Fuzzy-Rough Entropy Measure and Histogram Based Patient Selection for miRNA Ranking in Cancer

Jayanta Kumar Pal, Shubhra Sankar Ray, Sung-Bae Cho, and Sankar K. Pal

Abstract—MicroRNAs (miRNAs) are known as an important indicator of cancers. The presence of cancer can be detected by identifying the responsible miRNAs. A fuzzy-rough entropy measure (FREM) is developed which can rank the miRNAs and thereby identify the relevant ones. FREM is used to determine the relevance of a miRNA in terms of separability between normal and cancer classes. While computing the FREM for a miRNA, fuzziness takes care of the overlapping between normal and cancer expressions, whereas rough lower approximation determines their class sizes. MiRNAs are sorted according to the highest relevance (i.e., the capability of class separation) and a percentage among them is selected from the top ranked ones. FREM is also used to determine the redundancy between two miRNAs and the redundant ones are removed from the selected set, as per the necessity. A histogram based patient selection method is also developed which can help to reduce the number of patients to be dealt during the computation of FREM, while compromising very little with the performance of the selected miRNAs for most of the data sets. The superiority of the FREM as compared to some existing methods is demonstrated extensively on six data sets in terms of sensitivity, specificity, and F score. While for these data sets the F score of the miRNAs selected by our method varies from 0.70 to 0.91 using SVM, those results vary from 0.37 to 0.90 for some other methods. Moreover, all the selected miRNAs corroborate with the findings of biological investigations or pathway analysis tools. The source code of FREM is available at http://www.jayanta.droppages.com/FREM.html

Index Terms—miRNA, cancer, bioinformatics, fuzzy-rough entropy, information measure, rough set, fuzzy set, soft computing, feature selection, pattern recognition, histogram

1 Introduction

MicroRNAs (miRNA) are one type of non-coding RNAs which are not directly associated with protein coding. They can be found in every eukaryotic cell and some viruses [1]. MiRNAs work on messengerRNAs (mRNA) and inhibit the protein translation process [2]. By controlling the protein generation, miRNAs can regulate many functions in human body (e.g., cell division, cell death etc.). Various articles pointed out the miRNAs as one of the major biological marker of cancer in the human body [3], [4]. As a consequence different types of investigations such as detection of abnormal miRNAs in the body [4], [5], [6], and ranking of miRNAs [7] as per their relevance to any cancer are carried out by various researchers. There are several other issues such as detection of miRNA-mRNA interacting pair [8] and involvement of miRNAs in drug resistance [9]. Among all these investigations selection of cancer specific miRNAs has an important role on cancer research as irrelevant miRNAs decrease classification accuracy and increase both the biological and computational costs.

- *J.K. Pal is with the Center for Soft Computing Research, Indian Statistical Institute, Kolkata, West Bengal 700108, India. E-mail: jkp_it08@yahoo.com.*
- *S.S. Ray and S.K. Pal are with the Center for Soft Computing Research, Machine Intelligence Unit, Indian Statistical Institute, Kolkata, West Bengal 700108, India. E-mail: {shubhra, sankar}@isical.ac.in.*
- *S.B. Cho is with the Department of Computer Science, Yonsei University, Seoul 03722, South Korea. E-mail: sbcho@yonsei.ac.kr.*

Manuscript received 3 May 2016; revised 25 Oct. 2016; accepted 26 Oct. 2016. Date of publication 1 Nov. 2016; date of current version 30 Mar. 2018. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TCBB.2016.2623605

Before going into the details of technical issues in miRNA ranking/selection let us discuss about the generation of miRNA expression and the process of miRNA creation. In the methodology of expression generation [10] each miRNA is treated as a microbead which is marked by a color (fluorescent dye) code. The amount of a particular miRNA can be scanned as the intensity value of the color and it is stored as the expression of that miRNA.

At the first step of miRNA creation, primary miRNA transcripts (pri-miRNA) of length $\sim$ 1000 nucleotide (nt) are generated from the miRNA genes in the nucleolus [1], [11]. Then the pri-miRNA gets cleaved by enzyme Drosha and its partner DGCR8/Pasha, resulting in precursor miRNA (pre-miRNA) of length $\sim$ 60-100 nt. This pre-miRNA is transported from nucleus to cytoplasm through the pores of the nuclear membrane with the help of two proteins viz., RanGTP and exprotin-5. Pre-miRNAs are then cleaved by Dicer enzyme and $\sim$22 nt mature miRNA duplex are generated. The miRNA duplex contains a guide strand and a passenger strand. The passenger strand degrades and the guide strand generates simplex mature miRNA.

The ranking and selection of miRNAs can be performed by several methodologies based on the expression values and their corresponding labels. Here the existing algorithms for gene selection [12], [13], [14], [15] can also be used. In [7] miRNAs are ranked by using fold change of expressions. In that investigation paired samples are used and miRNAs are ranked by considering their common deregulation in multiple cancers. In paired samples, both the normal and cancer

tissues are collected from the same patient. So, this method is unsuitable for miRNA selection using unpaired samples and it needs to be modified in such a way that it can handle unpaired samples, as well. In [7], those miRNAs are given importance which are globally deregulated in eight types of cancer. However, there are some miRNAs, which are very important for a particular cancer only.

In [16], hypothesis test based algorithms are applied to select initial subset of miRNA and then fold change of each miRNAs is checked (if fold change is greater than 2) and 51 miRNAs are selected as abnormally expressed miRNAs.

Apart from the fold change base methodologies two other important techniques viz., classifier based method [17] and information theory based measure [18] are successfully used in gene ranking and are also useful in miRNA ranking. A SVM classifier based recursive feature elimination technique (SVMRFE) is developed for gene ranking in [17]. In the investigation 'Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy' (MRMR) [18], mutual information between genes and corresponding class labels are used to determine the relevance of genes. In the same investigation (MRMR) redundancy is calculated by computing mutual information between two gens. Combination of SVMRFE and MRMR are used to rank genes in [19] and the method is based on the tradeoff between the ranking of these two methods. In [20], three algorithms are developed for gene ranking by utilizing the expression overlapping between different classes. Among those three methods one is a parametric method and is based on prior probabilities of classes and class conditional probabilities of different samples. The other two are non-parametric version of the first method, where there is no need to consider sample distributions for the ranking process. A null space based gene selection method is available in [21]. Here, the utility of feature extraction using null linear discriminant analysis is extended to make it useful for gene selection. In [22], Correlation-based Feature Selection (CFS) is integrated with Particle Swarm Optimization to select important genes from microarray gene expression data. A centroid based method is described in [23] for selecting relevant genes. The gene selection task is designed as a L1-regularized optimization problem, based on linear discriminant analysis criterion. The centroids of different classes are computed using a kernel-based approach to determine the between-class separability and within-class compactness of the classes. In [24] a feature selection method is developed by utilizing the structural relationships among the features. The features are initially divided into several groups and then from each group one representative feature is selected. The selection problem is initially formulated as a binary constrained optimization problem and later transformed into a convex optimization problem. It is then solved using a block coordinate gradient descent optimization algorithm, useful for high dimensional feature selection. A hybrid algorithm for gene selection is presented in [25] by combining cellular learning automata (CLA) with ant colony optimization (ACO). First, the number of geans is reduced by using Fisher criterion, and then CLA and ACO are applied to find out the set of genes with high classification accuracy. In [26], the utility of some interaction based feature selection methods are demonstrated in the context of high dimensional data. In this investigation the open challenges to study with big data is also discussed. A nonparametric ReDiscovery Curve (RDCurve) based method is developed in [27] to study the stability of various ranking methods.

The miRNA ranking task deals with a set of expression values and a set of labels (normal or cancer). So, set theoretic approach such as information measures between two sets (i.e., expression values and labels) can be a right approach to handle this task. Note that, the normal and cancer expressions of a miRNA can overlap with each other and expressions of two different miRNAs can also overlap. In this regard fuzzy set theoretic approach can be useful to handle both the cases. The investigation in [28], shows the methodology to compute mutual information, V information and χ^2 information in fuzzy domain. These information measures are then used to rank the genes according to the highest relevance and lowest redundancy. In the remaining part of the article fuzzy mutual information, fuzzy V information and fuzzy χ^2 information base ranking methods are called as FMI, FVI and FCI, respectively.

Identification of important miRNAs can be performed by two methodologies, selection of most relevant group and ranking of miRNAs according to their relevance to any cancer. The two different approaches may lead to different outcomes in terms of the set of important miRNAs. In the former one the miRNAs will be selected automatically as per their performance as a group and in the latter one those will be ranked according to their individual performance. From the ranked miRNAs some of the top ones are selected as per the users choice.

A method for selecting most relevant group of miRNAs is developed in one of our previous studies viz., FMIMS [29] and a ranking method is proposed in the present investigation. In FMIMS, initially miRNAs are ranked on the basis of higher interclass distance and lower intraclass distance between normal and cancer class. Then the miRNAs are divided into several groups by using a SVM based method and the most relevant group among them is selected by using fuzzy mutual information.

In this article we propose a fuzzy-rough entropy measure (FREM) to rank miRNAs according to their relevance to normal and cancer classes. We considered both the overlap of expressions between classes and class sizes for computing the entropy. The computation mainly involves four steps (for each miRNA), viz., (i) computing membership values of the expressions corresponding to normal and cancer classes, (ii) defining lower and upper approximations for normal and cancer classes, (iii) computing relative frequencies of lower approximate and overlapping regions (see Section 2.3.1), and (iv) calculating the entropy of the miRNA using relative frequencies. Rank the miRNAs in ascending order according to the entropy value and select a subset from the top ranked ones. From the selected subset redundant miRNAs can be removed (redundancy removal) according to the user's need. While, the incorporation of fuzzy membership values in the entropy is useful for handling the overlapping nature of miRNA expressions in normal and cancer group, the application of lower and upper approximations of rough set helps to determine the exactness in the size of the groups. Further, we have extended our study to select two subsets from the normal

and cancer patients, respectively, for each miRNA by two histograms. Those selected patients are used to rank miRNAs by the proposed FREM. The patient selection process provides comparable results with those obtained by considering all patients. The novelty of our study lies in developing: 1) a fuzzy-rough entropy measure to rank the miRNAs by considering their overlapping expressions of normal & cancer patients and the class sizes and, 2) a histogram based method to select a subset of patients from a large set, whenever required. In the first case, the concept of fuzzy rough entropy is used in miRNA ranking for the first time, and it is also designed in a unique way using the relative frequency. In the second case, selection of a certain percentage of patients from each bin of histogram using miRNA expressions is itself a new concept.

The rest of the article is organized as follows. A brief description of fuzzy-rough set and the details of the proposed methodology are provided in Section 2. The experimental results using various data sets and classifiers are reported in Section 3. In Section 4 the issues and motivation behind the development of our algorithm and the biological relevance of selected miRNAs are discussed. Finally, Section 5 concludes this investigation.

2 Methods

In this section we will discuss the proposed methods in detail. Before the description of our methods the concept of fuzzy-rough set and a way to calculate the fuzzy membership value are described in brief as these are used in the proposed methodology.

2.1 Fuzzy-Rough Set

Consider U as the universal set and X as a set inside U. In the universal set, granules of elements are formed by indiscernibility relation R (a relation which creates a group of similar objects with respect to that relation) [28], [30] and the family of all such granules/groups obtained by R is represented as U/R. The indiscernibility among the elements of U results in an inexact or rough definition of set X (rough set) where X is defined by two exactly definable sets R-lower approximation ($\underline{R}X$) and R-upper approximation ($\overline{R}X$) of X, inside the universal set U, as

$$\underline{R}X = \bigcup \{Y \in U/R : Y \subseteq X\} \tag{1}$$

$$\overline{R}X = \bigcup \{Y \in U/R : Y \cap X \neq \phi\}. \tag{2}$$

In Eq. (1), Y is a group in U, $\underline{R}X$ refers to the union of all the groups which are a subset of X. U/R represents the granules/groups formed by the relation R in the universe U. The $\overline{R}X$ (Eq. (2)) implies the union of all the groups whose intersection with X returns a nonempty set. The set $X \subseteq U$ is represented by $\langle \underline{R}X, \overline{R}X \rangle$ in the approximation space $\langle U, R \rangle$ and the region $\left(\overline{R}X - \underline{R}X\right)$ is known as the boundary region of the set X. In these definitions when X is a crisp set and R is a fuzzy equivalence relation the pair $\langle \underline{R}X, \overline{R}X \rangle$ is called the fuzzy-rough set of X [30]. For a fuzzy-rough set the lower and upper approximations of X are represented as

$$\underline{R}X = \{(u, \underline{M}(u)) | u \in U\} \tag{3}$$

$$\overline{R}X = \{(u, \overline{M}(u)) | u \in U\}, \tag{4}$$

where u is an element a in U, $\underline{M}(u)$ & $\overline{M}(u)$ refer to the membership values of u in $\underline{R}X$ & $\overline{R}X$, respectively. Consider, $m_Y(u)$ represents the fuzzy membership of an element u in a group Y ($Y \in U/R$) and $\mu_X(u)$ implies the membership of u to the set X. For fuzzy-rough set $0 \leq m_Y(u) \leq 1$, $\sum_Y m_Y(u) = 1$ and $\mu_X \in \{0, 1\}$.

2.2 Computing Membership Value

The fuzzy membership values can be computed by any function such that the membership value of an element with respect to a group is non-decreasing with its closeness (in terms of distance) to the center of that group. The steps for computing fuzzy membership values of each element, say u_j, in a group (Y_i) are provided below. Here $i = 1, 2, \ldots, N$ & $j = 1, 2, \ldots, l$ and N & l are the total number of groups and total number of elements (miRNA expression for this investigation) in the universe U, respectively. The steps for computing membership values are as follows.

1. Determine the center of a group by computing the average of its expression values. Repeat the process for all the groups.
2. Calculate the membership values of an element (i.e., miRNA expression) in both the groups using the Eq. (5) where, u_j is the jth expression value of a miRNA and c_i as the average of the expression values in group Y_i (i.e., the center of a group). Here, the membership value of u_j in Y_i is calculated as

$$m_{Y_i}(u_j) = \frac{1}{\sum_{i'=1}^{N} [(c_i - u_j)/(c_{i'} - u_j)]^2}. \tag{5}$$

   Some properties of Eq. (5) are (i) $m_{Y_i}(u_j) \in [0, 1]$, (ii) $\sum_{i=1}^{N} m_{Y_i}(u_j) = 1$, $\forall j$, (iii) the value of $m_{Y_i}(u_j)$ increases as the closeness of u_j to c_i (center of the ith group) increases (i.e., $(c_i - u_j)^2$ decreases) and (iv) if u_j coincides with any center of a group (i.e., the distance of the sample from the center of that group is 0) then the value of $m_{Y_i}(u_j)$ for the corresponding group will be 1, and it will be 0 for the other groupes. Note that in Eq. (5), $\frac{0}{0}$ is considered as 1 as $\lim_{n \to 0} \frac{n}{n} = 1$.
3. Follow steps 1-2 for all the miRNAs.

2.3 Proposed FREM

Let us now describe our fuzzy-rough entropy measure in detail. As mentioned earlier the method of computing FREM consists of four steps for each miRNA. In the first step, fuzzy membership values of a miRNA expression for belonging to normal and cancer groups are computed. Here patient information (normal and cancer) is used to create these groups where the elements (i.e., expressions) of these groups can overlap with each other. The membership function used in our investigation is provided in Eq. (5). Note that, in our problem domain some expression value can have higher membership to the other group than their own group for any miRNA. In the second step, for each

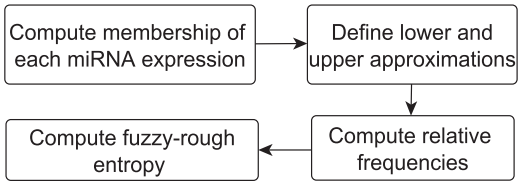


Fig. 1. Block diagram of computing FREM.

expression the membership values belonging to upper and lower approximations in both the classes (normal and cancer) are computed using their group membership values and crisp decision variables. For every element the assignment of membership to lower and upper approximations of a class/set is performed in such a way that the membership value of an element (i.e., expression) does not depend on the memberships of the other elements, as the goal is to identify those miRNAs which are capable in differentiating normal and cancer expressions in most of the cases. In the third step relative frequencies of lower approximation and overlapping region (see Section 2.3.1) of both the classes are computed using the corresponding membership values. The fourth and the final step deals with the computation of entropy using the relative frequencies. Two block diagrams for computing FREM and the ranking procedure of miRNAs using it are shown in the Figs. 1 and 2, respectively.

In a part of our investigation two subsets of patients are selected for each miRNA using two histograms which can help in reducing the number of patients to be handled in the ranking procedure. Here, the expression values of each group (normal or cancer) are divided into $\sqrt{n}$ number of bins to construct the histogram, where n is the number of patients in a group. Finally, a percentage of patients are selected from each bin and those are used for miRNA ranking. The percentage is kept fixed for all the miRNAs and the bins of the histograms. The detailed steps of our methods are discussed in the following sections.

2.3.1 Defining Lower and Upper Approximations

Let u_j be an element in the universe U (i.e., $u_j \in U$), $m_{Y_i}(u_j)$ represent the membership of u_j (varies in the range [0, 1]) in the group Y_i, and $\mu_{X_i}(u_j)$ represent the membership of u_j in class/set X_i (takes value 0 or 1) where $1 \leq i \leq N$ ($N \geq 2$) and $1 \leq j \leq l$. The variables N and l are the total number of classes/sets and elements in U, respectively. In our study the value of N is 2 as there are two classes (normal and cancer). For the indiscernibility relation R among the elements in U, compute lower ($\underline{R}X_i$) and upper ($\overline{R}X_i$) approximations of set X_i as

$$\underline{R}X_i = \{(u_j, \underline{M}_i(u_j)) | u_j \in U\}, and \tag{6}$$

$$\overline{R}X_i = \{(u_j, \overline{M}_i(u_j)) | u_j \in U\}, \tag{7}$$

$$\text{where} \quad \underline{M}_i(u_j) = \sum_{Y_i \in U/R} min(m_{Y_i}(u_j), \mu_{X_i}(u_j)), \tag{8}$$

$$\overline{M}_i(u_j) = \sum_{Y_i \in U/R} max(m_{Y_i}(u_j), \mu_{X_i}(u_j)), and, \quad j = 1, 2, \ldots, l \tag{9}$$

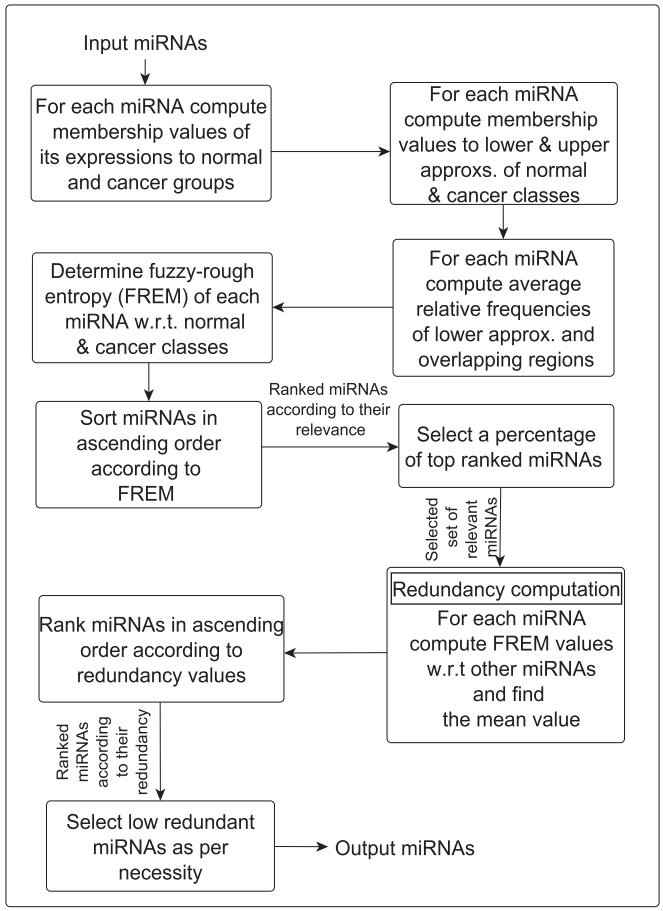


Fig. 2. Ranking and selection of miRNAs according to their relevance and redundancy using FREM.

In the proposed method $\underline{R}X$ contains only membership values of elements in group $Y_i \subseteq X_i$ (i.e., the membership values for the other elements are 0). For a particular miRNA, high membership belonging to lower bound implies that the expression values of that miRNA are correlated with the patient condition (normal or cancer). The $\overline{R}X_i$ contains the elements $Y_i \subseteq X_i$ with membership value 1 and the membership ($[0, 1]$) of the overlapping elements from other sets. Let us say, these overlapping elements create overlapping region in $\overline{R}X_i$. In the Eqs. (8) and (9) the symbol '$\sum$' refers to the fuzzy union operation.

2.3.2 Computation of Relative Frequency

The steps to determine the relative frequency of lower approximation and overlapping region are as follows.

1. Compute the cardinality of $\underline{R}X_i$ as

$$|\underline{R}X_i| = \sum_{j=1}^{l} M_i(u_j), \tag{10}$$

where $u_j \in U$. l is the total number of elements in U.

2. Compute the relative frequency of $\underline{R}X_i$ as

$$\underline{\lambda}_i = \frac{|\underline{R}X_i|}{l_i}, \tag{11}$$

where l_i is the total number of elements in $\underline{R}X_i$ with nonzero membership. Note that l_i also represents the size of class X_i.

3. Calculate the average relative frequency of lower approximations for all the classes as

$$\lambda_1 = \frac{1}{N} \times \sum_{i=1}^{N} \underline{\lambda}_i, \tag{12}$$

where N is the total number classes.

4. Compute total cardinality in the overlapping region of the universe U as

$$C = \sum_{i=1}^{N} [l_i - |\underline{R}X_i|]. \tag{13}$$

5. Calculate the average relative frequency in overlapping region (λ_2) as

$$\lambda_2 = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{l_i - |\underline{R}X_i|}{l_i} \right]. \tag{14}$$

2.3.3 Entropy Computation

The fuzzy-rough entropy measure is computed as

$$H = -\sum_{i=1}^{2} \lambda_i log_2 \lambda_i. \tag{15}$$

2.4 Ranking and Selection of miRNAs Using FREM

The block diagram for ranking and selection of miRNAs using FREM is shown in Fig. 2 and the detail steps of ranking and selection of miRNAs using FREM are presented in Table 1.

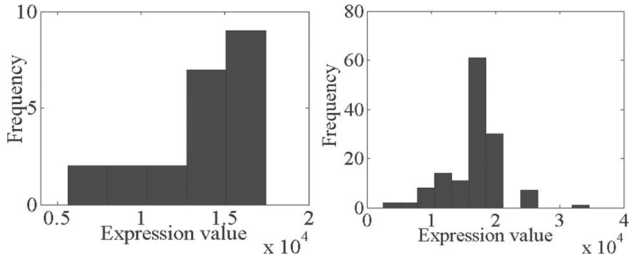
In our investigation the aim is to select the miRNAs with lower fuzzy-rough entropy value. While lower entropy in relevance computation refers to the lesser overlapping between normal and cancer expressions of a miRNA, in redundancy computation it refers to the less similarity between two miRNAs. In other words, relevant miRNAs indicate those miRNAs whose expressions are highly correlated with the class labels, whereas redundant miRNAs refer to those miRNAs which are highly correlated with one or more miRNAs. In our methodology, first the relevant miRNAs are selected to obtain the best accuracy and then redundant miRNAs are removed to reduce the biochemical and computational costs. As the redundant miRNAs are almost similar in nature, presence of all of them in general does not provide additional capability, as compared to one of them, in classifying an unknown miRNA. Further, if a certain redundant miRNA helps in the classification process then removal of it may decrease the overall accuracy rather than increasing it. So, selection of relevant miRNAs (i.e., removal of irrelevant miRNAs) increases accuracy and removal of redundancies helps to decrease biochemical and computational costs rather than increase in accuracy.

2.5 Histogram Based Patient Selection and Ranking of miRNAs (HFREM)

In a part of our investigation we have used a subset of patients instead of using all the patients during the computation of FREM. Some times it may happen that the number of patients is large for a data set. In this scenario we can use a reduced set of patients for miRNA ranking to decrease the computational costs. However it is possible for any miRNA to be expressed with wide range of values corresponding to different patients. Therefore the frequency of occurrence of expression values is important to take care during patient selection. So, consideration of the said frequency in patient selection seems to be more appropriate than excluding the patients randomly or eliminating the high or low expression values. Consequently, a histogram based method can help in this regard. Histogram is a well known method for estimating the frequency of occurrence of any numerical data and widely used in the area of image processing. In few studies [31], [32], related to the analysis of DNA and gene, it is used to visualize various biological information. In our investigation histogram is used for selecting patients by considering the frequency of occurrence of their expression values. For a particular miRNA, each group of expressions (collected either from normal or cancer patient) are divided into $\sqrt{n}$ number of bins (where n is the number of patients in each class) to construct two different histograms. Figs. 3a and 3b show such histograms for the miRNA hsa-let-7c using its 22 normal and 136 cancer patients. For normal patients the number of bins is five having patients 2, 2, 2, 7 and 9 in different bins. For cancer patients the number of bins is 12 and the numbers of patients are 2, 2, 8, 14, 11, 61, 30, 0, 7, 0, 0 and 1 in various bins. Now for the same miRNA

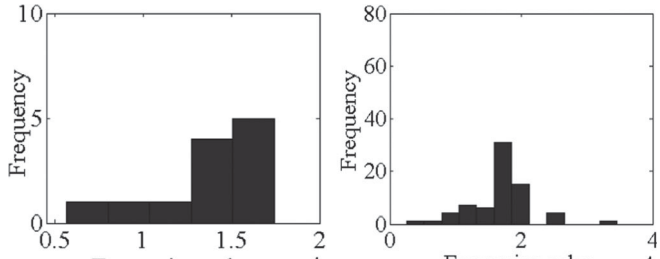
TABLE 1
Summary of the Proposed Method for Ranking of miRNAs

Step 1	For a particular miRNA compute the membership values of its expressions to normal & cancer groups using Eq. (5).
Step 2	Compute memberships to lower & upper approximate regions of normal & cancer classes of that miRNA by using Eqs. (6), (7), (8), and (9).
Step 3	Calculate average relative frequencies of lower approximate and overlapping regions of the miRNA using Eqs. (10), (11), (12), (13), and (14).
Step 4	Compute fuzzy-rough entropy of the miRNA using Eq. (15)
Step 5	Repeat Steps 1-4 for all the miRNAs and sort them in ascending order according to FREM.
Step 6	Select the desired percentage from the top ranked miRNAs and consider them as relevant ones. From them remove the redundant ones, if required, using the following steps.
Step 7	Consider two miRNAs, along with their expressions, as two different groups.
Step 8	Compute the membership (Eq. (5)) of each expression to both the groups, as formed in Step 7.
Step 9	Compute memberships of expressions to lower & upper approximate regions (using Eqs. (6), (7), (8), and (9)) of the two classes corresponding to two miRNAs..
Step 10	For a particular miRNA, compute its FREM values with respect to all others in the similar way, as mentioned in Steps 3-4.
Step 11	Compute the average of all those FREM values. This determines the redundancy of the concerned miRNA.
Step 12	Repeat Steps 7-11 for all the selected miRNAs in Step 6, and remove the ones with higher redundancy, as per the need.



(a) Normal expressions (b) Cancerous expressions

Fig. 3. Histograms of expressions for miRNA hsa-let-7c using normal and cancer patients in pancreas data.



(a) Normal expressions (b) Cancerous expressions

Fig. 4. Histograms of expressions for miRNA hsa-let-7c using normal and cancer patients in pancreas data after selecting 50 percent patients from each bin as shown in Fig. 3.

(hsa-let-7c) we select 50 percent of patients from each bin of both the histograms, (Figs. 3a and 3b) and the results are shown in Figs. 4a and 4b. It can be observed from the figures that while the number of patients in each bin is changing after the selection procedure the pattern of the histogram is remaining the same after the process. In this histogram based patient selection process the number of selected patients corresponding to each miRNA can be different. The steps for the procedure are as follows.

S1) Divide the expression values of each type of patients into $\sqrt{n}$ number of bins. Where n is the number of patients in each class (normal or cancer).
S2) Select a parentage (say p) of top expressions values, arranged in descending order, from each bin of a histogram to construct the group containing a reduced set of expressions corresponding to each miRNA.
S3) Compute FREM for each miRNA by considering two classes, i.e., normal and cancer (see Sections 2.2, 2.3, 2.3.1, 2.3.2, and 2.3.3).
S4) Select a subset of miRNAs from the whole set (see Section 2.4)

2.6 Illustrative Example

Let us take an example to show how the relevance of a miRNA can be determined by computing FREM. Consider miR1 is a miRNA and there are three normal and four cancer patients in the data set. Let us say the values of expressions corresponding to the normal samples P_1, P_2 & P_3 are 10, 9 & 11 and those of the cancer patients P'_1, P'_2, P'_3 & P'_4 are 14, 13, 20 & 17. The steps for computing FREM are given below. The computed values are rounded off up to the two decimal positions.

2.6.1 Compute Membership Values of miRNA Expressions

For a particular miRNA miR1, the fuzzy membership values of all the expressions corresponding to all the patients for normal group are computed using Eq. (5) as $\frac{1.00}{P_1}+\frac{0.87}{P_2}+\frac{0.83}{P_3}+\frac{0.37}{P'_1}+\frac{0.50}{P'_2}+\frac{0.29}{P'_3}+\frac{0.13}{P'_4}$, and for the cancer group those are $\frac{0.00}{P_1}+\frac{0.13}{P_2}+\frac{0.17}{P_3}+\frac{0.67}{P'_1}+\frac{0.50}{P'_2}+\frac{0.71}{P'_3}+\frac{0.87}{P'_4}$. Here $'+'$ indicates the fuzzy union operator.

2.6.2 Defining Lower and Upper Approximations

In this section we will compute the membership values of the expressions belonging to lower and upper approximations in both the classes. The steps are as follows:

i) As the set of crisp decision variables represent the normal and cancer classes/sets, the membership values (crisp) of the patients (P_1, P_2, P_3, P'_1, P'_2, P'_3 & P'_4) representing the normal class are $\frac{1}{P_1}+\frac{1}{P_2}+\frac{1}{P_3}+\frac{0}{P'_1}+\frac{0}{P'_2}+\frac{0}{P'_3}+\frac{0}{P'_4}$ and for the cancer class those values are $\frac{0}{P_1}+\frac{0}{P_2}+\frac{0}{P_3}+\frac{1}{P'_1}+\frac{1}{P'_2}+\frac{1}{P'_3}+\frac{1}{P'_4}$.

ii) Using fuzzy-rough set, the membership values of the expressions (corresponding to different patients) belonging to lower and upper approximation (see Eqs. (6), (7), (8), and (9)) of normal class are calculated as

$$min\left\{\left(\frac{1.00}{P_1}+\frac{0.87}{P_2}+\frac{0.83}{P_3}+\frac{0.37}{P'_1}+\frac{0.50}{P'_2}+\frac{0.29}{P'_3}+\frac{0.13}{P'_4}\right),\left(\frac{1.00}{P_1}+\frac{1.00}{P_2}+\frac{1.00}{P_3}+\frac{0.00}{P'_1}+\frac{0.00}{P'_2}+\frac{0.00}{P'_3}+\frac{0.00}{P'_4}\right)\right\}=\frac{1.00}{P_1}+\frac{0.87}{P_2}+\frac{0.83}{P_3}+\frac{0.00}{P'_1}+\frac{0.00}{P'_2}+\frac{0.00}{P'_3}+\frac{0.00}{P'_4},$$

and

$$max\left\{\left(\frac{1.00}{P_1}+\frac{0.87}{P_2}+\frac{0.83}{P_3}+\frac{0.37}{P'_1}+\frac{0.50}{P'_2}+\frac{0.29}{P'_3}+\frac{0.13}{P'_4}\right),\left(\frac{1.00}{P_1}+\frac{1.00}{P_2}+\frac{1.00}{P_3}+\frac{0.00}{P'_1}+\frac{0.00}{P'_2}+\frac{0.00}{P'_3}+\frac{0.00}{P'_4}\right)\right\}=\frac{1.00}{P_1}+\frac{1.00}{P_2}+\frac{1.00}{P_3}+\frac{0.37}{P'_1}+\frac{0.50}{P'_2}+\frac{0.29}{P'_3}+\frac{0.13}{P'_4},$$

respectively.

iii) Similarly, for cancer class all the membership values of the patients belonging to lower approximation are $\frac{0}{P_1}+\frac{0}{P_2}+\frac{0}{P_3}+\frac{0.67}{P'_1}+\frac{0.50}{P'_2}+\frac{0.71}{P'_3}+\frac{0.87}{P'_4}$ and those values belonging to upper approximation are $\frac{0.00}{P_1}+\frac{0.13}{P_2}+\frac{0.17}{P_3}+\frac{1}{P'_1}+\frac{1}{P'_2}+\frac{1}{P'_3}+\frac{1}{P'_4}$.

2.6.3 Calculation of Relative Frequency

The relative frequency of miR1 belonging to lower approximation (see Eq. (11)) in normal and cancer classes are $\frac{(1+.87+.83)}{3}=0.90$ and $\frac{(0.67+0.50+0.71+0.87)}{4}=0.69$, respectively. So, the average relative frequency belonging to the lower approximation is $\frac{(0.90+0.69)}{2}=0.79$ (see Eq. (12)). Similarly the average relative frequency belonging to the overlapping

region is $\frac{1}{2} \times \left(\frac{(3-(1+.87+.83))}{3} + \frac{(4-(0.67+0.50+0.71+0.87))}{4} \right) = 0.21$ (see Eq. (14)).

2.6.4 Entropy Computation

The entropy (Eq. (15)) of miR1 can be calculated as $-[0.79 \times log_2(0.79) + 0.21 \times log_2(0.21)] = 0.74$. The value is used to determine its relevance where a lower entropy indicates higher relevance. In a similar way, entropy of the other miRNAs can be computed to determine their relevance.

2.7 Properties of FREM

Here we present some basic properties of the proposed entropy measure and its components, as follows:

i) The variable λ_1 (see Eq. (12)) can take any value from greater than 0 to 1. In other words

$$\lambda_1 \in (0, 1]. \tag{16}$$

In our investigation $m_{Y_i}(u_j) > 0$ for at least one group Y_i (considering there is at least one element in the universe of discourse). So in this case $\underline{M}_i(u_j) = 0$, only when $\mu_{X_i}(u_j) = 0$ for all X_i (see Eq. (8)). This is only possible when $X_i = \phi$ for all i, which is not possible as we are not considering any empty classes. Therefore for any class X_i, $\mu_{X_i}(u_j) \neq 0$ and $l_i > 0$ for at least one value of i. So from Eqs. (8) and (10) it can be said that $|\underline{R}X_i| > 0$ for least one i. From Eq. (11) it can be observed that $\lambda_i > 0$ as $l_i > 0$ for at least one X_i and hence λ_1 (see Eq. (12)) is always greater than 0.

As mentioned earlier l_i is the number of elements in $\underline{R}X_i$ with nonzero membership, therefore $|\underline{R}X_i| \leq l_i$ (see Eq. (11)). Further, from Eq. (11) it can be said that $\lambda_i \leq \frac{1}{N}$. Hence $\lambda_1 \leq 1$ (see Eq. (12)).

ii) The summation of the two variables λ_1 (Eq. (12)) and λ_2 (Eq. (14)) results in 1

$$\lambda_1 + \lambda_2 = 1. \tag{17}$$

Combining Eqs. (11), (12) and (14) we get

$$\begin{aligned}
\lambda_1 + \lambda_2 &= \frac{1}{N} \sum_{i=1}^{N} \frac{|\underline{R}X_i|}{l_i} + \frac{1}{N} \sum_{i=1}^{N} \left[\frac{l_i - |\underline{R}X_i|}{l_i} \right] \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{|\underline{R}X_i|}{l_i} + \frac{1}{N} \sum_{i=1}^{N} \left[1 - \frac{|\underline{R}X_i|}{l_i} \right] \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{|\underline{R}X_i|}{l_i} + \frac{1}{N} \times N - \frac{1}{N} \sum_{i=1}^{N} \left[\frac{|\underline{R}X_i|}{l_i} \right] \\
&= 1.
\end{aligned}$$

iii) The value of H (see Eq. (15)) varies from 0 to 1, i.e.,

$$0 \leq H \leq 1. \tag{18}$$

In our problem

$$H = -[\lambda_1 log_2(\lambda_1) + \lambda_2 log_2(\lambda_2)]. \tag{19}$$

Here $\lambda_1 + \lambda_2 = 1$, therefore according to the theory of entropy $0 \leq H \leq 1$.

TABLE 2
Summary of the Data Sets

Cancer Type	*Total No. of miRNAs*	*No. of Normal Patients*	*No. of Cancer Patients*
Breast	309	5	93
Colorectal	352	8	58
Lung	866	19	17
Melanoma	864	22	35
Pancreas	847	22	136
Nasopharyngeal	887	19	31

3 Experimental Results

Six data sets, viz., breast [4], colorectal [5], lung [6], melanoma [16], pancreas [33] and nasopharyngeal [34] are used to test the performance of FREM. The summary of the used data sets is presented in Table 2. From the table it can be observed that the breast, colorectal, lung, melanoma, pancreas cancer and nasopharyngeal carcinoma data sets consist of 98 (five normal and 93 cancer), 66 (eight normal and 58 cancer), 36 (19 normal and 17 cancer), 57 (22 normal and 35 cancer), 158 (22 normal and 136 cancer) & 50 (19 normal and 31 cancer) samples (patients) and 309, 352, 866, 866, 847 and 887 miRNAs, respectively.

The performance of the miRNAs selected by different methods is evaluated in terms of sensitivity, specificity, accuracy, F score and Mathews Correlation Coefficient (MCC) [35] using SVM and Naive Bayes classifiers. The training and testing are preformed on the basis of leave one out cross validation principle. The measures sensitivity, specificity and F scores are defined as

$$Sensitivity = \frac{true\ positives(TP)}{true\ positives(TN) + false\ negatives(FN)}, \tag{20}$$

$$Specificity = \frac{true\ negatives(TN)}{true\ negatives(TN) + false\ positives(FN)}\ and \tag{21}$$

$$F = \frac{2 \times Sensitivity \times Specificity}{Sensitivity + Specificity}. \tag{22}$$

Here, the true positive refers to the number of correctly detected cancer miRNA expressions and false negative refers to the number of undetected cancer miRNA expressions, for a cancer sample. True negative implies the number of correctly detected normal miRNA expressions and false positive implies the wrongly detected cancer miRNA expressions (i.e., detected as cancer expressions, but actually they are normal expressions), for a normal sample. The accuracy is calculated as

$$Accuracy(\%) = \frac{No.\ of\ correctly\ classified\ samples}{Total\ samples} \times 100. \tag{23}$$

TABLE 3
F Scores of the Selected miRNAs for Various Data Sets Using SVM and Naive Bayes Classifiers

Cancer Type	Total Samples/ Patients	Total miRNAs (no. and F score)			Selected top ranked miRNAs (1%) (no. and F score)		
		No.	SVM	Naive Bayes	No.	SVM	Naive Bayes
Breast	98	309	0.60	0.69	3	0.88	0.92
Colorectal	66	352	0.61	0.66	4	0.91	0.91
Lung	36	866	0.52	0.53	9	0.82	0.83
Melanoma	57	864	0.61	0.62	9	0.89	0.89
Pancreas	158	847	0.62	0.68	8	0.91	0.94
Nasopharyngeal	50	887	0.31	0.31	9	0.70	0.72

The MCC is defined as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{24}$$

where, the TP, TN, FP and FN are the same to those defined in Eqs. (20) and (21). The value of MCC lies between -1 to +1. A MCC value greater than zero indicates that the prediction capability is better than random prediction.

3.1 Performance Evaluation

In this section we evaluate the efficacy of our method based on the aforementioned measures. As removal of redundant miRNAs is optional in the proposed methodology, we used the selected miRNAs (i.e., most relevant miRNAs) in various performance evaluation process. Table 3 shows the F score values obtained by SVM and Naive Bayes classifiers for all the input miRNAs and 1 percent of the top ranked miRNAs corresponding to different data sets. It is clear from the table that the F score values are considerably improved for the selected miRNAs as compared to the total set of miRNAs. For example using colorectal cancer data set, with 66 patients, only 4 miRNAs are selected out of 352 miRNAs which improve the F score value from 0.61 to 0.91 and 0.66 to 0.91 for SVM and Naive Bayes classifiers, respectively. Similar results are obtained for other data sets also. In addition to the F score, we evaluated the classification performance achieved by these top 1 percent miRNAs in terms of sensitivity, specificity, accuracy & MCC value. The results regarding this evaluation are reported in Table 4.

It is observed that the sensitivity, specificity and accuracy achieved by our method vary from 0.60 to 0.92, 0.81 to 0.91 and 71.33 to 91.02 percent, respectively, using SVM classifier. Similarly, using Naive Bayes classifier, these three measures (sensitivity, specificity and accuracy) vary from 0.64 to 0.99, 0.81 to 0.93 and 72.78 to 94.11 percent, respectively, depending on the different data sets. Using SVM and Naive Bayes classifiers the corresponding MCC values range from 0.45 to 0.82 and 0.46 to 0.88 for various data sets. The effectiveness of the redundancy removal technique is demonstrated on top 10 percent miRNAs, as an example. According to our method first we selected the relevant miRNAs then removed a percentage (e.g., 20 percent) of redundant miRNAs from the selected set. The results regarding the redundancy removal technique are reported in Table 5. From the table it can be observed that even after 20 percent removal of the redundant miRNAs (i.e., 80 percent are remaining) the F score value are remaining almost the same for all the data sets, using both the classifiers. For example, in colorectal cancer data set, 28 miRNAs out of 35 miRNAs are selected after redundancy removal. As seen, the F score value with the 80 percent miRNAs are changed from 0.84 to 0.83 for SVM and that is remaining the same for Naive Bayes classifier.

TABLE 5
Classification Performance After Redundancy Removal from the Selected Set of Relevant miRNAs

Cancer Type	Before Redundancy Removal			After Redundancy Removal		
	No. of miRNAs	F score		No. of miRNAs	F score	
		SVM	Naive Bayes		SVM	Naive Bayes
Breast	31	0.77	0.83	26	0.77	0.83
Colorectal	35	0.84	0.85	28	0.83	0.85
Lung	86	0.72	0.68	70	0.71	0.66
Melanoma	86	0.82	0.84	70	0.82	0.84
Pancreas	85	0.81	0.91	68	0.79	0.91
Nasopharyngeal	89	0.54	0.67	71	0.52	0.68

3.2 Results Using HFREM

The results using our histogram based patient selection and ranking of miRNAs using HFREM are reported here. The related curves (using SVM classifier) are shown in the Fig. 5. We varied the percentage of selected patients in each bin (see Section 2.5) from 10 to 100 percent in steps of 10. For most of the data sets the F-scores are close to the best value when the number of selected patients from each bin of the histogram is 60 percent of the bin size or more, except for breast cancer where the F score is close to the best near 90 percent of the bin size. For example, in Nasopharyngeal

TABLE 4
Classification Performance of the Selected miRNAs for Various Data Sets Using SVM and Naive Bayes Classifiers

Classifier	Performance	Breast	Colorectal	Lung	Melanoma	Pancreas	Nasopharyngeal
SVM	Sensitivity	0.85	0.91	0.87	0.89	0.92	0.60
	Specificity	0.90	0.91	0.81	0.88	0.90	0.83
	Accuracy (%)	87.74	90.78	83.81	88.79	91.02	71.33
	MCC	0.75	0.81	0.68	0.77	0.82	0.45
Naive Bayes	Sensitivity	0.99	0.96	0.86	0.92	0.95	0.64
	Specificity	0.85	0.87	0.79	0.85	0.93	0.81
	Accuracy(%)	92.50	91.59	82.90	88.70	94.11	72.78
	MCC	0.86	0.83	0.66	0.78	0.88	0.46

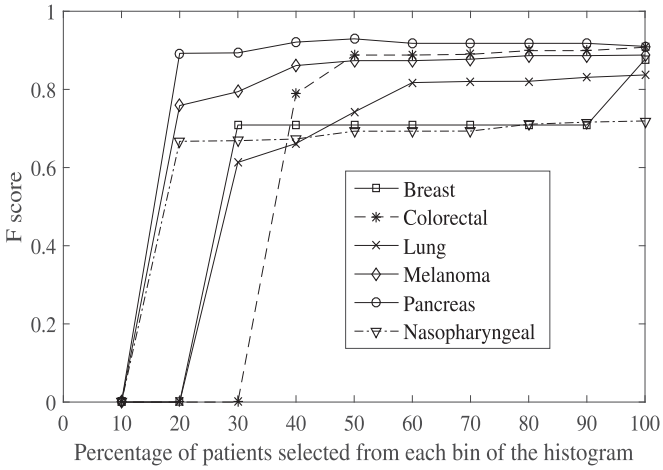


Fig. 5. Testing the performance of HFREM by selecting variable percentage of patients from the bins of the histograms.

data the F score increases from 0.67 to 0.70 when the percentage of selected patients is increased from 60 to 100 percent from each bin of the two histograms (normal & cancer).

3.3 Comparison with Other Approaches

In this section we compare the performance of our method with several well known methods for gene selection. The methods considered are SVMRFE [17], MRMR [18], SVMRFE with MRMR [19] and the method in [28] using fuzzy mutual information (FMI), fuzzy V information (FVI) and fuzzy χ^2 information (FCI) measures.

3.3.1 Comparison Using Variable Number of miRNAs

The performance of FREM with that of related algorithms is compared by varying the number of selected miRNAs. Here the percentage of top miRNAs is varied from 1 to 50. Comparison is made in terms of classification capability (F score) of the selected miRNAs. The experimental results using SVM are shown in Figs. 6a, 6b, 6c, 6d, 6e, and 6f. We have choosen 1, 5 and 10 percent of top ranked miRNAs, initially, and then increased the percentage in steps of 10. It can be observed from the figures that our algorithm performs best with 1 percent of the top ranked miRNAs as compared to higher percentage of miRNAs. Our algorithm also outperforms other methods with 1 percent of the top ranked miRNAs. For example, in the case of breast cancer data set FREM achieves F score 0.88 whereas the next highest F score (0.77) is achieved by FMI.

3.3.2 Comparison of Sensitivity, Specificity, and F Score

In addition to the comparisons in terms of F score, as shown in Section 3.3.1 we also compared FREM with related methods in terms of sensitivity and specificity using top 1 percent of the ranked miRNAs and the results are reported in Table 6. The best results are marked by bold font in the table.

It is observed that for FREM sensitivity varies from 0.60 (Nasopharyngeal) to 0.92 (Pancreas), specificity ranges from 0.81 (Lung) to 0.91 (Colorectal) and F score varies between 0.70 (Nasopharyngeal) and 0.91 (Colorectal & Pancreas) for different data sets. Further, FREM performs the best in terms of sensitivity and specificity for all the data sets except for the nasopharyngeal data set in the case of sensitivity. While for nasopharyngeal data set the sensitivity for FREM is 0.60 as compared to 0.85 of SVMRFE, the specificity value corresponding to SVMRFE is 0.36 which is lower than the performance of a random prediction. In contrast, the specificity value of FREM using this data set is 0.83 which is the highest and much superior to that of SVMRFE. The second highest specificity for Nasopharyngeal data set is jointly achieved by FMI and FCI (0.64) which is also considerably less than that of FREM.

3.3.3 Comparison with Some Recent Methods

The proposed FREM is also compared with some recent investigations such as the methods (Overlap-prob technique, Overlap technique & Overlap2 technique) in the article [20], null space based feature selection (NSBFS) [21], correlation & particle swarm optimization (PSO) [22] and consistency based ranking & interact [26]. In the investigation in [26], it is shown that consistency-based ranking & interact method perform better than many algorithms. We evaluated all the mentioned methods in terms of sensitivity, specificity and F score using SVM classifier and the results are reported in Table 7. It is observed from the table that FREM performs the best in terms of sensitivity, specificity and F score for all the cases except (i) for breast data where Overlap-prob, Overlap, techniques show the best performance in terms of specificity, (ii) lung data where Overlap-prob, Overlap and Overlap2 techniques perform the best in terms of sensitivity, and (iii) for nasopharyngeal cancer where Overlap technique performs the best in terms of specificity than FREM. In other words, in 15 out of 18 cases (3 measures $\times$ 6 data sets $\times$ 1 classifier) our method shows superior performance to the other methods. For each of the measures and data sets the best result obtained by any method is marked by bold font. Overlap technique is found as the second best method which performs superior in 4 out of 18 cases where, one of those results is jointly best with FREM.

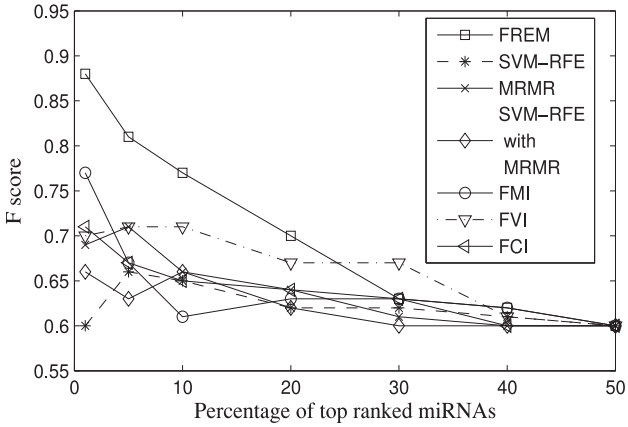
3.3.4 Cross Validation by Excluding the Test Sample During Ranking

In a part of our experiment we performed miRNA ranking by only using the training samples, i.e, we performed leave one out cross validation by removing the test sample from the set of samples used for miRNA ranking. This procedure helps in preventing the selection bias which could lead to underestimated error rates.

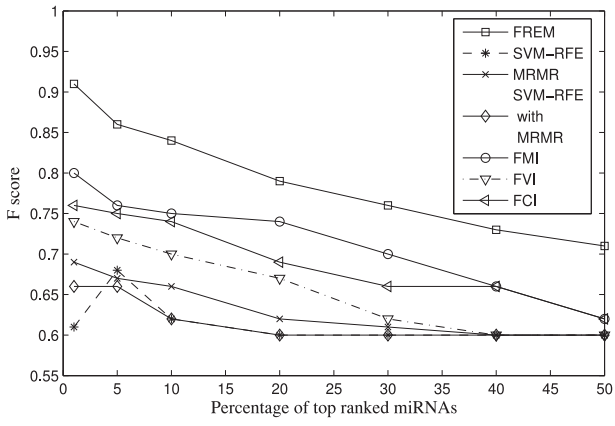
The cross validation procedure is used on top two selection methods i.e., FREM and overlap technique. The results corresponding to this experiment are reported in Table 8. From the table it is clear that the classification performance of FREM is the best in all the cases except for specificity in lung and nasopharyngeal cancer data. Further, by comparing the results of Table 8 with those of Table 7 it is observed that for FREM and Overlap the values of different measures in Table 8 are decreased by amounts of 0.06 and 0.26, respectively, on an average.

4 Discussion

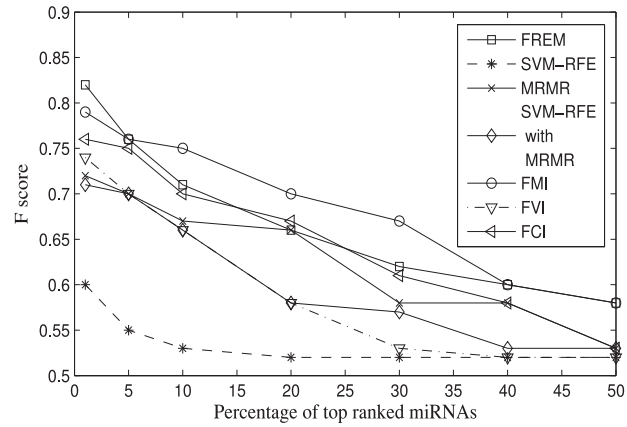
As mentioned earlier, the objective of the investigation is to rank the miRNAs as per their relevance to a particular cancer. The method is based on fuzzy rough entropy measure. In our



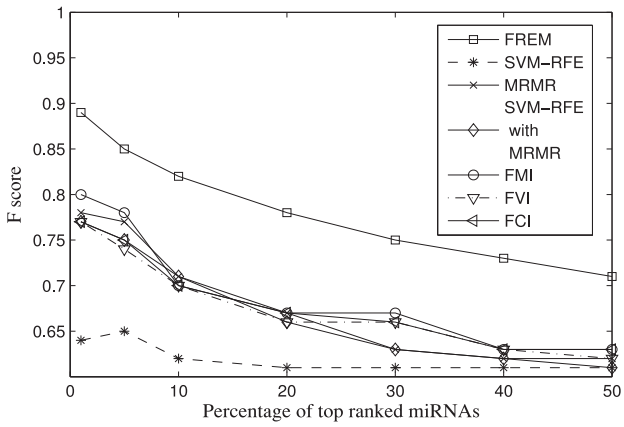
(a) Comparison using Breast cancer



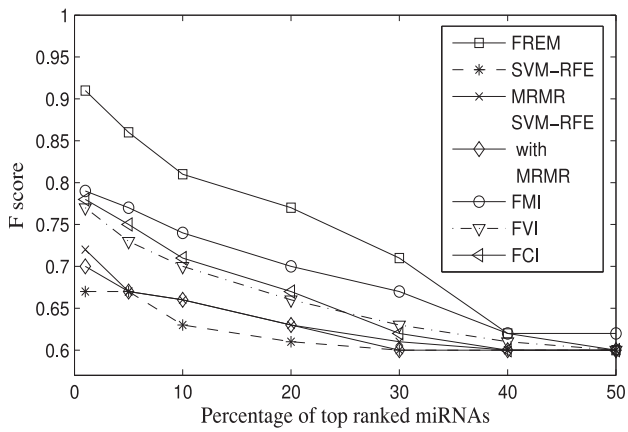
(b) Comparison using Colorectal cancer



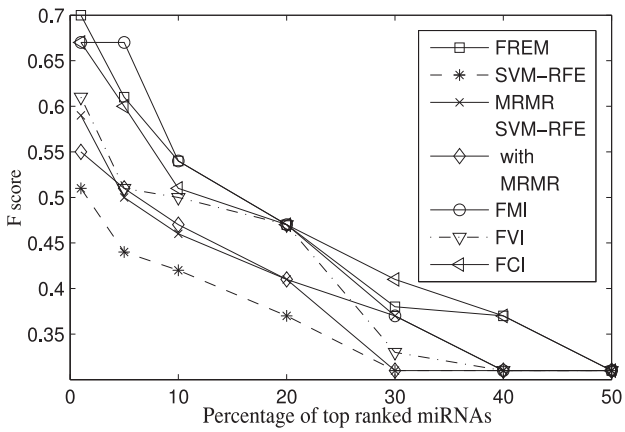
(c) Comparison using Lung cancer



(d) Comparison using Melanoma



(e) Comparison using Pancreas cancer



(f) Comparison using Nasopharyngeal carcinoma

Fig. 6. Comparison among various selection methods in terms of F score for different percentages of miRNAs using SVM classifier.

problem domain all the patients have a particular class label. Therefore, the decision variable corresponding to any patient can be represented by the membership value either 1 or 0 with respect to a particular class. So the set of decision variables turns out to be a crisp set and two crisp sets/class (normal and cancer) construct the universal set. As we do not know which miRNAs are responsible for cancer, labels of miRNAs are unknown. However, we can use patient information (normal and cancer) to create two groups of expressions corresponding to each miRNAs. The elements (i.e., miRNA expressions) of these two groups can overlap with each other and can have fuzzy membership (varying between [0, 1]). As the decision variables are crisp and the expression values constitute fuzzy groups, fuzzy-rough set theoretic approach would be more appropriate than only crisp or fuzzy approach to handle the problem of miRNA ranking.

In information theory, entropy measure is used to calculate the uncertainty of an element belonging to a class. In our problem domain, it can be utilized for checking the uncertainty caused by a miRNA in belonging to normal or cancer class. However, entropy measure in crisp domain cannot handle the uncertainty arising from overlapping

TABLE 6
Comparison of Sensitivity, Specificity, and F Score Achieved by Some Existing miRNA/Gene Selection Methods Using SVM Classifier

Method	Measures	Breast	Colorectal	Lung	Melanoma	Pancreas	Naso-pharyngeal
FREM	Sensitivity	**0.85**	**0.91**	**0.87**	**0.89**	**0.92**	0.60
	Specificity	**0.90**	**0.91**	**0.81**	**0.88**	**0.90**	**0.83**
	F Score	**0.88**	**0.91**	**0.83**	**0.89**	**0.91**	**0.70**
SVMRFE	Sensitivity	0.60	0.67	0.60	0.66	0.65	**0.85**
	Specificity	0.60	0.56	0.60	0.62	0.70	0.36
	F Score	0.60	0.61	0.60	0.64	0.67	0.51
MRMR	Sensitivity	0.73	0.70	0.73	0.80	0.69	0.74
	Specificity	0.65	0.68	0.72	0.76	0.74	0.50
	F Score	0.69	0.69	0.72	0.78	0.72	0.59
SVMRFE with MRMR	Sensitivity	0.59	0.66	0.71	0.78	0.69	0.72
	Specificity	0.75	0.66	0.71	0.76	0.71	0.50
	F Score	0.66	0.66	0.71	0.77	0.70	0.55
FMI	Sensitivity	0.75	0.77	0.82	0.88	0.76	0.70
	Specificity	0.80	0.83	0.77	0.72	0.82	0.64
	F Score	0.77	0.80	0.79	0.80	0.79	0.67
FVI	Sensitivity	0.65	0.71	0.77	0.82	0.72	0.61
	Specificity	0.77	0.78	0.71	0.72	0.82	0.61
	F Score	0.70	0.74	0.74	0.77	0.77	0.61
FCI	Sensitivity	0.67	0.74	0.77	0.82	0.76	0.70
	Specificity	0.75	0.78	0.74	0.72	0.79	0.64
	F Score	0.71	0.76	0.76	0.77	0.78	0.67

The best results are marked by bold font.

class boundaries which can be tackled by incorporation of fuzziness in the entropy measure. Moreover incorporation of lower and upper approximation of rough set theory would further help in determining the exactness in the size of the class. The miRNAs are sorted in ascending order according to their FREM value and the miRNA with the lowest value is considered as the best one. After ranking a percentage from the top of the list is selected for further operations.

TABLE 7
Comparison of Sensitivity, Specificity, and F Score with Some Recent Investigations Using SVM Classifier

Method	Measures	Breast	Colorectal	Lung	Melanoma	Pancreas	Naso-pharyngeal
FREM	Sensitivity	**0.85**	**0.91**	0.87	**0.89**	**0.92**	**0.60**
	Specificity	0.90	**0.91**	**0.81**	**0.88**	**0.90**	0.83
	F Score	**0.88**	**0.91**	**0.83**	**0.89**	**0.91**	**0.70**
Overlap-prob technique	Sensitivity	0.75	0.80	**0.99**	0.57	0.91	0.53
	Specificity	**0.99**	0.87	0.23	0.66	0.89	0.75
	F Score	0.85	0.83	0.38	0.61	0.90	0.62
Overlap technique	Sensitivity	0.76	0.87	**0.99**	0.82	0.90	0.53
	Specificity	**0.99**	0.90	0.23	0.86	**0.90**	**0.85**
	F Score	0.86	0.88	0.37	0.84	0.90	0.66
Overlap2 technique	Sensitivity	0.82	0.89	**0.99**	0.82	0.90	0.58
	Specificity	0.70	0.91	0.23	0.86	**0.90**	0.82
	F Score	0.75	0.90	0.38	0.84	0.90	0.68
NSBFS	Sensitivity	0.65	0.62	0.59	0.66	0.73	0.44
	Specificity	0.40	0.71	0.67	0.58	0.78	0.74
	F Score	0.49	0.67	0.63	0.62	0.76	0.56
PSO	Sensitivity	0.82	0.85	0.80	0.81	0.86	0.47
	Specificity	0.93	0.90	0.71	0.82	0.84	0.78
	F Score	0.86	0.87	0.75	0.81	0.85	0.58
Consistency based ranking	Sensitivity	0.58	0.53	0.78	0.60	0.50	0.33
	Specificity	0.75	0.59	0.61	0.63	0.52	0.71
	F Score	0.66	0.56	0.69	0.62	0.51	0.45
Interact	Sensitivity	0.80	0.76	0.56	0.83	0.86	0.47
	Specificity	0.65	0.81	0.52	0.80	0.89	0.80
	F Score	0.72	0.78	0.54	0.81	0.87	0.59

The best results are marked by bold font.

TABLE 8
Classification Performance of the Selected miRNAs Using SVM

Data set	Performance	Methods	
		FREM	Overlap technique
Breast	Sensitivity	**0.85**	0.82
	Specificity	**0.80**	**0.80**
	F score	**0.83**	0.81
Colorectal	Sensitivity	**0.89**	0.25
	Specificity	**0.78**	0.75
	F score	**0.83**	0.37
Lung	Sensitivity	**0.75**	0.20
	Specificity	0.71	**0.76**
	F score	**0.73**	0.32
Melanoma	Sensitivity	**0.85**	0.50
	Specificity	**0.81**	0.40
	F score	**0.83**	0.44
Pancreas	Sensitivity	**0.90**	0.33
	Specificity	**0.87**	0.64
	F score	**0.89**	0.44
Nasopharyngeal	Sensitivity	**0.54**	0.20
	Specificity	0.75	**0.92**
	F score	**0.63**	0.33

The selection is performed using the training set only. The best results are marked by bold font.

After the selection of relevant miRNAs, redundant miRNAs can be removed from the selected set to reduce the cost of biochemical tests required to generate miRNA expressions for an unknown patient. In calculating redundancy we can apply the same approach of FREM by considering two different miRNAs as two classes and then computing the entropy. The average entropy of a miRNA with respect to all other is used as redundancy value of that miRNA. This principle and methodology can be used to multi-class problems as well. In a part of our investigation we developed a histogram based patient selection technique to test the performance of FREM with reduced set of patients; thereby making the histogram based fuzzy-rough entropy measure (HFREM) suitable for ranking miRNAs with large number of patients.

TABLE 9
Selected miRNAs for Different Data Sets

Breast	Colorectal	Lung
hsa-miR-193a	hsa-miR-30a-5p	hsa-let-7d hsa-miR-423-5p
hsa-miR-30d	hsa-miR-378	hsa-let-7f hsa-miR-140-3p hsa-miR-98
hsa-miR-142-3p	hsa-miR-195 hsa-miR-422a	hsa-miR-195 hsa-miR-126 hsa-miR-20b hsa-let-7e
Melanoma	**Pancreas**	**Nasopharyngeal**
hsa-miR-17	hsa-miR-200c	hsa-miR-638
hsa-miR-664	hsa-miR-30c	hsa-miR-762
hsa-miR-145	hsa-miR-181b	hsa-miR-1915
hsa-miR-422a	hsa-miR-30b	hsa-miR-135a
hsa-miR-216a	hsa-miR-130a	hsa-miR-1275
hsa-miR-186	hsa-miR-216b	hsa-miR-940
hsa-miR-1301	hsa-miR-148a	hsa-miR-572
hsa-miR-328	hsa-miR-130b	hsa-miR-29c
hsa-let-7d		hsa-miR-548q

TABLE 10
Pathway Analysis of Selected miRNAs Using DIANA/Starbase

Cancer Type	miRNA	Target genes	p value related to gene targeting
Breast	hsa-miR-193a	PTK2, ETS1, ETV6, KRAS, PLAUR, CCND1 , PLAU	0.001
	hsa-miR-30d	GJA1,IRS1,RARB, CSNK1A1, RUNX2	0.002
	hsa-miR-142-3p	CASK, CCDC6, GHR, KDM6A, PCGF3, ZBTB10, RHOBTB3, JMJD1C	0.0002
Colorectal	hsa-miR-30a-5p	B3GNT5, B4GALT6, CPSF6, PHTF2	0.001
	hsa-miR-378	MAPK1, TCF7L2, PIK3R3	0.006
	hsa-miR-195	BTRC, CCNE1, CDS2, FBXW7, GOLGA1, KIF21A, PHF19, PURA, RAB11FIP2	2.17×10^{-9}
	hsa-miR-422a	CLDN12, RNF20	0.090
Lung	hsa-let-7d	COL1A1, LRIG2, LRIG3	0.001
	hsa-let-7f	CCND1	0.001
	hsa-miR-140-3p	E2F1, E2F2, MYC, CDKN1B, CDKN2B, LAMC1, BIRC3, TRAF1	0.002
	hsa-miR-98	E2F1, E2F2, MYC, CDKN1B, CDKN2B, LAMC1, BIRC3, TRAF1	0.001
	hsa-miR-195	CDK6, CCND1, E2F3	3.13×10^{-07}
	hsa-miR-126	E2F1, PIK3R2, CCNE2	0.0009
	hsa-miR-20b	BMP2, BMPR2, DMTF1, HIF1A, INTS6, MMP2, CDKN1A, CTNNB1, TIMP2	0.0002
	hsa-let-7e	E2F2, COL4A2, BCL2L1, TP53, COL4A6, COL4A1	0.04
Melanoma	hsa-miR-17	APP, JAK1, CDKN1A	0.020
	hsa-miR-664	FZD4, FZD5, TCF4, WNT7A, ADCY7, PRKCA, NRAS, KITLG, WNT9B, GNAI1, PRKACB	0.020
	hsa-miR-145	IGF1R, CDKN1A	0.080
	hsa-miR-422a	E2F2, PTEN, IGF1R	0.030
	hsa-miR-216a	PTEN	0.008
	hsa-miR-186	WNT5A, KIT, GNAI1	0.040
	hsa-miR-1301	CDK6, PIK3R1	0.030
	hsa-miR-328	IGF1R, FGF11	0.080
	hsa-let-7d	ITGB3, NAP1L1	0.007
Pancreas	hsa-miR-200c	TGFA, EGFR, IKBKB, PIK3CA, AKT3, E2F3	0.050
	hsa-miR-30c	RAB8A, ATP1B2, GNAQ, ATP2B1, PLA2G12A, RAB27B, RAB11A, RAB3D, ATP2A2, RAP1B, PLA2G2C	0.040
	hsa-miR-181b	BCL2, SMAD7	0.010
	hsa-miR-30b	RAB8A, RAB3D, ATP2B1, PLA2G2D, RAB11A, PLA2G12A, ATP2A2,RAB27B, ATP1A2, RAP1B, PLA2G2C, GNAQ	0.030
	hsa-miR-130a	E2F2,MAPK1, SMAD4, STAT3, TGFB2, TGFBR1, TGFBR2, RALBP1	0.005
	hsa-miR-148a	E2F3,SMAD2,TGFB2, PIK3R3,RALBP1	0.006
	hsa-miR-130b	E2F1, SMAD4, RB1, TGFBR2	0.001

We checked the involvement of the top 1 percent miRNAs (see Table 9), ranked by the proposed method, in related cancers. The relevance of the selected miRNAs to a cancer is investigated by using two publicly available pathway analysis tools, DIANA [36] and Starbase [37]. These tools are capable to find out the the target genes of a miRNA responsible for cancer development. These tools also provide a merged p-value [36] of a miRNA corresponding to its target genes. The miRNAs identified by FREM, their target genes and p-values are reported in Table 10. Using these pathway analysis tools, while 7 out of 8 miRNAs are found to be relevant for pancreas cancer data set, for lung cancer data set 8 out of 9 miRNAs are found to be relevant. Interestingly all the selected miRNAs are found as relevant ones for breast, colorectal, and melanoma data set. As an example we can consider lung cancer data set where the miRNA hsa-let-7f targets CCND1 gene (p-value 0.001) and causes uncontrolled cell division. Similar results can be observed for other data sets also.

As none of the miRNAs in nasopharyngeal carcinoma, hsa-miR-423-5p in lung cancer and hsa-miR-216b in pancreas cancer data sets is identified as involved in any pathways by pathway analysis tools, the role of these miRNAs in cancer is searched through various biological investigations. The miRNA hsa-miR-423-5p is identified as a responsible miRNA for lung cancer in the investigation in [6] and hsa-miR-216b is reported as a relevant one for pancreas cancer in [38], which corroborate with our investigation. Similarly miRNAs (out of 9) hsa-miR-548q, hsa-miR-1915, hsa-miR-572, hsa-miR-762, hsa-miR-638 and hsa-miR-135a are found as relevant miRNAs for nasopharyngeal cancer in the investigation [34]. As reported in [39] the miRNA hsa-miR-940 is responsible for nasopharyngeal cancer which controls the Nestin protein level to regulate the cell growth and death. Further, two remaining miRNAs, hsa-miR-29c and hsa-miR-1275 are reported as the important biomarkers for nasopharyngeal cancer in the investigations [40] and [41], respectively, which shows similarity with our study.

5 Conclusion

A fuzzy-rough entropy measure is proposed for ranking miRNAs with respect to their relevance to cancers. The computation of FREM consists of three steps, viz., calculating membership of the expression values belonging to the lower and upper approximations of both the classes (normal and cancer), computing relative frequency in lower approximation and overlapping region, and determining entropy. Relevance of each miRNA is determined by the FREM between its two classes, normal and cancer. Then miRNAs are sorted in the ascending order and a percentage of top miRNAs is selected from the ranked list. The redundant miRNAs are removed from the selected set according to the necessity. The selected miRNAs are found to be relevant according to two publicly available pathway analysis tools or related biological investigations.

The classification accuracy (in terms of F score) of the miRNAs selected by FREM varies from 0.70 to 0.91 (when SVM is used as the classifier) and 0.72 to 0.94 (when Naive Bayes classifier is used as the classifier) which are superior to some existing methods in most of the cases. Superiority of FREM is also observed in terms of sensitivity, specificity and F score when it is compared with several classical and

recent methods. Moreover, we tested our method by performing the ranking process only on the training samples. In other words, the test sample is kept outside from the set of samples used during ranking. Using SVM classifier sensitivity, specificity and F score values obtained by this procedure vary from 0.54 to 0.90, 0.71 to 0.87 and 0.63 to 0.89, respectively, depending on various data sets.

The histogram based patient selection method is developed to reduce the number of patients before computing relevance of the miRNAs using FREM. For most of the data sets the corresponding F scores are almost the same with those obtained after removal of 40 percent patients using our histogram based process. In the future, numerous experiments on patients, involving miRNAs, are likely to be appended to the same existing miRNAs and under that situation this technique seems to be a promising tool.

As the miRNAs selected by the FREM are also pointed out as important by pathway analysis tool or related biological investigations, biologists may use this method for prior prediction of miRNAs involved in cancer. The principle and methodology of FREM can also be applied for other diseases where, related miRNA or gene expressions are available. The method is also applicable for multiclass problems (e.g., ranking of miRNAs in terms of their relevance in normal pancreas, pancreatitis and pancreas cancer), where more than two classes are available for classification. The experimental results on multiple data sets and similarity of the findings with those of biological experiments reveal the importance of the selected miRNAs using FREM.

Acknowledgments

S. K. Pal acknowledges the J. C. Bose fellowship and the Raja Ramanna fellowship of the Govt. of India.

References

[1] S. S. Ray and S. Maiti, "Noncoding RNAs and their annotation using metagenomics algorithms," *Wiley Interdisciplinary Rev.: Data Mining Knowl. Discovery*, vol. 5, no. 1, pp. 1–20, 2015.

[2] M. A. Valencia-Sanchez, J. Liu, G. J. Hannon, and R. Parke, "Control of translation and mRNA degradation by miRNAs and siRNAs," *Genes Develop.*, vol. 20, no. 5, pp. 515–524, 2006.

[3] G. A. Calin, et al., "Frequent deletions and down-regulation of micro-RNA genes mir15 and mir16 at 13q14 in chronic lymphocytic leukemia," *Proc. Nat. Academy Sci. United States America*, vol. 99, no. 24, pp. 15 524–15 529, 2002.

[4] C. Blenkiron, et al., "MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype," *Genome Biol.*, vol. 8, no. 10, pp. R214.1–R214.16, 2007.

[5] G. M. Arndt, et al., "Characterization of global microRNA expression reveals oncogenic potential of mir-145 in metastatic colorectal cancer," *BMC Cancer*, vol. 9, no. 1, pp. 1–17, 2009.

[6] A. Keller, et al., "miRNAs in lung cancer-studying complex fingerprints in patient's blood cells by microarray experiments," *BMC Cancer*, vol. 9, no. 1, pp. 1–10, 2009.

[7] R. Navon, H. Wang, I. Steinfeld, A. Tsalenko, A. Ben-Dor, and Z. Yakhini, "Novel rank-based statistical methods reveal microRNAs with differential expression in multiple cancer types," *PLoS One*, vol. 4, no. 11, 2009, Art. no. e8003.

[8] Y. Li, C. Liang, K.-C. Wong, K. Jin, and Z. Zhang, "Inferring probabilistic miRNA–mRNA interaction signatures in cancers: A role-switch approach," *Nucleic Acids Res.*, vol. 42, no. 9, 2014, Art. no. e76.

[9] F. H. Sarkar, Y. Li, Z. Wang, D. Kong, and S. Ali, "Implication of micrornas in drug resistance for designing novel cancer therapy," *Drug Resistance Updates*, vol. 13, no. 3, pp. 57–66, 2010.

[10] P. Einat, "Methodologies for high-throughput expression profiling of microRNAs," in *MicroRNA Protocols*, S.-Y. Ying, Ed. Berlin, Germany: Springer, 2006, pp. 139–157.

[11] S. S. Ray, J. K. Pal, and S. K. Pal, "Computational approaches for identifying cancer miRNA expressions," *Gene Expression*, vol. 15, no. 5–6, pp. 243–253, 2013.

[12] S. S. Ray, A. Ganivada, and S. K. Pal, "A granular self-organizing map for clustering and gene selection in microarray data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 9, pp. 1890–1906, Sep. 2016.

[13] A. Sharma, S. Imoto, and S. Miyano, "A top-R feature selection algorithm for microarray gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 3, pp. 754–764, Nov. 2012.

[14] M. Sehhati, S. Mehridehnavi, H. Rabbani, and M. Pourhossien, "Stable gene signature selection for prediction of breast cancer recurrence using joint mutual information," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 6, pp. 1440–1447, Nov./Dec. 2015.

[15] L. Yu, Y. Han, and M. E. Berens, "Stable gene selection from microarray data via sample weighting," *IEEE/ACM Trans. Comput, Biol. Bioinf.*, vol. 9, no. 1, pp. 262–272, Jan./Feb. 2012.

[16] P. Leidinger, et al., "High-throughput miRNA profiling of human melanoma blood samples," *BMC Cancer*, vol. 10, no. 1, pp. 1–11, 2010.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, 2002.

[18] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[19] P. A. Mundra and J. C. Rajapakse, "SVM-RFE with MRMR filter for gene selection," *IEEE Trans. Nanobiosci.*, vol. 9, no. 1, pp. 31–37, Mar. 2010.

[20] A. Sharma, S. Imoto, and S. Miyano, "A between-class overlapping filter-based method for transcriptome data analysis," *J. Bioinf. Comput. Biol.*, vol. 10, no. 5, pp. 1 250 010:1–1 250 010:20, 2012.

[21] A. Sharma, S. Imoto, S. Miyano, and V. Sharma, "Null space based feature selection method for gene expression data," *Int. J. Mach. Learn. Cybern.*, vol. 3, no. 4, pp. 269–276, 2012.

[22] A. Chinnaswamy and R. Srinivasan, "Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data," in *Proc. 6th Int. Conf. Innovations Bio-Inspired Comput. Appl.*, 2016, pp. 229–239.

[23] S. Guo, D. Guo, L. Chen, and Q. Jiang, "A centroid-based gene selection method for microarray data classification," *J. Theoretical Biol.*, vol. 400, pp. 269–276, 2016.

[24] M. F. Ghalwash, X. H. Cao, I. Stojkovic, and Z. Obradovic, "Structured feature selection using coordinate descent optimization," *BMC Bioinf.*, vol. 17, no. 1, pp. 1–14, 2016.

[25] F. V. Sharbaf, S. Mosafer, and M. H. Moattar, "A hybrid gene selection approach formicroarray data classification using cellular learning automata and ant colony optimization," *Genomics*, vol. 107, no. 6, pp. 231–238, 2016.

[26] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Feature selection for high-dimensional data," *Progress Artif. Intell.*, vol. 5, no. 2, pp. 65–75, 2016.

[27] X. Lu, A. Gamst, and R. Xu, "RDCurve: A nonparametric method to evaluate the stability of ranking procedures," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 7, no. 4, pp. 719–726, Oct.-Dec. 2010.

[28] P. Maji and S. K. Pal, "Fuzzy-rough sets for information measures and selection of relevant genes from microarray data," *IEEE Trans. Syst. Man Cybern.-Part B: Cybern.*, vol. 40, no. 3, pp. 741–752, Jun. 2010.

[29] J. K. Pal, S. S. Ray, and S. K. Pal, "Identifying relevant group of miRNAs in cancer using fuzzy mutual information," *Med. Biol. Eng. Comput.*, vol. 54, no. 4, pp. 701–710, 2016.

[30] D. Sen and S. K. Pal, "Generalized rough sets, entropy, and image ambiguity measures," *IEEE Trans. Syst. Man Cybern.-Part B: Cybern.*, vol. 39, no. 1, pp. 117–128, Feb. 2009.

[31] A. M. Costa, J. T. Machado, and M. D. Quelhas, "Histogram-based dna analysis for the visualization of chromosome, genome and species information," *Bioinf.*, vol. 27, no. 9, pp. 1207–1214, 2011.

[32] W. Pope, et al., "Differential gene expression in glioblastoma defined by ADC histogram analysis: Relationship to extracellular matrix molecules and survival," *Amer. J. Neuroradiology*, vol. 33, no. 6, pp. 1059–1064, 2012.

[33] A. S. Bauer, et al., "Diagnosis of pancreatic ductal adenocarcinoma and chronic pancreatitis by measurement of microRNA abundance in blood and tissue," *PLoS One*, vol. 7, no. 4, Apr. 2012, Art. no. e34151.

[34] X.-H. Zheng, et al., "Plasma microRNA profiling in nasopharyngeal carcinoma patients reveals mir-548q and mir-483-5p as potential biomarkers," *Chinese J. Cancer*, vol. 33, no. 7, pp. 330–338, May 2014.

[35] J. K. Pal, S. S. Ray, and S. K. Pal, "A weighted threshold for detection of cancerous miRNA expressions," *Fundamenta Informaticae*, vol. 127, no. 1–4, pp. 289–305, 2013.
[36] I. S. Vlachos, et al., "Diana miRPath v.2.0: Investigating the combinatorial effect of micrornas in pathways," *Nucleic Acids Res.(Web Server Issue)*, vol. 40, pp. W498–504, May 2012.
[37] J.-H. Li, S. Liu, H. Zhou, L.-H. Qu, and J.-H. Yang, "Starbase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale clip-seq data," *Nucleic Acids Res. (Database Issue)*, vol. 14, pp. D92–D97, Dec. 2014.
[38] M.-Z. Ma, et al., "Candidate microRNA biomarkers of pancreatic ductal adenocarcinoma: Meta-analysis, experimental validation and clinical significance," *J. Exp. Clinical Cancer Res.*, vol. 32, pp. 1–14, 2013.
[39] J. Ma, et al., "Depletion of intermediate filament protein nestin, a target of microRNA-940, suppresses tumorigenesis by inducing spontaneous DNA damage accumulation in human nasopharyngeal carcinoma," *Cell Death Disease*, vol. 5, no. 8, 2014, Art. no. e1377.
[40] J. Peng, et al., "Profiling miRNAs in nasopharyngeal carcinoma ffpe tissue by microarray and next generation sequencing," *Genomics Data*, vol. 2, pp. 285–289, Dec. 2014.
[41] X. Zeng, et al., "Circulating mir-17, mir-20a, mir-29c, and mir-223 combined as non-invasive biomarkers in nasopharyngeal carcinoma," *Plos One*, vol. 7, no. 10, 2012, Art. no. e46367.

Jayanta Kumar Pal received the BTech degree in information technology and the MTech degree in computer science and engineering in 2008 and 2010, respectively, from the West Bengal University of Technology, India. He is currently working as a research fellow in the Center for Soft Computing Research, Indian Statistical Institute, Kolkata. His research interests include soft computing and bioinformatics.

Shubhra Sankar Ray received the MSc degree in electronic science and the MTech degree in radio physics and electronics from the University of Calcutta, Kolkata, India, in 2000 and 2002, respectively and the PhD degree in engineering from Jadavpur University, Kolkata, in 2008. He was a post-doctoral fellow in the Saha Institute of Nuclear Physics, Kolkata, from 2008 to 2009. His current research activities include bioinformatics, granular computing, neural networks, genetic algorithms, and soft computing. Three of his publications are listed as curated paper in the Saccharomyces Genome Database, Stanford University, California. He received the Microsoft Young Faculty Award in 2010.

Sung-Bae Cho received the BS degree in computer science from Yonsei University, Seoul, Korea, and the MS and PhD degrees in computer science from the Korea Advanced Institute of Science and Technology (KAIST), Taejeon, Korea. He was an invited researcher in the Human Information Processing Research Laboratories, Advanced Telecommunications Research (ATR) Institute, Kyoto, Japan, from 1993 to 1995, and a visiting scholar with the University of New South Wales, Canberra, Australia, in 1998. He was also a visiting professor with the University of British Columbia, Vancouver, Canada, from 2005 to 2006. Since 1995, he has been a professor in the Department of Computer Science, Yonsei University. His research interests include neural networks, pattern recognition, intelligent man-machine interfaces, evolutionary computation, and artificial life. He received the outstanding paper prizes from the IEEE Korea Section in 1989 and 1992, and another one from the Korea Information Science Society in 1990. He also received the Richard E. Merwin prize from the IEEE Computer Society in 1993. He was listed in Who's Who in Pattern Recognition from the International Association for Pattern Recognition in 1994, and received the best paper awards at the International Conference on Soft Computing in 1996 and 1998. Also, he received the best paper award at the World Automation Congress in 1998, and listed in Marquis Who's Who in Science and Engineering in 2000 and in Marquis Who's Who in the World in 2001. He is a senior member of the IEEE and a member of the Korea Information Science Society, INNS, the IEEE Computational Intelligence Society, and the IEEE Systems, Man, and Cybernetics Society.

Sankar K. Pal (M'81-SM'84-F'93-LF'2015) received the PhD degree in radio physics and electronics from the University of Calcutta, in 1979, and the PhD degree in electrical engineering along with DIC from Imperial College, University of London, in 1982. He is a distinguished scientist and former director of the Indian Statistical Institute, and a former INAE chair professor. Currently, he is a DAE Raja Ramanna fellow and J. C. Fellow of the Goverment of India. He founded the Machine Intelligence Unit and the Center for Soft Computing Research: A National Facility in the Institute in Calcutta. He joined his Institute in 1975 as a CSIR senior research fellow where he became a full professor in 1987, distinguished scientist in 1998, and the director for the term 2005-2010. He worked with the University of California, Berkeley and the University of Maryland, College Park, in 1986-1987, the NASA Johnson Space Center, Houston, Texas in 1990-1992 and 1994, and the US Naval Research Laboratory, Washington DC, in 2004. Since 1997, he has been serving as a distinguished visitor of the IEEE Computer Society (USA) for the Asia-Pacific Region, and held several visiting positions in Italy, Poland, Hong Kong, and Australian universities. He is a fellow of the World Academy of Sciences (TWAS), International Association for Pattern Recognition (IAPR), International Association of Fuzzy Systems (IFSA), International Rough Set Society (IRSS), and all the four National Academies for Science/Engineering in India. He is a co-author of 19 books and more than 400 research publications such as *Pattern Recognition and Machine Learning*, *Image Processing*, *Data Mining and Web Intelligence*, Soft Computing, *Neural Nets*, *Genetic Algorithms*, *Fuzzy Sets*, *Rough Sets*, *Cognitive Machine,* and *Bioinformatics*. He visited more than 40 countries as a keynote/ invited speaker or an academic visitor. He received the 1990 S.S. Bhatnagar Prize (which is the most coveted award for a scientist in India), 2013 Padma Shri (one of the highest civilian awards) by the President of India and many prestigious awards in India and abroad including the 1999 G.D. Birla Award, 1998 Om Bhasin Award, 1993 Jawaharlal Nehru Fellowship, 2000 Khwarizmi International Award from the President of Iran, 2000-2001 FICCI Award, 1993 Vikram Sarabhai Research Award, 1993 NASA Tech Brief Award, 1994 IEEE Trans. Neural Networks Outstanding Paper Award, 1995 NASA Patent Application Award (USA), 1997 IETE-R.L. Wadhwa Gold Medal, 2001 INSA-S.H. Zaheer Medal, 2005-06 Indian Science Congress-P.C. Mahalanobis Birth Centenary Gold Medal from the Prime Minister of India for Lifetime Achievement, 2007 J.C. Bose Fellowship of the Government of India, 2013 Indian National Academy of Engineering (INAE) chair professorship, IETE Diamond Jubilee Medal 2013, IEEE Fellow Class Golden Jubilee Medal 2014, and INAE S.N. Mitra Award 2015. He is/was an associate editor of the *IEEE Transactions Pattern Analysis and Machine Intelligence* (2002-2006), the *IEEE Transactions Neural Networks* (1994-1998 and 2003-2006), *Neurocomputing* (1995-2005), the *Pattern Recognition Letters* (1993-2011), the *International Journal Pattern Recognition & Artificial Intelligence*, the *Applied Intelligence, Information Sciences*, the *Fuzzy Sets and Systems*, the *Fundamenta Informaticae*, the *LNCS Transactions Rough Sets*, the *International Journal of Computational Intelligence and Applications*, *IET Image Processing*, the *Ingeniería y Ciencia*, and the *Journal Intelligent Information Systems*; editor-in-chief of the *International Journal of Signal Processing, Image Processing and Pattern Recognition*; a book series editor of *Frontiers in Artificial Intelligence and Applications*, IOS Press, and *Statistical Science and Interdisciplinary Research*, World Scientific; a member, executive advisory editorial board of the *IEEE Transactions on Fuzzy Systems*, the *International Journal on Image and Graphics*, and the *International Journal of Approximate Reasoning*; and a guest editor of the *IEEE Computer*, the *IEEE SMC*, and the *Theoretical Computer Science*. He is a life fellow of the IEEE.