Stacked Sparse autoencoder for unsupervised features learning in PanCancer miRNA cancer classification

1st Imene Zenbout
IFA department, NTIC faculty, Constantine 2 University
CRBT, CERIST
Constantine , Algeria
imene.zenbout@univ-constantine2.dz

2nd Abdelkrim Bouramoul
IFA department, NTIC faculty, Constantine 2 University
Misc laboratory
Constantine, Algeria
abdelkrim.bouramoul@univ-constantine2.dz

3rd Souham Meshoul
Princess Nourah bint Abderahmen University
Riyadh, Saudi Arabia
sbmeshoul@pnu.edu.sa

***Abstract*—The recent progress in cancer diagnosis is genomic data analysis oriented. miRNA is playing an important role as cancer biomarkers to move with cancer diagnosis and therapy towards personalized medicine with the ultimate goal to augment survival rate and disease prevention. The recent explosion in genomic data generation has motivated the use of miRNA to enhance diagnosis, prognosis and treatment. In this work we have explored the integrated Atlas PanCancer miRNA profiles, using deep features learning based on unsupervised Stacked Sparse AutoEncoder (SSAE). The proposed SSAE model learns features representation from the used data. The consistency of the learned features has been tested using classification of samples according to 31 cancer types. The model performance has been compared to state-of-the-art unsupervised features learning models. The obtained results exhibit the competitiveness and promising performance of our model, where an accuracy rate of about 95% has been achieved.**

***Index Terms*—Deep learning, Bioinformatics, features learning, Sparse autoencoders, miRNA, PanCancer.**

I. Introduction

The recent and tremendous advance in high sequencing technologies [1] have forstred the role of genomic data across all the transcriptomic as a key answer to different biological related questions and precisely in disease genetics. With these new genomic and genetic data availability and transparency, miRNA role moved from noisy particles to a highly engaged genomic instances in gene regulation and post protein function. This has led to a direct involving of miRNA in the occurrence or the suppression of cancer [2].

microRNA (miRNA) are classified as non-coding regulatory genes [3], that can be found in small fragments of non-coding RNA regions (about 21-23 nucleotide) [3], [4]. Since the discovery of miRNA in 1993 by R.C.Lee [5], the generation of miRNA data using high throughput technologies [6], [7] to explore the direct role of miRNA and cancer diagnosis and gene impact become intensive. The particularity of miRNA profiles is their ability to be a direct tool in cancer analysis, therapy and post treatment [8], which represents the main motivation of this work. The miRNA data share the same issue with gene expression data which is the very small sample size with regard to the high profiles dimensionality .i.e there is some profiles that are irrelevant in cancer diagnosis and related decisions compared to the low number of patient samples. Obviously, this lends itself to a dimensionality reduction problem where it is required to extract the miRNA signature representation that can be a relevant predictors in cancer diagnosis.

In this work we propose a deep unsupervised features learning model, based on stacking three sparse autoencoders to learn new features from the initial noisy miRNA profiles inputs. The learned features through the different abstraction levels, have been used to train classifiers to predict the cancer type of a specific sample according to 31 different cancer type. The proposed unsupervised and supervised models have been trained on the Atlas PanCancer [9] data set. The particularity of this data set is that it combines different cancer type. This may help us to draw information from the well explored cancer type that have a big number of samples and/or a high correlation between the different miRNA profiles and apply these information to classify, or understand the cancer type with poor exploration rate. The features learning model has been compared to some of the most known unsupervised features learning and dimensionality reduction models, here we used pricipal component analysis PCA and kernel principle component analysis KPCA. The rest of the paper is organized as follows: A literature review in section II. Section III is devoted to a brief introduction to sparse autoencoders. Section IV describes the data set and the preprocessing steps. Our proposal is presented in section V along with the set of experimental results and discussion.

II. miRNA cancer Classification

Recently, the exploration of noncoding regions rule in cancer diagnosis and therapy is attracting a large community of scientists. The miRNA data set analysis using statistical and machine learning become one of the trending problems in bioinformatics [3]. In cancer diagnosis and classification, we cite the work of J.Lu et all [10] where the authors analysed mammalian miRNA using k-nearest neighbors and probabilistic neural network algorithm. Kotlarchyk et al [11], used ensemble methodology to classify different cancer type based on miRNA profiles. A statiscical support vector machine- k-nearest neighbors is proposed by D. Ting-ting et al [12], where they used t-statics to select relevent miRNA feautures and a combination of kNN and SVM as classifiers to distinguish between positive and negative samples in different cancer type data set. For multiclass cancer classification, P.Yongjun [13] used subset-based ensenmble method features selection, by generating multiple miRNA subset based on the correlation among miRNAs and then using classifiers to learn valuable knoweledge from each subset to finally combine the results of each classifier by averaging probabilities.A fuzzy normalization based approach is proposed by M Anidha et al [14], where the authors used relevant information gain and F-score to select the most important features in cancer diagnosis, yet in this work the experiments were for binary classification tasks only. A web advisor consisting of semi-supervised classifiers, with pearson correlation, Kappa statistics and recursive feature elimination for selecting the best miRNA profiles, was conducted by N.Cheerla et al [8] ,to perdict cancer type and treatment recommendation based on the Atlas PanCancer data set. In paper [15], the authors used Beep belief nets and active learning to apply multi-level gene/miRNA feature selection, and to visualize the impact between genes and miRNAs, and select the most discriminating miRNAs profiles, the paper tested the performance of the proposed approach in classifying 3 cancer types. Whereas L.Fu et al [16] used stacked auto-encoders to enhance cancer diagnosis and treatment, by building both miRNA-miRNAs and human disease-disease similarities network and then use stacked autoencoder to extract the best features set from the similarity results in order to employit in predicting cancer type. Convolution net works CNN were also used by A. L.Rincon et al [17], to classify the PanCancer data types, where the authors applied Evolutionary algorithm to optimize the architecture of the CNN model.

III. Sparse Auto Encoders

An autoencoder is a symmetric neural network, which copies the input of the network to its output passing through a bottleneck layer that represents the latent features space(figure1). A sparse autoencoder is an autoencoder with applying a sparsity value $\sigma(h)$ on the training of the encoder part, in addition to the reconstruction loss [18]. This sparsity value will deactivate the low value nodes, which led to the extraction of more relevant features representation.

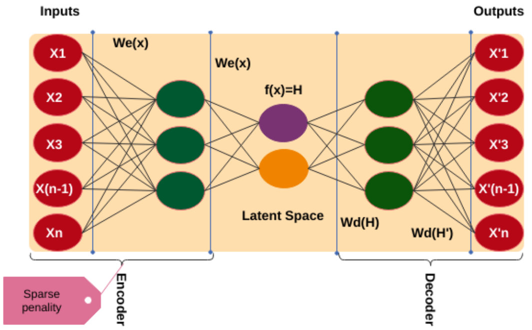


Fig. 1: Sparse Autoencoder Architecture

TABLE I: Data Set description befor/after preprocessing, number of training/ testing sets

	Before	After	Cancer type
Number of patients	10824	10783	31
Number of regions	743	494	31
Training Samples		**Testing Samples**	
7548		3235	

$$L(x, g(h)) + \sigma(h) \tag{1}$$

$g(h)$ is the decoder output and $h = f(x)$ is the encoder output. A detailed description of the autoencoder architecture is in sectionV. Sparse autoencoders have been intensively used for feature learning problems in different domains, emotion detection and robotics [21], medical imaging [20] also and not only medical diagnosis [22].

IV. Data Collection and preprocessing

We collected the *Atlas PanCancer* [9] miRNA Data set used for predicting cancer type from the TCGA data base repository(10/12/2018 18:14). The miRNA data set have been generated using next generation sequencing on around 33 types of cancer in the US hospitals. The initial miRNA data set consist of more than 10 thousand patients and around 800 short non-coding RNAs profiles. We have applied a preprocessing to the data matrix by eliminating the miRNA instances with more than 20% zero values, also we used a log transformation to eliminate the skewed data and finally data imputation to replace the missing values, After we have divided our final data matrix to 70% samples used to train the supervised model, and 30% samples to evaluate the performance of the trained classifier. Table I exhibit the data set description before and after preprocessing and table II illustrate the distribution of samples on the different cancer types.

V. SSAE features learning

We can denote the tackled problem as a matrix X of a dimension $N * M$ where N represents the number of samples and M represents the set of non-coding regions, where each x_{ij} corresponds to a miRNA value i of a sample j. The proposed architecture(Figure2) consist of two phases, a dimensionality reduction phase and a predictive phase. In phase one we have used unsupervised features learning

TABLE II: Distribution of samples among cancer types

	Cancer Type	Number of Samples
1-	BRCA	1164
2-	KIRC	570
3-	THCA	569
4-	HNSC	565
5-	LUAD	555
6-	PRAD	544
7-	UCEC	542
8-	LGG	527
9-	LUSC	511
10-	OV	486
11-	STAD	474
12-	SKCM	452
13-	COAD	429
14-	BLCA	429
15-	LIHC	421
16-	KIRP	321
17-	CESC	311
18-	SARC	260
19-	ESCA	195
20-	LAML	188
21-	PCPG	186
22-	PAAD	182
23-	READ	155
24-	TGCT	138
25-	THYM	126
26-	KICH	89
27-	MESO	87
28-	UVM	80
29-	ACC	79
30-	UCS	56
31-	DLBC	47

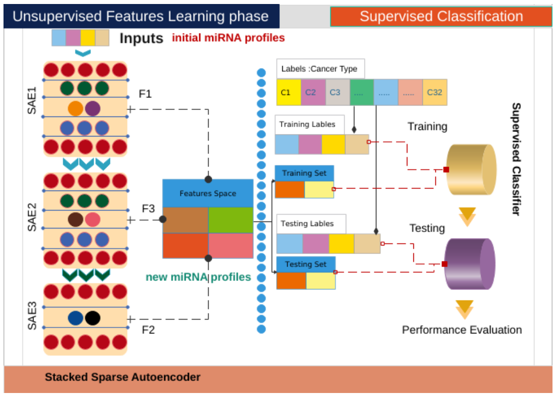


Fig. 2: Stacked Sparse Autoencoder architecture for miRNA based cancer classification

to train a stacked sparse autoencoders (SSAE), where we have piled three sparse autoencoders$[SAE_1, SAE_2, SAE_3]$, in which the input of SAE_i is the output of SAE_{i-1}, where the particularity of the output of autoencoders is that the data is a reconstruction of the input with less noise. The features vectors generated from the three AEs has been concatenated to train a predictive models. These models are trained using supervised learning to predict the cancer type. The two steps in our analytical architecture have been implemented using python 3.5 and *Keras* [23] with tensorflow backened. The experimental results have been processed on HP-bs0xx with Intel Core i7-7500U CPU @ 2.70GHz 4 and 8 GB memory.

A. First Phase

In this step, we have used SSAE to extract a new features representation, that is more accurate in multi-class cancer diagnosis. The first sparse autoencoder SAE_1 takes the features vector S of the matrix X of range M, and fed it to the encoder, in the bottleneck layer a new latent space F_1 of range K, where $K < M$ is generated and based on this latent space the decoder try to reconstruct the input S as close as possible at the output of the decoder where $S \approx S'$. The output S' of SAE_1 become the input of SAE_2 and the same steps are followed to generate a latent space F_1 and the decoder try to reconstruct S' at the output of the decoder S'' where $S' \approx S''$. Equally S'' is the new input of SAE_3 and the bottleneck of the third sparse autoencoder generate the last latent features space vector F_3. The consistency of each autoencoder and their final architecture settings has been evaluated by calculating the reconstruction error loss between the input of the encoder and the output of the decoder for each SAE_i. In our proposal we have used the $mean\_abselout\_error$ loss function(eq2). The three generated features representation$[F_1, F_2, F_3]$ from each sparse auto encoder have been concatenated in one features vector $F4$ to be used to train the classifiers.

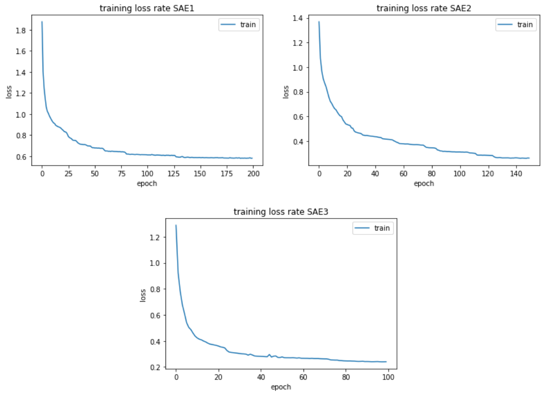
$$mae = 1/n\sum_{i=1}^{n}|x_i - x_i'|/n = \sum_{i=1}^{n}|e_i|/n \tag{2}$$

While zooming on the architecture of each autoencoder in phase one (tableIII), we describe it as follows:

* SAE_1**:** We have used a deep architecture, where the encoder consist of two fully connected layers (494,250 node) with a L2 regularization as sparse penalty, a latent space layer with 50 node that will further generate the new space features F_1, and a symmetric decoder to reconstruct the encoder input with 250, 494 node for each layer respectively.
* SAE_2**:** Equally, we have used a deep autoencoder with two fully connected layers of 494, 150 nodes for each to represent the encoder, and we applied on it a sparse L2 regularization penalty, a 50 nodes bottleneck layer to generate the new F_2 features representation, and a symmetrical decoder.
* SAE_3**:** In the last step we used the simple representation of a sparse autoencoder, since our data have been purified from the biggest amount of noise in the two previous sparse auto encoders, we need to avoid falling into the curse of overfitting and underfitting problem where our autoencoder will only copy the input to the output without learning a new features representation. So our SAE_3 is composed of only one fully connected sparse layer to represent the encoder(494 nodes), a bottleneck layer composed of 50 nodes that represent the last features vector F_3, and a 494 nodes fully connected layer for the mirror decoder.

In order to tune each layer weights of the autoencoders(table III), we have used a Relu nonlinear function. While the bottleneck layer has been tuned using a Softplus activation

Fig. 3: Training performance of the SSAE across each autoencoder



function. We trained the stacked autoencoder using $mini - batch\_gardient\_descent$ training and $Adam$ optimizer as follows:

1- We trained SAE_1 through 200 epochs on a batch size equals to 180 samples from the initial input data set that represents the value of non-coding regions of all the available patients, to obtain a experimental reconstruction loss value of 0.56.

2- SAE_2 have been trained on 150 epochs with a batch size of 150 using the reconstructed input from SAE_1, the experimental reconstruction loss after training was 0.32.

3- The output of SAE_2 have been used to train SAE_3 on 100 epochs with a batch size of 130, the reconstruction loss after training was 0.21.

The figure3, demonstrate the training process of each encoder, where we can see that SAE_1 converged toward the best performance around 150 epochs while SAE_2 was able to stabilize around the epoch 125, whereas SAE_3 converged rapidly to its best performance around the epoch 80. After training the three autoencoders we have extracted the latent space of each autoencoder and concatenate the three vectors as the new miRNA features space to be used in the second phase.

B. Second Phase

The second phase is for classification, where we have used three classifiers to predict the class of a cancer sample according to 31 cancer types. Support vectors machine (SVM), Decision trees(DT), Random Forest(RF), and K-nearest neighbors were the chosen models to be trained to fulfill the diagnosis task. The performance of the model have been assessed through hold-out cross validation where we split our data into 70% training and 30% testing. Besides, to evaluate the performance of our SSAE in learning new features representations, we have compared the performances of the trained classifiers with other classifiers trained on features generated by some of state of the art unsupervised dimensionality reduction methods, namely Principal component analysis (PCA), and Kernel principal component analysis(KPCA). The

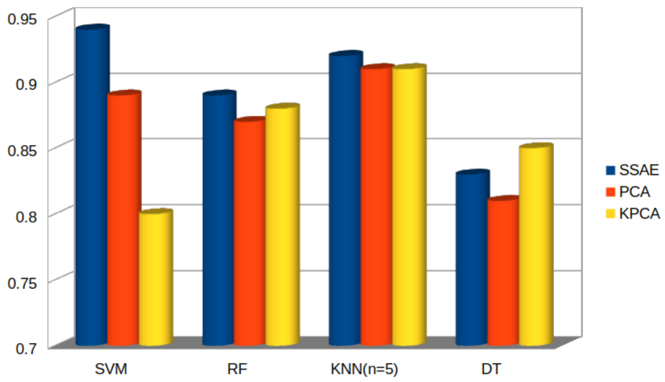


Fig. 4: Accuracy score of the classifiers on SSAE and the other dimensionality reduction methods

overall accuracy score of each classifier (figure4), shows that the predictive models trained on the features representation extracted from SSAE are more powerful to predict the class of each sample. Hence SVM/SSAE scored the highest accuracy in discriminating between the different cancer types with a performance that reaches approximately 95%. While in DT we can see that KPCA was able to overcome our approach with a difference of 0.02. In KNN and RF the performance of the classifiers on each dimensionality reduction approach was so close with a superiority to our approach as an accuracy of 92% and 89% respectively.

Since Accuracy is not enough to evaluate a classifier, also since our problem is a multi-class classification problem we have choose other metrics to evaluate the performance of our models all along the trained classifiers. We have used micro/macro and weighted average values to evaluate the consistency of each classifiers on the prediction of each class, tables[IV,V,VI,VII]. TableIV, represent the case with the best performance in each classifier. We conclude from the results that the SVM/SSAE scored the best performed model, the micro average score reflect the ability of the model in predicting positive samples with a high rate (95%) for both micro average precision and micro average recall. Equally the macro average and the weighted average results are very promising despite the fact that our data are size variant. Tables[V,VII] exhibit the overall performance of the classifiers, where our features representation learning model was able to slightly overcome those trained on PCA and KPCA. In tableVI, were the case our DT/SSAE model was not able to perform better than the DT/KPCA classifier. The collection of results tables exhibit the high consistency of the SSAE features. Where all along most of the classifiers our model was able to score the highest values possible, and in all the experiments we have tested, PCA features were not able to perform better, than ours, yet KNN/PCA was so close to KNN/SSAE with equal micro average and weighted average values, here, only a small difference was captured by the macro average values.

Compared to the results published in [8] and [17], we can say that our model was very powerful in discriminating

TABLE III: Stacked Spares Autoencoders description

	SAE_1					SAE_2					SAE_3		
Architecture	Enc		LS	Dec		Enc		LS	Dec				
	L_1	L_2	L	L'_1	L'_2	L_1	L_2	L	L'_1	L'_2	L_1	L_2	L_3
	494	250	50	250	494	494	150	50	150	494	494	50	494
Epochs	200					150					100		
Batch size	180					150					130		
Activation function	[Relu-Softplus]												
Regularizers	L2(0.001)					L2(0.0001)					L2(0.00001)		
Loss function	mae					mae					mae		
Reconstruction error	0.56					0.23					0.19		

TABLE IV: SVM classifier micro/macro/weighted average score ; P:Precision, R:Recall, $f1-s : f1-score$

	Metric	micro-Av	macro-Av	weighted-Av	Acc
SSAE	P	0.95	0.95	0.95	**0.947**
	R	0.95	0.92	0.95	
	$f1-s$	0.95	0.94	0.95	
PCA	P	0.89	0.94	0.92	0.894
	R	0.89	0.79	0.89	
	$f1-s$	0.89	0.83	0.89	
KPCA	P	0.80	0.63	0.78	0.803
	R	0.80	0.59	0.80	
	$f1-s$	0.80	0.59	.0.77	

TABLE V: RF classifier micro/macro/weighted average score ; P:Precision, R:Recall, $f1-s : f1-score$

	Metric	micro-Av	macro-Av	weighted-Av	Acc
SSAE	P	0.90	0.92	0.90	**0.899**
	R	0.90	0.85	0.90	
	$f1-s$	0.90	0.86	0.89	
PCA	P	0.87	0.90	0.88	0.874
	R	0.87	0.81	0.87	
	$f1-s$	0.87	0.82	0.86	
KPCA	P	0.88	0.89	0.88	0.881
	R	0.88	0.83	0.86	
	$f1-s$	0.88	0.84	.0.87	

TABLE VI: DT classifier micro/macro/weighted average score ; P:Precision, R:Recall, $f1-s : f1-score$

	Metric	micro-Av	macro-Av	weighted-Av	Acc
SSAE	P	0.84	0.86	0.84	0.838
	R	0.84	0.79	0.84	
	$f1-s$	0.84	0.81	0.83	
PCA	P	0.82	0.84	0.82	0.818
	R	0.82	0.77	0.82	
	$f1-s$	0.82	0.79	0.82	
KPCA	P	0.85	0.87	0.85	**0.851**
	R	0.85	0.80	0.85	
	$f1-s$	0.85	0.82	.0.85	

TABLE VII: KNN classifier micro/macro/weighted average score ; P:Precision, R:Recall, $f1-s : f1-score$

	Metric	micro-Av	macro-Av	weighted-Av	Acc
SSAE	P	0.92	0.91	0.92	**0.923**
	R	0.92	0.91	0.92	
	$f1-s$	0.92	0.91	0.92	
PCA	P	0.92	0.92	0.92	0.919
	R	0.92	0.90	0.92	
	$f1-s$	0.92	0.90	0.92	
KPCA	P	0.92	0.90	0.92	0.918
	R	0.92	0.90	0.92	
	$f1-s$	0.92	0.90	.0.92	

between the 31 cancer types, despite the fact that some of the cancer types samples are very low in count. Cheerla et al [8] addressed this problem by eliminating the types that have smaller number of patients, so they worked on only 21 cancer type using semi-supervised learning to augment the accuracy score to 97%. For A.L.Rincon et al [17], the authors also dealt with 29 cancer types to reach a training accuracy 96%. Also we assume that by integrating more characteristics like stage and gender to our analytical strategy we may improve the results of the 31 predicted cancer type.

VI. Conclusion

In this paper we have implemented a stacked sparse unsupervised auto encoder to learn new features representation that may help in promoting cancer genetic diagnosis based on the short non-coding RNA regions, which plays a significant role in silencing, regulating and managing the transcription biological process in human body. The learned features have been evaluated through a supervised models, where our proposed unsupervised features learning model was able to generate a new discriminant data representation leading to a competitive method with regard to the state-of the art methods. We believe that the collection of new samples or moving toward semi-supervised classification or integrating some clinical information may enhance the results obtained in this work, also the use of the PanCancer data set may give to our model the flexibility and the easy use on other cancer types generated from different genomic data banks for further research aspects.

References

[1] F.Cristiano, P.Veltri. ”Methods and techniques for miRNA data analysis”. in Microarray Data Analysis. Humana Press, New York, NY, 2015. pp 11–23.

[2] S.Tam, M.S.Tsao,J.D.Mcpherson.” Optimization sof miRNA-seq data preprocessing”. Briefings in bioinformatics, 2015, pp 950–963.

[3] S.Sing, et al. ”Machine learning techniques in exploring microRNA gene discovery, targets, and functions” in Bioinformatics in MicroRNA Research. Humana Press, New York, NY, 2017. pp 211–224.

[4] P.H.Gunaratne, C.Coarfa , B.Soibam , A.Tandon . ”miRNA Data Analysis: Next-Gen Sequencing”. in Fan JB. (eds) Next-Generation MicroRNA Expression Profiling Technology. Methods in Molecular Biology (Methods and Protocols),2012 Humana Press

[5] R.C.LEE,R.L.Feinbaum, V.Ambros. ”The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. cell”, 1993, pp 843–854.

[6] J.Xuan, Y.Yu , T.Qing a, L.Guo , L.Shi. Next-generation sequencing in the clinic: promises and challenges. Cancer letters, 2013,pp 284–95.

[7] K.R.Kukurba, S.B.Montgomery.”RNA sequencing and analysis”. Cold Spring Harbor Protocols, 2015.

[8] N.Cheerla, O.Gevaert, ”MicroRNA based Pan-Cancer diagnosis and treatment recommendation”. BMC bioinformatics, 2017.

[9] J.Liu, et al. ”An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics”. Cell, 2018, pp 400–416.

[10] J.Lu, et al. ”MicroRNA expression profiles classify human cancers”. nature, 2005.

[11] A. Kotlarchyk,Khoshgoftaar, T., Pavlovic, M., Zhuang, H., A.S Pandya. Identification of microRNA biomarkers for cancer by combining multiple feature selection techniques. Journal of Computational Methods in Sciences and Engineering, 2011. pp 283–298.

[12] D,Ting-ting, S.Chang-ji, D.Yan-shou,B. Yi-duo. ”Analysis of miRNA expression profile based on SVM algorithm”.in IOP Conference Series: Earth and Environmental Science. IOP Publishing, 2018.

[13] P.Yongjun, P.Minghao, R. Keun Ho. ”Multiclass cancer classification using a feature subset-based ensemble from microRNA expression profiles”. Computers in biology and medicine, 2017, pp 39–44.

[14] M.Anidha; K.Premalatha, ”An application of fuzzy normalization in miRNA data for novel feature selection in cancer classification”. Biomed. Res, 2017, 28.9: 4187-4195.

[15] R.Ibrahim, N.A.Yousri, M.A.Ismail, N. M.El-Makky,”Multi-level gene/MiRNA feature selection using deep belief nets and active learning”. in 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2014. pp 3957–3960.

[16] L.Fu, Q.Peng.”A deep ensemble model to predict miRNA-disease association”. Scientific reports, 2017.

[17] A. L.RINCON, et al. ”Evolutionary optimization of convolutional neural networks for cancer miRNA biomarkers classification”. Applied Soft Computing, 2018, pp 91–100.

[18] I.Goodfellow, Y.Bengio, A.Courville. ”Deep learning”. MIT press, 2016.

[19] M.Tschannen, O.Bachem, M.Lucic, ”Recent advances in autoencoder-based representation learning”, arXiv preprint arXiv:1812.05069, 2018.

[20] Y-D.Zhang, et al. ”Seven-layer deep neural network based on sparse autoencoder for voxelwise detection of cerebral microbleed”. Multimedia Tools and Applications, 2018, pp 10521–10538.

[21] L Chen, et al. ”Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction”. Information Sciences, 2018, pp 49–61.

[22] C.Zhang, et al. ”Deep Sparse Autoencoder for Feature Extraction and Diagnosis of Locomotive Adhesion Status”. Journal of Control Science and Engineering,2018.

[23] C.François et al.”Keras”.https://keras.io.2015