World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:14, No:2, 2020

A Hybrid Feature Selection and Deep Learning Algorithm for Cancer Disease Classification

Niousha Bagheri Khulenjani, Mohammad Saniee Abadeh

Open Science Index, Computer and Information Engineering Vol:14, No:2, 2020 waset.org/Publication/10011084

***Abstract*—**Learning from very big datasets is a significant problem for most present data mining and machine learning algorithms. MicroRNA (miRNA) is one of the important big genomic and non-coding datasets presenting the genome sequences. In this paper, a hybrid method for the classification of the miRNA data is proposed. Due to the variety of cancers and high number of genes, analyzing the miRNA dataset has been a challenging problem for researchers. The number of features corresponding to the number of samples is high and the data suffer from being imbalanced. The feature selection method has been used to select features having more ability to distinguish classes and eliminating obscures features. Afterward, a Convolutional Neural Network (CNN) classifier for classification of cancer types is utilized, which employs a Genetic Algorithm to highlight optimized hyper-parameters of CNN. In order to make the process of classification by CNN faster, Graphics Processing Unit (GPU) is recommended for calculating the mathematic equation in a parallel way**.** The proposed method is tested on a real-world dataset with 8,129 patients, 29 different types of tumors, and 1,046 miRNA biomarkers, taken from The Cancer Genome Atlas (TCGA) database.

***Keywords*—**Cancer classification, feature selection, deep learning, genetic algorithm.

I. Introduction

MiRNAs are small non-coding RNA molecules with an approximately 22 nucleotides in length. As a one of the important regulators of cell division, they play important roles in diverse biological processes, including cell proliferation, differentiation, and apoptosis. Several studies have demonstrated that over expression of miRNAs are related to many important diseases such as cancer and also show responsible for cancer development [1], [2].

The earliest discovered evidences show miRNA involvement in human diseases and cancers [1]. Various conducted studies have proved this hypothesis that miRNA's expression is deregulated in human cancers through different manners [2]. miRNA could be used in order to diagnosing and predicting cancers instead of impractical current methods. miRNA biomarkers could be recognized directly from biological specifications, including; blood, urine, etc. [3]. Another advantage of using miRNA as a manner of diagnosis is its ability to predict and find probability of cancers in individuals, which could be useful for survival.

There are only few databases that could be used as miRNA expression data for cancer research to extract features, known as cancer biomarkers. Therefore, these databases would be of utmost practical importance [4], [5]. In other words, it has been mentioned that each miRNA is seen in some distinct cancers that can be recognized as a clue for cancer diagnosis and classification. According to this idea, plenty of researchers have been trying to track miRNA's clues and its differences in cancers.

This study aims to introduce a framework in order to diagnosis cancers by using deep learning approach. One of the main purposes of the proposed approach is to reduce dimension of miRNA datasets and classify it more accurately. The proposed method is introduced in three parts. At the first, the dataset is normalized and then a feature selection algorithm is used for dimensionality reduction. Afterward, a CNN that its parameters are optimized by the genetic algorithm is applied for cancer classification.

II. Dataset

The considered dataset containing miRNA sequence isoforms was obtained from TCGA. This dataset includes 29 types of cancer and 1046 miRNA expressions for each patient whose characteristics can be found in Table I.

Non-coding molecules with about 22 nucleotides heights are known as miRNAs, which act as post-transcriptional regulators primarily (Fig. 1) [6]. The biogenesis of miRNAs begins in the nucleus with the transcription by RNA pol II or pol III synthesizing a double stranded RNA. To form pre-miRNA, a pri-miRNA is processed by the Drosha, and then the pre-miRNA is sent to the cytoplasm. Dicer processes pre-miRNA to a double-stranded RNA form. One of which, called the guide strand, is the base of miRNAs. The other one is degraded. Stability of miRNAs in different fluids is the reason behind finding them as appropriate makers of disease. One of the functions of the miRNAs is promoting degradation; this could provide a comprehension of the mechanisms which could be used in order to diagnoses diseases.

III. Methods

In this paper, a methodology for the classification of miRNA biomarker dataset is purposed. Because of various types of cancers and high number of features due to the number of samples, the data are very challenging. In this way, it is tried to reduce the complexity of the data and to classify it more accurately. Therefore, a method with three parts is proposed in this research. At the first part, it is tried to preprocess the data for normalization. In the second part, a

Niousha Bagheri Khulenjani is with the Faculty of Electrical & Computer Engineering, Tarbiat Modares University, Tehran, Iran (e-mail: newsha_ bagheri@modares.ac.ir).

Mohammad Saniee Abadeh is with the Faculty of Electrical & Computer Engineering, Tarbiat Modares University, Tehran, Iran (corresponding author, e-mail: saniee@modares.ac.ir).

feature selection method is used to reduce the dimensions and select significant features. At the last part, an evolutionary algorithm is utilized to classify the dataset.

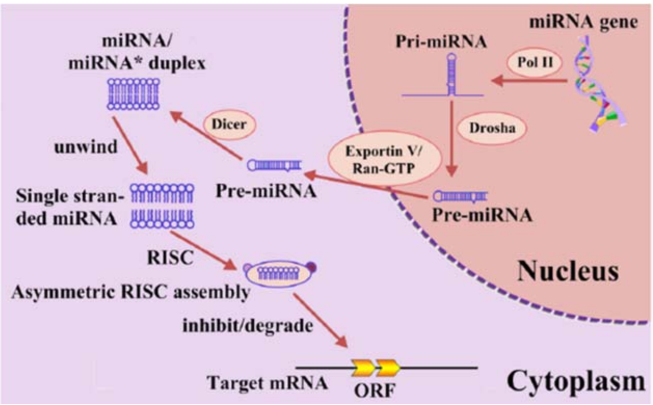

Fig. 1 The role of miRNAs [6]

TABLE I
CANCER MIRNA DATASET INCLUDING 29 TYPES OF TUMORS

Cancers	Classes	Number of samples
Bladder Urothelial Carcinoma	BLCA	415
Kidney renal clear cell carcinoma	KIRC	261
Skin Cutaneous Melanoma	SKCM	452
Uterine Corpus Endometrial Carcinoma	UCEC	418
Uveal Melanoma	UVM	80
Adrenocortical carcinoma	ACC	80
Breast invasive carcinoma	BRCA	778
Cholangiocarcinoma	CHOL	35
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	47
Esophageal carcinoma	ESCA	200
Liver hepatocellular carcinoma	LIHC	374
Pheochromocytoma and Paraganglioma	PCPG	184
Head and Neck squamous cell carcinoma	HNSC	488
Kidney renal papillary cell carcinoma	KIRP	292
Lower Grade Glioma	LGG	530
Lung squamous cell carcinoma	LUSC	341
Mesothelioma	MESO	87
Testicular Germ Cell Tumors	TGCT	155
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	308
Kidney Chromophobe	KICH	66
Lung adenocarcinoma	LUAD	458
Pancreatic adenocarcinoma	PAAD	179
Prostate adenocarcinoma	PRAD	500
Thymoma	THYM	124
FFPE Pilot Phase II	FPPP	45
Sarcoma	SARC	262
Uterine Carcinosarcoma	UCS	57
Stomach adenocarcinoma	STAD	399
Thyroid carcinoma	THCA	514

A. Z-Score Normalization

This part aims to prepare data for changing the values of numeric columns in the dataset to a common scale. Therefore, for normalizing the features range, Z-Score [7] is applied. The normalized value of attributes for z-score normalization is computed as:

$$Z = \frac{v-\mu_i}{\sigma_i} \qquad (1)$$

In (1), μ, σ and v stand for the mean, the standard deviation and real value of each attributes, respectively, which after normalization for all feature values become 0 and 1.

B. Feature Selection

After normalizing data in the previous part, the normalized data is applied to feature selection algorithm. Due to the number of miRNA's features data compared to the number of samples in the cancer classes, a feature selection method is used to reduce dimensions of the dataset and improve the accuracy of classification.

In this study, the Fisher Complexity measure for selecting significant features is applied [8]. The Fisher Complexity method ranks features based on their importance, high speed and performance in terms of classification accuracy. In addition, this method helps the model to concentrate on significant features and eliminate others. Fisher Discriminant Ratio computes correlations between attributes for a multi-class miRNA dataset and can be computed using (2):

$$F1 = max\frac{\sum_{C_j=1}^{C}\sum_{C_k=C_j+1}^{C} P_{C_j} P_{C_k} (\mu_{C_j}^{f_i} - \mu_{C_k}^{f_i})^2}{\sum_{C_j}^{C} P_{C_j} \sigma_{C_j}^2} \qquad (2)$$

where, μ and σ are mean and variance of classes, respectively, and P_{C_j} is the proportion of corresponding C_j. The Fisher Discriminant Ratio evaluates the class separability by each attribute. The greater Fisher Ratio for a feature cues makes classes more separable. In other words, high values for this measure indicates less overlap and complexity of the dataset. In this case, the model selects 300 attributes of the dataset among 1046 attributes with maximum feature ratio.

C. Evolutionary Deep Algorithm

After normalization and dimension reduction, the data are ready for classification. The CNN classifier for classifying of cancer types is utilized, which uses a Genetic Algorithm to highlight the optimized hyper-parameters of CNN [9].

D. CNN

The final step of the proposed method is classifying the samples into 29 cancer classes.

The convolution algorithm includes four main layers of the convolution layer, the ReLU layer, the pooling layer, and the fully connected layer. The convolution, ReLU and fully connected layers are considered as the hidden layer. After feeding input data to the network, a series of convolution operations are performed by the filters in each layer. Then, an activation function *rectified linear unit (*ReLU) is used. After the convolution layer, it is common to use a pooling layer which reduces the number of parameters and computation and also avoids overfitting. After the first part (hidden layer), there is classification part that contains a fully connected layer with Softmax function. In a fully connected layer, the neurons have complete connections to all neurons in the previous layer. At the end, the Softmax function for final classifying is applied.

The Softmax function is calculated from (3), where z is an input vector to the output layer and j represents the output units.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{h=1}^{k} e^{z_h}} \quad (3)$$

The used CNN in this paper has three convolutional layers and a fully connected layer. To train this network, there are three important hyper parameters including; number of filters (w_i), size of convolution's window (wd_i), and size of pooling window (h_i). Therefore, each layer of this algorithm could be described by these parameters. In total, the algorithm has three convolutional layers, that for each one, three parameters are considered. Also, w_4 is utilized to describe a fully connected layer; consequently, there are 10 hyper parameters for this network (w_1, w_2, w_3, wd_1, wd_2, wd_3, h_1, h_2, h_3, and w_4). Dropout layer with probability of 0.7 is considered at the last step of this network.

E. Genetic Algorithm

In the previous part, 10 hyper parameters for CNN classifier are introduced. Finding the optimized values for the hyper parameters is the goal of this section. As all 10 hyper parameters are integers, an optimizing problem could be finding a series of integer numbers. The Genetic Algorithm applied in this approach is shown in Fig. 2. Each series of integer numbers is representative of an individual in Genetic Algorithm or a CNN.

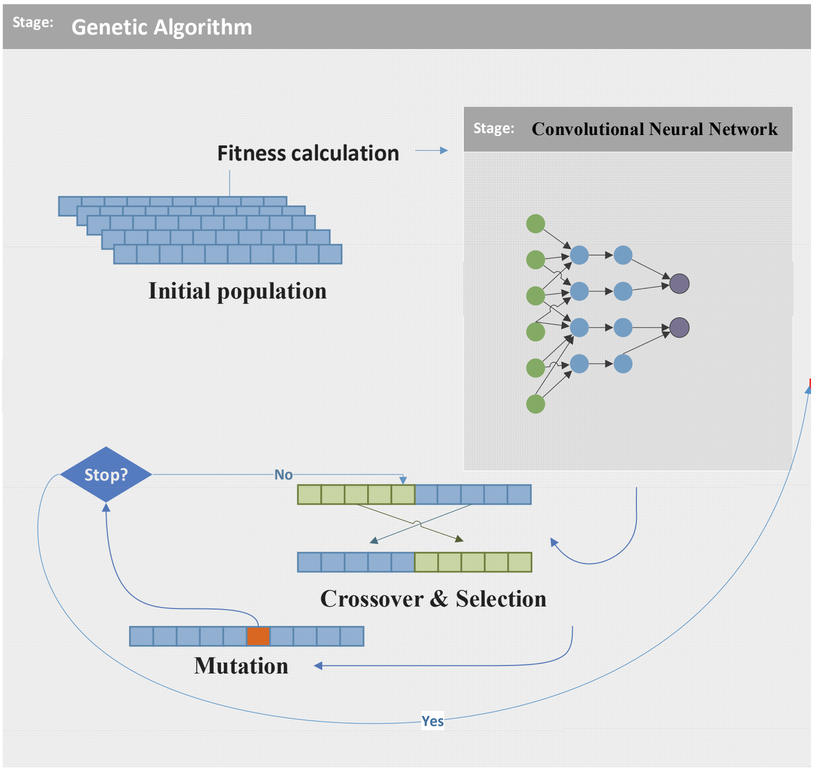

Fig. 2 The process of Genetic Algorithm in the Evolutionary Deep approach

In this process, there are some steps set a value for chromosomes. Due to dependency of CNN's layers, arbitrary series of numbers could not be taken as CNN's parameters. Therefore, the maximum and minimum of each parameter are calculated and the value is randomly chosen from this range. Hence, it is vital to consider this limitation in initial population, crossover, and mutation. In this method, a modifying chromosome function is added to Genetic Algorithm to avoid other possible invalid solutions. At the first, a group of solutions (individuals) are incidentally built and modified. In the purposed method, 50 individuals' chromosomes are considered as the initial population, one of which demonstrates setting of CNN's hyper parameters or 10 integer numbers. Fitness function is equal to the mean of calculated accuracy of CNN.

IV. Result

This section presents the performance of the proposed method. In the proposed method, features that are not involved in predicting labels are omitted. In this way, feature selection helps the model focus on genes with higher ability to distinguish classes. Hence, convolution neural network can classify miRNA data easier and faster.

The proposed algorithm is verified with a dataset containing 29 types of Cancer miRNA Expression. All the obtained results are the mean of 10-fold cross-validation. In 10-fold cross-validation, The dataset is divided into 10 subsets, D_1, …, D_{10}, which are approximately equal in size and disjoint in each other. In loop i of the algorithm, all subsets except D_i are used for training; it is employed for testing the trained classifier. This loop is repeated until all subsets have been used for testing. Finally, the average of the 10 tests is calculated as the result of the classifier. So, the proposed method is evaluated by widely used performance metrics. Also, for a better evaluation of this method, it is compared with state-of-art machine learning methods.

A. Performance Metrics

All the reported results are the mean of 10-fold cross-validation. Here, the widely used criteria are fallowed to evaluate proposed method, including accuracy (Acc), precision (Prec), recall (Rec) and F1 score [10]. They are calculated as:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$Prec = \frac{TP}{TP+FP} \tag{5}$$

$$Rec = \frac{TP}{TP+FN} \tag{6}$$

$$Fsc = 2 * \frac{Prec*Rec}{Prec+Rec} \tag{7}$$

F1-score is defined in terms of precision and recall. The recall is the ratio of relevant points that have been selected by the model to the overall total of the relevant points. Also, the precision is the ratio of the relevant points that have been selected by the model to the total selected points, where TP is the True Positives, FP is the False Positives, TN is the False Negative and FN is the False Negative.

TABLE II
ACCURACY OF CLASSIFIERS

Classifiers	Accuracy
Ada Boost Classifier [12]	0.2889
Bagging Classifier [13]	0.8851
Gradient Boosting Classifier [14]	0.9253
Random Forest Classifier [15]	0.8727
Logistic Regression [16]	0.6766
Logistic Regression CV	0.9405
Passive Aggressive Classifier [17]	0.7072
Ridge Classifier CV [18]	0.7706
SGD Classifier [19]	0.68
Bernoulli NB	0.7749
Gaussian NB	0.7375
Multinomial NB	0.2005
K Neighbors Classifier [20]	0.8521
Nearest Centroid [21]	0.5179
Radius Neighbors Classifier	0.1426
SVC [22]	0.1358
Decision Tree Classifier [23]	0.8148
Extra Tree Classifier [24]	0.5509
Proposed method	0.9465

B. Discussion

All codes and algorithms are implemented in the Python programming language. GTX1080 TI (GPU) graphics card is used to massively exploit the parallelism calculation of CNN.

These evaluation metrics are used in order to compare the performance of the proposed method with other state-of-the-art researches. The proposed method is compared with 18 other methods from in Table II. The implementations are token from the Scikit-Learn Python Packages [11]. These accuracies reported are the average of 10-fold classification. It can be concluded from the results that the proposed method achieved the highest accuracy in comparison with other methods.

Lopez-Rincon et al. purposed the most accurate method on this dataset [9]. TABLE III shows the mean of 10-fold cross-validation test accuracy for each of 29 classes compared with the method recently proposed in [9]. According to the results, our proposed method can achieve more efficient results for 13 cancer classes compared to Lopez-Rincon [9] method.

The value of the precision and F1 score for miRNA dataset are 0.949 and 0.9477, respectively. On the other hand, [9] reports that execution of their method on this dataset takes 14 days using 744 core system. While, in our proposed method, by using an ordinate system having GTX1080 TI, this process takes only seven days. The reasons behind this matter are implementing feature selection in order to decrease the dataset's dimensions from 1046 to 300 features, and in addition, utilizing GPU for CNN calculations in parallel.

TABLE III
ACCURACY EACH CANCER CLASS

Classes	Accuracy of [9]	Accuracy of proposed Method
ACC	0.8889	**0.9375**
BLCA	1	0.978
BRCA	0.9518	**0.98**
CESC	0.875	0.7757
CHOL	0.6	**1**
DLBC	0.8	**0.9667**
ESCA	0.92	**1**
FPPP	0.8333	**0.953**
HNSC	0.9836	0.919
KICH	1	0.9833
KIRC	0.9655	**0.9808**
KIRP	1	**1**
LGG	1	0.9698
LIHC	1	**1**
LUAD	1	0.8667
LUSC	1	0.9412
MESO	0.7272	**1**
PAAD	1	0.91
PCPG	1	**1**
PRAD	1	0.8667
SARC	1	**1**
SKCM	1	0.9889
STAD	0.9268	**0.9875**
TGCT	1	0.9677
THCA	0.98	0.9398
THYM	1	0.9233
UCEC	0.9459	0.9112
UCS	1	0.911
UVM	1	0.8912

V. Conclusion

In this paper, a three-part algorithm is introduced to classify miRNA cancer biomarkers.

Firstly, data normalization and then a feature selection method have been used to select important features. The feature selection is a very important process, because in some datasets, certain features may not help predicting labels, and may sometimes lead to overfitting. In other words, this algorithm eliminates obscure features and selects important features that can better describe the data [8], [25].

Next, a CNN classifier for classification of cancer types is utilized, which uses Genetic Algorithm to highlight the optimized hyper-parameters of CNN. Also, in this research to simplify complexity of CNN's calculation, Graphic Process Unit (GPU) is recommended, which calculate CNN's equations in parallel.

In short, this paper presents a deep learning framework to classify cancer miRNA biomarkers accurately. To achieve this goal, Genetic Algorithm is used, which optimized hyper-parameters of CNN. Moreover, to select most informative genes, feature selection algorithm is used to reduce the high dimensionality of data. Results show that the purposed method is capable of classifying various cancers in miRNA dataset more accurately in comparison with 18 state-of-the-art methods.

For the future, it is our goal to test this method on other cancer miRNA databases, which include many more types of cancers and also number of patients. Moreover, it would be also interesting to investigate the miRNA data of healthy individuals along with cancer patients' data to compare and validate the extracted biomarkers efficiency.

References

[1] Calin, G.A., et al., *Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia.* Proceedings of the National Academy of Sciences, 2002. **99**(24): p. 15524-15529.

[2] Peng, Y. and C.M. Croce, *The role of MicroRNAs in human cancer.* Signal transduction and targeted therapy, 2016. **1**: p. 15004.

[3] Sauter, E.R. and N. Patel, *Body fluid micro (mi) RNAs as biomarkers for human cancer.* Journal of Nucleic Acids Investigation, 2011. **2**(1): p. e1-e1.

[4] Bartels, C.L. and G.J. Tsongalis, *MicroRNAs: novel biomarkers for human cancer.* Clinical chemistry, 2009. **55**(4): p. 623-631.

[5] Cortez, M.A., et al., *MicroRNAs in body fluids—the mix of hormones and biomarkers.* Nature reviews Clinical oncology, 2011. **8**(8): p. 467.

[6] Liu, B., et al., *Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy.* Journal of theoretical biology, 2015. **385**: p. 153-159.

[7] Jain, S., S. Shukla, and R. Wadhvani, *Dynamic selection of normalization techniques using data complexity measures.* Expert Systems with Applications, 2018. **106**: p. 252-262.

[8] Sarbazi-Azad, S., M.S. Abadeh, and M.I.N. Abadi. *Feature Selection in Microarray Gene Expression Data Using Fisher Discriminant Ratio.* in *2018 8th International Conference on Computer and Knowledge Engineering (ICCKE)*. 2018. IEEE.

[9] Lopez-Rincon, A., et al., *Evolutionary optimization of convolutional neural networks for cancer miRNA biomarkers classification.* Applied Soft Computing, 2018. **65**: p. 91-100.

[10] Alshammari, T., et al., *Evaluating machine learning techniques for activity classification in smart home environments.* Int. J. Comput. Electr. Autom. Control Inf. Eng, 2018. **12**: p. 48-54.

[11] Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* Journal of machine learning research, 2011. **12**(Oct): p. 2825-2830.

[12] Hastie, T., et al., *Multi-class adaboost.* Statistics and its Interface, 2009. **2**(3): p. 349-360.

[13] Breiman, L., *Pasting small votes for classification in large databases and on-line.* Machine learning, 1999. **36**(1-2): p. 85-103.

[14] Friedman, J.H., *Greedy function approximation: a gradient boosting machine.* Annals of statistics, 2001: p. 1189-1232.

[15] Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

[16] Cox, D.R., *The regression analysis of binary sequences.* Journal of the Royal Statistical Society: Series B (Methodological), 1958. **20**(2): p. 215-232.

[17] Crammer, K., et al., *Online passive-aggressive algorithms.* Journal of Machine Learning Research, 2006. **7**(Mar): p. 551-585.

[18] Tikhonov, A.N. *On the stability of inverse problems.* in *Dokl. Akad. Nauk SSSR*. 1943.

[19] Hearst, M.A., et al., *Support vector machines.* IEEE Intelligent Systems and their applications, 1998. **13**(4): p. 18-28.

[20] Altman, N.S., *An introduction to kernel and nearest-neighbor nonparametric regression.* The American Statistician, 1992. **46**(3): p. 175-185.

[21] Tibshirani, R., et al., *Diagnosis of multiple cancer types by shrunken centroids of gene expression.* Proceedings of the National Academy of Sciences, 2002. **99**(10): p. 6567-6572.

[22] Boser, B.E., I.M. Guyon, and V.N. Vapnik. *A training algorithm for optimal margin classifiers.* in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*. 2003.

[23] Breiman, L., et al., *Classification and regression trees–crc press.* Boca Raton, Florida, 1984.

[24] Geurts, P., D. Ernst, and L. Wehenkel, *Extremely randomized trees.* Machine learning, 2006. **63**(1): p. 3-42.

[25] Potharaju, S.P. and M. Sreedevi, *Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance.* Clinical Epidemiology and Global Health, 2019. **7**(2): p. 171-176.