 doi:10.20944/preprints202009.0013.v1

A Survey of Techniques Leveraging miRNA as Biomarkers for Generalized and Type Specific Cancer Diagnosis.

Abhik Banerjee,
abhik.banerjee.1999@gmail.com,
Department of CSE,
Netaji Subhash Engineering College,
Kolkata-152

Shilpi Bose.
bose.shilpi08@gmail.com,
Department of CSE,
Netaji Subhash Engineering College,
Kolkata-152.

Dr. Chandra Das.
daschandraresearch@gmail.com,
Department of CSE,
Netaji Subhash Engineering College,
Kolkata-152

Abstract: MicroRNAs are used as biomarkers for classification of cancer subtypes since certain miRNAs are differentially expressed in normal and patient samples. Moreover, miRNAs target mRNAs and can heavily influence Gene Expressions. Thus, deregulation of miRNAs is linked to various disorders. Thus, miRNAs can be used for prognosis and developing personalized health solutions for patients. Given the importance of miRNAs, there has been substantial work done in the field. In this paper, recent works in the field of using miRNAs expressions of patients were considered. A total of 20 papers were surveyed which utilized feature selection ensembles, fuzzy logic as well as deep learning. 10 papers have been reported which offer insight into how miRNAs can be utilized for subtype-specific or generalized cancer diagnosis.

1. Introduction.

Since 1993, when the first MicroRNA was discovered in C.Elegans, a lot of research has been done in finding miRNAs and their properties. MicroRNAs are non-coding RNAs of length 22nt~.

 doi:10.20944/preprints202009.0013.v1

They regulate gene expression and also influence MessengerRNAs and gene expression [1][2][3]. Because of this reason, over the last two decades, a considerable amount of work has been done on miRNAs. miRNAs influence mRNA in the protein translation process and are regarded as mRNA 'fine-tuners' [6]. Their deregulation can be the cause of disorders and these can be detected from their expression in a patient. This has been validated by several studies done on miRNA. It has been also found that miRNAs are differentially expressed across patients suffering from disorders and normal people. This has led to research in using miRNAs as potential biomarkers for diseases like cancer. The usefulness of miRNAs is further increased when one considers that these expressions can be reproduced across patients [7]. While there are miRNAs which generalize their expression across multiple disorders, not all miRNAs may be important for cancer prognosis.

Given the fact that differentially expressed microRNAs can be used in diagnosis, it has also been conjectured that microRNAs can be used for personalized treatments of cancer subtypes as well. miRNAs are also used for target prediction and also find disease association. Since there are miRNAs which may be cancer-specific biomarkers [8][9][10][11], the use of all miRNAs may be costly both biologically and computationally. This has led to considerable work being done in finding methods for dimensionality reduction. Feature selection is preferred in this context most often over feature extraction since with feature selection one can get a subset of the original miRNAs from the dataset. This is useful for validation and references.

There have been many approaches to the problem. While some works utilize the full dataset for miRNA biomarker-based cancer classification, other approaches such as using ensembles[12-15], graph-based feature selection[16], fuzzy logic-based have been proposed [17]. Alteration of Gene Selection Algorithms can be used for this purpose[18-21].

The remainder of the paper is organized in the following manner:- Section 2 lists the chosen surveyed works which use feature selection, Section 3 lists works which utilize Fuzzy logic for miRNA biomarker classification, Section 4 lists a work where self-training and co-learning have been used while Section 5 reviews work leveraging Deep Neural Nets for biomarker classification in cancer. Results Summary of the papers have been presented in Section 6 and the scope of future works is discussed in Section 7. Section 8 concludes the paper.

2. miRNA/Gene Expression-integration and Feature Selection-based Cancer Prediction using miRNAs.

a. Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection (Lopez-Rincon et al.)

This work by Alejandro and team aims to find a robust method to identify miRNAs which may lead to better classification of cancer results without the use of prevalent invasive techniques. Their work succeeds in identifying 50 miRNAs which hold significant information with regards to cancer subtypes. While 35 of those miRNAs are

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 September 2020 doi:10.20944/preprints202009.0013.v1

already well-known and recorded as biomarkers for cancer prediction, Alejandro and collaborators introduce 15 new miRNAs which were previously undiscovered.

Given its highly efficient results, Next-Generation Sequencing technique miRNASeq BCGSC IlluminaHiSeq miRNASeq 55 Level 3 is leveraged to extract the expression of 1,046 miRNA features across 8,129 patients. This subset is taken from The Cancer Genome Atlas (TCGA) database [23]. The data is categorised into 29 labels or cancer subtypes. These include -

I. Adrenocortical carcinoma [ACC]
II. Bladder Urothelial Carcinoma [BLCA]
III. Breast invasive carcinoma [BRCA]
IV. Cervical squamous cell carcinoma [CESC]
V. Cholangiocarcinoma [CHOL]
VI. Esophageal carcinoma [ESCA]
VII. FFPE Pilot Phase II [FPPP]
VIII. Head and Neck squamous cell carcinoma [HNSC]
IX. Kidney Chromophobe [KICH]
X. Kidney renal clear cell carcinoma [KIRC]
XI. Kidney renal papillary cell carcinoma [KIRP]
XII. Lower Grade Glioma [LGG]
XIII. Liver hepatocellular carcinoma [LIHC]
XIV. Lung adenocarcinoma [LUAD]
XV. Lung squamous cell carcinoma [LUSC]
XVI. Lymphoid Neoplasm Diffuse Large B-cell Lymphoma [DLBC]
XVII. Mesothelioma [MESO]
XVIII. Pancreatic adenocarcinoma [PAAD]
XIX. Pheochromocytoma and Paraganglioma [PCPG]
XX. Prostate adenocarcinoma [PRAD]
XXI. Sarcoma [SARC]
XXII. Skin Cutaneous Melanoma [SKCM]
XXIII. Stomach adenocarcinoma [STAD]
XXIV. Testicular Germ Cell Tumors [TGCT]
XXV. Thymoma [THYM]
XXVI. Thyroid carcinoma [THCA]
XXVII. Uterine Carcinosarcoma [UCS]
XXVIII. Uterine Corpus Endometrial Carcinoma [UCEC]
XXIX. Uveal Melanoma [UVM]

The authors use the idea that a feature which is considered to be important or is assigned a higher weight by a classifier is most probably more informative compared to those in the dataset which receive a lower weightage by the classifier. Ensemble Feature Selection comes into play in this regard as the work uses a set of well-known classifiers to first classify the dataset. This means that the classifiers are run on a normalized dataset of all 1,046 miRNAs collected from patients. After this step, each miRNA is assessed based on its relative importance in the classification of patient labels by a specific classifier. The

 doi:10.20944/preprints202009.0013.v1

aggregate results are taken for each miRNA. As mentioned before, this procedure successfully identifies 35 known miRNAs in literature as well as gives 15 new unused/undiscovered miRNAs which are used by the classifier ensemble to classify the labels.

As is evident from the list of labels or cancer subtypes that the dataset, which the authors use, contains, the method can be used to identify miRNAs which can be used as general biomarkers rather than cancer subtype-specific biomarkers. The following are the classifiers which the authors have identified and listed for primary classification and feature selection task:-

I. BaggingClassifier
II. GradientBoostingClassifier
III. LogisticRegression
IV. PassiveAggressiveClassifier
V. RandomForestClassifier
VI. RidgeClassifier
VII. SGDClassifier
VIII. SVC (Support Vector Machine with Linear Kernel)

The algorithm given by Lopez-Rincon et al. can be surmised in the following steps:

I. Normalization of the Dataset and then division of the dataset into N folds. At each step LOOCV (Leave One Out Cross Validation) is used to train the aforementioned classifiers and then test them.
II. For each fold, all the 1,046 miRNA expression features are run as a classifier as training. The one remaining fold is used for testing.
III. The weights from the classification task are obtained from the classifier for each feature and the top 100 features are chosen.
IV. Steps b and c are repeated for all folds and all classifiers.
V. All the miRNA features from step c and assigned an aggregate score which depends on how many times they were selected by each classifier to be in top 100 features used for classification.
VI. Once the miRNAs are sorted in as the score in step c, the top 100 among those features are chosen and then the classification task is rerun on these 100 features only for each classifier.

This showed a relative increase in accuracy of classification of the samples by the classifiers on the top 100 miRNAs. The authors choose to list the top 50 of those features sorted by the frequency of their appearance post-classification on the reduced set of features.

MicroRNAs like hsa-let-7b, hsa-let-7f-1, hsa-let-7i, hsa-mir-29c referred in [43] as well as hsa-135-a-1 and hsa-103-1 were picked up by the procedure. These are well-renowned biomarkers that are used for cancer prediction.

b. Integrating MicroRNA and mRNA Expression Data for Cancer Classification (Oğul H. and Altındağ O.)

mRNA and microRNAs are known to affect gene regulatory networks. A considerable amount of research has been done to find out which is more effective - classification via mRNA expressions or by microRNA expression profile study. In this paper, the author put forward the notion that both of these sequences are important and that mRNA can complement microRNA expression data in that the information contained in both can be harnessed simultaneously for discriminating between healthy and cancerous patient samples.

The paper is based on the works of Peng et al.[26] and Lu et al. [25]. Lu et al [25] in their work stated that the study of microRNA expression data after proper feature selection can yield better results than that of messengerRNA expression data. This was contradicted by Peng et al.[26]. Ogul and Altindag present an approach whereby pairwise mRNA and miRNAs from each patient is obtained and used for classification tasks. The results show that classifiers tend to perform better on a pairwise combined dataset of mRNA and miRNA compared to the individual expressions.

For proving the conjecture, the authors choose to use three types of datasets. The first dataset is mRNA dataset is a subset of the data provided by Ramaswamy et al.[27] as Global Cancer Map (GCM) mRNA dataset. The subset contains a record of 16,063 gene expressions across 89 samples with 11 classes of data. The next dataset is taken from Lu et al. [25]'s bead-based flow cytometric miRNA. The same 89 samples are chosen from this data with 217 miRNA expression profiles. This helps in pairing up of the samples' mRNA and miRNA and create the third combined dataset. This mRNA-miRNA contains 16,280 features from respective datasets with the same samples and class labels.

The multi-class tumour classification task is performed on all three datasets to assess the performance of the combined dataset. The classifiers used for this are as follows:-

I. C4.5 Decision Tree (DT)
II. Artificial Neural Networks (ANN)
III. Support Vector Machines (SVM)
IV. Naïve Bayes multinomial classifier (NBM)
V. K-Nearest Neighbors (KNN).

The full datasets are used as a baseline performance measure for the proposed approach whereby all the features are run on the classifiers across the classifiers utilizing LOOCV for training-testing purposes. Since there are 5 classifiers chosen, the authors choose to measure the performance dataset wise across the classifier by assigning the datasets scores for every category of classifier they perform the best with.

Next, the authors perform feature selection on the datasets. To select the 100 most informative features from each dataset, the following attribute selection schemes are used:-

I. SVM-based attribute selection.

II. Information Gain-based attribute selection.
III. Gain Ratio-based attribute selection.
IV. Chi-square Test-based attribute selection.
V. CFS subset attribute selection.

For each attribute selection scheme, the three reduced datasets are run on the five aforementioned classifiers with the same dataset scoring scheme. The overall score of each dataset is just the aggregate of the scores obtained in each feature selection scheme. With this approach, the combined dataset is seen to have performed far better for classification tasks than the individual datasets with the best accuracy obtained on the Artificial Neural Network. This proved that the classifiers had a tendency to discriminate better between samples if they learn the expression profiles of both mRNA and miRNA for a given sample.

c. A New Direction of Cancer Classification: Positive Effect of Low-Ranking MicroRNAs (Li et al.)

The miRNA's which have higher measure scores such as F-Score, Information Gain with respect to class labels and ReliefF Scores among others are chosen for classification in most studies. These scores give an idea of the feature's discriminative power. Generally, the better the score, the better the feature can contribute to classification. However, this can often lead to important miRNA's being left out. The work by Li et al. dwells on whether the miRNA's which have a poor score can also contribute to the classification task in a positive manner. It is a well-known fact that a single miRNA can influence the expression of multiple genes. This can lead to relations between multiple miRNAs. The authors explore this possibility and try to leverage it in their proposed work.

The authors use a correlation-based feature selection approach with multiple methods like PSO, best-first search, tabu search and re-ranking search along with Pearson's correlation, Chi-square distribution, information gain, and gain ratio feature selection methods for references. The features are sorted in the descending order of their scores. The top-ranking features are chosen for the classification task using SVM, KNN, Bayes Network and Decision Trees. Next, low ranking features are taken and the same procedure is run on them.

The results from the work of Li and team present that the classification accuracy is actually increased when considering low-ranking features along with the high-ranking ones obtained from their feature selection approach. The datasets used to benchmark the performance are colon, pancreas, uterus, T cell acute lymphoblastic leukaemia (ALL), and B cell ALL. The accuracy when considering low-ranking features along with high-ranking features is 94.52% as opposed to the 89.04% accuracy using information gain as the feature selection measure. This work by Li and team thus proves that low-ranking miRNAs might also hold information about the class labels - information which can be utilized for a better result.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 September 2020 doi:10.20944/preprints202009.0013.v1

d. An SVM-wrapped Multiobjective Evolutionary Feature Selection Approach for Identifying Cancer-MicroRNA Markers (Mukhopadhyay A., Maulik U.).

With the inspiration coming directly from nature, Genetic algorithms are regarded as a robust and high-performant method to optimize a given set of conditions. There are many Genetic Algorithms like MOGA, Micro-GA, SPEA, PAES [37] which can be used for this task. Among these NSGA-II [38] is often used as a multi-objective optimization algorithm which can provide efficient solutions to the problem. This paper proposes the use of NSGA-II coupled with an SVM wrapper for selection of miRNA's which can be leveraged in cancer diagnosis.

NSGA-II algorithm is able to provide a better solution in terms of computation time when compared to NSGA. The procedure adopted for selecting a set of miRNAs as the final subset of the initial set is summarized below:-

I. The dataset samples are Standard Normalized. After dividing the training and test sets, the miRNAs are used in multi-objective optimization using NSGA-II where the objectives include:-
   i. Maximization of:-
      - f1 = Sensitivity = tp / tp + fn.
      - f2 = Specificity = tn / tn + fp.
   ii. Minimization of:-
      - f3 = |S|.

II. The SVM wrapper used along with NSGA-II produces a final set of miRNA sets which can act as the solution to the aforementioned objective functions.

III. The subset which maximizes
   i. F = 2 × (Precision × Recall) / (Precision + Recall) - where precision is Precision = tp / (tp + fp) and recall is Sensitivity
   Is selected for the as the set of miRNAs to be tested.

Dataset used by the authors is the same given by Lu et al.[25]. The dataset contains 217 mammalian miRNAs from which the authors have extracted breast, colon, kidney, lung, prostate and uterus cancer subtypes. Signal to Noise Ratio of each miRNA feature is used as a measure of its initial fitness. The top 100 miRNAs sorted in the descending order of their SNR score are selected for the process.

The resulting final dataset from the process is tested with the classifier results of LASSO [39][40] and SCAD[41]. The miRNAs selected by the proposed work perform better than those selected by both of the aforementioned procedures. The paper reports an accuracy of 89.80% on the miRNAs selected by the approach under discussion as opposed to 87.76% and 85.71% by the respective methods.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 September 2020 doi:10.20944/preprints202009.0013.v1

3. Fuzzy Set Approaches for miRNA Classification.

a. An application of fuzzy normalization in miRNA data for novel feature selection in cancer classification (Anidha M., Premalatha K.).

This work by authors Anidha M. and Premalatha K. aims to explore what happens when you have multiple fuzzy boundaries to make sure that every feature in the dataset has a smooth transition rather than just having a crisp or single transition from membership. The authors propose the use of 3 fuzzy membership functions rather than just one to obtain a membership of a miRNA feature. This results in every feature in the dataset definitely being placed in either one of the rules. However, since there are multiple membership functions assessing a gene expression value, there is also a chance of all of them ascertaining that the feature under discussion is a deemed member. This can easily be complemented by choosing the rule which has deemed the highest membership for a given expression value. The proposed approach, thus, provides a very smooth transition courtesy of the multiple membership functions at play. This approach is used to normalize the feature values before the actual feature selection is done.

For selecting the features, the authors employ F-Score along with an entropy-based mean score. The top miRNAs being picked from this approach are then tested for their effective discriminative power to distinguish between multiple class labels.

The normalized features are evaluated first using F-Score that gives a good measure of the miRNA's class separability and spread of each class value. This means that a feature with higher F-Score can be used to discriminate between better class labels. Next, the features are assessed using Relevant Information Gain. This helps in measuring redundancy between features. Finally, a ratio of F-Score and Relevant Information Gain is used to determine whether a feature should be selected.

The authors demonstrated the results using SVM and ANN classifiers. The datasets used for testing this approach included Angulo Dl's Lung Cancer, Takeuchi SU's Lung Cancer, WangY's Breast Cancer, VandeVijver_SU's Breast Cancer, Soutiriou's Breast Cancer, Soutiriou ER's Breast Cancer, and WangQ ST Neurobi (Early Stage) datasets.

The highest accuracy for this approach was found to be 100 % on both SVM and ANN on Angulo Dl's Breast Cancer Dataset by selecting the top 100 miRNA's evaluated from their ratio of F-Score and Relevant Information Gain.

b. Identifying Relevant Group of miRNAs in Cancer using Fuzzy Mutual Information (Pal J.K., Ray S. S., Pal S.K.).

There are many other techniques which are based on Fuzzy Theory that can be employed for feature selection. Likes of these include Fuzzy V Score, Fuzzy Chi-Square Score and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 September 2020 doi:10.20944/preprints202009.0013.v1

Fuzzy Mutual Information. The authors in this propose the use of Fuzzy Mutual Information measure to identify miRNAs which contain information useful for the classification of cancer labels. Their work focuses on two fields - firstly, selecting a subset of miRNAs from the total set which can contribute most. Secondly, having an efficient and fast method to weed out redundancy among the subset chosen so that only cream crops of miRNAs remain in the final set.

The common problem with omics dataset is the unbalanced distribution and population of sample types. For this reason, a proper measure has to be employed which can assess the class separability and distribution. To tackle this problem, the authors choose to create a class-wise representative and then compare those representatives.

The algorithm proposed by Pal J.K., Ray S.S. and Pal S.K. can be surmised in the following steps:-

I. Calculation of average inter-class as well as the intra-class distance between all the labels of a miRNA. The ratio of inter-class to intra-class scaled distance gives a measure of class separability.

II. Step 1 is repeated for all miRNAs available. After the class separability measure is obtained, for each miRNA the class representative is calculated. The class representatives for a given miRNA is essentially the mean location of all the sample expressions of a specific class scaled by the standard deviation.

III. The miRNAs are sorted in the descending order of their class separability obtained from step 1. The top-ranking miRNA is selected and an SVM is trained using the class representatives. Those miRNAs from the sorted miRNAs which are correctly classified by the SVM are grouped into the miRNA under discussion.

IV. Step 3 is repeated until all the miRNAs are assigned to a group. Thus, the first stage of the proposed method concludes.

V. For each miRNA in a given group, the Fuzzy Mutual Information is calculated compared to the class label. This is done for every miRNA in the group. The average of these values is the relevance score of this group. The group with the highest score is chosen as the best group.

VI. To discard the redundant miRNAs in the chosen group, pairwise Fuzzy Information Gain Relevance is calculated. The average redundancy of a miRNA across all pairs containing that miRNA is the score of that miRNA. The top N miRNAs are chosen from the group after ranking them as per their scores.

The proposed approach is applied to breast, renal, colorectal, lung, melanoma and prostate cancer datasets. The miRNAs picked up by the method proposed by the authors are assessed on SVM and kNN classifiers. The miRNAs from the first and second groups obtained by the aforementioned procedure record a substantially better performance compared to the accuracy obtained on using all the miRNAs. The performance is also compared to the performance of SVM with Recursive Feature Elimination (SVMRFE), maximum relevance & minimum redundancy (MRMR) approach and a combined SVMRFE & MRMR approach. The performance of features selected via Fuzzy Mutual

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 September 2020 doi:10.20944/preprints202009.0013.v1

Information based approach is better than SVMRFE, MRMR and SVMRFE & MRMR approach.

The biological relevance of the miRNAs picked by the authors' method is also assessed in the paper. Whereas 12 out of the 15 miRNAs are recorded to be generalized and specific biomarkers of cancer subtypes, 3 novel miRNAs are also picked up by the approach.

c. Fuzzy mutual information-based grouping and new fitness function for PSO in selection of miRNAs in cancer (Pal J.K., Ray S.S., Pal S.K.).

This work by Pal, Ray and Pal borrows certain elements from their previously mentioned work titled "Identifying Relevant Group of miRNAs in Cancer using Fuzzy Mutual Information". The features are ranked using Fuzzy Mutual Information and then grouped in a similar manner so as to obtain a robust set of features. However, the novelty of the work lies in the use of FMI as a fitness function for Particle Swarm Optimization and subsequent usage of the module/framework to create groupings of miRNAs from a given set of miRNAs of a cancer type.

The use of FMI as fitness for PSO has been right alongside the usage of FMI for grouping using SVM for group formation. The resulting miRNAs are tested on colorectal, lung, pancreas, cancers, nasopharyngeal carcinoma and melanoma datasets with SVM and kNN classifiers using Leave One Out Cross Validation (LOOCV). The reported results of the proposed work show that the PSO fitted with FMI is able to outperform SVMRFE, MRMR and SVMRFE with MRMR methods on the aforementioned datasets.

Particle Swarm Optimization is an algorithm which is very much utilized in swarm intelligence and relies on each particle in the swarm finding its optimal best location and always keeping a track of the overall best position attained by any other particle which may be present in the swarm. This idea is utilized in the work done by the authors where all the miRNAs are grouped into "particles". Thus, instead of using each miRNA as a particle and the whole dataset as a swarm to optimize a certain set of criteria, the groups of miRNAs are taken as particles and all the groups together form a swarm.

The idea is to find a particle, which is essentially a grouping of miRNAs, which has been able to attain the best possible position, the best possible position being a condition where the group has minimum redundancy, i.e., the redundancy between one miRNAs and another having same information content for discrimination is reduced and the overall relevance is maximized. The optimality or fitness of a grouping or particle is decided by the average Fuzzy Mutual Information of all miRNAs in that "particle".

The main focus of the work done by the authors is to assess the performance of Fuzzy Mutual Information based grouping and selection of miRNAs and find the biological relevance of the results. Using the procedure, the authors were able to detect miRNAs which could be used as cancer subtype-specific biomarkers. The miRNA interaction has been validated using DIANA and Starbase.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 September 2020 doi:10.20944/preprints202009.0013.v1

4. Use of Co-Training and Self-Learning.

a. miRNA and Gene Expression based Cancer Classification using Self-Learning and Co-Training Approaches (Ibrahim et al.)

In the paper "miRNA and Gene Expression based Cancer Classification using Self-Learning and Co-Training Approaches" by Ibrahim et al., the use of self-learning and co-training as a method for classification of cancer subtypes has been explored. The paper under discussion employs semi-supervised learning to leverage the publicly available unlabelled as well as a labelled dataset to build a classifier which can predict and classify between cancer subtypes.

The method begins with the labelled datasets where the classifiers are trained and tested. Once this phase completes, both self-learning and co-training based methods of classification use the unlabelled datasets to train the classifiers on the data that it has confidently been able to predict (a confidence threshold is chosen for this task). This procedure continues till the unlabelled dataset has been exhausted and the classifiers are completely trained.

The authors of the paper use three cancer dataset types. These include breast cancer dataset, lung cancer and hepatocellular carcinoma (HCC) dataset on Random Forests and SVM-based classifiers. The results showed considerable improvement in classification tasks - "*The results show around 20% improvement in F1-measure in breast cancer, around 10% improvement in precision in metastatic HCC cancer and 3% improvement in F1-measure in squamous lung cancer over the Random Forests and SVM classifiers.*" ("miRNA and Gene Expression based Cancer Classification using Self-Learning and Co-Training Approaches" by Ibrahim et al.,).

The algorithm takes into account that miRNAs can influence more than one Genes and multiple Genes may, thus, share or be targeted by the same miRNA's. This fact is used in mapping Unlabelled miRNA's to their target Genes in the Gene Expression dataset and vice versa during co-training. Since a single miRNA may be replaced by multiple Gene Expressions, an average of those is considered for training the classifier for miRNA tasks. A similar method is adopted when mapping miRNAs to a single Gene Expression.

The paper details the self-learning task as follows:-

I. A classifier is trained on the Labeled Dataset.
II. This trained classifier is used to infer on the Unlabelled Dataset.
III. The predictions of subtypes which are above a certain confidence threshold are removed from the Unlabelled dataset and added on to the labelled datasets with the labels predicted by the classifier.
IV. The classifier is trained from scratch on this new dataset.
V. Step 3 and 4 are repeated until the Unlabelled dataset has been exhausted.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 September 2020 doi:10.20944/preprints202009.0013.v1

In the co-training adaptation, the following steps are taken:-

I. Two separate classifiers are trained on miRNA and Gene Expressions dataset for cancer respectively.
II. Now the respective classifiers are used to classify the unlabelled data. At this stage, the classifier trained on miRNA classifies unlabelled miRNA and same for Gene Expression.
III. The inferred results above a certain threshold are chosen. Up till this point, the process is similar to self-learning. From here on out, the classifiers start to "co-learn".
IV. Once the new labelled dataset is constructed, the miRNA to Gene relations are retrieved using miRanda[43] and the replacement in the newly formed labelled dataset. The classifiers are now re-trained on the new dataset.
V. This is where the classifier trained on miRNA trains to classify on the new labelled dataset formed from previously labelled data and new labelled data obtained by mapping miRNAs to target Genes' Expression.
VI. The process in step 3, 4 and 5 are repeated until there is no appreciable increase in performance or the unlabelled dataset runs out.

"miRNA and Gene Expression based Cancer Classification using Self-Learning and Co-Training Approaches" by Ibrahim et al. explores one of the first ventures of co-training and self-learning in the domain of cancer prediction using miRNA as well as Gene Expression data.

5. Deep Learning-based approach

a. Evolutionary Optimization of Convolutional Neural Networks for Cancer miRNA Biomarkers Classification (Lincoin et al.).

Convolutional Neural Networks has been shown to have better performance than many other network types depending on the problem. CNNs are one of the most decorated neural network types in the field of Computer Vision. Sequence data is often run on CNNs as a benchmark against other neural network types and architectures.

This paper by Lincoin et al. explores the efficiency of a Convolutional Neural Network Architecture when the said architecture is used in conjunction with an Evolutionary Algorithm to assess the hyperparameter tuning. The architecture is used on miRNA expression data for potential biomarker classification of patients.

Evolutionary algorithms are used for many complex problems and are favoured because of their performance. In the paper under discussion, Evolutionary Algorithm has been used to control hyperparameters such as the number of output channels, the width of the convolutional kernel, the width of the max-pooling after the convolution. The authors combine 1 convolutional layer and 1 max-pooling layer into a single convolution. Thus, at each step, the convolution preserves the spatial information and the max-pooling

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 September 2020 doi:10.20944/preprints202009.0013.v1

reduces the dimensions while conserving the information content of the miRNA. The number of convolutional layers has been fixed to 3 in this case.

The approach of the authors to the problem of using miRNA biomarker-based cancer classification can be surmised in the following steps:-

I. The data is normalized and the initial hyperparameters are set for the CNN.
II. The Evolutionary algorithm is given to work with the initial population of where the hyperparameters make up each individual of the population. The CNN is trained and tested with 10-fold cross-validation and the accuracy is checked for each individual of the population.
III. In the next generation, two individuals (essentially a collection of hyperparameters in an array) are taken and a random cut is applied on. The children produced are via interchanging the hyperparameters. The children thus obtained from the next generation. The population is checked based on the hyperparameters and trimmed is necessary.
IV. Steps 2 and 3 are repeated until the stopping criterion is reached. The last remaining generation is assessed based on the accuracy metric and the best individual in the population is selected.

The dataset used by the authors is taken from the Cancer Genome Atlas and contains a total of 8,129 patient samples across 29 subtypes of cancer with 1,046 miRNAs. This dataset is normalized and is used for 10-fold cross-validation. No feature selection has been applied on the original dataset and all the miRNAs are utilized as input to the CNN for classification. As a reference, the authors use 21 classifiers to test the accuracy on the dataset with all 1,046 miRNAs. The reported accuracy obtained by the authors from their approach is 96.6%.

b. Neural Network Cascade Optimizes MicroRNA Biomarker Selection for Nasopharyngeal Cancer Prognosis (Zhu W., Kan X.).

This paper uses a novel neural network architecture to explore whether a linear relation can be indirectly attained between miRNAs and mortality rate due to a certain disease (case in point, nasopharyngeal carcinoma). The neural network architecture termed as “neural network cascade” is used to obtain a score from the miRNAs and the relation between the score and the mortality rate is ascertained.

The authors utilized miRNA expression dataset taken from Gene Expression Omnibus. A total of 873 miRNA features of 312 nasopharyngeal carcinoma (NPC) were selected for the proposed method. The features were normalized before being used. Furthermore, Spearman’s Correlation coefficient was calculated for each miRNA to assess the relationship between the miRNA and patient survivability. This was a measure to rank the miRNA’s based on their efficiency of discriminating between two classes of patients present in the dataset - alive and dead. The miRNAs were ranked based on the coefficient

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 September 2020 doi:10.20944/preprints202009.0013.v1

and then potential biomarkers were selected from among the sorted miRNAs. The authors chose the top 9 miRNAs from the sorted miRNAs.

The neural network architecture proposed by the authors in the paper consists of several small neural networks. Each neural network unit functions independently from the rest. The neural networks are cascaded to form a pyramid-type structure. The units of ANN which were used have the following architecture that constitutes the Neural Network Cascade:-

I. At the base of the pyramid, an ANN with a 1-11-1 unit structure takes in the miRNA expression value. There are a total of 9 miRNAs selected. Hence, at the base, there are 9 such ANN units - each taking in the normalized expression value from each miRNA.

II. At the next layer of the pyramid, three consecutive ANN units are taken together and fed into one ANN with a 3-11-1 unit structure. This unit takes in the transformed expression value of each miRNA and then computes the overall score from these units. Note that there is no specific formula here to calculate the score given by this ANN unit. Since there are 9 ANN units, 3 such ANN units are fed into these ANN units in this layer. There are 3 such ANN units in this layer.

III. In the final layer of the cascade, a 3-11-1 ANN unit takes in the output of the ANN unit from the previous layer and generates the score that is to be used to assess the relationship with the survivability of the patient.

The scores obtained from the NNC after the miRNA expression values of the samples were fed was then subjected to a Student's t-test. The scores vs the survivability of the patient were the subject of the test. The area under the ROC curve was used to assess the prediction performance. The neural network cascade covered 0.951 area.

6. Result Summary.

This section summaries the results from the papers which were surveyed and mentioned in this paper. Note that while ideally, every paper should have been mentioned, not all works use the same classifiers or have reported every intermediate result. As such the best result reported by the authors is considered for the table given below in order to summarize their works and have a comparison with other works. The conditions where these results were obtained (as reported) have also been surmised.

Author	Title	Classifiers				
		SVM	kNN	NB	ANN	DT
1. Lopez-Rincon et al. [22]	Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection	0.9546	-	-	-	-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 September 2020 doi:10.20944/preprints202009.0013.v1

2. Oğul et al. [24]	Integrating MicroRNA and mRNA Expression Data for Cancer Classification	0.933	0.921	0.91	0.966	0.708
3. Li et al. [28]	A New Direction of Cancer Classification: Positive Effect of Low-Ranking MicroRNAs	0.9452	-	-	-	-
4. Mukhopadhyay et al. [29]	An SVM-wrapped Multiobjective Evolutionary Feature Selection Approach for Identifying Cancer-MicroRNA Markers	0.8980	-	-	-	-
5. Anidha et al. [30]	An application of fuzzy normalization in miRNA data for novel feature selection in cancer classification	1.0	-	-	1.0	-
6. Pal et al. [31]	Identifying Relevant Group of miRNAs in Cancer using Fuzzy Mutual Information	0.86	0.83	-	-	-
7. Pal et al. [33]	Fuzzy mutual information-based grouping and new fitness function for PSO in selection of miRNAs in cancer	0.86	-	0.93	-	-

1. Lopez-Rincon et al. [22] - Best accuracy reported on
2. Oğul et al.[24] - Best accuracy reported on SVM for SVM based attribute selection and CFS based attribute selection (both at 93.3%), on kNN for SVM based attribute selection at 92.1%, on NB for SVM based attribute selection at 91%, on ANN for SVM based attribute selection at 96.6% and on Decision Tree at 70.8 also with SVM based attribute selection.
3. Li et al. [28] - Best accuracy was reported for SVM classifier used with Information Gain with Low-ranking miRNAs taken into account.
4. Mukhopadhyay et al. [29] - Best accuracy reported on SVM using MOGA with Top 5 miRNA features picked up is 89.80%.

5. Anidha et al.[30] - Best accuracy reported on SVM and ANN for top 50 and top 100 genes picked up is 100%.
6. Pal et al [31] - Best accuracy reported on SVM for Breast cancer dataset at 86% and on kNN for Renal cancer dataset at 83%.
7. Pal et al [33] - Best accuracy reported on SVM at 86% and Naive Bayes at 93% both for Pancreas Cancer Dataset.

Works utilizing self-learning/co-training [34] and Neural Networks [35][36] did not offer accuracy measures for other classifiers. Performance for the former was measured in terms of Precision, Recall and F1-Score instead of Accuracy while the later works offered classifier design for use with miRNA as a biomarker for cancer patients.

7. Conclusion.

Next-Generation Sequencing techniques for biomarkers have enabled high-throughput and massively parallel processing[42] . This has accelerated the field of research for using miRNAs for personalized healthcare by leaps and bounds. MicroRNAs can be utilized to design personalized care for cancer patients. Given their immense importance in the regulation of biochemical processes, microRNAs have enjoyed the attention for research.

However, a large part of the human genome still remains unexplored. That fact is 217 mammalian miRNAs were reported until the early 2000s[25]. The major obstacle in the research of miRNA is the scarcity of proper data. The characteristic dimensions of “small n and large p” (where n represents samples and p represents genes) have also been part of miRNA data. With the advent of Deep Learning, one might attempt to develop techniques which leverage deep neural network architectures which can be used for miRNA biomarker-based predictions - being limited by the amount of data available. Other limitations include unbalanced data and lack of proper labels. While some approaches can accommodate unlabelled data with unsupervised learning, the majority require properly labelled data.

Out of the 20 research papers surveyed, 10 were selected and their results were reported. The papers reviewed for this survey were primarily based on how to use machine learning techniques for selecting a subset of miRNAs to be used as biomarkers. A few papers reported good results with combined datasets of mRNA and miRNA.

References.

1. Bagga S, Bracht J, Hunter S, Massirer K, Holtz J, Eachus R, Pasquinelli AE: Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. Cell 2005, 122:553-563.
2. Du T, Zamore PD: microPrimer: the biogenesis and function of microRNA. Development 2005, 132:4645-4652.
3. Xu P, Guo M, Hay BA: MicroRNAs and the regulation of cell death. Trends in Genetics 2004, 20:617-624.
4. Cheng AM, Byrom MW, Shelton J, Ford LP: Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. Nucl Acids Res 2005, 33:1290-1297.
5. Calin GA, Sevignani C, Dumitru CD, Hyslop T, Noch E, Yendamuri S, Shimizu M, Rattan S, Bullrich F, Negrini M, Croce CM: Human microRNA genes are frequently located at fragile sites and genomic regions

involved in cancers. Proceedings of the National Academy of Sciences of the United States of America 2004, 101:2999-3004.

6. Bartel,D.P. and Chen,C.Z. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. Nat. Rev. Genet., 5, 396–400.
7. Wach S, Nolte E, Theil A, Stohr C, Rau TT, Hartmann A, Ekici A, Keck B, Taubert H, Wullich B. MicroRNA profiles classify papillary renal cell carcinoma subtypes. Brit J Cancer 2013; 109: 714-722.
8. M. Ouzounova et al., "MicroRNA mir-30 family regulates non-attachment growth of breast cancer cells," BMC Genomics, vol. 14, p. 139, 2013.
9. A. H. Lund, "mir-10 in development and cancer," Cell Death and Differentiation, vol. 17, pp. 209–214, 2010.
10. H. Y. Chen et al., "miR-103/107 promote metastasis of colorectal cancer by targeting the metastasis suppressors DAPK and KLF4," Journal of Cancer Research, vol. 72, pp. 3631-3641, 2012.
11. S. S. Ray, J. K. Pal, and S. K. Pal, "Computational approaches for identifying cancer miRNA expressions," Gene Expression, vol. 15, pp. 243–253, 2013.
12. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics. 2009;26(3):392–398.
13. Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer; 2008. p. 313–325.
14. Seijo-Pardo B, Porto-Diaz I, Bolon-Canedo V, Alonso-Betanzos A. Ensemble feature selection: Homogeneous and heterogeneous approaches. Knowledge-Based Systems. 2017;118:124–139.
15. Zhao Z and Liu H. Spectral feature selection for supervised and unsupervised learning. In Proceedings of the 24th international conference on Machine learning (ICML '07). 2007. Association for Computing Machinery, New York, NY, USA, 1151–1157.
16. Wang C, Hu L, Guo M, Liu X, Zou Q. imDC: an ensemble learning method for imbalanced classification with miRNA data. Genetics and Molecular Research. 2015;14(1):123–133.
17. J. K. Pal, S. S. Ray, S. Cho and S. K. Pal, "Fuzzy-Rough Entropy Measure and Histogram Based Patient Selection for miRNA Ranking in Cancer," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 15, no. 2, pp. 659-672, 1 March-April 2018, doi: 10.1109/TCBB.2016.2623605.
18. S. S. Ray, A. Ganivada, and S. K. Pal, "A granular self-organizing map for clustering and gene selection in microarray data," IEEE Trans. on Neural Networks and Learning Systems, vol. 27, pp. 1890–1906, Sep 2016.
19. A. Sharma, S. Imoto, and S. Miyano, "A top-r feature selection algorithm for microarray gene expression data," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 9, pp. 754–764, Nov. 2012.
20. M. Sehhati et al., "Stable gene signature selection for prediction of breast cancer recurrence using joint mutual information," IEEE/ACM Trans. on Computational Biology and Bioinformatics,vol. 12, pp. 1440–1447, Nov./Dec. 2015.
21. L. Yu, Y. Han, and M. E. Berens, "Stable gene selection from microarray data via sample weighting," IEEE/ACM Trans. on Computational Biology and Bioinformatics, vol. 9, pp. 262–272, Jan./Feb. 2012.
22. Lopez-Rincon, A., Martinez-Archundia, M., Martinez-Ruiz, G.U. et al. "Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection." BMC Bioinformatics 20, 480 (2019). https://doi.org/10.1186/s12859-019-3050-8.
23. Calore F, Lovat F, Garofalo M. "Non-coding RNAs and cancer." International journal of molecular sciences. 2013;14(8):17085–17110.
24. Oğul, H. & Altindag, O. (2013). Integrating MicroRNA and mRNA expression data for cancer classification. ICPRAM 2013 - Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods. 503-507.

25. Lu, J., Getz, G., Miska, E., Alvarez-Saavedra, E., Lamb, J., Peck, D., et al., 2005. MicroRNA expression profiles classify human cancers. Nature, 435, 83-838.
26. Peng, S., Zeng, X., Li. X., Peng, X., Chen, L., 2009. Multi-class cancer classification through gene expression profiles: microRNA versus mRNA. J. Genet. Genomics, 36, 409-416.
27. Ramaswamy, S., Tamayo, P., Rifkin, R., et al. 2001. Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci., 98, 15149-15154.
28. Li, F., Piao, M., Piao, Y., Li, M., & Ryu, K. H. (2014). A New Direction of Cancer Classification: Positive Effect of Low-Ranking MicroRNAs. Osong public health and research perspectives, 5(5), 279–285. https://doi.org/10.1016/j.phrp.2014.08.004
29. Mukhopadhyay A. and Maulik U., "An SVM-Wrapped Multiobjective Evolutionary Feature Selection Approach for Identifying Cancer-MicroRNA Markers," in IEEE Transactions on NanoBioscience, vol. 12, no. 4, pp. 275-281, Dec. 2013, doi: 10.1109/TNB.2013.2279131.
30. Anidha, M. & Premalatha, Kandasamy. (2017). An application of fuzzy normalization in miRNA data for novel feature selection in cancer classification. Biomedical Research (India). 28. 4187-4195.
31. Pal JK, Ray SS, Pal SK. Identifying relevant group of miRNAs in cancer using fuzzy mutual information. Med Biol Eng Comput. 2016;54(4):701-710. doi:10.1007/s11517-015-1360-1
32. Pal JK, Ray SS, Cho S. and Pal SK, "Fuzzy-Rough Entropy Measure and Histogram Based Patient Selection for miRNA Ranking in Cancer," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 15, no. 2, pp. 659-672, 1 March-April 2018, doi: 10.1109/TCBB.2016.2623605.
33. Pal JK, Ray SS, Pal SK. Fuzzy mutual information based grouping and new fitness function for PSO in selection of miRNAs in cancer. Comput Biol Med. 2017;89:540-548. doi:10.1016/j.compbiomed.2017.08.013.
34. Ibrahim, Rania & Yousri, Noha & Ismail, Mohamed & El-Makky, Nagwa. (2014). miRNA and Gene Expression based Cancer Classification using Self- Learning and Co-Training Approaches. Proceedings - 2013 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2013. 10.1109/BIBM.2013.6732544.
35. Rincon, Alejandro & Tonda, Alberto & Elati, Mohamed & Schwander, Olivier & Piwowarski, Benjamin & Gallinari, Patrick. (2018). Evolutionary Optimization of Convolutional Neural Networks for Cancer miRNA Biomarkers Classification. Applied Soft Computing. 65. 10.1016/j.asoc.2017.12.036.
36. Zhu, W., & Kan, X. (2014). Neural network cascade optimizes microRNA biomarker selection for nasopharyngeal cancer prognosis. PloS one, 9(10), e110537. https://doi.org/10.1371/journal.pone.0110537
37. Handl, J. & Kell, Douglas & Knowles, Joshua. (2007). Multiobjective optimization in bioinformatics and computational biology. IEEE/ACM Trans Comput Biol Bioinform (TCBB) 4:279-292. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM. 4. 279-92. 10.1109/TCBB.2007.070203.
38. K. Deb, A. Pratap, S. Agarwal and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," in *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182-197, April 2002, doi: 10.1109/4235.996017
39. R. Tibshirani, “Regression shrinkage and selection via the lasso,” Journal of the Royal Statistical Society, Series B, vol. 58, pp. 267–288, 1994.
40. P. Bradley and O. L. Mangasarian, “Feature selection via concave minimization and support vector machines,” in Machine Learning Proceedings of the Fifteenth International Conference (ICML 98. Morgan Kaufmann, 1998, pp. 82–90.
41. J. Fan and R. Li, “Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties,” Journal of the American Statistical Association, vol. 96, no. 456, pp. 1348–1360, 2001.
42. Kim RY, Xu H, Myllykangas S, Ji H. Genetic-based biomarkers and next-generation sequencing: the future of personalized care in colorectal cancer. *Per Med*. 2011;8(3):331-345. doi:10.2217/pme.11.16

43. Ferracin M, Veronese A, Negrini M. Micromarkers: miRNAs in cancer diagnosis and prognosis. Expert Rev Mol Diagn. 2010;10(3):297-308. doi:10.1586/erm.10.11.
44. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS (2004) Human MicroRNA Targets. PLoS Biol 2(11): e363. https://doi.org/10.1371/journal.pbio.0020363.
45. Barrett T1, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. (2013) NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. 41:D991–D995.