Identifying Drug Resistant miRNAs Using Entropy Based Ranking

Jayanta Kumar Pal, Shubhra Sankar Ray, and Sankar K. Pal

Abstract—MicroRNAs play an important role in controlling drug sensitivity and resistance in cancer. Identification of responsible miRNAs for drug resistance can enhance the effectiveness of treatment. A new set theoretic entropy measure (SPEM) is defined to determine the relevance and level of confidence of miRNAs in deciding their drug resistant nature. Here, a pattern is represented by a pair of values. One of them implies the degree of its belongingness (fuzzy membership) to a class and the other represents the actual class of origin (crisp membership). A measure, called granular probability, is defined that determines the confidence level of having a particular pair of membership values. The granules used to compute the said probability are formed by a histogram based method where each bin of a histogram is considered as one granule. The width and number of the bins are automatically determined by the algorithm. The set thus defined, comprising a pair of membership values and the confidence level for having them, is used for the computation of SPEM and thereby identifying the drug resistant miRNAs. The efficiency of SPEM is demonstrated extensively on six data sets. While the achieved F-score in classifying sensitive and resistant samples ranges between 0.31 & 0.50 using all the miRNAs by SVM classifier, the same score varies from 0.67 to 0.94 using only the top 1 percent drug resistant miRNAs. Superiority of the proposed method as compared to some existing ones is established in terms of F-score. The significance of the top 1 percent miRNAs in corresponding cancer is also verified by the different articles based on biological investigations. Source code of SPEM is available at http://www.jayanta.droppages.com/SPEM.html

Index Terms—miRNA, cancer, drug resistance, fuzzy set, $\mathbf{Z}^+$ number, entropy, histogram, granular probability, bioinformatics

1 Introduction

One of the important challenges in cancer research is to provide effective treatment to the patients. Despite substantial improvement of chemotherapeutic agents in last decades, drug resistance is still a challenging issue in cancer treatment [1]. Resistance towards a drug indicates the situation when the applied drug is unable to improve the condition of a patient. Like cancer development, the drug resistance is also a result of the malfunctioning of intercellular processes [1]. The main reasons of drug resistance can be pointed out as, (i) over expression of multi drug resistance transporters (i.e., not allowing the medicine to enter into the effected cells), (ii) defects in the apoptotic machinery (i.e., repairing the cell DNA after its damage by the drug), (iii) alteration of drug metabolism (i.e., reducing the efficacy of drugs), (iv) alteration in drug targets & DNA repair (i.e., enhancement in damaged DNA repairing process), and (v) disruption of redox homeostasis (i.e., imbalance between cellular oxidants and antioxidants, and thereby causing further tumor development). All these issues are the result of improper cell signalling caused by abnormal activities of MicroRNAs (miRNAs). MiRNAs are single-stranded non-coding RNAs of length ~ 20 nucleotides [2], [3] which directly works on messenger RNAs (mRNAs) and inhibit protein translation process. Several investigations [4], [5], [6] pointed out the presence of deregulated miRNAs in the patients who developed drug resistance. Medical practitioners can provide more accurate and effective treatment to the patients by identifying these miRNAs.

Several investigations reveal the role of different miRNAs in drug resistance during cancer treatment. The investigation by Kurokawa et al. [5] shows the involvement of different miRNAs in the drug resistance of colon cancer. Various biochemical procedures are used to detect and validate the miRNAs causing the resistance. Hierarchical clustering on the expressions of miRNAs, involved in drug resistance, resulted in two clusters, one for sensitive and another for resistant patients. Hummel et al. [6] examined esophageal adenocarcinoma (EAC) and esophageal squamous cell carcinoma (ESCC) patients having both the samples (i.e., drug sensitive and resistant). The samples were compared using microarray technology and quantitative real-time polymerase chain reaction (RT-PCR) and the responsible miRNAs were pointed out. Three miRNAs (miR-134, miR-487b & miR-655) causing drug resistance in lung cancer by targeting MAGI2 gene were revealed by Kitamura et al. [7]. A comparative study between drug sensitive and resistant ovarian cancer patients were performed in [4] and 11 miRNAs were identified as the drug resistant ones. Here, the drug resistant miRNAs were validated by RT-PCR method. The selected miRNAs were further analyzed by Ingenuity Pathway Analysis (IPA) tool and Kyoto Encyclopedia of Genes and Genomes (KEGG) database for identifying pathway targets and networks, respectively.

- *J.K. Pal is with the Center for Soft Computing Research, Indian Statistical Institute, Kolkata 700108, India, and also with the Department of Computer Science and Engineering, University of Calcutta, Kolkata 700073, India. E-mail: jkp_it08@yahoo.com.*
- *S.S. Ray and S.K. Pal are with the Center for Soft Computing Research and Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India. E-mail: {shubhra, sankar}@isical.ac.in.*

Manuscript received 30 Dec. 2018; revised 22 June 2019; accepted 19 July 2019. Date of publication 6 Aug. 2019; date of current version 3 June 2021. (Corresponding author: Jayanta Kumar Pal.)
Digital Object Identifier no. 10.1109/TCBB.2019.2933205

Breast cancer patients were examined in [8] and evidences were found for 27 miRNAs causing drug resistance. In that investigation miRNA microarray screening and RT-PCR were used for experiments and validations and MiR-489 was identified as the most important miRNA for the development of drug resistance in breast cancer patients.

Although plenty of experiments are already performed by biochemical methods to identify the drug resistant miRNAs, investigations based on computational prediction techniques for this purpose are yet to be developed according to the best of our knowledge. In these circumstances, computational methods designed for feature/gene selection (e.g., SVM-RFE, MRMR) can be used to identify/select drug resistant miRNAs. Next, we will discuss about some of those methods.

A method called support vector machine based recursive feature elimination (SVM-RFE) was developed in [9] to rank the genes as per their relevance. In some investigations such as "Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy" (MRMR) [10] and "SVM-RFE with MRMR filter for gene selection" [11], both the relevance and redundancy of the features/genes are considered for ranking. While, the former method uses mutual information theory, the latter one combines the methods SVM-RFE and MRMR. A gene selection algorithm was developed in [12] by extending the concept of null linear discriminant analysis. A selection method based on the interaction between features was developed by Boln-Canedo et al. [13]. Although these methods exploited various characteristics of the data sets in order to reduce the unnecessary features, they don't consider the class overlapping information during gene selection. One approach to handle the mentioned drawback was addressed in a gene selection method developed by Sharma et al. [14]. However the method fails to measure the degree of class overlapping. The problem can be solved by the concept of fuzzy set and it was adopted by Maji and Pal [15] where three information measures were defined by fuzzy set theoretic approach to determine the relevance and redundancy of the genes. Another important issue about the gene/miRNA selection is to handle the unequal class sizes. In one of our previous studies [16] we developed a method based on fuzzy rough entropy measure which utilizes the concept of rough lower approximation to handle the issue of different class sizes and also the concept of fuzzy set to exploit the class overlapping information.

From the aforesaid studies it may be realized that we are handling two attributes of a pattern (element), i.e., the feature value/s as its measurement and the class label as the information of its origin. Here class overlapping refers to the issue of same or closer feature values of two patterns having different class labels. However, these different class labels are completely distinct from each other and there is no chance of overlapping. So, the set of feature value/s can constitute fuzzy set and the set of labels can constitute crisp set. However separation of these two attributes is not possible. Literature survey shows mainly two approaches to deal with this situation. These are, (i) considering two sets, i.e., one for features (fuzzy) and another for labels (crisp) [15] and (ii) using the concept of fuzzy-rough & rough-fuzzy set [17]. In fuzzy-rough or rough-fuzzy set, an element has two different memberships, one for the granules and another for the set [16]. Therefore with the help of existing techniques we can deal with only one membership of a pattern, at a time. To overcome this problem a set is defined in this article and the concept is further extended to define an entropy measure. The main contributions of this article can be summarized as: (i) defining a set to represent an element by a pair of membership values, where one implies the belongingness of the element (fuzzy membership) to any class/set and another implies its actual class of origin (crisp membership), (ii) extending the said concept by incorporating the confidence level of having a particular pair of membership values and thereby defining set S^+, (iii) formulating granular probability measure to determine the confidence level, (iv) developing a method to create granules using histograms where the histograms are generated using class spread information and (v) finally defining an entropy based on the concept of set S^+. The newly defined entropy additionally provides the level of confidence for having the computed value of entropy. Here, the miRNAs are ranked as per the lower entropy values (i.e., higher relevance).

TABLE 1
Summary of the Data Sets

CancerType	*Total No. of miRNAs*	*No. of sensitive patients*	*No. of resistant patients*
Colon [5]	723	4	4
Esophageal [6]	847	6	12
Squamous cell [6]	847	6	12
Lung [7]	751	4	4
Ovarian [4]	664	37	28
Breast [8]	2020	5	5

The rest of the article is organized as follows. A brief description about the used data sets are presented in Section 2 Some preliminary concepts of set and the details of the developed methods are provided in Section 3. The experimental results on different data sets are reported in Section 4. Finally, Section 5 concludes this investigation.

2 Data Set

In this investigation six data sets are used to evaluate the performance of SPEM. These are colon cancer [5], esophageal adenocarcinoma [6], squamous cell carcinoma [6], lung cancer [7], ovarian cancer [4] and breast cancer [8]. All the data sets are available in gene expression omnibus for public access. A brief overview of the used data set along with their source article is provided in Table 1. Note that, in these data sets some miRNAs are used multiple times for expression generation by different expression detectors. The total number of miRNAs in Table 1 also includes these repetitions.

3 Methods

This section deals with the development of various techniques and their judicious integration for miRNA, ranking and thereby identifying drug resistant ones. Let us, first, briefly describe the concepts of crisp & fuzzy sets, fuzzy membership and Z^+ number before describing the developed methods.

i) Crisp set: Consider U as a universal set and $\chi \subset U$. χ is called crisp set if $\{\zeta \in \chi : \mu_\chi(\zeta) \in \{0,1\}\}$ where $\mu_\chi(\zeta)$ represents the membership of ζ in χ.

ii) Fuzzy set: Let U be a universal set and $\widetilde{\chi} \subset U$. $\widetilde{\chi}$ is called fuzzy set if $\{\zeta \in \widetilde{\chi} : \mu_{\widetilde{\chi}}(\zeta) \in [0,1]\}$ where $\mu_{\widetilde{\chi}}(\zeta)$ represents the membership of ζ in $\widetilde{\chi}$.

iii) Fuzzy membership: Consider $\widetilde{\chi}_i$ is a fuzzy set and ψ_i is its center, where $1 \leq i \leq \rho$ and $\rho \geq 2$ represents the total number of sets in the universal set U. In this

article the membership of an element (ζ) in $\widetilde{\chi}_i$ is represented by $\mu_{\widetilde{\chi}_i}(\zeta)$ and is defined as

$$\mu_{\widetilde{\chi}_i}(\zeta) = 1/\sum_{j=1}^{\rho}[\mathbb{D}(\psi_i,\zeta)/\mathbb{D}(\psi_j,\zeta)], \quad (1)$$

where $\mathbb{D}$ represents any function for computing distance between two elements. In our investigation we defined the mentioned function as $\mathbb{D}(a,b) = (a-b)^2$. Eq. (1) has some properties such as: a) $\mu_{\widetilde{\chi}_i}(\zeta) \in [0,1]$, b) $\sum_{i=1}^{\rho} \mu_{\widetilde{\chi}_i}(\zeta) = 1$, c) when $\mathbb{D}(\psi_i,\zeta)$ decreases, $\mu_{\widetilde{\chi}_i}(\zeta)$ increases and d) if $\mathbb{D}(\psi_i,\zeta) = 0$ then $\mu_{\widetilde{\chi}_i}(\zeta) = 1$ and $\mu_{\widetilde{\chi}_j}(\zeta) = 0,\ \forall j$, where $i \neq j$.

iv) Z^+ number: Let ξ be a random variable and $\xi \in \mathbb{R}$, where $\mathbb{R}$ represents the set of real numbers. A 'Z^+' number [18] corresponding to ξ is represented as $Z^+ = (\mu, \rho)$ where μ is a fuzzy membership function which imposes a fuzzy constraint on the values that ξ may take and ρ is the probability distribution/density of ξ.

3.1 Proposed Methods

Our goal is to identify the drug resistant miRNAs and also to determine the level of confidence in decision-making (whether drug resistant or not) with the identified miRNAs. In this regard, a measure called set S^+ based entropy (SPEM) is defined to rank the miRNAs as per their relevance. First, the patients corresponding to a miRNA are represented by a pair of memberships and level of confidence for having that particular pair. Then the entropy and the confidence level for having that entropy, for a particular miRNA, is determined using the three mentioned values corresponding to every patient. The miRNAs are ranked according to the obtained entropy value in ascending order and a portion from the top of the list is selected as the drug resistant ones. A block diagram of the method is shown in Fig. 1. Here, S^+ is developed by extending another novel concept of set called S. First the rationale behind the development of set S is stated and then the definition of the set is provided. Then the requirement of S^+ is discussed and its definition is provided. Finally the entropy computation using the concept of set S^+ is described and used for identifying drug resistant miRNAs.

Consider $\widetilde{\chi}_1$ & $\widetilde{\chi}_2$ are two fuzzy sets in the universe U, and ζ is an element ($\zeta \in U$). Say, $\mu_{\widetilde{\chi}_1}(\zeta) = \epsilon_1$ & $\mu_{\widetilde{\chi}_2}(\zeta) = \epsilon_2$ represent the membership values of ζ in $\widetilde{\chi}_1$ & $\widetilde{\chi}_2$, respectively. In classification problems, a pattern/element can have feature/s as its measurement and a label as the information of its origin. The label is always fixed for any element in the same category. However, the feature value/s corresponding to an element can show proximity with that/those of any other elements from different categories. For example, consider the problem of classifying drug sensitive and resistant patients using miRNA expressions. In this case we have two sets, one is the sensitive patients and the another is resistant patients. Here a sensitive patient (pattern/element) can have an expression value (feature) which is more closer to that of resistant patients, but still its label is always fixed (i.e., healthy). Here the mentioned characteristic of feature value/s needs to be handled by fuzzy set, whereas the labels should be dealt with crisp set. Let us assume a similar condition, i.e., ζ has a label (χ_1 or χ_2) despite having memberships in both $\widetilde{\chi}_1$ & $\widetilde{\chi}_2$. As the labels are suitable for dealing with crisp set, we denoted them by the notation of crisp set (i.e., χ_1 & χ_2 instead of $\widetilde{\chi}_1$ & $\widetilde{\chi}_2$).

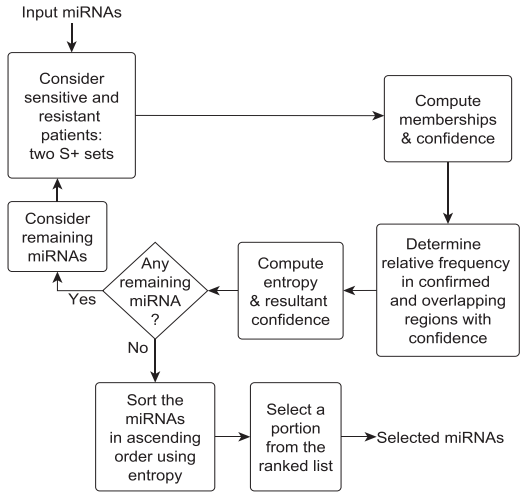


Fig. 1. Block diagram of SPEM.

Now, considering $\epsilon_1 > \epsilon_2$, there can be two instances. These are (i) $\epsilon_1 > \epsilon_2$ and the label of ζ is χ_1 and (ii) $\epsilon_1 > \epsilon_2$ and the label of ζ is χ_2. Since ϵ_1 is the membership of ζ in $\widetilde{\chi}_1$, therefore we can mark the former instance as normal case and the latter as abnormal case. Similar issues will also be there for $\epsilon_1 < \epsilon_2$. Therefore to represent the belongingness of a labeled element in a set, we need a pair of memberships. One of them represents the fuzzy membership of that element in a set and another represents the membership regarding the class label/origin of the element. So, using only fuzzy set we cannot utilize the information regarding the label or origin of an element. The situation demands a new approach to represent the membership of the labeled elements in an easy and convenient way. In this regard we propose the concept of set S and it is defined in the next section (Section 3.1.1).

3.1.1 Set S

Consider $\widehat{\chi}_i$ is a set S in the universe U where $1 \leq i \leq \rho$ and $\rho \geq 2$ represents the total number of sets in U. Say ζ_j is a labeled element ($\zeta_j \in U$ & $1 \leq j \leq \aleph$) and comprises of two variables ν_j & δ_j ($ie., \zeta_j = (\nu_j, \delta_j)$). Here $\nu \in \mathbb{R}$ (set of real numbers) represents the feature value/s of ζ, and δ_j ($\delta_j \in \{\chi_1, \chi_2, ..., \chi_i, ..., \chi_\rho\}$) represents its label. If the memberships of ν_j & δ_j in $\widehat{\chi}_i$ are represented by $\mu_{\widetilde{\chi}_i}(\nu_j)$ & $\mu_{\chi_i}(\delta_j)$, respectively, then the membership of ζ in $\widehat{\chi}_i$ can be written as

$$\mu_{\widehat{\chi}_i}(\zeta_j) = (\mu_{\widetilde{\chi}_i}(\nu_j), \mu_{\chi_i}(\delta_j)), \quad (2)$$

where, $\mu_{\widetilde{\chi}_i}(\nu_j) \in [0,1]$ and $\mu_{\chi_i}(\delta_j) \in \{0,1\}$.

3.1.2 Computation with Set S

In this section we will discuss the computation of membership values in S and also some basic operation on the set. The meaning of all the symbols used in the definition of S (Section 3.1.1) are also used in the following sections with the same meaning. The basic principle for computation with S is given as

$$f[\mu_{\widehat{\chi}_i}(\zeta_j)] = [f\{\mu_{\widetilde{\chi}_i}(\nu_j)\}, f\{\mu_{\chi_i}(\delta_j)\}], \quad (3)$$

where f represents a function.

- Determining $\mu_{\widetilde{\chi}_i}(\nu_j)$: The value of $\mu_{\widetilde{\chi}_i}(\nu_j)$ can be computed using any fuzzy membership function such as the function described in Eq. (1). The membership

$\mu_{\widetilde{\chi}_i}(\nu)$ represents the degree of belongingness of an element to any S set.

- Determining $\mu_{\chi_i}(\delta_j)$: The value of $\mu_{\chi_i}(\delta_j)$ can be defined as

$$\mu_{\chi_i}(\delta_j) = \begin{cases} 1, & \text{if } \delta_j = \chi_i \text{ (i.e., the label of } \zeta \text{ is } \chi_i) \\ 0, & \text{otherwise} \end{cases}. \tag{4}$$

Consider $\rho = 2$ and $\aleph = 1$. So, we have $\widehat{\chi_1}$ and $\widehat{\chi_2}$ as two sets (S) in the universe U and during membership computation we have to compute $\mu_{\chi_1}(\delta_1)$ and $\mu_{\chi_2}(\delta_1)$. Let us also consider, $\delta_1 = \chi_1$ (i.e., the label of ζ_1 is χ_1). Therefore, according to Eq. (4), $\mu_{\chi_1}(\delta_1) = \mu_{\chi_1}(\chi_1) = 1$ and $\mu_{\chi_2}(\delta_1) = \mu_{\chi_2}(\chi_1) = 0$.

- Cardinality: The cardinality of the set S, $\widehat{\chi}_i$, can be computed as

$$\begin{aligned} |\widehat{\chi}_i| &= \sum_{j=1}^{\aleph} \{\mu_{\widehat{\chi}_i}(\zeta_j)\} \\ &= \sum_{j=1}^{\aleph} \{\mu_{\widetilde{\chi}_i}(\nu_j), \mu_{\chi_i}(\delta_j)\}, ([\text{using Eq. 2}]) \\ &= \Big\{ \sum_{j=1}^{\aleph} \mu_{\widetilde{\chi}_i}(\nu_j), \sum_{j=1}^{\aleph} \mu_{\chi_i}(\delta_j) \Big\}, ([\text{using Eq. 3}]). \end{aligned} \tag{5}$$

In Eq. (5) $\sum_{j=1}^{\aleph} \mu_{\widetilde{\chi}_i}(\nu_j)$ represents the total belongingness of $\zeta_j(\forall j)$ in $\widehat{\chi}_i$ and $\sum_{j=1}^{\aleph} \mu_{\chi_i}(\delta_j)$ represents the size of $\widehat{\chi}_i$.

- Conditional cardinality: This measure enables us to determine the cardinality of elements in a set with particular label. It is represented as

$$|(\widehat{\chi}_i | \mu_{\chi_i}(\delta_j) = x)| = \Big\{ \sum_{\substack{1 \le j \le \aleph \\ \mu_{\chi_i}(\delta_j) = x}} \mu_{\widetilde{\chi}_i}(\nu_j), \sum_{\substack{1 \le j \le \aleph \\ \mu_{\chi_i}(\delta_j) = x}} \mu_{\chi_i}(\delta_j) \Big\}, \tag{6}$$

where, $x \in \{0, 1\}$.

In this equation $x = 1$ represents the cardinality of a set with the membership of those elements whose class label are the same as that particular class/set. The condition $x = 0$ implies the cardinality computed by the membership of the elements with the origin different from the considered calss/set. In other words, the conditional cardinality helps us to separately compute the total membership of the elements in a class/set having the same class label, and also having different class labels.

- Union: The union operation can be performed as

$$\begin{aligned} \bigcup_{i=1}^{\rho} \widehat{\chi}_i &= \max_{1 \le i \le \rho} \{\mu_{\widehat{\chi}_i}(\zeta_j)\}, \forall j \\ &= \{\max_{1 \le i \le \rho} \{\mu_{\widetilde{\chi}_i}(\nu_j)\}, \max_{1 \le i \le \rho} \{\mu_{\chi_i}(\delta_j)\}\}, \forall j. \end{aligned} \tag{7}$$

- Intersection: The intersection operation can be carried out as

$$\begin{aligned} \bigcap_{i=1}^{\rho} \widehat{\chi}_i &= \min_{1 \le i \le \rho} \{\mu_{\widehat{\chi}_i}(\zeta_j)\}, \forall j \\ &= \{\min_{1 \le i \le \rho} \{\mu_{\widetilde{\chi}_i}(\nu_j)\}, \min_{1 \le i \le \rho} \{\mu_{\chi_i}(\delta_j)\}\}, \forall j. \end{aligned} \tag{8}$$

- Subset: Let $\widehat{\Upsilon}$ be a S set in U. For any i, $\widehat{\Upsilon}$ will be called a subset of $\widehat{\chi}_i$ (i.e., $\widehat{\Upsilon} \subseteq \widehat{\chi}_i$) if they satisfy the conditions

$$\text{i) } \mu_{\widetilde{\Upsilon}}(\nu_j) \le \mu_{\widetilde{\chi}_i}(\nu_j), \ \forall \ j \text{ and} \tag{9}$$

$$\text{ii) } \mu_{\Upsilon}(\nu_j) \le \mu_{\chi_i}(\nu_j), \ \forall \ j. \tag{10}$$

3.1.3 Confidence in Decision-Making

So far we have designed a set to handle an element having feature value/s and a label. However, the pair of membership values in that approach (see Eq. (2)) do not estimate the level of confidence for having the mentioned pair. To determine the aforesaid confidence, one can compute the probability of occurrence of an element inside the set. Definitely, frequent occurrence increases the confidence level. However, in real life situations it is very rare to have multiple elements with exactly the same feature value/s. For example, in a set of drug resistant cancer patients, it is nearly impossible to get multiple patients having the same expression value. Given the circumstances we can consider the granules inside a set, formed by any similarity between the elements. The elements belonging from a bigger granule size, can be considered as more probable ones to occur in the set. In this regard we estimated the probability of occurrence of an element according to the granule size from which it belongs to. The probability estimated by the mentioned procedure is named as granular probability and its value is used as the confidence.

3.1.4 Granular Probability

Let $\mathbb{G}_k$ be a granule inside $\widehat{\chi}_i$ and η_k be the number of element/s in it, where $1 \le k \le \tau$ ($\tau < \aleph$). If $\{\zeta_j = (\nu_j, \delta_j) | \mu_{\chi_i}(\delta_j) = 1\}$ and $\zeta_j \in \mathbb{G}_k$, the granular probability (ϱ_k) of ζ_j is computed as

$$\varrho_k = \frac{\eta_k}{\sum_{j=1}^{\aleph} \mu_{\chi_i}(\delta_j)}, \tag{11}$$

where, $\varrho_k \in [0, 1]$ and $\sum_{k=1}^{\tau} \varrho_k = 1$ and the value of τ can be different for various sets. Consequently ϱ_k represents the granular probability for having $\mu_{\widehat{\chi}_i}(\zeta_j)$ corresponding to any ζ_j. As $\mu_{\widehat{\chi}_i}(\zeta_j) = (\mu_{\widetilde{\chi}_i}(\nu_j), \mu_{\chi_i}(\delta_j))$ and the granular probability implies the measure of confidence level, we can say that ϱ_k represents the confidence level of having the pair $\mu_{\widetilde{\chi}_i}(\nu_j)$ and $\mu_{\chi_i}(\delta_j)$. The level of confidence is included with the aforementioned pair in the concept of S, to make it more informative. In our method, the confidence level (i.e., granular probability) of any pair of memberships corresponding to an element is fixed for all sets, as it is computed by satisfying the condition $\mu_{\chi_i}(\delta_j) = 1$. The newly defined set is referred as S^+ and is described in detail in Section 3.1.5.

Note that ϱ_k is computed by satisfying the condition $\mu_{\chi_i}(\delta_j) = 1$ and it remains the same for any ζ_j in every $\widehat{\chi}_i^+$. We are using the term granule instead of subset for referring $\mathbb{G}_k$ as it contain only those ζ_j which fulfill the condition $\mu_{\chi_i}(\delta_j) = 1$ and also we are not considering class/set overlapping. We also developed a histogram based method, suitable for dealing with miRNA expression values, to create the granules (see Section 3.1.7).

3.1.5 Set S^+

Let $\widehat{\chi}_i^+$ be a S^+ set in the universe U where, $1 \le i \le \rho$ and $\rho \ge 2$ represents the total number of sets in U. The

membership of an element ζ_j in $\widehat{\chi}_i^+$ can be represented as

$$\mu_{\widehat{\chi}_i^+}(\zeta_j) = (\mu_{\widetilde{\chi}_i}(\nu_j), \mu_{\chi_i}(\delta_j), \varrho_k), \tag{12}$$

where $\mu_{\widetilde{\chi}_i}(\nu_j) \in [0,1]$, $\mu_{\chi_i}(\delta_j) \in \{0,1\}$ and $\varrho_k \in [0,1]$. As in set S (see Section 3.1.1), here also $\mu_{\widetilde{\chi}_i}(\nu_j)$ & $\mu_{\chi_i}(\delta_j)$ represents the memberships of ν_j & δ_j in $\widehat{\chi}_i^+$, respectively. The entity ϱ_k implies the level of confidence (Eq. (11)) of having the pair of membership values $\mu_{\widetilde{\chi}_i}(\nu_j)$ and $\mu_{\chi_i}(\delta_j)$. Here $\zeta_j \in \mathbb{G}_k$ and $\mathbb{G}_k$ is a granule inside $\widehat{\chi}_i^+$.

3.1.6 Computation with set S^+

The operations on set S^+ can be performed as

$$f[\mu_{\widehat{\chi}_i^+}(\zeta_j)] = [f\{\mu_{\widetilde{\chi}_i}(\nu_j)\}, f\{\mu_{\chi_i}(\delta_j)\}, \varrho_r], \tag{13}$$

where f and ϱ_r represent a function and the resultant granular probability (or confidence level), respectively, after any operation. The membership values $\mu_{\widetilde{\chi}_i}(\nu_j)$ and $\mu_{\chi_i}(\delta_j)$ can be computed as the same method followed in set S (see Section 3.1.2). They also follow the same properties as in the previous definitions.

- Cardinality: The cardinality of the set χ_i^+ can be computed as

$$|\widehat{\chi}_i^+| = \{|\widehat{\chi}_i|, \varrho_r\}, \tag{14}$$

where, $|\widehat{\chi}_i|$ can be computed by Eq. (5). As $|\widehat{\chi}_i|$ implies the total membership (i.e., belongingness) of the elements in a set (see Eq. (5)), we need a value to depict the overall membership of the set and thereby computing ϱ_r to represent the overall granular probability of the set. In this regard we compute relative frequency (see Eq. (16)) of $\widehat{\chi}_i^+$ to determine the value of ϱ_r. We defined three types of ϱ_r such as $\varrho_{\min}$, ϱ_{avg} and $\varrho_{\max}$ which can be used to represent the confidence of both the cardinality and relative frequency values. These three measures are suitable to be used as per the necessity and computed as

$$\varrho_r = \sum_{\substack{1\le k\le \tau \\ s\ge\epsilon}} \varrho_k, \text{ where, } s = \begin{cases} \epsilon_k^-, & \text{if } r = \min \\ \epsilon_k, & \text{if } r = avg \\ \epsilon_k^+, & \text{if } r = \max \end{cases}, \tag{15}$$

where given the conditions $\zeta_j = (\nu_j, \delta_j)$, $\mu_{\chi_i}(\delta_j) = 1$ & $\zeta_j \in \mathbb{G}_k$. The variables ϵ, ϵ_k^-, ϵ_k and ϵ_k^+ are defined as

$$\epsilon = \frac{1}{\aleph} \sum_{1\le j\le \aleph} \mu_{\widetilde{\chi}_i}(\nu_j), \tag{16}$$

$$\epsilon_k^- = \min_{1\le l\le \eta_k} \mu_{\widetilde{\chi}_i}(\nu_j), \tag{17}$$

$$\epsilon_k = \frac{1}{\eta_k} \sum_{1\le l\le \eta_k} \mu_{\widetilde{\chi}_i}(\nu_j) \text{ and} \tag{18}$$

$$\epsilon_k^+ = \max_{1\le l\le \eta_k} \mu_{\widetilde{\chi}_i}(\nu_j). \tag{19}$$

The probability $\varrho_{\max}$ should be used when we want to find out the probability of occurrence of the granule/s with at least one ζ_j having $\mu_{\widetilde{\chi}_i}(\nu_j) \ge \epsilon$. Similarly, ϱ_{avg} and $\varrho_{\min}$ are useful to find out the probability of the granule/s satisfying the condition $\epsilon_k \ge \epsilon$ and the condition $\epsilon_k^- \ge \epsilon$, respectively. After defining ϱ_r as $\varrho_{\min}$, ϱ_{avg} and $\varrho_{\max}$, we can represent $|\widehat{\chi}_i^+|$ as $|\widehat{\chi}_{i_r}^+|$ (see Eq. 20) where r implies '$\min$' or 'avg' or '$\max$' as per the necessity. In Eq. (20) $|\widehat{\chi}_i|$ is defined in Eq. (5).

$$|\widehat{\chi}_{i_r}^+| = \{|\widehat{\chi}_i|, \varrho_r\}. \tag{20}$$

- Conditional cardinality: The procedure to compute conditional cardinality is given as

$$\begin{aligned} |(\widehat{\chi}_{i_r}^+|\mu_{\chi_i}(\delta_j) = x)| &= \Big\{|(\widehat{\chi}_i|\mu_{\chi_i}(\delta_j) = x)|(\varrho_r|\mu_{\chi_i}(\delta_j) = x)\Big\} \\ &= \Big\{|(\widehat{\chi}_i|\mu_{\chi_i}(\delta_j) = x)| \sum_{1\le k\le \tau s\ge \epsilon_x} \varrho_k\Big\}, \end{aligned} \tag{21}$$

where r and s have the same meaning as in Eq. (15) and ϵ_x ($x = \{0,1\}$) can be called as conditional relative frequency and is defined as

$$\epsilon_x = \begin{cases} \frac{1}{\sum_{1\le j\le \aleph} \mu_{\chi_i}(\delta_j)} \times \sum_{1\le j\le \aleph \mu_{\chi_i}(\delta_j)=1} \mu_{\widetilde{\chi}_i}(\nu_j), & \text{if } x = 1 \\ \frac{1}{\left(\aleph - \sum_{1\le j\le \aleph} \mu_{\chi_i}(\delta_j)\right)} \times \sum_{1\le j\le \aleph \mu_{\chi_i}(\delta_j)=0} \mu_{\widetilde{\chi}_i}(\nu_j), & \text{if } x = 0 \end{cases}. \tag{22}$$

Here the '$\min$', avg and $\max$ are suitable to use for the same causes as described in ϱ_r computation. The only difference is that, here we are using ϵ_x (Eq. (22)) instead of ϵ (Eq. (16)) in determination of $(\varrho_r|\mu_{\chi_i}(\delta_j) = x)$. Conditional relative frequency also have its confidence measure which is same as that of conditional cardinality, i.e., $\varrho_r|\mu_{\chi_i}(\delta_j) = x$.

- Union: The union of different sets (S^+) can be determined as

$$\bigcup_{i=1}^{\rho} \widehat{\chi}_i^+ = \Big[\max_{1\le i\le \rho}\{\mu_{\widetilde{\chi}_i}(\nu_j)\}, \max_{1\le i\le \rho}\{\mu_{\chi_i}(\delta_j)\}, \varrho'\Big], \forall j. \tag{23}$$

In Eq. (23) the values of $\mu_{\widetilde{\chi}_i}(\nu_j)$ & $\mu_{\chi_i}(\delta_j)$ of an element in the resultant set are determined according to the rule of set S (see Eq. (7)). Here, we need to find out is the value of ϱ'. Consider $\widehat{\chi}_1^+$ & $\widehat{\chi}_2^+$ are two S^+ sets, and $\widehat{\chi}_3^+$ is their union. The values of $\mu_{\chi_3}(\delta_j)$ can be different than those of $\widehat{\chi}_1^+$ & χ_2^+. As granular probability depends on the value of $\mu_{\chi_3}(\delta_j)$ (see Section 3.1.4), ϱ' should be recomputed using Eq. (11). The rule is applicable for union operation between any two or more S^+ sets.

- Intersection: For the same reason described in union operation, ϱ' can be computed using Eq. (11) in the case of intersection and the intersection operation can be represented as

$$\bigcap_{i=1}^{\rho} \widehat{\chi}_i^+ = \Big[\min_{1\le i\le \rho}\{\mu_{\widetilde{\chi}_i}(\nu_j)\}, \min_{1\le i\le \rho}\{\mu_{\chi_i}(\delta_j)\}, \varrho'\Big], \forall j. \tag{24}$$

- Subset: If $\widehat{\Upsilon}^+$ is a subset of $\widehat{\chi}_i^+$ (i.e., $\widehat{\Upsilon}^+ \subseteq \widehat{\chi}_i^+$), then $\mu_{\widetilde{\Upsilon}}(\nu_j)$ and $\mu_{\Upsilon}(\nu_j)$ will hold the conditions in Eqs. (9) and (10), respectively. Like union and intersection operations, ϱ' has to be computed for the subset (using Eq. (11)).

3.1.7 Histogram Based Granulation Method

One of the best techniques to obtain the frequency of occurrence, of different expression values, is histogram. More frequent occurrence of any expression indicates more confidence to have that value in various patients. In this regard a histogram based method is proposed for computing granular probability in order to determine the level of confidence. In this investigation one miRNA is considered at a time for various computations. As mentioned earlier, granules in a set contain those patients (ζ_j) which hold the condition $\mu_{\chi_i}(\delta_j) = 1$ (see Section 3.1.4). So, we will get one histogram corresponding to every class/set, i.e., one for drug sensitive class and another for drug resistant class. First the histogram is generated and then each bin of the histogram is considered as one granule. The details of the method is described as follows.

S1) Determine the initial point of the histogram: The initial point (say $\mathbb{A}$) for constructing bins of the histogram is considered as the average of all the elements/patients (ζ_j) having χ_i as the class label (i.e., $\mu_{\chi_i}(\delta_j) = 1$). Here, $\mathbb{A}$ is defined as

$$\mathbb{A} = \frac{1}{\sum_{1\le j\le \aleph} \mu_{\chi_i}(\delta_j)} \times \sum_{1\le j\le \aleph \mu_{\chi_i}(\delta_j)=1} \nu_j. \tag{25}$$

Here, $\mathbb{A}$ can also be considered as the center of the class/set having label χ_i.

S2) Determine left bins: The left bins are constructed using those elements which satisfies the condition $\mathbb{A} > \nu_j$. In other words, the left bins are constructed by those elements whose value (ν_j) is less than that of the average value ($\mathbb{A}$). The detail steps of this process are as follows.

a) Initiate the starting point of the left bins (s^-) at $\mathbb{A}$ (i.e., initiate $s^- = \mathbb{A}$).

b) Determine the endpoint of the bin (e^-) as

$$e^- = s^- - \sigma^-, \tag{26}$$

where σ^- is defined as

$$\sigma^- = \lambda_{\widehat{\chi}_i^+} \times \sqrt{\left\{\frac{1}{l_1} \times \sum_{\substack{1\le j\le \aleph \\ \mathbb{A} > \nu_j}} (\mathbb{A} - \nu_j)^2\right\}}, [\forall \mu_{\chi_i}(\delta_j) = 1], \tag{27}$$

where $\lambda_{\widehat{\chi}_i^+}$ is a weight factor and is defined as

$$\lambda_{\widehat{\chi}_i^+} = \frac{\sqrt{\sum_{1\le j\le \aleph \mu_{\chi_i}(\delta_j)=0} (\mathbb{A} - \nu_j)^2 / \left(\aleph - \sum_{1\le j\le \aleph \mu_{\chi_i}(\delta_j)=1} \mu_{\chi_i}(\delta_j)\right)}}{\sqrt{\sum_{1\le j\le \aleph \mu_{\chi_i}(\delta_j)=1} (\mathbb{A} - \nu_j)^2 / \sum_{1\le j\le \aleph \mu_{\chi_i}(\delta_j)=1} \mu_{\chi_i}(\delta_j)}}. \tag{28}$$

Here (Eq. (27)), l_1 represents the number of elements in $\widehat{\chi}_i^+$ satisfying the conditions $\mathbb{A} > \nu_j$ & $\mu_{\chi_i}(\delta_j) = 1$. The numerator part of Eq. (28) determines the average deviation of the elements, having class label other than χ_i, form the center of the set/class with label χ_i. The denominator part of the equation computes the standard deviation of the elements having class label χ_i. Here $\lambda_{\widehat{\chi}_i^+} > 1$ indicates the elements with label χ_i are more closer to each other than the other elements in the universe. In other words, higher value of $\lambda_{\widehat{\chi}_i^+}$ implies better class separability. As our goal is to separate different elements with various labels, we incorporated $\lambda_{\widehat{\chi}_i^+} > 1$ during granulation.

c) Include the elements (ζ_j) inside the bin which satisfy the condition $e^- \le \nu_j < s^-$.

d) Update s^- with the value of e^- and recompute e^- using Eqs. (26) and (27) for the construction of next bins.

e) Repeat Step S2)c for the remaining elements.

f) Repeat Steps S2)d - S2)e until all ζ_j (satisfying $\mathbb{A} > \nu_j$) are included in any bin.

S3) Determining right bins: The elements satisfying the condition $\mathbb{A} < \nu_j$ are used to construct the right bins. The steps for constructing the bins are as follows.

a) Initiate the starting point(s^+) of the first bin at $\mathbb{A}$ (i.e., initiate $s^+ = \mathbb{A}$).

b) Determine the endpoint of the bin (e^+) as

$$e^+ = s^+ + \sigma^+, \tag{29}$$

where σ^+ is defined as

$$\sigma^+ = \lambda_{\widehat{\chi}_i^+} \times \sqrt{\left\{\frac{1}{l_2} \times \sum_{\substack{1\le j\le \aleph \\ \mathbb{A} < \nu_j}} (\mathbb{A} - \nu_j)^2\right\}}, [\forall \mu_{\chi_i}(\delta_j) = 1]. \tag{30}$$

In Eq. (29), l_2 represents the number of elements in $\widehat{\chi}_i^+$ satisfying the conditions $\mathbb{A} < \nu_j$ & $\mu_{\chi_i}(\delta_j) = 1$.

c) Include the elements (ζ_j) inside the bin which satisfy the condition $s^+ \le \nu_j < e^+$.

d) For constructions of next bins, update the value of s^+ with that of e^+ and recompute e^+ using Eqs. (29) and (30) to obtain the new start and end point of a bin.

e) Repeat Step S3)c for the remaining elements.

f) Repeat Steps S3)d - S3)e until all ζ_j (satisfying $\mathbb{A} < \nu_j$) are included in any bin.

3.1.8 Computation of Entropy Using Set S^+

The main aim of any classification problem is to discriminate different elements according to their label or origin. However an element with a particular label can have various memberships in different sets (see Section 3.1) due to the overlapping between sets. This causes uncertainty in decision-making and there is a need to measure the uncertainty. In this regard a method called set S^+ based Entropy Measure (SPEM) is developed to handle the aforesaid issue. In our problem, we want to determine the uncertainty in decision-making caused by the patients (ζ_j) having an expression value (ν_j) with fuzzy membership ($\mu_{\widetilde{\chi}_i}(\nu_j)$) and a label ($\delta_j$) with crisp membership ($\mu_{\chi_i}(\delta_j)$). So, the mentioned entropy is determined using these membership values to quantify the uncertainty. The veriable ϱ_k (see Section 3.1.4), which indicates the confidence of having the pair of membership values $\mu_{\widetilde{\chi}_i}(\nu_j)$ and $\mu_{\chi_i}(\delta_j)$, is used to determine the confidence level in the entropy value. Therefore the entropy measure using S^+ will provide the measure of uncertainty in decision-making and also provide the level of confidence in the measured uncertainty. The detail steps of the method are given as follows.

i) Determine the total relative frequency in the confirmed zone: In this investigation the term confirmed zone is used to indicate the memberships of different elements in their origins (i.e., the elements satisfying the condition $\mu_{\chi_i}(\delta_j) = 1$). Note that in Eq. (22), $x = 1$ (i.e., $\epsilon_x = \epsilon_1$) implies that the computation is performed using those $\mu_{\widetilde{\chi_i}}(\nu_j)$ values where the condition $\mu_{\chi_i}(\delta_j) = 1$ holds true. Therefore we can compute the total relative frequency in the confirmed zone (γ_c) using ϵ_1 as

$$\begin{aligned}\gamma_c &= \frac{1}{\rho} \times \sum_{1 \le i \le \rho} \epsilon_{\widehat{\chi_i^+}}, [\epsilon_{\widehat{\chi_i^+}} \text{ is the value of } \epsilon_1 \text{ in } \widehat{\chi}_i^+] \\ &= \frac{1}{\rho} \times \sum_{1 \le i \le \rho} \Big[\frac{1}{\sum_{1 \le j \le \aleph} \mu_{\chi_i}(\delta_j)} \times \sum_{\substack{1 \le j \le \aleph \\ \mu_{\chi_i}(\delta_j) = 1}} \mu_{\widetilde{\chi_i}}(\nu_j) \Big], \qquad (31)\end{aligned}$$

[using Eq. 22]

As mentioned earlier, the conditional relative frequency has the same confidence measures as the conditional cardinality (see Section 3.1.6). Therefore we can compute the confidence of γ_c by using the values of $(\varrho_r | \mu_{\chi_i}(\delta_j) = 1)$ (see Eq. (21)). Note that we computed γ_c by taking the average of $\epsilon_{\widehat{\chi_i^+}}$ (i.e., conditional relative frequency). So, the confidence corresponding to γ_c is also determined by the average of $(\varrho_r | \mu_{\chi_i}(\delta_j) = 1)$ using Eq. (32).

$$\varrho_c^r = \frac{1}{\rho} \times \sum_{1 \le i \le \rho} (\varrho_r | \mu_{\chi_i}(\delta_j) = 1), \qquad (32)$$

where r represents '$\min$' or 'avg' or '$\max$' according to the necessity. The three notations '$\min$' or 'avg' or '$\max$' are used for the same cause as they are used during the computation of confidence in conditional cardinality or relative frequency.

ii) Determine the total relative frequency in the overlapping zone: The overlapping zone is referred to those memberships which represents class/set overlap information. The total relative frequency in the confirmed zone is computed using $\epsilon_{\widehat{\chi_i^+}}$ (see Eq. (31)). So, total relative frequency (γ_o) in overlapping zone can be determined as

$$\begin{aligned}\gamma_o &= \frac{1}{\rho} \times \sum_{1 \le i \le \rho} (1 - \epsilon_{\widehat{\chi_i^+}}) \\ &= \frac{1}{\rho} \times \Big[\rho - \sum_{1 \le i \le \rho} \Big\{ \frac{1}{\sum_{1 \le j \le \aleph} \mu_{\chi_i}(\delta_j)} \times \sum_{1 \le j \le \aleph \mu_{\chi_i}(\delta_j) = 1} \mu_{\widetilde{\chi_i}}(\nu_j) \Big\} \Big] \\ &= 1 - \Big[\frac{1}{\rho} \times \sum_{1 \le i \le \rho} \Big\{ \frac{1}{\sum_{1 \le j \le \aleph} \mu_{\chi_i}(\delta_j)} \times \sum_{1 \le j \le \aleph \mu_{\chi_i}(\delta_j) = 1} \mu_{\widetilde{\chi_i}}(\nu_j) \Big\} \Big] \\ &= 1 - \gamma_c, [\text{using Eq. } 31]. \qquad (33)\end{aligned}$$

As mentioned earlier, the confidence level of any memberships related to an element is constant for all sets (see Section 3.1.4), the confidence level corresponding to γ_o will be the same as that of γ_c.

iii) Compute entropy: The entropy is computed as

$$H = -\big[\gamma_c \log_2(\gamma_c) + \gamma_o \log_2(\gamma_o)\big]. \qquad (34)$$

The confidence measure of H is the same as in Eq. (32).

3.1.9 Ranking Method

The steps for ranking are as follows.

S1) Consider the set of sensitive and resistant patients of a miRNA as two S^+ sets $\widehat{\chi}_1^+$ and $\widehat{\chi}_2^+$, respectively.
S2) Consider all patients as the elements (ζ_j).
S3) Determine $\mu_{\widetilde{\chi_i}}(\nu_j)$, $\mu_{\chi_i}(\delta_j)$ & ϱ_k using Eqs. (1), (4) & (11), respectively, for both $\widehat{\chi}_1^+$ and $\widehat{\chi}_2^+$.
S4) Compute total relative frequency in confirmed zone (Eq. (31)).
S5) Determine total relative frequency in overlapping zone (Eq. (33)).
S6) Compute the confidence corresponding to relative frequency values (in Steps 4 & 5) using Eq. (32) considering '$r = \max$'.
S7) Compute entropy using Eq. (34).
S8) Repeat Steps S1-S7 for all the miRNAs.
S9) Sort miRNAs in ascending order according to the entropy.
S10) Select a portion of miRNAs from the top of the list.

In ranking process we want to find out the probability of occurrence of those granules where $\mu_{\widetilde{\chi_i}}(\nu_j) | \mu_{\chi_i}(\delta_j) = 1$ has at least one value greater or equal to ϵ_1. Therefore we considered $r = \max$ (Step S6) during confidence estimation.

4 Experimental Results

We evaluated the performance of the ranking method using five performance measures such as sensitivity, specificity, accuracy, F-score and Matthews correlation coefficient (MCC). Here, leave-one-out cross validation is used for all evaluations, where, each sample is selected as the test sample once at a time and the rest $N - 1$ samples are used for training, where N is the total number of samples. Hence at the end of the test phase we obtain N number of evaluation results. Finally, the average of these results is considered as the achieved performance. Here, drug sensitive and resistant samples are considered as the negative and positive samples, respectively.

4.1 Performance Evaluation

This section deals with the performance evaluation of the top 1 percent miRNAs (see Table 2) ranked by SPEM. The improvement in classification performance (F-score) using the selected miRNAs is shown in Table 3. It is observed from the table that the F scores are considerably increased when the selected miRNAs are used instead of all of them. For example, using SVM classifier, F-score value corresponding to colon cancer increased from 0.59 to 0.88 when top 1 percent (7) miRNAs are used instead of all (723) miRNAs. Note that a miRNA with higher rank can have lower confidence than that of a lower ranked miRNA. For example, in colon cancer, hsa-miR-767-3p and hsa-miR-492 are ranked as 4th and 5th miRNAs, respectively, as per their importance in identifying drug resistance. However the confidence of the 4th ranked miRNA (0.62) is less than that of the 5th ranked miRNA (1.00).

4.2 Comparison with Other Methods

We compared SPEM with some well known existing methods using SVM classifier. The methods used in this investigation for comparison are feature selection for high-dimensional

TABLE 2
Selected miRNAs and Their Confidence Level Corresponding to Different Data Sets

Colon		Esophageal	
miRNA	Confidence	miRNA	Confidence
hsa-miR-485-5p	1.00	hsa-miR-23a	1.00
hsa-miR-645	0.87	hsa-miR-200b	1.00
hsa-miR-376a	0.87	hsa-miR-203	1.00
hsa-miR-767-3p	0.62	hsa-miR-27b	1.00
hsa-miR-492	1.00	hsa-miR-19b	0.96
hsa-miR-124	0.87	hsa-miR-200b	1.00
hsa-miR-892a	0.75	hsa-miR-455-5p	0.96
		hsa-miR-17	1.00
Lung		**Ovarian**	
miRNA	Confidence	miRNA	Confidence
hsa-miR-27a	1.00	hsa-miR-592	0.92
hsa-miR-410	0.87	hsa-miR-744	0.90
hsa-miR-572	1.00	hsa-miR-181a-2	0.82
hsa-miR-516-3p	1.00	hsa-miR-210	0.82
hsa-miR-520D-3P	0.87	hsa-miR-760	0.86
hsa-miR-99b	0.87	hsa-miR-548c-3p	0.84
hsa-miR-483-5p	0.87		
hsa-miR-27a	1.00	hsa-miR-382	0.77
Squamous cell		**Breast**	
miRNA	Confidence	miRNA	Confidence
hsa-miR-99b	1.00	hsa-miR-4264	0.80
		hsa-miR-1268a	1.00
		hsa-miR-4771	1.00
hsa-miR-18a	1.00	hsa-miR-489	1.00
		hsa-miR-4687-3p	1.00
		hsa-miR-4724-5p	0.60
hsa-miR-125a-5p	1.00	hsa-miR-1304-5p	1.00
		hsa-miR-214-3p	0.80
		hsa-miR-4637	1.00
hsa-miR-1226	1.00	hsa-miR-198	0.80
		hsa-miR-5009-3p	1.00
hsa-miR-935	1.00	hsa-miR-4467	1.00
		hsa-miR-4302	0.90
hsa-miR-146a	0.96	hsa-miR-4325	1.00
		hsa-miR-4802-5p	0.90
hsa-miR-297	0.96	hsa-miR-1228-5p	1.00
		hsa-miR-1258	0.90
hsa-miR-1269	1.00	hsa-miR-4769-5p	1.00
		hsa-miR-4261	1.00
		hsa-miR-5006-3p	0.90

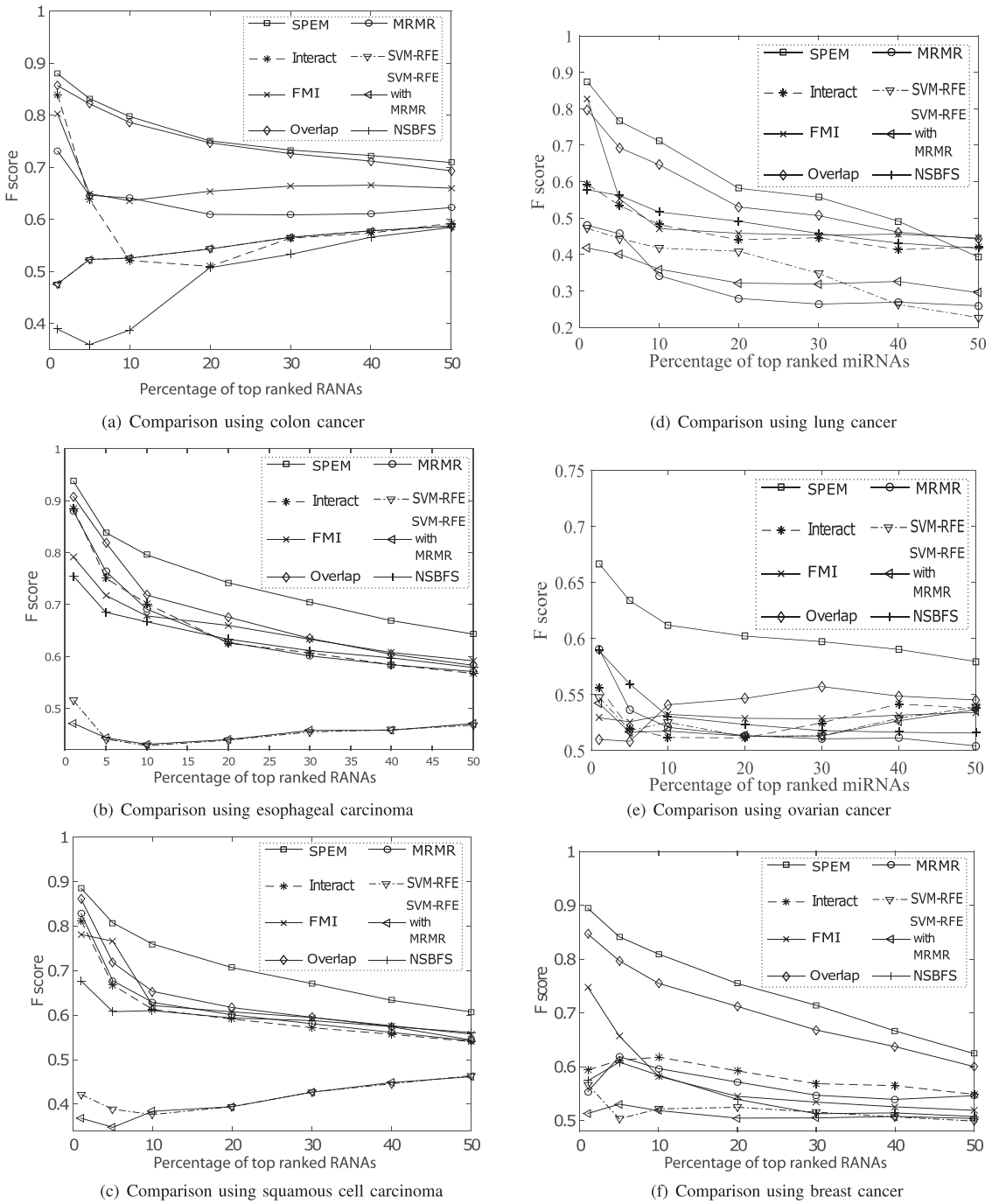
TABLE 3
F-Scores of the Selected miRNAs for Various Data Sets Using SVM and Naive Bayes Classifiers

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
Colon	8	723	0.59	0.46	7	0.88	0.82
Esophageal	18	847	0.52	0.60	8	0.94	0.95
Squamous cell	18	847	0.50	0.57	8	0.89	0.83
Lung	8	751	0.31	0.17	8	0.87	0.72
Ovarian	65	664	0.52	0.22	7	0.67	0.54
Breast	10	2020	0.50	0.45	20	0.89	0.85

TABLE 4
Comparison of Classification Performances Obtained by Top 1 Percent miRNAs Using Different Methods

Method	Measures	Colon	Esophageal	Squamous	Lung	Ovarian	Breast
SPEM	Sensitivity	**1.00**	**0.94**	0.90	**0.84**	**0.71**	**0.88**
	Specificity	**0.79**	**0.94**	**0.87**	**0.91**	**0.62**	**0.91**
	Accuracy(%)	**89.28**	**93.75**	**88.54**	**87.50**	**66.98**	**89.50**
	F-score	**0.88**	**0.94**	**0.88**	**0.87**	**0.67**	**0.89**
	MCC	**0.80**	**0.87**	**0.77**	**0.75**	**0.34**	**0.79**
Interact	Sensitivity	**1.00**	0.87	0.81	0.56	0.61	0.62
	Specificity	0.73	0.89	0.81	0.62	0.51	0.57
	Accuracy(%)	84.57	88.54	81.25	59.37	56.03	59.50
	F-score	0.84	0.88	0.81	0.59	0.56	0.59
	MCC	0.75	0.77	0.62	0.19	0.12	0.19
FMI	Sensitivity	0.82	0.79	0.78	0.81	0.55	0.71
	Specificity	**0.79**	0.79	0.60	0.84	0.51	0.79
	Accuracy(%)	80.00	79.16	69.27	82.81	52.97	75.00
	F-score	0.80	0.79	0.68	0.83	0.53	0.75
	MCC	0.61	0.58	0.39	0.66	0.06	0.50
Overlap	Sensitivity	**1.00**	0.91	**0.92**	0.81	0.51	0.83
	Specificity	0.75	0.91	0.81	0.78	0.51	0.76
	Accuracy(%)	87.50	91.79	86.45	79.68	50.99	75.19
	F-score	0.86	0.91	0.86	0.69	0.51	0.75
	MCC	0.77	0.84	0.73	0.60	0.57	0.69
MRMR	Sensitivity	0.75	0.86	0.82	0.44	0.59	0.62
	Specificity	0.71	0.90	0.83	0.53	0.59	0.50
	Accuracy(%)	73.21	88.02	82.81	48.43	59.06	56.00
	F-score	0.73	0.88	0.83	0.48	0.59	0.55
	MCC	0.46	0.76	0.66	-0.03	0.18	0.12
SVM-RFE	Sensitivity	0.54	0.53	0.52	0.41	0.55	0.61
	Specificity	0.43	0.50	0.35	0.56	0.54	0.53
	Accuracy(%)	48.21	51.56	43.75	48.43	54.77	57.00
	F-score	0.48	0.51	0.42	0.47	0.55	0.57
	MCC	-0.03	0.03	-0.13	-0.32	0.09	0.14
SVM-RFE with MRMR	Sensitivity	0.54	0.51	0.50	0.53	0.56	0.58
	Specificity	0.43	0.44	0.29	0.34	0.53	0.46
	Accuracy(%)	48.21	47.39	39.58	43.75	54.25	52.00
	F-score	0.48	0.47	0.37	0.42	0.54	0.51
	MCC	-0.03	-0.05	-0.21	-0.13	0.08	0.04
NSBFS	Sensitivity	0.43	0.74	0.69	0.56	0.58	0.52
	Specificity	0.36	0.77	0.67	0.59	0.60	0.64
	Accuracy(%)	39.28	75.52	67.70	57.81	58.94	58.00
	F-score	0.39	0.75	0.68	0.58	0.59	0.57
	MCC	-0.21	0.51	0.35	0.16	0.18	0.16

data (interact) [13], fuzzy-rough sets for information measures and selection of relevant genes from microarray data (FMI) [15], between-class overlapping filter-based method for transcriptome data analysis (overlap) [14], feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy (MRMR) [10], gene selection for cancer classification using support vector machines (SVM-RFE) [9], SVM-RFE with MRMR filter for gene selection (SVM-RFE with MRMR) [11] and null space based feature selection method for gene expression data (NSBFS) [12]. First, we compared the classification performances using the top 1 percent miRNAs obtained by SPEM and the aforesaid existing methods. The comparison is performed in terms of sensitivity, specificity, accuracy, F-score & MCC using SVM classifier, and the results are reported in Table 4. It is observed from the table that sensitivity, specificity, accuracy, F-score & MCC of our method ranges between 0.71 & 1.00, 0.62 & 0.94, 66.98 percent & 93.75 percent, 0.67 & 0.94, and 0.34 & 0.87, respectively, depending on various data sets. For all the data sets our method performs the best, except for the squamous cell carcinoma data set where the sensitivity of SPEM (0.90) is slightly less than that of overlap method [14] (0.92). Although, in squamous cell carcinoma data, overlap method performs better than SPEM in terms of sensitivity by 0.02 unit, for the other measures (i.e., specificity, F score, accuracy and MCC) overlap method performs inferior to SPEM. Moreover, the utility of SPEM can be observed by studying the percentage of its superior performance compared to the other methods using different data sets and measures. It performs better in 29 out of 30 cases



(a) Comparison using colon cancer

(b) Comparison using esophageal carcinoma

(c) Comparison using squamous cell carcinoma

(d) Comparison using lung cancer

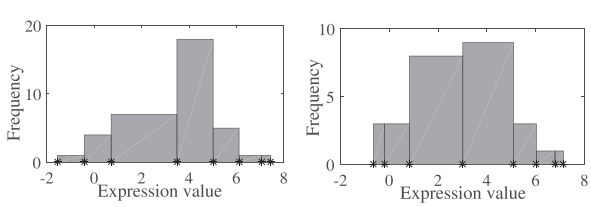
(e) Comparison using ovarian cancer

(f) Comparison using breast cancer

Fig. 2. Comparison among ranking methods in terms of F-score for different percentages of miRNAs using SVM classifier.

(5 measures $\times$ 6 data sets $\times$ 1 classifier). In other words, the developed method provides better results for 96.67 percent cases and the chance of obtaining inferior result is only 3.33 percent. In Table 4, the best results are marked by bold fonts.

Apart from the comparison provided in Table 4, we also compared the proposed SPEM with the existing methods using variable number of miRNAs. Here the percentage of miRNAs is varied from 1 to 50 by selecting from the top of the list. Comparisons are made in terms of F-score and are shown in Figs. 2a, 2b, 2c, 2d, 2e, 2f. We have chosen 1, 5 and 10 percent of top ranked miRNAs, initially, and then increased the percentage in steps of 10. It is evident from the figures that our algorithm performs the best using 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



(a) Sensitive (b) Resistant

Fig. 3. Histogram of hsa-miR-501-3p expressions corresponding to drug sensitive and resistant ovarian cancer patients.

TABLE 5
Biological Significance of the Selected miRNAs

Colon		Esophageal	
miRNA	Significance	miRNA	Significance
hsa-miR-485-5p	Drug resistance [19]	hsa-miR-23a	Drug resistance [6]
hsa-miR-645	Drug resistance [20]	hsa-miR-200b	Drug resistance [6]
hsa-miR-376a	Potential therapeutic target [21]	hsa-miR-203	Drug resistance [6]
hsa-miR-767-3p	Cancer development [22]	hsa-miR-27b	Drug resistance [23]
hsa-miR-49[2]	Drug resistance [24]	hsa-miR-19b	Drug resistance [23]
hsa-miR-124	Increasing sensitivity in radiotherapy [25]	hsa-miR-455-5p	Drug resistance [6]
hsa-miR-892a	Drug resistance [26]	hsa-miR-17	Drug resistance [6]
Lung		Ovarian	
miRNA	Significance	miRNA	Significance
hsa-miR-27a	Drug resistance [27]	hsa-miR-592	Cancer prognosis [28]
hsa-miR-410	Drug resistance [29]	hsa-miR-744	Tumor suppressor [30]
hsa-miR-572	Drug resistance [31]	hsa-miR-181a-2	Drug resistance [32]
hsa-miR-516-3p	None	hsa-miR-210	Apoptosis resistance [33]
hsa-miR-520d-3p	Radiosensitivity [34]	hsa-miR-760	Cancer aggressiveness [35]
hsa-miR-99b	Tumor suppression [36]	hsa-miR-548c-3p	Potential therapeutic target [37]
hsa-miR-483-5p	Promoting metastasis [38]	hsa-miR-382	Tumor suppression [39]
Squamous cell		Breast	
miRNA	Significance	miRNA	Significance
hsa-miR-99b	Drug resistance [6]	hsa-miR-4264	Cancer development [40]
		hsa-miR-1268a	Drug resistance [41]
		hsa-miR-4771	Cancer development [40]
hsa-miR-18a	Drug resistance [6]	hsa-miR-489	Drug resistance [42]
hsa-miR-125a-5p	Drug resistance [6]	hsa-miR-4687-3p	Potential therapeutic target [43]
hsa-miR-1226	Drug resistance [6]	hsa-miR-4724-5p	None
		hsa-miR-1304-5p	Cancer development [44]
		hsa-miR-214-3p	Drug resistance [45]
hsa-miR-935	Drug resistance [6]	hsa-miR-4637	None
		hsa-miR-198	Tumor suppression [46]
		hsa-miR-5009-3p	None
hsa-miR-146a	Drug resistance [48]	hsa-miR-4467	Drug resistance [47]
		hsa-miR-4302	Cancer development [49]
		hsa-miR-4325	None
hsa-miR-297	Cancer development [51]	hsa-miR-4802-5p	Cancer development [50]
		hsa-miR-1228-5p	Drug resistance [52]
		hsa-miR-1258	Cancer development [53]
hsa-miR-1269	None	hsa-miR-4769-5p	Potential therapeutic target [54]
		hsa-miR-4261	Cancer development [55]
		hsa-miR-5006-3p	Cancer development [56]

SEPM achieves the best F-score (0.67) in ovarian cancer data, whereas for the same data MRMR achieves the second best F-score (0.59).

4.3 Other Experiments

In this investigation two histograms (one for sensitive and another for resistant) are generated for every miRNA. Two such histograms, generated by the proposed method (see Section 3.1.7), are shown in Fig. 3a & 3b as examples.

We checked the top 1 percent miRNAs selected by our method for their involvement in related cancers. To fulfill the purpose we searched the existing biological investigations extensively, to find out the significance of the selected miRNAs and the findings are reported in Table 5. The table shows the activity of a miRNA corresponding to a particular cancer. The articles where the mentioned activities of the miRNAs are reported are also shown in the table. It is observed that many selected miRNAs using the SPEM are already identified as drug resistant miRNAs by various investigations. For example, in [19] inactivity of hsa-miR-485-5p is identified as one of the causes of drug resistance in colon cancer. Similar observations are found in other cancers as well. Some miRNAs which are not found as drug resistant ones can be considered as novel prediction for further biological investigation.

In the literature, many investigations can be found [57], [58], [59] on miRNA-Disease association prediction. In [57] a method based on inductive matrix completion is developed where the missing miRNA-disease association is predicted using the known associations and by integrating miRNA and disease similarities. Similar approaches for miRNA disease association prediction are available in [58], [59]. Investigations on individual miRNAs and their relations with a particular cancer is also available in the literature [4], [5], [6]. However, studies on miRNA-drug interaction association is scanty. So, emphasis in this domain may provide new insights in miRNA-drug-resistance association and can be considered as future research topic.

5 Conclusion

A new set is defined in this investigation to represent the membership of a pattern/element in a supervised system where a pattern/element has feature value/s as its measurement and a class label as its origin. In this system, the feature values of the elements in the same category can be different but the corresponding class labels of those elements is always the same. Moreover, the feature values of the elements from different categories can show proximity with each other but their class labels are always distinct. A set (S) is constructed for each category based on these information of the constituting elements (patterns). Here the membership of each element is represented by a pair. One among them is a fuzzy membership indicating its degree of belongings to the concerned category and another is crisp membership representing its actual class of origin. Some basic operations such as cardinality, conditional cardinality, union and intersection on S are defined. The said concept of S is further extended by including an additional information of confidence in having a particular pair of membership values in S and the extended set is named as S^+. The confidence is determined by a novel measure called granular probability. The granules required for estimating the probability are formed by a new method called histogram based granulation, where the width and number of bins corresponding to the histogram are determined automatically in the procedure. Finally the concept of S^+ set is used to compute the entropy of a miRNA, which additionally provides the level of confidence along with the measured entropy. The miRNAs are ranked as per the lowest entropy (highest relevance) and a portion from the ranked list is selected for further operations.

To evaluate SPEM we used top 1 percent of the ranked miRNAs and checked the classification performance in detecting

the presence of drug resistant miRNA using six different data sets. The classification performance (F-score) thus measured, varies from 0.67 to 0.94 using SVM and 0.54 to 0.95 using Naive Bayes classifiers. Superiority of the proposed SPEM to some existing methods is established in terms of sensitivity, specificity, accuracy, F-score and MCC. While comparing different methods, using the top 1 percent miRNAs ranked by them, we obtained the best results for 29 out of 30 cases using SPEM. We also tested our method using variable number of miRNAs and observed that the results for top 1 percent miRNAs are better than those obtained by large number of miRNAs. The method also shows superior performance to other methods for variable number of miRNAs for most of the cases.

The miRNAs identified by SPEM are verified for drug resistance characteristics by checking various articles based on biological investigations. We found that many miRNAs selected by our method directly corroborate those articles. The selected miRNAs which are not found as drug resistant ones in biological investigations can be considered for further biological validation to confirm their drug resistant characteristic.

The concept of set S and set S^+ can be useful for any supervised system. The technique for generating histogram is useful for any numerical data where estimation of frequency of occurrence is required. Similarly, the proposed granulation technique is useful where estimation of frequency of occurrence is required for creating granules. In our study the granular probability measure is designed for the patterns having one feature value. This technique can further be extended to make it suitable for handling patterns with multiple feature values. SPEM can is also applicable for other diseases where related miRNA or gene expressions are available. The method can also be used for multiclass classification problems. The experimental results on various data and similarity of the findings with the biological experiments reveal the importance of our investigation.

Acknowledgments

S. K. Pal acknowledges an INSA Distinguished Professorship.

References

[1] X. An, C. Sarmiento, T. Tan, and H. Zhub, "Regulation of multidrug resistance by microRNAs in anti-cancer therapy," *Acta Pharmaceutica Sinica B*, vol. 7, no. 1, pp. 38–51, 2017.

[2] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 4, pp. 905–915, Jul./Aug. 2017.

[3] J. Wu and Z. Zhou, "Sequence-based prediction of microRNA-binding residues in proteins using cost-sensitive laplacian support vector machines," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 3, pp. 752–759, May/Jun. 2013.

[4] A. Vecchione, et al., "A microRNA signature defines chemoresistance in ovarian cancer through modulation of angiogenesis," *Proc. Nat. Acad. Sci. United States Amer.*, vol. 110, no. 24, pp. 9845–9850, 2013.

[5] K. Kurokawa, et al., "Role of miR-19b and its target mRNAs in 5-fluorouracil resistance in colon cancer cells," *J. Gastroenterology*, vol. 47, no. 8, pp. 883–895, 2012.

[6] R. Hummel, et al., "MicroRNA signatures in chemotherapy resistant esophageal cancer cell lines," *World J. Gastroenterology*, vol. 20, no. 40, pp. 14 904–14 912, 2014.

[7] K. Kitamura, et al., "MiR-134/487b/655 cluster regulates TGF-β-induced epithelial-mesenchymal transition and drug resistance to gefitinib by targeting MAGI2 in lung adenocarcinoma cells," *J. Exp. Med.*, vol. 13, no. 2, pp. 444–453, 2014.

[8] X. Chen, et al., "Suppression of SPIN1-mediated PI3K-Akt pathway by miR-489 increases chemosensitivity in breast cancer," *J. Pathology*, vol. 239, no. 4, pp. 459–472, 2016.

[9] 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.

[10] 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.

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

[12] 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.

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

[14] 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. 1–20, 2012.

[15] 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.

[16] J. K. Pal, S. S. Ray, S. B. Cho, and S. K. Pal, "Fuzzy-rough entropy measure and histogram based patient selection for miRNA ranking in cancer," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 2, pp. 659–672, Mar./Apr. 2018.

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

[18] L. A. Zadeh, "A note on Z-numbers," *Inf. Sci.*, vol. 181, no. 14, pp. 2923–2932, 2011.

[19] W. Xiong, et al., "Microarray analysis of circular RNA expression profile associated with 5-fluorouracil-based chemoradiation resistance in colorectal cancer cells," *BioMed Res. Int.*, vol. 2017, pp. 1–8, Jun. 2017, Art. no. 8421614.

[20] S. T. Guo, et al., "Micro-RNA is an oncogenic regulator in colon cancer," *Oncogenesis*, vol. 6, no. 5, 2017, Art. no. e335.

[21] Z.-H. Mo, X.-D. Wu, S. Li, B.-Y. Fei, and B. Zhang, "Expression and clinical significance of microRNA-376a in colorectal cancer," *Asian Pacific J. Cancer Prevention*, vol. 15, no. 21, pp. 9523–9527, 2014.

[22] L. Yan, W. Zhao, H. Yu, Y. Wang, Y. Liu, and C. Xie, "A comprehensive meta-analysis of microRNAs for predicting colorectal cancer," *Med.*, vol. 95, no. 9, 2016, Art. no. e2738.

[23] K. Tanaka, et al., "miR-27 is associated with chemoresistance in esophageal cancer through transformation of normal fibroblasts to cancer-associated fibroblasts," *Carcinogenesis*, vol. 36, no. 8, pp. 894–903, 2015.

[24] Y. Lu, et al., "MicroRNA profiling and prediction of recurrence/relapse-free survival in stage I lung cancer," *Carcinogenesis*, vol. 33, no. 5, pp. 1046–1054, 2012.

[25] Y. Zhang, et al., "MiR-124 radiosensitizes human colorectal cancer cells by targeting PRRX1," *PLOS One*, vol. 9, no. 4, pp. 1–9, 2014.

[26] X. An, C. Sarmiento, T. Tan, and H. Zhu, "Regulation of multidrug resistance by microRNAs in anti-cancer therapy," *Acta Pharmaceutica Sinica B*, vol. 7, no. 1, pp. 38–51, 2017.

[27] J. Li, Y. Wang, Y. Song, Z. Fu, and W. Yu, "miR-27a regulates cisplatin resistance and metastasis by targeting RKIP in human lung adenocarcinoma cells," *Molecular Cancer*, vol. 13, no. 1, pp. 193–201, 2014.

[28] W. Wang, J. Yang, Y.-Y. Xiang, J. Pi, and J. Bian, "Overexpression of hsa-miR-320 is associated with invasionand metastasis of ovarian cancer," *J. Cellular Biochemistry*, vol. 118, no. 11, pp. 3654–3661, 2017.

[29] X. Ke, et al., "MiR-410 induces stemness by inhibiting Gsk3β but upregulating β-catenin in non-small cells lung cancer," *Oncotarget*, vol. 8, no. 7, pp. 11 356–11 371, 2017.

[30] K. L. Gorringe, et al., "Are there any more ovarian tumor suppressor genes? a new perspective using ultra high-resolution copy number and loss of heterozygosity analysis," *Genes Chromosomes Cancer*, vol. 48, no. 10, pp. 931–942, 2009.

[31] X. Ge, L. Zheng, M. Huang, Y. Wang, and F. Bi, "MicroRNA expression profiles associated with acquired gefitinib-resistance in human lung adenocarcinoma cells," *Molecular Med. Rep.*, vol. 11, no. 1, pp. 333–340, 2015.
[32] T. Boren, et al., "MicroRNAs and their target messenger RNAs associated with ovarian cancer response to chemotherapy," *Gynecologic Oncology*, vol. 113, no. 2, pp. 249–255, 2009.
[33] L. Li, et al., "Hypoxia-induced miR-210 in epithelial ovarian cancer enhances cancer cell viability via promoting proliferation and inhibiting apoptosis," *Int. J. Oncology*, vol. 44, no. 6, pp. 2111–2120, 2014.
[34] X. Chen, et al., "Plasma miRNAs in predicting radiosensitivity in non-small cell lung cancer," *Tumor Biol.*, vol. 39, no. 9, pp. 11 927–11 936, 2016.
[35] Y. Liaoa, et al., "MiR-760 overexpression promotes proliferation in ovarian cancer by downregulation of PHLPP2 expression," *Gynecologic Oncology*, vol. 143, no. 3, pp. 655–663, 2016.
[36] J. Kang, et al., "microRNA-99b acts as a tumor suppressor in non-small cell lung cancer by directly targeting fibroblast growth factor receptor 3," *Exp. Therapeutic Med.*, vol. 3, no. 1, pp. 149–153, 2012.
[37] X. Sun, et al., "MiR-548c impairs migration and invasion of endometrial and ovarian cancer cells via downregulation of twist," *J. Exp. Clinical Cancer Res.*, vol. 35, no. 1, pp. 1–10, 2016.
[38] Q. Song, et al., "miR-483–5p promotes invasion and metastasis of lung adenocarcinoma by targeting RhoGDI1 and ALCAM," *Cancer Res.*, vol. 3, no. 11, pp. 3031–3042, 2014.
[39] H. Tan, et al., "miR-382 inhibits migration and invasion by targeting ROR1 through regulating EMT in ovarian cancer," *Int. J. Oncology*, vol. 48, no. 1, pp. 181–190, 2015.
[40] X. Wang, D. Jiang, C. Xu, and G. Zhu, "Differential expression profile analysis of miRNAs with HER-2 overexpression and intervention in breast cancer cells," *Int. J. Clinical Exp. Pathology*, vol. 10, no. 5, pp. 5039–5062, 2017.
[41] S. Zhong, et al., "MicroRNA expression profiles of drug-resistance breast cancer cells and their exosomes," *Oncotarget*, vol. 7, no. 15, pp. 19 601–19 609, 2016.
[42] L. Jiang, et al., "Mir-489 regulates chemoresistance in breast cancer via epithelial mesenchymal transition pathway," *FEBS Lett.*, vol. 588, no. 11, pp. 2009–2015, 2014.
[43] D. J. Schultz, et al., "Genome-wide miRNA response to anacardic acid in breast cancer cells," *PLoS One*, vol. 12, no. 9, 2017, Art. no. e0184471.
[44] H. Zhao, J. Shen, L. Medico, D. Wang, C. B. Ambrosone, and S. Liu, "A pilot study of circulating miRNAs as potential biomarkers of early stage breast cancer," *PLoS One*, vol. 5, no. 10, 2010, Art. no. e13735.
[45] X. Yu, et al., "MiR-214 increases the sensitivity of breast cancer cells to tamoxifen and fulvestrant through inhibition of autophagy," *Molecular Cancer*, vol. 2015, no. 14, pp. 208–223, 2015.
[46] Y. Hu, Z. Tang, B. Jiang, J. Chen, and Z. Fu, "miR-198 functions as a tumor suppressor in breast cancer by targeting CUB domain-containing protein 1," *Oncology Lett.*, vol. 13, no. 3, pp. 1753–1760, 2017.
[47] Y.-W. Wang, W. Zhang, and R. Ma, "Bioinformatic identification of chemoresistance-associated microRNAs in breast cancer based on microarray data," *Oncology Rep.*, vol. 39, no. 3, pp. 1003–1010, 2018.
[48] K. Sugimura, et al., "Let-7 expression is a significant determinant of response to chemotherapy through the regulation of IL-6/STAT3 pathway in esophageal squamous cell carcinoma," *Clinical Cancer Res.*, vol. 18, no. 18, pp. 5144–5153, 2012.
[49] E. Forma, et al., "Association between the c.*229C > T polymorphism of the topoisomerase IIβ binding protein 1 (TopBP1) gene and breast cancer," *Oncology Rep.*, vol. 40, no. 5, pp. 3493–3502, 2013.
[50] F. Qian, et al., "Genetic variants in microRNA and microRNA biogenesis pathway genes and breast cancer risk among women of African ancestry," *Human Genetics*, vol. 135, no. 10, pp. 1145–1159, 2016.
[51] B. Hui, X. Chen, L. Hui, R. Xi, and X. Zhang, "Serum miRNA expression in patients with esophageal squamous cell carcinoma," *Oncology Lett.*, vol. 10, no. 5, pp. 3008–3012, 2015.
[52] M. Y. Shah, X. Pan, L. N. Fix, M. A. Farwell, and B. Zhang, "5-fluorouracil drug alters the microRNA expression profiles in MCF-7 breast cancer cells," *J. Cellular Physiology*, vol. 226, no. 7, pp. 1868–1878, 2011.
[53] L. Zhang, P. S. Sullivan, J. C. Goodman, P. H. Gunaratne, and D. Marchetti, "MicroRNA-1258 suppresses breast cancer brain metastasis by targeting heparanase," *Cancer Res.*, vol. 71, no. 3, pp. 1868–1878, 2011.
[54] B. G. Kim, et al., "Transcriptome-wide analysis of compression-induced microRNA expression alteration in breast cancer for mining therapeutic targets," *Oncotarget*, vol. 7, no. 19, pp. 27 468–27 478, 2016.
[55] O. Berillo, M. Régnier, and A. Ivashchenko, "Binding of intronic miRNAs to the mRNAs of host genes encoding intronic miRNAs and proteins that participate in tumourigenesis," *Comput. Biol. Med.*, vol. 43, no. 10, pp. 1374–1381, 2013.
[56] R. Chacolla-Huaringa, J. Moreno-Cuevas, V. Trevino, and S.-P. Scott, "Entrainment of breast cell lines results in rhythmic fluctuations of microRNAs," *Int. J. Molecular Sci.*, vol. 18, no. 7, 2017, Art. no. 1499.
[57] X. Chen, L. Wang, J. Qu, N.-N. Guan, and J.-Q. Li, "Predicting miRNA-disease association based on inductive matrix completion," *Bioinf.*, vol. 34, no. 24, pp. 4256–4265, 2018.
[58] X. Chen, D. Xie, L. Wang, Q. Zhao, Z.-H. You, and H. Liu, "BNPMDA: Bipartite network projection for miRNA-disease association prediction," *Bioinf.*, vol. 34, no. 18, pp. 3178–3186, 2018.
[59] X. Chen, J. Yin, J. Qu, and L. Huang, "MDHGI: Matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction," *PLoS Comput. Biol.*, vol. 14, no. 8, 2018, Art. no. e1006418.

Jayanta Kumar Pal received the BTech degree in information technology and the MTech degree in computer science and engineering from the West Bengal University of Technology, India, in 2008 and 2010, respectively. He is working toward the PhD degree in the Department of Computer Science and Engineering, University of Calcutta, Kolkata. He is currently working as a research fellow with 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 (Eng.) degree from Jadavpur University, Kolkata, in 2008. He was a post-doctoral fellow with 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 a curated paper in the Saccharomyces Genome Database, Stanford University, California. He was a recipient of the Microsoft Young Faculty Award in 2010.

Sankar K. Pal (M'80-SM'84-F'93-LF'15) received the PhD degrees from the University of Calcutta, Kolkata, and Imperial College, London. He joined the Indian Statistical Institute, Kolkata, in 1975 as a CSIR-SRF where he became a full professor in 1987, a distinguished scientist in 1998, and the director in 2005. He founded the Machine Intelligence Unit and the Center for Soft Computing Research at his institute in Kolkata. He was an INAE chair professor and J.C. Bose National fellow. Currently, he is an INSA distinguished professor. He worked with the University of California at Berkeley, Berkeley, California, and University of Maryland at College Park, College Park, Maryland, NASA JSC, Houston, Texas, and US Naval Research Laboratory, Washington, DC. He held several visiting positions in Italy, Poland, Hong Kong, and Australia. He has coauthored 20 books and more than 400 research publications in the areas of pattern recognition, machine learning, image/video processing, data mining, web intelligence, soft computing, bioinformatics, and cognitive machines. He received several national/ international awards including the S.S. Bhatnagar Prize (India), the Padma Shri (India), the G.D Birla Award (India), and the Khwarizmi International Award (Iran). He has been an IEEE CS Distinguished Visitor since 1987. He is/was on the editorial boards of 20 journals including several IEEE Transactions. He has visited 45 countries as a keynote/ invited speaker. He is a fellow of TWAS, IAPR, IFSA, and all four National Academies for Science/Engineering in India. He is a life fellow of the IEEE.