ISSN 2320 - 2602
Volume 9 No.7, July 2020
International Journal of Advances in Computer Science and Technology
Available Online at http://www.warse.org/IJACST/static/pdf/file/ijacst17972020.pdf
https://doi.org/10.30534/ijacst/2020/17972020

A SURVEY ON CANCER CLASSIFICATION AND PREDICTION FROM GENE EXPRESSION DATA USING DEEP LEARNING

Deepthi V[1], Vineetha S[2]
[1]Computer Science and Engineering RIT, Kottayam, India, deepthiharidas143@gmail.com
[2]Computer Science and Engineering RIT, Kottayam, India, svineetha@rit.ac.in

ABSTRACT

Nowadays cancer related deaths are increasing. Breast Cancer is one of the major cause for it. Chance of occurring breast cancer among women is very high as compared to that of the men. Cancer is mainly caused due to gene mutation. Gene expression among patients who undergone the treatment is analysed to deepens the knowledge of the disease progression

and prognosis. Because of high dimensionality and data complexity, cancer detection from gene expression data is very difficult. Nowadays, there are several deep learning approaches for detection and identification of cancer causing gene. This paper gives an overview of prediction and classification of cancer from gene expression using deep learning.

Key words: Deep Learning, Dimensionality Reduction, Cancerous cell, Gene Expression.

1. INTRODUCTION

Gene expression analysis leads to many important discoveries in the field of medicine. These analysis helps in identifying the variant gene from gene expression that in turn helps in diagnosis of cancer. Gene expression analysis measure the activity level of genes within a given tissue and thus provides information with respect to the activities in the related cells. Gene expression data are derived by measuring the number of messenger ribonucleic acid or mRNA produced during Transcription process. Several researches are still studying the problem of Classification, prediction of Cancer based on gene expression using deep learning and unsupervised methods. Since these methods are used rarely in gene expression because of high dimensionality and lack of data availability. Several technologies are introduced in order to reduce dimensionality.

2. APPLICATION OF CANCER PREDICTION AND PROGNOSIS

Goals are entirely different in case of cancer prediction and prognosis from that of cancer detection and diagnosis. In cancer prediction and prognosis mainly concerned with prediction of :

- Cancer susceptibility.
- Cancer recurrence.
- Person’s chance of betterment or survivability.
- Motivate people to compare the variations in gene expressions along with clinical data.
- Help to plan how to carry out the treatment

First case trying to predict the possibility of developing cancer ahead to the occurrence of the disease. Second case trying to predict the possibility of redeveloping of cancer after probable settlement of the disease. Third case is trying to predict life expectancy, survivability, progression etc after the disease diagnosis. In the latter two cases prognostic predictions success dependent on the quality of the diagnosis. Anyhow disease prognosis can only come after a medical diagnosis and a prognostic prediction.

3. DIFFERENT APPROACHES

Several methods are used to detect cancer from gene expression data. Some methods like Recursive Feature Elimination, Univariate Association Filtering to select features from gene expression set. Recursive Feature Elimination using SVM is used to find gene expression set. Informative gene are selected by different classification approaches along with correlation based feature selector. Studying gene characteristics helps to deepens the understanding cancer Classification, detection etc. It has also been Applied to various applications such as discovery of drug, cancer prediction and diagnosis during cancer treatment. Several studies are still doing on cancer detection and classification.

Perl CGI : J. Herrero et. al. [1] present a web tool for pre-processing the microarray gene expression data. Data is analysed first and then undergo appropriate transformation. Software used here is Perl CGI(Common Gateway interface) script . The pre-processing of microarray data is performed by Perl CGI, which acts as a main gateway to gene expression analysis tool. The novelty of this is the presence of Pre-Analysis module that determine which transformation are the most appropriate one. Pre-analysis module also perform several checks and plots different histograms. After these pre-analysis module the real pre-processing occur. The sever accepts input flat files. So pre-processing is applied to the dataset. Preprocessing include filtering, transformation etc. At end user can download processed data in sever file format. The main pros of this is to take the advantage of server capabilities and can guarantee the use of latest version of software.

Single gene classifier : Xiaosheng Wang et. al. [2] Use a single gene to construct classification models for molecular classification of cancer. Identifying a single gene from the dataset having large number of genes with large differences in their expression between the classes is not very difficult. But it is difficult to find gene from the dataset of very few genes with large difference in their expression between the classes, as it is very difficult to build the single gen classifier. Because the gene selected may be the noise-gene with the greatest apparent degree of differential expression. Other than the standard methods, Single gene classifiers are mainly used for the selection of noise genes. In the training set some noise genes could have good t-test or WMW test statistics. Using such genes, single gene classifiers were build, the performance of such classifiers are poorer than the classifiers built with longer gene list. Classifiers using longer gene list prevent from falling into the trap of noise genes. Single gene classifiers present in this paper have a advantage in time efficiency for development and evaluation in cross validation. Optimal Complexity depends upon the number of genes selected in the classifiers, criteria of gene selection and classification rules employed. Training set was developed by applying the entropy based discretization method. This will find the optimal cut point for the single gene to be selected on the basis of t or WMW statistics. This finding is also included in the selection of single gene. By using discretization in gene selection methods can cause missing of informative genes but it is less in t-test or WMW gene selection approach. Within training set, t-test or WMW test, the statistically significant genes are identified. If there are multiple genes with the smallest p value (very rare),then one with the smallest order number in the dataset is chosen. Once single gene is selected for the training set, classification rule is constructed based on a single cut-point for the expression levels of that gene. Classifier performance is evaluated by comparing the single gene model with other standard classifiers and other standard models.

Unsupervised Learning Approach: Rasool Fakoor et. al. [3] present a technique for the detection and classification of types of cancer based on gene expression data. This paper uses unsupervised feature learning for cancer detection and cancer type analysis from gene expression data. The advantage of this method over previous method is that it can apply to data from different cancer data types. And also help to detect and classify cancer type based on gene expression. There are mainly 2 phases in the feature learning approach.

- Due to high dimensionality, data is pre-processed by applying PCA to reduce dimensionality.
- Develop a sparse encoder which has been used to learn features

After applying PCA, only linear function of the input data is extracted. Inorder to extract the non-linearity relations between expression of different gene, different feature learning method is used in the second phase (i.e. sparse autoencoder). Three different of sparse encoders have been used to learn features. By using PCA and sparse feature technique provide potential to overcome problems of traditional approaches with feature dimensionality and limited size data sets. This method improve the accuracy in cancer classification. Here result can't be improved by adding unlabeled data into feature sets. Because dataset is specialized microaaray to which not a lot of samples available.

MLP-DAE model: Rui Xie et. al. [4] A deep learning model as show in fig.1 based on the Multilayer Perceptron (MLP) and Stacked Denoising Auto-Encoder (DAE) is used to predict gene expression from genotype. MLP maps input of the model i.e, single nucleotide polymorphisms (SNP) genotype, onto output after pre-processing. 2 autoencoder are used here i.e, autoencoder1 and autoencoder2 which serve as hidden layer in this model and trained using backpropogation algorithm.

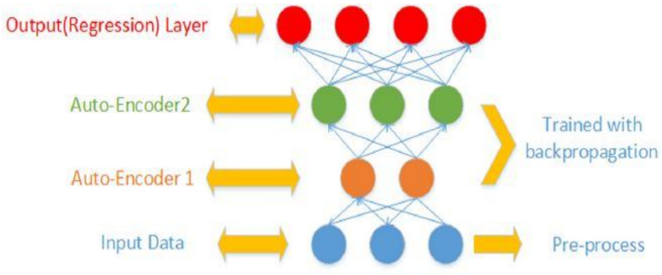

Figure 1: Model of MLP and DAE

Autoencoder helps to reduce dimension by extracting meaningful information. Training is carried out by 2 steps. In first step training autoencoder with a stochastic gradient descent algorithm. Second step utilises 2 autoencoders as two hidden layers and train them with MLP. For optimization backpropagation algorithm is applied. After training, optimal model is selected by cross validation and performance is evaluated on independent dataset. Dropout technique is used to improve performance.

SDAE model: Padideh Danaee et. al. [5] Uses a Deep Learning Approach for cancer detection and Relevant Gene Identification. Detection of cancer from gene expression data is very challenging due to high dimensionality and complexity of data. In this paper SADE is used for dimensionality reduction. For reducing the dimensionality of the data linearly PCA is also used but it doesnot extract some nonlinear relationships of the data. Then an autoencoder is applied to capture nonlinear relationships. But single autoencoder cant extract all useful information from high dimensional noisy data. So SADE with multi-layered architecture is used so that it can extract meaningful patterns from these data without losing meaningful information. Dimensionality is reduced incrementally In this RNA-seq expression data from The Cancer Genome Atlas database is analysed for tumour and healthy breast samples.

In SDAE as shown in fig.2 , output produced is close to the input but in the lower dimension. Autoencoder consist of encoder and Decoder. Encoder is non-linear function and convert matrix of higher dimension to lower dimension. Input layer encodes the data to generate a hidden or latent layer. Decoder takes the hidden representation from previous layer and decodes the data close to the input Extracting meaningful features from gene expression data helps in the classification of cancer cells. Overfitting problem in the learning phase of SDAE is avoided by dropout technique. 5-fold cross-validation is used for splitting dataset into test and train set. The highest accuracy was attained by using SDAE features applied to SVM classification.

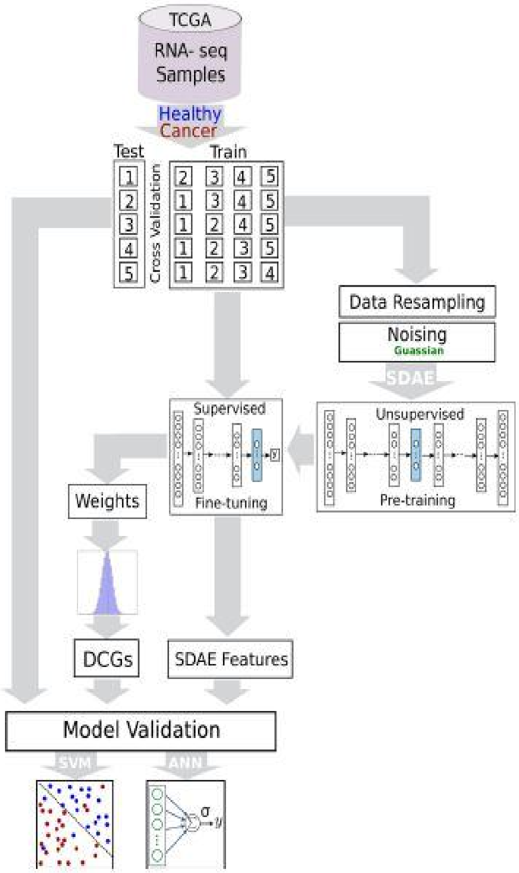

Figure 2: SDAE

MLP-SAE model: Xie R et. al. [6] present a deep learning model (MLP-SAE) as shown in fig.3 based on Multilayer Perceptron (MLP) and Stacked Denoising Autoencoder SAE. Here SAE is used for feature extraction and MLP is for backpropogation. As gene expression changes immediately reflect into the phenotype of organisms. MLP and Autoencoder have their own similarities and differences. Both consist of input layer, output layer and hidden layer. Activation function can also applied to both autoencoder and MLP. But the autoencoder reproduce input data by supervised learning. MLP used to predict the target value from the given input. So in this model and SAE are used together to predict gene expression from MLP genetic variations. Model consist of one input layer, one output layer and 2 hidden layer. Input is SNP genotype from yeast. SNP genotype is the measurement of genetic variation of single nucleotide polymorphism between members of a species. Input undergo processing then enter into the model. Output layer produce output as predicted gene expression value. Regression model is the output layer of the model. MLP-SAE model is trained and optimised by backpropagation algorithm. Training is carried out by two steps. After training cross validation is performed for selecting the optimal model. Performance also evaluated.

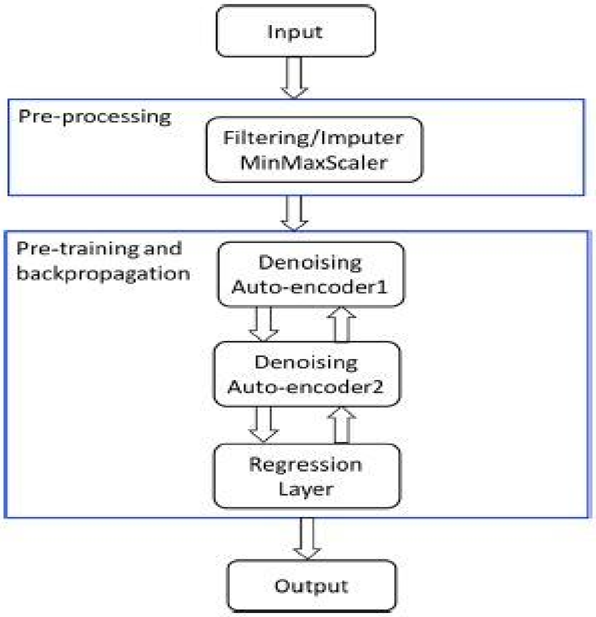

Figure 3: MLP-SAE Model

Summary Stage model vs Joint model: Rohit J. Katea et. al. [7] introduce a Stage-Specific Predictive Models for Breast Cancer Survivability. Cancer stages identified based on tumor size and extend of spreading. This case use summary stages i.e, Insitu, Localised, Regional and Distant stages. In Insitu summary stages, cancerous cells are restricted to layer of cells in which they are developed. In Localized summary stages, cancerous cell is limited to the organ in which it began. In Regional summary stages, it starts to spread into nearby lymph node tissue and organs. In Distant summary stages it spread to distant lymph node, tissues and organ. Survivability in this case is defined as surviving for 5years after the diagnosis. In this study naive Bayes, logistic regression and decision tree machine learning classification methods are used to predict breast cancer survivability. Here it checks the difference in the performance of Machine Learning Models. 3 different machine learning methods are compared with traditional joint models. The summary stage model is trained by data from respective summary stage. Joint model is trained by data that incidence of all summary stages. The performance on all summary stages together found to be better than the individual summary stages. Joint model offers no advantage over separate summary stage specific models.

Ensemble Classifier: Sara Tarek et. al. [8] present an ensemble classifier which increases the performance of classification and provide better results. K-NN algorithm is used as the base classifier. Classification of cancer is mainly based on the gene expression. This classifier can overcome three main drawbacks of the existing system, i.e, they can enhance the accuracy of the result, next is that this ensemble technique can be applied to more cancer types, and prevent the effect of over-fitting. 3 datasets are used here. 4 modules are there in this system: Pre-processing Module, Gene Selection Module, Classification Ensemble Module and Post-processing Module. Dataset is prepared in the Pre-processing module by Filtering, Thresholding, Logarithmic transformation and Data normalisation. Three different algorithms are used by the proposed system mainly for feature selection in First Module i.e, Gene Selection module: BAHSIC (Backward Elimination Hilbert Schmidt Independence Criterion), EVD (Extreme Value Distribution Based Gene Selection) and SVD Entropy (Singular Value Decomposition Entropy Gene Selection). Proposed system mainly contain of 5 base classifiers. The second module consist of sub-modules: first one is Error Estimation Module In this module an error estimator BRE is used. It is comparatively more accurate and faster error estimator. If sample is correctly classified semi-BRE is used which updates the error count and goes for the next sample. Majority Voting Module is the final step which combine the predictions made by ensemble member classifiers. Jasmir et. al. [9] Build a classification model as shown in fig.4 which classify breast cancer based on recurrence and non-recurrence event. Algorithm used is Multilayer perceptron algorithm with backpropagation rule is used. Data is utilized from different data centres. A computer Aided Diagnosis (CAD) system based on deep convolution neural network (cnn). This framework can predict and correctly diagnose the disease. Network is trained using Multilayer Perceptron Algorithm. In evaluation phase, 10-fold cross validation is used to check the accuracy of the model.

AlexNet Function: Karim Faez et.al. [10] design a system for the early diagnosis of prostate cancer. System is designed with One Deep Learning and three ANN algorithms to improve the existing system accuracy. The dataset is divided into 2 set, the training set and the test set. System is designed based on the AlexNet function. AlexNet function is a CNN based method of deep learning that is trained on more than a million images from Image Net database. Designed system is compared with SVM and ANN. It has higher accuracy than the existing system. Disadvantage is the selection of more accurate deep learning algorithm for problem solving. From all comparison deep learning algorithm predicts prostate cancer more accurately.

DNN model: Ahn et.al. [11] build a universal classifier based on the assumption that there is a universal nature that differentiate normal tissue and cancerous tissue by taking the advantages of neural network. The classification of cancerous cell and normal cells on the basis of gene expression is very challenging and is overcome by supervised Machine Learning method like SVM, SOM etc. Dataset is made by using microarray and RNA-seq data. Dataset is divided into test set and train set. For training 5 gene set is prepared. Each gene set is trained for cancer and normal prediction model. Feed-Forward network is established by using tensorflow to identify cancer and normal data. ReLU provides higher performance. Predicted model is compared with SVM based mode build by SVM function (using e1071 package in R), logistic regression, ridge lasso and elastic net (glmnet package in R). Cross validation was set by pROC packages. Comparison showed that DNN have better performance. DNN algorithm is applied to DNN model to find the single gene’s contribution to the outcome. If output is positive, then there is high

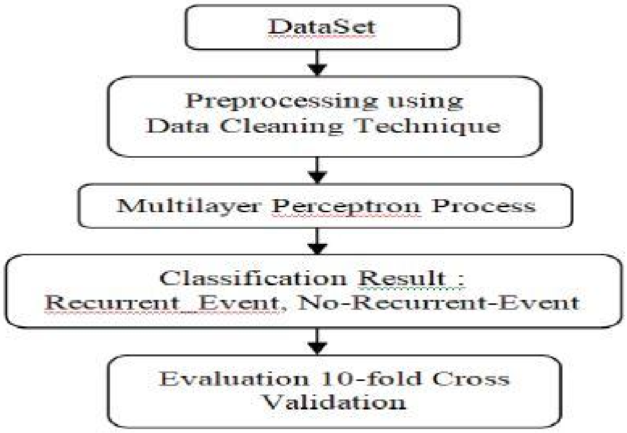

Figure 4: Flow chart of classification process

probability of sample to be cancerous. If the output is close to zero, then there is little effect of gene expression on the DNN outcome.

DeepCC: Feng Gao et. al. [12] present a novel deep learning-based framework for the classification of cancer molecular subtype. DeepCC has several advantages mainly platform independent, robust to missing data and used for simple sample prediction. DeepCC mainly contain 2 steps

- In first step, high-throughput gene expression data is transformed to functional spectra. This functional spectrum is the enrichment score of all gene set.
- In second step Classification is performed based on deep learning. By taking this functional spectra as input and classifier is trained by using deep learning. Feature selection is challenging due to high dimensional data.

METHOD	TECHNIQUE	ADVANTAGE	DISADVANTAGE
Perl CGI Script	Software presented perform functions for microarray data preprocessing. Novelty of this system is Pre-analysis module	1) Web server can take advantage of server capabilities and guarantees the use of latest version of software 2) Fast implementation 3) Avoid problem derived from particular file format	1) Whole data processing required too much time.
Entropy Based Discretization Method	Single gene classifier is developed and trained by this method for molecular classification of cancer	1) Helps to identify single noise causing gene in the data set.	1) Discretization method may cause missing of informative genes 2) Cause overfitting problem
Unsupervised and Deep Learning Methods	1) Mainly used for Detection and Classification of different types of cancer from gene expression data 2) Dimensionality is reduced by PCA and Autoencoder.	1) Deal with gene expression of different types of cancer	1) Result can't be improved by adding unlabelled data to the dataset because dataset is specialized microarray
MLP-SAE	Deep learning model which predict gene expression from genotype of organisms	1) Dropout Technique is used to improve the performance	1) Processing time is high
Deep learning SDAE	Deep learning Architecture SDAE is used to extract meaning features from gene expression data that enable classification of cancer cells	1) Avoid overfitting problem in learning phase by using dropout technique	1) Deep learning approaches requires large data set which is not available for cancer tissues
MLP-SAE	Deep learning model is build on MLP and SAE to extract gene expression from genotype	1) Avoid overfitting problem 2) Model is applicable to all organisms	1) Data processing requires time
Machine Learning Methods	3 Machine Learning Methods are used to build a model that predict the breast cancer survivability for each stages and compared them with traditional joint model	1) Model is applicable to all cancer data type	1) It is limited to the third stage
Ensemble Classifier	Using K-NN as the base classifier, system can classify the cancer based on gene expression	1) Enhancement in the Result Accuracy 2) Avoid overfitting problem 3) Model is applicable to all cancer data type	1) K-NN have scalability problem
Computer Aided Diagnosis System	Using Multilayer perceptron a classification model is trained to classify cancer based on recurrence event and non-recurrence event	1) Avoid overfitting problem 2) Model is applicable to all cancer data type	1) Still missing values of data is present even after data preprocessing
Alex Net Function	System is designed based on Alex Net function using one deep learning and 3 ANN algorithm which help in early diagnosis of prostrate cancer	1) High Result Accuracy 2) Avoid overfitting problem	1) Selection of more accurate deep learning algorithm for problem solving
DNN Classification Model	Model can classify cancerous cell and normal cell from gene expression samples by using neural networks	1) Enhancem-ent in the Result Accuracy 2) Avoid overfitting problem 3) Based on the dominance of single gene that whole sample treated as cancerous	1) Can't determine which gene contribute to DNN outcome in a individual sample

Supervised Classification Framework	Cancer Classification Framework is developed based on functional spectra quantifying activities of biological pathways	1) Platform independent 2) Robust to missing data values 3) Used for simple sample prediction	1) Availability of Clinical data

Table 1: Comparison Table

Dataset is downloaded from Bioconductor and it is in their processed form. The frame in DeepCC is implemented on the basis of MXNeT it is a open-source deep learning software framework used to train and deploy neural network.

Table above describes different technique used. From these it is possible to understand advantages and disadvantages of these techniques.

4. CONCLUSION

There are many approaches for the prediction and classification of cancer cells and normal healthy cells using gene expression by deep learning methods. But its very challenging due to the dimensionality of gene expression. As a result complexity of network is very high. To find the global patterns in the data of gene expression several unsupervised learning techniques are used. Several techniques like PCA, autoencoder etc. are used to reduce dimensionality. SADE convert high dimensional, noisy gene expression data to lower dimension meaningful data. Identification of gene from gene expression that cause cancer is very important in the field of cancer diagnosis and treatment. Different deep learning (DL) architectures have several advantages in using huge data so can able to predict and classify cancer more effectively.

ACKNOWLEDGEMENT

The authors would like to acknowledge the contribution and support from the Computer Science and Engineering Department of Rajiv Gandhi Institute of Technology.

REFERENCES

1. J. Herrero, R. Diaz-Uriarte and J.Dopazo, "**Gene Expression Data Preprocessing**", Bioinformatics(19), no.5 2003,pages 655-656, 2003.
2. Xiaosheng Wang and Richard Simon, "**Microarray-based Cancer Prediction Using Single Genes**" , BMC Bioinformatics(12), no.1 391, 2011.
3. Rasool Fakoor, Faisal Ladhak, Azade Nazi, Manfred Huber "**Using Deep Learning To Enhance Cancer Diagnosis And Classification**", Proceedings Of the 30th International Conference On Machine Learning, JMLR: WCP vol 28, 2013.
4. Rui Xie, Andrew Quitadamo y, Jianlin Cheng and Xinghua Shi "**A Predictive Model of Gene Expression Using A Deep Learning Framework**", IEEE International conference on Bioinformatics and Biomedicine (BIBM)pp 676-681, 2016.
5. Padideh Danaee, Reza Ghaeini, and David A. Hendrix, "**A Deep Learning Approach For Cancer Detection And Relevant Gene Identification**", Pacific Symposium on Biocomputing, World Scientific, pp 219-229, 2017.
6. Xie R, Wen J, Quitadamo A, Cheng J, Shi X. "**A Deep AutoEncoder Model For Gene Expression Prediction**", BMC genomics, no.9:845, 2017.
7. Rohit J. Katea, Ramya Nadigb, "**Stage-Specific Predictive Models for Breast Cancer Survivability**", International Journal of Medical Informatics, Volume 97, 2017.
8. Sara Tarek , Reda Abd Elwahab, Mahmoud Shoman, "**Gene Expression Based Cancer Classification**", Egyptian Informatics Journal,science direct, 2017.
9. Jasmir, Siti Nurmaini,Reza Firsandaya Malik, Dodo Zaenal Abidin,Ahmad Zarkasi, Yesi Novaria Kunang, Firdaus, "**Breast Cancer Classification Using Deep Learning**", International Conference on Electrical Engineering and Computer Science (ICECOS), 2018.
10. DeHengame Abbasi Mesrabadi, Karim Faez, "**Improving Early Prostate Cancer Diagnosis By Using Artificial Neural Networks And Deep Learning**", 2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS),IEEE, 2018.
11. Ahn, T., Goo, T., Lee, C. H., Kim, S., Han, K, Park, S., Park, T. "**Deep Learning-Based Identification Of Cancer Or Normal Tissue Using Gene Expression Data**", IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1748-1752), 2018.
12. Feng Gao, Wei Wang, Miaomiao Tan, Lina Zhu, Yuchen Zhang, Evelyn Fessler, Louis Vermeulen and Xin Wang, "**DeepCC: A Novel Deep Learning-Based Framework For Cancer Molecular Subtype Classification**", Gao et al. Oncogenesis , 2019.