A New Evolutionary Rough Fuzzy Integrated Machine Learning Technique for microRNA selection using Next-Generation Sequencing data of Breast Cancer

Jnanendra Prasad Sarkar*†
Larsen & Toubro Infotech Ltd.
Pune, India

Indrajit Saha*
Dept. of Computer Science and Engg.,
National Institute of Technical Teachers' Training and Research
Kolkata, India

Somnath Rakshit*
Centre of New Technologies,
University of Warsaw
Warsaw, Poland

Monalisa Pal
Machine Intelligence Unit,
Indian Statistical Institute
Kolkata, India

Michal Wlasnowolski‡
Faculty of Mathematics and Information Science, Warsaw University of Technology
Warsaw, Poland

Anasua Sarkar
Department of Computer Science and Engineering, Jadavpur University
Kolkata, India

Ujjwal Maulik
Department of Computer Science and Engineering, Jadavpur University
Kolkata, India

Dariusz Plewczynski§
Center of New Technologies,
University of Warsaw
Warsaw, Poland

ABSTRACT

MicroRNAs (miRNA) play an important role in various biological process by regulating gene expression. Their abnormal expression may lead to cancer. Therefore, analysis of such data may discover potential biological insight for cancer diagnosis. In this regard, recently many feature selection methods have been developed to identify such miRNAs. These methods have their own merits and demerits as the task is very challenging in nature. Thus, in this article, we propose a novel wrapper based feature selection technique with the integration of Rough and Fuzzy sets, Random Forest and Particle Swarm Optimization, to identify putative miRNAs that can solve the underlying biological problem effectively, i.e. to separate tumour and control samples. Here, Rough and Fuzzy sets help to address the vagueness and overlapping characteristics of the dataset while performing clustering. On the other hand, Random Forest is applied to perform the classification task on the clustering results to yield better solutions. The integrated clustering and classification tasks are considered as an underlying optimization problem for Particle Swarm Optimization method where particles encode features, in this case, miRNAs. The performance of the proposed wrapper based method has been demonstrated quantitatively and visually on next-generation sequencing data of breast cancer from The Cancer Genome Atlas (TCGA). Finally, the selected miRNAs are validated through biological significance tests. The code and dataset used in this paper are available online[1].

CCS CONCEPTS

• **Computing methodologies** → **Feature selection**; • **Applied computing** → **Bioinformatics**; **Transcriptomics**; *Systems biology*;

KEYWORDS

Breast Cancer, Clustering, Fuzzy Set, Feature Selection, Particle Swarm Optimization, Random Forest, Rough Set

ACM Reference Format:
Jnanendra Prasad Sarkar, Indrajit Saha, Somnath Rakshit, Monalisa Pal, Michal Wlasnowolski, Anasua Sarkar, Ujjwal Maulik, and Dariusz Plewczynski. 2019. A New Evolutionary Rough Fuzzy Integrated Machine Learning Technique for microRNA selection using Next-Generation Sequencing data of Breast Cancer. In *Genetic and Evolutionary Computation Conference Companion (GECCO '19 Companion), July 13–17, 2019, Prague, Czech Republic.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3319619.3326836

1 INTRODUCTION

MicroRNAs (miRNA) are small non-coding molecules of single-stranded RNA, 22-25 nucleotide long. MicroRNAs (miRNAs) bind

*Equally contributed
†Additional Affiliation: Department of Computer Science and Engineering, Jadavpur University, Kolkata, India
‡Additional Affiliation: Center of New Technologies, University of Warsaw, Warsaw, Poland
§Additional Affiliation: Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland

GECCO '19 Companion, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6748-6/19/07…$15.00
https://doi.org/10.1145/3319619.3326836

[1]http://www.nitttrkol.ac.in/indrajit/projects/mirna-pso-rfcm-rf-berastcancer/

partially with complementary sites in target messenger RNAs (mRNAs) and thus regulate gene expression of animals and plants, where dysregulation causes tumor formations [8]. MicroRNAs are also found to have important roles in other diseases like diabetes [2], infectious disease [22], various neurodegenerative disorder [11]. In recent decade, various analytical processes have been studied on miRNA. Those are identifying set of miRNAs derived from common primary transcripts [21], co-expression analysis between neighbouring miRNAs [4], prediction of miRNA targets [10] and specificity of miRNA in particular tissue [23]. Individual miRNA can target many mRNAs based on sequence complementarity. However, a significant fraction of these interactions may depend on cell type, context [3] and also on the binding of additional co-factors [12]. Moreover, a smaller subset of target interactions actually cause tumour development. Therefore, it is important to identify the potential set of miRNAs, which is very challenging task. Therefore, analysis of miRNAs in various aspects has become major focused area of research in recent decades. Recent studies reveal that some miRNAs are differently expressed in both normal and cancerous tumour tissues of all types. Additionally, it is also seen that some miRNAs are differently expressed in specific tumour tissue. So, it suggests that there might be any link between miRNAs and oncogenesis. Also, diagnosis of cancer might be possible from onco-miRNA signature. Therefore, in addition to wet laboratory experiments, computational methods can also be useful to detect onco-miRNA signature and an alternative method for medical diagnosis. In this regards, different machine learning techniques like K-Nearest Neighbour (K-NN) [1], Support Vector Machine (SVM) [7], Decision tree (DT) [20], Naive Bayesian classifier (NB) [9] etc. are used for analysis. However, performance of these heterogeneous methods depends very much on the selection of features. This fact motivated us to propose a novel method for identifying potential set of features.

In this regard, it is important to address inherent vagueness, uncertainty and overlapping characteristics within the dataset. Fuzzy C-Means (FCM) [5] using Fuzzy set theory can handle overlapping characteristics. However, it is very sensitive to noisy data. Thus, variants of FCM [16, 19] have been developed for the same to handle subtle vagueness and uncertainty by incorporating Rough set theory [18] and known as Rough Fuzzy C-Means (RFCM) [16]. According to Rough set theory, a point can either belong to a particular cluster with membership value 1 or to the boundary region of multiple clusters. The boundary region is considered as overlapping region of more than one clusters. Hence, we have used both Rough and Fuzzy set theories together to handle vagueness, uncertainty and overlapping characteristics of the dataset. However, Rough Fuzzy integrated technique yields clusters having set of crisp and rough points. Therefore, to get better clustering results, well-known machine learning method called Random Forest (RF) [6] is applied on rough points after being trained on crisp points. The integrated clustering (RFCM) and classification (RF) tasks are considered as an underlying optimization problem for Particle Swarm Optimization (PSO) [14] in order to identify potential set of features, in this case miRNAs, to perform better separation of tumour and control samples. Here PSO encodes miRNAs as elements of a particle. The proposed wrapper based technique is abbreviated as PSO-RFCM-RF.

The publicly available Breast Invasive Carcinoma (BRCA) dataset[2] is used to demonstrate the performance of proposed method. The breast cancer is believed to be most widely diagnosed cancer type and mostly found within female population. In female body, it mostly begins in cells of the lobules which are known as milk-producing glands. Thereafter, it might get spread outside milk ducts. Unlike non-invasive, invasive cancers grow into healthy tissues. Sometimes, both non-invasive and invasive cancers are found in same specimen. Even today, the exact cause for this cancer is not fully known, whereas, the proper analysis of various biomolecules may bring more insights. Thus, the analysis of miRNA expression and its proper selection may help to achieve that goal. In this regard, the proposed PSO-RFCM-RF is used and the selected miRNAs are validated quantitatively as well as through biological significance tests.

2 EVOLUTIONARY ROUGH FUZZY INTEGRATED MACHINE LEARNING TECHNIQUE

This section describes the proposed wrapper based feature selection technique.

Algorithm 1 Steps of the RFCM

Input:
X, the dataset
η, the fuzzy exponent
ϵ, a small real threshold value between [0,1]
K, the number of cluster
f_{LW}, relative weight for lower approximation of rough clustering, $0 < f_{LW} < 1$
Output: $[\mu]\ where,\ 1 \leq l \leq K\ and\ 1 \leq i \leq n$

1: Select random K points from dataset as K cluster means
2: **repeat**
3: Compute μ_{li} for all n points using Equation 3
4: Compute the difference between highest two computed membership, μ_{li} of each and every n data points
// *Let μ_{li} and μ_{hi}, highest and second highest computed membership values of x_i among all K clusters, where $1 \leq l, h \leq K\ and\ h \neq l$*
5: Compute the value of threshold Δ
// *Δ is the mean of $(\mu_{li} - \mu_{hi}), \forall i = 1, 2, \ldots, n$*
6: **if** $(\mu_{li} - \mu_{hi}) > \Delta$ **then**
7: $\mu_{li} \leftarrow 1, \mu_{hi} \leftarrow 0\ \forall h = 1, 2, \ldots, K\ and\ h \neq l$
// *x_i is exactly classified to $\underline{B}(C_l)$, also to $\overline{B}(C_l)$ as per RST*
8: **else**
9: Keep μ_{hi} unchanged $\forall h = 1, 2, \ldots, K$
// *x_i can belong to $Upper\ Approximation$ of multiple clusters. Hence, x_i belong to $\overline{B}(C_l)$ and $\overline{B}(C_h)$*
10: **end if**
11: Compute new mean with the help of Equation 4
12: **until** $|Current\ J_{RFCM} - Previous\ J_{RFCM}| \leq \epsilon$
13: **return** $[\mu]\ where,\ 1 \leq l \leq K\ and\ 1 \leq i \leq n$

The proposed clustering and classification integrated wrapper based feature selection technique uses Rough and Fuzzy sets to cluster a dataset $X = \{x_i \mid 1 \leq i \leq n\}$. The steps of clustering are described in Algorithm 1, where it produces crisp and rough sets of points. Crisp set of points are crisply classified into *lower approximation* region whereas rough points belong to boundary region of multiple clusters. According to Rough set theory [18], *lower approximation* ($\underline{B}(X)$) and *upper approximation* ($\overline{B}(X)$) are

[2] https://cancergenome.nih.gov/

Algorithm 2 Steps of RFCM-RF

Input:
X, the dataset
η, the fuzzy exponent
ϵ, a small real threshold value between [0,1] used to terminate RFCM
K, the number of cluster
f_{LW}, relative weight for lower approximation of rough clustering, $0 < f_{LW} < 1$
$\mathbb{T}$, Number of tree for RF
Output: F, the final class label vector of X

1: Using Algorithm 1 produce crisp dataset, $\mathbb{L} = \{x_i \in \underline{B}(C_l) \mid 1 \leq l \leq K \; and \; 1 \leq i \leq n\}$ and corresponding cluster label vector, λ_1
2: Classify $\mathbb{L}^* = (X - \mathbb{L})$ using RF, trained by $\mathbb{L}$ and λ_1 to get label vector, λ_2
3: Combine λ_1 and λ_2 to get final cluster label vector, F, where F should be in order of X
4: **return** F

defined in Equation 1, where U is non-empty set called *universe* and B determines the *equivalence* or *indiscernibility* relation. An *indiscernibility* class containing x is denoted as $B(x)$. The difference between *upper* and *lower approximation* regions, (i.e., $BN(X) = \overline{B}(X) - \underline{B}(X)$), is called boundary region of X. If $BN(X)$ is empty then X is called crisp set of points, otherwise it is called as rough set of points.

$$\underline{B}(X) = \bigcup_{x \in U} \{B(x) | B(x) \subseteq X\}; \; \overline{B}(X) = \bigcup_{x \in U} \{B(x) | B(x) \cap X \neq \phi\} \tag{1}$$

Algorithm 1 optimises the objective function as defined in Equation 2.

$$J_{RFCM} = \begin{cases} f_{LW} \times \mathcal{A} + f_{BN} \times \mathcal{B}, \; if \underline{B}(C_l) \neq \emptyset, BN(C_l) \neq \emptyset \\ \mathcal{A}, \; if \underline{B}(C_l) \neq \emptyset, BN(C_l) = \emptyset \\ \mathcal{B}, \; if \underline{B}(C_l) = \emptyset, BN(C_l) \neq \emptyset \end{cases} \tag{2}$$

$$\mathcal{A} = \sum_{l=1}^{K} \sum_{x_i \in \underline{B}(C_l)} (\mu_{li})^{\eta} \, D(c_l, x_i); \; \mathcal{B} = \sum_{l=1}^{K} \sum_{x_i \in BN(C_l)} (\mu_{li})^{\eta} \, D(c_l, x_i)$$

$$\mu_{li} = \frac{1}{\sum_{h=1}^{K} (\frac{D(c_l, x_i)}{D(c_h, x_i)})^{\frac{2}{\eta - 1}}}; \; \sum_{l=1}^{K} \mu_{li} = 1 \text{ for } 1 \leq l \leq K; \; 1 \leq i \leq n, \tag{3}$$

where, C_l is lth cluster and $D(c_l, x_i)$ measures Euclidean distance of the point, x_i from the center of cluster, $c_l \in C_l$. η is weighting coefficient while μ_{li} as defined in Equation 3 represents the fuzzy membership value or the degree of belongingness of the ith point to the lth cluster. According to rough set theory, the degree of belongingness of the points is 1 for a particular cluster within *lower approximation* region. Therefore, Equation 2 can be rewritten as $\mathcal{A} = \sum_{l=1}^{K} \sum_{x_i \in \underline{B}(C_l)} D(c_l, x_i)$. The cluster center is updated by Equation 4.

$$c_l = \begin{cases} f_{LW} \times \mathcal{G}_{LW} + f_{BN} \times \mathcal{G}_{BN}, \\ \quad if \underline{B}(C_l) \neq \emptyset, BN(C_l) \neq \emptyset \\ \mathcal{G}_{LW}, \; if \underline{B}(C_l) \neq \emptyset, BN(C_l) = \emptyset \\ \mathcal{G}_{BN}, \; if \underline{B}(C_l) = \emptyset, BN(C_l) \neq \emptyset \end{cases} \tag{4}$$

where,

$$\mathcal{G}_{LW} = \frac{\sum_{x_i \in \underline{B}(C_l)} x_i}{| \underline{B}(C_l) |}; \; \mathcal{G}_{BN} = \frac{\sum_{x_i \in BN(C_l)} \{(\mu_{li})^{\eta}\} x_i}{\sum_{x_i \in BN(C_l)} \{(\mu_{li})^{\eta}\}}$$

However, using RFCM, it is difficult to determine the definite belongingness of rough points in a particular cluster. Thus, Random Forest (RF) is used to classify those rough points with the help of crisp points that are used to train the RF. It refines the performance of the clustering. The steps of RFCM-RF are described in Algorithm 2. Furthermore, the RFCM-RF is considered as an optimization problem for Particle Swarm Optimization (PSO) while identifying the potential set of features, in this case miRNAs. Here PSO is used as a global optimizer to achieve the optimal solution, in this case, the set of miRNAs while performing clustering and classification tasks in integrated fashion, i.e. RFCM-RF. The method is described in Algorithm 3.

Algorithm 3 Steps of PSO integrated RFCM-RF

Input:
X, the dataset
η, the fuzzy exponents
ϵ, a small real threshold value between [0,1] used to terminate RFCM
K, the number of cluster
f_{LW}, relative weight for lower approximation of rough clustering, $0 < f_{LW} < 1$
$\mathbb{T}$, Number of tree for RF
$\mathcal{N}_{par}$, Number of particles
$\mathcal{N}_{itr}$, Number of iteration for PSO
L, Length of particle
α, Inertia weight $\in [0.5, 1]$
β_1, β_2, Cognitive and Social constant
Output: $\mathcal{S}_{best}$, Best feature subset

1: $\hat{X} \leftarrow Preprocess(X)$
2: $\mathcal{P}^{(t)} \leftarrow InitalPopulation(\hat{X}, \mathcal{N}_{par}, L)$
3: **for** $i = 1$ to $\mathcal{N}_{itr}$ **do**
4: $[\mathcal{P}_{l_{best}}{}^{(t)}, \mathcal{P}_{g_{best}}{}^{(t)}] \leftarrow FitnessRFCMRF(\hat{X}, \mathcal{P}^{(t)})$
5: $\mathcal{V}^{(t+1)} \leftarrow Velocity(\mathcal{P}^{(t)}, \mathcal{V}^{(t)}, \mathcal{P}_{l_{best}}{}^{(t)}, \mathcal{P}_{g_{best}}{}^{(t)})$ // using Equation 5
6: $\mathcal{P}^{(t+1)} \leftarrow Position(\mathcal{V}^{(t+1)}, \mathcal{P}^{(t)})$ // using Equation 6
7: $\mathcal{S}_{best} \leftarrow BestFeatueSet(\mathcal{P}^{(t)}, \mathcal{P}_{g_{best}}{}^{(t)})$
8: **end for**
9: **return** $\mathcal{S}_{best}$

PSO works with a population of candidate solution called Swarm where candidate solutions are represented as particles ($\mathcal{P}_j$, where $j = 1, 2, \ldots, \mathcal{N}_{par}$ and $\mathcal{N}_{par}$ is number of particles). Element of each such particle is composed of position and length ($\mathcal{L}$). The movement of a particle is tracked by updating velocity ($\mathcal{V}_j$) and position as defined in Equation 5.

$$\mathcal{V}_j{}^{(t+1)} = \alpha \times \mathcal{V}_j{}^{(t)} + \beta_1 \times (\mathcal{P}_{l_{best}}{}^{(t)} - \mathcal{P}_j{}^{(t)}) + \beta_2 \times (\mathcal{P}_{g_{best}}{}^{(t)} - \mathcal{P}_j{}^{(t)}) \tag{5}$$

$$\mathcal{P}_j{}^{(t+1)} = \mathcal{P}_j{}^{(t)} + \mathcal{V}_j{}^{(t+1)} \tag{6}$$

Where, t is time of different iterations, α is the inertia weight $\in [0.5, 1]$, β_1 is cognitive constant and β_2 is social constant. Moreover, $\mathcal{P}_{l_{best}}$ and $\mathcal{P}_{g_{best}}$ represent local best particle of current iteration and global best particle till current iteration respectively. PSO algorithm terminates after fix number of iterations. In *InitialPopulation* step, a particle is prepared after random selection of elements (in this case miRNAs) from pre-processed dataset. The encoded particle is then used to compute fitness using objective function mentioned in Algorithms 1 and 2. The fitness value ranges from 0 to 100 where, higher value denotes better result. Based on fitness value, local and global best particles are identified to update the *Velocity*. Thereafter, new position of the particle is computed using updated velocity. Finally, the algorithm gets terminated after a fix number of iterations

producing the optimal feature set. The entire process is run for 50 times in our experiment.

The proposed technique, PSO-RFCM-RF is random in nature, which has a probability of having false positive or false negative while selecting miRNAs. To reduce this probability of false positive or false negative, PSO-RFCM-RF is executed for 50 times followed by ranking of miRNAs based on occurrence in 50 different sets of features. Maximum number of occurrence of any miRNAs over 50 runs indicates that it is significant for producing the better fitness value by reducing error while assigning points to a cluster. After 50 runs, PSO-RFCM-RF ensures that miRNAs are ranked according to their occurrence and finally set of miRNAs are selected from sorted list. Here the number of runs is an important factor for reducing false negative. Mathematically, it has been found that 50 runs are sufficient to reduce false negative. Suppose, at each run, PSO-RFCM-RF selects random $\mathbb{S}$ = 10 miRNAs from the entire collection of miRNAs, $\mathbb{D}$ and the total number of runs = $\mathbb{R}$. For $i = 1, 2, \ldots, \mathbb{D}$, let $\mathbb{V}_i$ is the Bernouilli distributed indicator variable where $\mathbb{V}_i$ =1 if miRNA, m_i never gets selected. The probability of selecting m_i in a single run is = $\mathbb{S}/\mathbb{D}$ and probability that it does not get selected is = (1 - $\mathbb{S}/\mathbb{D}$). Hence the expectation of $\mathbb{V}_i$ can mathematically be defined as in Equation 7.

$$\mathbb{E}[\mathbb{V}_i] = Pr(\mathbb{V}_i) = (1 - \frac{\mathbb{S}}{\mathbb{D}})^{\mathbb{R}} \tag{7}$$

Let us assume, $\mathbb{V} = \sum_{i=1}^{\mathbb{D}} \mathbb{V}_i$ is the random variable which counts the number of miRNAs that do not belong to the final set of the miRNAs at least once. By linearity relation of the expectation, Equation 8 can be written as below.

$$\mathbb{E}[\mathbb{V}] = \sum_{i=1}^{\mathbb{D}} \mathbb{E}[\mathbb{V}_i] = \mathbb{D} * (1 - \frac{\mathbb{S}}{\mathbb{D}})^{\mathbb{R}} \tag{8}$$

Therefore, it can be written that,

$$\mathbb{E}[(\mathbb{D} - \mathbb{V})] = \mathbb{D} - \mathbb{E}[\mathbb{V}] \tag{9}$$

Substituting the parameters $\mathbb{S}$, $\mathbb{R}$ and $\mathbb{D}$ with the values as 10, 50 and 244 respectively, the expected number of miRNAs reported at least once after 50 runs is 212 and the expected number of new miRNAs added in a further iteration is 1. However, in our experiment, the sorted number of miRANs is 60. Hence, it is proved that 50 runs are sufficient to get a stable set of miRNAs to reduce the probability of false negative. This justifies the process of selection of miRNAs and determining 50 runs in the proposed technique.

3 COMPLEXITY ANALYSIS

3.1 Space Complexity Analysis

PSO-RFCM-RF mostly needs space to store data, population, centers of the K clusters, fuzzy membership matrix and fitness value of each particle of population. Additional space is required for processing of RF. Therefore, overall space complexity can be computed as $O(nm + Km + Kn + \mathbb{T}nm + Km\mathcal{N}_{par})$. After simplifying, the worst case space complexity is $O(n^2)$ when $n = m$.

3.2 Time Complexity Analysis

Worst case time complexity of PSO-RFCM-RF mainly depends on two parts, (a) time to compute objective function, which is time complexity of RFCM-RF and (b) time required for standard PSO algorithm. For first part, majority of the time, RFCM-RF spends on computation of fuzzy membership matrix, searching for two highest two membership values, computation of centers of each cluster and additionally for RF processing. Considering all these activities, the overall time complexity can be derived as $O(4Knm + Kn + \mathbb{T}nmlog(n))$. For second part, PSO takes time overall as $O(nlog(n)\mathcal{N}_{par})$ for each iteration. Therefore, time complexity of PSO-RFCM-RF can be considered as $O(4Knm + Kn + \mathbb{T}nmlog(n)) + nlog(n)\mathcal{N}_{par}$. After simplification, worst case time complexity is $O(n^2)$, when $n = m$, for single iteration.

Table 1: Statistics of Patients in BRCA data

Data Category	Number of Patients	Avg. Age of Patient (in years)	Avg. days to followup
Tumour	762	57.98	1288.29
Control	87	58.64	835.17

Table 2: Top 10 miRNAs with their up/down regulation, p-value and PubMed ID

miRNA	Regulation (Up/Down)	*p-value*	PubMed ID
hsa-mir-139	Down	1.92e-50	26497851
hsa-mir-21	Up	3.25e-49	29552160
hsa-mir-183	Up	4.17e-47	26170234
hsa-mir-96	Up	7.57e-47	24366472
hsa-mir-486	Down	3.36e-41	25027758
hsa-mir-10b	Down	1.43e-47	16103053
hsa-mir-145	Down	1.49e-46	25124875
hsa-mir-144	Down	4.87e-32	29387244
hsa-mir-15a	Up	5.23e-20	28979704
hsa-mir-182	Up	8.31e-44	19574223

Table 3: Results produced by different feature selection methods using 10-folds cross-validation on BRCA data

Methods	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F-measure (%)
PSO-RFCM-RF	**93.25**	**89.35**	**93.02**	**93.35**	**91.15**
SNR-RF	76.39	84.21	78.98	74.00	74.39
t-test-RF	76.82	84.64	79.41	74.43	74.82
RankSum-RF	76.39	84.21	78.98	74.00	74.39
JMI-RF	75.74	85.27	77.59	72.25	74.32
mRMR-RF	75.74	85.27	77.59	72.25	74.32
MIFS-RF	76.67	85.61	79.00	74.33	74.69

4 EXPERIMENTAL RESULTS

4.1 Dataset Preparation

We have used NGS based miRNA expression data of Breast Invasive Carcinoma (BRCA) from The Cancer Genome Atlas (TCGA)[3]

[3]https://cancergenome.nih.gov/

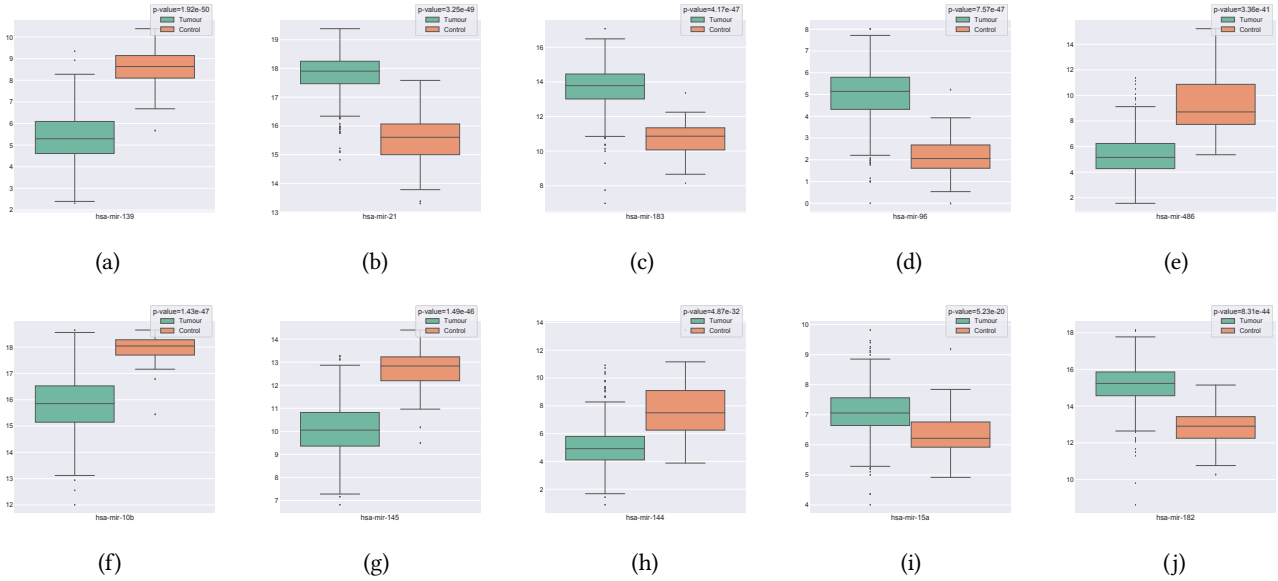


Figure 1: Box plots showing the change in expression values for the selected top 10 miRNAs identified by PSO-RFCM-RF, (a) hsa-mir-139 (b) hsa-mir-21 (c) hsa-mir-183 (d) hsa-mir-96 (e) hsa-mir-486 (f) hsa-mir-10b (g) hsa-mir-145 (h) hsa-mir-144 (i) hsa-mir-15a (j) hsa-mir-182

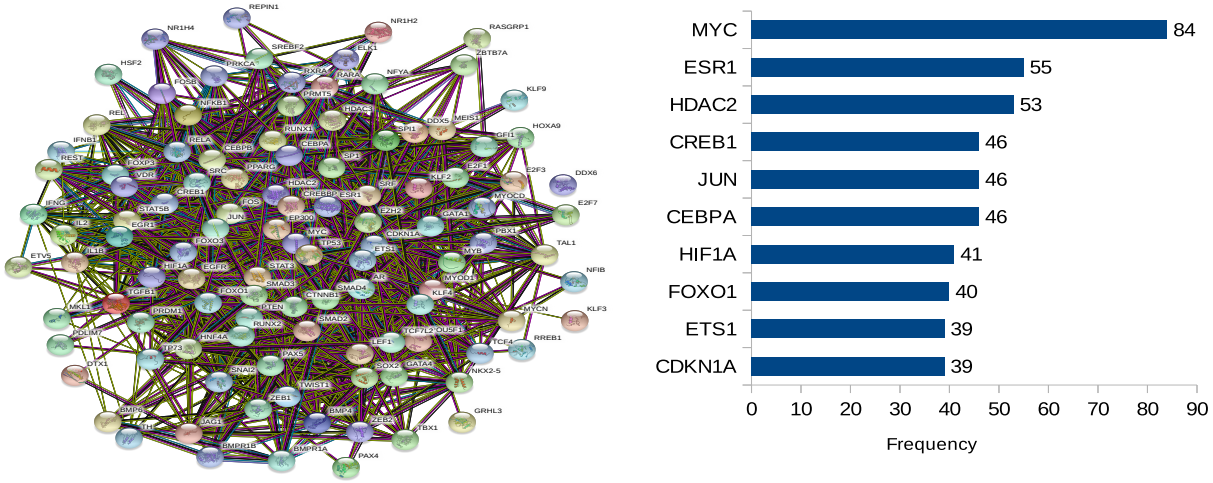
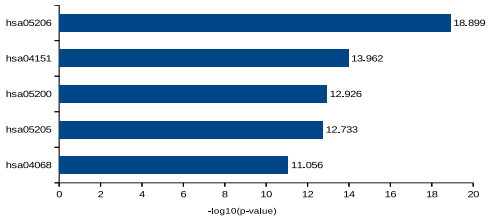


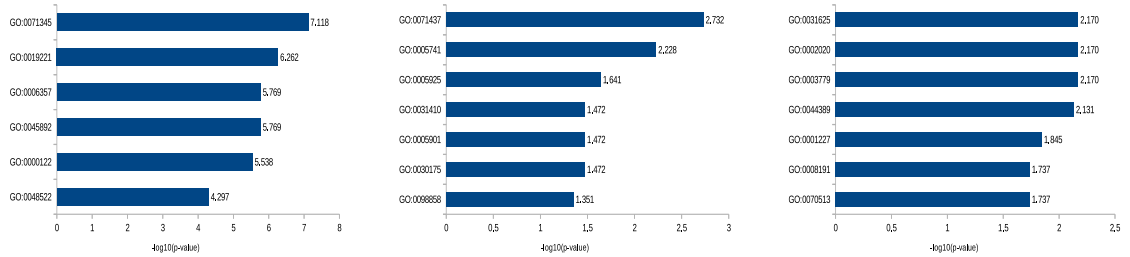
Figure 2: PPI Network for proteins (TFs) that target the top 10 miRNAs, as obtained from the TransmiR Database and the barplot shows degree of top 10 connected proteins

prepared by Illumina sequencing technology where the expression value has been computed in form of reads per million count (RPM). The dataset contains 1046 miRNA expression values for 762 tumour patients and 87 control data as shown in Table 1. Control data comprises patients who are not affected by cancer. Each patient is encoded with barcode like "TCGA-S3-A6ZH-01A-22R-A32K-13". Barcode is read as "TCGA": Project, "S3": Tissue source site (TSS), "A6ZH": Participant, "01": Sample type; "A": Vial, "22": Order of portion; "R": Molecular type of analyte, "A32K": Plate, "13": Center. A few preprocessing activities have been performed for the dataset before using in experiments. In the collected dataset, there are many miRNAs which have zero expression values. Thus, such miRNAs are excluded from the dataset which in turn, reduces the number of miRNAs from 1046 to 244. Moreover, the expression values of miRNAs are also normalized by log function with base 2.



ID	Term
hsa05206	MicroRNAs in cancer
hsa04151	PI3K-Akt signaling pathway
hsa05200	Pathways in cancer
hsa05205	Proteoglycans in cancer
hsa04068	FoxO signaling pathway
hsa05202	Transcriptional misregulation in cancer

Figure 3: Bar plot of the significant KEGG Pathways for selected top 10 miRNA



ID	Term
GO:0071345	Cellular response to cytokine stimulus
GO:0019221	Cytokine-mediated signaling pathway
GO:0045892	Regative regulation of transcription, DNA-templated
GO:0006357	Regulation of transcription from RNA polymerase II promoter
GO:0000122	Negative regulation of transcription from RNA polymerase II promoter
GO:0048522	Positive regulation of cellular process

(a)

ID	Term
GO:0071437	Invadopodium
GO:0005741	Mitochondrial outer membrane
GO:0005925	Focal adhesion
GO:0031410	Cytoplasmic vesicle
GO:0005901	Caveola
GO:0030175	Filopodium
GO:0098858	Actin-based cell projection

(b)

ID	Term
GO:0031625	Ubiquitin protein ligase binding
GO:0002020	Protease binding
GO:0003779	Actin binding
GO:0044389	Ubiquitin-like protein ligase binding
GO:0001227	Transcriptional repressor activity, RNA polymerase II transcription regulatory region sequence-specific binding
GO:0008191	Metalloendopeptidase inhibitor activity
GO:0070513	Death domain binding

(c)

Figure 4: Bar plot of GO Enrichment analysis for, (a) Biological Process (b) Cellular Component and (c) Molecular Function of selected top 10 miRNA identified by PSO-RFCM-RF

4.2 Input Parameters and Performance Metrics

Input parameter values are set experimentally and those are Fuzzy Exponent, η = 2; Number of particles, $\mathcal{N}_{par}$ = 50; Number of Iterations, $\mathcal{N}_{itr}$ = 50; Length of particles, L = 10; Cognitive constant, β_1 = 2; Social constant, β_2 = 2; Inertia Weight, α = 0.9; Number of trees for RF, $\mathbb{T}$ = 50; Relative weight for *lower approximation* of RST, ω_{low} = 0.95 and *boundary approximation*, ω_{up} = 0.05. All the algorithms have been implemented in Matlab and executed on an Intel Core i5-2410M CPU at 2.30 GHz Machine with 8GB RAM and Windows 7 operating system. Moreover, the proposed technique is validated using statistical metrics like *Accuracy*, *Precision*, *Sensitivity*, *Specificity* and *F-measure* respectively.

4.3 Results

PSO-RFCM-RF technique executes 50 times, where PSO maintains the population size as 50. Each particle in the population of PSO denotes a possible solution. To evaluate each such particle, RFCM-RF method is applied to compute fitness value. In each particle, randomly 10 elements, in this case miRNAs, are selected for evaluating the fitness. Based on the fitness value, local and global best solutions are identified. Such 50 global best solutions are considered after 50 individual run, where each solution contains 10 miRNAs. Based on the occurrence of each miRNA in 50 solutions, top 10 miRNAs are selected and reported in Table 2. Thereafter, these miRNAs are used to perform the classification task using RF with 10-fold cross validation and the results are reported in Table 3. Here, the selection of top 10 miRNAs has been done in order to avoid the false negative. Moreover, it is found from Table 3, PSO-RFCM-RF produces average percentage values of Accuracy, Precision, Sensitivity, Specificity and F-measure as 93.25, 89.35, 93.02, 93.35 and 91.15 respectively on such 10 miRNAs better as compared to the other well-known feature selection techniques viz. Signal-to-Noise Ratio (SNR), t-test, RankSum, Joint Mutual Information (JMI), Minimum Redundancy Maximum Relevance (mRMR) and Mutual Information-based Feature Selection (MIFS) which have been applied on top 10 miRNAs as identified by them. Apart from this, the Figure 1 shows the change of expression of selected top 10 miRNAs using box plot and the corresponding *p-value* is reported after performing Kruskal-Wallis

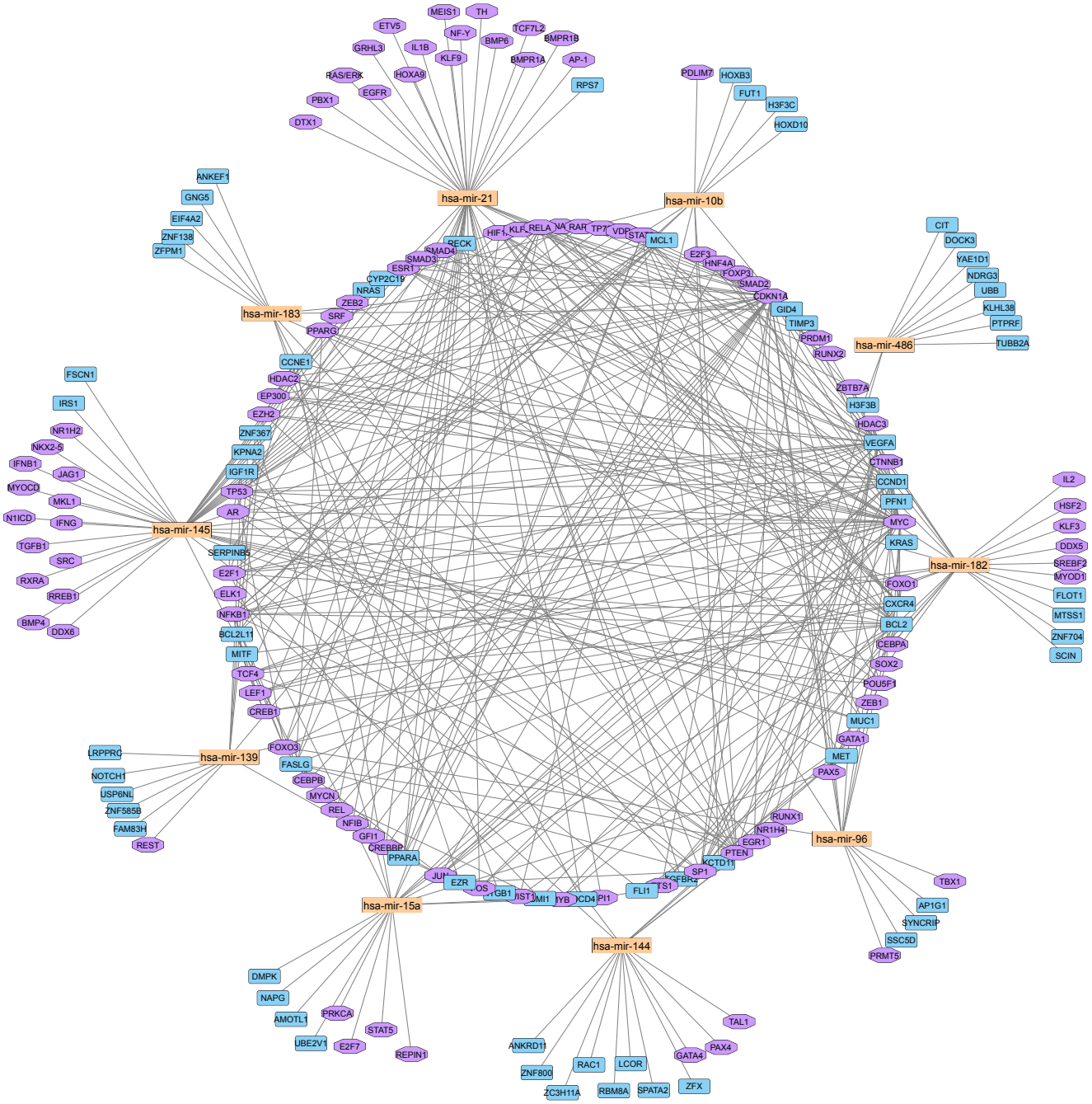


Figure 5: Network Plot of selected top 10 miRNAs that are associated with genes and TFs (Nodes marked with orange, blue and purple colour represent the miRNA, associated genes and associated TFs respectively)

H-test at 5% significance level. The same is also reported in Table 3 with the up/down regulation and PubMed ID. The *p-value* shows that the miRNAs are significantly differentially expressed with *p-value* less than 0.05 and PubMed ID are showing association of these miRNAs to the breast cancer. Therefore, it is evident that the selected 10 miRNAs are quantitatively putative and statistically significant.

4.4 Biological Significance

The biological significance of the top 10 selected miRNAs is evaluated with the help of Protein-Protein Interaction (PPI) [25], KEGG pathway analysis [13] and Gene Ontology (GO) enrichment analysis [17] for Biological Process, Cellular Component and Molecular Function.

The Protein-Protein Interaction analysis for the selected miRNAs has been conducted using STRING [24] database and shown in Figure 2. In this diagram, each node represents protein produced by a single protein-coding gene locus, whereas each edge represents protein-protein associations. The degree of interaction of each node is computed and the top 10 proteins as transcription factors (TFs) are shown in Figure 2 with the help of a bar plot. It is observed from the analysis that some important human transcription factors (TFs) like *MYC*, *Estrogenreceptor*1, *HDAC*2, *CREB*1 etc. for selected miRNAs are found as a part of Protein-Protein Interaction network. These TFs are found in breast cancer and can be regarded as targets for molecular therapies.

KEGG pathway analysis has been performed using DIANA tool [26] and the analytical findings uncover the pathway of targeted genes associated with the identified top 10 miRNAs. The targeted genes are extracted from miRTarBase database. Each pathway contains a particular score of adjusted *p-values* where, lower value signifies the higher probability of the pathway to be enriched with set of associated genes. Based on the values of adjusted *p-values*, top five pathways for selected top 10 miRNAs are shown in Figure 3. The analysis reveals the presence of PI3K-Akt signaling pathways which plays a significant role to stimulate the cell growth in human body. Over activation of this might cause an abnormal cell proliferation which are found at high rate in case of breast cancer [27]. We also found Proteoglycans which plays an important role in contributing to the various other cancer types. Similarly, FOXO signaling pathway is regarded as the target for the modulation of cancer [28].

Gene Ontology (GO) Enrichment analysis has been done using Enrichr tool [15]. This analysis discovers the various biological and cellular processes associated with selected 10 miRNAs as reported in Figures 4(a), (b) and (c). The biological process includes Cellular response to cytokine stimulus (GO:0071345), Cytokine-mediated signaling pathway (GO:0019221), Negative regulation of transcription, DNAtemplated (GO:0045892) etc. Similarly Cellular Components are found like Invadopodium (GO:0071437), Mitochondrial outer membrane (GO:0005741), Focal adhesion (GO:0005925), Cytoplasmic vesicle (GO:0031410) etc. and Molecular Functions like Ubiquitin protein ligase binding (GO:0031625), Protease binding (GO:0002020), Actin binding (GO:0003779) etc. are found as a part of GO enrichment analysis.

Additionally, Figure 5 shows the network analysis which has been performed using Cytoscape tool. The network analysis establishes the relationship among selected top 10 miRNAs with associated genes and transcription factors (TF). The associated genes and TFs are found as a part of KEGG pathway analysis and analysis of Protein-Protein Interaction. In the figure, the orange nodes represent the miRNA, whereas blue nodes signify associated genes and nodes with purple colour are associated TFs. From Figure 5, it is evident that hsa-mir-145 is associated with TF, *Estrogenreceptor*1 (*ESR*1) and Gene, *CYP2C*19 which play a crucial role in breast cancer. Similarly, hsa-mir-182 is associated with *MYC*, *FOXO*1 which also have important role to grow breast cancer.

5 CONCLUSION

In this article, a novel wrapper based feature selection technique has been proposed with the integration of clustering and classification tasks for selecting putative set of miRNAs. For this purpose, Rough and Fuzzy sets have been used to handle vagueness, uncertainty and overlapping characteristics of dataset while Random Forest and Particle Swarm Optimization have been used to improve the final results and to find the potential set of miRNAs by exploring the search space better. The results of the PSO-RFCM-RF have been demonstrated qualitatively and visually. It outperforms the existing techniques and provides putative miRNAs. Furthermore, the biological significance analysis has also been conducted to establish the biological relevance of those miRNAs in breast cancer. The results are statistically and biologically significant.

ACKNOWLEDGMENT

This work has been supported by Polish National Science Centre (2014/15/B/ST6/05082), Foundation for Polish Science (TEAM to DP) and the grant from Department of Science and Technology, Govt. of India and Polish Government under Indo-Polish/Polish-Indo project No.: DST/INT/POL/P-36/2016. The work was co-supported by grant 1U54DK107967-01 "Nucleome Positioning System for Spatiotemporal Genome Organization and Regulation" within 4DNucleome NIH program, and by European Commission as European Cooperation in Science and Technology COST actions: CA18127 "International Nucleome Consortium" (INC), and CA16212 "Impact of Nuclear Domains On Gene Expression and Plant Traits". The work was partially supported as RENOIR Project by the European Union Horizon 2020 research and innovation programme under the Marie Sklodowska- Curie grant agreement No 691152 and by Ministry of Science and Higher Education (Poland), grant Nos. W34/H2020/2016, 329025/PnH/2016.

REFERENCES

[1] N. S. Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46, 3 (1992), 175–185.
[2] C. H. Bang-Berthelsen, L. Pedersen, T. Fløyel, P. H. Hagedorn, T. Gylvin, and F. Pociot. 2011. Independent component and pathway-based analysis of miRNA-regulated gene expression in a model of type 1 diabetes. *BMC Genomics* 12, 1 (2011), 97.
[3] D. P. Bartel. 2009. MicroRNAs: target recognition and regulatory functions. *Cell* 136 (2009), 215–233.
[4] S. Baskerville and D. P. Bartel. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* 11, 3 (2005), 241–247.
[5] J. C. Bezdek. 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Kluwer Academic, MA, USA.
[6] L. Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.
[7] C. Cortes and V. Vapnik. 1995. Support-vector networks. *Machine Learning* 20, 3 (1995), 273–297.
[8] C. M. Croce. 2009. Causes and consequences of microRNA dysregulation in cancer. *Nature Reviews Genetics* 10 (2009), 704–714.
[9] H. George and J. P. Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* 69 (1995), 338–345.
[10] A. Grimson, K. K. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell* 27, 1 (2007), 91–105.

[11] J. G. Hunsberger, E. B. Fessler, F. L. Chibane, Y. Leng, D. Maric, A. G. Elkahloun, and D. M. Chuang. 2013. Mood stabilizer-regulated miRNAs in neuropsychiatric and neurodegenerative diseases: identifying associations and functions. *American Journal of Translational Research* 5, 4 (2013), 450–464.

[12] A. Jacobsen, J. Wen, D. S. Marks, and A. Krogh. 2010. Signatures of RNA binding proteins globally coupled to effective microRNA target sites. *Genome Research* 20 (2010), 1010–1019.

[13] M. Kanehisa and S. Goto. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28 (2000), 27–30.

[14] J. Kennedy and R. Eberhert. 1995. Particle swarm Optimization. *In Proceedings of IEEE International Conference on Neural Networks* 4 (1995), 1942–1948.

[15] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and A. Ma'ayan. 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* 44 (2016), W90–W97.

[16] P. Maji and S. Paul. 2013. Rough-Fuzzy Clustering for Grouping Functionally Similar Genes from Microarray Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 10, 2 (2013), 286–299.

[17] A. Michael and et. al. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics* 25 (2000), 25–29.

[18] Z. Pawlak. 1992. *Rough Sets: Theoretical Aspects of Resoning About Data.* Kluwer Academic Publishers, Norwell, MA, USA.

[19] G. Peters, F. Crespo, P. Lingras, and R. Weber. 2013. Soft clustering - Fuzzy and rough approaches and their extensions and derivatives. *International Journal of Approximate Reasoning* 54, 2 (2013), 307–322.

[20] J. R. Quinlan. 1986. Induction of Decision Trees. *Machine Learning* 1, 1 (1986), 81–106.

[21] A. Rodriguez, S. Griffiths-Jones, J. L. Ashurst, and A. Bradley. 2004. Identification of mammalian microRNA host genes and transcription units. *Genome Research* 14, 10A (2004), 1902–1910.

[22] H. Song, Q. Wang, Y. Guo, S. Liu, R. Song, X. Gao, L. Dai, B. Li, D. Zhang, and J. Cheng. 2013. Microarray analysis of microRNA expression in peripheral blood mononuclear cells of critically ill patients with influenza A (H1N1). *BMC Infectious Diseases* 13, 1 (2013), 257.

[23] Y. Sun, S. Koo, N. White, E. Peralta, C. Esau, N. M. Dean, and R. J. Perera. 2004. Development of a micro-array to detect human and mouse microRNAs and characterization of expression in human organs. *Nucleic Acids Research* 32 (2004), e188.

[24] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, Lars J. Jensen, and Christian von Mering. 2017. The STRING database in 2017: quality-controlled protein – protein association networks, made broadly accessible. *Nucleic Acids Research* 45 (2017), D362–D368.

[25] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, L. J. Jensen, and C. vonMering. 2017. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research* 45 (2017), D362–D368.

[26] I. Vlachos, K. Zagganas, M. D. Paraskevopoulou, G. Georgakilas, D. Karagkouni, T. Vergoulis, T. Dalamagas, and A. Hatzigeorgiou. 2015. DIANA-miRPath v3.0: Deciphering microRNA function with experimental support. *Nucleic Acids Research* 43 (2015), W460–W466.

[27] S. X. Yang, E. Polley, and S. Lipkowitz. 2016. New insights on PI3K/AKT pathway alterations and clinical outcomes in breast cancer. *Cancer Treatment Review* 45 (2016), 87–96.

[28] X. Zhang, N. Tang, T. J. Hadden, and A. Rishi. 2011. Akt, FoxO and regulation of apoptosis. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1813, 11 (2011), 1978–1986.