Deep learning-based ovarian cancer subtypes identification using multi-omics data

Team Details:

Garima Singh (1806143)

Mrinal (1806149)

Published: 24 August 2020

Authors:

- Long-Yi Guo
- Ai-Hua Wu,
- Yong-xia Wang,
- Li-ping Zhang,
- Hua Chai &
- Xue-Fang Liang

Second School of Clinical Medicine, Guangzhou University of Chinese Medicine, Guangzhou, 510020, China

Center for Reproductive Medicine, Guangdong Hospital of Traditional Chinese Medicine, Guangzhou, 510120, China

PMID: 32863885

PMCID: PMC7447574

DOI: 10.1186/s13040-020-00222-x

Project Report

November 11th, 2020

Index

- Details of selected paper
- Data set description
- Sample dataset
- Model proposed in paper
- Model implemented by us
- Differences in models
- Results obtained
- Sample output
- Code walk-through
- Conclusion
- References

Details of selected paper

Title: Deep learning-based ovarian cancer subtypes identification using multi-omics data

Published: 24 August 2020

Authors:

- Long-Yi Guo
- Ai-Hua Wu,
- Yong-xia Wang,
- Li-ping Zhang,
- Hua Chai &
- Xue-Fang Liang

1. Second School of Clinical Medicine, Guangzhou University of Chinese Medicine, Guangzhou, 510020, China
2. Center for Reproductive Medicine, Guangdong Hospital of Traditional Chinese Medicine, Guangzhou, 510120, China
3. School of Data and Computer Science, Sun Yat-sen University, Guangzhou, 510000, China

PMID: 32863885

PMCID: PMC7447574

DOI: 10.1186/s13040-020-00222-x

Data set description

- We used the multi-omics ovarian data for our training and testing purpose. Here, multi-omics dataset refers to the dataset containing multiple omes which in our cases are genomes.
- The data type of our dataset is real. The given data basically contains the mutation in genes of the patients suffering from ovarian cancer over a period of time.

SOURCE :

The Cancer Genome Atlas(TCGA) shared the data of patients suffering from various types of cancers. The data is available on their website https://portal.gdc.cancer.gov. The data has been shared from Surgical Oncology Research Lab of Boston and UCLA School of Medicine. The data is generated from the RNA-seq and mRNA information of the patients.

In this study we utilized the multi-omics ovarian data for training and three datasets in GEO were used as the independent tests. The details about these four datasets were introduced in following:

TCGA dataset

We downloaded the multi-omics ovarian cancer data from TCGA public datasets((https://portal.gdc.cancer.gov). The *R* package *TCGA-assemble2* [13] was used for data collection and we obtained 298 samples concluded three types of omics data: mRNA-seq data (UNC Illumina HiSeq_RNASeq V2), miRNA-seq data (BCGSC Illumina HiSeq) and copy number variation (CNV) data (BROAD-MIT Genome wide SNP_6). All these data were obtained from the TCGA level 3 data. And the CNV feature was extracted by averaging the copy numbers of all CNV variations on one gene. After that the features and samples which missing more than 20% would be excluded. For the remaining data, the missing values were imputed based on the median values by using *R* package "*imputeMissings*" [14].

Test datasets

In GSE26712 we downloaded the RNA-seq and the clinical information of 185 ovarian cancer patients shared from Surgical Oncology Research Lab of Boston, and in GSE32062 we got 260 ovarian cancer case samples offered by Obstetrics and Gynecology, Niigata University. GSE53963 contains mRNA information of 174 samples from UCLA School of Medicine. All of these test datasets can be downloaded in Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov)

Sample dataset

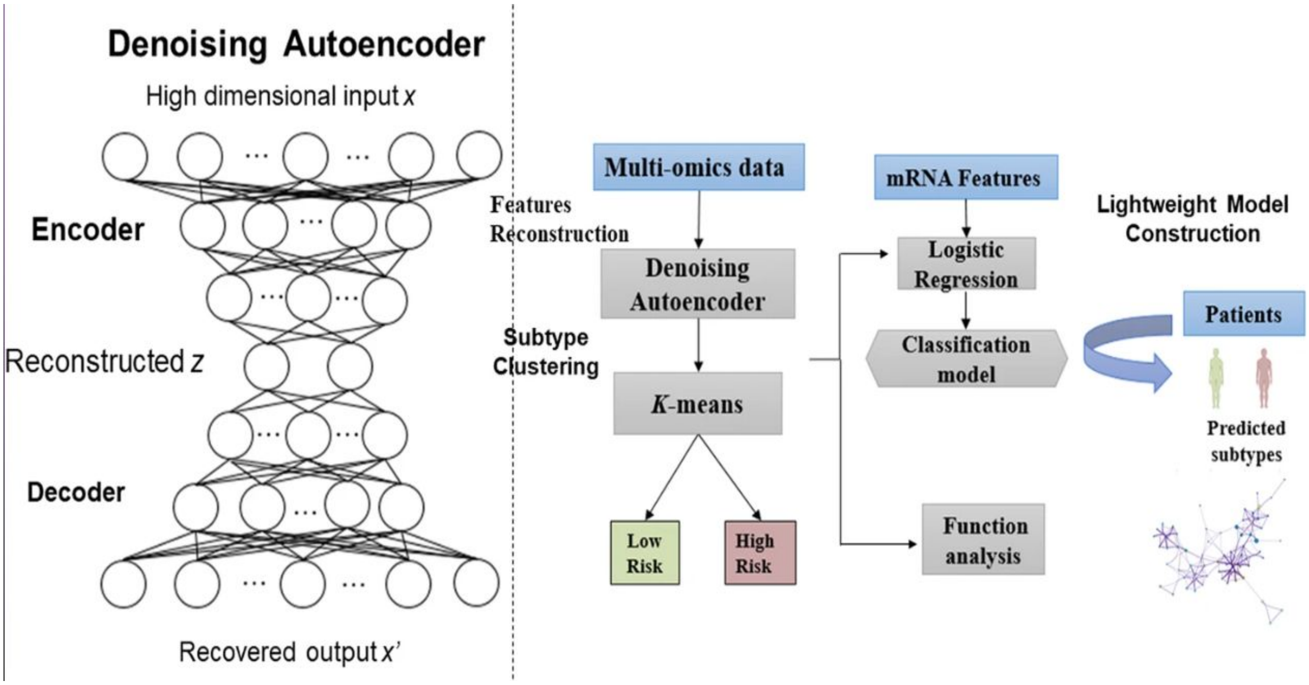
TCGA.2W.A8YY.01	TCGA.4J.AA1J.01
-0.369700774	-0.711083223
-0.376244799	-0.747875378

TCGA(The Cancer Genome Atlas) dataset contains the reports of various cancer patients including their imaging, cancer type and gene mutation and then the dataset is prepared on all the factors.

We have considered the data of two patients suffering from cancer whose data was prepared based on three types of omics data which includes mRNA, miRNA and CNV. These data are further studied to classify various types of cancers and also for the treatment purpose of the patient.

Model proposed in paper

The model proposed in the paper is a logistic regression classification model with mRNA data. The cluster subtype used to build the model is *k*-means.



The architecture of proposed deep learning framework

In Fig. 1 we show the architecture of proposed deep learning framework, firstly the multi-omics ovarian cancer features *x* (mRNA, miRNA and CNV) are inputted into the DAE for generating the low dimensional representation *z*. And then the reconstructed features *z* are used to cluster the patients using the *k*-means. Based on the clustered subtypes from *k*-means, we further built the light-weighted logistic regression classification model with mRNA expression data to reducing the features required for patients' classification. The available code of this deep learning framework was shared in https://github.com/Hua0113/DAE_km.

Evaluations of ovarian cancer subtypes identification

We implemented different methods for comparing the performances of different cluster methods: k-means, hierarchical clustering, k-means using the reconstructed features by PCA (PCA-kmeans), SparseK, iCluster, k-means using the reconstructed features by KPCA (KPCA- kmeans), k-means using the reconstructed features by AE (AE-k means) and DAE-kmeans. The silhouette score is used to measure the cluster performance and the log rank *p*-value to measure the differences of the different subtypes of cancers. The higher silhouette score means the method achieved better performance for clustering, and the lower log-rank p-value means the greater differences in cancer subtypes.

Model implemented by us

The model used by us is a logistic regression classification model which uses K-means clustering techniques. We have chosen to align with the paper specifications and techniques as closely as possible but changed certain values to improve accuracy score.

The multi-omics data of patients are inputted into the Denoising Autoencoder for generating *z*. With the help of generated *z*, the patients are clustered using *k*-means. The optimal number of clusters was determined using *silhouette score*. In the model, we tested the *k* from [2, 8] and we finally used *k=2* as it had the highest *silhouette score*.

After obtaining the labels clustered by *k*-means, we built a light-weighted mRNA model for reducing the number of genes needed to identify cancer subtypes by using a logistic regression algorithm.

Differences

The paper used DAE-*k* means a clustering method which we have already seen. We used *k*-means which has a lower silhouette score compared to the one used in paper. The optimal number of clusters are again determined using *silhouette score*. In the model, we tested the *k* from [2, 8] and we finally used *k=2* as it had the highest *silhouette score*.

After obtaining the labels clustered by *k*-means, we built a light-weighted mRNA model for reducing the number of genes needed to identify cancer subtypes by using a logistic regression algorithm.

Results obtained

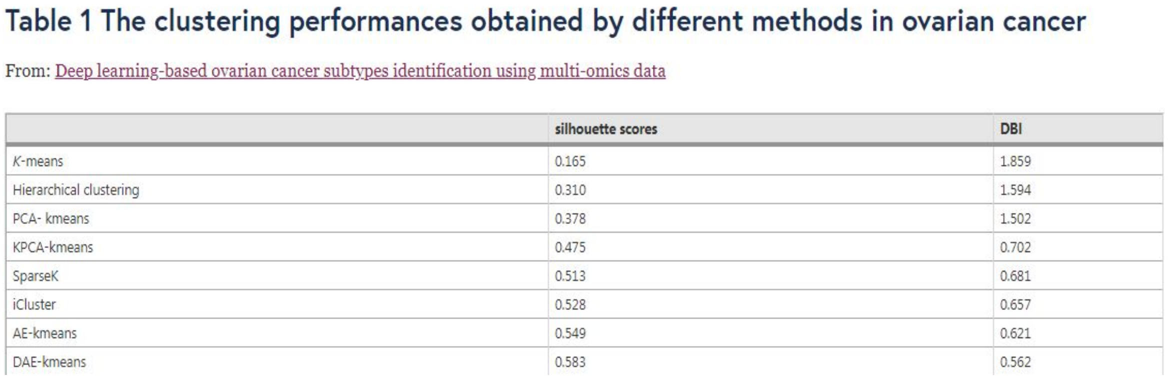


Table 1 The clustering performances obtained by different methods in ovarian cancer

From: Deep learning-based ovarian cancer subtypes identification using multi-omics data

	silhouette scores	DBI
K-means	0.165	1.859
Hierarchical clustering	0.310	1.594
PCA- kmeans	0.378	1.502
KPCA-kmeans	0.475	0.702
SparseK	0.513	0.681
iCluster	0.528	0.657
AE-kmeans	0.549	0.621
DAE-kmeans	0.583	0.562

In Table 1 we show the clustering performances obtained from different methods by using ovarian cancer multi-omics data which contained mRNA, miRNA and CNV. We used the silhouette scores and Davies Bouldin scores (DBI) to evaluate the clustering performances of the methods. It is obvious that without any dimensionality reduction method, *K*-means achieved lowest silhouette score and highest DBI among these methods. And the methods based on traditional dimensionality reduction methods (PCA, KPCA) performed only better than *k*-means and hierarchical clustering, but worse than SparseK, iCluster and two deep learning-based methods. The results in Table 1 prove the power of deep learning, and DAE-kmeans perform better than any other methods indicated the superiority of our method.

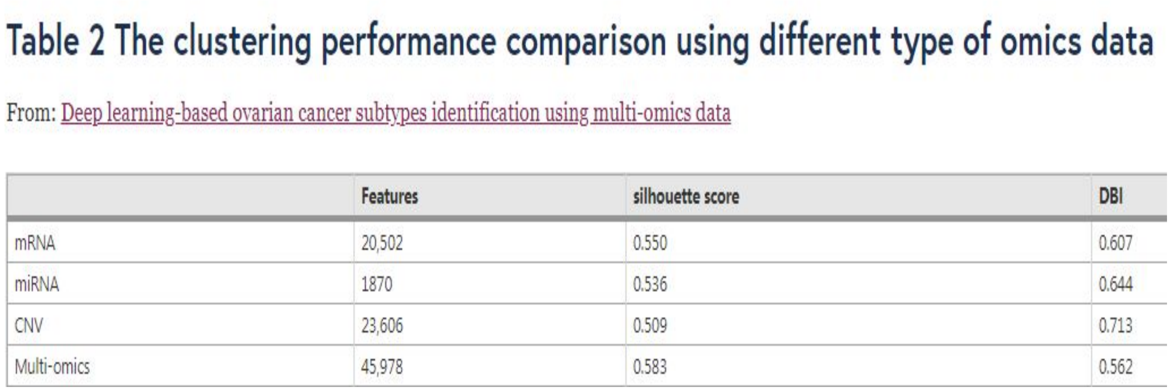


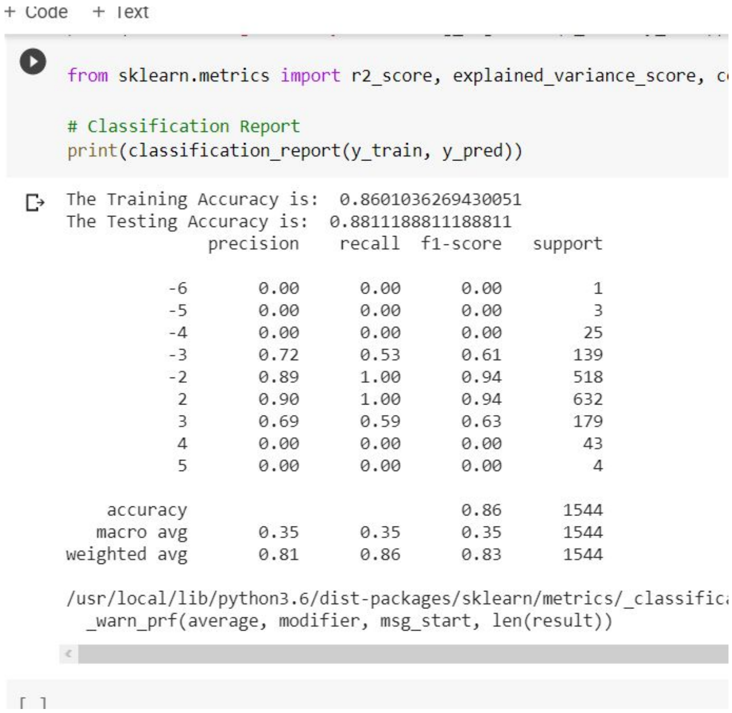
Table 2 The clustering performance comparison using different type of omics data

From: Deep learning-based ovarian cancer subtypes identification using multi-omics data

	Features	silhouette score	DBI
mRNA	20,502	0.550	0.607
miRNA	1870	0.536	0.644
CNV	23,606	0.509	0.713
Multi-omics	45,978	0.583	0.562

In Table 2 we give the clustering performance comparison using different type of omics data. The results indicated that when using single type of omics data, the mRNA performed best with the silhouette score 0.550, and the CNV achieved worst performance with silhouette score value of 0.509. The miRNA performed better than CNV but worse than mRNA. It is obviously that clustering using multi-omics in our deep learning framework achieved 6% higher silhouette score and 7.41% lower DBI, compared with which obtained by using mRNA data.

Sample Output



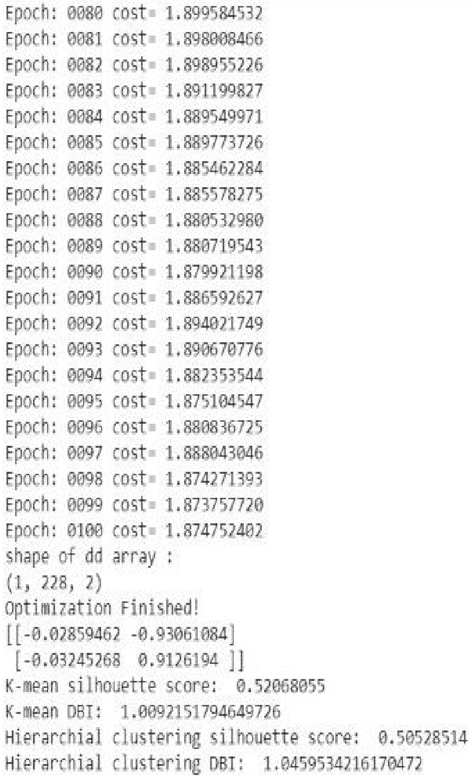
```
from sklearn.metrics import r2_score, explained_variance_score, c

# Classification Report
print(classification_report(y_train, y_pred))
```

```
The Training Accuracy is:  0.8601036269430051
The Testing Accuracy is:  0.8811188811188811
              precision    recall  f1-score   support

          -6       0.00      0.00      0.00         1
          -5       0.00      0.00      0.00         3
          -4       0.00      0.00      0.00        25
          -3       0.72      0.53      0.61       139
          -2       0.89      1.00      0.94       518
           2       0.90      1.00      0.94       632
           3       0.69      0.59      0.63       179
           4       0.00      0.00      0.00        43
           5       0.00      0.00      0.00         4

    accuracy                           0.86      1544
   macro avg       0.35      0.35      0.35      1544
weighted avg       0.81      0.86      0.83      1544

/usr/local/lib/python3.6/dist-packages/sklearn/metrics/_classific
  _warn_prf(average, modifier, msg_start, len(result))
```

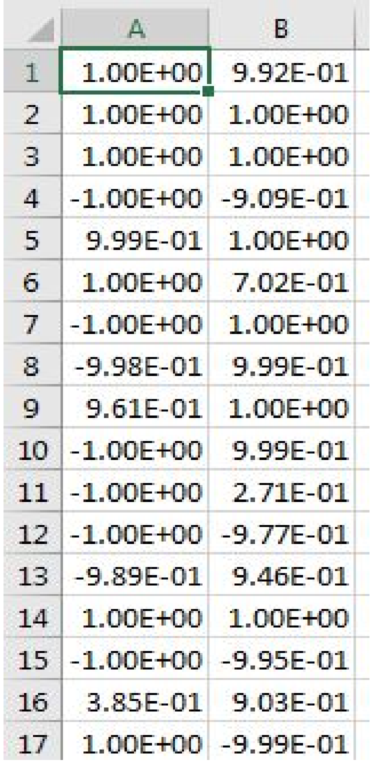
We had our dataset of patients, which we split for training and testing of our model. We randomly split our dataset and applied our Logistic Regression model on our training and testing dataset.

Further, we calculated the feature importance of various columns to select the best suited one which in turn would improvise our model.

Finally, we calculated the training and testing accuracy of our model.



```
Epoch: 0080 cost= 1.899584532
Epoch: 0081 cost= 1.898008466
Epoch: 0082 cost= 1.898955226
Epoch: 0083 cost= 1.891199827
Epoch: 0084 cost= 1.889549971
Epoch: 0085 cost= 1.889773726
Epoch: 0086 cost= 1.885462284
Epoch: 0087 cost= 1.885578275
Epoch: 0088 cost= 1.880532980
Epoch: 0089 cost= 1.880719543
Epoch: 0090 cost= 1.879921198
Epoch: 0091 cost= 1.886592627
Epoch: 0092 cost= 1.894021749
Epoch: 0093 cost= 1.890670776
Epoch: 0094 cost= 1.882353544
Epoch: 0095 cost= 1.875104547
Epoch: 0096 cost= 1.880836725
Epoch: 0097 cost= 1.888043046
Epoch: 0098 cost= 1.874271393
Epoch: 0099 cost= 1.873757720
Epoch: 0100 cost= 1.874752402
shape of dd array :
(1, 228, 2)
Optimization Finished!
[[-0.02859462 -0.93061084]
 [-0.03245268  0.9126194 ]]
K-mean silhouette score:  0.52068055
K-mean DBI:  1.0092151794649726
Hierarchial clustering silhouette score:  0.50528514
Hierarchial clustering DBI:  1.0459534216170472
```



	A	B
1	1.00E+00	9.92E-01
2	1.00E+00	1.00E+00
3	1.00E+00	1.00E+00
4	-1.00E+00	-9.09E-01
5	9.99E-01	1.00E+00
6	1.00E+00	7.02E-01
7	-1.00E+00	1.00E+00
8	-9.98E-01	9.99E-01
9	9.61E-01	1.00E+00
10	-1.00E+00	9.99E-01
11	-1.00E+00	2.71E-01
12	-1.00E+00	-9.77E-01
13	-9.89E-01	9.46E-01
14	1.00E+00	1.00E+00
15	-1.00E+00	-9.95E-01
16	3.85E-01	9.03E-01
17	1.00E+00	-9.99E-01

Different cost values for different epochs helped us to run our model making various assumptions and then select the assumption corresponding to best cost value among the given epochs.

We also calculated silhouette values for some clustering methods. DAE-K means it has values. So, we used DAE-K which means clustering in our model.

Code walk-through

This is the python code for denoising autoencoders (DAE) and the k-means using the reconstructed features. The silhouette scores and Davies Bouldin scores (DBI) were used to evaluate the clustering performances.

Sources:

- https://jovian.ai/garimasingh128/deep-learning-based-ovarian-cancer-subtypes-identification-using-multi-omics-data
- https://colab.research.google.com/drive/1_ruztGS5i1h9ugO28h_foaK_0Sczsq_H?usp=sharing
- https://colab.research.google.com/drive/1dj2WfGTwTJYjATTJXxSUNyGZuRD2FW5O?usp=sharing

Conclusion

- We designed a novel deep learning-based framework for ovarian cancer subtype identification, and a logistic regression method was used to build the light-weighted classification model.
- Compared to identifying subtypes using single omics data, the multi-omics data analysis can utilize more information. Hence, we proposed a model which in turn would help to robustly identify ovarian cancer subtypes.
- Ovarian cancer ranks 5th in cancer death among women. It has a high mortality rate. Also the risk of getting ovarian cancer is quite high. So, identifying molecular subtypes of ovarian cancer is important.

It is important to know more about the ovarian cancer heterogeneity between different patients for choosing different treatment programs and predicting clinical outcomes. In this study we proposed a novel deep learning framework for integrating multi-omics data with denoising autoencoders for identifying the ovarian cancer subtypes. Two subtypes from the molecular level were identified in ovarian cancer, and the results show our proposed method is competitive and reliable. The method comparison results indicated our method out-performed than the traditional and deep learning-based methods. More importantly, the classification model was proved by three independent test datasets collected from GEO. All the *p*-values less than 0.05 show that the differences between the classified cancer subgroups are significant.

By combining the results in DEG and WGCNA analysis, we selected 34 target genes related to ovarian cancer. And using these 34 identified genes, 19 KEGG pathways were enriched including PI3K-Akt signaling pathway and human papillomavirus infection pathway. The literature review shows 19 (56%) biomarkers and 8(42.1%) KEGG pathways identified based on the classification subtypes have been proved to be associated with ovarian cancer.

References

- https://biodatamining.biomedcentral.com/articles/10.1186/s13040-020-00222-x
- https://linkinghub.elsevier.com/retrieve/pii/S0090825810002623
- https://clincancerres.aacrjournals.org/content/14/16/5198

Availability of data and materials

All the data analyzed during the current study are available in the TCGA and GEO datasets.

-------------------End of Report-------------------

THANK YOU