparison with other methods is done considering only MLP as classifier. The methods with which we compare our model are listed in Table 3 and the comparison result is provided in Table 4. This comparison shows that we outperform all the state-of-the-art methods for three out of five datasets. For the rest two, we are only beaten by a small margin but never by a single method, i.e., no single method outperforms our method in more than 1 dataset. Therefore, in conclusion, we can say that our method outperforms all methods individually.

6 Biological significance

Selection of genes (i.e., biomarkers), which are responsible for cancer, is of utmost importance. Therefore, to prove the utility of the proposed model in identifying the biomarkers, we provide the biological significance of the selected genes. It is to be noted that we give the biological significance of the genes selected by E_U for $n = 5$ as it yields better result than any other value of n. Table 5 lists the gene ids, their corresponding names, and the papers which refer to those genes as cancerous. Whereas Table 6 records the transcription factor (TF), KEGG pathway, and gene ontology (GO) of the genes. The three aspects of GO—molecular function (MF), cellular component (CC), and biological process (BP)—are also listed.

7 Conclusion

FS in gene expression data is a real challenge to the researchers due to high dimension of the datasets. To address this, we have proposed two simple models—ensemble of filter rankings along with GA, namely, E_U and E_I. Three different types of filter methods viz., ReliefF (distance based), chi-square (dependence based), and SU (information based) are taken into consideration at the first stage, and union or intersection of the top-n features are done. The subset is used as an input to GA. The classifier independence of our model can be seen from the results obtained by using the three classifiers—MLP, K-NN, and SVM. To the best of our knowledge, this is the first time when this kind of 2-stage model has been proposed.

We have also compared our results with some recent works, and it is observed that our approach achieves better accuracy with lower number of features in comparison to others. No single method has been able to outperform our results neither in terms of accuracy nor in number of selected features. Biological significance of the genes selected by our model has been extensively studied as well as papers, which have marked those genes as cancerous have been referred. Large number of match proves the effectiveness of our model in cancerous gene identification. Finally, it is worth mentioning that the proposed method is a general model, which can also be applied to other high-dimensional datasets (like RNA sequence). A detailed biological study of the genes obtained from our work can also be done to ascertain better cancer gene identification systems.

Compliance with ethical standards

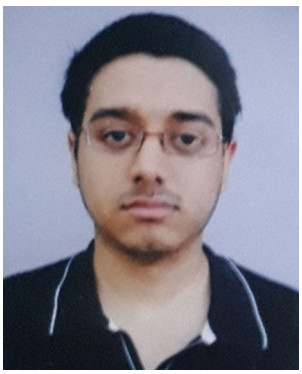
Competing interests None of the authors has any competing interests in the manuscript.

References

1. Vaidya AR (2015) Neural mechanisms for undoing the “curse of dimensionality”. J Neurosci 35:12083–12084
2. Jain A, Zongker D (1997) Feature selection: evaluation, application, and small sample performance. IEEE Trans Pattern Anal Mach Intell 19:153–158

3. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
4. Mitra P, Murthy CA, Pal SK (2002) Unsupervised feature selection using feature similarity. IEEE Trans Pattern Anal Mach Intell 24: 301–312
5. Kashef S, Nezamabadi-pour H (2015) An advanced ACO algorithm for feature subset selection. Neurocomputing 147:271–279. https://doi.org/10.1016/j.neucom.2014.06.067
6. Duval B, Hao J-K, Hernandez Hernandez JC (2009) A memetic algorithm for gene selection and molecular classification of cancer. Proc 11th Annu Conf Genet Evol Comput - GECCO '09 201 . doi: https://doi.org/10.1145/1569901.1569930
7. Mohamed NS, Zainudin S, Othman ZA (2017) Metaheuristic approach for an enhanced mRMR filter method for classification using drug response microarray data. Expert Syst Appl 90:224–231
8. Hall MA (1999) Correlation-based feature selection for machine learning
9. Shannon CE, Weaver W (1964) The mathematical theory of communication. University of Illinois Press, Urbana, pp 10–61
10. Wang Z, Zhang Y, Chen Z et al (2016) Application of ReliefF algorithm to selecting feature sets for classification of high resolution remote sensing image. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2016, pp 755–758. https://doi.org/10.1109/IGARSS.2016.7729190
11. Uğuz H (2011) A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowl-Based Syst 24:1024–1032
12. Jin X, Xu A, Bie R, Guo P (2006) Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. In: International workshop on data mining for biomedical applications. Springer-Verlag Berlin, Heidelberg, pp 106–115
13. Zheng Z, Wu X, Srihari R (2004) Feature selection for text categorization on imbalanced data. ACM Sigkdd Explor Newsl 6:80–89
14. Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23:2507–2517
15. Kwon O-W, Chan K, Hao J, Lee T-W (2003) Emotion recognition by speech signals. In: Eighth European Conference on Speech Communication and Technology
16. Dash M, Liu H (1997) Feature selection for classification. Intell Data Anal 1:131–156
17. Yang J, Honavar V (1998) Feature subset selection using a genetic algorithm. IEEE Intell Syst their Appl 13:44–49
18. Jain I, Jain VK, Jain R (2017) Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. Appl Soft Comput 62:203–215. https://doi.org/10.1016/j.asoc.2017.09.038
19. Forsati R, Moayedikia A, Jensen R et al (2014) Enriched ant colony optimization and its application in feature selection. Neurocomputing 142:354–371. https://doi.org/10.1016/j.neucom.2014.03.053
20. Rashedi E, Nezamabadi-Pour H, Saryazdi S (2010) BGSA: binary gravitational search algorithm. Nat Comput 9:727–745. https://doi.org/10.1007/s11047-009-9175-3
21. Mollaee M, Moattar MH (2016) A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification. Biocybern Biomed Eng 36:521–529
22. Chuang L-Y, Yang C-S, Wu K-C, Yang C-H (2011) Gene selection and classification using Taguchi chaotic binary particle swarm optimization. Expert Syst Appl 38:13367–13377
23. Banka H, Dara S (2015) A hamming distance based binary particle swarm optimization (HDBPSO) algorithm for high dimensional feature selection, classification and validation. Pattern Recogn Lett 52:94–100
24. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A (2015) Distributed feature selection: an application to microarray data classification. Appl Soft Comput 30:136–150
25. Apolloni J, Leguizamón G, Alba E (2016) Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. Appl Soft Comput J 38:922–932. https://doi.org/10.1016/j.asoc.2015.10.037
26. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A (2012) An ensemble of filters and classifiers for microarray data classification. Pattern Recogn 45:531–539
27. Hsu H-H, Hsieh C-W, Lu M-D (2011) Hybrid feature selection by combining filters and wrappers. Expert Syst Appl 38:8144–8150
28. Park CH, Kim SB (2015) Sequential random k-nearest neighbor feature selection for high-dimensional data. Expert Syst Appl 42: 2336–2342
29. Mohapatra P, Chakravarty S, Dash PK (2016) Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system. Swarm Evol Comput 28:144–160
30. Sun S, Peng Q, Zhang X (2016) Global feature selection from microarray data using Lagrange multipliers. Knowl Based Syst 110:267–274
31. García-Torres M, Gómez-Vela F, Melián-Batista B, Moreno-Vega JM (2016) High-dimensional feature selection via feature grouping: a variable neighborhood search approach. Inf Sci (NY) 326:102–118
32. Huang J, Cai Y, Xu X (2007) A hybrid genetic algorithm for feature selection wrapper based on mutual information. Pattern Recogn Lett 28:1825–1844
33. Kira K, Rendell LA (1992) A practical approach to feature selection. In: Proceedings of the ninth international workshop on. Mach Learn:249–256
34. Kononenko I (1994) Estimating attributes: analysis and extensions of RELIEF. In: European Conference on Machine Learning. Lecture Notes in Computer Science book series (LNCS), vol 784. Springer-Verlag Berlin, Heidelberg, pp 171–182
35. Spears WM, De Jong KD (1995) On the virtues of parameterized uniform crossover. Naval Research Lab, Washington DC
36. BioInformatics Laboratory http://www.biolab.si/supp/bi-cancer/projections/info/BC_CCGSE3726_frozen.html
37. Schölkopf B, Smola A, Müller K-R (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 10:1299–1319
38. Tipping ME, Bishop CM (1999) Probabilistic principal component analysis. J R Stat Soc Ser B (Statistical Methodol) 61:611–622
39. Lawrence ND (2006) The Gaussian process latent variable models for visualisation of high dimensional data. In: Proceedings of the 16th International Conference on Neural Information Processing Systems. MIT Press, Cambridge, pp 329–336
40. Li C-G, Guo J (2006) Supervised isomap with explicit mapping. In: innovative computing, information and control, 2006. ICICIC'06. First International Conference on. IEEE, pp 345–348
41. Van Der Maaten L, Postma E, Van den Herik J (2009) Dimensionality reduction: a comparative. J Mach Learn Res 10: 66–71
42. Pinto da Costa JF, Alonso H, Roque L (2011) A weighted principal component analysis and its application to gene expression data. IEEE/ACM Trans Comput Biol Bioinforma 8:246–252
43. Dhir CS, Lee J, Lee S-Y (2012) Extraction of independent discriminant features for data with asymmetric distribution. Knowl Inf Syst 30:359–375
44. le Rolle A-F, Chiu TK, Fara M et al (2015) The prognostic significance of CXCL1 hypersecretion by human colorectal cancer epithelia and myofibroblasts. J Transl Med 13:199

45. Kropotova ES, Zinovieva OL, Zyryanova AF et al (2014) Altered expression of multiple genes involved in retinoic acid biosynthesis in human colorectal cancer. Pathol Oncol Res 20:707–717
46. Bongaerts BWC (2008) Alcohol consumption as a risk factor for colorectal cancer: an epidemiological study on genetic susceptibility and molecular endpoints. Maastricht University, Maastricht, pp 127–144
47. Chiang S-C, Han C-L, Yu K-H et al (2013) Prioritization of cancer marker candidates based on the immunohistochemistry staining images deposited in the human protein atlas. PLoS One 8:e81079
48. Papadaki C, Sfakianaki M, Lagoudaki E et al (2014) PKM2 as a biomarker for chemosensitivity to front-line platinum-based chemotherapy in patients with metastatic non-small-cell lung cancer. Br J Cancer 111:1757–1764
49. Liang B, Shao Y, Long F, Jiang S-J (2016) Predicting diagnostic gene biomarkers for non-small-cell lung cancer. Biomed Res Int 2016:1–8
50. Lonergan KM, Chari R, Coe BP et al (2010) Transcriptome profiles of carcinoma-in-situ and invasive non-small cell lung cancer as revealed by SAGE. PLoS One 5:e9162
51. Jiang C, Huang T, Wang Y et al (2014) Immunoglobulin G expression in lung cancer and its effects on metastasis. PLoS One 9: e97359
52. Van den Broeck A, Vankelecom H, Van Eijsden R et al (2012) Molecular markers associated with outcome and metastasis in human pancreatic cancer. J Exp Clin Cancer Res 31:68
53. Goonesekere NCW, Andersen W, Smith A, Wang X (2017) Identification of genes highly downregulated in pancreatic cancer through a meta-analysis of microarray datasets: implications for discovery of novel tumor-suppressor genes and therapeutic targets. J Cancer Res Clin Oncol 144(2):309–320
54. Bittanti S, Garatti S, Liberati D (2005) From DNA micro-arrays to disease classification: an unsupervised clustering approach. IFAC Proc 38:319–324
55. Labaj W, Papiez A, Polanski A, Polanska J (2017) Comprehensive analysis of MILE gene expression data set advances discovery of leukaemia type and subtype biomarkers. Interdiscip Sci Comput Life Sci 9:24–35
56. Liberati D, Bittanti S, Garatti S (2005) Unsupervised mining of genes classifying leukemia. In: Encyclopedia of data warehousing and mining. IGI Global, pp 1155–1159
57. Khabbaz M, Kianmehr K, Alshalalfa M, Alhajj R (2010) An integrated framework for fuzzy classification and analysis of gene expression data. Strategic advancements in utilizing data mining and warehousing technologies, pp 151–153
58. Tong DL (2010) Genetic algorithm-neural network: feature extraction for bioinformatics data. Doctorate Thesis (Doctorate), Bournemouth University
59. Chen Z, Gerke T, Bird V, Prosperi M (2017) Trends in gene expression profiling for prostate cancer risk assessment: a systematic review. Biomed Hub 2:1
60. Kelly KA, Setlur SR, Ross R et al (2008) Detection of early prostate cancer using a hepsin-targeted imaging agent. Cancer Res 68:2286–2291
61. Noel EE, Ragavan N, Walsh MJ et al (2008) Differential gene expression in the peripheral zone compared to the transition zone of the human prostate gland. Prostate Cancer Prostatic Dis 11:173–180
62. D'Antonio KEB (2009) Analysis of novel targets in the pathobiology of prostate cancer. University of Pittsburgh
63. Kelemen A, Abraham A, Chen Y (2008) Computational intelligence in bioinformatics. Springer, Heidelberg
64. Lazzarini N, Bacardit J (2017) RGIFE: a ranked guided iterative feature elimination heuristic for the identification of biomarkers. BMC Bioinformatics 18:322
65. Xu J, Mu H, Wang Y, Huang F (2018) Feature genes selection using supervised locally linear embedding and correlation coefficient for microarray classification. Comput Math Methods Med 2018. https://doi.org/10.1155/2018/5490513
66. Massoner P, Lueking A, Goehler H et al (2012) Serum-autoantibodies for discovery of prostate cancer specific biomarkers. Prostate 72:427–436
67. Tsai Y-S, Aguan K, Pal NR, Chung I-F (2011) Identification of single-and multiple-class specific signature genes from gene expression profiles by group marker index. PLoS One 6:e24259



Manosij Ghosh is presently an undergraduate student of Computer Science & Engineering department at Jadavpur University. His research interests comprise of Evolutionary Algorithms and Bioinformatics.

Sukdev Adhikary received his Bachelor degree in CSE from IIEST and is currently pursuing his Masters in Jadavpur University. His areas of interest include Filter methods and Bioinformatics.

Kushal Kanti Ghosh is presently an undergraduate student of Computer Science & Engineering department at Jadavpur University. His research interests comprise of Bioinformatics and Machine Learning.

Shemim Begum received her B. Tech degree from University of Burdwan in 2003. She received her M. Tech degree from University of Calcutta in 2005. She is pursuing PhD from Jadavpur University.

Aritra Sardar is presently an undergraduate student of Computer Science & Engineering department at Jadavpur University. His areas of interests comprise of machine learning and Bioinformatics.

Ram Sarkar received his B. Tech from University of Calcutta in 2003. He received his Master and PhD degrees from Jadavpur University in 2005 and 2012 respectively. He is a senior member of the IEEE.