Capturing biological patterns from gene expression data of Polycystic Ovarian Syndrome using unsupervised dimensionality reduction algorithms

A V Lakshmy[1], N Sowmya Manojna[2]

1 Department of Computer Science and Engineering, Indian Institute of Technology, Madras; cs16b101@smail.iitm.ac.in
2 Department of Biotechnology, Indian Institute of Technology, Madras; sowmyamanojna@smail.iitm.ac.in

Abstract

Polycystic Ovarian Syndrome (PCOS) is a disorder caused due to endocrine dysfunction, that affects women of reproductive age. Although the aetiology of PCOS isn't known, patients diagnosed with PCOS are generally found to exhibit elevated levels of the androgen and lower levels of progesterone. In order to understand the pathophysiology of PCOS, we have explored PCOS gene expression data comprising 9 datasets from the NCBI GEO database. We have used unsupervised linear dimensionality reduction techniques such as Principal Component Analysis (PCA), Independent Component Analysis (ICA) & Non-negative Matrix Factorization and non-linear dimensionality reduction techniques such as Variational Auto Encoders (VAE) & Denoising Auto Encoders (DAE) to extract biologically important signals from the data. The VAE network was trained using the `binary cross-entropy` loss function coupled with a Kullback-Leibler divergence penalty, while the DAE network was trained using a `MSE` cost function. Our model has identified 5 genes - FAM163A, FOLR2, S100A6, AKR1A1 and MCL1, that correspond to the latent dimensions that maximally separate the PCOS data points from the control data points. These genes were found to participate in key pathways related to PCOS such as insulin secretion, vitamin and mineral absorption, insulin resistance, androgen and prostaglandins productions. Additionally, we also worked on understanding the ability of the different dimensionality reduction algorithms in identifying key features in the biological data, their stability and the similarity in the features identified by each algorithm across different latent dimensions.

Introduction

Most cells in a given organism contain the same set of genes, but only a select set of these genes are expressed at any given time. This gene expression is a highly regulated process. When cellular systems are perturbed, changes in the biological processes reflect on the gene expression data.

Gene expression analysis involves studying the behavior and functioning of individual genes in an organism, by analyzing their gene expression (mRNA) levels. The dynamic functional view provided by gene expression analysis enables us to acquire a better understanding of individual genes, gene networks and biological pathways. This in turn proves very useful in understanding the pathophysiology of diseases, analyzing the effects of mutations, evaluating the efficiency of drugs, and so on. However, due to the enormity of the gene

expression data and the large number of parameters that need to be estimated in the process, direct analysis of the data is both tedious and computationally intensive. Hence, many gene expression analysis studies make use of dimensionality reduction methods, such as Principal Component Analysis (PCA) [1] [2], before analyzing the gene expression data.

However, deriving insights from the latent (transformed) space alone, is difficult due to the limitations imposed by the transformation. Additionally, we need to decode the latent space representation back to the original space in order to completely understand the biological significance. Unsupervised neural network based approaches, such as autoencoders, can help deal with these shortcomings, by performing both dimensionality reduction (encoding) and decoding. Recent works such as [3] [4] [5] employ Denoising Autoencoders (DAEs), while the works of [3] [6] [7] explore Variational Autoencoders (VAEs). Depending on the algorithm used and the range of latent space dimensionalities explored, different kinds of biological patterns can be captured [3].

Polycystic Ovarian Syndrome (PCOS) is an endocrine disorder that affects almost one-tenth of women of reproductive age globally [8]. The exact cause of PCOS is still uncertain [9]. Many prior works have analysed differential expressions of certain genes in samples collected from different tissues of PCOS patients as well as non-PCOS controls, in order to gain insights into the pathophysiology of PCOS. Some of these include ovarian theca cells [10], ovarian tissue cells [11], omental adipose tissue [12], granulosa cells [13], cumulus cells [14], and endometrial tissue [15]. Many of these studies have employed PCA as a dimensionality reduction technique to analyse the gene expression data.

In our study, we have focused on analyzing gene expression data of PCOS patients as well as non-PCOS controls, in order to infer biological patterns associated with PCOS, by applying unsupervised dimensionality reduction techniques, including Principal Component Analysis (PCA), Independent Component Analysis (ICA), Non-negative Matrix Factorization (NMF), Denoising Autoencoders (DAEs) and Variational Autoencoders (VAEs), across different latent dimensional representations [3].

Materials and methods

Gene Expression Data Collection

All the gene expression data used in our study was obtained from NCBI GEO. [16] A total of 9 datasets related to PCOS was available on GEO. The gene expression datasets and the associated annotation files were accessed and parsed programmatically from the search result.

The gene expression levels for each dataset was extracted from their respective SOFT files and the mapping between the gene IDs used in the dataset and their Entrez Gene IDs was extracted from their respective annotation files. The distribution of the genes (i.e.) Normal/Control, PCOS or after drug application was extracted from the Sample Subset tab of each dataset.

GEO Sample ID	# PCOS Samples	# Control Samples
GDS1050	9	4
GDS1051	9	4
GDS2084	8	7
GDS3104	16	13
GDS3841	12	11
GDS4132	10	10
GDS4133	10	13
GDS4399	7	3
GDS4987	14	15

Table 1. Individual Datasets Summary

Data Pre-Processing

All genes that did not map to an Entrez ID were removed. In cases where multiple gene IDs had the same Entrez ID, the average expression of the multiple genes was taken.

The gene expression levels were then normalized to the range $[0, 1]$ using Min-Max Scaling. Unique Entrez IDs across all the datasets was taken and the entries from the 9 datasets were combined into a single dataset. As imputation across 24,981 columns was computationally intensive, we only considered columns that had complete data.

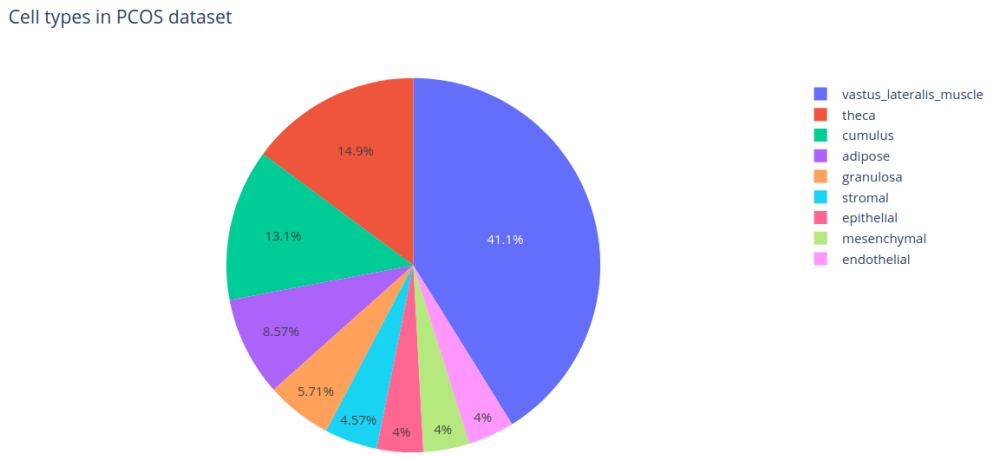

Fig 1. Cell type distribution in the final combined dataset.

Exploratory Data Analysis

Once the mapping and consolidation of the dataset was done, an exploratory analysis of the data using heterogeneous networks was carried out. The association data between the nodes was taken from MSigDB [17]. From the results below, we observe that most of our genes in the final dataset comprise of genes related to the human immune system.

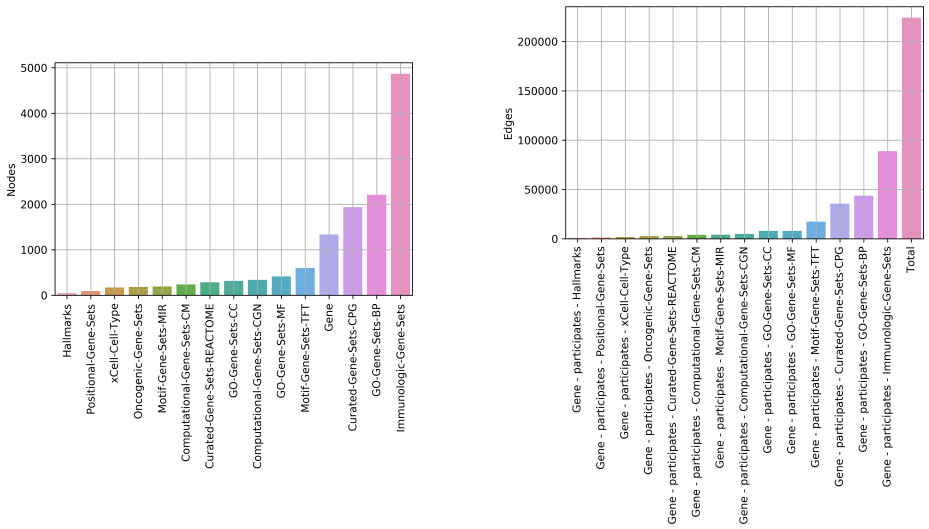

Fig 2. Analysis of the metanodes and metaedges present in the heterogeneous network of the PCOS data, on the left and right respectively.

Techniques Used for Unsupervised Dimensionality Reduction

Following the approach given by [3], we have employed linear as well as non-linear unsupervised dimensionality reduction methods to analyze the PCOS gene expression data.

Linear dimensionality reduction techniques

- Principal Component Analysis (PCA): PCA identifies linear directions along which the variance of the data is maximum. The directions are then arranged based on the decreasing order of variance. The data is then projected onto the reduced set of directions.
- Independent Component Analysis (ICA): The dimensionality reduction done by ICA is very similar to that of PCA, but, it makes a key assumption that the latent dimensions are mutually independent and non-Gaussian. It is like a rotation of PCA.
- Non-negative Matrix Factorization (NMF): NMF dimensionality reduction is used on samples that have non-negative values. The coefficients in the linear combination of the initial data are necessarily non-negative. Hence, the dimensions that don't contribute to the latent dimensions have a zero coefficient.

Non-linear dimensionality reduction techniques

Autoencoders are unsupervised deep neural networks, which comprise of three main parts: an **encoder** that performs dimensionality reduction on the input data, one or more **hidden layer(s)** which capture the latent or hidden representations, and a **decoder** that decodes the latent dimensional representation back to the original dimension, as the output.

- Variational Autoencoders (VAE): VAEs are stochastic autoencoding frameworks. The hidden layers in a VAE learn and take into account both the mean and the variance of the data. The decoder then samples from this distribution generator (with variations) to produce the output.
- Denoising Autoencoders (DAE): In a DAE, some random noise is deliberately added to the input data, and the network is trained so as to remove the noise and capture the relevant signals.

Architecture of Auto-Encoders

The architecture of the Variational and Denoising Auto Encoders used is as follows:

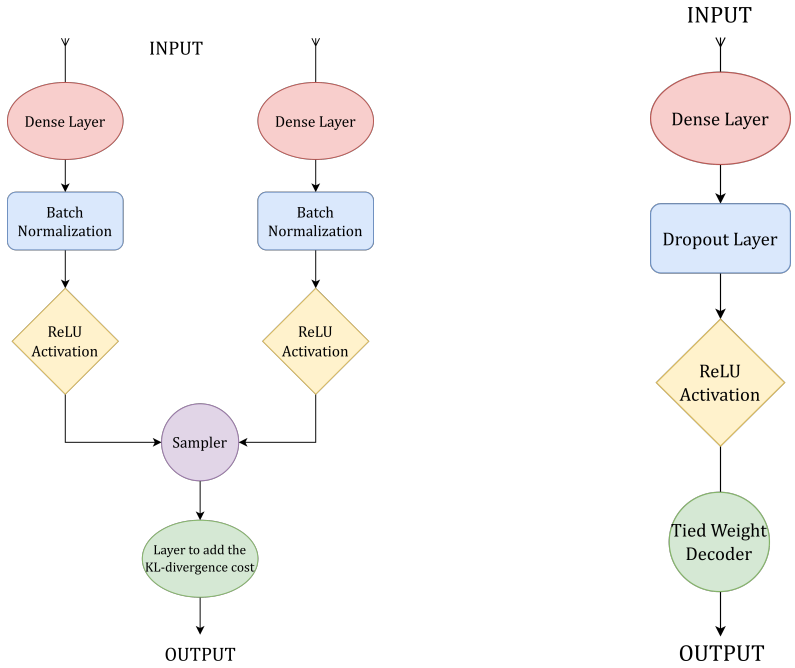

Fig 3. Architecture of the Variational and Denoising Autoencoders used in our model on the left and right respectively. [3]

Computational Resources

The analysis and implementation of linear dimensionality reduction algorithms (i.e.) PCA, ICA and NMF, was performed on an Intel Core i5 CPU, 64-bit computer laptop. The analysis and implementation of the non-linear dimensionality reduction algorithms (i.e.) Denoising Autoencoders and Variational Autoencoders were performed on Google Colaboratory using the Google GPUs as backend.

Results

PCA, ICA, NMF across Latent Space Dimensionalities

PCA, ICA and NMF were applied on the merged PCOS dataset, across a set of 27 latent dimensionalities, ranging from 2 to 150.

Reconstruction Performance

In order to assess the ability of the models to capture the fundamental features of the data in the latent dimensions, we calculated the reconstruction cost of all the models. We use both L_2 loss and `binary cross-entropy` loss functions. In case of `binary cross-entropy`, to ensure that input values were integers, the data was typecasted to `int`. The reconstruction performance of the algorithms are as follows:

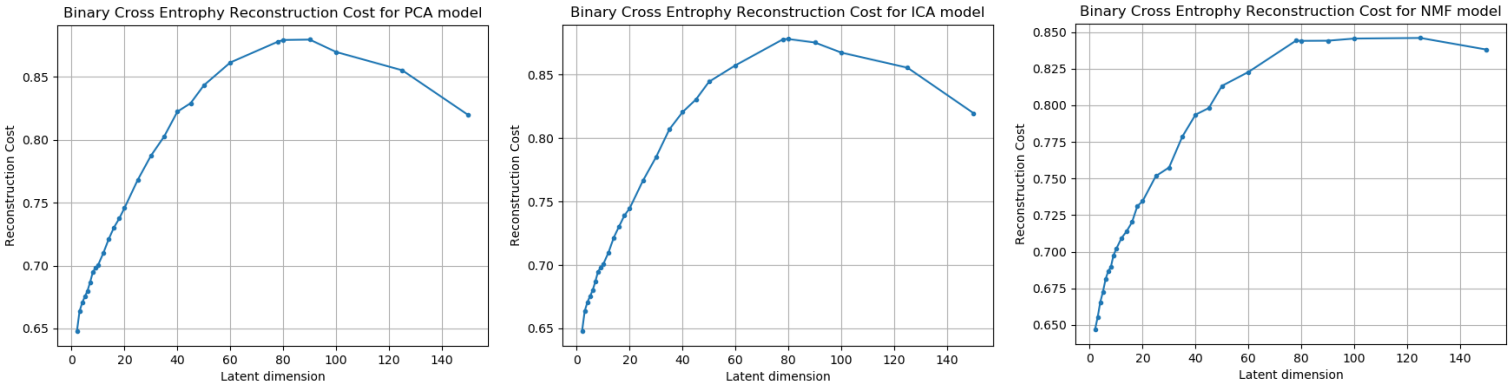

Fig 4. Cost - `binary cross-entropy` loss across different algorithm and different latent dimensions.

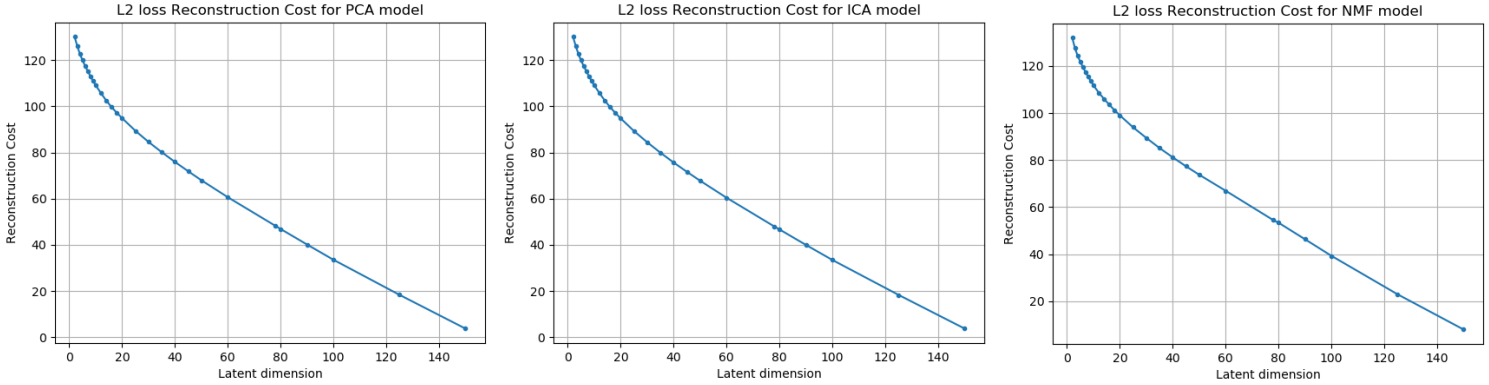

Fig 5. Cost - L_2 loss across different algorithm and different latent dimensions.

From the figures above, we can see that as the latent dimensions are increased, the L_2 loss between the input and the reconstructed output decreases. However, we also see that the `binary cross-entropy` cost is low at both extremes and peaks at an intermediate latent dimension. This variation could be potentially due to the typecasting performed. Hence, we are using the L_2 loss alone.

From the L_2 reconstruction cost, we can see that the model is able to capture the biological signals from the data and the loss decreases as the number of latent dimension increases.

Strongly Associated Dimensions

The strongly associated components are the components that led to a sudden increase in the features captured from the data by the algorithm. In order to identify the strongly associated dimensions, we have used the correlation between the input data and the reconstructed data from the latent dimensions. [3]

The correlation of the input and the reconstructed output is as follows:

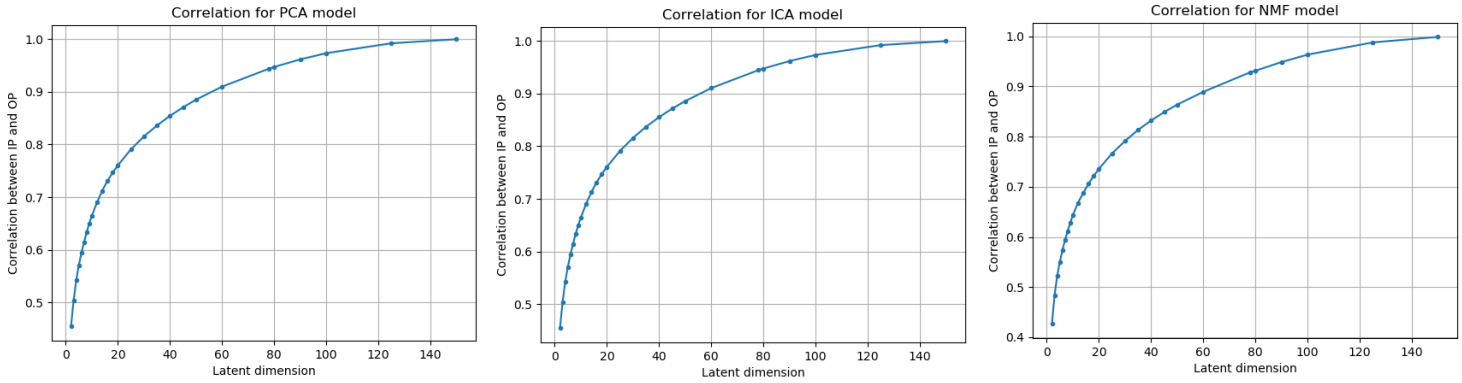

Fig 6. Correlation between the input and the reconstructed output across different algorithm and different latent dimensions.

The correlation gain plots across different algorithms is as follows:

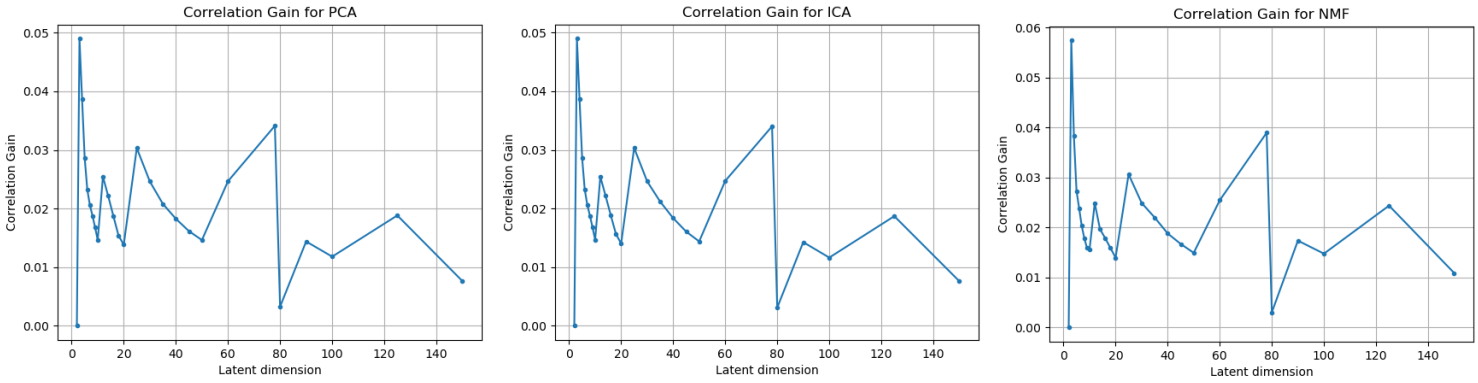

Fig 7. Correlation gain observed across different algorithms and different latent dimensions.

From the images above, we can see that the latent dimension 3 has the largest correlation peak indicating that the addition of the dimension has captured a large amount of biological data when compared to the immediate prior dimensions. Similar peaks are observed at dimensions 10, 25, 60, 78 and 125. The combination of these dimensions make up the strongly associated dimensions.

DAE, VAE across Latent Space Dimensionalities

DAE and VAE models were trained on the merged PCOS dataset, across the 27 latent dimensions, ranging from 2 to 150.

Reconstruction Performance

The `MSE` cost function was used in case of DAE, while the `binary cross-entropy` loss function coupled with Kullback-Leibler divergence penalty was used in case of VAE. The reconstruction performance of the Variational Autoencoders are as follows:

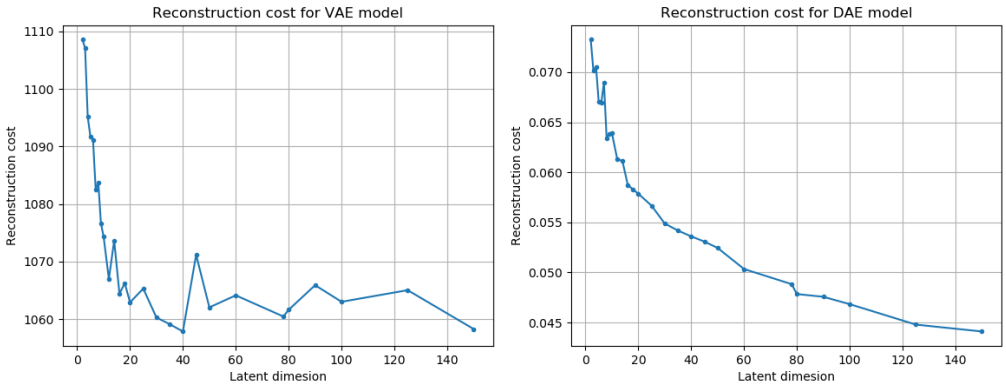

Fig 8. Reconstruction Performance of the autoencoder models - VAE and DAE, to the left and right respectively.

Strongly Associated Dimensions

The strongly associated dimension analysis using correlation is as follows:

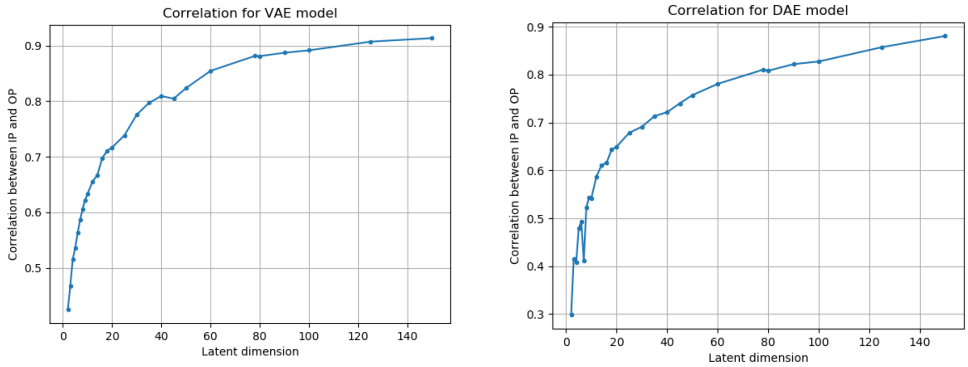

Fig 9. Correlation between the input and the reconstructed output in the auto-encoder models.

The correlation gain for the two models is as follows:

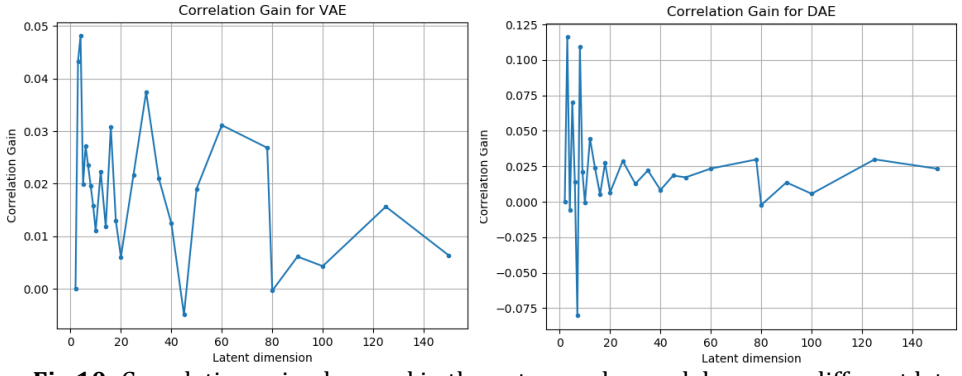

Fig 10. Correlation gain observed in the autoencoder models, across different latent dimensions.

Although the number of peaks in the correlation gain is higher in the autoencoder models, the latent dimension 3 still contributes to the largest peak as observed before.

Consolidated Correlation Analysis

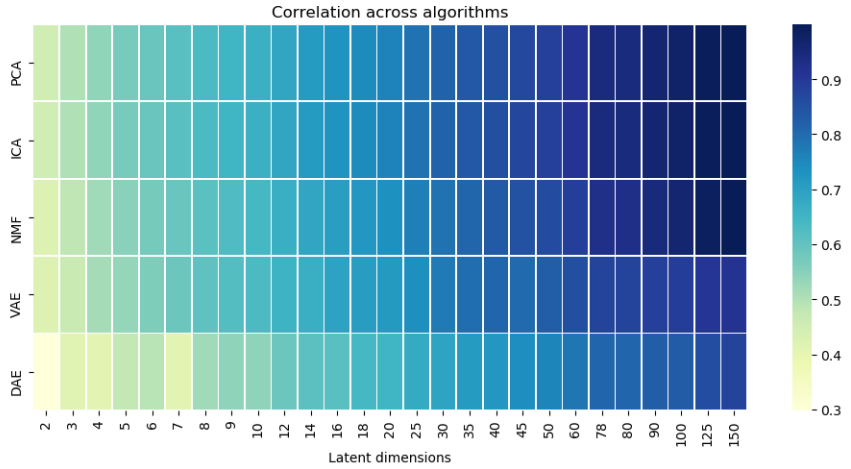

Fig 11. Consolidated correlation analysis across different algorithms.

Stability and Correlation Analysis

To analyse the stability of the 5 different models (PCA, ICA, NMF, VAE, DAE), as well as the correlation between the models within and across different latent dimensionalities, Singular Value Canonical Correlation Analysis (SVCCA) scores were calculated from the model weights [3] [18].

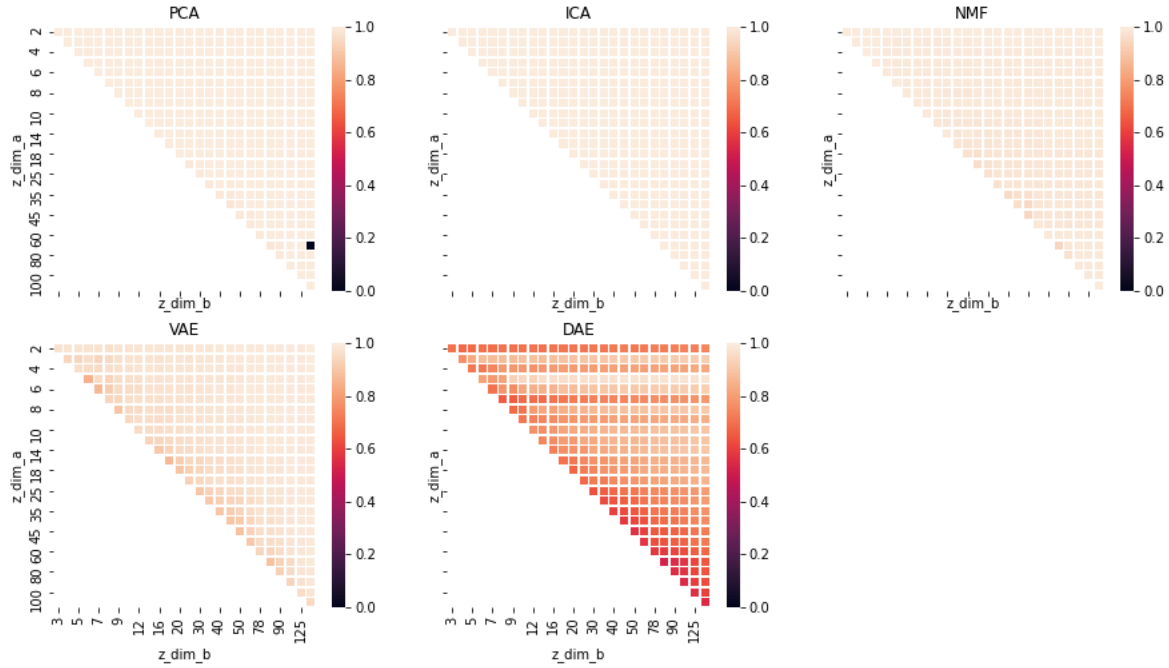

Fig 12. Stability (Mean SVCCA scores) across latent dimensionalities

As observed from the above figure, for PCA and ICA, the SVCCA scores between the weights are very high, across the different latent dimensions explored. This is because PCA is a

deterministic algorithm, and even if we increase the number of latent space dimensions, only the solution with the maximum variance is considered. Since ICA is like a rotation of PCA, the results of ICA are also comparable with those of PCA. However, there is a very low correlation between latent dimensions 78 and 150 for PCA. For NMF as well as VAE, the SVCCA scores are high across all the latent dimensions. For DAE, we can see that the SVCCA scores have a lot of variation (ranging from 0.6 to 1.0) across the different latent dimensionalities. This is indicative of the relatively lower stability of the DAE model due to the random initializations.

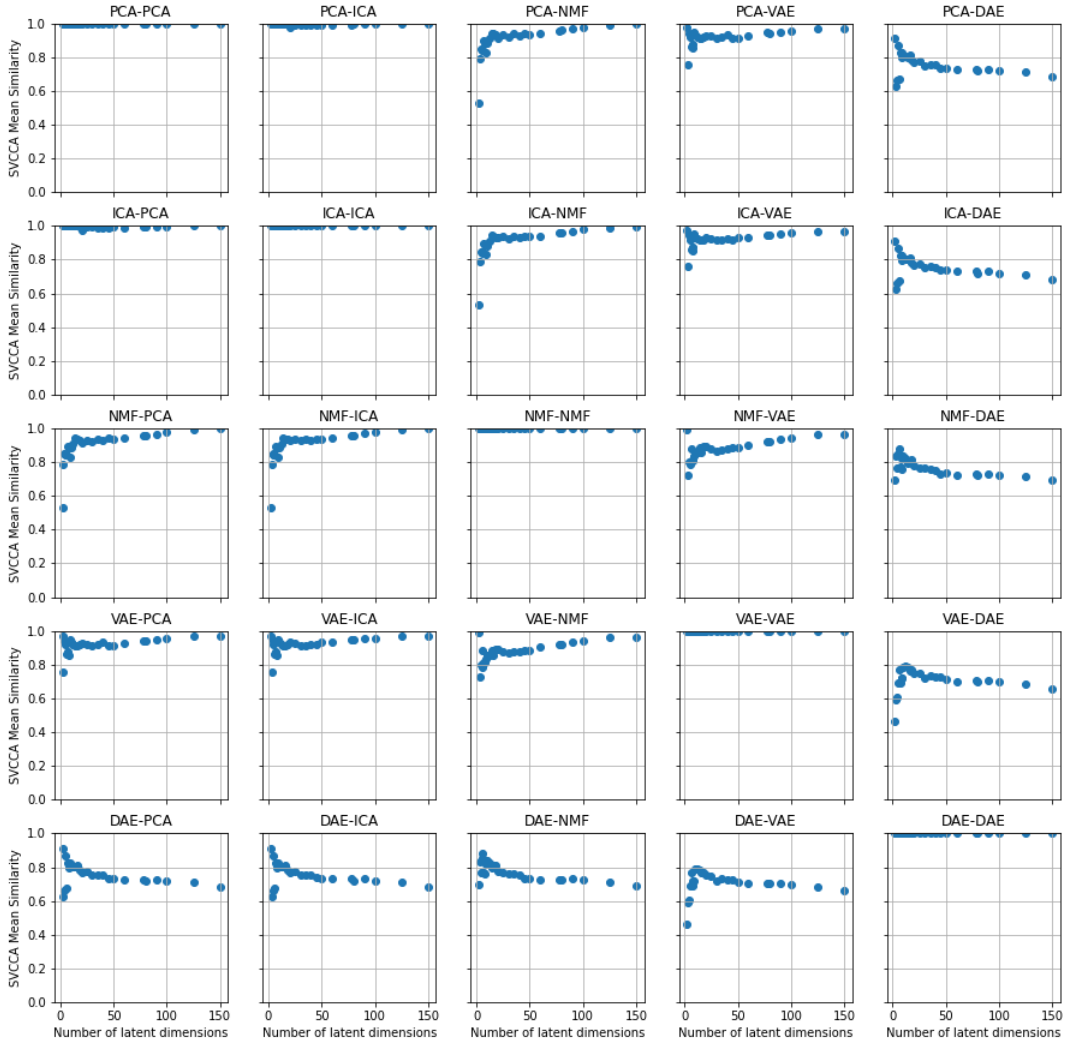

Fig 13. Stability (Mean SVCCA scores) across algorithms

From the above plot, we observe that the SVCCA scores for PCA with ICA (and vice-versa) are very high across latent dimensions. The SVCCA scores for VAE and NMF with PCA and ICA increase sharply initially, peak at latent dimension 3, drop slightly thereafter, but subsequently remain quite high. This result is in agreement with what has been observed in the earlier section. The SVCCA scores for DAE with other models increase initially and then the scores drop for higher dimensions. Also, the SVCCA scores of DAE with other models is relatively lower (in the 0.6-0.8 range) as compared to the other model combinations (in the 0.8 - 1.0 range). This is again indicative of the relatively lower stability of the DAE model.

T-tests Analysis

Tailed T-tests were performed by comparing the weights for PCOS vs Control samples to identify the gene ID, latent dimensionality and model combination which gives maximum separation between the two sets.

Model	Gene ID	Latent Dimension	$-\log_{10}(P)$	Tissues most expressed in [16]
PCA	148753	30	5.665	Testis, Adrenal, Gall bladder
ICA	2350	40	4.677	Placenta, Gall bladder, Urinary bladder
NMF	6277	60	2.450	Colon, Lung, Stomach
DAE	10327	2	1.747	Kidney, Duodenum, Small intestine
VAE	4170	5	1.879	Bone marrow, Gall bladder, Appendix

Table 2. Best Separation Between PCOS and Control Samples Across Algorithms and Latent Dimensions

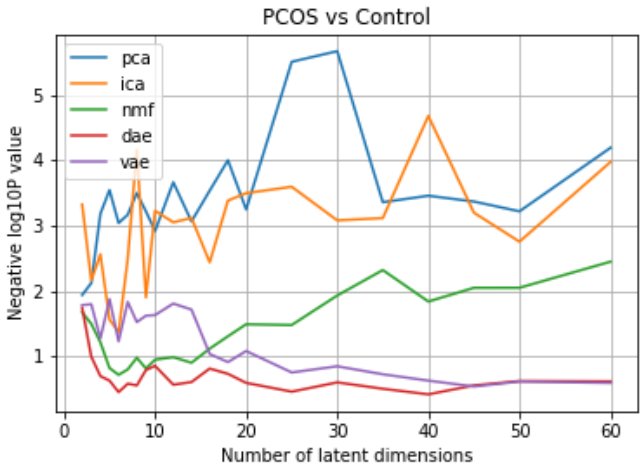

Fig 14. T-test results for PCOS vs Control across algorithms

From the above table and figure, we can observe that PCOS vs Control samples were best separated by PCA, with respect to gene ID 148753, in latent dimension 30, with over-expression of the gene in the PCOS samples. This gene ID corresponds to Family with sequence similarity 163 member A [16]. This gene has a biased expression in human tissues such as testis, adrenal and gall bladder.

PCOS vs Control samples were also well-separated by ICA, with respect to gene ID 2350, in latent dimension 40. This gene ID corresponds to folate receptor beta [16]. This gene has a broad expression in human tissues such as placenta, gall bladder and urinary bladder.

For the PCA, ICA and NMF models, we observe a general increase in the $-\log_{10}(P)$ values (with fluctuations) as the number of latent dimensions increase. However, for the DAE and VAE, we can see an overall decreasing trend in the $-\log_{10}(P)$ values as the number of latent dimensions increase. This indicates that PCA, ICA and NMF are able to better distinguish between PCOS and Control samples at relatively higher latent dimensionalities, but DAE and VAE are able to capture this difference better at lower latent dimensionalities.

Discussion

From the results obtained above, it is interesting to note that our model is able to identify the following genes - FAM163A, FOLR2, S100A6, AKR1A1 and MCL1, that correspond to the latent dimensions that maximally separate the PCOS data points from the control data points. Out of these 5 genes, we got literature evidence for the role of the following genes - FOLR2, S100A6 in PCOS.

The gene FOLR2 (Gene ID 2350), is known to code for the protein - folate receptor beta. The folate receptor beta protein is known to play a key role in the pathways associated with vitamin digestion and absorption. [19], [20], [21] Perturbations of this pathway, results in low folate production, affecting the utilization of Vitamin B12 and other B vitamins. This in turn affects the extent of the maturation egg released during ovulation. Incomplete maturation, prevents the ovaries from releasing adequate amount of progesterone that is essential in maintaining normal menstrual cycle in women.

Additionally, vitamin B12 deficiency is also associated with insulin resistance and obesity. [22], [23], [24]. The high levels of insulin in the blood results in enhanced production of androgen by the theca cells. This starts a cascade of actions affecting the progesterone levels, extent of maturation of the eggs. [25], [26], [27]. This role of folic acid in insulin resistance and PCOS has been further ratified here [28].

The gene S100A6 (Gene ID 6277) is also known to code for the S100 Calcium binding protein. This protein is known to play a key role in the stimulation of calcium dependent insulin release and prolactin secretion. [19], [20], [21] They are also known to participate in the estrous cycle. [29]

Conclusion

We analysed gene expression data of PCOS patients as well as non-PCOS controls using unsupervised dimensionality reduction techniques such as PCA, ICA, NMF, DAE and VAE, across a range of different latent dimensions. We found that the different models in different latent dimensional representations were able to extract interesting biological patterns about the pathophysiology of PCOS, such as the folate receptor gene and the S100 Calcium binding protein. Hence, we concluded that, in order to derive new and meaningful insights into biological data from gene expression analysis, it is essential to explore different dimensionality reduction algorithms as well as latent dimensions.

Challenges

There were very few PCOS gene expression datasets available on GEO, with each dataset comprising only about 10-30 samples. Moreover, many gene IDs in these datasets did not have a mapping to an Entrez ID and hence had to be discarded. Hence, the final combined dataset which we obtained after data pre-processing was rather small. This limited our analyses in terms of the number of latent dimensional representations that we could explore.

Future Work

In this project, we used only one kind of architecture for the autoencoder models (DAE, VAE). In future, different architectures or other optimizations for the DAE, VAE models could be explored.

Acknowledgments

We would like to thank Professor Manikandan Srinivasan for imparting knowledge on different algorithms, their applications and giving us the opportunity to work on this project. We would also like to thank the TAs of the course for providing constructive feedback on our initial report and guide us towards performing a more well-rounded analysis.

Author Contributions

Lakshmy: Performed stability and SVCCA analysis across all the models and across all latent dimensions. Performed T-tests comparing PCOS vs Control samples to identify the gene/feature, latent dimensionality and model combination which gives maximum separation between the two sets.

Sowmya: Obtained gene expression datasets and annotations. Performed data pre-processing steps and carried out heterogeneous network analysis. Performed gene expression data analysis on individual datasets and combined dataset using PCA, ICA, NMF and the autoencoder models - DAE and VAE. Analyzed the reconstruction cost and strongly associated latent dimensions across all the models. Worked on the literature survey that ratifies the findings of our model.

References

1. Bartenhagen C, Klein HU, Ruckert C, Jiang X, Dugas M. Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. BMC bioinformatics. 2010;11(1):567.
2. Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. Bioinformatics. 2001;17(9):763–774.
3. Way GP, Zietz M, Rubinetti V, Himmelstein DS, Greene CS. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. Genome Biology. 2020;21(1):1–27.
4. Tan J, Hammond JH, Hogan DA, Greene CS. Adage-based integration of publicly available pseudomonas aeruginosa gene expression data with denoising autoencoders illuminates microbe-host interactions. MSystems. 2016;1(1).
5. Tan J, Doing G, Lewis KA, Price CE, Chen KM, Cady KC, et al. Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks. Cell systems. 2017;5(1):63–71.
6. Simidjievski N, Bodnar C, Tariq I, Scherer P, Andres Terre H, Shams Z, et al. Variational autoencoders for cancer data integration: design principles and computational practice. Frontiers in genetics. 2019;10:1205.
7. Way GP, Greene CS. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. BioRxiv. 2017; p. 174474.
8. Azziz R, Carmina E, Chen Z, Dunaif A, Laven JS, Legro RS, et al. Polycystic ovary syndrome. Nature reviews Disease primers. 2016;2(1):1–18.
9. Norman RJ, Dewailly D, Legro RS, Hickey TE. Polycystic ovary syndrome. The Lancet. 2007;370(9588):685–697.

10. Gilling-Smith C, Willis DS, Beard RW, Franks S. Hypersecretion of androstenedione by isolated thecal cells from polycystic ovaries. The Journal of Clinical Endocrinology & Metabolism. 1994;79(4):1158–1165.

11. Jansen E, Laven JS, Dommerholt HB, Polman J, van Rijt C, van den Hurk C, et al. Abnormal gene expression profiles in human ovaries from polycystic ovary syndrome patients. Molecular Endocrinology. 2004;18(12):3050–3063.

12. Cortón M, Botella-Carretero JI, Benguria A, Villuendas G, Zaballos A, San Millán JL, et al. Differential gene expression profile in omental adipose tissue in women with polycystic ovary syndrome. The Journal of Clinical Endocrinology & Metabolism. 2007;92(1):328–337.

13. Kaur S, Archer KJ, Devi MG, Kriplani A, Strauss III JF, Singh R. Differential gene expression in granulosa cells from polycystic ovary syndrome patients with and without insulin resistance: identification of susceptibility gene sets through network analysis. The Journal of Clinical Endocrinology & Metabolism. 2012;97(10):E2016–E2021.

14. Haouzi D, Assou S, Monzo C, Vincens C, Dechaud H, Hamamah S. Altered gene expression profile in cumulus cells of mature MII oocytes from patients with polycystic ovary syndrome. Human reproduction. 2012;27(12):3523–3530.

15. Fung JN, Mortlock S, Girling JE, Holdsworth-Carson SJ, Teh WT, Zhu Z, et al. Genetic regulation of disease risk and endometrial gene expression highlights potential target genes for endometriosis and polycystic ovarian syndrome. Scientific reports. 2018;8(1):1–19.

16. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic acids research. 2012;41(D1):D991–D995.

17. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. Cell systems. 2015;1(6):417–425.

18. Raghu M, Gilmer J, Yosinski J, Sohl-Dickstein J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In: Advances in neural information processing systems; 2017. p. 6076–6085.

19. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000;28(1):27–30.

20. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Science. 2019;28(11):1947–1951.

21. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: Integrating viruses and cellular organisms. Nucleic Acids Research. 2020;.

22. Valdés-Ramos R, Ana Laura GL, Beatriz Elina MC, Alejandra Donaji BA. Vitamins and type 2 diabetes mellitus. Endocrine, Metabolic & Immune Disorders-Drug Targets (Formerly Current Drug Targets-Immune, Endocrine & Metabolic Disorders). 2015;15(1):54–63.

23. Cigerli O, Parildar H, Unal AD, Tarcin O, Kut A, Eroglu H, et al. Vitamin deficiency and insulin resistance in nondiabetic obese patients. Acta Endocrinologica (Bucharest). 2016;12(3):319.

24. Kaya C, Cengiz SD, Satıroğlu H. Obesity and insulin resistance associated with lower plasma vitamin B12 in PCOS. Reproductive biomedicine online. 2009;19(5):721–726.

25. Nestler JE. Insulin regulation of human ovarian androgens. Human Reproduction. 1997;12(suppl_1):53–62.

26. Diamanti-Kandarakis E, Pappalou O, Kandaraki EA. The Role of Androgen Excess on Insulin Sensitivity in Women. In: Hyperandrogenism in Women. vol. 53. Karger Publishers; 2019. p. 50–64.

27. Baptiste CG, Battista MC, Trottier A, Baillargeon JP. Insulin and hyperandrogenism in women with polycystic ovary syndrome. The Journal of steroid biochemistry and molecular biology. 2010;122(1-3):42–52.

28. Regidor PA, Schindler AE, Lesoine B, Druckman R. Management of women with PCOS using myo-inositol and folic acid. New clinical data and review of the literature. Hormone molecular biology and clinical investigation. 2018;34(2).

29. Hong EJ, Park SH, Choi KC, Leung PC, Jeung EB. Identification of estrogen-regulated genes by microarray analysis of the uterus of immature rats exposed to endocrine disrupting chemicals. Reproductive Biology and Endocrinology. 2006;4(1):49.