Advanced Machine Learning Models for Large Scale Gene Expression Analysis in Cancer Classification: Deep Learning Versus Classical Models

Imene Zenbout[(⊠)] and Souham Meshoul

Computer Science Department, Faculty of NTIC,
University Constantine 2 - Abdelhamid Mehri Biotechnology Research Center
(CRBt) & CERIST, Constantine, Algeria
{imene.zenbout,souham.meshoul}@univ-constantine2.dz

Abstract. Analysis of large gene expression datasets for cancer classification is a crucial task in bioinformatics and a very challenging one as well. In this paper, we explore the potential of using advanced models in machine learning namely those based on deep learning to handle such task. For this purpose we propose a deep feed forward neural network architecture. In addition, we also investigate other classical yet very popular machine learning classifiers namely, support vector machine, naive bayes, k-nearest neighbours and shallow neural networks. The main objective is to appreciate the extent to which they are able to deal with the increasing size of these datasets. We conducted our experimental study using a high-performance computing platform with 32 compute nodes, each consisting of two Intel (R) Xeon (R) CPU E5-2650 2.00 GHz processors. Each processor is made up of 8 cores. Five data sets available at the omnibus library have been used to test the five models . Experimental results show the effectiveness of deep learning and its ability to deal with large scale data.

Keywords: Gene expression · Machine learning · Deep learning
Neural network · Classification · Cancer classification · Big data

1 Introduction

In the last decades, the remarkable advances in microarrays technology opened huge opportunities in genomic research and especially in cancer researches to move from clinical decisions and standard medicine toward personalized medicine. The analysis of gene expression level may reveal a lot of informations about the cancer type, its outcomes also allow the possibility to predict about the best therapy in order to improve the survival rate.

Gene expression microarrays is a new breakthrough technology developed in the late 1990s [1] that can measure the gene expression level of thousands

of genes corresponding to different samples or experiments simultaneously [2]. Many solution schemes for cancer classification and therapy process on molecular and cellular levels may be concluded from the analysis and the comparison of the generated data through different experiments [3]. Microarrays technology has two variants in the market [3], (1) cDNA microarrays-*On Spotted array*- and (2) oligonucleotide microarrays-*On GeneChip*-. cDNA microarrays are cheaper and more flexible as custom-made arrays, it was developed at Stanford University. While oligonucleotide arrays (developed at Affymetrix) are more automated, stable, and easier to be compared through different experiments [3,4]. The data produced by microarrays technology represent the result of thousands of genes for few experiments where this matrix can be used to evaluate the variation of gene through samples or the interaction of genes in different samples.

Since DNA microarray technology allows to analyse the gene data quickly and at one time in order to get the expression pattern of a huge amount of genes simultaneously [5], gene expression data are unique in their nature due to three reasons: (1) their high dimensionality (more than thousands of genes), (2) the publicly available data are very small just hundred or fewer of samples, (3) a big partial of the genes are irrelevant in cancer classification and analysis, where the problem is to find the difference between cancerous gene expression tissues and non-cancerous tissues. For these reasons, and in order to handle those kind of data researchers proposed that feature selection and/or dimensionality reduction is a relevant process in order to take advantage of the data and to converge toward accurate classifiers. Several machine learning methods have been used in caner classification, yet recently deep learning start to be investigated as well in this process due to its ability to work on raw and high dimensional data.

The paper investigates the use of advanced machine learning to handle large scale gene expression data to enhance cancer classification. Also it explores the potential of deep learning based classifiers to manage such datasets. Hence, we propose a simple feed forward neural network and implement four yet powerful classical classifiers namely, support vector machine (SVM), k-nearest neighbours (KNN), bayes naive (BN) and shallow neural network (SNN). We tested the four classifiers along with the deep classifier on publicly available five cancer datasets in the omnibus library. the cancer types are: Leukemia cancer, inflammatory breast cancer, lung cancer, bladder cancer and thyroid cancer

The remainder of the paper is organized as the following: the first Sect. 2 highlights the used classification methods. Then Sect. 3 presents an overview on the recent works related to machine learning and deep learning for gene expression and cancer classification. In Sect. 4 we explained our proposed deep feed forward neural network for the discussed problem. Then the used datasets are described in Sect. 5. Section 6 deals with the experimental study and presents the obtained results and our discussion. Finally in Sect. 7 conclusions are drawn.

2   Classification Methods

Many classification methods have been introduced through time. In the following we present four main methods.

2.1   K-Nearest Neighbours

K-nearest neighbours (KNN) classifier is the simplest supervised classifier that attempts to find the class membership of an unknown instance in the testing dataset $\{X\}$ on the basis of the majority vote of the k-nearest neighbours [6]. KNN is a *lazy learning* or an instance based learning, where the function is approximated locally and all the computation is postponed until classification [5]. When classifying a sample x, the KNN classifier finds in the testing set $\{X\}$ the most similar k examples to x and then chooses the most appropriate label class among this examples, by calculating the similarities between the attributes of the object x and the k samples. The simplest or the most used way to calculate the similarity between x and y is the geometric distance [7].

2.2   Support Vector Machine

Support Vector Machine (SVM) is also a supervised machine learning tool, that was introduced and implemented in 1995 [8] for pattern recognition. SVM was widely used for both classification and regression tasks [9]. The concept of SVM is based on [8,10–12]:

The $\{X\}$ instances of the training data set are plotted in some high-dimensional features space, where the task is to find the support vectors that maximise the margin (also the optimal hyperplane) not between the vector and the data but between the classes in the space (see Fig. 1).
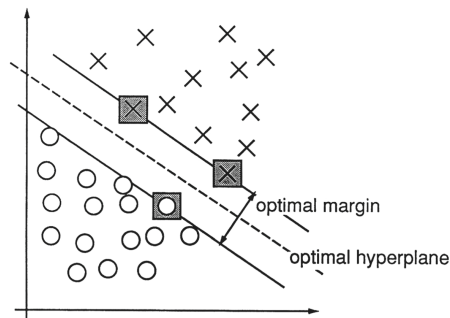
Fig. 1. An SVM example represents the maximum margin between classes in two dimensional space [8]

2.3   Naive Bayes Classifier

Naive Bayes classifier (NB) as well is one of the first simple supervised machine learning. It is a probabilistic model based on the Bayesian formula to calculate the probability of class A given the values B_i of all attributes for an instance to be classified [13]. NB classifiers follow the assumption that all attributes of a

given example are independent of each other, which facilitates the learning phase because every parameter can be learned separately, especially in the scalable data [14]. Naive bayes classifier have been intensively used in different fields such as document classification [14], Medical application like EGG signal analysis [15], music emotion classification [13] based on lyrics (text) analysis, and for image classification [16] as well.

2.4   Deep Learning

Deep Learning (DL) is the new breakthrough in machine learning and Artificial intelligence. DL migrates with machine learning technique from hand-designed features toward data-driven features-learning, where deep learning can learn complex models through simple features learned from raw data [17].

Deep Neural Networks (DNN) were the best showcase of deep learning with the aspect of multilayer that offers the possibility to explore the hierarchical representation of data by increasing the level of abstraction [18]. This properties allowed DNN to demonstrate state-of-the-art performance in different domains [19–21].

In deep learning we can find: (1) deep neural networks (DNN), (2) convolution neural network (CNN) and (3) recurrent neural network (RNN). DNN is the simplest representation of multilayer neural network. It may be either a multilayer perceptron , auto encoders (AE), stacked auto encoders (SAE), deep belief networks (DBN) or boltzman machine. While (2), convolution neural networks are built upon three majors layers convolution layers, max-pooling layers and and non-linear layer. At each convolutional layer a group of local weighted sums called features are obtained. At each pooling layer, maximum or average sub sampling of non-overlapping regions in feature maps is performed which allows CNNs to identify more complex features [17,18]. RNNs, they are designed to use sequential information, and they have a basic structure with cyclic connection. Past information is implicitly stored in the hidden units called *state vectors* using an explicit memory *long short term memory*, and the current output is computed based on all the previous input through this state vector [17].

3   Machine Learning in Gene Expression Cancer Analysis Related Work

Both supervised and unsupervised methods have been used in gene expression data analysis. in 1998 a cluster analysis based on graphical visualisation method to reveal correlated patterns between genes were proposed in [22]. Supervised machine learning served microarrays data analysis intensively and effectively [5]. Neural network were proposed in [23] for Cancer classification and diagnostic prediction. Li et al. [24] proposed a genetic algorithm/k-nearest neighbours approach in order to select effective genes that can be highly discriminative in cancer sample classification, by splitting the set of genes into several subsets and then calculate the frequency of genes' membership to the subset. After a

number of iterations the genes with high frequency are the most relevant to the classification. The latter was used recently in [25] in order to select the most discriminative genes to classify the TCGA data of 31 different cancer type. SVM also was used in the field [10], where in [26] a new SVM ensemble based on Adaboost (ADASVM) and consistency based feature selection (CBFS) was proposed for leukemia cancer classification, SVM was used to overcome the problems of regular ensemble methods based on decision trees and neural network. Where the authors cited in the former the issue of the tree size and overfitting problem in the latter. Another approach based on Battcharya distance was implemented in [27] for colon cancer and leukemia cancer. The features were selected based on their ranking score, where the genes with larger Battcharya distance are the most effective in classification. Then the subset with the lowest error classification rate is selected as the marker genes. In [28] a shallow neural network was proposed for colon cancer classification with a variation on parameter setting that uses the Monte-Carlo algorithm with SVM theory.

Recently researchers start to apply deep learning in the context [29]. Table 1 illustrates the top recent researches in the literature, where we compared the works based on the used features selection model, the classification model and its accuracy.

Table 1. Deep learning cancer classification recent research. **H/L** the highest and lowest accuracy score of the classifier depends on the dataset

Reference	Feature selection	Classification method	Accuracy	
[30]	PCA+ Sparse AE	Softmax classifier	L	H
			35.0%	97.5%
	PCA+ Stacked AE		L	H
			33.71%	95.15%
[31]	Adversarial net + CNN +RBM	Segmoid+CNN	——	
[32]	SDAE	SVM	98.04%	
		ANN	96.95%	
[33]	Desq	(KNN,SVM,DT,RF, GBDTs)+ANN	H	L
			98.80%	98.41%

Fakoor et al. [30] present the use of deep learning for cancer classification through unsupervised features learning. The proposed approach is a two phases process. The feature learning phase, where Principal Component Analysis (PCA) was used for dimensionality reduction. Since PCA is a linear representation of data, some raw features were added to capture the non-linearity of the features. Then sparse auto encoders (Stacked auto encoders in the second test) were used for the unsupervised features selection. In the second phase, the set of learned features with some of the labelled data were passed to the classifier to learn the

classifier, as well fine-tuning was used to tune the weights of the features and generalize the features set to adapt to different cancer types.

Bhat et al. [31] used adversarial model based on convolutional neural network and restricted boltzmann machine for gene selection and classification of Inflammatory Breast Cancer. The proposed generative adversarial network (GAN) is a combination of two network. The first network represent a generator that tries to mimic examples (wrong inputs) from the training data set and fed them among the real inputs to the second network. The latter works as a discriminator that tries to distinguish the true inputs from the false ones and classify the samples as accurately as possible. The process continues until the discriminator can no longer distinguish noise input from the real ones. The learnt features are passed to a sigmoid layer for supervised classification.

Danaee et al. [32] proposed stacked denoising auto encoders (SDAE) for breast cancer classification. The paper used SDAE to addresses the high dimensionality and noisy gene expression issues and to select the most discriminative genes in breast cancer classification. The selected genes have been evaluated by ANN and SVM.

In [33], a deep learning approach that combines five classical classification methods was proposed for the classification of lung cancer, stomach cancer and inflammatory breast cancer. The paper used DeSeq for features selection, then the selected features were passed through the five classifiers namely, KNN, SVM, Decision Trees (DTs), Random Forest(RF) and GBDTs in the first classification stage. The output of the first stage is used as the input for a five layer neural network to classify the samples.

4   Deep Forward Neural Network for Cancer Classification

The tackled cancer classification problem can be formulated as follows: Given a matrix $\{X\}$ of NxM dimension where N represent the number of samples and M is the number of genes, each $x_{i,j}$ represents the expression level of the gene j related to the sample i, and each sample X is associated to a class that can be either cancerous or not cancerous for binary classification. It can also refer to the the corresponding subtype of the cancer for multiclass classification. Then the problem can be binary classification or multiclass classification.

The architecture is a multilayer feed forward neural network organized as the following:

– The input layer receives the set of features that represent the gene expression values of each sample.
– Seven hidden layers have been used. Four are fully connected layers, and between the layers we added three dropout layers that applies a dropout penalty to avoid overfitting.
– An output layer with a softmax classifier is used to assign the set of received features from the Seventh hidden layer to their corresponding class.
– We applied a regularization *l2()* on the input data at the input layer level.
– For the activation of layers we used the non-linear *tanh* and *relu* functions.

Algorithm 1: Proposed architecture pseudo-code

Data: X,y
Apply one of [KPCA, FRE, UFS] for dimensionality reduction;
X_train, $X\_test < -$Split(X);
y_train, $y\_test < -$Split(y);
Build the Deep forward classifier;
Initialized the Deep forward classifier;
Define the number of epochs and the batch size;
while *iteration less than or equal to the number of epochs* **do**
    while *batch_size less than or equal to the number of samples* **do**
        X_batch, $y\_batch < - next\_batch(X\_train, y\_train)$;
        Train model(X_batch, y_batch);
        Update $batch\_size$;
    end
    Evaluate model(X_test, y_test);
    Reset $batch\_size$;
end

The pseudo-code (Algorithm 1) outlines the different steps of our proposed classifier building. We used batch training to train the network with *adam*optimizer and a *categorical crossentropy* loss. Also, we applied *hold-out* cross validation (70% training data, 30% testing data) to asses the performance of the classifier. The used performance metrics are accuracy and the loss function where the objective is to maximize the accuracy and minimize the loss without dropping in overfitting and underfitting issues.

For dimensionality reduction we used three methods namely, Kernel Principal Component analysis (KPCA) for non-linear problems, Recursive Feature Elimination (RFE) and Univariate Feature Selection (UFS). In this way we can evaluate the performance of the proposed classifier on different reduced data space.

5   Datasets

The datasets (Table 2) are publicaly available in the GEO bank (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi). They represent the expression level of patient genes that define if the samples are cancerous or not cancerous, the type and the stage of the disease. We applied data preprocessing and imputation on some of the data sets in order to handle the missing values of some genes that appear in few samples.

– *Leukimea Cancer (DS1):* The data set is stored under the key *GSE15061* [34], it represents a case study of the transformation of leukemia cancer from AML to MDS stage. the samples are all bone marrow distributed as 164 MDS patients, 202 AML patients and 69 non leukemia. The total set is 870 samples with 54613 genes.

– *Inflamatory Breast Cancer (DS2):* Stored under the key *GSE45581* [35]. The samples are the expression of IBC tumor cells and non-IBC cells. The dataset is a total of 45 samples of Inflammatory Breast Cancer (IBC) and non-IBC with 40991 genes.
– *Lung Cancer (DS3):* The dataset is stored under the key *GSE2088* [36]. It represents a set of 48 samples of squamous cell carcinoma (SSC), 9 samples of adenocarcinoma and 30 normal lung cancer samples. The total set is 87 samples of 40368 genes.
– *Bladder Cancer (DS4):* The access key is *GSE31189* [37], it represents the gene expression of human urothelial cells, it contains 52 samples of urothelial bladder cancer patient and 40 non-cancer samples. The set is 92 samples represented through 54675 genes.
– *Thyroid Cancer (DS5): GSE82208* [38], this data set has been used to differentiate between malignant and benign follicular tumours. The set is a collection of 27 samples of follicular thyroid cancer (FTC) and 25 follicular thyroid adenomas (FTA) with the dimensionality of 54675.

Table 2. The data sets description (* preprocessed data set)

Data set	Genes	Samples	Classes
DS1	54613	870	MDS, AML, non-leukemia
DS2	40991	45	IBC, non-IBC, Normal
DS3(*)	40368	87	Normal, Squamous carcinoma=SSC, Adenocarcinoma
DS4	54671	92	Cancerous , Normal
DS5	54671	52	FTC, FTA

6   Results and Discussion

For the aforementioned classical machine learning models (SVM, BN, KNN) we used the scikit-learn python package models, for the shallow network and deep neural network architecture we used sequential model of keras package with tensorflow back-end.

The experimental results (Table 3) shows the variation of the classification accuracy rate, depending on the classifier and the dimensionality reduction method. The obtained results demonstrate the usefulness of supervised machine learning in tumour classification. Yet the results also prove that the deep classifier was able to achieve better performance and score a higher accuracy (up to 100% in different cases) than the classical models.

The proposed DNN model was able to achieve the highest possible accuracy between the classifiers in many situations for the five datasets. Citing the dataset DS4, with the new feature space obtained by univariate feature selection, deep learning overcomes the other classifiers. While in DS1, DS2 respectively DS3,

the deep classifier achieved the highest accuracy score in both RFE and UFS. Whereas in DS5, for the three dimensionality reduction models deep learning was able to conquer the other classifiers.

Table 3. Comparative study results in terms of accuracy. Bold values represent the best obtained score.

Datasets	FS	SVM	KNN	BN	DNN	Shallow net
DS1	KPCA	0.44	0.**0.47**	0.40	0.45	0.44
	RFE	0.64	085	0.66	**0.90**	0.88
	UFS	0.63	0.79	0.57	**0.80**	0.79
DS2	KPCA	0.29	0.64	**0.86**	0.64	0.36
	RFE	0.28	0.42	0.64	**0.78**	0.71
	UFS	0.29	0.57	0.79	**0.85**	0.51
DS3	KPCA	0.59	**1.0**	**1.0**	0.81	0.70
	RFE	0.70	0.96	**1.0**	**1.0**	0.96
	UFS	**1.0**	**1.0**	0.96	**1.0**	0.96
DS4	KPCA	0.60	0.57	**0.82**	0.68	0.57
	RFE	0.57	0.60	**0.78**	0.64	0.60
	UFS	0.57	0.93	0.92	**0.96**	0.79
DS5	KPCA	0.38	0.56	0.81	**0.87**	0.81
	RFE	0.87	0.87	0.87	**1.0**	0.93
	UFS	0.81	**0.88**	0.81	**0.88**	0.87

Compared to SVM and shallow networks, BN and KNN performance was very promising as well. Both classifiers were able to achieve the highest score in three out of five datasets. The Bayes naive classifier performance was at its best with kernel principle components and recursive feature elimination in DS2, DS3, DS4. While KNN performed better with KPCA and UFS in DS1,DS3 and DS5. The overall performance of SVM and shallow network was good yet in the studied cases, it was not good enough compared to the deep classifier performance.

For the case where the proposed classifier was not able to achieve the best accuracy, we believe that an improved architecture (in its density, depth and parameters setting) and a better feature selection model would improve its performance. It is worth noting that the worst cases for the deep network (DS1,DS2,DS3, and DS4) was where we used KPCA as a dimensionality reduction method. This let us to make the assumption that the new feature space was not quite discriminative in order to train the deep classifier to perform accurately.

7   Conclusion

In the era of information and massive datasets, classification and machine learning have been intensively applied by computational, statistical and data analysis

researchers to mine, organize, and categorize huge data sets in order to extract a valuable knowledge and acceptable patterns in a variety of field for decades.

Recently with the advances in biological data generation and the migration of biological and medical community toward personalized medicine and cancer advanced treatment systems, scientists start to apply classification and machine learning in order to classify and extract biomarker genes that may help in the therapy process. Through this paper we have seen that machine learning was widely used from the first and classical models to the new deep learning innovation. Therefore we think it may be a key for new achievements in medical informatics. Also the experimental results and the theoretical research mainly in cancer classification problem, have proved to us that every classification model have its strength and weakness and the variation between the performance of each classifier, mainly classical models, depends on the data and the experimental environment. Also we have seen that deep learning is very effective and powerful to handle biological large scale data sets, and was able to conquer other models in their discrimination and classification accuracy. In our future contributions we will try to use deep models for the selection and identification of relevant biomarkers for cancer diagnosis, therapy process.

References

1. Bumgarner, R.: Overview of DNA microarrays: types, applications, and their future. Curr. Protoc. Mol. Biol. 22.1.1–22.1.11 (2013)
2. Zhang, X., Zhou, X., Wang, X.: Basics for bioinformatics. In: Jiang, R., Zhang, X., Zhang, M.Q. (eds.) Basics of Bioinformatics, pp. 1–25. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38951-1_1
3. Xu, Y., Cui, J., Puett, D.: Omic data, information derivable and computational needs. In: Xu, Y., Cui, J., Puett, D. (eds.) Cancer Bioinformatics, pp. 41–63. Springer, New York (2014). https://doi.org/10.1007/978-1-4939-1381-7_2
4. Harrington, C.A., Rosenow, C., Retief, J.: Monitoring gene expression using dna microarrays. Curr. Opin. Microbiol. **3**(3), 285–291 (2000)
5. Bhola, A., Tiwari, A.: Machine learning based approaches for cancer classification using gene expression data. Mach. Learn. Appl.: Int. J. **2**, 01–12 (2015)
6. Kriti, Virmani, J., Agarwal, R.: Evaluating the efficacy of gabor features in the discrimination of breast density patterns using various classifiers. In: Dey, N., Ashour, A., Borra, S. (eds.) Classification in BioApps, LNCVB, vol. 26, pp. 105–131. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-65981-7_5
7. Kubat, M.: Similarities: nearest-neighbor classifiers. An Introduction to Machine Learning, pp. 43–64. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20010-1_3
8. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)

9. Cleophas, T.J., Zwinderman, A.H.: Support vector machines. In: Cleophas, T.J., Zwinderman, A.H. (eds.) Machine Learning in Medicine, pp. 155–161. Springer, Dordrecht (2013). https://doi.org/10.1007/978-94-007-6886-4_15

10. Vanitha, C.D.A., Devaraj, D., Venkatesulu, M.: Gene expression data classification using support vector machine and mutual information-based gene selection. Procedia Comput. Sci. **47**(Supplement C), 13–21 (2015). Graph Algorithms, High Performance Implementations and Its Applications (ICGHIA 2014)

11. Kubat, M.: Inter-class boundaries: linear and polynomial classifiers. An Introduction to Machine Learning, pp. 65–90. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20010-1_4

12. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, New York (2014)

13. An, Y., Sun, S., Wang, S.: Naive Bayes classifiers for music emotion classification based on lyrics. In: 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), pp. 635–638, May 2017

14. McCallum, A., Nigam, K., et al.: A comparison of event models for Naive Bayes text classification. In: AAAI-98 Workshop on Learning for Text Categorization, Madison, WI, vol. 752, pp. 41–48 (1998)

15. Sharmila, A., Geethanjali, P.: Dwt based detection of epileptic seizure from EEG signals using naive bayes and k-NN classifiers. IEEE Access **4**, 7716–7727 (2016)

16. Karthick, G., Harikumar, R.: Comparative performance analysis of Naive Bayes and SVM classifier for oral X-ray images. In: 2017 4th International Conference on Electronics and Communication Systems (ICECS), pp. 88–92, February 2017

17. Yann, L., Yoshua, B., Geoffrey, H.: Deep learning. Nature **521**, 436–444 (2015)

18. Min, S., Lee, B., Yoon, S.: Deep Learning in Bioinformatics. ArXiv e-prints, March 2016

19. Elleuch, M., Maalej, R., Kherallah, M.: A new design based-SVM of the CNN classifier architecture with dropout for offline arabic handwritten recognition. Procedia Comput. Sci. **80**(C), 1712–1723 (2016)

20. Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L., Somogyi, R.: Large-scale temporal gene expression mapping of central nervous system development. Proc. Natl. Acad. Sci. **95**(1), 334–339 (1998)

21. Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J.: Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat. Biotechnol. **33**(8), 831–838 (2015)

22. Michaels, G.S., Carr, D.B., Askenazi, M., Fuhrman, S., Wen, X., Somogyi, R.: Cluster analysis and data visualization of large-scale gene expression data. Pac. Symp. Biocomput. **3**, 42–53 (1998)

23. Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., et al.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. **7**(6), 673–679 (2001)

24. Li, L., Darden, T.A., Weingberg, C., Levine, A., Pedersen, L.G.: Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. Comb. Chem. High Throughput Screen. **4**(8), 727–739 (2001)

25. Li, Y., Kang, K., Krahn, J.M., Croutwater, N., Lee, K., Umbach, D.M., Li, L.: A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. BMC Genomics **18**(1), 508 (2017)

26. Begum, S., Chakraborty, D., Sarkar, R.: Cancer classification from gene expression based microarray data using SVM ensemble. In: 2015 International Conference on Condition Assessment Techniques in Electrical Systems (CATCON), pp. 13–16, December 2015

27. Ang, J.C., Haron, H., Hamed, H.N.A.: Semi-supervised SVM-based feature selection for cancer classification using microarray gene expression data. In: Ali, M., Kwon, Y.S., Lee, C.-H., Kim, J., Kim, Y. (eds.) IEA/AIE 2015. LNCS (LNAI), vol. 9101, pp. 468–477. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19066-2_45

28. Chen, H., Zhao, H., Shen, J., Zhou, R., Zhou, Q.: Supervised machine learning model for high dimensional gene data in colon cancer detection. In: 2015 IEEE International Congress on Big Data, pp. 134–141, June 2015

29. Urda, D., Montes-Torres, J., Moreno, F., Franco, L., Jerez, J.M.: Deep learning to analyze RNA-seq gene expression data. In: Rojas, I., Joya, G., Catala, A. (eds.) IWANN 2017. LNCS, vol. 10306, pp. 50–59. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59147-6_5

30. Fakoor, R., Ladhak, F., Nazi, A., Huber, M.: Using deep learning to enhance cancer diagnosis and classification. In: Proceedings of the International Conference on Machine Learning (2013)

31. Bhat, R.R., Viswanath, V., Li, X.: Deepcancer: detecting cancer through gene expressions via deep generative learning. CoRR abs/1612.03211 (2016)

32. Danaee, P., Ghaeini, R., Hendrix, D.A.: A deep learning approach for cancer detection and relevent gene identification, pp. 219–229. World Scientific (2016)

33. Xiao, Y., Wu, J., Lin, Z., Zhao, X.: A deep learning-based multi-model ensemble method for cancer prediction. Comput. Methods Programs Biomed. **153**, 1–9 (2018)

34. Mills, K.I., Kohlmann, A., Williams, P.M., Wieczorek, L., Liu, W.M., Li, R., Wei, W., Bowen, D.T., Loeffler, H., Hernandez, J.M., Hofmann, W.K., Haferlach, T.: Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of AML transformation of myelodysplastic syndrome. Blood **114**(5), 1063–1072 (2009)

35. Woodward, W.A., Krishnamurthy, S., Yamauchi, H., El-Zein, R., Ogura, D., Kitadai, E., Niwa, S.I., Cristofanilli, M., Vermeulen, P., Dirix, L., Viens, P., van Laere, S., Bertucci, F., Reuben, J.M., Ueno, N.T.: Genomic and expression analysis of microdissected inflammatory breast cancer. Breast Cancer Res. Treat. **138**(3), 761–772 (2013)

36. Fujiwara, T., Hiramatsu, M., Isagawa, T., Ninomiya, H., Inamura, K., Ishikawa, S., Ushijima, M., Matsuura, M., Jones, M.H., Shimane, M., Nomura, H., Ishikawa, Y., Aburatani, H.: ASCL1-coexpression profiling but not single gene expression profiling defines lung adenocarcinomas of neuroendocrine nature with poor prognosis. Lung Cancer **75**(1), 119–125 (2012)

37. Urquidi, V., Goodison, S., Cai, Y., Sun, Y., Rosser, C.J.: A candidate molecular biomarker panel for the detection of bladder cancer. Cancer Epidemiol. Prev. Biomark. **21**(12), 2149–2158 (2012)

38. Wojtas, B., Pfeifer, A., Oczko-Wojciechowska, M., Krajewska, J., Czarniecka, A., Kukulska, A., Eszlinger, M., Musholt, T., Stokowy, T., Swierniak, M., Stobiecka, E., Chmielik, E., Rusinek, D., Tyszkiewicz, T., Halczok, M., Hauptmann, S., Lange, D., Jarzab, M., Paschke, R., Jarzab, B.: Gene expression (mRNA) markers for differentiating between malignant and benign follicular thyroid tumours. Int. J. Mol. Sci. **18**(6) (2017)