Communications and Network, 2024, 16, 135-148
https://www.scirp.org/journal/cn
ISSN Online: 1947-3826
ISSN Print: 1949-2421

A Convolutional Deep Neural Network Approach for miRNA Clustering

Ghada Ali Mohamed Shommo[1], Hadia Abbas Mohammed Elsied[2], Amira Kamil Ibrahim Hassan[1], Sara Elsir Mohamed Ahmed[3], Lamia Hassan Rahmatalla Mohamed[4], Wafa Faisal Mukhtar[1]

[1]Faculty of Computer Science & Information Technology, Sudan University of Science & Technology, Khartoum, Sudan
[2]Faculty of Business Studies, Department of MIS, Sudan University of Science & Technology, Khartoum, Sudan
[3]School of Management, Ahfad University for Women, Omdurman, Sudan
[4]Department of Management Information Systems, College of Business Administration, King Faisal University, Hofuf, Saudi Arabia
Email: ghada@sustech.edu, hadiaabbas@gmail.com, amirakamil32@yahoo.com, saraelsir1@yahoo.com, lhmohamed@kfu.edu.sa, wafaafaisal@sustech.edu

How to cite this paper: Shommo, G.A.M., Elsied, H.A.M., Hassan, A.K.I., Ahmed, S.E.M., Mohamed, L.H.R. and Mukhtar, W.F. (2024) A Convolutional Deep Neural Network Approach for miRNA Clustering. *Communications and Network*, **16**, 135-148.
https://doi.org/10.4236/cn.2024.164007

Received: July 24, 2024
Accepted: October 21, 2024
Published: October 24, 2024

Abstract

The regulatory role of the Micro-RNAs (miRNAs) in the messenger RNAs (mRNAs) gene expression is well understood by the biologists since some decades, even though the delving into specific aspects is in progress. Clustering is a cornerstone in bioinformatics research, offering a potent computational tool for analyzing diverse types of data encountered in genomics and related fields. MiRNA clustering plays a pivotal role in deciphering the intricate regulatory roles of miRNAs in biological systems. It uncovers novel biomarkers for disease diagnosis and prognosis and advances our understanding of gene regulatory networks and pathways implicated in health and disease, as well as drug discovery. Namely, we have implemented clustering procedure to find interrelations among miRNAs within clusters, and their relations to diseases. Deep clustering (DC) algorithms signify a departure from traditional clustering methods towards more sophisticated techniques, that can uncover intricate patterns and relationships within gene expression data. Deep learning (DL) models have shown remarkable success in various domains, and their application in genomics, especially for tasks like clustering, holding immense promise. The deep convolutional clustering procedure used is different from other traditional methods, demonstrating unbiased clustering results. In the paper, we implement the procedure on a Multiple Myeloma miRNA dataset publicly available on GEO platform, as a template of a cancer instance analysis, and hazard some biological issues.

Keywords

miRNA, Deep Clustering, DeepTrust, Convolutional Neural Network, Recurrence Plot

1. Introduction

Genes are expressed in different sizes and directions during cellular processes, and each gene's expression level is crucial for proper cell functioning [1]. Measuring gene expression levels is a powerful tool for understanding cell structure, function, and biological dynamics. Gene arrays are also used to simultaneously capture messenger RNA (miRNA) expression levels of thousands of genes. Gene arrays provide snapshots of gene expression patterns in a cell, and temporal changes in expression levels, represented by gene expression samples, provide valuable information about the dynamics of biological systems [2].

Using gene expression data for analysis presents several data privacy and security challenges. Gene expression data can be highly sensitive, because it contains information about an individual's genetic makeup. There are various ethical and legal frameworks governing the use of genetic data.

MicroRNAs (miRNAs) are small, non-coding RNA (genes) molecules that are crucial in post-transcriptional gene regulation. They involve various biological processes, including development, differentiation, and disease progression.

A critical aspect of miRNA research is identifying and clustering miRNAs based on their sequence similarities, which can provide insights into their evolutionary relationships and functional associations.

miRNAs play specific role in gene regulatory networks, such as gene silencing and regulation, post-transcriptional modulation. Its involvement in complex regulatory networks can affect multiple gene expressions, and interact with transcription factors.

Influencing cell proliferation and survival, miRNAs can regulate cell cycle progression, apoptosis, and cellular stress responses.

miRNAs have shown research significance. Their stable presence in body fluids such as blood enabled them to serve as biomarkers for diagnosing various diseases, such as cancer, cardiovascular, and neurodegenerative disorders. They are used as inhibitors to control disease progression, and hence hold potential to develop novel strategy for disease therapy.

Traditional clustering methods, such as hierarchical clustering and K-means clustering, have been widely used. However, these methods often need help to capture the complex relationships and patterns within miRNA sequences, leading to suboptimal clustering results [3].

Static measurements may not capture the complete picture of cellular processes, so temporal structures in gene expression time series are widely studied to elucidate the dynamics of cellular responses to various stimuli, such as changes in temperature, immune responses and other cellular systems [1]. Convolutional neural networks (CNNs) have achieved great success in many exploratory and predictive vision tasks, including image classification, object detection, and face recognition. Convolutional neural networks have become essential in deep learning, especially in complex tasks, due to their ability to learn hierarchical features from raw input data automatically [1].

Combining convolutional neural networks (CNNs) with microRNAs clustering (miRNAs) involves using CNN architecture to analyze miRNA data. If the miRNA data includes sequences, CNNs can be used to analyze the sequences and explore the messages. Convolutional layers can learn patterns and features from miRNA sequences, capturing important information for downstream tasks [3]. By combining deep clustering with guided clustering techniques, you can leverage the strengths of both approaches to enhance the clustering results and uncover novel biological insights from integrated miRNA and mRNA data.

The main problem with ML algorithms, is that although they have proven their efficiency with low dimensional data, their accuracy and efficiency have degraded when applied on high dimensional and huge number of datasets. Besides they suffer from high computational complexity issue, for which trials for being overcome were not guaranteed, either by dimensionality reduction (DR), or using Kenel methods for instance. Therefore, to obtain better clustering results, it is worth to apply a DR method on high-dimensional datasets that allow features conservation. DL on the other hand is more effective in representation learning (RL) and feature extraction from image [4].

We propose an enhancement approach for clustering miRNAs using a Convolutional Deep Neural Network (CDNN) to address this limitation. Deep learning techniques, particularly convolutional neural networks (CNNs), have shown great promise in capturing intricate patterns in biological sequences, making them suitable for miRNA clustering. By leveraging the hierarchical and compositional nature of miRNA sequences, we design a CDNN architecture that can effectively learn the representations of miRNAs and their relationships.

The proposed CDNN architecture consists of multiple convolutional layers followed by max-pooling layers to extract features from miRNA sequences. These features are then fed into fully connected layers to perform clustering based on learned representations. To train the CDNN, we utilize a large dataset of annotated miRNA sequences, leveraging supervised and unsupervised learning strategies to enhance the network's ability to capture meaningful patterns and relationships. The training process involves minimizing a clustering loss function that encourages miRNAs with similar sequences to be grouped together, while pushing dissimilar miRNAs apart in the feature space.

To evaluate the effectiveness of our proposed approach, we conducted experiments using real miRNA datasets obtained from public repositories. We compared the clustering performance of the CDNN-based approach with traditional methods, such as hierarchical clustering and K-means clustering, using standard evaluation metrics, including Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). Our results demonstrate that the CDNN-based approach achieves superior clustering accuracy and robustness, outperforming traditional methods across different miRNA datasets. Furthermore, the CDNN exhibits a high degree of tolerance to noise and variability in miRNA sequences, making it a promising tool for handling real-world data.

2. Related Work

2.1. miRNA Clustering Using Machine Learning Techniques

MicroRNAs (miRNAs) play a pivotal role in cellular processes directly correlating to the genesis and progression of various diseases, including cancer [5]. The potential for miRNAs as therapeutic targets and disease biomarkers has triggered the growth of research into miRNA clustering, facilitating the discovery of miRNA families and their biogenesis. Concomitant (concurrent) with this has been the rise of advanced computational techniques, such as Convolutional Deep Neural Networks (CDNNs), which promise tremendous potential for clustering miRNAs [6].

miRNA clustering involves grouping them based on their sequences, expression patterns, or genes (mRNAs) they target. It provides insights into their regulatory mechanisms, biological functions, potential diseases and biomarkers therapeutic applications.

Effective computational tools are required for in-depth miRNA analysis, which led to the formation of diverse clustering approaches. Traditional methods like hierarchical clustering, k-means, and DBSCAN, among others, have been employed. However, they have limitations, such as incapacity to handle large datasets and misclassifications [7].

These clustering techniques only use sequence characteristics to cluster miRNAs and ignore functional properties. It is essential to cluster miRNA and its related functions in terms of both functional and sequence properties.

2.2. Deep Clustering

Deep learning is a subset of machine learning and artificial intelligence known for its ability to learn unlabelled and unstructured data [8]. This marvel of technology holds a profound capacity for clustering expansive genomic data, further redefining the bioinformatics ecosystem.

Different approaches have been used in Deep Clustering in the literature. These are the pipeline-model approach that first: 1) learn data representation using different deep neural network (DNN) architectures, and 2) next apply a machine learning (ML)-based clustering algorithms [4]. Deep Embedding Clustering (DEC), and Deep Clustering Network (DCN) are examples of approaches that uses multilayer perceptrons (MLP) architecture, and k-means clustering [9] [10]. Clustering Using CNN (CCNN) [11] and clustering using pairwise constraints clustering CNN (NNCPC), are examples of approaches that use CNN architecture [12], and k-means clustering.

Another approach is Single-Model approach that perform end-to end clustering without being preceded with representation learning step [13].

Deep Neural Networks, specifically the convolutional variant (CNN), have been increasingly used for miRNAs clustering due to their inherent capability to extract hierarchical features from input data automatically. CNN's unique architecture of convolutional and pooling layers works excellently in sifting through the overwhelming dimensionality and complexity of miRNAs sequences [14].

Xie *et al.* has introduced Deep Embedded Clustering (DEC) algorithm to learn feature representation and assign cluster [9]. Gui *et al.* have introduced deep clustering framework that uses convolutional auto encoders for image clustering and learning representations [15]. Yang *et al.*, have also proposed a method that improve K-means performance being integrated by deep learning [16].

Deep clustering algorithms integrate feature learning and clustering into a unified framework, promising higher accuracy and robustness [9]. Autoencoder-based clustering algorithm, an iteration of deep learning, offers a two-fold operation: encoding, which compresses the input into a lower-dimensional space and decoding, which reconstructs the original input data [17]. This methodology facilitates the identification of subtle patterns and inherent structures within genomic data.

Several publications have highlighted the methodological and computational benefits of CNNs for clustering miRNAs. In a work by [18], they successfully applied a convolutional neural network for clustering miRNA sequences and unravelling their latent taxonomy, which significantly impacted research related to the diagnosis and therapy of diseases [7]. Similarly, a study by [19] exhibited the efficacy of their novel deep learning model, DeepMirTar, to perform a binary classification for accurately predicting miRNA-target interactions. Their model outperformed traditional machine learning methods, such as SVM and Random Forest. The Convolutional Deep Neural Network (CNN) offers a potential solution to curbing the limitations of conventional miRNA clustering methods. CNNs have revolutionized numerous machine learning applications due to their ability to process large dimensional data efficiently, making them suitable for high dimensional miRNA data [20]. Their use of multiple layers for feature learning and abstract representation enhances precision and reduces misclassifications.

CDNNs are a category of Neural Networks that have shown remarkable potential in bioinformatics, specifically in sequence analysis [8]. CDNNs can automatically and adaptively learn spatial hierarchies of features from raw input data, providing a potent asset in the clustering of miRNAs.

The primary advantage of CDNNs and their suitability in classifying miRNAs is their capacity to learn abstraction from data, a skill particularly useful when dealing with complex biological data, including miRNAs. Translating raw sequence data into more abstract, high-level features, CDNNs greatly enhance the clustering process by reducing data dimensionality and capturing discriminative features [19]. Initial research into the use of CDNNs for miRNA clustering points towards positive trends. A study by [6] employed CDNNs to perform unsupervised learning of miRNA sequences, demonstrating promising results in biomarker detection, which serves to reinforce the potential of CDNNs for superior clustering of miRNA sequences and motivates further detailed exploration of this approach.

The convolutional deep neural network presents an innovative and enhanced approach to miRNA clustering. It addresses the limitations of conventional methods, effectively dealing with both functional and sequence properties of miRNAs. Hence, it provides comprehensive bioinformatics solutions that can contribute to the understanding and treatment of genetic diseases.

DeepTrust Clustering (DPCl) is a method that transforms gene expression time series into images and applies deep clustering techniques to group genes effectively. By converting time series data into images, DPCl leverages advancements in deep learning for image processing, enhancing pattern recognition and learning. This approach improves data representation and clustering performance by transforming data into a higher-dimensional space through image conversion [2].

An experiment investigated whether DNN architecture can serve a comparable function. The Pan-Cancer Analysis Project, collected data from thousands of patients with primary tumors that occurred in various body sites and covered 12 tumor types, provided the random subset of the dataset used. The experiment showed five types of cancer patients with reasonably high distinctive patterns. Patients with BRCA, COAD, and LUAD are particularly distinctly clustered, whereas patients with PRAD and KIRC are somewhat mixed and not well separated. When utilizing Convolutional Autoencoder (CAE)-based Latent Features (LFs), the Agglomerative Clustering (AC) final output is marginally superior to when using one alone. According to the optimal base clustering algorithm (in this case, the AC algorithm). The cause is that Long Short-Term Memory (LSTM)-Autoencoders (AEs) learned Latent Features (LF) are of higher quality than raw GE data, which ultimately improves the Gene Expression (GE) profiles' separability a little bit. Only some of these patterns are easily discernible in the raw GE profiles, as the t-SNE plot illustrates [4].

Deep convolutional clustering algorithms combine Convolutional Neural Networks (CNNs) with clustering techniques to extract and leverage spatial hierarchies in data, which is particularly useful for image data but can also be adopted for other types of structured data, including biological data such as miRNA expression profiles [9] [15].

DPCl is a framework uses architecture that transform time series data to image for data representation, and then apply deep convolutional clustering algorithm that uses convolutional neural networks (CNNs), and next apply k-means clustering. The conversion of expression data to image to enrich data representation. The method has shown an outperformance compared to traditional machine learning clustering algorithms.

3. Methods

3.1. DPCl

DPCl involves converting gene expression and time series into images and applying deep clustering techniques to create reliable gene clusters. This study implies DPCl algorithm, as illustrated in **Figure 1**, on a miRNA expression dataset with multiple samples.

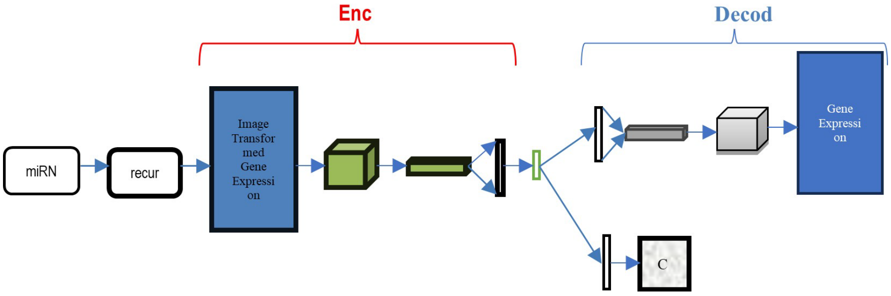


Figure 1. DPCl algorithm architecture.

3.2. The Dataset

We concentrated on the Multiple Myeloma information on GEO page GSE16558. We focused on GPL8965, which contains miRNA expression profiles that correspond to various stages of myeloma pathology. The total number of miRNA expression profiles targeted in this study were 296.

3.3. Image-Transformation Using Recurrence Plot

We make use of recurrence plots for encoding miRNA expression dataset that contains multiple samples for each miRNA as images. The recurrence plot (RP) is a graphical tool for displaying the temporal properties of dynamical systems. To be more exact, an RP is a phase space representation of the trajectories of dynamical systems [21] [22].

An RP is a binary N × N image defined as:

$$R_{i,j} = \begin{Bmatrix} 1 & \mathcal{E} - \left\| \vec{x}_{(i)} - \vec{x}_{(j)} \right\| \geq 0 \\ 0 & \text{otherwise} \end{Bmatrix} \tag{1}$$

where $R_{i,j}$ is the pixel value of the *i*th row and *j*th column, $\mathcal{E}$ is the radius of the $\mathcal{E}$-tube defining the largest acceptable distance between trajectories to be considered as recurrent and $\vec{x}_{(i)}$ is the *i*th element in input data

$$R_{i,j} = \left\| \vec{x}_{(i)} - \vec{x}_{(j)} \right\| \tag{2}$$

Equation 1 becomes Equation 2 if thresholding operation is omitted. The new image is known as a global recurrence plot since it is unthresholded and consequently non-binary. We restrict ourselves to global recurrence plots Equation 2. The dynamic behavior traits are reflected in patterns on RP.

One of the common problems with agglomerative clustering is determining the number k of clusters. We chose to use the elbow approach [22] and found that eight miRNA clusters (k) were needed to favor a meaningful result.

The convolutional autoencoder used in the DPCl algorithm's parameters was set as follows: 1) network structure that contains three connected convolutional layers with (32, 64, 128) filters, 2) (5, 5, 3) kernel size and 3) same stride length (2) for all convolutional layers. The dimensions of the embedded space are equal to 8, which is the number of miRNAs clusters. The decoder part of the network is symmetric to the encoder part. In the decoder part, we used convolutional transpose layers with stride. We used the ReLU [23] activation function on all convolutional layers to add non-linearity to our model and avoid the vanishing gradient problem. We trained the model for 300 epochs using ADAM optimizer [24].

3.4. Clustering Recurrence Plots

Several DNN architectures are used in deep clustering. Also, since augmenting a vanilla autoencoder with convolutional units takes the spatial structures into account, it is straightforward and improves visual imagery performance [23] [25]. Using the autoencoder's activations on its bottleneck layer, embeddings, as the inputs is a simple strategy for convolutional autoencoder-based clustering, which can be achieved in two steps: i) loading of cluster centroids, ii) iterative clustering through modification of centroids. The loading is created by mapping n gene expression recurrence plots into a lower-dimensional latent space Z, which is done by training a convolutional autoencoder. Each recurrence plot is passed through the autoencoder, and standard k-means is performed in the embedding space Z after the training is finished. These operations result with K initial centroids μ_j where $j = 1,\ldots,K$. After the auto encoder is trained, the decoder part is detached from the network, we are only interested in the generation of better embeddings from the encoder part.

4. Results and Discussion

We used data from a study that implemented a holistic procedure to evaluate our deep clustering results to discover miRNA-mRNA modules [3]. This study utilized both miRNA and mRNA expression datasets and miRNA target prediction databases based on sequence data or experimentally validated and data bases that use both sequence and expression the highest score is 0.82008266 when applied. In contrast, our deep clustering technique involved only the miRNA expression dataset.

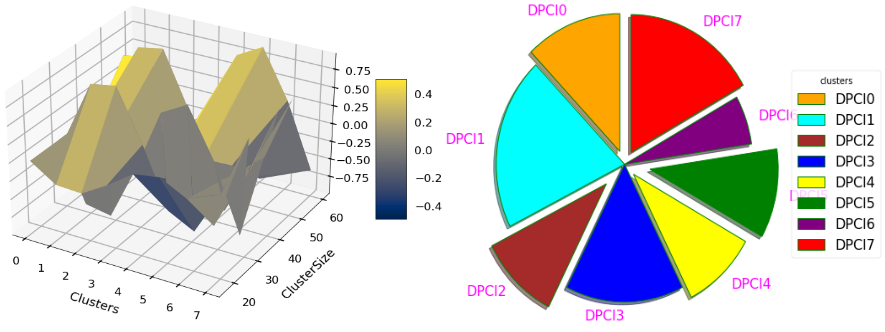
Table 1 represents the results after applying DPCl, and if compared with the results obtained from the holistic procedure which showed that 40% of the miRNAs were assigned to cluster 7 as in Table 2. Our method distributed these miRNAs among all 8 clusters, demonstrating unbiased clustering results as shown in Figure 2, unlikely the biased distribution shown in Figure 3.

Table 1. miRNA clusters generated by DPCl.

cls_name	DPCl0	DPCl1	DPCl2	DPCl3	DPCl4	DPCl5	DPCl6	DPCl7
cls_size	36	60	30	44	28	31	17	50

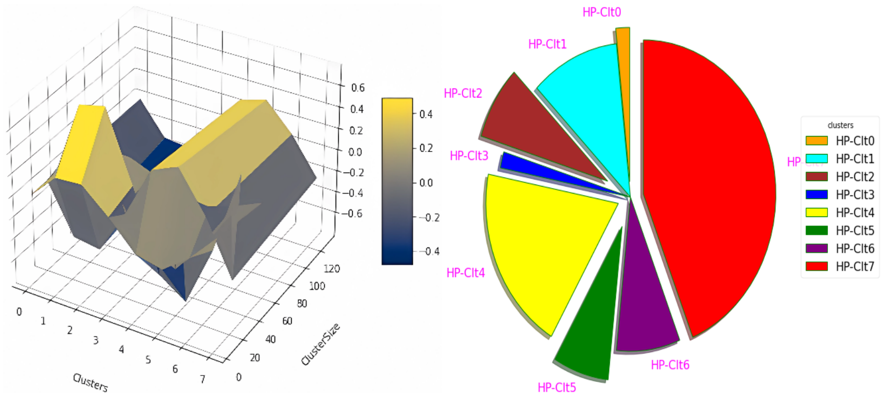
Table 2. miRNA clusters generated by generic clustering method.

cls_name	HP-Clt0	HP-Clt1	HP-Clt2	HP-Clt3	HP-Clt4	HP-Clt5	HP-Clt6	HP-Clt7
cls_size	4	32	23	5	58	20	23	130



(a) using pie chart (b) using pie chart

Figure 2. Distribution of clusters using DPCl. Which demonstrate the unbiased clustering of these miRNAs among all 8 clusters, as shown in a and b charts.



(a) using landscape (b) using pie chart

Figure 3. Distribution of HP Clusters, which demonstrate the biased clustering of these miRNAs among all 8 clusters, as shown in a and b charts.

Table 3 describes how the HP-clusters resembled in rows are redistributed using DC resembled as columns. It is evident that the miRNAs in DPCl1 are mostly from HP-Clt2, HP-Clt3 and HP-Clt5 respectively.

Table 3. Redistribution of the HP-clusters into the DP-clusters.

HP/DPCl	DPCl0	DPCl1	DPCl2	DPCl3	DPCl4	DPCl5	DPCl6	DPCl7	Total
HP-Clt0	0	0	0	2	1	0	0	2	5
HP-Clt1	3	1	2	10	7	2	4	3	32
HP-Clt2	0	22	0	0	0	0	0	1	**23**
HP-Clt3	0	5	0	0	0	0	0	0	**5**
HP-Clt4	6	5	9	9	4	10	2	13	58
HP-Clt5	0	**20**	0	0	0	0	0	0	**20**
HP-Clt6	1	0	3	4	1	2	1	11	23
HP-Clt7	26	7	16	19	15	17	10	20	130
Total	36	60	30	44	28	31	17	50	

This table shows how Holistic Procedure (HP) clusters have been redistributed among DPCl clusters (DC). HP-Clst7 is the largest cluster and has been distributed among ALL DC clusters.

Since the miRNAs in HP-Clt2 which were 23, they were almost on the HP-Clt1 except only one, but when looking at HP-Clt3 and HP-Clt5 we can find that they were all included in DPCl1. The cluster number of course does not mean anything particularly, but it is clear that the miRNAs in these clusters are closely related to each other. The miRNAs are identified in **Tables 4-6** respectively.

Table 4. HP-Clt2.

HP-Clt2
hsa-miR-122-5p, hsamiR-124-3p, hsa-miR-214-3p, hsa-miR-24-3p, hsa-miR-30a-3p, hsa-miR-323b-5p, hsa-miR-325, hsa-miR-371a-3p, hsa-miR-373-3p, hsa-miR-502-5p, hsa-miR-510-5p, hsa-miR-516b-5p, hsa-miR-518a-3p, hsa-miR-520c-3p, hsa-miR-526b-5p, hsa-miR-532-5p, hsa-miR-542-5p, hsa-miR-548a-3p hsa-miR-551b-3p, hsa-miR-575, hsa-miR-596, hsa-miR-622

Table 5. HP-Clt3.

HP-Clt3
hsa-miR-548d-3p, hsa-miR-553, hsa-miR-580-3p, hsa-miR-653-5p, hsa-miR-656-3p

Table 6. HP-Clt5.

HP-Clt5
hsa-miR-206, hsa-miR-299-5p, hsa-miR-337-3p, hsa-miR-379-5p, hsa-miR-381-3p, hsa-miR-424-5p, hsa-miR-514a-3p, hsa-miR-515-5p, hsa-miR-517a-3p, hsa-miR-518b, hsa-miR-544a, hsa-miR-562, hsa-miR-563, hsa-miR-597-5p, hsa-miR-600, hsa-miR-617, hsa-miR-660-5p, hsa-miR-95-3p, hsa-miR-98-5p, hsa-miR-99a-5p

Considering HP-Clt7 and focusing on two miRNA disease studies:1) A study conducted in August 2020 by Caixia Li *et al.* on human patients with COVID-19 elucidated differentially expressed miRNAs [26]. 2) Karina *et al.* focused on a group of miRNAs called mir-17-92 and their relationship with the E2F-RB pathway, which contributes to various types of cancers such as lung, breast, bladder, and brain [27]. They also demonstrated the relationship of these miRNAs with colorectal cancer [28]. Please refer to **Table 7** to differentiate between the miRNAs in these two studies. **Table 8** shows that miRNAs hsa-miR-16 and hsa-miR-146b fell into the same cluster. Jose' Marı'a Galva'n-Roma' *et al.* showed that both of these miRNAs could be used as biomarkers for CAP prognosis [29] and were also differentially expressed in the COVID-19 study mentioned above.

DPCl has discovered how possible interrelations among members inside one cluster could be found.

Additionally, Cristina Morsiani showed that miR-92a-3p and miR-18a-5p, which fell into the same cluster (specify), are potential biomarkers for blood circulation in liver transplant recipients. These miRNAs were upregulated in recipients with certain complications, as shown in [30].

The DPCl regathered these miRNAs in the same cluster, although they were in different clusters using HP. This show shows how DPCl could discover potential biomarkers for diseases.

Table 7. COVID miRNAs.

HP-Clst	miRNAs	DC-Clst
4	hsa-miR-17-5p	2
4	hsa-miR-18a-5p	3
6	hsa-miR-618	7
7	hsa-miR-30c-5p	2
7	hsa-miR-627-5p	3
7	hsa-miR-183-5p	4
7	hsa-miR-146b-5p	5
7	hsa-miR-16-5p	5
7	hsa-miR-21-5p	6

Table 8. miR-17-92 Cluster.

HP-Clst	miRNAs	DC-Clst
1	hsa-miR-92a-3p	3
4	hsa-miR-17-3p	7
7	hsa-miR-19a-3p	0
7	hsa-miR-19b-3p	3
7	hsa-miR-20a-5p	6
4	hsa-miR-17-5p	2
4	hsa-miR-18a-5p	3

On the other hand, although DPCl has the ability to discover interrelations among cluster members, it failed in getting into more depth to express these relations. For instance, miRNAs hsa-miR-16 and hsa-miR-183 fell into different clusters [31] using HP, but have been redistributed into different clusters using DPCl. Dan Cao in [31] identify that miRNAs associated with active tuberculosis (ATB) demonstrated that among the differentially expressed miRNAs, hsa-miR-16 was significantly decreased while hsa-miR-183 was significantly increased. From this study we find that both of them were potential biomarkers, but have different level of expression which has been expressed by DPCl just distributing them into different clusters.

5. Conclusion

To conclude our paper, we aimed to get much deeper in clustering results that use machine learning, by adopting a deep learning strategy to dig inside the interrelation among miRNA elements inside the same clusters. The strategy uses deep clustering technique that transforms expression data to images by applying CNN. Therefore, we have exploited DPCl algorithm originally proposed for time series data, and applied it on miRNA expression samples data.

This work has been carried out to focus on miRNA clustering from a data analytics perspective. Therefore, since the data has been derived from a biological database, it is worth to incorporate biological perspective to enhance the effectiveness of the procedure. Besides, results have shown how interrelations among miRNAs in one cluster could open research questions in investigating disease etiologist.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Dundar, E., Jin, A. and Culurciello, J. (2014) Preprint Repository arXiv Achieves Milestone Million Uploads. *Physics Today*, **2014**, No. 12. https://doi.org/10.1063/pt.5.028530

[2] Ozgul, O.F., Bardak, B. and Tan, M. (2021) A Convolutional Deep Clustering Framework for Gene Expression Time Series. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **18**, 2198-2207. https://doi.org/10.1109/tcbb.2020.2988985

[3] Shommo, G. and Apolloni, B. (2021) A Holistic miRNA-mRNA Module Discovery. *Non-Coding RNA Research*, **6**, 159-166. https://doi.org/10.1016/j.ncrna.2021.09.001

[4] Karim, M.R., Beyan, O., Zappa, A., Costa, I.G., Rebholz-Schuhmann, D., Cochez, M., *et al.* (2020) Deep Learning-Based Clustering Approaches for Bioinformatics. *Briefings in Bioinformatics*, **22**, 393-415. https://doi.org/10.1093/bib/bbz170

[5] Miotto, R., Wang, F., Wang, S., Jiang, X. and Dudley, J.T. (2017) Deep Learning for Healthcare: Review, Opportunities and Challenges. *Briefings in Bioinformatics*, **19**, 1236-1246. https://doi.org/10.1093/bib/bbx044

[6] Zhou, X., Menche, J., Barabási, A. and Sharma, A. (2014) Human Symptoms-Disease Network. *Nature Communications*, **5**, Article No. 4212. https://doi.org/10.1038/ncomms5212

[7] An, J., Lai, J., Lehman, M.L. and Nelson, C.C. (2012) Mirdeep*: An Integrated Application Tool for miRNA Identification from RNA Sequencing Data. *Nucleic Acids Research*, **41**, 727-737. https://doi.org/10.1093/nar/gks1187

[8] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436-444. https://doi.org/10.1038/nature14539

[9] Xie, J., Girshick, R. and Farhadi, A. (2016) Unsupervised Deep Embedding for Clustering Analysis. 33*rd International Conference on Machine Learning ICML* 2016, Vol. 1, 740-749.

[10] Yang, J., Parikh, D. and Batra, D. (2016) Joint Unsupervised Learning of Deep Representations and Image Clusters. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 5147-5156. https://doi.org/10.1109/cvpr.2016.556

[11] Lukic, Y., Vogt, C., Durr, O. and Stadelmann, T. (2016) Speaker Identification and

Clustering Using Convolutional Neural Networks. 2016 *IEEE* 26*th International Workshop on Machine Learning for Signal Processing* (*MLSP*), Vietri sul Mare, 13-16 September 2016, 1-6. https://doi.org/10.1109/mlsp.2016.7738816

[12] Hsu, Y.-C. and Kira, Z. (2015) Neural Network-Based Clustering Using Pairwise Constraints. 1-12. http://arxiv.org/abs/1511.06321

[13] Johnson, S.C. (1967) Hierarchical Clustering Schemes. *Psychometrika*, **32**, 241-254. https://doi.org/10.1007/bf02289588

[14] Young, T., Hazarika, D., Poria, S. and Cambria, E. (2018) Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. *IEEE Computational Intelligence Magazine*, **13**, 55-75. https://doi.org/10.1109/mci.2018.2840738

[15] Guo, X., Liu, X., Zhu, E. and Yin, J. (2017) Deep Clustering with Convolutional Autoencoders. In: Liu, D.R., *et al.*, Eds., *Neural Information Processing*, Springer International Publishing, 373-382. https://doi.org/10.1007/978-3-319-70096-0_39

[16] Yang, M., Fu, B., Sidiropoulos, X. and Hong, N.D. (2017) Towards k-Means-Friendly Spaces: Simultaneous Deep Learning and Clustering. *International Conference on Machine Learning*, Sydney, 6-11 August 2017, 3861-3870.

[17] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. and Manzagol, P.A. (2010) Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, **11**, 3371-3408.

[18] Xue, H., Li, J., Xie, H. and Wang, Y. (2018) Review of Drug Repositioning Approaches and Resources. *International Journal of Biological Sciences*, **14**, 1232-1244. https://doi.org/10.7150/ijbs.24612

[19] Ren, Y., Pu, J., Yang, Z., Xu, J., Li, G., Pu, X., *et al.* (2024) Deep Clustering: A Comprehensive Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 1-21. https://doi.org/10.1109/tnnls.2024.3403155

[20] Otis, J.S., Niccoli, S., Hawdon, N., Sarvas, J.L., Frye, M.A., Chicco, A.J., *et al.* (2014) Pro-Inflammatory Mediation of Myoblast Proliferation. *PLOS ONE*, **9**, e92363. https://doi.org/10.1371/journal.pone.0092363

[21] Marwan, N., Wessel, N., Meyerfeldt, U., Schirdewan, A. and Kurths, J. (2002) Recurrence-Plot-Based Measures of Complexity and Their Application to Heart-Rate-Variability Data. *Physical Review E*, **66**, Article ID: 026702. https://doi.org/10.1103/physreve.66.026702

[22] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning Representations by Back-Propagating Errors. *Nature*, **323**, 533-536. https://doi.org/10.1038/323533a0

[23] Masci, J., Meier, U., Cireşan, D. and Schmidhuber, J. (2011) Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. In: Honkela, T., *et al.*, Eds., *Artificial Neural Networks and Machine Learning—ICANN* 2011, Springer, 52-59. https://doi.org/10.1007/978-3-642-21735-7_7

[24] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. MIT Press.

[25] Ben-Dor, A. and Yakhini, Z. (1999) Clustering Gene Expression Patterns. *Proceedings of the* 3*rd Annual International Conference on Computational Molecular Biology*, Lyon, 11-14 April 1999, 33-42. https://doi.org/10.1145/299432.299448

[26] Pairo-Castineira, E., Clohisey, S., Klaric, L., Bretherick, A.D., Rawlik, K., Pasko, D., *et al.* (2020) Genetic Mechanisms of Critical Illness in Covid-19. *Nature*, **591**, 92-98. https://doi.org/10.1038/s41586-020-03065-y

[27] Conkrite, K., Sundby, M., Mukai, S., Thomson, J.M., Mu, D., Hammond, S.M., *et al.* (2011) *miR*-17~92 Cooperates with *RB* Pathway Mutations to Promote Retinoblastoma. *Genes & Development*, **25**, 1734-1745.

https://doi.org/10.1101/gad.17027411

[28] Mohajeri Khorasani, A., Mohammadi, S., Raghibi, A., Haj Mohammad Hassani, B., Bazghandi, B. and Mousavi, P. (2024) miR-17-92a-1 Cluster Host Gene: A Key Regulator in Colorectal Cancer Development and Progression. *Clinical and Experimental Medicine*, **24**, Article No. 85. https://doi.org/10.1007/s10238-024-01331-1

[29] Galván-Román, J.M., Lancho-Sánchez, Á., Luquero-Bueno, S., Vega-Piris, L., Curbelo, J., Manzaneque-Pradales, M., *et al.* (2020) Usefulness of Circulating microRNAs miR-146a and miR-16-5p as Prognostic Biomarkers in Community-Acquired Pneumonia. *PLOS ONE*, **15**, e0240926. https://doi.org/10.1371/journal.pone.0240926

[30] Morsiani, C., Collura, S., Sevini, F., Ciurca, E., Bertuzzo, V.R., Franceschi, C., *et al.* (2023) Circulating miR-122-5p, miR-92a-3p, and miR-18a-5p as Potential Biomarkers in Human Liver Transplantation Follow-Up. *International Journal of Molecular Sciences*, **24**, Article No. 3457. https://doi.org/10.3390/ijms24043457

[31] Cao, D., Wang, J., Ji, Z., Shangguan, Y., Guo, W., Feng, X., *et al.* (2020) Profiling the mRNA and Mirna in Peripheral Blood Mononuclear Cells in Subjects with Active Tuberculosis. *Infection and Drug Resistance*, **13**, 4223-4234. https://doi.org/10.2147/idr.s278705