EVALUATION OF miRNA-BASED CLASSIFIERS FOR CANCER DIAGNOSIS

[1] Eliza Razak, [2] Faridah Yusof, [3]Raha Ahmad Raus
International Islamic University Malaysia, Jalan Gombak, 53100 Kuala Lumpur, Selangor, Malaysia.
For Correspondence; yfaridah@iium.edu.my

ABSTRACT: *Cancers account for the major deadliest noncommunicable diseases across all segments of the population and responsible for around 13% of all deaths world-wide. Cancer prevalence rate has noticeably quickened its pace in Malaysia and the world as we know it. Conventional diagnostic imaging and invasive biopsy examinations are still the gold standard for the diagnosis of cancer. However, these conventional methods suffer from low diagnosis sensitivity compounded by work-intensive analysis. There have indeed been a number of miRNA studies to tackle the challenges associated with cancer biomarker discovery. However, the existing diagnosis techniques using miRNA suffer from low diagnosis accuracy, sensitivity, and specificity. The low diagnosis accuracy and sensitivity of the existing techniques stems from the fact that there is extremely low miRNA count in body fluids and the presence of a huge number of irrelevant miRNAs in the expression data. There is also an inevitable problem of cross contamination between cells and exosomes in sample preparation steps. This paper describes the state-of-the-art miRNA-based classifiers for cancer miRNA expression classification. To lower the computational complexity, we employ a heuristic-based miRNA selection approach to select relevant miRNAs that are directly responsible for cancer diagnosis. Among the classifiers, Random Forest (RF) has achieved an average accuracy of 97% over 11 independent datasets. The experimental results are quite encouraging and the predictive framework managed to classify cancer accurately even with much noise contaminated in the datasets.*

Keywords: miRNA, Cancer diagnosis, Marker selection, miRNA-based classifiers.

1. INTRODUCTION

Cancer is a major deadliest noncommunicable disease that involves rapid and uncontrolled cell growth with too little apoptosis wherein cells divide and grow exponentially, generating malignant tumors that it may not only spread to the neighboring tissues but may also spread to the whole body through different circulatory routes. Cancer prevalence rate has noticeably quickened its pace in Malaysia and the world as we know it. Despite the fact that cancer is preventable and curable in early stages, the vast majority of patients are diagnosed with cancer very late. Some cancers have a predictable and distinguishable pattern which can be picked up by pattern recognition and machine learning techniques. Therefore, a computerized cancer recognition system is required to prevent people from dying as a consequence of this unfortunate disease. Currently, in actual clinical practice, conventional diagnostic imaging, invasive biopsy examinations are still the gold standard for the diagnosis of cancer and protein biomarkers such as carcinoembryonic antigen (CEA), alpha-foetoprotein (AFP), prostate specific antigen (PSA) have been widely used in cancer management [1] . However, these conventional methods suffer from low diagnosis sensitivity compounded by work-intensive analysis. Recently, there has been a tremendous increase in interest concerning circulating microRNAs (miRNAs) as a potent cancer biomarker to improve cancer management [2]. Because of higher sensitivity and specificity and with minimally invasive sampling procedures, miRNA has gained great interest in medical field. In fact, miRNAs are small non coding extracellular RNAs, approximately 18 to 22 nucleotides long, which are produced through a series of complex biogenesis pathway [3, 4] . Up to now, over two thousand human miRNA have been identified [5, 6] . miRNA can be found in the form of free circulating miRNA, encapsulated inside carrier vesicles like exosomes, microsomes and lipoproteins such as HDL and LDL or co-fractionated with protein complex Ago2 [7] . Therefore, miRNAs can be identified in body fluid such as blood, plasma, serum, urine and saliva. Interestingly, these miRNAs are found to regulate many cellular processes such as post-transcriptional gene expression, cell development, proliferation, differentiation, metabolism, aging, apoptosis and angiogenesis [8] . Ectopic microRNA expression and disturbed signaling pathways have been associated with tumorigenesis, progression, and angiogenesis as well as local recurrence and distant metastases [9] .

There have been some attempts to recognize cancer using machine learning techniques. [10] came up with a rule-based distance measure to diagnose hereditary breast cancer from miRNA expression data. Perell, Vincent et al. [11] adduced an innovative logistic regression technique based on multinomial logistic regression and lasso (least absolute shrinkage and selection operator) method. The log likelihood of each cancer class is determined in accordance with the weighted sum of the input vectors. Their technique showed the potential to eradicate the spectrum bias of logistic regression and boost the prediction accuracy. [12] originated an avant-garde SVM classifier with a recursive feature elimination technique called SVM-RFE in order to ascertain cancer-related miRNAs biomarkers. Their system was shown to be capable of distinguishing different cancer types. From their findings, we can draw conclusion that the correct combination of marker selection and classifier is important to obtain a good classification accuracy. [13] came up with innovative hypergraph-based ANN model which can identify the higher-order correlation among different variables involved in a particular cancer stage. They employed hypergraph method to regularize the classifier which can penalize over fitting and improve the classification accuracy.

However, the existing diagnosis techniques using miRNA suffer from low diagnosis accuracy, sensitivity, and specificity. The low diagnosis accuracy and sensitivity of the existing techniques stems from the fact that there is high-level of noise in miRNA expression data (resulting from low miRNA sample count in body fluids and contamination in

sample preparation and the presence of a huge number of irrelevant miRNAs in the expression data). This paper describes the state-of-the-art miRNA-based classifiers for cancer miRNA expression classification. To lower the computational complexity, we employ a heuristic-based miRNA selection approach to select relevant miRNAs that are directly responsible for cancer diagnosis.

2. MATERIAL AND METHODS

In this section, we describe the data sets and the methodology to model predictive framework with different classifiers. This study proposes a two-layered framework that consists of marker selection, and classification. Figure1 describes the overall predictive framework. In addition, we explain the performance metrics used to evaluate the generalization ability of the five different classifiers.

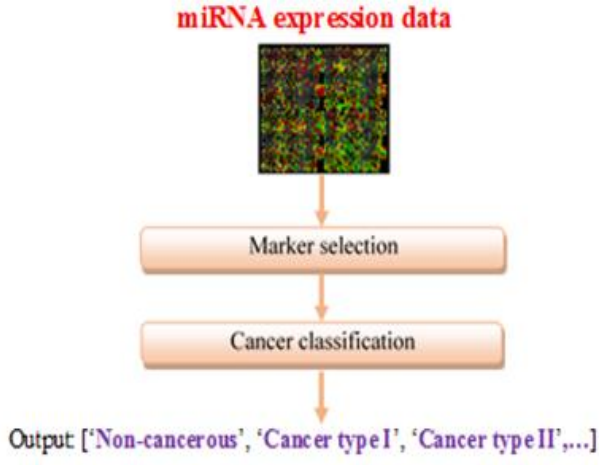

Fig. 1: Predictive framework

Data Set

The proposed framework has been tested on eleven publicly available datasets as listed in Table I. Each dataset contains more than 20 samples with more than 300 miRNAs.

Marker Selection

This section deals with the first layer of the cancer classification framework which is the marker selection process. Ideally, the best subset that contains the least number of miRNA markers that most contribute to the classification accuracy should be chosen, while discarding the rest of the miRNA. The aim of marker selection is to pick a subset of miRNA markers, $\mathbb{S} \subset G$, that can sufficiently discriminate cancer type $\boldsymbol{Y}$, given $|\mathbb{S}| = \mathcal{s}$ where ($\mathcal{s} \ll$ g). Information gain (IG) of a miRNA to the class label $\boldsymbol{Y}$ is defined as:

$$IG(i;\mathbb{S}) \stackrel{\text{def}}{=} H(\boldsymbol{Y}|\boldsymbol{X}_{\mathcal{s}}) - H\left(\boldsymbol{Y};\boldsymbol{X}_{\mathbb{S}\cup\{i\}}\right) \quad (1)$$

The gain ratio (GR) between Y and s is defined to be:

$$GR(i;\mathbb{S}) \stackrel{\text{def}}{=} \frac{IG(i;\mathbb{S})}{H(\boldsymbol{Y})} \quad (2)$$

For each iteration, the miRNA i that maximizes GR is chosen according to Equation 3.

$$arg\max_{i} GR(i;\mathbb{S}) \quad (3)$$

The cut-off condition can be set either based on $|\mathbb{S}|$ or $\frac{d\,GR(i;\mathbb{S})}{d|\mathbb{S}|} \approx 0$.

Classification

Classifiers perform statistical inferences on the miRNA expression vector and estimate the likelihood of the vector belonging to each cancer class. This study employed five different miRNA based classifiers which are Random Forest (RF), Hoeffding Tree, KStar (K*), JRip, and AdaBoostM1.

Random Forest

Random forests (RF) is a one of the most prominent decision tree (DT) algorithms in the literature. It is formulated as a recursive partitioning scheme of the expression vector space. Finally, input expression vectors are classified by traversing the top of the tree down to a leaf in accordance with the decisions along the route. Random Forest classifier is not sensitive to outliers and noise [14], it is very suitable for classification of miRNA expression data. Random Forest classifier able to reduce bias in supervised classification [15] . The Random Forest algorithm will be used to perform final inference on miRNA expression vectors and produce their corresponding cancer types as output.

Hoeffding Tree

Hoeffding Tree also known as very fast decision tree (VFDT) is a type of decision tree algorithm which can be used to perform miRNA expression data classification [16] . It utilizes the Hoeffding bound to measure error rates in terms of some distance metric which is absolute distance and to build nodes and branches [17] . Initially, there is only one node at the top, which is branched into a few sub-nodes, which are then branched further into lower-level sub-nodes. Each node divides the expression vector space from the parent node into two or more sub-spaces in accordance with a definite discrete function of expression values. As the name suggests, this algorithm is fast, efficient and produces accurate classification outputs.

KStar (K*)

KStar (K*) classifier is a popular strain of instance-based classifier. Instance-based cancer classification is arguably one of the easiest ways to classify cancer. Instance-based learning is also called memory-based learning or "lazy" learning because there is no training phase. In the prediction phase, given an expression vector $\boldsymbol{x} \in \mathbb{R}^n$, an instance-based cancer classifier sequentially loops through all the samples and chooses the class label of the sample that is most similar to $\boldsymbol{x}$ in terms of some distance metric, which is entropy distance, as output [18] . A (K*) simply chooses the majority of the class labels of the ƙ samples that are nearest to $\boldsymbol{x}$.

Ripper

Repeated Incremental Pruning to produce error reduction (Ripper) is a direct method that extracts the rules directly from the data. Classes are determined through grow and pruning phases [19] . Initially, the rule set $= \{\}$, and the grow phase continuously adds markers i to the rule set using learn-one-rule function until the stop threshold has reached. Following the growth phase, pruning phase performs repeated incremental pruning resulting in error reduction [20] . Finally,

the best rule is produced that can sufficiently discriminate classes of cancer, given an expression vector $\boldsymbol{x} \in \mathbb{R}^n$.

AdaBoostM1

Adaptive boosting (AdaBoost) is an ensemble meta-algorithm armed with multiple weak classifiers to classify cancer. Ensemble learning is a process by which weak classifiers are purposefully commingled to craft a single strong classifier [21] . The weak classifiers are called 'base learners'. Committee voting, which simply lets the base learners 'vote' for the cancer class $\mathcal{C}$ given expression vector $\boldsymbol{x}$ and chooses the majority. A weighted voting scheme, where weights are gradually incremented for consistently accurate base learners, causes both bias and variance reduction [22] .

Table 1: Datasets

Datasets	#miRNAs	#Samples
Meningioma (MG)	331	34
Breast Cancer (BC)	851	53
Gastric Cancer (GC)	704	41
DLBCL	174	55
Hepatocellular carcinoma (HCC)	856	146
Pancreatic Cancer (PaC)	1205	31
Malignant schwannoma (MS)	339	24
Prostate Cancer (PS)	373	142
Ovarian Cancer (OC)	379	84
Colorectal Cancer (CRC)	377	157
Multiple Myeloma (MM)	366	57

Validation

In order to assess the feasibility and validity of the proposed predictive framework, leave-one-out cross-validation (LOOCV) was applied [23] . The results will be then averaged to produce an estimate of the accuracy of the system. The following performance metrics were used to gauge the performance of the system: Accuracy and F-measure [24, 25] . F-Measure, $\mathbb{F}$, combines both precision and recall as a unified index.

3. PERFORMANCE ANALYSIS

All the 12 cancer datasets listed in Table 1 were used to test the performance of proposed classifiers. We performed leave-one-out cross-validations (LOOCV) where an N-sized dataset was partitioned into N equal-sized sub-datasets. Out of the N sub-datasets, a single sub-dataset was retained as the validation data for testing the model, and the remaining N- 1 sub-datasets were used as training data. The whole cross-validation process was then repeated N- 1 more times such that each of the N sub-datasets got used exactly once as the validation data. The results were then averaged over all the N trials. Figure 2 benchmarks the maximum accuracy rates of five classifiers for 11 datasets. The vertical axis represents the accuracy of the classifiers in percentage while the horizontal axis represents the tested cancer datasets.

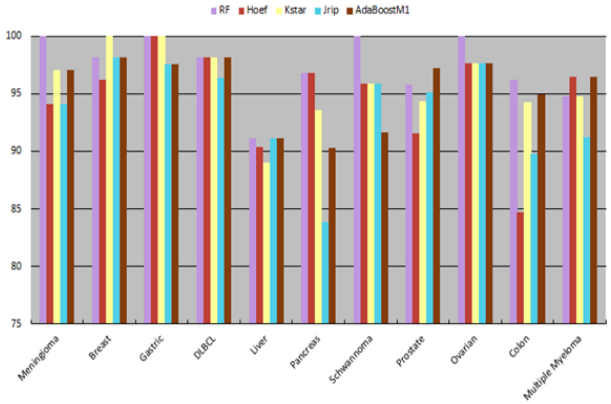

Figure 2: Benchmarking on LOOCV accuracy rates of five classifiers over 11 datasets

In addition, Table 2 lists the same set of results of five classifiers together with selected miRNAs for 11 datasets in a tabular format. The miRNAs for each trail were selected using gain ratio (GR) marker selection technique outlined in section 2.2.

Table 2: Maximum accuracy rates of the state-of-the-art classifiers over 11 datasets

Dataset	Accuracy (RF)	# miRNA	Accuracy (Hoef)	# miRNA	Accuracy (K*)	# miRNAs	Accuracy (JRip)	# miRNA	Accuracy (AdaB)	# miRNA
MG	100%	2	94.12%	10	97.06%	1	94.12%	1	97.06%	5
BC	98.11%	3	96.23%	1	100%	2	98.11%	1	98.11%	1
GC	100%	1	100%	1	100%	1	97.56%	1	97.56%	1
DLBCL	98.18%	2	98.18%	1	98.18%	5	96.36%	1	98.18%	1
HCC	91.10%	5	90.41%	15	89.04%	10	91.10%	20	91.10%	10
PaC	96.77%	12	96.77%	10	93.55%	15	83.87%	10	90.32%	5
MS	100%	6	95.83%	1	95.83%	1	95.83%	10	91.67%	1
PC	95.77%	8	92.25%	20	94.37%	15	95.07%	10	97.18%	10
OC	100%	3	97.62%	5	97.62%	1	97.62%	1	97.62%	1
CRC	96.18%	14	84.71%	20	94.27%	15	89.81%	15	94.90%	15
MM	94.74%	9	96.49%	10	94.74%	5	91.23%	1	96.49%	5
	μ_{RF} = 97.35%		μ_{A1DE} = 94.72%		μ_{K^*} =95.88%		μ_{JRip} = 93.70%		μ_{AdaB} = 95.47%	

In the training phase, the marker selection algorithms uncover markers based on a certain threshold. In the prediction phase, given an expression vector $\boldsymbol{x} \in \mathbb{R}^n$, all the miRNA are removed except the relevant miRNA markers. Finally, the output expression vector is $\boldsymbol{x} \in \mathbb{R}^s$ with reduced dimensionality. Therefore, the marker selection process is an imperative stepping stone to accurate and reliable cancer classification.

The maximum LOOCV accuracy of Random forest (RF) is 100% for 4 out of 11 datasets namely, meningioma, gastric cancer, malignant schwannoma and ovarian cancer. Similarly, KStar (K*) achieved maximum accuracy of 100% for 2 out of 11 datasets namely, gastric cancer and breast cancer. Likewise, Hoeffding (Hoef) achieved maximum accuracy of 100% for gastric cancer data set with no misclassification. However, the highest accuracy rates of JRip and AdaBoostM1 (AdaB) over 11 datasets are 98.11% and 98.18% respectively. As shown in Table 2, the average maximum LOOCV accuracy of RF classifier (μ_{RF}) is 97.35%, μ_{K^*} is 95.88% and μ_{Hoef} is 94.72% across all the 11 datasets. Interestingly, μ_{AdaB} is very close to K* and higher than μ_{Hoef} across all the 11 datasets which is 95.47% while μ_{JRip} is the lowest among all five classifies**.** Based on the F-measures results, out of five classifiers, RF was found to have best classification performance. Figure 3-7 illustrates the weighted F-measure of state-of-the-art classifiers for varying number of miRNAs for 11 datasets. The vertical axis represents the weighted F-measure of the classifiers while the horizontal axis represents the number of miRNAs. A weighted F-measure is simply F-measures of distinct cancer class labels weighted by the prior probability values of the class labels. The RF obtained perfect F-measures which is 1.0 for 4 out of 11 cancer datasets and that it outperformed the rest of the classifiers. The results show that accuracy does increase with the number of selected miRNAs, albeit without perfect monotonicity. Results also show that at certain instances, accuracy decreases with an increase in the number of miRNAs, departing from monotonicity. This may be because all the proposed classifiers are sensitive to the presence of irrelevant attributes. Furthermore, the disruptions in monotonicity might be because of the intrinsic imperfection in the proposed marker selection process [26] . The accuracy rate of the cancer recognition system using the Random forest classifier with the gain ratio marker selection process seems to be higher than the other four cancer classifiers.

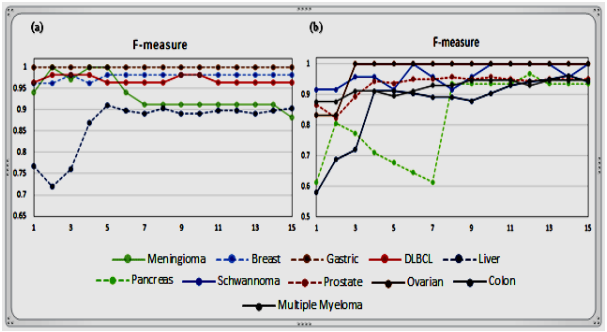

Fig. 3: Weighted F-measure vs. no. of miRNAs of Random forest classifier with gain ration marker selection for 11 datasets.

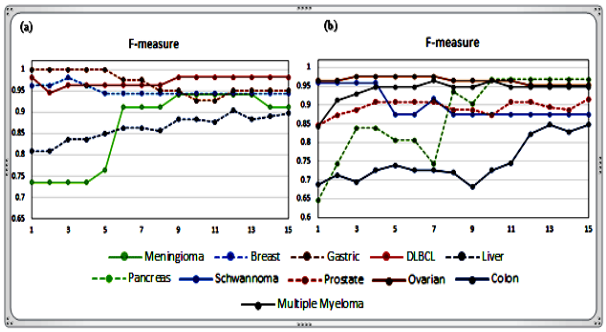

Fig. 4: Weighted F-measure vs. no. of miRNAs of Hoeffding tree classifier with gain ration marker selection for 11 datasets.

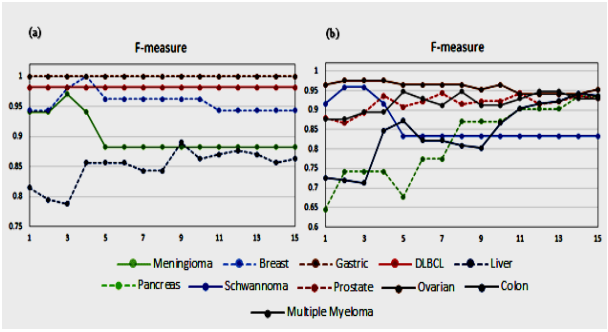

Fig. 5: Weighted F-measure vs. no. of miRNAs of K* classifier with gain ration marker selection for 11 datasets.

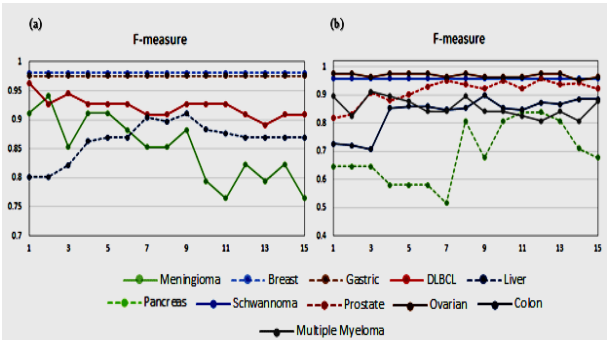

Fig. 6: Weighted F-measure vs. no. of miRNAs of JRip classifier with gain ration marker selection for 11 datasets.

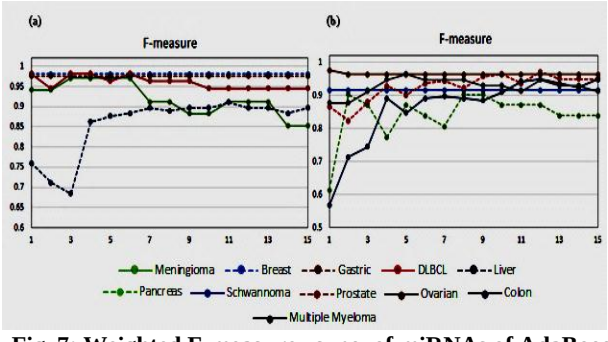

Fig. 7: Weighted F-measure vs. no. of miRNAs of AdaBoostM1 classifier with gain ration marker selection for 11 datasets.

4. CONCLUSION

This paper benchmarked the performance of five classification algorithms to classify cancers from miRNA expression data. While comparing the performance metrics among the classifiers, Random forest results are more

encouraging. Random Forest (RF) has achieved an average accuracy of 97% over 11 independent datasets. The experimental results are quite encouraging and the predictive framework managed to classify cancer accurately even with much noise contaminated in the datasets.

5. REFERENCES

[1] Mäbert, K., M. Cojoc, C. Peitzsch, I. Kurth, S. Souchelnytskyi and A. Dubrovska. "Cancer biomarker discovery: current status and future perspectives." International journal of radiation biology 90(8): 659-677 (2014).

[2] Croce, C. M., C.-g. Liu, G. A. Calin and C. Sevignani. Diagnosis and Treatment of Cancers with microRNA Located in or Near Cancer-Associated Chromosomal Features, US Patent 20,150,368,647 (2015).

[3] Pereira, D. M., P. M. Rodrigues, P. M. Borralho and C. M. Rodrigues. "Delivering the promise of miRNA cancer therapeutics." Drug discovery today 18(5): 282-289 (2013).

[4] Xie, B., Q. Ding, H. Han and D. Wu. "miRCancer: a microRNA–cancer association database constructed by text mining on literature." Bioinformatics: btt014 (2013).

[5] Eastlack, S. C. and S. K. Alahari. "MicroRNA and Breast Cancer: Understanding Pathogenesis, Improving Management." Non-Coding RNA 1(1): 17-43 (2015).

[6] Rane, J. K., M. Scaravilli, A. Ylipää, D. Pellacani, V. M. Mann, M. S. Simms, M. Nykter, A. T. Collins, T. Visakorpi and N. J. Maitland. "MicroRNA expression profile of primary prostate cancer stem cells as a source of biomarkers and therapeutic targets." European urology 67(1): 7-10 (2015).

[7] Mo, M.-H., L. Chen, Y. Fu, W. Wang and S. W. Fu. "Cell-free circulating miRNA biomarkers in cancer." Journal of Cancer 3: 432 (2012).

[8] Zhong, X., G. Coukos and L. Zhang. miRNAs in human cancer. Next-Generation MicroRNA Expression Profiling Technology, Springer: 295-306 (2012).

[9] Farazi, T. A., J. I. Hoell, P. Morozov and T. Tuschl. MicroRNAs in human cancer. MicroRNA Cancer Regulation, Springer: 1-20 (2013).

[10] Tanić, M., K. Yanowski, E. Andrés, G. Gómez-López, M. R.-P. Socorro, D. G. Pisano, B. Martinez-Delgado and J. Benítez (2015). "miRNA expression profiling of formalin-fixed paraffin-embedded (FFPE) hereditary breast tumors." Genomics data 3: 75-79.

[11] Perell, K., M. Vincent, B. Vainer, B. L. Petersen, B. Federspiel, A. K. Møller, M. Madsen, N. R. Hansen, L. Friis-Hansen and F. C. Nielsen. "Development and validation of a microRNA based diagnostic assay for primary tumor site classification of liver core biopsies." Molecular oncology 9(1): 68-77 (2015).

[12] Saha, S., S. Mitra and R. K. Yadav. "A multiobjective based automatic framework for classifying cancer-microRNA biomarkers." Gene Reports 4: 91-103 (2016).

[13] Kim, S.-J., J.-W. Ha and B.-T. Zhang. "Constructing higher-order miRNA-mRNA interaction networks in prostate cancer via hypergraph-based learning." BMC Systems Biology 7(1): 1-16 (2013).

[14] Deng, H. and G. Runger. "Gene selection with guided regularized random forest." Pattern Recognition 46(12): 3483-3489 (2013).

[15] Erho, N., A. Crisan, I. A. Vergara, A. P. Mitra, M. Ghadessi, C. Buerki, E. J. Bergstralh, T. Kollmeyer, S. Fink and Z. Haddad. "Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy." PloS one 8(6): e66855 (2013).

[16] Kourtellis, N., G. D. F. Morales, A. Bifet and A. Murdopo. "VHT: Vertical Hoeffding Tree." arXiv preprint arXiv: 1607.08325 (2016).

[17] Desai, H., D. Vasiyani and J. Gandhi. "A Survey on Data Stream and Its Various Techniques." (2015).

[18] Shehata, M., F. Khalifa, A. Soliman, A. Takieldeen, M. A. El-Ghar, A. Shaffie, A. C. Dwyer, R. Ouseph, A. El-Baz and R. Keynton. 3D diffusion MRI-based CAD system for early diagnosis of acute renal rejection. Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on, IEEE (2016).

[19] Feng, H., Y. Chen, K. Zou, L. Liu, Q. Zhu, Z. Ran, L. Yao, L. Ji and S. Liu. A New Rough Set Based Classification Rule Generation Algorithm (RGA). International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Springer (2014).

[20] Asadi, S. and J. Shahrabi. "RipMC: RIPPER for Multiclass Classification." Neurocomputing 191: 19-33 (2016).

[21] Dubout, C. and F. Fleuret. "Adaptive sampling for large scale boosting." Journal of Machine Learning Research 15(1): 1431-1453 (2014).

[22] Cheki, M., H. Jazayeri-Rad and P. Karimi. "Enhancing the noise tolerance of fault diagnosis system using the modified adaptive boosting algorithm." Journal of Natural Gas Science and Engineering 29: 303-310 (2016).

[23] Natrella, M. G. Experimental statistics, Courier Corporation (2013).

[24] Hernández-Orallo, J., P. Flach and C. Ferri. "A unified view of performance metrics: Translating threshold choice into expected classification loss." Journal of Machine Learning Research 13(Oct): 2813-2869 (2012).

[25] Koyejo, O. O., N. Natarajan, P. K. Ravikumar and I. S. Dhillon. Consistent binary classification with generalized performance metrics. Advances in Neural Information Processing Systems (2014).

[26] Frank, R. L., E. Lenzmann and L. Silvestre. "Uniqueness of radial solutions for the fractional Laplacian." Communications on Pure and Applied Mathematics (2015).