iSOM-GSN: AN INTEGRATIVE APPROACH FOR TRANSFORMING MULTI-OMIC DATA INTO GENE SIMILARITY NETWORKS VIA SELF-ORGANIZING MAPS

Anonymous authors
Paper under double-blind review

ABSTRACT

One of the main challenges in applying graph convolutional neural networks on gene-interaction data is the lack of understanding of the vector space to which they belong and also the inherent difficulties involved in representing those interactions on a significantly lower dimension, viz Euclidean spaces. The challenge becomes more prevalent when dealing with various types of heterogeneous data. We introduce a systematic, generalized method, called iSOM-GSN, used to transform "multi-omic" data with higher dimensions onto a two-dimensional grid. Afterwards, we apply a convolutional neural network to predict disease states of various types. Based on the idea of Kohonen's self-organizing map, we generate a two-dimensional grid for each sample for a given set of genes that represent a gene similarity network. We have tested the model to predict breast and prostate cancer using gene expression, DNA methylation and copy number alteration, yielding prediction accuracies in the 94-98% range for tumor stages of breast cancer and calculated Gleason scores of prostate cancer with just 11 input genes for both cases. The scheme not only outputs nearly perfect classification accuracy, but also provides an enhanced scheme for representation learning, visualization, dimensionality reduction, and interpretation of the results.

1 INTRODUCTION

Large scale projects such as "The Cancer Genome Atlas" (TCGA) generate a plethora of multi-dimensional data by applying high-resolution microarrays and next generation sequencing. This leads to diverse multi-dimensional data in which the need for devising dimensionality reduction and representation learning methods to integrate and analyze such data arises. An earlier study by Shen et al. proposed algorithms iCluster (Shen et al., 2009a) and iCluster+ (Shen et al., 2009b), which made use of the latent variable model and principal component analysis (PCA) on multi-omic data and aimed to cluster cancer data into sub-types; even though it performed well, it did not use multi-omics data. In another study, (Lyu and Haque, 2018) attempted to apply heatmaps as a dimensionality reduction scheme on gene expression data to deduce biological insights and then classify cancer types from a Pan-cancer cohort. However, the accuracy obtained by using that method was limited to 97% on Pan-cancer data, lacking the benefits of integrated multi-omics data.

In a recent study (Choy et al., 2019) used self-Organizing maps (SOMs) to embed gene expression data into a lower dimensional map, while the works of (Bustamam et al., 2018; Mallick et al., 2019; Paul and Shill, 2018; Loeffler-Wirth et al., 2019) generate clusters using SOMs on gene expression data with different aims. In addition, the work of (Hopp et al., 2018) combines gene expression and DNA methylation to identify subtypes of cancer similar to those of (Roy et al., 2018), which identifies modules of co-expressing genes. On the other hand, the work of (Kartal et al., 2018) uses SOMs to create a generalized regression neural network, while the model proposed in (Yoshioka and Dozono, 2018; Shah and Luo, 2017) uses SOMs to classify documents based on a word-to-vector model. Apart from dimensionality reduction methods, attempts have been made by applying supervised deep machine learning, such as deepDriver (Luo et al., 2019), which predicts candidate driver genes based on mutation-based features and gene similarity networks. Although these works have been devised to use embedding and conventional machine learning approaches, the use deep neural networks on multi-omics data integration is still in its infancy. In addition, these methods lack

Gleason Score	Number of Samples	Group
3+4	147	34
4+3	101	43
4+5,5+4	139	9

Table 1: Distribution of the different Gleason groups considered for PRCA.

in adequacy to generalize them multi-omics data to predic disease states. More specifically, none of these models combine the strength of SOMs for representation learning combined with the CNN for image classification as we do in this work.

In this paper, a deep learning-based method is proposed, and is used to predict disease states by integrating multi-omic data. The method, which we call iSOM-GSN, leverages the power of SOMs to transform multi-omic data into a gene similarity network (GSN) by the use of gene expression data. Such data is then combined with other genomic features to improve prediction accuracy and help visualization. To our knowledge, this the first deep learning model that uses SOMs to transform multi-omic data into a GSN for representation learning, and uses CNNs for classification of disease states or other clinical features. The main contributions of this work can be summarized as follows:

- A deep learning method for prediction of tumor aggressiveness and progression using iSOM-GSN.
- A new strategy to derive gene similarity networks via self-organizing maps.
- Use of iSOM-GSN to identify relevant biomarkers without handcrafted feature engineering.
- An enhanced scheme to interpret and visualize multi-dimensional, multi-omics data.
- An efficient model for graph representation learning.

2 Materials and Methods

2.1 Datasets

We considered two datasets as part of our study: The Cancer Genome Atlas (TCGA) Prostate Adenocarcinoma (PRCA) (National Cancer Institute, 2013) and The Cancer Genome Atlas (TCGA) Breast Invasive Carcinoma (BRCA) (National Cancer Institute, 2015). Our aim here is to classify patients based on Gleason scores for PRCA (Hamzeh et al., 2017), and tumor stage for BRCA (Firoozbakht et al., 2017). The total number of samples for PRCA and BRCA were 499 and 570 respectively. Both datasets had approximately 60,000 features for gene expression data alone. Thus, a variance threshold of 0.2% was applied to these data, which removes all features that have at least 80% zero values; this step reduced the feature set size to 16,000.

The data were then normalized on a common scale for all omics, including DNA methylation and CNA data. The gene names were preserved in HUGO format and the names considered irrelevant by HUGO were removed. All the three types of data were then combined, based on patient ID, which yielded data for 387 and 392 patients for PRCA and BRCA respectively, containing all three required omic data.

Since imbalance was observed across all classes in the PRCA dataset, we considered only three distinct Gleason scores. It is worthwhile to note that samples with Gleason score 7 were considered as two different classes, i.e., 3+4 and 4+3, for example, since these two groups are clinically different. More details on class distribution are shown in Tables 1 and 2.

MultisgCV was used to further process the data (Lawrence et al., 2013). The MutsigCV algorithm identifies significantly mutated genes by building a patient-specific mutation model based on gene expression and DNA methylation data. This method takes the whole genome or exome sequence as input and identifies genes that are mutated more often. The top 14 mutated genes from Multisig were considered for the rest of the experiment.

Tumor Stage	Number of Samples
2A	179
2B	129
3A	84

Table 2: Distribution of the different tumor groups considered for BRCA.

2.2 PROPOSED METHOD

We consider the problem of integrating multiple types of omics data. For this purpose, we propose a three-step approach, which we call iSOM-GSN, and whose main steps are depicted in Figure 1. First, we create a GSN by extracting features from one data type, in our case, gene expression data. Then, for each sample, we integrate all data types by considering features extracted from the first step. Finally, we apply a CNN to perform classification with training and test split at 70:30 ratio to test the model.

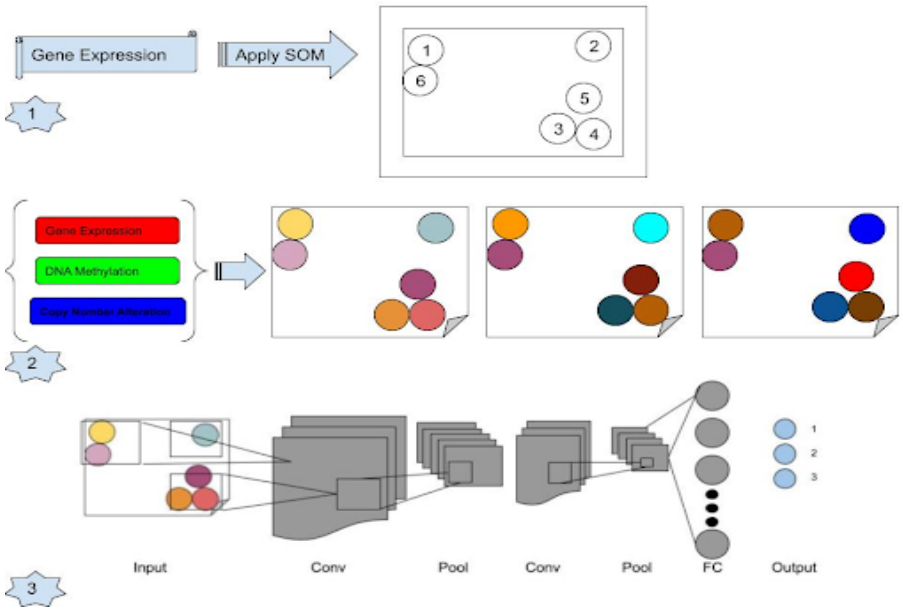

Figure 1: Block diagram of the main components of iSOM-GSN.

We assume the input data is a set of matrices $S = \{s_{oj}^{(i)}\}$, where $i = \{1, 2, 3, \ldots, n\}$ represents the samples, $j = \{1, 2, 3, \ldots, m\}$ represents the genes, and $o = \{1, 2, 3, \ldots, p\}$ represents types of data (omics). Here, n is the number of samples, m is the number of genes, and p is the number of types of omics.

2.3 GENE SIMILARITY NETWORK

The first step consists of creating a gene similarity network (GSN) by applying a self-organizing map learning algorithm. In this step, we consider only one type of data, i.e., gene expression. Let $S_1 = \{s_{1j}^{(i)}\}_{i,j=1}^{n,m}$ denote one omic data where $j = \{1, 2, 3, \ldots, m\}$ represents the set of genes and $i = \{1, 2, 3, \ldots, n\}$ represents the set of samples. S_1 is the input to the SOM.

A SOM is a lower-dimensional representation of complex, higher-dimensional data in such a way that distances among vectors in the original space are preserved in the new representation. A SOM is learned via an unsupervised clustering algorithm, which takes sample vectors as inputs, and groups them based on the similarities derived by the features. In our case, the input vectors to the SOM are the samples with gene expression values of all samples as features. The following are the main steps that are followed to construct a SOM.

1. Initialize m neurons with random weights assigned to each neuron c_k, where $k = 1, 2, \ldots, m$, where m is the number of genes under consideration, in our case 14.
2. Calculate the Euclidean distance between each gene g_j and its neuron c_k, and identify the winning neuron, i.e., the neuron that has the smallest distance to its respective neuron. The Euclidean distance is calculated as follows:

$$d_j = \sqrt{\sum_{i=0}^{i=n} (s_{1j}^{(i)} - c_{ji})^2} \tag{1}$$

   where $s_{1j}^{(1)} = g_j$ represents the gene vector for i^{th} sample and c_{j1} represents neuron vector.
3. Suppose that c_k is the winning neuron, i.e., it is the closest to gene g_j. Then, we update the weight of c_k using Equation (2). The winning neuron is also known as best matching unit (BMU).
4. Update the weights of the neurons that are in proximity to the BMU, c_k. To account for this, we use a neighbourhood function that is defined by Equation (3).
5. Repeat steps 2 - 4 for e iterations or until desired convergence (i.e., the weights remain unchanged or the change is less than a threshold).
6. Finally, obtain c_m neurons, which represent g_m genes in the two-dimensional space, represented by Equation (5).

$$c_k(t+1) = c_k(t) + \theta_j(t) L(t) (s_{1j}(t) - c_k(t)) \,, \tag{2}$$

   where $L(t)$ is the learning rate regulation function defined in Equation (4).

$$\Theta(t) = exp\left(\frac{d_j^2}{2\sigma^2(t)}\right) \quad t = 1, 2, \ldots, e \tag{3}$$

$$L(t) = L_0 exp\left(\frac{-t}{\lambda}\right) \quad t = 1, 2, \ldots, e \tag{4}$$

   where L_0 is initial learning rate.

$$X = (x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m) \,, \tag{5}$$

   where (x_j, y_j) represents the coordinates of g_j.

As a result of running the training algorithm, a SOM is obtained in which the genes are organized based on their similarity, representing a GSN. This network is represented as a two-dimensional lattice whose coordinates are denoted as in Equation (5). Figure 2 shows the two SOMs derived from the two datasets, BRCA and PRCA. Observing the evolution of the SOM learning algorithms through the different epochs for both datasets (see Figures 10-13 of the Supplementary Material), show how complex, high-dimensional relationships among related genes are revealed and visualized in a simple way on a two-dimensional map.

2.4 Integrating Multiple Data Types

The second step of iSOM-GSN is to integrate multiple data types. We use the GSN generated in the first step as a template image; in the example, the genes are indexed with numbers by followng the mapping listed in Table 3. We then expand a circular region around the points with a predefined radius and color the circles as shown in Figure 3. We color each circle by considering each data view by using the RGB color scheme, where Red is represented by gene expression, Green by DNA methylation and Blue by CNA.

In our case, $S_{1j}^{(i)}$ represents gene expression, $S_{2j}^{(i)}$ DNA methylation, and $S_{3j}^{(i)}$ copy number alteration (CNA).

For each sample $s^{(i)}$, gene g_j is colored as in the RGB palette, by considering the rule of Equation (6):

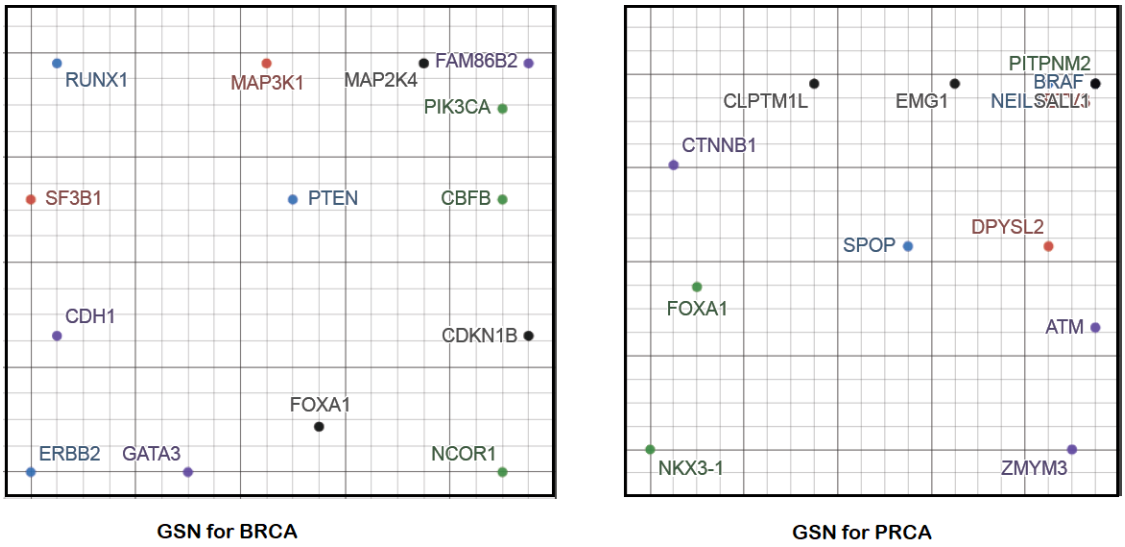

Figure 2: Comparison of PRCA and BRCA genes.

Index	Gene Name BRCA	Gene Name PRCA
0	RUNX1	SPOP
1	PIK3CA	FOXA1
2	GATA3	CTNNB1
3	FOXA1	CLPTM1L
4	SF3B1	DPYSL2
5	PTEN	NEIL1
6	CBFB	PITPNM2
7	CDH1	ATM
8	MAP2K4	EMG1
9	MAP3K1	ETV3
10	ERBB2	BRAF
11	NCOR1	NKX3-1
12	FAM86B2	ZMYM3
13	CDKN1B	SALL1

Table 3: Indices of gene names for the BRCA and PRCA datasets.

$$x_{pq} = \begin{cases} RGB_j^{(i)} & \text{If point } (p,q) \text{ is within certain radius of } g_j, \\ 0 & \text{otherwise} \quad . \end{cases} \tag{6}$$

where $R_j^{(i)} = S_{1j}^{(i)}, \quad G_j^{(i)} = S_{1j}^{(i)} and \quad B_j^{(i)} = S_{1j}^{(i)}$

As a result, we obtain a set of matrices, one per each sample, defined as follows:

$$X^{(i)} = \{x_{pq}^{(i)}\}\,. \tag{7}$$

Figure 3 represents a sample image created after integrating multiple omics for the BRCA dataset. As can be observed, various shades of colors for different genes represent their values with respect to the three different types of omic data.

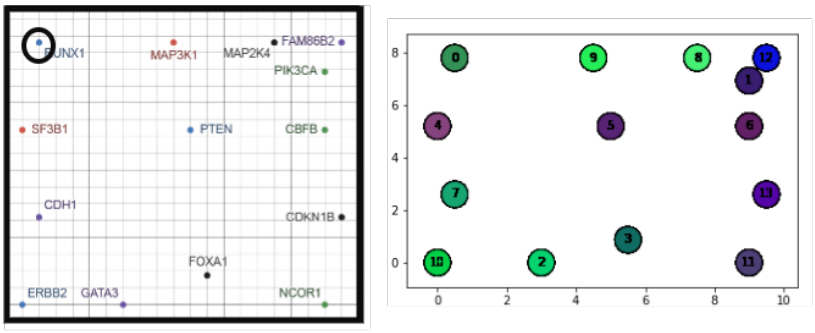

Figure 3: Sample image after integrating multiple data views for BRCA for 14 genes.

2.5 Convolutional Neural Network

The last step of iSOM-GSN is to feed the images generated in the previous step to the CNN, to predict the state of the disease as the final output. The architecture of the CNN is proven to be the most effective in learning visual representations. The CNN is also known to perform better than the human eye in many visual processing problems. The usage of the CNN in any method is just a variation in how the convolution and pooling layers are combined, and how the network is trained.

A more detailed, schematic diagram of the entire network design is depicted in Figure 7 of the Supplementary Material. The network includes two convolutional layers and two fully-connected layers with a small number of neurons. Our choice of a smaller network design is motivated both from our desire to reduce the risk of over-fitting as well as to simplify the nature of the classification. All three color channels, i.e., RGB, are processed directly by the network. The subsequent convolutional and fully connected layers are then defined as follows:

- 32 filters of size 3×3 pixels are applied to the input in the first convolutional layer, followed by a rectified linear operator (ReLU), a Max-pooling layer taking the maximal value of 2×2 regions with two-pixel strides and a local response normalization layer.
- The output of the previous layer is then processed by the second convolutional layer, containing 32 filters of size 3×3 pixels. Again, this is followed by ReLU, a Max-pooling layer and a local response normalization layer with the same hyper-parameters as before.
- First fully connected layer that receives the output of the second convolutional layer and contains 128 neurons, followed by ReLU and a dropout layer.
- Second fully connected layer that receives the output of the first fully connected layer and output three neurons, followed by ReLU and a dropout layer.

Finally, the output of the last fully-connected layer is fed to a Soft-max layer that assigns a probability to each class. The prediction itself is performed by applying Soft-max to choose the class with maximal probability for the given test image.

Aside from using a lean network architecture, i.e., fewer layers, we apply two additional methods to further limit the risk of over-fitting. First, we apply dropout learning, i.e., randomly setting the output value of the network neurons to zero. The network includes three dropout layers with a dropout ratio of 0.5 (50% chance of setting a neuron's output value to zero). Second, we use data augmentation by taking a random input image, and scaling and mirroring it in each forward-backward training pass. Training is done using the Adam optimizer (Kingma and Ba, 2014).

3 Experimental Results

To assess the performance of iSOM-GSN, the data was divided into training and test datasets with a ratio 70:30. Minmax scaling was then applied on the test dataset followed by ranging the training dataset accordingly. Note that the test data is scaled using the same criterion applied to the training

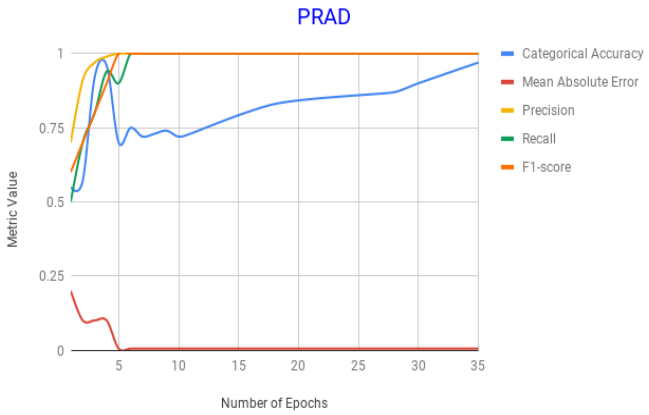

Figure 4: Evaluation of performance of the classifier per epoch.

dataset, and without any information about the classes. We then calculated the main performance measures that include categorical accuracy, precision, recall, F1-score and mean absolute error.

3.1 Results

iSOM-GSN has been run on two multi-omic datasets namely PRCA (National Cancer Institute, 2013) and BRCA (National Cancer Institute, 2015) using the model and parameters described earlier in this paper. Figure 4 depicts the plot of how various performance measures convolve with an increasing number of epochs. In general, it can be seen that the predictive performance is almost perfect with respect to various parameters. However, regarding the number of genes obtained after filtering, both datasets used the exact number of genes retained for effective classification.

In addition, Figure 6 of the Supplementary Material depicts the plot of the receiver operating characteristic (ROC), area under curve (AUC). In general terms, it can be seen that the predictive performance is in the range 94-98% with respect to various parameters. However, regarding the number of genes obtained after filtering, both datasets used the same number of genes retained for effective classification. This shows that only 14 genes are enough to classify any clinical variable using the proposed model, and those genes are significantly mutated.

3.2 Biological Validation

We illustrate the ability to discover and visualize patterns of genomic interactions in biological comprehensive context for classification. When the goal is to identify potential biomarkers and factors that characterize biological and clinical aspects, the proposed approach comes to the rescue. As part of a feature selection step, we narrowed down the relevant features to just 14 genes, which are sufficient for classification and achieve 96% accuracy. These genes are listed in Table 5. All genes are identified either as tumor suppressor genes or oncogenes for known pathways as shown in Table 5. TP53 and PTEN are the common genes that are highly mutated in both PRCA and BRCA datasets. These gene are tumor suppressor genes, which are well known to express high gene expression values and are considered as biomarkers for cancer in general (https://www.genecards.org/, 2019). On the other hand, SALL1 and PITPNM2 are genes that are not known as cancer related genes.

While validating the genes related to BRCA data set, we have found in that genes MAP3K1 and MAP2K4 are strong predictors of MEK inhibitors which are frequently found in breast, prostate and colon cancers. These can be potential targets for drugs (Xue et al., 2018). In another study, we have found co-occurring mutations of PIK3CA and MAP3K1 are functionally significant in breast cancer and MAP3K1 mutational status may be considered as a predictive biomarker for efficacy in PI3K pathway inhibitor trials (Avivar-Valderas et al., 2018). We have also found that expression of genes GATA-3 and FOXA1 are sufficient to differentiate breast carcinoma from other and hence are excellent bio-markers (Davis et al., 2016). This is also supported by the findings reported in

PRCA	Pathways	BRCA	Pathways
SPOP	Signaling by Hedgehog and Hedgehog Pathway	RUNX1	Transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds and Embryonic and Induced Pluripotent Stem Cell Differentiation Pathways and Lineage-specific Markers.
TP53	Apoptosis Modulation and Signaling and Glioma.	PIK3CA	Glioma and Development Dopamine D2 receptor transactivation of EGFR
FOXA1	Embryonic and Induced Pluripotent Stem Cell Differentiation Pathways and Lineage-specific Markers	TP53	Apoptosis Modulation and Signaling and Glioma.
CTNNB1	Beta-Adrenergic Signaling and Blood-Brain Barrier Pathway: Anatomy	SF3B1	Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3 and mRNA Splicing - Major Pathway
MED12	Gene Expression and RNA Polymerase II Transcription Initiation And Promoter Clearance	PTEN	Glioma and Metabolism of proteins
PITPNM2	Glycerophospholipid biosynthesis and Metabolism	CBFB	Regulation of nuclear SMAD2/3 signaling and ATF-2 transcription factor network
PTEN	Glioma and Metabolism of proteins	CDH1	Arf6 trafficking events and Integrated Breast Cancer Pathway
ATM	Apoptotic Pathways in Synovial Fibroblasts and Integrated Cancer Pathway	MAP2K4	Apoptosis Modulation and Signaling and Tacrolimus/Cyclosporine Pathway, Pharmacodynamics
NKX3-1	Endometrial cancer and Pathways in cancer	MAP3K1	Apoptosis Modulation and Signaling and Tacrolimus/Cyclosporine Pathway, Pharmacodynamics
ZMYM3	Diseases associated with ZMYM3 include Dystonia 3, Torsion, X-Linked and Myasthenic Syndrome, Congenital, 6, Presynaptic.	NCOR1	Signaling by NOTCH1 and Transcriptional activity of SMAD2/SMAD3-SMAD4 heterotrimer
SALL1	Transcriptional regulation of pluripotent stem cells and Developmental Biology	CDKN1B	CDK-mediated phosphorylation and removal of Cdc6 and PI3K-AKT-mTOR signaling pathway and therapeutic opportunities

Figure 5: Most relevant genes found to predict Gleason groups of PRCA and stages of BRCA.

(Hisamatsu et al., 2015), which claim that these two genes are associated with a less aggressive phenotype and they give better prognosis in patients with HR-positive or HER2-negative breast cancer. Thus, we confirm that the GSN's formed are helpful to find potential novel bio-markers.

3.3 Comparison with Other Approaches

A practical challenge for validating this framework is the unavailability of independent datasets with all data types. An intrinsic question arises as to what extent a single data type (e.g., gene expression) is effective in classifying, based on our framework. The closest method to compare our work with is the one that uses gene expression and DNA methylation (Hopp et al., 2018), though they use the two types of data separately, and apply it to a single disease. To that end, we ran iSOM-GSN on a single omic data at a time. We discovered that gene expression data alone yielded 90% accuracy, whereas DNA methylation and CNA, alone, yielded 87% and 89% classification accuracy, respectively. This demonstrates the advantages of combining multi-omics, the strengths of the SOMs for representation learning combined with the power of deep CNNs for image-based data classification.

4 Conclusions

This paper presents a framework that uses a self-organizing map and a convolutional neural network used to conduct data integration, representation learning, dimensionality reduction, feature selection and classification simultaneously to harness the full potential of integrated high-dimensional large scale cancer genomic data. We have introduced a new way to create gene similarity networks, which can lead to novel gene interactions. We have also provided a scheme to visualize high-dimensional, multi-omics data onto a two-dimensional grid. In addition, we have devised an approach that could also be used to integrate other types of multi-omic data and predict any clinical aspects or states of diseases, such as laterality of the tumor, survivability, or cancer sub types, just to mention a few.

This work can also be extended to classify Pan-cancer data. Omics can be considered as a vector and more than three types of data (i.e., beyond RGB images) can be incorporated for classification. Apart from integrating multi-omics data, the proposed approach can be considered as an unsupervised clustering algorithm, because of the competitive learning nature of SOMs. We can also apply iSOM-GSN on other domains, such as predicting music genre's for users based on their music preference. As a first step, we have applied the SOM to a Deezer dataset and the results are encouraging 14. Applications of iSOM-GSN can also be in drug response or re-purposing, prediction of passenger or oncogenes, revealing topics in citation networks, and other prediction tasks.

REFERENCES

Avivar-Valderas, A., McEwen, R., Taheri-Ghahfarokhi, A., Carnevalli, L. S., Hardaker, E. L., Maresca, M., Hudson, K., Harrington, E. A., and Cruzalegui, F. (2018). Functional significance of co-occurring mutations in PIK3CA and MAP3K1 in breast cancer. *Oncotarget*, 9(30):21444–21458. PMC5940413[pmcid].

Bustamam, A., Rivai, M. A., and Siswantining, T. (2018). Implementation of spectral clustering on microarray data of carcinoma using self organizing map (SOM). *AIP Conference Proceedings*, 2023(1):020240.

Choy, C. T., Wong, C. H., and Chan, S. L. (2019). Embedding of genes using cancer gene expression data: Biological relevance and potential application on biomarker discovery. *Frontiers in Genetics*, 9:682.

Davis, D. G., Siddiqui, M. T., Oprea-Ilies, G., Stevens, K., Osunkoya, A. O., Cohen, C., and Li, X. B. (2016). GATA-3 and FOXA1 expression is useful to differentiate breast carcinoma from other carcinomas. *Human Pathology*, 47(1):26 – 31.

Firoozbakht, F., Rezaeian, I., D'agnillo, M., Porter, L., Rueda, L., and Ngom, A. (2017). An integrative approach for identifying network biomarkers of breast cancer subtypes using genomic, interactomic, and transcriptomic data. *Journal of Computational Biology*, 24(8):756–766. PMID: 28650678.

Hamzeh, O., Alkhateeb, A., Rezaeian, I., Karkar, A., and Rueda, L. (2017). Finding transcripts associated with prostate cancer Gleason stages using next generation sequencing and machine learning techniques. In Rojas, I. and Ortuño, F., editors, *Bioinformatics and Biomedical Engineering*, pages 337–348, Cham. Springer International Publishing.

Hisamatsu, Y., Tokunaga, E., Yamashita, N., Akiyoshi, S., Okada, S., Nakashima, Y., Taketani, K., Aishima, S., Oda, Y., Morita, M., and Maehara, Y. (2015). Impact of GATA-3 and FOXA1 expression in patients with hormone receptor-positive/HER2-negative breast cancer. *Breast Cancer*, 22(5):520–528.

Hopp, L., Löffler-Wirth, H., Galle, J., and Binder, H. (2018). Combined SOM-portrayal of gene expression and DNA methylation landscapes disentangles modes of epigenetic regulation in glioblastoma. *Epigenomics*, 10(6):745–764. PMID: 29888966.

https://www.genecards.org/ (2019). Gene cards. *Human Gene Database*.

Kartal, S., Oral, M., and Ozyildirim, B. M. (2018). Pattern layer reduction for a generalized regression neural network by using a self–organizing map. *International Journal of Applied Mathematics and Computer Science*, 28(2):411 – 424.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Lawrence, M. S., Stojanov, P., and Polak, P. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499:214 EP –.

Loeffler-Wirth, H., Kreuz, M., Hopp, L., Arakelyan, A., Haake, A., Cogliatti, S. B., Feller, A. C., Hansmann, M.-L., Lenze, D., Möller, P., Müller-Hermelink, H. K., Fortenbacher, E., Willscher, E., Ott, G., Rosenwald, A., Pott, C., Schwaenen, C., Trautmann, H., Wessendorf, S., Stein, H., Szczepanowski, M., Trümper, L., Hummel, M., Klapper, W., Siebert, R., Loeffler, M., and Binder, H. (2019). A modular transcriptome map of mature B cell lymphomas. *Genome Medicine*, 11(1):27.

Luo, P., Ding, Y., Lei, X., and Wu, F.-X. (2019). deepdriver: Predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. *Frontiers in Genetics*, 10:13.

Lyu, B. and Haque, A. (2018). Deep learning based tumor type classification using gene expression data. *bioRxiv*.

Mallick, P., Ghosh, O., Seth, P., and Ghosh, A. (2019). Kohonen's self-organizing map optimizing prediction of gene dependency for cancer mediating biomarkers. In Abraham, A., Dutta, P., Mandal, J. K., Bhattacharya, A., and Dutta, S., editors, *Emerging Technologies in Data Mining and Information Security*, pages 863–870, Singapore. Springer Singapore.

National Cancer Institute (2013). TGCA. *cBioPortal for Cancer Genomics*. Dataset is available at: https://www.cbioportal.org/study/summary?id=prad_tcga.

National Cancer Institute (2015). TGCA. *cBioPortal for Cancer Genomics*. Dataset is available at: https://www.cbioportal.org/study/summary?id=brca_tcga_pub2015.

Paul, A. K. and Shill, P. C. (2018). Incorporating gene ontology into fuzzy relational clustering of microarray gene expression data. *Biosystems*, 163:1 – 10.

Roy, S., Manners, H. N., Jha, M., Guzzi, P. H., and Kalita, J. K. (2018). *Soft Computing Approaches to Extract Biologically Significant Gene Network Modules*, pages 23–37. Springer Singapore, Singapore.

Shah, S. and Luo, X. (2017). Exploring diseases based biomedical document clustering and visualization using self-organizing maps. In *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–6.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009a). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics (Oxford, England)*, 25(22):2906–2912. 19759197[pmid].

Shen, R., Olshen, A. B., and Ladanyi, M. (2009b). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912.

Xue, Z., Vis, D. J., Bruna, A., Sustic, T., van Wageningen, S., Batra, A. S., Rueda, O. M., Bosdriesz, E., Caldas, C., Wessels, L. F. A., and Bernards, R. (2018). MAP3K1 and MAP2K4 mutations are associated with sensitivity to MEK inhibitors in multiple cancer models. *Cell Research*, 28(7):719–729.

Yoshioka, K. and Dozono, H. (2018). The classification of the documents based on Word2Vec and 2-layer self organizing maps. *International Journal of Machine Learning and Computing*, 8:252–255.

SUPPLEMENTARY MATERIAL

SOURCE CODE

The source code has been posted on a Github project, available at the following anonymous website: https://gitlab.com/NF2610/isom_gsn.

ADDITIONAL FIGURES AND TABLES

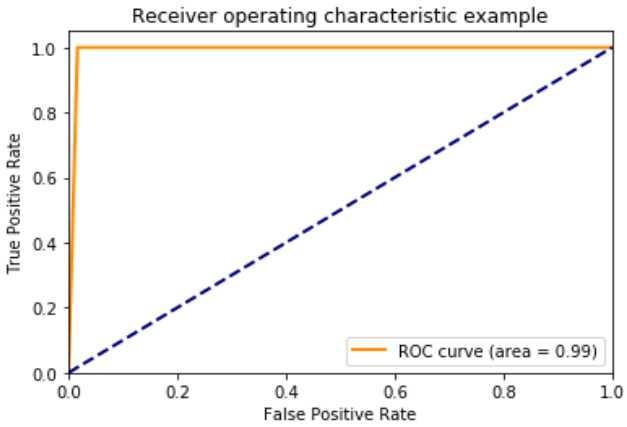

Figure 6: ROC plots for the proposed model run on the PRCA dataset.

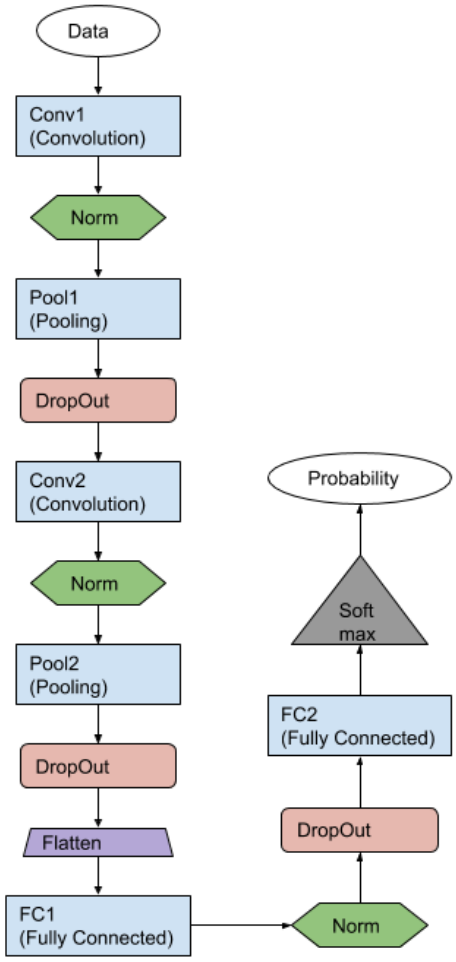

Figure 7: Schematic diagram of the convolutional neural network architecture.

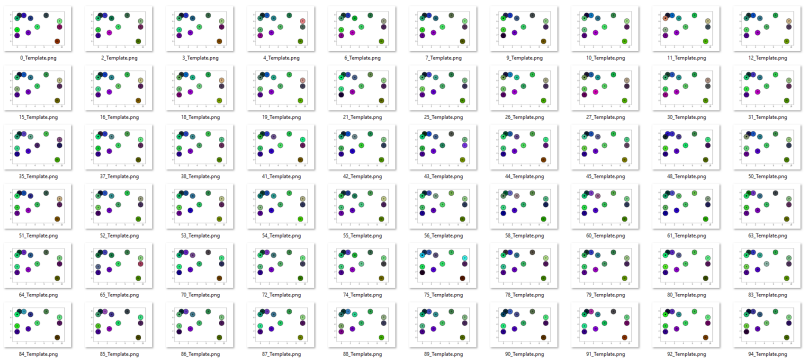

Figure 8: Sample SOMs obtained from the training dataset.

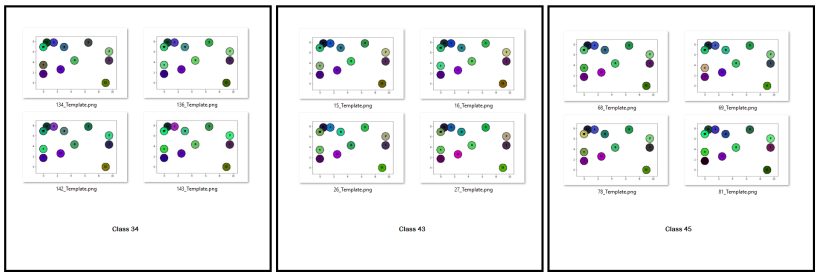

Figure 9: Sample SOMs obtained from the training dataset for each class.

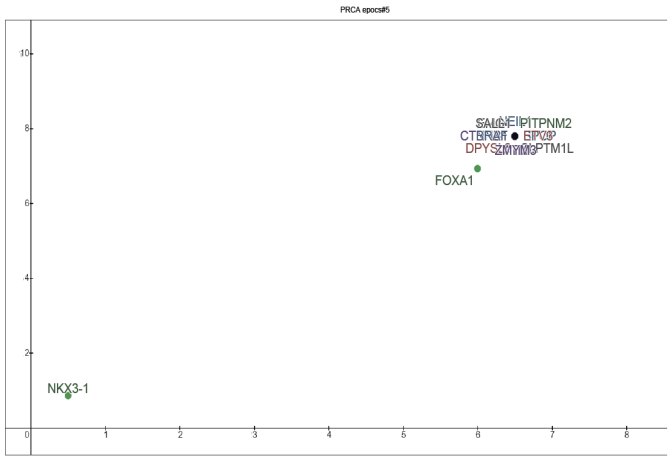

Figure 10: GSN after 5 Epochs for PRCA.

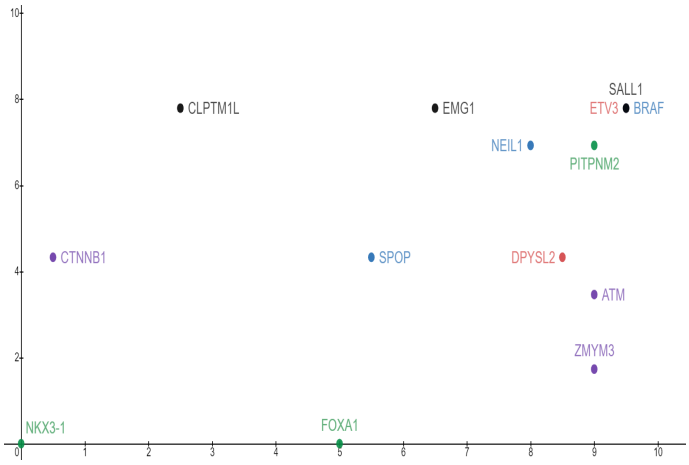

Figure 11: GSN after 1500 Epochs for PRCA.

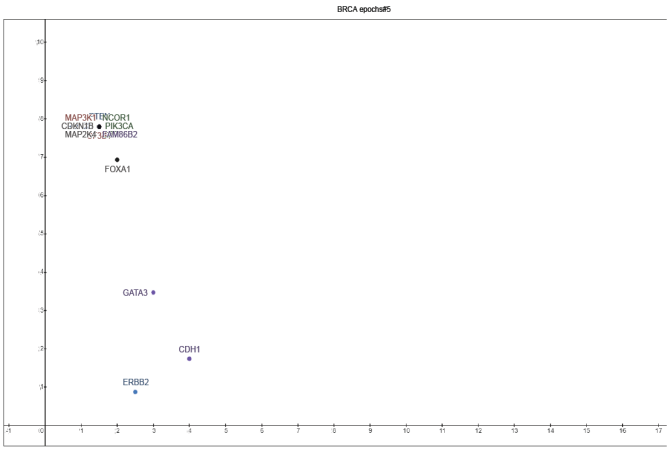

Figure 12: GSN after 5 Epochs for BRCA.

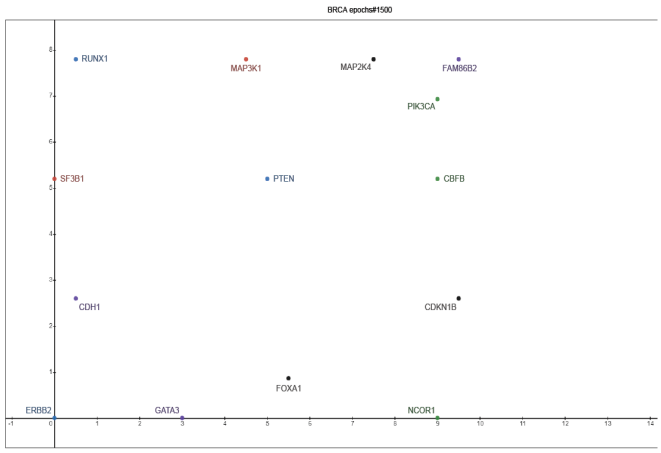

Figure 13: GSN after 1500 Epochs for BRCA.

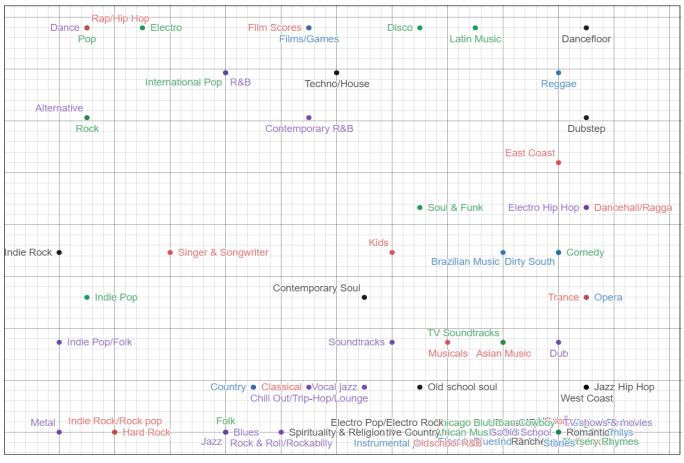

Figure 14: Sample organization of Genres after 1000 epochs