

Multi-class Classification of Gene Expression Data Using Deep Learning for Cancer Prediction

Soumiya Hamena and Souham Meshoul

Abstract—Deep Learning is a machine learning model that has shown superior performance for a wide range of application. Deep learning has revived the old domain of artificial neural networks and allowed to renew with connexionism in an unprecedented and strongest way. In fact, Deep Learning has advanced rapidly since the early 2000s in various fields and have provided remarkable results in various machine learning applications, including speech recognition, computer vision, and natural language processing. Accordingly, Deep Learning algorithms are promising avenue of research in the automated extraction of complex data representations at high levels of abstraction. By another way, one of the major challenges in bioinformatics is the construction of accurate classification models based on huge and high dimensional data sets such as gene expression data. Gene expression is evaluated by measuring the number of RNA transcripts in a tissue sample. Cancer classification using gene expression data is applied to solve the problems relating to cancer diagnosis and drug discovery. In this paper, we propose an approach that proposes a deep learning based model to achieve multi-class classification of gene expression data with the aim to predict the type of cancer. The validation of the approach is achieved using Keras platform. Very encouraging results have been obtained.

Index Terms—Gene expression, deep learning, machine learning, neural networks, bioinformatics, Keras.

I. INTRODUCTION

Bioinformatics has become a very important subject of research in recent years, covering many topics ranging from the storage of DNA data and the mathematical modeling of biological sequences, analyzing complex human diseases mechanisms, to the understanding and modeling of the evolutionary history of life [1]. Starting from the sequencing of the first small phage genome (5.386 bases length) and going towards the sequencing of 1.000 human genomes (three billion bases length each) [2]. In the last few years, we assist to the proliferation of data from genomic sequences. Meanwhile of the emergence of high throughput sequencers which have become an indispensable tool in biological research. Henceforth, future research and medical discoveries mainly depend on our capacity to analysis and process the

huge amount of genomics data. With the advent of these data, new bioinformatics challenges have arisen and there is still work to be done to teach biologist swimming in this ocean of sequences to keep them safe [2]. To cope with this evolution, bioinformatics must exploit high-level expertise. Indeed, Deep Learning has become the innovation of the early decade facing such as issue.

In the era of Big Data, transforming massive biomedical data into precious knowledge has been one of the most important research topics in bioinformatics; Deep learning has progressed rapidly since the early 2000s in various fields. As a result, application of deep learning in bioinformatics to extract information useful and valuable from data has been emphasized in both academia and industry [3]. Deep learning that has evolved since the acquisition of big data, the spread of parallel and distributed computing and sophisticated learning algorithms has led to breakthroughs in several fields such as image recognition, speech recognition, and natural language processing [4]. Therefore, bioinformatics will widely exploit and benefit from deep learning.

Also, Gene expression profiling with DNA microarrays has emerged as a powerful approach to study the cancer transcriptome [5]. In fact, the transcriptome is defined as the set of transcripts present in a cell at a given time and under given conditions. It is image of the functional state of the genome. DNA microarrays allow the transcriptome to be studied by simultaneous observation of the expression of several thousand genes in a given cell or tissue, thus measuring the modifications of the various cellular states [6]. Many of published studies have presented analyses of human cancer samples, identifying gene expression signatures for most major cancer types and subtypes, and uncovering gene expression patterns that correlate with various characteristics of tumors including tumor grade or differentiation state, metastatic potential, and patient survival [7]. In fact, many studies have focused on the distinction of two classes of disease, but few approaches have been designed for classification in presence of multiple tumor classes. In fact, new subtypes of cancer continue to be discovered, which has imposed an increasing number of multi-class problems. In addition, clinical experiments to study tumors in phase, grade, survival and drug sensitivity will certainly generate multi-class microarray data sets [8].

In this context, we propose an approach that proposes a deep learning based model to achieve multi-class classification of gene expression data with the aim to predict the type of cancer. For this, we used Keras which is a python library for deep learning that wraps the efficient numerical libraries Theano and Tensorflow, to develop and evaluate deep neural network model for multi-class classification of gene expression data.

Manuscript received March 13, 2018; revised October 4, 2018.

The authors are with Computer Science and Application Department/FNTIC, Constantine, Algeria (e-mail: Soumiya.hamena@univ-constantine2.dz, Souham.meshoul@univ-constantine2.dz).

II. BACKGROUND AND RELATED WORK

A. Data Classification

Classification is one of the most common applications for data mining. Data mining using labeled data is known as supervised learning. If the designated attribute is categorical, the task is called classification. If the designated attribute is numerical, the task is called regression. Also, classification is one form of prediction, where the value to be predicted is a label. A very popular way of doing this, is to use a neural network. This is a complex modeling technique based on a model of human neuron. A neural network is given a set of inputs and is used to predict outputs. The general approach to classification as a two-step process. In the first step, we constructed classification model based on available data. In the second step, we decide if the model's accuracy is acceptable and if so, we use the model to classify new data [9]. Data classification can be achieved in different ways such as: Decision Tree induction, k-Neighbors Classifier, Bayes classification methods, rule based classification, Artificial Neural Networks, classification by Backpropagation, Support Vector Machines.

Decision Tree: the structure of a decision tree consists of a set of nodes whose each internal node corresponds to an attribute. Each branch corresponds to a possible value of this attribute. Each leaf of the tree corresponds to a class and provides a classification. Each path in the tree corresponds to a rule. A rule is generated for each path from the root tree to a leaf. The attribute-value pairs of a path form a conjunction and the terminal node represents the predicted class. Then, the tree is built in descending way by choosing at each level an attribute to partition the learning data. The process of generating the tree is done in two steps which are the recursive descending construction of the tree, then the removal of branches that introduce noise to improve the classification quality in order to reduce the error rate.

K-Neighbors Classifier: this is one of the simplest supervised classification methods that is based on analogy. Therefore, one does not need to construct a model, it is the learning sample associated with a function of distance and a function of choice of the class according to the classes of the nearest neighbors, which constitutes the model. In fact, the main tasks to be followed are the choice of the relevant attributes for the classification task in question and the data. Then, the choice of the distance by attribute and the mode of combination of the distances according to the type of the fields and the preliminary knowledge of the problem. And finally, the choice of the number k of neighbors determined, a heuristic frequently used is to take k equal to the number of attributes plus 1.

Bayes classification methods: the Bayesian classification is a probabilistic classification approach based on the Bayes rule or the Bayesian approach to probabilities. One advantage of this approach is that it naturally allows the integration of a priori knowledge.

Support Vector Machines: the principle of this method resides in the resolution of a problem of nonlinear optimization and more exactly quadratic whose solution is unique because the problem is convex. It is among the most efficient methods of classification since it allows the processing of large data. However, the constructed model is

difficult to interpret.

B. Microarray Data Classification

The appearance of DNA microarray was the beginning of quantitative biology on a large scale. This technique made it possible to simultaneously measure the expression of thousands of genes transcribed in a single experiment. It is question of placing thousands of genes sequences in known locations on glass slide called a gene chip. A sample containing DNA or RNA is placed in contact with the gene chip [10]. Complementary base pairing between the sample and the gene sequences on the chip produces light that is measured. Areas on the chip producing light identify genes that are expressed in the sample. These experiments consist mainly in either analyzing each gene several times under different conditions or monitoring each gene in a single environment in different types of tissue [11]. The problem that arises is to manage the microarray data with its dimension which is large. Classification and manipulation of the algorithms is becoming increasingly difficult and complex to study and analyze the expression of the genes.

In microarray gene classification problem, initially we have a set of genes and the corresponding class label, the objective is to determine the similarity between the genes of the same class, so that when a test gene is given, the corresponding class label is found. We can formally describe the task as follow: given a dataset, the goal is to develop a classifier that maps an object to its class label. The model of the classifier is developed by studying the samples described by a set of features named input features. Each sample is assumed to belong to a predefined class, named the output class label. In this paper, we address multiple-class classification problem for gene expression data with the aim to predict cancer.

C. Deep Neural Networks

The concept of deep learning comes from research on artificial neural network [12]. The basic structure of deep neural networks consists of an input layer, multiple hidden layers and an output layer. The layers are made of nodes. A node is just a place where computation occurs. Also, a node combines input from the data with a set of weight. Once input data are given to the deep neural network, output values are computed sequentially along the layers of the network. At each layer, the input vector which contains the output values of each unit in the lower layer is multiplied by the weight vector for each unit in the current layer to produce the weighted sum. After, the sum is passed by the activation function of a node to compute the output values of the layer [13].

III. THE PROPOSED MODEL

In our work, we propose a multi-class classification using Keras deep learning for cancer prediction. Deep learning methods are a modern update to artificial neural networks that exploit abundant cheap computation. They build much larger and more complex neural networks. The task consists in defining the input layer, hidden layers and output layer. We formulated cancer prediction as a classification problem with multiple classes. The nodes of the neural network are

fully connected where each connection is parameterized by a real value weight. We can formally present our approach as follows: solving gene expression data multi class classification problem of cancer prediction involves developing a model M that helps identifying the type of cancer for each sample. Given S a set of labeled data $S = \{S_1, S_2, \dots, S_n\}$, where S_i refers to a sample which is defining using gene expression features as a D -dimensional vector, where each S_i for $j=1 \dots D$, refers to a features and refers to the class of cancer or the type of cancer $S_i = \{G_1, G_2, \dots, G_d, C_i\}$. Therefore, the model to be developed should be able to predict the type of cancer of new sample.

We used two activation functions for the layers namely (RELU) and (SIGMOID). RELU (Rectified Linear Units) is used by the most recent deep learning network for the hidden layers; it has output 0 if the input is less than 0, and raw output if otherwise. So, if the input is greater than 0, the output is equal to the input. That is, RELU is more like a real neuron in your body [14]:

$$f(x) = \max(x, 0)$$

SIGMOID function has been widely applied in machine learning [15]:

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

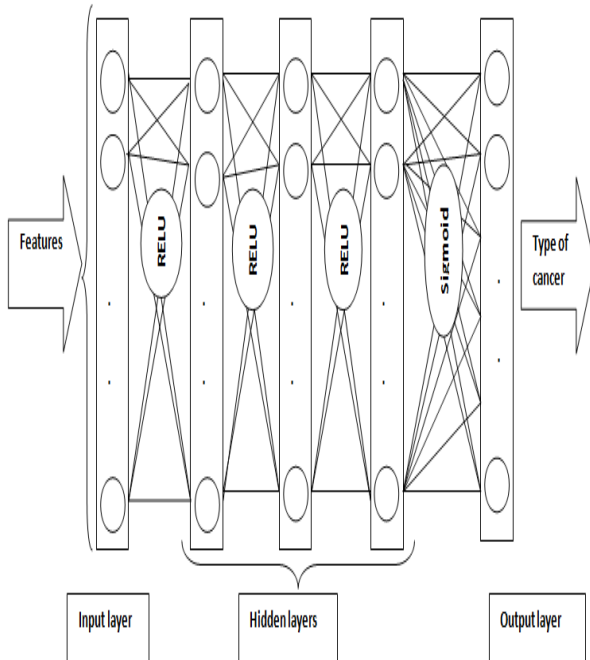


Fig. 1. Architecture of the proposed model.

The first input layers are the sets of features, that is, the number of nodes in input is the number of genes expression features (Table I). The output layer must create one output value for each class. So the number of nodes in the output is the number of classes (Table I). Often, the best network structure is found through an experimental process, in our model we used a fully connected network structure with three hidden layers whose nodes number is initialized with the number of nodes input. We initialized the network weights to a small random number. Also, we used the RELU on the first

two hidden layers and the SIGMOID on the third. Finally, we used the efficient Adam gradient descent optimization algorithm with a logarithmic loss function [16], which is called “categorical_crossentropy” in keras. The architecture of the proposed model is shown on Fig. 1.

IV. EXPERIMENTAL STUDY

A. Description of Used Datasets

The proposed model has been assessed using 7 different gene expression datasets available in [17]. Table 1 gives the details of gene expression datasets used that consist of a matrix of gene expression vectors obtained from DNA microarrays for a number of patients. Tissue indicate the tissue from where samples are taken such as: blood, colon and lung, Total Samples indicate the number of samples, Num of classes indicate the number of classes, No of genes indicates the number of gene expression values, class labels indicates the different type of cancer and the last column indicates distribution of samples within the classes. The model has been trained and validated on Theano using Keras deep learning library in Python with Anaconda.

To assess the performance of the model, we used the following performance measure:

Accuracy: is the number of correct predictions made divided by the total number of predictions made.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FN + FP + TN)}$$

where:

TP : is the number of real data of the positive class that the model can predict correctly.

TN : is the number of real data of the negative class that model can predict correctly.

FP : is the number of real data of the negative class but the model incorrectly predicts.

FN : is the number of real data of the positive class but the model incorrectly predicts.

Loss: a loss value implies how well or bad a model is behaving after each iteration. Also, one would expect the reduction of loss after each iteration.

$$\text{Loss} = -\frac{1}{N} \sum_{i=0}^N \sum_{j=0}^J y_j * \log(\hat{y}_j) + (1 - y_j) * \log(1 - \hat{y}_j)$$

where:

\hat{y}_j : is the predicted value for the j -th label of a given sample.

y_j : is the corresponding true value.

N : is the number of classes or labels.

B. Experimental Results

For each dataset, we executed the proposed model by calculating the accuracy and the loss as the performance measurement value. We also perform some methods of multiple classification using Python for the same dataset in order to achieve a comparative study, the methods are:

Decision Tree Classifier (CART), k-Neighbors Classifier (KNN), Gaussian Naïve Bayes (NB) and Support Vector Machines (SVM). We have divided the datasets into two, 80% of which we have used to train our model and 20% that we have test the model. The results obtained are shown in Table II.

TABLE I: USED DATASETS

Datasets	Tissue	Num of classes	Total Samples	No. of Genes	Class labels	Class wise samples
Tomlins-2006-v1	Prostate	5	104	2315	EPI	27
					MET	20
					PCA	32
					PIN	13
					STROMA	12
Liang-2005	Brain	3	37	1411	GBM	28
					ODG	6
					NORMAL	3
Khan-2001	Multi-tissues	4	83	1069	EWS	29
					BL	11
					NB	18
					RMS	25
					PT1	11
Lapoint-2004-v2	Prostate	4	110	2496	PT2	39
					PT3	19
					NORMAL	41
					PS	13
					CC	3
Risinger-2003	Endometrium	4	42	1771	E	19
					N	7
					EPI	27
					MET	20
					PCA	32
Tomlin-2006-v2	Prostate	4	92	1288	PIN	13
					DLBCL	42
					FL	9
					CLL	11

TABLE II: EXPERIMENTAL RESULTS

Datasets	Proposed model		CART		KNN		NB		SVM	
	acc	loss	acc	loss	acc	loss	acc	loss	acc	loss
Tomlins-2006-v1	0.92	1.15e-04	0.46	0.15	0.66	0.19	0.73	0.17	0.40	0.21
Liang-2005	1.00	2.59e-05	0.73	0.31	0.70	0.20	0.70	0.39	0.71	0.39
Khan-2001	1.00	1.9e-04	0.81	0.11	0.91	0.11	0.91	0.15	0.98	0.05
Lapoint-2004-v2	0.62	0.34	0.59	0.17	0.64	0.18	0.67	0.14	0.64	0.17
Risinger-2003	0.45	1.34e-01	0.43	0.32	0.51	0.20	0.73	0.23	0.36	0.35
Tomlin-2006-v2	0.73	1.08e-01	0.45	0.23	0.63	0.17	0.73	0.15	0.36	0.20
Alizadeh-2000-v2	1.00	1.19e-07	0.87	0.16	1.00	0.00	0.92	0.11	0.97	0.07

C. Performance Measures

Furthermore, we plotted the area under the ROC (Receiver Operating Characteristic) curves and the precision-recall curves for Tomlins-2006-v1, Khan-2001, Lapoint-2004-v2 and Tomlin-2006-v2 Datasets, respectively. A ROC plot presents Sensitivity (the true positive rate) on the Y axis and Specificity (the false positive rate) on X axis. A precision-recall curve presents precision on the Y axis and recall on the X axis. where:

$$true\ positive\ rate = \frac{TP}{TP + FN}$$

$$false\ positive\ rate = \frac{FP}{FP + TN}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1\ score = 2 * \frac{precision * recall}{precision + recall}$$

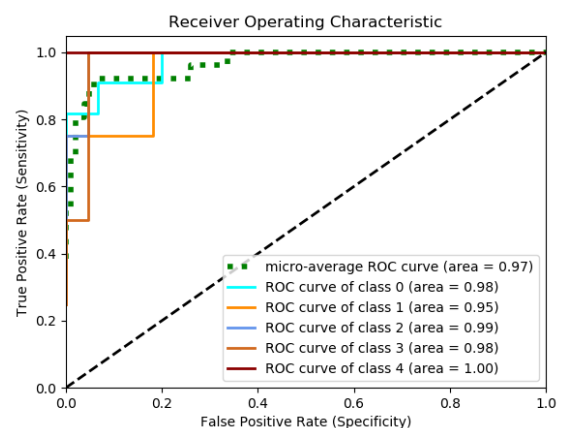


Fig. 2. Roc curve of Tomlins-2006-v1.

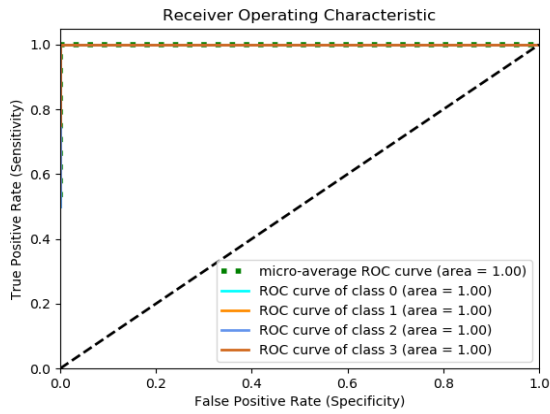


Fig. 3. Roc curve of Khan-2001.

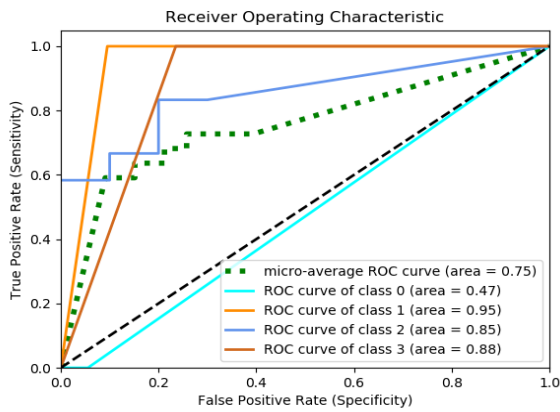


Fig. 4. Roc curve of Lapoint-2004-v2.

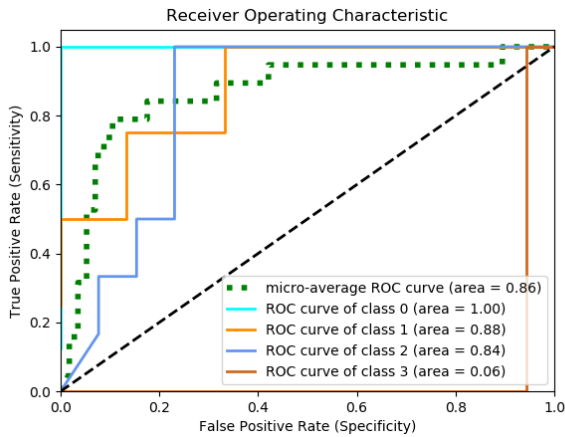


Fig. 5. Roc curve of Tomlin-2006-v2.

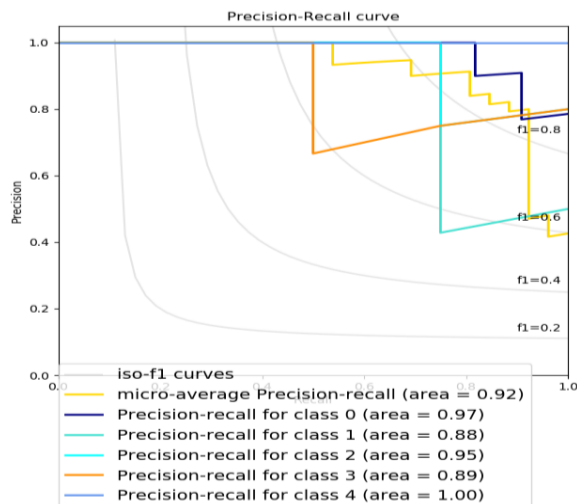


Fig. 6. Precision-recall curve and iso-f1 curves of Tomlins-2006-v1.

Fig. 2-Fig. 5 show plot of ROC curves (AUC) of Tomlins-2006-v1, Khan-2001, Lapoint-2004-v2 and Tomlin-2006-v2 Datasets, respectively. We have obtained a micro-average of 0.97, 1.00, 0.75 and 0.86, respectively.

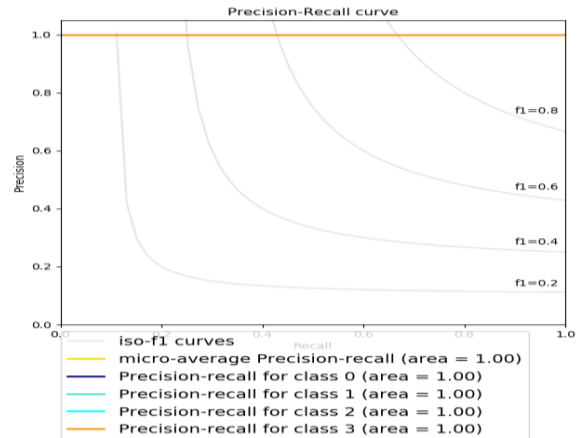


Fig. 7. Precision-recall curve and iso-f1 curves of Khan-2001.

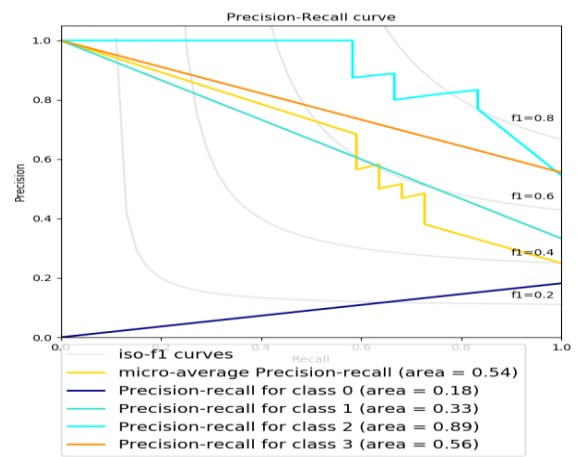


Fig. 8. Precision-recall curve and iso-f1 curves of Lapoint-2004-v2.

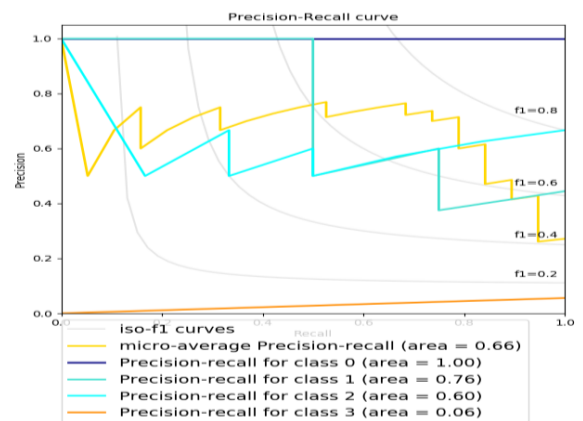


Fig. 9. Precision-recall curve and iso-f1 curves of Tomlin-2006-v2.

Fig. 6-Fig. 9 illustrate plot of Precision-recall curves and iso-f1 curves of Tomlins-2006-v1, Khan-2001, Lapoint-2004-v2 and Tomlin-2006-v2 Datasets, respectively; from which we obtained a micro-average of 0.92, 1.00, 0.54 and 0.66, respectively.

V. CONCLUSION

In this paper, we proposed a multi-class classification of gene expression data using deep learning to predict the type

of cancer. It is important to emphasize that the type of treatment of cancer will depend on the type of cancer and how advanced it is. However, the integration of gene expression data with other genomic data such as: transcriptomic and proteomic could improve upon the classification model. Training and validating of the proposed model using several biological datasets have demonstrated their effectiveness.

REFERENCES

- [1] R. Jiang *et al.*, "Basics of Bioinformatics: Lecture notes of the graduate summer school on bioinformatics of China," Beijing and Springer-Verlag Berlin Heidelberg, 2013.
- [2] N. Rodriguez-Ezpeleta, *Bioinformatics for High Throughput Sequencing*, New York: Springer, 2012.
- [3] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in Bioinformatics*, 2017, vol. 18, no. 5, pp. 851-869.
- [4] Y. Cun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436-444, 2015.
- [5] A. L. Tarca, R. Romero, and S. Draghici, "Analysis of microarray experiments of gene expression profiling," 2006, vol. 195, issue 2, pp. 373-388.
- [6] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of Tumors using gene expression data," *Journal of the American Statistical Association*, 2002, vol. 97, no. 457.
- [7] P. Danaee, R. Ghaeini, and D. A. Hendrix, "A deep learning approach for cancer detection and relevant gene identification," *Pacific Symposium on Biocomputing*, 2017, vol. 22, pp. 219-229.
- [8] S. Yang and D. Q. Naiman, "Multiclass cancer classification based gene expression comparison," *Stat Appl Genet Mol Biol*, 2014, vol. 13, no. 4, pp. 477-496.
- [9] M. Bramer, *Principles of Data Mining*, United Kingdom: Springer, 2016.
- [10] R. K. Singh and M. SivabaLakrishnan, "Feature selection of gene expression data for cancer classification: A review," *Procedia Computer Science*, vol. 50, 2015, pp. 52-57.
- [11] D. A. V. Ca, D. Devaraj, and M. Venkatesulu, "Gene expression data classification using support vector machine and mutual information

- based gene selection," *Procedia Computer Science*, vol. 47, 2015, pp. 13-21.
- [12] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, 2015, pp. 85-117.
- [13] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in Bioinformatics*, vol. 18, issue 5, 2017, pp. 851-869.
- [14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097-1105.
- [15] G. Cybenko, "Approximation by superpositions of sigmoidal function," *Mathematics of Control, Signal and Systems*, 1989, vol. 2, pp. 303-314.
- [16] K. Diederik and P. BaJimmy, "ADAM: A method for stochastic optimization," presented at 3rd International Conference for Learning Representations, San Diego, 2015.
- [17] M. C. P. Souto *et al.*, "Clustering cancer gene expression data: A comparative study," *BMC Bioinformatics*, vol. 9, p. 497, 2008.



Soumiya Hamena received the Dipl. master degree from Mentouri Constantine University (UMC). She is also currently pursuing the PhD at Constantine2 Abdelhamid Mehri University. Now her research areas are data mining and machine learning with applications to bioinformatics.



Souham Meshoul received the State Engineer degree, the MS degree and state doctorate degree from Mentouri Constantine University (UMC). Currently she is a full professor at Constantine2 Abdelhamid Mehri University where she also serves as head of the NTIC college scientific council. She has been involved in many research projects in Algeria and abroad in Europe and Kingdom of Saudi Arabia. Her research interests span several areas namely computational intelligence, optimization, data mining, big data analytics, machine learning with applications to bioinformatics, image analysis and biometrics.