Targeted unsupervised features learning for gene expression data analysis to predict cancer stage

Imene Zenbout
imene.zenbout@univ-constantine2.dz
Abedlhamid Mehri University Constantine-2-,NTIC Faculty,Department of Fundamental Informatics and its Applications
Constantine, Algeria
CRBT , CERIST
Algeria

Abdelkrim Bouramoul
abdelkrim.bouramoul@univ-constantine2.dz
Abedlhamid Mehri University Constantine-2-,NTIC Faculty,Department of Fundamental Informatics and its Applications, Misc Laboratory
Constantine, Algeria

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

ABSTRACT

The intensive explosion in the generation of large scale cancer gene expression data brought several computational challenges, yet opened great opportunities in exploring different pathways in order to improve cancer prognosis, diagnosis and treatment. In this paper, we propose a targeted unsupervised learning model, based on deep autoencoders (TAE) to learn significant cancer representation based on the gene expression omnibus(GEO) integrated expO data set, for the ultimate goal of constructing an accurate cancer stage predictive model. Where, the trained model was tested on two gene expression cancer data sets namely, lung cancer for clinical stage and intensive breast cancer (IBC) for pathological stage. In which, the model extracted new features space for the two cancer type based on the knowledge built from the expO data set. The generated features were used to train classifiers to predict the cancer stage of each sample. We evaluated the effectiveness of our proposal by comparison to the principal component analysis (PCA) unsupervised dimensionality reduction, as well as to the supervised univariate features selection method. The experimental results, show a promising performance of our analysis model to build a collaborative knowledge from different cancer type to enhance the prediction rate of different cancer stage.

KEYWORDS

Unsupervised learning, Feature selection, Gene expression, Deep Learning, Auto encoder, Bioinformatics

ACM Reference Format:
Imene Zenbout, Abdelkrim Bouramoul, and Souham Meshoul. 2019. Targeted unsupervised features learning for gene expression data analysis to predict cancer stage. In *10th International Conference on Computational Systems-Biology and Bioinformatics (CSBio 2019), December 4–7, 2019, Nice, France.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3365953.3365958

1 INTRODUCTION

Due to the remarkable advance in high throughput sequencing technologies namely, microarrays[6] and next generation sequencing(NGS)[17]. Many avenues have been opened to enhance oncology research[3, 14], where the community science work on answering the major challenges around cancer causes, evolution and behaviors, as well as to choose the appropriate way to deal with each case separately.

Microarrays and NGS have been used to generate various omic data for clinical and research questions. In this paper we focus on gene expression data (GE), that represents the abundance of mRNA level in the selected samples(tissue, bone marrow, blood ...etc) .These technologies allow the expression measuring of thousands of genes simultaneously, under different experiments [28]. The generated results are gathered in a series matrix that represents the set of GE values for each patient. Where, the major challenge when handling this kind of data is their high dimensional genes space(more than ten thousands) compared to the low number of samples (a hundreds or less) [5], whereas a major part of these genes are irrelevant in building diagnosis, prognosis and predictive cancer models. This constraint, complicates the task of valuable knowledge extraction from this noisy dimensional space. In this context, the extraction of important genes and eliminating the irrelevant one can play a key role in enhancing the medical decisions. Many computational and/or statistical models have been introduced for cancer gene expression analysis, some works in the literature are discussed in section2.

GE data sets are publicly available on specific websites because of the number of projects funded by the oncology research consortium, that aim to collect the possible amount

CSBio 2019, December 4–7, 2019, Nice, France

ACM ISBN 978-1-4503-7215-2/19/12. . . $15.00
https://doi.org/10.1145/3365953.3365958

of cancer samples and create reference data banks such as *GEO* and*TCGA*, with the ultimate goal to improve cancer related clinical decisions and promote oncology research. Integrated cancer datasets projects *(expO for GEO, PanCancer for TCGA)* have been lunched, trying to collect the biggest number of cancer samples regardless of the cancer type, along with the patients medical record and clinical outcomes. The goal behind these data sets is to help in understanding and solving the problems of poorly explored cancer types(for further information you can visit http://www.intgen.org/).

In this work, we aim to investigate the GEO expO dataset, following the stategy of building a targeted features learning model trained on the expO data set that contains various cancer types to select discriminative features that help in cancer stage prediction. The trained model is applied to extract new features from other single cancer type cohort data sets, namley:lung cancer, and intensive breast cancer. For the feature learning model we applied, deep unsupervised learning through deep autoencoders(AE). The use of deep learning as a computational approach was due to its caracteristiques in handeling, raw and noisy data[26], which is often the case of the gene expression data [6], as well as the performance of AE in the generation of reduced data spaces[22]. The generated features space from the two data sets is to be further applied to train a supervised classifiers to predict cancer stage (clinical and/or pathological).

The rest of the paper is organized as the following. The second section 2, discusses a set of related works about machine and deep learning in cancer gene expression analysis. Section 3 explains the notion of unsupervised features learning autoencoder. The data collection and preprocessing is detailed in section 4. Then, the proposed model and the set of experimentation are explained in section 5.

2 DEEP LEARNING IN CANCER GENE EXPRESSION DATA ANALYSIS

Many statistical methods have been applied to analyse gene expression data to extract valuable knowledge in the purpose of enhancing medical decision toward cancer patients. Machine learning approaches were a prioritised choice for a big majority of researches, where a lot of papers were introduced in the context, (see table1). For instance the work of Friedman et al[12], in which the authors applied Bayesian network in order to analyse the set of GE data. In the same path Li et al[18] have used Genetic Algorithm as a feature selection model combined with K-nearest-neighbours to select the set of relevant genes in Colon and Leukemia cancer. Support-Vector-Machine also appeared in several preposition, where we can cite the work of Ang et al[1] and S, Begum et al[24]. In the former, the paper used Battacharya distance to distinguish the relevant cancer genes from those with no impact on cancer diagnosis, in addition to SVM classifier. The later applied a linear kernel SVM through an ensemble method to analyse the Leukaemia cancer.

Whereas, with the remarkable advance in deep learning starting from the famous event ImageNet in 2012 until nowadays[16], the bioinformatic and the computational biology communities raised their assumptions; whether deep learning may play a significant role in biological data analysis, and in our point of interest gene expression and cancer data analysis[6]. For answering this, a variety of contributions were proposed, here we will cite some reviews like [2, 17, 21] for further research. For GE analysis, several deep models were applied to accomplish different analytical related tasks, citing the paper of Fakoor et al [11], the authors used an unsupervised learning method based on stacked autoencoders(SAE)to enhance cancer diagnosis. The controbution was divided in two phases.In the first phase, principal component analysis(PCA) and unsupervised stacked auto encoder(SAE) were used for dimensionality reduction. Then in the second phase, a supervised logistic model was trained by the reduced space to classify the cancer samples. Besides, Bhat et al [4] used Generative Adversarial Nets(GAN) for the selection of significant genes signature in breast, by using a combination of Boltzman machines and Convolutional Neural Network. Auto encoder, again was used in the work of Padideh et al [22] through Staked-Denoising-autoecoder(SDAE) for TCGA breast cancer data set analysis. The authors first used Kernel-PCA(KPCA) for dimensionality reductionused, then SDAE for unsupervised features learning, and a supervised classifier to assess the performance of the learned features space. A cancer gene characteristic selection model was proposed by Jian et al [20] to identify the most effective genes in cancer prognosis and diagnosis using samples learning instead of features learning combined with deep sparse filtering. The contributors used samples learning to transform the samples space of GE datasets in order to keep the traceability of the most relevant cancer genes, so that the new transformed space is employed for the gene selection process by a deep forward sparse filtering network.

3 DEEP AUTOENCODER FOR FEATURE LEARNING

Autoencoder is a popular and powerful neural network model that can learn hidden representation from unlabeled data[7]. The main idea of autoencoder is the conjunction of two functions; the *encoder* function $F(x)$ and the *decoder* function $G(F(x))$. Where the encoder maps the d-dimensional input space S into some reduced k-dimensional hidden space $H = F(S)$, while the decoder *reconstruct* this new representation H to its prior d-dimension in a way that the output is as close as possible to the original input of the encoder $S' = G(H)$ [25]. The assessment of the autoencoder performance is through minimizing the reconstruction error value by comparing the encoder inputs to the reconstructed decoder outputs i.e. the smallest is the value of reconstruction error the more S and S' are identical. A simple autoencoder is represented through a single layer percepteron for both the encoder (equation 1) and the decoder (equation 2) [13].

$$F_e(x) = \sigma_e(\omega_e x) = H \tag{1}$$

Table 1: Review table of Machine learning and deep learning models used in gene expression cancer data analysis

Approach	Classification Models	Features Selection	Cancer Type	Ref
Machine Learning	KNN	Genetic Algorithm	Colon, Leukemia	[18]
	ADASVM	CBFS	Leukaemia	[24]
	SVM	Battchareya Distance	Colon, Leukemia	[1]
	KNN	Genetic Algorithm	PanCancer TCGA	[19]
	Shallow-nets + monte-carlo Algorithm	/	Colon	[8]
Deep Learning	supervised classifier	KPCA+ SDAE	Breast	[22]
	GAN+ CNN+ RBM	/	Breast	[4]
	Logistic Classifier	PCA+SAE	13 types	[11]
	KNN+ SVM+ DT+ RF+ GBDTs, DL	DEsq	Lung , Stomach, IBC	[27]
	/	Sample learning, Sparse filtering+ FFNN	Five types	[20]

$$G_d(H) = \sigma_d(\omega_d H) \tag{2}$$

Where $\sigma()$ is a non-linear activation function such as sigmoid, tanh, or relu. The $\{\omega_e, \omega_d\}$ represent the weight parameter of the layers. A deep autoencoder is an asymmetric (figure1) multi layer representation for the encoder and the decoder with a bottleneck layer which represent the hidden layer H. Regarding the set of the deep autoencoder parameters, stochastic gradient descent(SGD) based optimization procedure can be used to train the model with an objective of minimizing the reconstruction error[13, 15, 25].

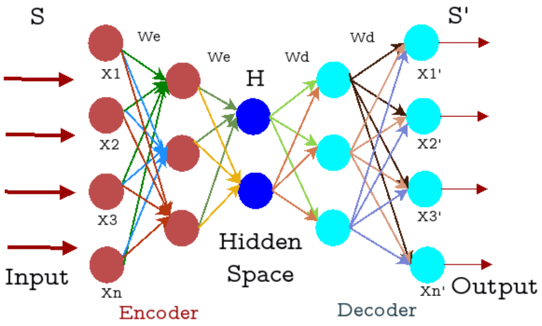

Figure 1: Deep AutoEncoder Architecture

4 DATA COLLECTION AND PREPROCESSING

We have collected three gene expression data sets(Table 2) from the omnibus bank, The expO data sets have been used for the unsupervised training, while the two candidate data sets have been applied for the evaluation and supervised classification.

- The expO dataset defined under the accession key *GSE2109*(date 09-01-2019 at 14:31), is composed of 2185 samples collected from patients with different type of cancer from different United State hospitals, the mRNA ubundance was measred using *Affyrmatix* microarrays. Along with the gene expression series matrix, the data set contains various patient information such as age, history of the disease, clinical and pathological Stage ...etc. As a preprocessing step We have performed a log transfromation on the whole data matrix to handle skewed data. After, we have applied Principal component analysis PCA unsupervised dimensionality reduction in order to reduce the data space for computational needs. This data have been used on two phases in the first phase, the whole data was used whether the sample have its stage lable or not. In the second phase, only the samples with a corresponding stage label were gathered in a new expression matrix.
- As For the second lableled datasets, the Lung cancer[10] (date 24-01-2019 at 20:30)and the intensive breast cancer (IBC)[23] (date 25-01-2019 at 4:29). The two series matrix were also generated by the Affyrmatix microarrays. The former is composed of 150 instances represent the clinical stage of the disease(Stage 1/stage 2). While, the later is composed of 366 pathological stage sample. We have used log transformation for both data set to treat the skewed data as well. But for the IBC data set we have applied data imputation, by replacing the missing mRNA expression values with the mean score.

5 PROPOSED ARCHITECTURE

The proposed architecture consists of two major phases. The first phase is an unsupervised features learning model based

Table 2: Data description

Use	Name	Accession key	Number Of Samples	
Features Learning Data	expO	*GE2109*	Total Number: 2185	
			Clinical	Stage 1: 192
				Stage 2: 054
				Stage 3: 060
				Stage 4: 014
			Pathological	Stage 1: 288
				Stage 2: 359
				Stage 3: 369
				Stage 4: 173
Classification Data	Lung Cancer	*GSE43580*	Clinical	Stage 1: 075
				Stage 2: 075
	IBC	*GSE86166*	Pathological	Stage 1: 138
				Stage 2: 209
				Stage 3: 209

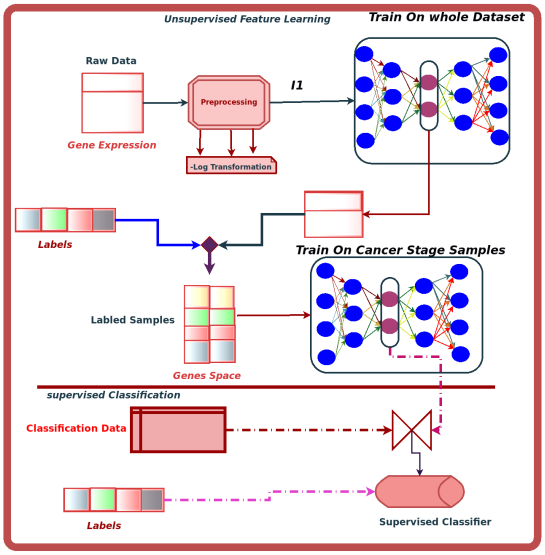

Figure 2: The proposed architecture for expO dataset analysis and cancer staging

on two steps, the first step is a multi-layer autoencoder E_1, in charge of taking the high dimensional unlabeled data S and try to map it into a reduced vector space H. From the generated vector H we selected only the labeled Samples X_y and the set of features is than used to train another autoencoderE_2 to direct the features set toward the remote target of prediction. The trained middle layer of E_2 is then used in the second phase as a features selector for new data sets. Finally, each sample from the generated features space by the targeted auto encoder (TAE) E_2 is associated to its corresponding labels in order to train a supervised classifier to predict the output of the samples .

5.1 The unsupervised autoencoder:

The first autoencoder E_1 is built to be trained on the whole samples, regardless of their cancer type, gender, or the causes of the classification. The architecture of E_1 is composed of:

- Four fully connected layers that build the encoder(2000, 1000,500,250).
- The hidden layer, where 50 instances were selected to represent the new features space.
- Four fully connected layers represent the decoder(250, 500,1000,2000).

5.2 The targeted unsupervised autoencoder(TAE):

The second autoencoder E_2 is applied for features selection tuning; here we have used the features generated by the bottleneck layer of E_1 and select only the labeled samples(Cancer stage)for training. The selected samples are then employed to train E_2. E_2 is represented through a simple multi-layer perceptron:

- A one fully connected layer encoder with 50 inputs.
- A hidden layer, that extract the 45 features used in prediction.
- A symmetrical fully connected layer for the decoder with 50 outputs.

5.3 The supervised phase:

The selected features using E_2 have been used to train the classifier to predict the class y of each sample X, for the case of multi-classification we have used *One Versus Rest* classifier to build a predictive model for each class.

For the autoencoder $E1$ we have taken the gene expression dataset as the input space set S which is a high dimensional space of range k, where each attribute i represent a candidate gene in a x sample of the data set. The encoder F_e takes the samples with the set of attributes and progress it through its

Table 3: Unsupervised phase experimental parameters

	E_1	E_2
Parameter	Value	Value
epochs	150	50
batch size	32	25
learning rate	0.02	0.02
Dropout	[0.001-0.1]	0.1
activation function	tanh-relu	tanh
Reconstruction loss	**2.637**	**0.003/0.004**

layers, where in each layer a non linear function(tanh, relu...) is applied on the nodes to update the models weights, so at the end the vector H is generated by $H = F_e(x) = \sigma(w_e x)$. The decoder G_d as mentioned in *section3*, tries to reverse the procedure and generate an output S' that is approximately equals to the encoder input S where $G_d(H) = S' \approx S$. In our contribution we focus only on the compacted features space H, not on the reconstructed instances S', despite we evaluate the consistency of the H vector by comparing the encoder input S and the decoder output S' through calculating the reconstruction loss error, in this paper we used the $mean\_absolute\_error$ eq3.

$$mae = 1/n \sum_{i=1}^{n} |x_i - x_i'|/n = \sum_{i=1}^{n} |e_i|/n \qquad (3)$$

The same procedure has been applied to E_2, only here the input is the set of labeled H vector samples. The encoder F_e generate a new feature vector H_f and the decoder G_e generate the reconstructed input H'.

5.4 Experimental results

The proposed architecture (autoencoder+Classifiers) has been implemented using Python-Anaconda, based on Keras package with a tensorflow-backend[9]. We used the default Keras configuration to initialise the autoencoders($E1/E2$) layers. After a series of executions we find that our models converge to their best performance with the following parameters (Table3). We fixed the number of epochs as 150/50 rounds for E_1 and E_2 respectively ,and a learning rate equals to 0.02 for both autoencoders. The overfitting issue have been solved through applying a drop-out penalty.

Training: In this hierarchy, we followed a two phase training topology(figure3), the unsupervised training phase and the supervised training phase as following:

In the first phase we trained the autoencoders E_1 using an *sgd* optimizer on the whole data space (labeled and unlabeled samples). After we extracted from the hidden layer the features representation space, that represents the samples X which posses a cancer stage label y (322 sample for clinical stage and 1833 for pathological samples). and for each pipeline we tuned the features by the second autoencoder through an *sgd* optimizer.

After the unsupervised training, the trained model E_2 has been used to extract the most relevant features in the new data sets(Clinical stage and Pathological stage), and use these features in association to their labels to train the supervised classifier. As for the classifiers we selected, support vector machine(SVM), Naive Bayes(NB), and k-nearest-neighbors(KNN), which allowed us to test the relevancy of the generated features space. The used classifiers have been trained and tested using k-fold cross validation, in our experiment we have used 10-fold, where 9 folds are used for training the classifier and the last fold is to evaluate its performance.

5.4.1 Evaluation and results discussion. In this level, we have applied different tests to measure the performance of the trained targeted auto-encoder(TAE) in learning significant representations. We have used two data sets namely Lung cancer for clinical stage prediction and IBC for pathological stage prediction. Where we followed the training pipeline illustrated in Figure 3 to build our classifiers. Mainly we used the trained TAEs (clinical/pathological) to extract a new features representation for both data set, after we used this new features representation to train the classifiers to predict the corresponding cancer stage. To add more relevancy and competitiveness to our work, we compared the results of the classifiers trained on TAE features to other features generated by state of the art models. Here we have used PCA as an unsupervised dimensionality reduction model to plot a features space of 45 instances and then train new classifiers on this space to predict cancer stage. Also, we applied Univariate features selection (UFS), that selects the highest score 45 features associated to cancer stage. This comparison helped in capturing the overall performance of our proposal compared to the other models. The results are detailed in (Table4, Table5). Although *Accuracy* is an important measure to identify the classification rate of the supervised classifier, its not sufficient to visualise its performance on the different class labels, that is why we have chose *precision*, *recall*, and $f_1 - score$ to assess the performance frequency of the classifier on each class.

The results represented in Table4, show that the classifiers trained on our TAE model achieved the highest performances compared to those trained on UFS and PCA. The set of features extracted through TAE have improved the ability of the classifiers to discriminate between samples of the two stages of cancer, with a precision of 100% for stage 1 and 86% for stage 2 of the decease, and 93% overall accuracy for SVM and KNN. The rate of results obtained from UFS was acceptable compared to PCA, where the classification performance was so low in differentiating between the samples of two classes, like in the PCA/SVM the score is 0% precision for stage 1 of cancer and 33% recall rate for satge 2 in BN with an overall accuracy in range[$40\% - 60\%$] for all classifiers.

Table 5, exhibit the performance of the different classifiers trained to predict the pathological stage for IBC cancer. As previously, the best results have been achieved using the classifiers trained on TAE features. Among the nine classifiers, the one with the best performances on the whole

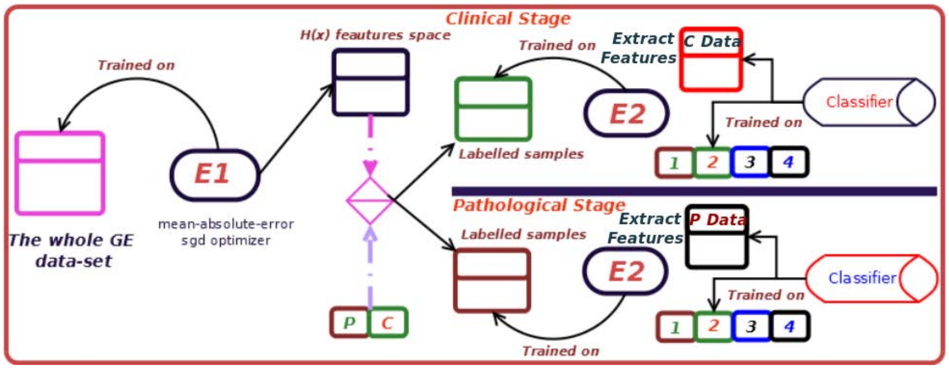

Figure 3: Training pipeline model

Table 4: Clinical Stage Lung cancer classification performance of different experiments FSM: Features Selection Model, C1/C2: Stage1/Stage2

		Precision		Recall		f_1-Score		Accuracy
FSM	Classifier	C_1	C_2	C_1	C_2	C_1	C_2	C_1/C_2
TAE	SVM	**1.00**	**0.86**	**0.89**	**1.00**	**0.94**	**0.92**	**0.93**
	BN	**1.00**	0.75	0.78	**1.00**	0.88	0.86	0.86
	KNN	**1.00**	**0.86**	**0.89**	**1.00**	**0.94**	**0.92**	**0.93**
UFS	SVM	**1.00**	0.75	0.78	**1.00**	0.88	0.86	0.86
	BN	0.85	0.84	0.81	0.88	0.83	0.86	0.84
	KNN	0.67	0.83	0.86	0.62	0.75	0.71	0.73
PCA	SVM	0.00	0.40	0.00	**1.00**	0.00	0.57	0.4
	BN	0.64	0.5	0.78	0.33	0.70	0.40	0.60
	KNN	0.5	0.36	0.22	0.67	0.31	0.47	0.40

set of metrics was the SVM/TAE classifier, in which the classification performance for all measures were above 50% for the three stages.

The TAE/BN classifier was very powerful in predicting satge 1 and stage 2 of the disease, yet its performance was very low for the third stage with a score less than 35%. TAE/KNN performance was acceptable for stage 1 and 2 but also very poor for stage 3. Compared to TAE, UFS classifiers performance is considered to be good in predicting stage 2 of cancer, yet poor for stage 3, with a precision in range [20% − 33%], a recall between [14% − 29%] and a $f_1 - score$ not greater than 25%. The PCA based classifiers showed a very weak classification rate of cancer stage, the accuracy rate along all the classifiers was lower than 45%. In PCA/SVM all the instances were classified as stage 2 IBC, while in PCA/BN and PCA/KNN showed moderate results in Stage 1 and Stage 2 of cancer. From the IBC result we can conclude that our TAE was able to conquer the other tested model in classifying stage three of the disease with at least a difference of 40%. We assume that by training the TAE on a bigger dataset , we may achieve better results. Also we assume that using new validation models like cross-validation on the current data set, as well as on new data cohort may improve the supervised phase performance.

6 CONCLUSION

In this paper we investigated the genetic signature impact of different cancer types on a specific cancer diagnosis task, and how integrated cancer types could direct the performance of a specific analytical model on cancers that tend to be hard to investigate; due to their complexity or the the lack of datasets. The use of the expO dataset as the integrated set, in addition to the powerfulness of deep auto-encoder in data integration, allowed us to build a good model that can select new features representation to be used in cancer staging both clinical and pathological. The experimental results showed very promising views to be explored in this direction, yet we aim to ameliorate the results of the unsupervised feature learning space by collecting new data sets from different sources to train the model in order to gain more efficiency in generating a powerful discriminating features. The second phase in the unsupervised representation learning may be tuned or rather transferred to extract new spaces that may be used for other diagnosis and/or prognosis procedures.

REFERENCES

[1] Jun Chin Ang, Habibollah Haron, and Haza Nuzly Abdull Hamed. 2015. Semi-supervised SVM-based feature selection for cancer classification using microarray gene expression data. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 468–477.

Table 5: Pathological Stage IBC cancer classification performance of different experiments FSM: Features Selection Model, C1/C2/C3: Stage1/Stage2/Stage3

		Precision			Recall			f_1-Score			Accuracy
FSM	Classifier	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3	$C_1/C_2/C_3$
TAE	SVM	0.64	0.57	**0.67**	0.64	0.62	**0.57**	**0.64**	0.59	**0.62**	0.61
	BN	0.57	**0.79**	0.33	**0.73**	**0.85**	0.14	**0.64**	**0.81**	0.20	**0.64**
	KNN	0.56	0.62	0.22	0.45	0.62	0.29	0.50	0.62	0.25	0.48
UFS	SVM	0.47	0.64	0.33	**0.73**	0.54	0.14	0.57	0.58	0.20	0.51
	BN	0.56	0.62	0.22	0.45	0.62	0.29	0.50	0.62	0.25	**0.64**
	KNN	0.38	0.46	0.20	0.27	0.46	0.29	0.32	0.46	0.24	0.35
PCA	SVM	0.00	0.42	0.00	0.00	**1.00**	0.00	0.00	0.59	0.00	0.41
	BN	**0.67**	0.45	0.20	0.36	0.69	0.14	0.47	0.55	0.17	0.45
	KNN	0.17	0.42	0.17	0.09	0.62	0.14	0.12	0.50	0.15	0.32

[2] Davide Bacciu, Paulo JG Lisboa, José D Martín, Ruxandra Stoean, and Alfredo Vellido. 2018. Bioinformatics and medicine in the era of deep learning. *arXiv preprint arXiv:1802.09791* (2018).

[3] Sam Behjati and Tarpey Patrick, S. 2013. What is next generation sequencing? *JArchives of disease in childhood* 98, 6 (2013), 236–8.

[4] R. R. Bhat, V. Viswanath, and X. Li. 2017. DeepCancer: Detecting Cancer via Deep Generative Learning Through Gene Expressions. In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*. 901–908.

[5] Amit Bhola and Arvind Tiwari. 2015. Machine Learning Based Approaches for Cancer Classification Using Gene Expression Data. *Machine Learning and Applications: An International Journal* 2 (12 2015), 01–12.

[6] Roger Bumgarner. 2013. Overview of DNA microarrays: types, applications, and their future. *Current protocols in molecular biology* 101, 1 (2013), 22–1.

[7] B Chandra and Rajesh K Sharma. 2015. Exploring autoencoders for unsupervised feature selection. In *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–6.

[8] Huaming Chen, Hong Zhao, Jun Shen, Rui Zhou, and Qingguo Zhou. 2015. Supervised machine learning model for high dimensional gene data in colon cancer detection. In *2015 IEEE International Congress on Big Data*. IEEE, 134–141.

[9] François Chollet et al. 2015. Keras. https://keras.io. (2015).

[10] IMPROVER DSC Collaborators, Adi L. Tarca, Christoph Zechner, Erhan Bilal, Florian Martin, Gustavo Stolovitzky, Heinz Koeppl, Jeremy J. Rice, Julia Hoeng, Kushal Kumar Dey, Manuel Peitsch, Mario Lauria, Marja Talikka, Michael Unger, Pablo Meyer, Preetam Nandy, Raquel Norel, Roberto Romero, Stephanie Boue, and Yang Xiang. 2013. Strengths and limitations of microarray-based phenotype prediction: lessons learned from the IMPROVER Diagnostic Signature Challenge. *Bioinformatics* 29, 22 (2013), 2892–2899.

[11] Rasool Fakoor, Faisal Ladhak, Azade Nazi, and Manfred Huber. 2013. Using deep learning to enhance cancer diagnosis and classification. In *Proceedings of the International Conference on Machine Learning*, Vol. 28. ACM New York, USA.

[12] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. 2000. Using Bayesian networks to analyze expression data. *Journal of computational biology* 7, 3-4 (2000), 601–620.

[13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

[14] Rajeshwar Govindarajan, Jeyapradha Duraiyan, Karunakaran Kaliyappan, and Murugesan Palanisamy. 2012. Microarray and its applications. *Journal of pharmacy & bioallied sciences* 4, Suppl 2 (2012), S310.

[15] Ozan Irsoy and Ethem Alpaydın. 2017. Unsupervised feature extraction with autoencoder trees. *Neurocomputing* 258 (2017), 63–73.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[17] Kimberly R Kukurba and Stephen B Montgomery. 2015. RNA sequencing and analysis. *Cold Spring Harbor Protocols* 2015, 11 (2015), pdb–top084970.

[18] Leping Li, Thomas A Darden, CR Weingberg, AJ Levine, and Lee G Pedersen. 2001. Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. *Combinatorial chemistry & high throughput screening*. (2001).

[19] Yuanyuan Li, Kai Kang, Juno M Krahn, Nicole Croutwater, Kevin Lee, David M Umbach, and Leping Li. 2017. A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. *BMC genomics* 18, 1 (2017), 508.

[20] Jian Liu, Cheng Yuhu, Wang Xuesong, Zhang1 Lin, and Jane Wang Z. 2018. Cancer Characteristic Gene Selection via Sample Learning Based on Deep Sparse Filtering. *SIENTIFIC REPORTS* (2018).

[21] S. Min, B. Lee, and S. Yoon. 2016. Deep Learning in Bioinformatics. *ArXiv e-prints* (March 2016). arXiv:cs.LG/1603.06430

[22] Danaee Padideh, Ghaeini Reza, and A Hendrix David. 2016. A deep learning approach for cancer detection and relevent gene identification. In *Pacific Symposium on Biocomputing*.

[23] Sangeetha Prabhakaran, Victoria T Rizk, Zhenjun Ma, Chia-Ho Cheng, Anders E Berglund, Dominico Coppola, Farah Khalil, James J Mulé, and Hatem H Soliman. 2017. Evaluation of invasive breast cancer samples using a 12-chemokine gene expression score: correlation with clinical outcomes. *Breast Cancer Research* 19, 1 (2017), 71.

[24] Begum S, Chakraborty D, and Sarkar R. 2015. Cancer classification from gene expression based microarray data using SVM ensemble. In *2015 International Conference on Condition Assessment Techniques in Electrical Systems (CATCON)*.

[25] Shuyang Wang, Zhengming Ding, and Yun Fu. 2017. Feature selection guided auto-encoder. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[26] Lecun . Yann, B. Yoshua, and H. Geoffrey. 2015. Deep Learning. *Nature* (May 2015).

[27] Xiao Yawen, Wu Jun, Lin Zongli, and Zhao Xiaodong. 2018. A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine* (2018).

[28] Xuegong Zhang, Xueya Zhou, and Xiaowo Wang. 2013. *Basics for Bioinformatics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–25.