

PORTRAIT

Holistic Data Visualization using Neural Networks

CHAYAN MAITRA



MACHINE INTELLIGENCE UNIT, INDIAN STATISTICAL INSTITUTE,
KOLKATA-700108, WEST BENGAL, INDIA

INDIAN STATISTICAL INSTITUTE

DOCTORAL THESIS

P O R T R A I T
Holistic Data Visualization using Neural Networks

CHAYAN MAITRA

under the supervision of
Prof. Rajat K. De

*A thesis submitted to the Indian Statistical Institute
in partial fulfillment of the requirements for
the degree of
Doctor of Philosophy (Computer Science)*

March 2026

To my family

ACKNOWLEDGEMENTS

Completing this doctoral journey has been one of the most challenging yet fulfilling experiences of my life. It would not have been possible without the constant support, encouragement, and guidance I received from my family, mentors, collaborators, seniors, colleagues, and friends. I take this opportunity to express my deepest gratitude to everyone who has contributed, directly or indirectly, to the realization of this work.

*First and foremost, I express my heartfelt gratitude to my family, who have been my strongest pillars of strength throughout this journey. My father, **Ram Mohan Maitra**, has always been a source of inspiration and motivation. My mother, **Kshama Maitra**, has been my greatest emotional support. Their unconditional love and constant encouragement gave me the confidence to overcome every obstacle that came my way. I am equally grateful to my elder sister, **Ratna Acharya**, who has always been my guiding light. Her words of encouragement and faith in my abilities have constantly inspired me to do better. A very special mention goes to **Anweshha Sengupta**, my life partner, whose unwavering support has been instrumental in completing this journey. Beyond being my emotional support, she has also been my collaborator, co-author, and critic, contributing intellectually to one of my works. I feel truly fortunate to have her as both my partner in life and in research. I also wish to express my deepest love and gratitude to my daughter, **Ahiri Maitra**, whose innocent smile and boundless joy have been my greatest source of happiness and motivation.*

*I owe my deepest respect and gratitude to my supervisor, **Prof. Rajat K. De**, whose mentorship has been invaluable. His deep insights, intellectual rigor, and constructive criticism have shaped my understanding of research. He gave me the freedom to explore ideas independently, while always being available to guide me in the right direction. His encouragement and patience allowed me to grow as a researcher and to think critically about every problem I encountered. I am truly indebted to him for his constant support and guidance throughout my doctoral studies.*

*I am especially thankful to **Dr. Dibyendu B. Seal**, my co-author and mentor in every sense of the word. He introduced me to the world of programming and guided me from the very basics to an advanced level. Much of what I have achieved in terms of technical expertise is a result of his patient teaching and constant encouragement. His belief in my abilities has always been a source of confidence. Beyond being a collaborator, he has been a constant motivator and well-wisher, standing by me in every situation. I consider his support one of the most important factors behind the successful completion of my Ph.D.*

*I also extend my gratitude to **Dr. Vivek Das**, my co-author, who has been a continuous source of motivation. His insightful discussions, positive attitude, and enthusiasm for research have greatly enriched my experience. His constructive feedback and encouragement helped me stay inspired, especially during challenging phases of my research.*

*My Ph.D. journey would not have been as fulfilling and enjoyable without the wonderful colleagues and friends I have had the privilege to work with. I express my sincere gratitude to my seniors, **Prasun Dutta**, **Subhadip Boral**, and **Ankita Dutta**, who guided me during the early and most uncertain phases of my Ph.D. Their constant advice, practical suggestions, and personal guidance made it easier for me to navigate through the complexities of academic research. Their experience and willingness to help have been invaluable.*

*My heartfelt thanks go to **Surochita Pal**, whose consistent technical support proved extremely valuable, particularly when coding became a crucial component of my work. Her readiness to help and share knowledge has made several difficult tasks manageable.*

*I would like to acknowledge my close friends and companions — **Sayan Saha**, **Sourav Basak**, **Kunal Biswas**, **Shramana Dey**, **Pallabi Dutta**, **Soumi Sarkar** and **Riddhasree Bhattacharya**. Their friendship, cooperation, and constant presence have made this long and demanding journey enjoyable. We have shared countless discussions, laughter, and memories, both academic and non-academic, which have made my Ph.D. life truly memorable. Their support during stressful times kept me motivated and grounded.*

*My sincere appreciation goes to my juniors — **Koustab Ghosh**, **Subhajit Patra**, **Sourav Ghosh**, and **Santu Mondal**. Their enthusiasm for learning and dedication towards research have been truly inspiring. I have had several mathematical and conceptual discussions with them, and those interactions have been mutually enriching. They not only motivated me to think deeper but also helped me refine many of my ideas.*

*Lastly, I would like to thank **Deepak Kumar** and **Subham Ghosh** for their constant help with technical and system-related support. Their timely assistance ensured that my work proceeded smoothly without unnecessary interruptions. Their cooperation was indispensable, particularly during crucial stages of coding and experimentation.*

This thesis is a culmination of years of effort, guidance, and collaboration from many wonderful individuals. To every person who has, in one way or another, contributed to this journey — whether by offering academic advice, moral support, or companionship — I offer my deepest gratitude. Every interaction, discussion, and shared moment has shaped me into the researcher and person I am today.

To all of you — thank you.

ABSTRACT

With the exponential growth of complex data across domains, effective visualization has become increasingly crucial for understanding relationships hidden within high-dimensional spaces. However, existing visualization techniques often struggle to effectively capture and represent such high-dimensional data. Motivated by this challenge, we have developed NeuroDAVIS, a neural network model designed to visualize high-dimensional data by extracting meaningful latent representations through deep feature extraction. While NeuroDAVIS has successfully addressed the visualization aspect, we have soon recognized the necessity of identifying the most relevant features that contribute to the visualization and downstream analysis. To address this issue, we have extended our framework and introduced NeuroDAVIS-FS, a feature selection model built upon the NeuroDAVIS architecture. This extension not only visualizes but also selects the most informative features from complex datasets.

As data collection techniques have evolved, it has become increasingly common to encounter datasets that capture multiple aspects—or modalities—of the same set of samples. Visualizing and understanding such multi-modal datasets require a more advanced approach that is capable of integrating and producing a joint embedding of these heterogeneous data sources. Inspired by this need, we have enhanced our model to handle multiple modalities, resulting in NeuroMDAVIS, an extended version of NeuroDAVIS, capable of joint visualization across modalities. In addition, we have also developed a feature selection framework, called NeuroMDAVIS-FS, that operates on multi-modal data, enabling the identification of key features across different data types. These key features have been found to be efficient in predicting patient survival. These key features have been found to be effective in predicting the survival of lung cancer patients.

Despite these advancements, another significant limitation remained—while the learned embeddings provided meaningful visualization and feature selection, they did not have generalization capabilities. Some regions of the embedding space remained sparse or unused, limiting the model’s ability to generate new, realistic samples. In order to overcome this issue, we have further extended our approach to a generative framework, giving rise to G-NeuroDAVIS. This generative variant produces a smooth, well-distributed latent embedding that is capable of not only capturing the complete data distribution but also generating high-quality synthetic samples. Through this progression—from NeuroDAVIS to NeuroDAVIS-FS, NeuroMDAVIS, NeuroMDAVIS-FS and finally to G-NeuroDAVIS—we present a comprehensive family of models that together address visualization, feature selection, multi-modal integration, joint visualization, and data generation in a unified framework.

Contents

Dedicate	ii
Acknowledgment	iii
Abstract	v
List of Figures	x
List of Tables	xxi
List of Publications	xxii
1 Introduction	1
1.1 Introduction	1
1.2 Dimension Reduction	2
1.2.1 Feature extraction and Feature selection	3
1.3 Data visualization	4
1.3.1 Methods for data visualization	5
t-SNE	5
Fit-SNE	6
UMAP	6
PHATE	7
IVIS	7
1.4 Integration and joint visualization	8
1.4.1 Methods for integration	8
1.5 Data Generation	9
1.5.1 Methods for data generation	10
1.6 Motivation of the thesis	10
1.7 Scope of the thesis	12
1.7.1 Chapter 2 - NeuroDAVIS: A neural network model for data visualization [1]	12
1.7.2 Chapter 3 - NeuroDAVIS-FS: Feature Selection Through Visualization Using NeuroDAVIS [2]	13
1.7.3 Chapter 4 - NeuroMDAVIS: Visualization of single-cell multi-omics data under deep learning framework [3]	13
1.7.4 Chapter 5 - NeuroMDAVIS-FS Identifies Prognostic Molecular Signatures in Lung Cancer towards Survival Analysis and Risk Stratification [4]	14

1.7.5	Chapter 6 - G-NeuroDAVIS: A generative model for data visualization through a generalized embedding [5]	14
1.7.6	Chapter 7 - Conclusions and Scope for Further Research	15
2	NeuroDAVIS: A neural network model for data visualization	17
2.1	Introduction	17
2.2	Methodology	18
2.2.1	Motivation	18
2.2.2	Architecture	21
2.2.3	Forward propagation	21
2.2.4	Learning	23
2.2.5	Projecting new observations	24
2.3	Proof of correctness	25
2.4	Results	27
2.4.1	Performance evaluation on synthetic 2D datasets	28
	Embedding, visualization and structure preservation	28
	Global structure preservation	29
	Inter-cluster distance preservation	32
	Cluster size preservation	33
2.4.2	Performance evaluation on high-dimensional datasets	35
	Numeric datasets	36
	Textual dataset	37
	Image datasets	38
	Biological datasets	42
2.5	Discussion and Conclusion	50
3	NeuroDAVIS-FS: Feature selection through visualization using NeuroDAVIS	53
3.1	Introduction	53
3.2	Problem statement and solution approach	54
3.2.1	Problem scenario	54
3.2.2	Solution approach	56
	Data description	56
	Methodology	57
3.3	Results	58
3.3.1	Comparison with the state-of-the-art method	59
	Results on <i>Breast Cancer</i> data	59
	Results on <i>Wine</i> data	60
	Results on <i>Digits</i> data	61
3.3.2	Case studies	62
	Results on Image datasets	63
	Results on Biological dataset	65
3.4	Discussion and Conclusion	65
4	NeuroMDAVIS: Visualization of single-cell multi-omics data under deep learning framework	69
4.1	Introduction	69

4.2	Methods	71
4.2.1	Architecture	71
4.2.2	Forward propagation	72
4.2.3	Learning	75
4.2.4	Projection of new observations	76
4.3	Results	77
4.3.1	Visualizing multi-omics data using NeuroMDAVIS	78
	Visualizing CITE-seq data	78
	Visualizing multiome data	82
4.3.2	Comparison against separate and shared visualization	83
4.3.3	Projection of new observations	86
4.3.4	Sensitivity analysis	87
4.4	Discussion and Conclusion	89
5	NeuroMDAVIS-FS Identifies Prognostic Molecular Signatures in Lung Cancer towards Survival Analysis and Risk Stratification	91
5.1	Introduction	91
5.2	Methodology	94
5.2.1	Data acquisition and pre-processing	94
5.2.2	NeuroMDAVIS-FS	95
5.2.3	Experimental Design	99
	Survival Analysis	99
	Risk Stratification	100
	Comparison with baseline model	100
5.3	Results	100
5.3.1	Feature Selection	102
5.3.2	Survival Analysis	102
5.3.3	Risk Stratification	103
5.3.4	Comparison with baseline	105
5.4	Discussion	106
5.4.1	Subtype-Specific Drivers and Tumor Suppressors	107
5.4.2	Metastasis and the Immune Microenvironment	107
5.4.3	Therapeutic Resistance and Prognostic Gain	108
5.5	Conclusion	108
6	G-NeuroDAVIS: A Generative Model for Data Visualization through a Generalized Embedding	111
6.1	Introduction	111
6.2	Methodology	113
6.2.1	Problem scenario and solution approach	113
	Problem scenario	113
	Solution approach	115
6.2.2	G-NeuroDAVIS: Architecture, Forward Propagation and Learning	115
	Architecture	115
	Forward propagation	117
	Learning	119
6.2.3	Conditional data generation in a supervised setting	121

6.3	Results	122
6.3.1	Data description	123
6.3.2	Result Analyses	123
	G-NeuroDAVIS in an unsupervised setting	124
	Sample Generation	134
	G-NeuroDAVIS in a supervised setting	137
	Complexity analyses	138
	Sensitivity analyses	139
	A case study on Celeb dataset	140
6.4	Discussion and Conclusion	141
7	Conclusions and Scope for Further Research	143
7.1	Research Contribution	143
7.1.1	Architecture	143
7.1.2	Results	145
7.2	Strengths	147
7.3	Limitations	148
7.4	Future Scope	149
	Appendices	151
A	NeuroDAVIS	152
A.I	Results on synthetic 2D datasets	152
A.II	Results on high-dimensional datasets	154
A.III	Downstream Analysis	166
A.IV	Projection of test samples	172
B	NeuroMDAVIS-FS	173
B.I	Results on LSCC cohort	173
B.II	Results on LUAD cohort	175
C	G-NeuroDAVIS	178
C.I	Embedding mapping	178
C.II	Sub-sample based analyses	180
C.III	G-NeuroDAVIS in a supervised setting	184
	Bibliography	185

List of Figures

2.1	Block diagrams of (A) Multi-layer neural network and (B) NeuroDAVIS. The dashed part in (B) represents the modified architecture that enables the regressors to get updated.	20
2.2	The network architecture of NeuroDAVIS, along with the mathematical notations.	22
2.3	The original distributions and the embeddings produced by NeuroDAVIS, t-SNE and UMAP for <i>Elliptic Ring</i> ((A)-(D)), <i>Olympic</i> ((E)-(H)), <i>Spiral</i> ((I)-(L)) and <i>Shape</i> ((M)-(P)) datasets.	30
2.4	Spearman rank correlation between pairwise distances in the original distribution, and pairwise distances in t-SNE, UMAP and NeuroDAVIS-produced embeddings of the 2D synthetic datasets <i>Elliptic Ring</i> , <i>Olympic</i> , <i>Spiral</i> and <i>Shape</i> . For <i>Elliptic Ring</i> dataset, median correlation coefficient values obtained are 0.98 (NeuroDAVIS), 0.27 (t-SNE) and 0.32 (UMAP). For <i>Olympic</i> dataset, median correlation coefficient values obtained are 0.96 (NeuroDAVIS), 0.61 (t-SNE) and 0.73 (UMAP). For <i>Spiral</i> dataset, median correlation coefficient values obtained are 0.98 (NeuroDAVIS), 0.96 (t-SNE) and -0.09 (UMAP). For <i>Shape</i> dataset, median correlation coefficient values obtained are 0.97 (NeuroDAVIS), 0.57 (t-SNE) and 0.40 (UMAP). The corresponding p – values obtained using the Mann-Whitney U test are 0.0001 (both NeuroDAVIS-t-SNE and NeuroDAVIS-UMAP) for <i>Elliptic Ring</i> , <i>Olympic</i> and <i>Shape</i> datasets. For <i>Spiral</i> dataset, p – values obtained are 0.009 (NeuroDAVIS-t-SNE) and 0.001 (NeuroDAVIS-UMAP).	31
2.5	The original distribution and the embeddings produced by NeuroDAVIS, t-SNE and UMAP for <i>World Map</i> dataset.	32
2.6	Spearman rank correlation between original distances among cluster centroids in the original dataset and distances between cluster centroids in the NeuroDAVIS, t-SNE and UMAP-generated embedding. The median correlation coefficient values are 0.96 (NeuroDAVIS), 0.59 (t-SNE) and 0.55 (UMAP) for <i>World Map</i> dataset. The corresponding p – value obtained using Mann-Whitney U test is 0.0001 (both NeuroDAVIS-t-SNE and NeuroDAVIS-UMAP).	33
2.7	Preservation of pairwise distances between clusters (continents) in <i>World Map</i> dataset by NeuroDAVIS, t-SNE and UMAP.	34
2.8	Embeddings produced by NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP for <i>Breast Cancer</i> ((A)-(F)), <i>Wine</i> ((G)-(L)), <i>Spam</i> ((M)-(R)) and <i>Coil20</i> ((S)-(X)) datasets	39

2.9	Spearman rank correlation coefficient between pairwise distances in original distribution, and pairwise distances in NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP embeddings of <i>Breast Cancer</i> , <i>Wine</i> , <i>Spam</i> and <i>Coil20</i> datasets. For <i>Breast Cancer</i> dataset, median correlation coefficient values obtained are 0.94 (NeuroDAVIS), 0.76 (t-SNE), 0.77 (UMAP), 0.77 (Fit-SNE), 0.99 (PCA) and 0.98 (RP). For <i>Wine</i> dataset, median correlation coefficient values obtained are 0.92 (NeuroDAVIS), 0.91 (t-SNE), 0.82 (UMAP), 0.92 (Fit-SNE), 0.99 (PCA) and 0.99 (RP). For <i>Spam</i> dataset, median correlation coefficient values obtained are 0.77 (NeuroDAVIS), 0.59 (t-SNE), 0.57 (UMAP), 0.61 (Fit-SNE), 0.86 (PCA) and 0.54 (RP), while for <i>Coil20</i> dataset, these values are 0.61 (NeuroDAVIS), 0.53 (t-SNE), 0.15 (UMAP), 0.56 (Fit-SNE), 0.78 (PCA) and 0.28 (RP). The corresponding p – values obtained using the Mann-Whitney U test are 0.0001 (both NeuroDAVIS-t-SNE and NeuroDAVIS-UMAP) for both <i>Breast Cancer</i> and <i>Spam</i> datasets. For <i>Wine</i> dataset, p – values obtained using Mann-Whitney U test are 0.73 (NeuroDAVIS-t-SNE), 0.002 (NeuroDAVIS-UMAP) and 1.00 (NeuroDAVIS-Fit-SNE), while for <i>Coil20</i> dataset, p – values obtained using Mann-Whitney U test are 0.0007 (NeuroDAVIS-t-SNE), 0.0001 (NeuroDAVIS-UMAP) and 0.002 (NeuroDAVIS-Fit-SNE).	40
2.10	(A)NeuroDAVIS (C) t-SNE (E) UMAP (G) Fit-SNE embeddings obtained on <i>FMNIST</i> dataset, (B) NeuroDAVIS (D) t-SNE (F) UMAP (H) Fit-SNE embeddings obtained on the first 50 PCA components of the original <i>FMNIST</i> dataset.	41
2.11	Spearman rank correlation coefficients between pairwise distances of the cluster centroids in high-dimension and the pairwise distances of the cluster centroids in the low-dimensional embedding of <i>FMNIST</i> dataset produced by NeuroDAVIS, t-SNE, UMAP, Fit-SNE and RP with and without PCA-based preprocessing.	42
2.12	2-dimensional embeddings of <i>Usoskin</i> data generated by NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PHATE, IVIS, PCA and RP respectively. . . .	43
2.13	2-dimensional embeddings of <i>Jurkat</i> data generated by NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PHATE, IVIS, PCA and RP respectively.	45

2.14	Spearman rank correlation between pairwise distances in original distribution and pairwise distances in NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PHATE, IVIS, PCA and RP-produced embeddings of <i>Usoskin</i> and <i>Jurkat</i> datasets. For <i>Usoskin</i> dataset, median correlation coefficient values obtained are 0.24 (NeuroDAVIS), 0.27 (t-SNE), 0.02 (UMAP), 0.29 (Fit-SNE), -0.12 (PHATE), 0.002 (IVIS), -0.05 (PCA) and 0.22 (RP). For <i>Jurkat</i> dataset, median correlation coefficient values obtained are 0.42 (NeuroDAVIS), 0.29 (t-SNE), 0.26 (UMAP), 0.29 (Fit-SNE), 0.23 (PHATE), 0.25 (IVIS), 0.32 (PCA) and 0.23 (RP). The corresponding p – values obtained using the Mann-Whitney U test for <i>Usoskin</i> dataset are 0.677 (NeuroDAVIS-t-SNE), 0.472 (NeuroDAVIS-Fit-SNE), 0.0001 (NeuroDAVIS-UMAP, NeuroDAVIS-PHATE, NeuroDAVIS-IVIS and NeuroDAVIS-PCA), and 0.0001 (NeuroDAVIS-t-SNE, NeuroDAVIS-UMAP, NeuroDAVIS-Fit-SNE, NeuroDAVIS-PHATE, NeuroDAVIS-IVIS, NeuroDAVIS-PCA and NeuroDAVIS-RP) for <i>Jurkat</i> dataset.	46
2.15	(A) and (B) show the results of performance comparison of NeuroDAVIS with that of t-SNE, UMAP, IVIS and Fit-SNE in terms of distance preservation capability and correlation between pairwise distances in high dimension to that in low dimension of <i>Jurkat</i> and <i>Usoskin</i> datasets, respectively. Distances in high dimension have been measured and distributed across 50 equal-width bins in sorted order. The corresponding distances in low dimensions have also been mapped into the same bins. The leading boxplots represent short distances, while the trailing ones represent long distances. The lower the spread of the boxplot, the better the preservation of distances. The median of the trailing boxplots representing long distances should be higher than that of the leading boxplots representing short distances. It signifies that the order of pairwise distances among observations in high dimension is similar to that among observations in low dimension. The last histogram for each dataset represents the count of distances in each bin. (C) and (D) show results of comparison of local shape preservation capability of NeuroDAVIS with that of t-SNE, UMAP, IVIS and Fit-SNE in terms of trustworthiness score on <i>Jurkat</i> and <i>Usoskin</i> datasets, respectively. (E) and (F) show results of comparison of robustness of NeuroDAVIS with that of t-SNE, UMAP, IVIS, and Fit-SNE in terms of subsampling-based correlation analysis on <i>Jurkat</i> and <i>Usoskin</i> datasets, respectively.	47
2.16	(A) 2-dimensional embeddings produced by NeuroDAVIS, t-SNE, UMAP, IVIS and Fit-SNE for the Mass Cytometry <i>Wong</i> data, coloured by cell-type; <i>MAIT</i> denotes mucosal-associated invariant <i>T</i> cells; <i>ILC</i> stands for innate lymphoid cells; and Tgd represents $\gamma\delta T$ cells. (B) Expression of <i>T</i> cell marker proteins <i>CD69</i> , <i>CD103</i> , <i>CD45RO</i> , <i>CCR7</i> on the NeuroDAVIS, t-SNE, UMAP, IVIS and Fit-SNE-generated embeddings.	48

3.1	For <i>Breast Cancer</i> dataset (A) The 2-dimensional embedding produced by NeuroDAVIS (ND). (B) Overlap of features between NeuroDAVIS-FS (NDFS) and the state-of-the-art. (C) Comparison of the developed models with state-of-the-art methods based on classification and clustering performance. For classification, K-nearest neighbor (k-NN) and Random Forest (RF) models are utilized and for evaluation of clustering performance, KMeans and Agglomerative clustering techniques are used.	59
3.2	For <i>Wine</i> dataset (A) The 2-dimensional embedding produced by NeuroDAVIS (ND). (B) Overlap of features between NeuroDAVIS-FS (NDFS) and the state-of-the-art methods. (C) Comparison of the developed models with the state-of-the-art based on classification and clustering performance. For classification, K-nearest neighbor (k-NN) and Random Forest (RF) models are utilized and for evaluation of clustering performance, KMeans and Agglomerative clustering techniques are used.	61
3.3	For <i>Digits</i> dataset (A) The 2-dimensional embedding produced by NeuroDAVIS (ND). (B) Overlap of features between NeuroDAVIS-FS (NDFS) and the state-of-the-art methods. (C) Comparison of the developed models with the state-of-the-art based on classification and clustering performance. For classification, K-nearest neighbor (k-NN) and Random Forest (RF) models are utilized and for evaluation of clustering performance, KMeans and Agglomerative clustering techniques are used.	62
3.4	Visual representation of randomly selected samples (RS) through the top features selected by NDFS for <i>Coil20</i> dataset. The classification performance of k-NN and RF has been shown over the selected feature space.	63
3.5	Visual representation of randomly selected samples (RS) through the top features selected by NDFS for <i>Olivetti Faces</i> dataset. The classification performance of k-NN and RF has been shown over the selected feature space.	64
3.6	Visual representation of randomly selected samples through the worst 5% features selected by NDFS for <i>Coil20</i> (upper) and <i>Olivetti Faces</i> (lower) datasets.	64
3.7	Visual representation of top 25 features selected by NDFS for <i>Jurkat</i> dataset (left). The classification performance of k-NN and RF has been shown over the selected feature spaces (right).	65
4.1	(A) A graphical abstract reflecting the overall workflow of this chapter. (B) The architecture of NeuroMDAVIS, that have been developed for multi-omics data visualization.	72
4.2	The network architecture of NeuroMDAVIS, along with the mathematical notations.	75
4.3	2-dimensional embeddings produced by NeuroMDAVIS, t-SNE, PCA + t-SNE, UMAP, PCA + UMAP, Fit-SNE, PCA + Fit-SNE, and IVIS for <i>bmcite30k</i> (first row), and <i>kotliarov50k</i> (Second row) datasets.	79

4.4	(A) and (B) show classification performance on low-dimensional embeddings of <i>bmcite30k</i> dataset generated using NeuroMDAVIS, t-SNE, PCA + t-SNE, UMAP, PCA + UMAP, Fit-SNE, PCA + Fit-SNE, and IVIS, in terms of accuracy, precision, recall and F1-score using k -NN and Random Forest classifiers respectively. (C) and (D) show the same on low-dimensional embeddings of <i>kotliarov50k</i> dataset.	80
4.5	NeuroMDAVIS-produced sub embeddings (based on cell types), colored by high/low influenza vaccine responder (left). Area-Under-the-Curve (AUC)-score for high/low influenza vaccine responder classification obtained using k -NN classifier (with $k = 50$) on the NeuroMDAVIS, t-SNE, UMAP, IVIS and Fit-SNE projections of <i>kotliarov50k</i> dataset (right).	81
4.6	Clustering performance on low-dimensional embeddings generated using NeuroMDAVIS, t-SNE, PCA + t-SNE, UMAP, PCA + UMAP, Fit-SNE, PCA + Fit-SNE, and IVIS, with respect to ARI, NMI and FMI scores of (A) <i>bmcite30k</i> , and (B) <i>kotliarov50k</i> datasets, using k -means clustering algorithm.	81
4.7	2-dimensional embeddings produced by NeuroMDAVIS, t-SNE, PCA + t-SNE, UMAP, PCA + UMAP, Fit-SNE, PCA + Fit-SNE, and IVIS for <i>pbmc10k</i> dataset.	83
4.8	(A) and (B) show classification performance on projections of <i>pbmc10k</i> dataset generated by NeuroMDAVIS, t-SNE, PCA + t-SNE, UMAP, PCA + UMAP, Fit-SNE, PCA + Fit-SNE, and IVIS in terms of accuracy, precision, recall and F1-score using k -NN and Random Forest classifiers. (C) k -means clustering performance on embeddings of <i>pbmc10k</i> dataset produced by NeuroMDAVIS, t-SNE, PCA + t-SNE, UMAP, PCA + UMAP, Fit-SNE, PCA + Fit-SNE, and IVIS, in terms of ARI, NMI and FMI scores.	84
4.9	2-dimensional joint embedding produced by NeuroMDAVIS, separate embeddings produced by NeuroDAVIS, and shared embeddings produced by MultiMAP for <i>bmcite30k</i> (first row), and <i>pbmc10k</i> (Second row) datasets.	85
4.10	Classification and clustering performance on projections generated by NeuroMDAVIS, NeuroDAVIS, and MultiMAP, in terms of accuracy, precision, recall and F1-score, and ARI, NMI, and FMI using k -NN classifiers and k -means clustering, respectively, of (A) <i>bmcite30k</i> and (B) <i>pbmc10k</i> datasets.	85
4.11	(A), (B) and (C) show 2D embeddings produced by NeuroMDAVIS on both the training and test datasets for <i>bmcite30k</i> , <i>kotliarov50k</i> and <i>pbmc10k</i> respectively.	86
4.12	Classification and clustering performance on projections generated by NeuroMDAVIS over the train and test parts of <i>bmcite30k</i> , <i>kotliarov50k</i> , and <i>pbmc10k</i> datasets, in terms of accuracy, precision, recall and F1-score, and ARI, NMI, and FMI using k -NN classifiers and k -means clustering, respectively.	87

4.13	Sensitivity analysis for the hyper-parameters (A) ‘batch size’ and (B) ‘hidden layer neurons’ on <i>bmcite30k</i> , <i>kotliarov50k</i> , and <i>pbmc10k</i> datasets. (A) The batch sizes for the six embeddings for <i>bmcite30k</i> and <i>kotliarov50k</i> have been considered to be 80, 96, 112, 128, 144, and 160; and 32, 48, 64, 80, 96, and 112 for <i>pbmc10k</i> . (B) The number of hidden layer neurons for the six embeddings for <i>bmcite30k</i> , and <i>kotliarov50k</i> have been considered to be [32, [100, 64]], [32, [128, 64]], [16, [100, 64]], [16, [128, 64]], [64, [100, 64]], and [64, [128, 64]]; and [32, [100, 100]], [32, [128, 128]], [16, [100, 100]], [16, [128, 128]], [48, [100, 100]], and [48, [128, 128]] for <i>pbmc10k</i> , respectively. Here, the term $[a, [b_1, b_2]]$ conveys the structure of the hidden layers, where a is the number of nodes in the shared hidden layer, and b_1 and b_2 are the same in modality-specific hidden layer modules, respectively.	88
5.1	Graphical overview of the multi-modal survival analysis framework using modality-specific top features. (1) Survival analysis has been performed by stratifying patients into high and low-expression groups based on individual modality-specific features, followed by KM survival estimation and univariate log-rank testing. (2) Risk stratification has been conducted using multivariate Cox proportional hazards (CoxPH) regression to compute patient-specific risk scores, enabling separation into high and low-risk groups. (3) Model performance has been evaluated by comparing the concordance index (C-index) of a baseline clinical model with omics-augmented models across LSCC, LUAD, and the combined cohort, demonstrating improved prognostic performance with multi-modal integration.	95
5.2	Event-wise distribution of demographic and clinical variables. Kernel density plots show differences in Age and BMI between Event = 0 (censored) and Event = 1 (death), while bar plots depict event-stratified counts for Sex and Smoking status. Donut charts summarize the overall categorical composition for each variable (age groups, BMI classes, sex, and smoking categories). The left and right panels show the details of LSCC and LUAD cohorts, respectively. ‘Censored’ refers to an event that has not yet occurred within the study’s timeframe.	96
5.3	Kaplan-Meier curves based on high and low-expression values of each of the top features selected by NeuroMDAVIS-FS from (A) Protein, (B) RNA, and (C) CNV omics modalities, while applied on both the LSCC and LUAD cohort together. Patients have been stratified into high and low-expression groups based on the median value of each feature. Log-rank tests have been conducted to assess statistical significance in survival differences.	101
5.4	KM curves based on high and low risk scores have been obtained from a CoxPH regression applied using (A) top RNAs and (B) top proteins, selected by NeuroMDAVIS-FS. Hazard ratios of (C) top RNAs and (D) top proteins, selected by NeuroMDAVIS-FS. These results have been obtained by utilizing both the LSCC and LUAD cohorts together.	104

5.5	Comparison of predictive performance has been measured by the concordance index (C-index) across different feature configurations and cohorts. Results have shown for a baseline clinical model involving age, sex, BMI, and smoking status, as well as for models incrementally augmented with RNA features, protein features, and their joint combinations. Evaluations have been reported for the combined LSCC-LUAD cohort, and for the individual LSCC and LUAD cohorts.	105
6.1	The network architectures of G-NeuroDAVIS for visualization of high-dimensional datasets.	114
6.2	The network architecture of G-NeuroDAVIS, along with the mathematical notations.	119
6.3	2-dimensional embeddings produced by G-NeuroDAVIS (left) and VAE (right) for (A) <i>MNIST</i> and (B) <i>FMNIST</i> datasets respectively, in the context of unsupervised learning (Different class samples are represented by different colors).	125
6.4	2-dimensional embeddings produced by G-NeuroDAVIS (left) and VAE (right) for (A) <i>Coil20</i> and (B) <i>Olivetti Faces</i> datasets respectively, in the context of unsupervised learning (Different class samples are represented by different colors).	126
6.5	Visualization of interpolation paths in the embedding space, and their corresponding decoded outputs on <i>MNIST</i> dataset. (Left) Scatter plot showing the latent embedding colored by class labels, with interpolation paths between the start (green star) and end (red star) points using linear, normalized, Cauchy, and spherical interpolation methods. (Right) Decoded images along each interpolation path (rows) from start to end (columns), demonstrating how different interpolation strategies traverse the latent space and influence the generations. (A) Inter-cluster interpolation with embedding traversal, (B) Isolated inter-cluster interpolation, and (C) Cluster-intersecting interpolation.	128
6.6	Visualization of interpolation paths in the embedding space, and their corresponding decoded outputs on <i>FMNIST</i> dataset. (Left) Scatter plot showing the latent embedding colored by class labels, with interpolation paths between the start (green star) and end (red star) points using linear, normalized, Cauchy, and spherical interpolation methods. (Right) Decoded images along each interpolation path (rows) from start to end (columns), demonstrating how different interpolation strategies traverse the latent space and influence the generations. (A) Inter-cluster interpolation with embedding traversal, (B) Isolated inter-cluster interpolation, and (C) Cluster-intersecting interpolation.	129
6.7	Classification performance of G-NeuroDAVIS against VAE on (A) <i>MNIST</i> and (B) <i>FMNIST</i> dataset using k-NN and RF classifiers in terms of Accuracy, precision, Recall, and F1-Score. Results have been obtained by varying parameters of k-NN (<i>neighbours</i>) and RF (<i>estimators</i> / <i>trees</i>) respectively.	131

6.8	Classification performance of G-NeuroDAVIS against VAE on (A) <i>Coil20</i> and (B) <i>Olivetti Faces</i> dataset using k-NN and RF classifiers in terms of Accuracy, precision, Recall and F1-Score. Results have been obtained by varying parameter of k-NN (<i>neighbours</i>) and RF (<i>estimators / trees</i>) respectively.	132
6.9	Trustworthiness scores obtained on G-NeuroDAVIS-generated embedding of <i>MNIST</i> , <i>FMNIST</i> , <i>Coil20</i> , and <i>Olivetti Faces</i> datasets and compared against that obtained by using VAE. Parameter <i>n_neighbour</i> has been varied for robust results.	132
6.10	Samples generated using G-NeuroDAVIS (left) and VAE (right) in an unsupervised fashion on <i>MNIST</i> , <i>FMNIST</i> , <i>Coil20</i> , and <i>Olivetti Faces</i> dataset.	134
6.11	Samples generated using AE-guided RealNVP on (A) <i>MNIST</i> , (B) <i>FMNIST</i> , (C) <i>Coil20</i> , and (D) <i>Olivetti Faces</i> datasets. Samples generated using DCGAN on (E) <i>MNIST</i> and (F) <i>FMNIST</i> datasets. Samples generated using DDPM on (G) <i>MNIST</i> and (H) <i>FMNIST</i> datasets.	135
6.12	Comparison of G-NeuroDAVIS in terms of (A) improved precision and (B) improved recall scores against VAE, DCGAN, DDPM, and AE-guided RealNVP on <i>MNIST</i> , <i>FMNIST</i> , <i>Coil20</i> , and <i>Olivetti Faces</i> datasets. . . .	136
6.13	Samples generated using G-NeuroDAVIS in an supervised fashion on (A) <i>MNIST</i> , (B) <i>FMNIST</i> , (C) <i>Coil20</i> , and (D) <i>Olivetti Faces</i> dataset.	138
6.14	Comparison based on trainable parameters	139
6.15	(A) Run time and (B) Inference time comparison of G-NeuroDAVIS against VAE, DCGAN, DDPM, and AE-guided RealNVP on <i>MNIST</i> , <i>FMNIST</i> , <i>Coil20</i> , and <i>Olivetti Faces</i> dataset.	139
6.16	(A) k-NN and (B) RF classifiers performance with varying hyper-parameters of G-NeuroDAVIS on <i>MNIST</i> , <i>FMNIST</i> , <i>Coil20</i> , and <i>Olivetti Faces</i> dataset.	140
6.17	(A) 2-dimensional embeddings produced by G-NeuroDAVIS on <i>Celeb</i> dataset in the context of unsupervised learning (Blue dots indicate male faces, while Red dots represent female faces). (B) Samples generated using G-NeuroDAVIS in an unsupervised fashion on <i>Celeb</i> dataset.	141
A.I.1	The original distributions and the embeddings produced by NeuroDAVIS, t-SNE and UMAP after a nine-dimensional transformation ($x + y$, $x - y$, xy , x^2 , y^2 , x^2y , xy^2 , x^3 , y^3) of <i>Elliptic Ring</i> ((A)-(D)), <i>Olympic</i> ((E)-(H)), <i>Spiral</i> ((I)-(L)) and <i>Shape</i> ((M)-(P)) datasets.	152
A.I.2	Preservation of cluster sizes by NeuroDAVIS, t-SNE and UMAP in <i>World Map</i> dataset. ρ is the Pearson correlation coefficient. The higher the ρ value, the better the preservation.	153
A.II.1	Embeddings produced by t-SNE with varying perplexity parameter of <i>Breast Cancer</i> dataset.	154
A.II.2	Embeddings produced by UMAP with varying parameters <i>n_neighbours</i> and <i>min_dist</i> of <i>Breast Cancer</i> dataset.	155
A.II.3	Embeddings produced by t-SNE with varying perplexity parameter of <i>Wine</i> dataset.	156
A.II.4	Embeddings produced by UMAP with varying parameters <i>n_neighbours</i> and <i>min_dist</i> of <i>Wine</i> dataset.	157

A.II.5	Embeddings produced by t-SNE with varying perplexity parameter of <i>Spam</i> dataset.	158
A.II.6	Embeddings produced by UMAP with varying parameters <i>n_neighbours</i> and <i>min_dist</i> of <i>Spam</i> dataset.	159
A.II.7	Embeddings produced by t-SNE with varying perplexity parameter of <i>Coil20</i> dataset.	160
A.II.8	Embeddings produced by UMAP with varying parameters <i>n_neighbours</i> and <i>min_dist</i> of <i>Coil20</i> dataset.	161
A.II.9	Embeddings produced by t-SNE with varying perplexity parameter of <i>Usoskin</i> dataset.	162
A.II.10	Embeddings produced by UMAP with varying parameters <i>n_neighbours</i> and <i>min_dist</i> of <i>Usoskin</i> dataset.	163
A.II.11	Embeddings produced by t-SNE with varying perplexity parameter of <i>Jurkat</i> dataset.	164
A.II.12	Embeddings produced by UMAP with varying parameters <i>n_neighbours</i> and <i>min_dist</i> of <i>Jurkat</i> dataset.	165
A.III.1	(A) and (B) show performance of k-NN classifier with $k = 5$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of <i>Breast Cancer</i> , <i>Wine</i> , <i>Spam</i> and <i>Coil20</i> datasets, in terms of Accuracy and F1-scores respectively, (C) and (D) show performance of RF classifier with $n_estimators = 20$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of <i>Breast Cancer</i> , <i>Wine</i> , <i>Spam</i> and <i>Coil20</i> datasets, in terms of Accuracy and F1-scores respectively.	166
A.III.2	(A) and (B) show performance of k-NN classifier with $k = 15$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of <i>Breast Cancer</i> , <i>Wine</i> , <i>Spam</i> and <i>Coil20</i> datasets, in terms of Accuracy and F1-scores respectively, (C) and (D) show performance of RF classifier with $n_estimators = 40$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of <i>Breast Cancer</i> , <i>Wine</i> , <i>Spam</i> and <i>Coil20</i> datasets, in terms of Accuracy and F1-scores respectively.	167
A.III.3	(A) and (B) show performance of k-NN classifier with $k = 25$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of <i>Breast Cancer</i> , <i>Wine</i> , <i>Spam</i> and <i>Coil20</i> datasets, in terms of Accuracy and F1-scores respectively, (C) and (D) show performance of RF classifier with $n_estimators = 60$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of <i>Breast Cancer</i> , <i>Wine</i> , <i>Spam</i> and <i>Coil20</i> datasets, in terms of Accuracy and F1-scores respectively.	168
A.III.4	(A) and (B) show performance of k-NN classifier with $k = 35$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of <i>Breast Cancer</i> , <i>Wine</i> , <i>Spam</i> and <i>Coil20</i> datasets, in terms of Accuracy and F1-scores respectively, (C) and (D) show performance of RF classifier with $n_estimators = 80$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of <i>Breast Cancer</i> , <i>Wine</i> , <i>Spam</i> and <i>Coil20</i> datasets, in terms of Accuracy and F1-scores respectively.	169

A.III.5	(A) and (B) show performance of k-NN classifier with $k = 45$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of <i>Breast Cancer</i> , <i>Wine</i> , <i>Spam</i> and <i>Coil20</i> datasets, in terms of Accuracy and F1-scores respectively, (C) and (D) show performance of RF classifier with $n_estimators = 100$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of <i>Breast Cancer</i> , <i>Wine</i> , <i>Spam</i> and <i>Coil20</i> datasets, in terms of Accuracy and F1-scores respectively.	170
A.III.6	(A) and (B) show ARI and FMI scores for cell-type clustering using k-means on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PHATE, IVIS, PCA and RP-generated embeddings of <i>Usoskin</i> and <i>Jurkat</i> datasets, (C) and (D) show ARI and FMI scores for cell-type clustering using agglomerative hierarchical clustering on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PHATE, IVIS, PCA and RP-generated embeddings of <i>Usoskin</i> and <i>Jurkat</i> datasets.	171
A.IV.1	Embeddings produced by NeuroDAVIS on both the training and test parts of <i>FMNIST</i> dataset.	172
B.I.1	Kaplan-Meier curves based on high- and low-expression values of each of the top features selected by NeuroMDAVIS-FS from (A) Protein, (B) RNA, and (C) CNV omics modalities, while applied on the LSCC cohort. Patients have been stratified into high and low-expression groups based on the median value of each feature. Log-rank tests have been used to assess statistical significance in survival differences.	173
B.I.2	Kaplan-Meier curve based on high and low risk scores obtained from a CoxPH regression applied using (A) top RNAs, and (B) top proteins selected by NeuroMDAVIS-FS. Hazard scores of (C) top RNAs, and (D) top proteins selected by NeuroMDAVIS-FS	174
B.II.1	Kaplan-Meier curves based on high- and low-expression values of each of the top features selected by NeuroMDAVIS-FS from (A) Protein, (B) RNA, and (C) CNV omics modalities, while applied on the LUAD cohort. Patients have been stratified into high and low-expression groups based on the median value of each feature. Log-rank tests have been used to assess statistical significance in survival differences.	175
B.II.2	Kaplan-Meier curve based on high and low risk scores obtained from a CoxPH regression applied using (A) top RNAs, and (B) top proteins selected by NeuroMDAVIS-FS. Hazard scores of (C) top RNAs, and (D) top proteins selected by NeuroMDAVIS-FS	176
C.I.1	Mapping from 2-dimensional embedding, produced by G-NeuroDAVIS and VAE, in an unsupervised fashion, to the original high-dimensional space for <i>MNIST</i> dataset.	178
C.I.2	Mapping from 2-dimensional embedding, produced by VAE, to the original high-dimensional space for <i>FMNIST</i> dataset.	178
C.I.3	Mapping from 2-dimensional embedding, produced by G-NeuroDAVIS, in an unsupervised fashion, to the original high-dimensional space for <i>Coil20</i> dataset.	179
C.I.4	Mapping from 2-dimensional embedding, produced by G-NeuroDAVIS, in an unsupervised fashion, to the original high-dimensional space for <i>Olivetti Faces</i> dataset.	179

C.II.1	2-dimensional embeddings produced by (A) G-NeuroDAVIS and (B) VAE, in an unsupervised fashion, for different sub-samples with varying sizes of <i>MNIST</i> dataset, coloured by different digits.	180
C.II.2	2-dimensional embeddings produced by (A) G-NeuroDAVIS and (B) VAE, in an unsupervised fashion, for different sub-samples with varying sizes of <i>FMNIST</i> dataset, coloured by clothing's label.	181
C.II.3	Classification performance of (A) k-NN and (B) RF on G-NeuroDAVIS-generated embedding against that obtained using VAE for different sub-samples with varying sizes of <i>MNIST</i> dataset in terms of Accuracy, Precision, Recall and F1-Score.	182
C.II.4	Classification performance of (A) k-NN and (B) RF on G-NeuroDAVIS-generated embedding against that obtained using VAE for different sub-samples with varying sizes of <i>FMNIST</i> dataset in terms of Accuracy, Precision, Recall and F1-Score.	183
C.III.1	Embeddings produced by G-NeuroDAVIS, in an supervised setting, for (A) <i>MNIST</i> , (B) <i>FMNIST</i> , (C) <i>Coil20</i> and (D) <i>Olivetti Faces</i> dataset	184

List of Tables

2.1	Description of datasets used for evaluation of NeuroDAVIS	28
3.1	Comparison of existing state-of-the-art and NeuroDAVIS-FS (NDFS) . .	55
3.2	Descriptions of the datasets utilized for the model evaluation	57
4.1	Summary of datasets used in this chapter	77
4.2	Comparison of basic features of the visualization methods used in this chapter	78
5.1	Top 15 features selected by NeuroMDAVIS-FS from each of the omics modalities, viz., Protein, RNA, and CNV. Here both LSCC and LUAD samples have been utilized to train NeuroMDAVIS.	99
6.1	Descriptions of the datasets used for evaluation of G-NeuroDAVIS	123
B.1	Top 15 features selected by NeuroMDAVIS-FS from each of the omics modalities, viz., Protein, RNA, and CNV. Here only LSCC samples utilized to train NeuroMDAVIS.	174
B.2	Top 15 features selected by NeuroMDAVIS-FS from each of the omics modalities, viz., Protein, RNA, and CNV. Here only LUAD samples utilized to train NeuroMDAVIS.	176
B.3	Multi-omics features and their functional and clinical relevance in lung cancer	177

List of Publications

(on the work of the present thesis)

1. **Chayan Maitra**, Vivek Das, Dibyendu B. Seal, and Rajat K. De. “Multi-Omics Integration for Identification of Prognostic Molecular Signatures for Survival Stratification in Lung Cancer”. *MedRxiv*, 2026.
DOI: <https://doi.org/10.64898/2026.02.28.26347335>
2. **Chayan Maitra**, Dibyendu B. Seal, Vivek Das, and Rajat K. De. “NeuroMDAVIS: Visualization of single-cell multi-omics data under deep learning framework”. *Journal of Computational Biology*, 0(0):15578666261423965, 2026. PMID: 41744270.
DOI: <https://doi.org/10.1177/15578666261423965>
3. **Chayan Maitra** and Rajat K. De. “G-NeuroDAVIS: A Generative Model for Data Visualization through a Generalized Embedding”. *Neural Networks*, 193:107948, 2026.
DOI: <https://doi.org/10.1016/j.neunet.2025.107948>
4. **Chayan Maitra**, Anwesha Sengupta, and Rajat K. De. “NeuroDAVIS-FS: Feature Selection Through Visualization Using NeuroDAVIS”. In *Proceedings of the International Conference on Pattern Recognition*, pages 284–298. Springer Nature Switzerland, 2025.
DOI: https://doi.org/10.1007/978-3-031-78395-1_19
5. **Chayan Maitra**, Dibyendu B. Seal, and Rajat K. De. “NeuroDAVIS: A neural network model for data visualization”. *Neurocomputing*, 573:127182, 2024.
DOI: <https://doi.org/10.1016/j.neucom.2023.127182>

List of other Publications

1. Sourav Ghosh, **Chayan Maitra** and Rajat K. De. “CMET: Clustering guided METric for quantifying embedding quality”. (*Communicated*)
2. **Chayan Maitra**, Dibyendu B. Seal, Vivek Das, Yevgeniy Vorobeychik, and Rajat K. De. “UMINT-FS: UMINT-guided Feature Selection for multi-omics datasets”. In *Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 594-601. IEEE, 2023
DOI: <https://doi.org/10.1109/BIBM58861.2023.10385731>
3. **Chayan Maitra**, Dibyendu B. Seal, Vivek Das, and Rajat K. De. “Unsupervised neural network for single cell Multi-omics INTegration (UMINT): an application to health and disease”. *Frontiers in Molecular Biosciences* 10 (2023): 1184748.
DOI: <https://doi.org/10.3389/fmolb.2023.1184748>

Chapter 1

Introduction

1.1 Introduction

Understanding complex data has always relied on the ability to visualize its hidden patterns and relationships from high-dimensional representations [6]. In scientific analysis, engineering workflows, and machine learning pipelines, visualization serves as an essential bridge between high-dimensional structures and human interpretability [7]. With increasing scale, variability, and complexity becoming inherent properties of modern datasets, traditional approaches often fail to accurately capture the underlying data geometry [8]. Modern data ecosystems now frequently incorporate multiple modalities ranging from images and text to multi-omics profiles and sensor streams; each showing its own distributional characteristics and noise patterns. This diversity necessitates a more comprehensive approach that spans feature extraction, feature selection, integration of multi-modal information, and ultimately, visualization that is both coherent and generative.

This thesis presents a complete visualization framework that functions across the continuum of data complexity and aims to address this evolving landscape. It begins with examining single-modality data, where robust feature extraction methods lay the foundation for dimensionality reduction and representation learning. These representations, once refined through feature selection strategies, enable improved downstream tasks, such as clustering, classification, and exploratory analysis. By grounding the pipeline in feature quality, the methodology ensures that visualizations remain faithful to the underlying structure of the data rather than artifacts of noise or bias.

As the framework scales to multi-modal data, the challenges become more intricate.

Different modalities often reside in incomparable feature spaces, vary in sample availability, and require alignment strategies that preserve complementary information while mitigating redundancy. The thesis also incorporates mechanisms for cross-modal representation learning and joint embedding, enabling coherent visualization across distinct data types. By doing so, it highlights relationships that would remain inaccessible within single modality analyses, fostering a deeper understanding of system-level patterns.

Beyond representation and integration, the thesis extends naturally toward generative capabilities. Modern generative models, such as diffusion models and variational autoencoders, not only visualize but also synthesize new data samples aligned with the learned manifold. Incorporating such models transforms visualization into a bidirectional process: users not only can view existing structures but also explore hypothetical variations, simulate data under controlled perturbations, and validate insights through synthetic generation. This shift from visualizing what exists to imagining what could exist represents a powerful expansion of the analytical toolkit.

The present thesis is designed as a unified framework that bridges fundamental representation techniques with advanced multi-modal integration and state-of-the-art generative modeling. It provides a holistic perspective on how data can be explored, interpreted, and synthesized across increasing levels of complexity. By connecting single and multi-modal learning, feature engineering, and generative visualization, the thesis positions itself as a comprehensive solution for researchers aiming to understand and innovate within the rapidly expanding universe of high-dimensional data. Thus the thesis encompasses the full spectrum of data visualization, from basic visualization to joint visualization and ultimately generalized visualization, and we, thereby, title it PORTRAIT.

1.2 Dimension Reduction

In real world scenario, where data grows both in scale and complexity, many features often become redundant, noisy, or less informative. This increases computational cost, complicates statistical modeling, and can obscure meaningful structure in the data [9]. Dimensionality Reduction (DR) provides a principled way to address these challenges by transforming the original feature space into a more compact representation that mostly preserves the essential information required for interpretation or downstream analysis. In

essence, it aims to simplify the dataset while retaining the underlying patterns that define its behaviour. Reduced-dimensional representations are easier to model, analyze and interpret. It enables visualization, reveals latent structure, and facilitates downstream tasks such as anomaly detection and pattern recognition. In this sense, dimensionality reduction is not merely a computational convenience but a fundamental step toward extracting meaningful insight from high-dimensional data. Existing literature broadly categorizes dimensionality reduction techniques into two main types, viz., feature extraction and feature selection [9].

1.2.1 Feature extraction and Feature selection

Feature extraction methods transform the original high-dimensional data into a new lower-dimensional representation by constructing latent features that capture the most informative relations among the data points. These new features are typically combinations or transformations of the original variables, and aim to preserve essential properties such as variance, neighborhood relationships or statistical dependence. In contrast, feature selection techniques reduce dimensionality by identifying and retaining a subset of the most relevant original features while discarding redundant, irrelevant or noisy ones [10]. Since the selected features remain unchanged, these methods tend to preserve interpretability, and are particularly useful in domains where understanding the role of individual variables is crucial.

Feature extraction methods often rely on matrix factorization, latent variable modeling, or decomposition strategies to uncover hidden structure in the data. Some commonly used DR techniques include Principal Component Analysis (PCA) [11], Independent Component Analysis (ICA) [12], Multi-dimensional Scaling (MDS) [13], Isomap [14], Random Projection (RP) [15], Non-negative Matrix Factorization (NMF) [16], and Singular Value Decomposition (SVD) [17]. PCA projects the data into a transformed space spanned by the vectors representing the maximum variance, while ICA extracts subcomponents from a multivariate signal. MDS is a DR and visualization technique that can extract dissimilarity within structures in the data. Isomap, a non-linear DR method that combines the advantages of several other methods, can be considered an improved version of MDS. Both NMF and SVD are methods for matrix factorization, which have found significant usage in topic modeling and signal processing. All the above methods fall into the

category of unsupervised DR. Linear Discriminant Analysis (LDA) [18], on the contrary, is a supervised method used for DR and pattern recognition.

On the other hand, feature selection strategies are commonly guided by statistical relevance, information-theoretic criteria, or model-based importance measures. There are three broad categories of feature selection methods, viz., Filter-based methods, Wrapper methods and Embedded methods [19, 20]. Filter-based selection methods rely on statistical metrics, such as variance thresholds, mutual information, or correlation-based criteria, to identify features that contribute a meaningful signal [21, 22]. One major drawback of these methods is assumption of features being independent and not considering the inherent association among them. Wrapper methods use predictive models to evaluate subsets of features and iteratively search for combinations that optimize performance. Forward Selection, Backward Elimination, and Recursive Feature Elimination (RFE) are some examples of wrapper methods [23, 24]. These methods are computationally expensive and sensitive to the specific model. Finally, Embedded methods, such as Lasso, Random Forest, and Gradient Boosting Machines, use ensemble models for feature selection [25, 26]. They function based on regularization and integrate feature selection directly into the training process. However, they are also computationally expensive and prone to model overfitting.

Feature extraction and feature selection together form the two principal paradigms of dimensionality reduction, each offering distinct trade-offs between compactness, interpretability, and computational efficiency [9]. Their applicability largely depends on the nature of the data and the goals of the analysis. However, there is a basic difference between dimensionality reduction and data visualization. It lies in their primary objective and usage, even though they are closely related and often overlap. Any DR technique should preserve the virtues of both local and global structures present in the data. However, we are more accustomed to seeing methods preferring one over the other [27, 28].

1.3 Data visualization

Dimensionality reduction focuses on optimizing information representation for computation and analysis, whereas data visualization focuses on optimizing representation for human perception. On the other hand, visualization can be considered a more specialized

and insightful form of dimensionality reduction. While dimensionality reduction techniques primarily focus on transforming high-dimensional data into a lower-dimensional representation by removing noise, redundancy, and irrelevant features, thereby improving the efficiency and performance of computational models, visualization methods go a step further. Data visualization not only reduces dimensions but also translates complex, abstract data into an interpretable visual form that humans can intuitively analyze. It enables researchers to perceive underlying structures, patterns, and relationships among samples that may otherwise remain hidden in numerical form. Through visualization, one can identify natural groupings or clusters, detect outliers, and gain a deep understanding of how different samples relate to each other. In essence, while dimensionality reduction enables machines to process data effectively, visualization empowers humans to comprehend, interpret, and draw meaningful inferences from that data.

1.3.1 Methods for data visualization

Researchers have developed several methods to visualize high-dimensional datasets. In the following paragraphs, several important visualization techniques are discussed. As discussed in the Section 1.2.1, several linear methods, like PCA, ICA and MDS, have been used in literature to visualize high-dimensional datasets. Non-linear projection algorithms, like t-SNE [29], however, have an edge over linear algorithms in extracting complex latent structures within the data [30].

t-SNE

t-SNE tries to make the latent data distribution similar to the original distribution by minimizing the KL-Divergence loss. Nevertheless, there are severe downsides to t-SNE. t-SNE is highly sensitive to noise, and even randomly distributed points may be transformed into spurious clusters [31]. Moreover, t-SNE is popular for preserving local distances over global distances [32, 33], which hinders drawing realistic conclusions from t-SNE visualizations.

Laurens van der Maaten has developed a parametric version of t-SNE [34], which parameterizes the non-linear mapping from low-dimensional space to the data space with the help of a feed-forward neural network, inspired by the working principle of Restricted Boltzmann Machine (RBM). First, a stack of RBMs is trained to construct a pretrained

neural network, which is then finetuned using a loss that attempts to preserve the local structure of the data in the embedded space. The parametric scheme accelerates the performance of t-SNE in visualization of large-scale datasets. Other parametric versions of t-SNE include [35] and [36].

Fit-SNE

Fast interpolation-based t-SNE (Fit-SNE) [37] is another variant of t-SNE, which provides accelerated performance on large datasets too. It replaces the most time-consuming step in t-SNE, which involves estimation of the degree of repulsion between every pair of observations, by using a smaller number of interpolation nodes to compute interactions among the observations. Instead of computing the full t-SNE embedding directly, Fit-SNE interpolates between the high-dimensional and low-dimensional affinities using Lagrange polynomials. However, the parameters of both t-SNE and Fit-SNE are very sensitive since a small change in any one of them may hamper the overall embedding quality [38].

UMAP

Uniform Manifold Approximation and Projection (UMAP) [39] is a manifold learning technique to reduce data dimension and scales to large datasets pretty well. The mathematical foundations of UMAP is inspired from the work of Laplacian eigenmaps which is another DR method. Graph construction and graph layout are the two major steps in UMAP. In order to construct a graph, UMAP uses a fuzzy set theoretic approach and to optimize the graph layout. It tunes a spectral embedding by minimizing cross-entropy between the edge-weights. However, UMAP assumes that there exists a manifold structure within the data, which is not always realistic. Like t-SNE [40], it also gives precedence to local relationships over global relationships.

A study by Kobak and Linderman [41] has established the fact that initial embeddings are very crucial for both t-SNE and UMAP. Embedding initialization supported by *apriori* knowledge accelerates the performance of both these methods. Nevertheless, both t-SNE and UMAP have usually failed to produce effective visualization of complex datasets independently.

PHATE

Potential of Heat-diffusion for Affinity-based Transition Embedding (PHATE) [42] is a nonlinear visualization method designed to capture both local and global structure in high-dimensional data by modelling how information diffuses across the data manifold. Instead of relying solely on distances between points, PHATE constructs a diffusion-based affinity matrix that reflects the underlying geometry and progression patterns in the dataset. This makes it especially powerful for visualizing continuous trajectories, such as developmental processes, cellular differentiation, or temporal transitions. The resulting embeddings are smooth, interpretable, and preserve meaningful transitions that might be distorted by methods like t-SNE. PHATE is widely used in fields such as single-cell biology, where understanding gradual changes in data is as important as identifying discrete clusters. However, the effectiveness of PHATE on other real-world datasets is yet to be assessed.

Neural network (NN)-based methods have been in use as non-linear DR tools in almost every field, more so during the last decade [43, 44, 45]. Unsupervised NNs are trained to learn a non-linear function, while features extracted from an intermediate hidden layer with relatively low cardinality serve as a low-dimensional representation of the data. The earliest usage of NN methods for projection can be mapped back to Self-Organizing Map (SOM), also known as the Kohonen map [46]. SOM is an unsupervised method that projects data into lower dimensions while preserving topological structures. In recent years, Autoencoder (AE)-based methods have predominantly ruled the DR space [47, 48, 49], where multiple variations of AE have been developed to address the problem of ‘curse of dimensionality’ in several application domains, including computer vision and computational biology.

IVIS

A Siamese neural network architecture with a triplet loss function, called IVIS [50], has been developed for data visualization of high-dimensional biological datasets. The network consists of multiple identical sub-networks that are capable of finding similarity/dissimilarity between the inputs by comparing feature values, and are usually learned using a triplet loss or a contrastive loss function. In the case of learning by a triplet loss function, a positive vector and a negative vector are compared against a baseline vector

(anchor), while in the case of learning by a contrastive loss, a weight decay function is used to regularize the weights. IVIS consists of three identical base networks and an improved triplet loss function that is used to learn the shared weights.

All these NN-based methods have shown high potential with respect to DR and visualization tasks. However, not all methods can simultaneously preserve local and global structures of the data.

1.4 Integration and joint visualization

Multi-modal datasets have become widely accessible across many domains, with growing collections that combine images, text, audio, sensor data, and structured information. Large public repositories and open-data initiatives now provide rich, well-curated multi-modal resources that support research in machine learning, healthcare, natural language processing, and beyond. As data collection technologies advance, the availability and diversity of multi-modal datasets continue to expand, enabling deeper and more comprehensive analyses. Multi-modal data can be approached in two main ways. The first and most used approach is by integrating different modalities into a unified representation and then projecting them into 2 or 3 dimensions for visualization. The second approach is by visualizing them in separate low-dimensional spaces and then aligning them into a single integrated space. Both these approaches have their respective drawbacks. Joint visualization is a more generalized version of data integration, as it has additional dimensionality constraints. Integration focuses on fusing complementary information to create a synergistic cohesive model that captures cross-modal relationships. Joint visualization emphasizes presenting aligned modalities together so that their similarities, differences, and interactions become visually interpretable.

1.4.1 Methods for integration

Methods that focus on integrating multi-modal or multi-omics data, without producing embeddings for visualization, are designed to fuse heterogeneous information into a shared, harmonized representation that captures cross-modal relationships. Approaches such as MOFA or MOFA+ (Multi-Omics Factor Analysis) [51, 52], DIABLO [53], TotalVI [54], UMINT [55], and Multigrade [56] learn latent factors that summarize the underly-

ing biological variation across datasets, enabling researchers to identify shared patterns, correlated signals, or molecular drivers that span multiple data types. Similarly, statistical frameworks, like Canonical Correlation Analysis (CCA), Multi-set CCA (MCCA), RGCCA, and multi-view autoencoders, align modalities by maximizing cross-correlation or reconstructing shared latent spaces, and help revealing how different data layers complement or reinforce each other [57]. These integration-focused methods prioritize uncovering the structure, and interplay between modalities rather than generating a visual layout for human interpretation.

Another class of integration methods concentrates on merging sample-level relationships across datasets. For example, Similarity Network Fusion (SNF) constructs patient-specific or sample-specific networks for each modality and then combines them into a single fused network that captures consensus similarities [58]. These approaches are powerful for tasks such as clustering, biomarker discovery, and identifying cross-modal patterns, but they also fall short of producing low-dimensional embeddings for visualization. Instead, they serve as foundational tools that prepare harmonized, multi-modal representations, which can then be fed into downstream visualization techniques like UMAP or t-SNE. In essence, integration-only methods excel at data fusion, structure discovery, and latent representation learning, setting the stage for—but not directly providing—human-readable visualizations.

Interestingly, when it comes to visualizing observations from multi-omics experiments, there is no joint data visualization method that can combine views from several omics layers. MultiMAP [59], a recently developed method for dimensionality reduction and multi-omics integration (a more generalized version of UMAP), enables shared visualization of multi-omics datasets. It creates a non-linear manifold representing different omics modalities, allowing the identification of shared features across cells in a co-embedded space. However, the shared visualization produced by MultiMAP generates separate embeddings for each of the omics modalities.

1.5 Data Generation

The methods we have discussed so far are capable of producing a visualization that either captures the local or global relationships present in any dataset [27, 28]. In both the

cases, the embedding does not have the generalization capability that shows a shortfall regarding sample generation from those embeddings. Here, the generalization capability of an embedding refers to a compact, meaningful, and lower-dimensional representation of data, which extends beyond the training set and is capable of generating diverse samples.

1.5.1 Methods for data generation

In order to address this problem of sample generation, several methods exist in recent literature. Variational Autoencoders (VAE) [60], Generative Adversarial Networks (GANs) [61], Denoising Diffusion Probabilistic Models (DDPM) [62], RealNVP (Real-valued Non-Volume Preserving) [63], and Glow (Generative Flow with Invertible 1×1 Convolutions) [64] are capable of generating samples efficiently, but do not always produce a latent embedding. VAE is a neural network-based model consisting of an encoder and a decoder. They can learn a latent representation from input data by minimizing the mean squared loss, and the divergence between the learned latent distribution and a predefined prior distribution simultaneously. On the other hand, GANs, containing a generator and a discriminator network, are trained adversarially. The generator is able to map samples from a simple distribution (e.g., Gaussian noise) to the data space, while the discriminator learns to distinguish between real and generated samples. Through adversarial training, GANs produce high-quality samples that are similar to real data and can capture the underlying distribution. Diffusion-based models, like DDPM, generate samples by starting from noise and gradually improving it through a learned reverse diffusion process [62]. During training, it learns to predict and remove noise added over several time steps. Flow-based models are another generative model that is able to learn a bijective mapping between a known statistical distribution (e.g., Gaussian) and the data distribution [63, 64]. They achieve it by modeling the data density through a sequence of invertible transformations. RealNVP and Glow are examples of flow-based models which are also capable of generating high-quality images.

1.6 Motivation of the thesis

This section provides the motivation underlying the study presented in this thesis. The challenge of dimensionality reduction and visualization of high-dimensional datasets

has long persisted in data analysis and machine learning research [8]. Modern high-throughput technologies have generated massive datasets comprising multiple views of the same samples, often containing complex and heterogeneous data types. Effective visualization of such datasets requires algorithms that can uncover hidden patterns without distorting local and global structures, while simultaneously revealing the inherent non-linearity of the data. However, only a limited number of methodologies have been developed, which can successfully accomplish this task. Motivated by these challenges, we have introduced NeuroDAVIS, a novel unsupervised deep neural network model that has been designed to perform structure-preserving and efficient visualization of high-dimensional data. During the subsequent stages of this research, we have perceived that sometimes visualization through feature extraction is not competent enough to provide an interpretable understanding of the underlying data characteristics. In certain cases, feature selection has played a more significant role than feature extraction [10]. Motivated by this realization, we have developed a feature selection module built upon NeuroDAVIS visualization framework.

Later, we have realized that paired data from multiple resources are available. These multi modal data required proper methodology for visualization. In order to achieve the same, we have developed a neural network-based model, called NeuroMDAVIS, that has performed joint visualization of multi-omics datasets. This model has produced a unified low-dimensional embedding that has preserved both local and global structures across different modalities, enabling a more integrated view of the data. Nonetheless, while the joint visualization has provided an elegant feature extraction mechanism, it has failed to highlight key modality-specific features that are informative. Motivated by this limitation, we have developed a feature selection framework on top of the joint visualization method.

Despite the effectiveness of these visualization and feature selection methods, all of them have shared a common limitation that they have not been capable of producing a generalized embedding. A generalized embedding is needed to capture the complex structure of the data in a latent space that is both representative and generative. Such an embedding not only preserves the intrinsic geometry and relationships within the data but is also able to generate new, realistic samples that follow the same distribution as the original data. In other words, while traditional visualization methods have provided meaningful two or three-dimensional projections, they have remained non-generative in nature and

have lacked the ability to reconstruct data from learned representations. Motivated by this crucial limitation, we have developed G-NeuroDAVIS, a novel generative deep neural network model that has extended the capability of NeuroDAVIS. G-NeuroDAVIS has been designed to learn a generalized embedding that not only visualizes high-dimensional data effectively but also captures the latent distribution well enough to generate new data samples resembling the real ones.

Thus, through the progressive development from NeuroDAVIS to its feature selection extensions and finally to G-NeuroDAVIS, this research has established a comprehensive framework for unified visualization, interpretability, and generative modeling of high-dimensional and multi-modal data.

1.7 Scope of the thesis

This section describes the scope of the thesis in detail.

1.7.1 Chapter 2 - NeuroDAVIS: A neural network model for data visualization [1]

In this chapter, we have introduced a novel unsupervised deep neural network model, called NeuroDAVIS, for data visualization. NeuroDAVIS is capable of extracting important features from the data, without assuming any data distribution, visualizing them effectively in low dimensions, and preserving both local and global structures simultaneously. It has been shown theoretically that the neighbourhood relationship of the data in high dimensions remains preserved in lower dimensions. The performance of NeuroDAVIS has been evaluated on a wide variety of synthetic and real high-dimensional datasets, including numeric, textual, image and biological data. NeuroDAVIS has been highly competitive against t-SNE, UMAP, PCA and RP with respect to visualization quality, preservation of data size, shape, and both local and global structure. It has outperformed Fit-SNE, a variant of t-SNE, for most of the datasets. For the biological datasets, besides t-SNE, UMAP, Fit-SNE, PCA and RP, NeuroDAVIS has also performed well compared to other state-of-the-art algorithms, like PHATE and IVIS. Downstream classification and clustering analyses have also revealed favourable results for NeuroDAVIS-generated embeddings over the others.

1.7.2 Chapter 3 - NeuroDAVIS-FS: Feature Selection Through Visualization Using NeuroDAVIS [2]

This chapter aims to develop a feature selection method, called NeuroDAVIS-FS, which performs in an unsupervised learning setup without assuming any prior data distribution. Initially, it considers training using the model NeuroDAVIS, developed earlier for data visualization, and then selects features according to the trained model. The efficacy of the NeuroDAVIS-FS has been demonstrated on various datasets from different domains and found to be effective in comparison with state-of-the-art feature selection methods. In addition, two case studies on image and biological datasets with a very low sample-feature ratio have been executed and found to be effective for relevant feature selection.

1.7.3 Chapter 4 - NeuroMDAVIS: Visualization of single-cell multi-omics data under deep learning framework [3]

Single-cell technologies have favoured extensive advancements in cell-type discovery, cell state identification, the development of lineage tracing, and disease understanding, among others. Further, single-cell multi-omics data provide several views for the same set of cells. Analyzing such multi-omics data is often hindered by its large dimension. In this chapter, we introduce NeuroMDAVIS, a novel unsupervised deep neural network model, for joint visualization of biological datasets having multiple modalities. Joint visualization refers to transforming the feature space of each modality and combining them to produce a latent embedding that supports visualization of the multi-modal dataset in the newly transformed feature space. NeuroMDAVIS transforms the feature space of each modality and integrates them into a shared latent space, capturing both modality-specific and common information across different layers. The model effectively learns both local and global relationships within the data, providing a meaningful low-dimensional representation for further analysis. NeuroMDAVIS competes against state-of-the-art visualization models like t-SNE, UMAP, Fit-SNE, IVIS, and the manifold learning-based generalized version of UMAP, called MultiMAP. Downstream analyses have reflected superior classification and clustering performance across all the datasets. To the best of our knowledge, NeuroMDAVIS is the first model to offer joint visualization for multi-modal biological datasets. It competes with the state-of-the-art visualization methods providing a robust and efficient approach

for understanding complex multi-omics data.

1.7.4 Chapter 5 - NeuroMDAVIS-FS Identifies Prognostic Molecular Signatures in Lung Cancer towards Survival Analysis and Risk Stratification [4]

Lung cancer is characterized by profound intratumoral and inter-patient heterogeneity, spanning histological subtypes, molecular landscapes, and the tumor microenvironment. While multi-omics integration is essential for capturing this complexity, leveraging these data to define survival-associated subpopulations explicitly remains a significant challenge. In this study, we have developed NeuroMDAVIS-FS, an unsupervised deep learning framework designed to stratify lung cancer patients by survival risk and identify molecular determinants underlying improved clinical outcomes. Using the CPTAC cohort, we have integrated genomic (CNV), transcriptomic (RNA-seq), and proteomic profiles to extract modality-specific features. Candidate biomarkers have been validated through Kaplan-Meier (KM) survival analysis and univariate Cox proportional hazards (CoxPH) regression. A final multivariate CoxPH model has effectively stratified patients into high-risk and low-risk cohorts (Kaplan Meier p -value < 0.001). Notably, integrating these molecular features with baseline clinical models has significantly improved prognostic accuracy, increasing the concordance index by 43.79% in LUAD, 31.05% in LSCC, and 23.76% across the pan-lung cancer cohort. These results demonstrate that NeuroMDAVIS-FS identifies robust, biologically relevant features that surpass traditional clinical variables in predicting patient outcomes, offering a scalable path for precision oncology.

1.7.5 Chapter 6 - G-NeuroDAVIS: A generative model for data visualization through a generalized embedding [5]

In this chapter, a novel generative model, called G-NeuroDAVIS, has been developed, which is capable of visualizing high-dimensional data through a generalized embedding and thereby generating new samples. The model leverages advanced generative techniques to produce high-quality embeddings that capture the underlying structure of the data more effectively compared with the existing methods. G-NeuroDAVIS can be trained in both supervised and unsupervised settings. We have rigorously evaluated the model through a

series of experiments, demonstrating superior performance in several downstream tasks, which highlights the effectiveness of the learned representations. Results of an interpolation experiment have reflected a smooth and meaningful transition in the generated images across various paths, which in turn has depicted preservation of the underlying data structure. Furthermore, the conditional sample generation capability of the model has been described through both qualitative and quantitative assessments, revealing a marked improvement in generating realistic and diverse samples. G-NeuroDAVIS has outperformed VAE significantly in terms of embedding quality and downstream tasks like classification. Moreover, the superior sample generation capability of G-NeuroDAVIS has been demonstrated against VAE, DCGAN, DDPM, and AE-guided RealNVP. These results highlight the efficacy of G-NeuroDAVIS to serve as a robust tool in various applications that demand high-quality data generation and representation learning.

1.7.6 Chapter 7 - Conclusions and Scope for Further Research

In this chapter, we conclude the thesis and provide some possible directions for further research.

Chapter 2

NeuroDAVIS: A neural network model for data visualization

2.1 Introduction

In this chapter, we have introduced a novel unsupervised deep learning model, called NeuroDAVIS, which serves the purpose of data visualization while addressing the issue of simultaneous preservation of local and global structures. We have proved mathematically that local points remain local in the embedding produced by NeuroDAVIS, and demonstrated empirically on a wide variety of datasets that it also preserves global shape. The primary novelty of NeuroDAVIS lies in the input layer. It takes a sparse matrix as its input and produces an efficient embedding once trained successfully. NeuroDAVIS is a general-purpose Neural Network model that can be used for dimension reduction and visualization of high-dimensional datasets. It can extract a meaningful embedding from the data, which captures significant features that are indicative of the inherent data non-linearity. NeuroDAVIS-extracted features can be used for data reconstruction as well. Moreover, NeuroDAVIS is free from assumptions about the data distribution. It is a non-parametric method that is able to project new samples to an existing embedding. The performance of NeuroDAVIS has been evaluated on a wide variety of 2D synthetic and real high-dimensional datasets, including numeric, textual, image and biological data. Despite the limitations of t-SNE and UMAP regarding preservation of both local and global structures, they are still the current state-of-the-art in the universe of dimension reduction or visualization. NeuroDAVIS has been competitive against both t-SNE and UMAP, and

arguably preserves data shape, size, and local and global relationships better than both of them. It has been proved mathematically that the corresponding embeddings of local neighbors in high dimensions remain local in low dimensions, too. NeuroDAVIS applies to all kinds (modalities) of datasets. Furthermore, it is capable of producing impressive and interpretable visualizations independently, i.e., without any kind of preprocessing, like reducing the dimension or initializing the embedding by some other methods [41].

2.2 Methodology

This section develops the proposed unsupervised neural network model, called NeuroDAVIS, for visualization of high-dimensional datasets. NeuroDAVIS is a non-recurrent, feed-forward neural network, which is capable of extracting a low-dimensional embedding that can capture relevant features and provide efficient visualization of high-dimensional datasets. It can preserve both local and global relationships between clusters within the data. The motivation behind NeuroDAVIS is described below, followed by its architecture and learning.

2.2.1 Motivation

Preservation of local and global shapes is the two main components of a visualization challenge. Several approaches have been put forth so far, but none of them can suitably address both these issues together. The majority of them try to preserve local shape rather than the global shape, while some techniques accomplish the other [27]. Thus, developing a method that visualizes data at a lower dimension while successfully preserving both local and global distances has been our key goal.

The fundamental mathematical concept behind NeuroDAVIS is inspired by the well-studied regression problem. In a regression problem, a set of independent variables (regressors) is used to predict one or more dependent variables. In other words, during regression, only a suitable function from the independent feature space to the dependent feature space is learned, keeping both the feature spaces fixed. In a visualization problem, most of the methods first initialize a random embedding and then try to make it as informative as possible with respect to the original high-dimensional data. In other words, similar to a regression task, it also tries to learn a function from the latent space to

the original space. However, unlike regression, the latent space is not fixed, i.e., during the learning process, the latent space also gets updated. Thus, in order to visualize the data in hand, if a set of regressors can be learned in an unsupervised manner, then it serves the purpose of visualization. Motivated by this aspect, we have developed a neural network-based method, called NeuroDAVIS, that can simultaneously learn both the latent space and the function mapping from the latent space to the original space.

Visualization through regression

Let $\mathbf{X}^{(1)} = [x_{ik}^{(1)}]_{n \times d_1}$, $x_{ik}^{(1)} \in \mathbb{R}$ and $\mathbf{X}^{(2)} = [x_{ik}^{(2)}]_{n \times d_2}$, $x_{ik}^{(2)} \in \mathbb{R}$ be two data matrices having n samples characterized by d_1 and d_2 features respectively. An i^{th} sample from $\mathbf{X}^{(1)}$ is denoted as $\mathbf{x}_{i:}^{(1)}$ (i^{th} row of $\mathbf{X}^{(1)}$), and a k^{th} feature from $\mathbf{X}^{(1)}$ is denoted as $\mathbf{x}_{:k}^{(1)}$ (k^{th} column of $\mathbf{X}^{(1)}$). For a better readability, we use $\mathbf{x}_i^{(1)}$ instead of $\mathbf{x}_{i:}^{(1)}$ to represent an i^{th} sample, through the entire chapter. The task of regression is to realize a continuous function f from $\mathbb{R}^{d_1}$ to $\mathbb{R}^{d_2}$ such that the loss, usually $\sum_{i=1}^n \|f(\mathbf{x}_i^{(1)}) - \mathbf{x}_i^{(2)}\|^2$, gets minimized. For this purpose, one can use a multi-layer neural network model (Figure 2.1A). To ensure that the input data fits into the model appropriately, the input layer should have d_1 neurons. The input will then be processed through a number of nonlinear activation layers before reaching the output layer, which has d_2 neurons. A loss is then determined using computed and expected outputs, and the weights and biases are updated using conventional back-propagation learning. The crucial fact about this kind of learning is that it does not support the updation of the regressors. As a result, the multi-layer neural network learns a sufficiently complex function that is able to produce close predictions.

Let $\mathbf{X} = [x_i]_{n \times d}$, $x_i \in \mathbb{R}$ be another data matrix that needs to be visualized. In NeuroDAVIS, initially, random 2D or 3D regressors (depending upon whether the aim is to project to 2 dimensions or 3 dimensions), are created using a sparse input, to regress the original data, X . Now, there are multiple objectives with some relaxation. Throughout the learning process, an appropriate continuous function from the latent space to the original space, which can effectively produce the high-dimensional data, has to be learned, and at the same time, the regressors need to be learned, too.

As shown in the block diagram of NeuroDAVIS (Figure 2.1B), the dashed part is used to control the regressors. Input layer of NeuroDAVIS has n neurons, where n is the number of samples. Each neuron is used for controlling exactly one sample, and therefore, input

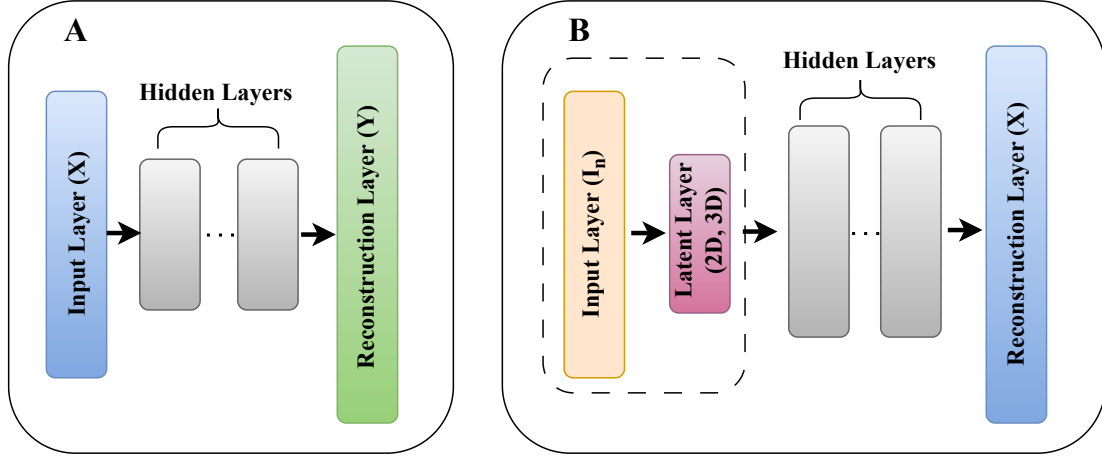


Figure 2.1: Block diagrams of (A) Multi-layer neural network and (B) NeuroDAVIS. The dashed part in (B) represents the modified architecture that enables the regressors to get updated.

to NeuroDAVIS is an identity matrix of order n . During the forward pass, each column of the identity matrix is fed to the input layer, thus ensuring that only certain weight values (initialized randomly) are fed as input to the latent layer. These random weights, generated by using the first two layers, are used as regressors. Similar to some multi-layer neural networks, like Autoencoder (AE) [43], NeuroDAVIS also tries to reconstruct the high-dimensional data at the reconstruction layer. A loss is then calculated, and the corresponding weight and bias values are updated using the standard back-propagation algorithm. Thus, in every epoch, as the weight values in the first layer get updated, the regressors get modified too.

The same task of data visualization can be achieved by an AE. At a glance, NeuroDAVIS looks similar to the standard AE since both of them produce an embedding by reconstructing the high-dimensional data. The major difference between these two networks lies in the input layer. AEs take the high-dimensional data as their input, encode it through a few hidden layers and reconstruct (decode) the input at the final layer. Unlike AEs, NeuroDAVIS takes an identity matrix as its input to control the samples, passes them through a few hidden layers and constructs the high-dimensional original data at the final layer. Moreover, NeuroDAVIS does not include an encoding layer.

2.2.2 Architecture

The architecture of NeuroDAVIS represents a novel neural network model as shown in Figure 2.2. NeuroDAVIS network consists of different types of layers, viz., an *Input layer*, a *Latent layer*, one or more *Hidden layer(s)* and a *Reconstruction layer*. Let $\mathbf{X} = [x_{ik}]_{n \times d}$ where $x_{ik} \in \mathbb{R}$ be a data matrix comprising n samples characterized by d features. Then the *Input layer* in NeuroDAVIS consists of n neurons. The number of neurons in the *Latent layer* is κ , where κ represents the number of dimensions to be used for visualization. In other words, κ is the number of dimensions at which the low-dimensional embedding is to be extracted. Usually, in real-life applications, a 2-dimensional ($\kappa = 2$) or 3-dimensional ($\kappa = 3$) visualization is desirable in order to understand the inherent structure of the data pictorially. The *Input layer* helps to create a random low-dimensional embedding of n samples (observations) at the *Latent layer*, which can be considered as an initial representative of the n original observations. The low-dimensional embedding at the *Latent layer* is projected onto the *Reconstruction layer* through one or more *Hidden layer(s)*. The *Reconstruction layer* tries to reconstruct the original d -dimensional space for the n samples from their random low-dimensional embedding. The number of neurons in the *Reconstruction layer* is thus d . In this work, we have used only two hidden layers. A *Hidden layer* tries to capture the non-linearity in the data and pass on the knowledge to the next layer in sequence. Thus, one can use multiple *Hidden layers* based on the complexity of the data. The number of *Hidden layers*, however, should not be large enough in order to avoid overfitting.

2.2.3 Forward propagation

As mentioned before, we have considered a data matrix of n samples $\mathbf{X} = [x_{ik}]_{n \times d}$ where $x_{ik} \in \mathbb{R}$. The input to NeuroDAVIS is an identity matrix $\mathbf{I}$ of order $n \times n$. For an i^{th} sample $\mathbf{x}_i$, we consider the i^{th} column vector $\mathbf{e}_i$ of $\mathbf{I}$. That is, on presenting $\mathbf{e}_i$ to the *Input layer*, an approximate version $\tilde{\mathbf{x}}_i$ of $\mathbf{x}_i$ is reconstructed at the *Reconstruction layer*. Let $\mathbf{a}_i^{(Inp)}$ and $\mathbf{h}_i^{(Inp)}$ correspond to the input to and output from the *Input layer*, and Let $\mathbf{a}_i^{(Lat)}$ and $\mathbf{h}_i^{(Lat)}$ correspond to the input to and output from the *Latent layer* on presentation of an i^{th} sample; also let $\mathbf{W}^{(Lat)}$ be the weight matrix between the *Input layer* and *Latent layer*, and $\mathbf{b}^{(Lat)}$ be the bias of the nodes in the *Latent layer*. Thus, for

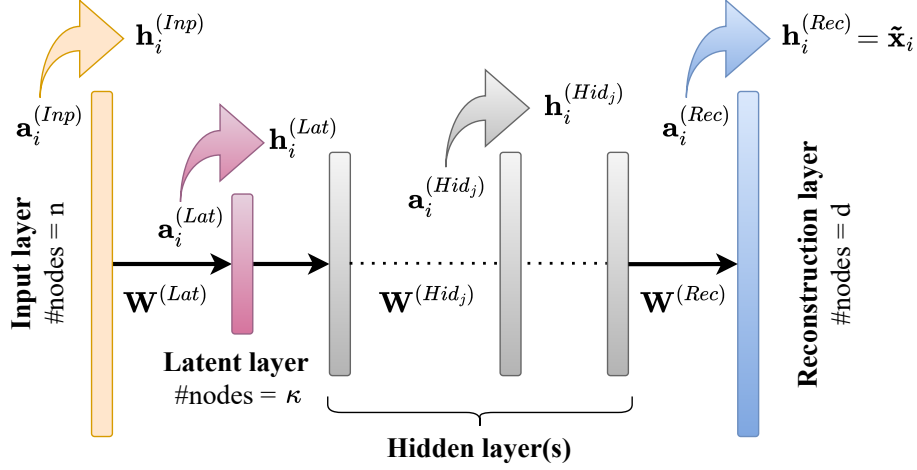


Figure 2.2: The network architecture of NeuroDAVIS, along with the mathematical notations.

the *Input layer*,

$$\begin{cases} \mathbf{a}_i^{(Inp)} = \mathbf{e}_i, \\ \mathbf{h}_i^{(Inp)} = \mathbf{a}_i^{(Inp)}, \quad \forall i \in \{1, 2, \dots, n\} \end{cases} \quad (2.1)$$

For the *Latent layer*, we have

$$\begin{cases} \mathbf{a}_i^{(Lat)} = \mathbf{W}^{(Lat)} \mathbf{h}_i^{(Inp)} + \mathbf{b}^{(Lat)}, \\ \mathbf{h}_i^{(Lat)} = \mathbf{a}_i^{(Lat)}, \quad \forall i \in \{1, 2, \dots, n\} \end{cases} \quad (2.2)$$

Here, $\mathbf{e}_i$ controls the weight parameters for the low-dimensional embedding of the i^{th} sample obtained at the *Latent layer*. It ensures that only the links connected to the i^{th} neuron of the *Input layer* will activate neurons in *Latent layer* on presentation of i^{th} sample.

Let the number of *Hidden layers* be l , and $\mathbf{a}_i^{(Hid_j)}$ and $\mathbf{h}_i^{(Hid_j)}$ be the input to and output from a j^{th} *Hidden layer* on presentation of an i^{th} sample. Also, let $\mathbf{W}^{(Hid_j)}$ and $\mathbf{b}^{(Hid_j)}$ be the weight matrix and bias of the nodes in j^{th} *Hidden layer*. Therefore, for the first *Hidden layer*, we have

$$\begin{cases} \mathbf{a}_i^{(Hid_1)} = \mathbf{W}^{(Hid_1)} \mathbf{h}_i^{(Lat)} + \mathbf{b}^{(Hid_1)}, \\ \mathbf{h}_i^{(Hid_1)} = \text{ReLU}(\mathbf{a}_i^{(Hid_1)}), \quad \forall i \in \{1, 2, \dots, n\} \end{cases} \quad (2.3)$$

For the $(l - 1)$ *Hidden layer(s)*, we have

$$\begin{cases} \mathbf{a}_i^{(Hid_j)} = \mathbf{W}^{(Hid_j)} \mathbf{h}_i^{(Hid_{j-1})} + \mathbf{b}^{(Hid_j)}, \\ \mathbf{h}_i^{(Hid_j)} = ReLU(\mathbf{a}_i^{(Hid_j)}), \end{cases} \quad \forall i \in \{1, 2, \dots, n\}, \quad \forall j \in \{2, 3, \dots, l\} \quad (2.4)$$

where $ReLU(\mathbf{y}) = \max(\mathbf{0}, \mathbf{y})$; $\max$ (maximum) being considered element wise.

A reconstruction of the original data is performed at the final layer, called *Reconstruction layer*. Let $\mathbf{a}_i^{(Rec)}$ and $\mathbf{h}_i^{(Rec)}$ be the input to and output from the *Reconstruction layer* on presentation of an i^{th} sample. Also, let $\mathbf{W}^{(Rec)}$ and $\mathbf{b}^{(Rec)}$ be the weight matrix and bias of the nodes in the *Reconstruction layer*. Thus, for the *Reconstruction layer*, we have

$$\begin{cases} \mathbf{a}_i^{(Rec)} = \mathbf{W}^{(Rec)} \mathbf{h}_i^{(Hid_l)} + \mathbf{b}^{(Rec)}, \\ \mathbf{h}_i^{(Rec)} = \mathbf{a}_i^{(Rec)}, \end{cases} \quad \forall i \in \{1, 2, \dots, n\} \quad (2.5)$$

Thus, NeuroDAVIS projects the latent embedding for a sample obtained at the *Latent layer* into a d dimensional space through the *Hidden layer(s)*.

2.2.4 Learning

As stated before, NeuroDAVIS has been used for dimensionality reduction and visualization of high-dimensional data. For each sample $\mathbf{x}_i$, NeuroDAVIS tries to find an optimal reconstruction $\tilde{\mathbf{x}}_i$ of the input data by minimizing the reconstruction error $\|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|$. $L2$ regularization has been used on the nodes' activities and edge weights to avoid overfitting. Usage of $L2$ regularization ensures minimization of model complexity. It also helps in dragging output and weight values towards zero. Regularization of the nodes' activities and weights has been controlled using regularization parameters λ_1 and λ_2 , respectively. The objective function thus becomes

$$\mathcal{L}_{\text{NeuroDAVIS}} = \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|_2^2 + \lambda_1 \sum_{i=1}^n \sum_{j=1}^l \|\mathbf{h}_i^{(Hid_j)}\|_2^2 + \lambda_2 \sum_{j=1}^l \|\mathbf{W}^{(Hid_j)}\|_F^2 \quad (2.6)$$

where $\|\cdot\|_F$ indicates the Frobenius norm of a matrix, defined as the square root of the sum of the squares of the matrix elements.

NeuroDAVIS has been trained using Adam optimizer [65]. Values of λ_1 and λ_2 have

been set experimentally for each dataset. Additionally, the number of epochs at which $\mathcal{L}_{NeuroDAVIS}$ saturates has been observed carefully to fix the optimal number of epochs needed for convergence.

At the onset of the forward pass, the weight values between the *Input layer* and *Latent layer* are initialized randomly. The learning of the NeuroDAVIS network is controlled by an identity matrix $\mathbf{I}$ fed as input to the *Input layer*. The usage of the identity matrix ensures updation of the weight values solely associated with the samples in the present batch. Thus, initially, a random low-dimensional embedding of n samples (observations) is created at the *Latent layer*. This low-dimensional representation at the *Latent layer* is then projected onto the *Reconstruction layer* by NeuroDAVIS. At the *Reconstruction layer*, the error value is calculated using Equation 2.6. This error is then propagated backwards, and the weight and bias values are updated for a better reconstruction of the samples of the present batch, in the next forward pass. This process is continued till convergence. On completion of the training phase, the transformed feature set extracted from the *Latent layer* serves as the low-dimensional embedding of the data, and is used for visualization for $\kappa = 2$ or 3 .

2.2.5 Projecting new observations

In the task of visualization, the major advantage of neural network models, is that they can be used to visualize new samples after successful training. NeuroDAVIS is not an exception and can project new samples to an existing embedding as well. Regression-based visualization methods consist of two major learning steps. First, the regressors used to regress the original data need to be learned. Second, an appropriate parametric function from the latent space to the original space also needs to be learned. NeuroDAVIS also contains two separate modules that serve the above-mentioned tasks. The first two layers, viz., *Input layer* and *Latent layer*, have been used to project the data into the latent space, and thus the weights between these two layers, i.e., $\mathbf{W}^{(Lat)}$, help to learn suitable regressors. On the other hand, the mapping from *Latent layer* to *Reconstruction layer* has been used to find a proper parametric function from the latent space to the original space.

Once NeuroDAVIS is trained over a dataset, these training observations can be visualized by the *Latent layer*. To visualize new observations from a similar distribution but are not included in the training dataset, a similar architecture with a resized *Input*

layer (based on the number of new observations) is considered. The weights $\mathbf{W}^{(Hid_j)}$, $\forall j \in \{1, 2, \dots, l\}$, already learned during training, remain unchanged during visualization of these new observations, while already learned $\mathbf{W}^{(Lat)}$ is discarded. Let the test data have n_1 samples. To visualize these data, an identity matrix of order n_1 is fed to the input layer, consisting of n_1 nodes, of the new model. These inputs are propagated up to the *Reconstruction layer* of the network with a newly initialized $\mathbf{W}^{(Lat)}$, creating an approximation of the test data points. A loss is then calculated using Equation 2.6 and the learning proceeds through back-propagation. This time only the new weights, $\mathbf{W}^{(Lat)}$, are updated, keeping the other weight values unaltered. After convergence, the embedding for the new samples can be obtained at the *Latent layer* of the new model.

2.3 Proof of correctness

In this section, we have established mathematically that neighbours in high dimensions remain preserved after projection into a low dimension.

Lemma 1. *For any real matrix $\mathbf{W}$ having Frobenius norm less than or equal to 1 and $\eta \leq 1$, $\|\mathbf{I} - \eta\mathbf{W}\mathbf{W}^T\|_2 \leq 1$.*

Proof. It is well known that

$$\|\mathbf{I} - \eta\mathbf{W}\mathbf{W}^T\|_2^2 = \lambda_{max}(\mathbf{I} - \eta\mathbf{W}\mathbf{W}^T)^2,$$

where $\lambda_{max}(A)$ represents the largest eigen value of A .

Now, $\mathbf{W}\mathbf{W}^T$ is a positive semi-definite matrix. Hence, all eigen values of $\mathbf{W}\mathbf{W}^T$ are non-negative. Also,

$$\lambda_{max}(\mathbf{W}\mathbf{W}^T) \leq \|\mathbf{W}\|_F \leq 1$$

Therefore, all eigen values of $\mathbf{W}\mathbf{W}^T$ lie in $[0, 1]$, which implies $\lambda_{max}(\mathbf{I} - \eta\mathbf{W}\mathbf{W}^T)^2 \leq 1$ for $\eta \leq 1$ □

Theorem 1. *Let $\mathbf{x}_u$ and $\mathbf{x}_v$ be two points in d -dimensional space ($\mathbf{x}_u, \mathbf{x}_v \in \mathbb{R}^d$) such that they belong to a δ -ball, i.e., $\|\mathbf{x}_u - \mathbf{x}_v\| < \delta$, δ being a predefined small positive number. Their corresponding low-dimensional embedding, generated by NeuroDAVIS, will come*

closer in each iteration during training if the weight matrix connected to the Reconstruction layer has a Frobenius norm less than 1.

Proof. Let us consider a simple NeuroDAVIS model with no hidden layer and no regularization. Let $\mathbf{y}_u$ and $\mathbf{y}_v$ be the corresponding initial low-dimensional embeddings of $\mathbf{x}_u$ and $\mathbf{x}_v$. Then

$$\mathbf{y}_u = \mathbf{W}^{(Lat)} \mathbf{e}_u + \mathbf{b}^{(Lat)} = \mathbf{W}_{:u}^{(Lat)} + \mathbf{b}^{(Lat)},$$

$\mathbf{e}_u$ being the u^{th} row of $\mathbf{I}$, and $\mathbf{W}_{:u}^{(Lat)}$ be the u^{th} column of $\mathbf{W}^{(Lat)}$. During each forward pass, NeuroDAVIS tries to reconstruct the original data point. Let $\tilde{\mathbf{x}}_u$ be a reconstruction of $\mathbf{x}_u$, i.e.,

$$\tilde{\mathbf{x}}_u = (\mathbf{W}_{:u}^{(Lat)} + \mathbf{b}^{(Lat)}) \mathbf{W}^{(Rec)} + \mathbf{b}^{(Rec)}$$

Without loss of generality, let the loss function be as follows:

$$\mathcal{L}_{NeuroDAVIS} = \sum_{i=1}^n \|\mathbf{x}_u - \tilde{\mathbf{x}}_u\|^2,$$

The corresponding weight and bias values are updated using

$$\mathbf{W}(t+1) = \mathbf{W}(t) - \eta \nabla_{\mathbf{W}(t)} \mathcal{L}$$

Therefore,

$$\begin{aligned} \frac{\partial \mathcal{L}_{NeuroDAVIS}}{\partial \mathbf{W}_{:u}^{(Lat)}} &= (\mathbf{x}_u - \tilde{\mathbf{x}}_u) \left(-\frac{\partial \tilde{\mathbf{x}}_u}{\partial \mathbf{W}_{:u}^{(Lat)}} \right) \\ &= -(\mathbf{x}_u - \tilde{\mathbf{x}}_u) \frac{\partial}{\partial \mathbf{W}_{:u}^{(Lat)}} (\mathbf{W}_{:u}^{(Lat)} \mathbf{W}^{(Rec)}) \\ &= -(\mathbf{x}_u - \tilde{\mathbf{x}}_u) \mathbf{W}^{(Rec)T} \end{aligned}$$

At t^{th} iteration, the updated weight is

$$\mathbf{W}_{:u}^{(Lat)}(t+1) = \mathbf{W}_{:u}^{(Lat)}(t) + \eta (\mathbf{x}_u - \tilde{\mathbf{x}}_u) \mathbf{W}^{(Rec)T}(t)$$

Thus,

$$\begin{aligned}
& \mathbf{y}_u(t+1) - \mathbf{y}_v(t+1) \\
&= \mathbf{W}_{:u}^{(Lat)}(t+1) - \mathbf{W}_{:v}^{(Lat)}(t+1) \\
&= [\mathbf{W}_{:u}^{(Lat)}(t) + \eta(\mathbf{x}_u - \tilde{\mathbf{x}}_u)\mathbf{W}^{(Rec)T}(t)] - [\mathbf{W}_{:v}^{(Lat)}(t) + \eta(\mathbf{x}_v - \tilde{\mathbf{x}}_v)\mathbf{W}^{(Rec)T}(t)] \\
&= [\mathbf{W}_{:u}^{(Lat)}(t) - \mathbf{W}_{:v}^{(Lat)}(t)] + \eta[(\mathbf{x}_u - \mathbf{x}_v) - (\tilde{\mathbf{x}}_u - \tilde{\mathbf{x}}_v)]\mathbf{W}^{(Rec)T}(t) \\
&\approx [\mathbf{W}_{:u}^{(Lat)}(t) - \mathbf{W}_{:v}^{(Lat)}(t)] - \eta(\tilde{\mathbf{x}}_u - \tilde{\mathbf{x}}_v)\mathbf{W}^{(Rec)T}(t) \quad [\text{as } \|\mathbf{x}_u - \mathbf{x}_v\| < \delta] \\
&\approx [\mathbf{W}_{:u}^{(Lat)}(t) - \mathbf{W}_{:v}^{(Lat)}(t)] - \eta[\mathbf{W}_{:u}^{(Lat)}(t) - \mathbf{W}_{:v}^{(Lat)}(t)]\mathbf{W}^{(Rec)}(t)\mathbf{W}^{(Rec)T}(t) \\
&\approx [\mathbf{W}_{:u}^{(Lat)}(t) - \mathbf{W}_{:v}^{(Lat)}(t)][\mathbf{I} - \eta\mathbf{W}^{(Rec)}(t)\mathbf{W}^{(Rec)T}(t)]
\end{aligned}$$

$$\begin{aligned}
\|\mathbf{y}_u(t+1) - \mathbf{y}_v(t+1)\| &\approx \|[\mathbf{W}_{:u}^{(Lat)}(t) - \mathbf{W}_{:v}^{(Lat)}(t)][\mathbf{I} - \eta\mathbf{W}^{(Rec)}(t)\mathbf{W}^{(Rec)T}(t)]\| \\
&\leq \|\mathbf{W}_{:u}^{(Lat)}(t) - \mathbf{W}_{:v}^{(Lat)}(t)\| \|\mathbf{I} - \eta\mathbf{W}^{(Rec)}(t)\mathbf{W}^{(Rec)T}(t)\| \\
&\leq \|\mathbf{W}_{:u}^{(Lat)}(t) - \mathbf{W}_{:v}^{(Lat)}(t)\| \quad [\text{By Lemma 1}]
\end{aligned}$$

Thus,

$$\|\mathbf{y}_u(t+1) - \mathbf{y}_v(t+1)\| \leq \|\mathbf{y}_u(t) - \mathbf{y}_v(t)\|$$

□

2.4 Results

The performance of NeuroDAVIS has been evaluated on a wide variety of datasets, including 2D synthetic datasets and high-dimensional (HD) datasets of different modalities, like numeric, text, image and biological data (single-cell RNA-sequencing (scRNA-seq) data). The reason for using 2D datasets is that the original cluster shapes and sizes for them are well known. Additionally, evaluating performance on HD datasets by generating low-dimensional embeddings and subjecting them to downstream analyses like clustering and classification ensures that the method is efficient and robust enough. The datasets used for evaluation have been described in Table 2.1. This section explains the experiments carried out in this Chapter to evaluate the performance of NeuroDAVIS and analyze

Table 2.1: Description of datasets used for evaluation of NeuroDAVIS

Category	Name	#Samples	#Features	#Classes	Source
2D	<i>Elliptic Ring</i>	1100	2	3	Synthetic
	<i>Olympic</i>	2500	2	5	
	<i>Spiral</i>	312	2	3	
	<i>Shape</i>	2000	2	5	
	<i>World Map</i>	2843	2	5	
HD	Numeric	<i>Breast Cancer</i>	569	30	[66]
		<i>Wine</i>	178	13	[67]
	Text	<i>Spam</i>	5572	513	[68]
		<i>Coil20</i>	1440	16385	[69]
	Image	<i>FMNIST</i>	60000	784	[70]
		<i>Usoskin</i>	622	25334	[71]
	scRNA-seq	<i>Jurkat</i>	3388	32738	[72]

the results. Here we have focused on the capability of NeuroDAVIS in low dimensional embedding, visualization, structure preservation, inter-cluster distance preservation and cluster size preservation.

2.4.1 Performance evaluation on synthetic 2D datasets

In order to demonstrate the effectiveness of NeuroDAVIS, we have initially applied it on four synthetic 2D datasets, viz., *Elliptic Ring*, *Olympic*, *Spiral* and *Shape*. The results have been compared with those obtained using t-SNE and UMAP. First, we have observed the embeddings produced by NeuroDAVIS and the other state-of-the-art methods. We have then analyzed the global shape, inter-cluster distance and cluster size preservation capabilities of NeuroDAVIS against those of the state-of-the-art methods. These experiments have helped us understand the above-mentioned properties since the ground truth is well known for 2D datasets.

Embedding, visualization and structure preservation

Elliptic Ring dataset (Figure 2.3(A)) consists of two small Gaussian balls within an outer Gaussian elliptic ring; *Olympic* dataset (Figure 2.3(E)) contains five circular rings representing the olympic logo; *Spiral* dataset (Figure 2.3(I)) contains three concentric

spirals, while *Shape* dataset (Figure 2.3(M)) consists of points representing the characters ‘S’, ‘H’, ‘A’, ‘P’ and ‘E’.

We have observed that the NeuroDAVIS-generated embedding of *Elliptic Ring* dataset (Figure 2.3(B)) contains two Gaussian balls inside the outer Gaussian ring, similar to the original data. Likewise, the distances between the rings in *Olympic* dataset and those between the concentric spirals in *Spiral* dataset, as well as their shapes, have been preserved in their NeuroDAVIS-generated embeddings (Figures 2.3(F) and 2.3(J)). Furthermore for *Shape* dataset, the readability of the characters ‘S’, ‘H’, ‘A’, ‘P’ and ‘E’ have not been compromised by NeuroDAVIS (Figure 2.3(N)), which is totally lost in the cases of t-SNE (Figure 2.3(O)) and UMAP (Figure 2.3(P)) embeddings. Thus, we can say that NeuroDAVIS has represented the clusters in the data, preserving both their shape and size, similar to their original distributions. Both t-SNE (Figure 2.3(C), 2.3(G), 2.3(K) and 2.3(O)) and UMAP (Figure 2.3(D), 2.3(H), 2.3(L) and 2.3(P)), however, have failed to preserve original distributions, and produced compact clusters instead, disrupting both local and global structures within the data.

We have further compared the pairwise distances in the original distribution to those in the embedding produced by t-SNE, UMAP and NeuroDAVIS, using a Spearman rank correlation. Figure 2.4 shows the results for ten different executions of the same. We have observed that the correlation coefficient values obtained by NeuroDAVIS have been better than those produced by t-SNE and UMAP for all the datasets.

We have then performed another experiment on these synthetic datasets, projecting the two-dimensional space into nine-dimensional space using the transformation $(x + y, x - y, xy, x^2, y^2, x^2y, xy^2, x^3, y^3)$ as performed in [30], and applying NeuroDAVIS to project the data back into a two-dimensional space. Results have been compared to those obtained by t-SNE and UMAP for the same transformation. Figure A.I.1 (in Appendices) shows that NeuroDAVIS has once again produced better embeddings, preserving both cluster shapes and sizes, as compared to t-SNE and UMAP.

Global structure preservation

In order to assess the performance of a dimension reduction/visualization method, we need to analyze the reduced dimensional embedding in terms of its capability to represent clusters in a way similar to their distribution in the high dimension. For this purpose, it is

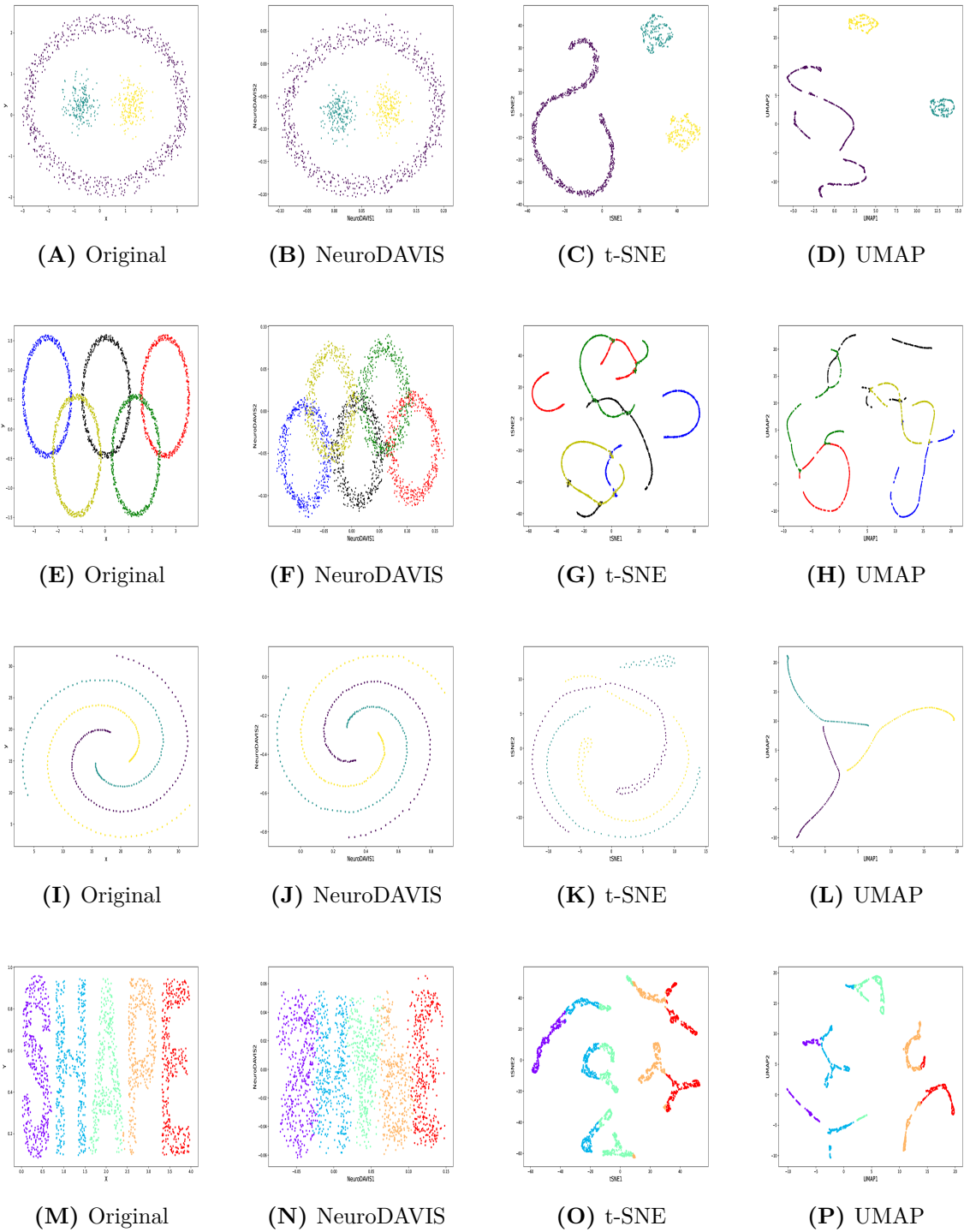


Figure 2.3: The original distributions and the embeddings produced by NeuroDAVIS, t-SNE and UMAP for *Elliptic Ring* ((A)-(D)), *Olympic* ((E)-(H)), *Spiral* ((I)-(L)) and *Shape* ((M)-(P)) datasets.

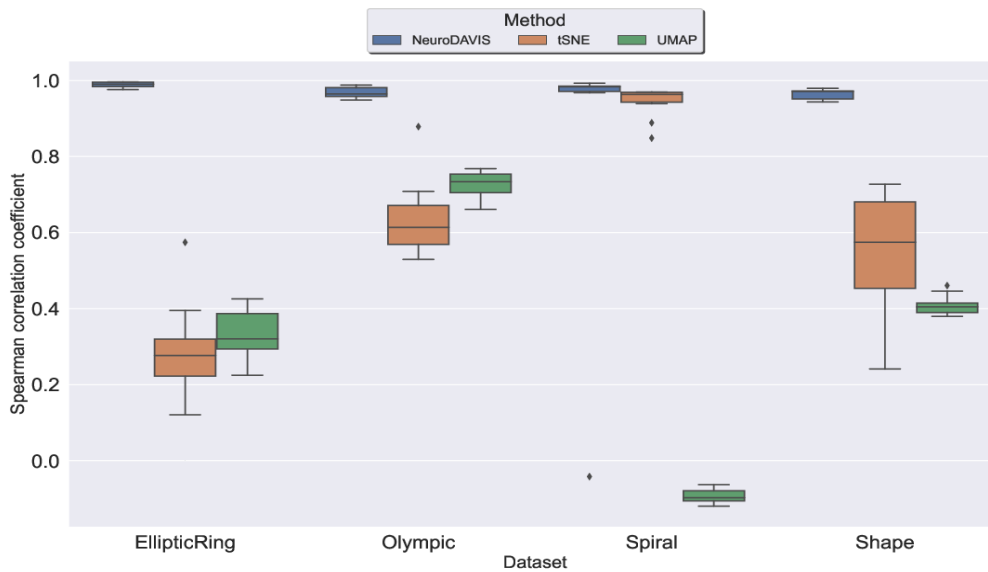


Figure 2.4: Spearman rank correlation between pairwise distances in the original distribution, and pairwise distances in t-SNE, UMAP and NeuroDAVIS-produced embeddings of the 2D synthetic datasets *Elliptic Ring*, *Olympic*, *Spiral* and *Shape*. For *Elliptic Ring* dataset, median correlation coefficient values obtained are 0.98 (NeuroDAVIS), 0.27 (t-SNE) and 0.32 (UMAP). For *Olympic* dataset, median correlation coefficient values obtained are 0.96 (NeuroDAVIS), 0.61 (t-SNE) and 0.73 (UMAP). For *Spiral* dataset, median correlation coefficient values obtained are 0.98 (NeuroDAVIS), 0.96 (t-SNE) and -0.09 (UMAP). For *Shape* dataset, median correlation coefficient values obtained are 0.97 (NeuroDAVIS), 0.57 (t-SNE) and 0.40 (UMAP). The corresponding p – values obtained using the Mann-Whitney U test are 0.0001 (both NeuroDAVIS-t-SNE and NeuroDAVIS-UMAP) for *Elliptic Ring*, *Olympic* and *Shape* datasets. For *Spiral* dataset, p – values obtained are 0.009 (NeuroDAVIS-t-SNE) and 0.001 (NeuroDAVIS-UMAP).

essential to study the inter-cluster separations within the data. Hence, we have performed a few more experiments to ascertain how well NeuroDAVIS preserves the inter-cluster distances in the low-dimensional space. Information on the ground truth is important for this kind of analysis. For this reason, we have created a synthetic 2D dataset representing the world map (a known structure) with five clusters/continents, viz., Eurasia, Australia, North America, South America and Africa (no Antarctica), following the tutorial ¹ (Figure 2.5(A)). The NeuroDAVIS-generated embedding for *World Map* dataset has been compared with its original data distribution, as well as the corresponding embeddings generated by t-SNE and UMAP (Figure 2.5). As evident from Figure 2.5, unlike t-SNE and UMAP, the shapes and sizes of the clusters in the NeuroDAVIS-generated embedding are quite similar to those in the original data. The Eurasia cluster has been split into two different halves in both t-SNE and UMAP embedding. However, NeuroDAVIS has been able to

¹<https://towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17>

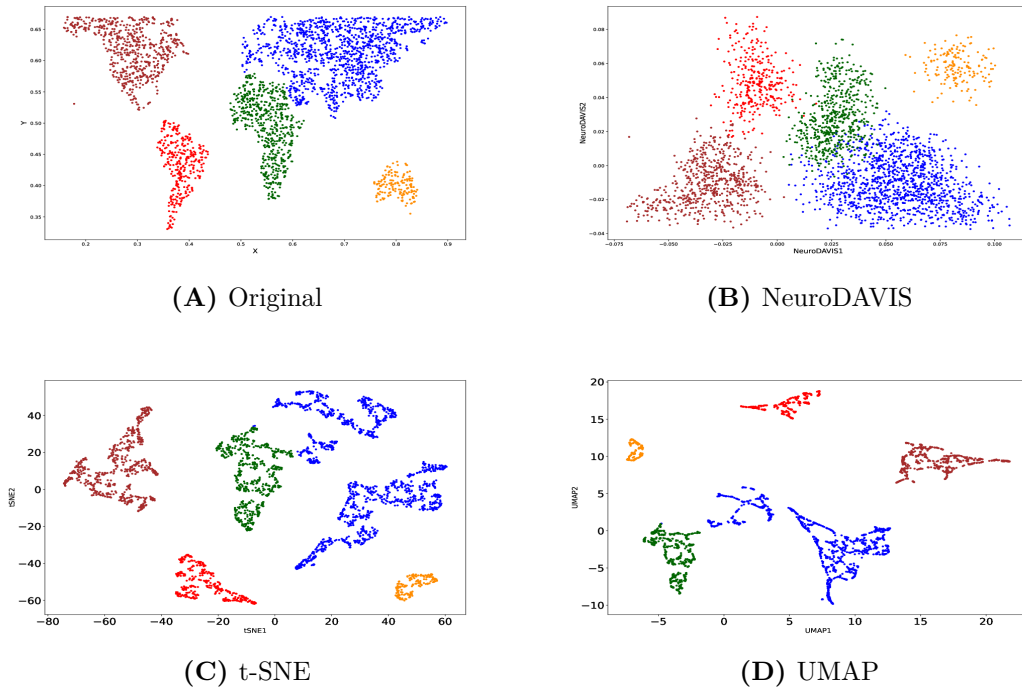


Figure 2.5: The original distribution and the embeddings produced by NeuroDAVIS, t-SNE and UMAP for *World Map* dataset.

keep it intact, similar to the original data.

Inter-cluster distance preservation

In order to check if the inter-cluster distances between the five continents have been preserved, we have calculated the Spearman rank correlation coefficients between the distances among the five cluster centroids obtained by NeuroDAVIS, and compared the result with those obtained using t-SNE and UMAP. As shown in Figure 2.6, repeated runs of the same experiment have confirmed that NeuroDAVIS has performed much better than t-SNE and UMAP, with respect to distance preservation.

We have further evaluated the NeuroDAVIS-embedding for *World Map* dataset for its ability to preserve pairwise euclidean distances between points in one cluster and points in the remaining clusters. Figure 2.7 containing results from repeated runs of this experiment, shows that for each of the clusters Eurasia, Australia, North America, South America and Africa, NeuroDAVIS has been able to preserve the original pairwise distances better than t-SNE and UMAP.

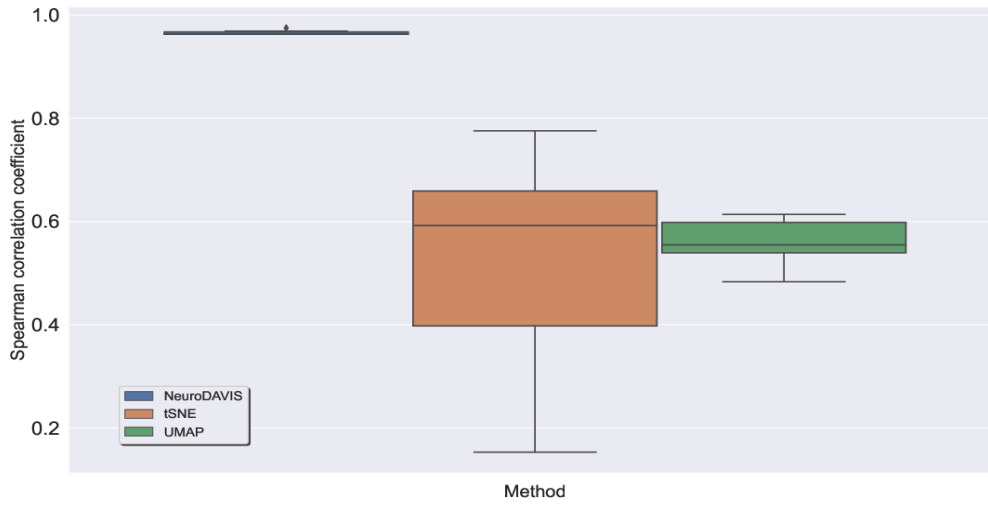
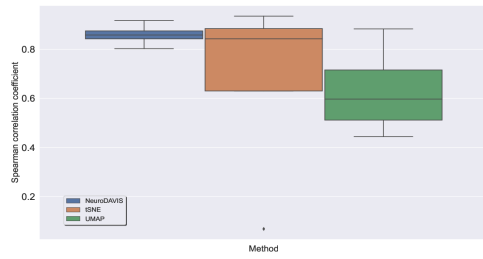


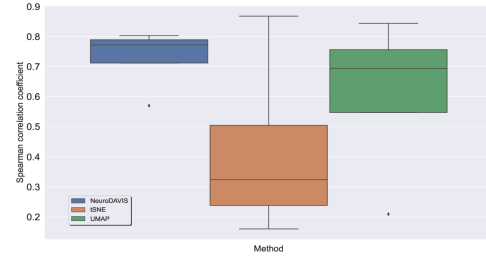
Figure 2.6: Spearman rank correlation between original distances among cluster centroids in the original dataset and distances between cluster centroids in the NeuroDAVIS, t-SNE and UMAP-generated embedding. The median correlation coefficient values are 0.96 (NeuroDAVIS), 0.59 (t-SNE) and 0.55 (UMAP) for *World Map* dataset. The corresponding p – value obtained using Mann-Whitney U test is 0.0001 (both NeuroDAVIS-t-SNE and NeuroDAVIS-UMAP).

Cluster size preservation

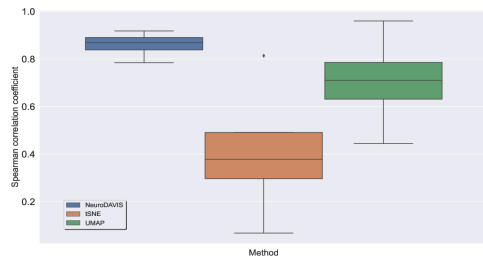
We wondered whether the sizes of the clusters in the original data were preserved in the NeuroDAVIS-generated embedding. Therefore, in order to ensure preservation of cluster sizes, we have carried out a few additional experiments on this *World Map* dataset. For each of the clusters in the dataset, we have estimated the area of the minimal rectangles bordering the clusters. This has been realized by considering the amount of spread of each cluster in either direction, as shown in Figure 2.5(A). We have then computed the Pearson correlation coefficients between the original area of the clusters and their reconstructed counterparts. Repeating this experiment multiple times, we have concluded that NeuroDAVIS has been able to preserve the original sizes of the clusters better than t-SNE and UMAP, as shown in Figure A.I.2 (in Appendices). All these experiments discussed above establish the superiority of NeuroDAVIS over t-SNE and UMAP in terms of global shape preservation.



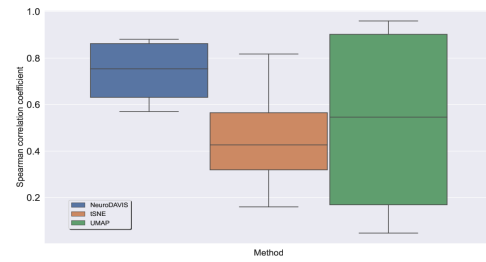
(A) Distances between Eurasia and others



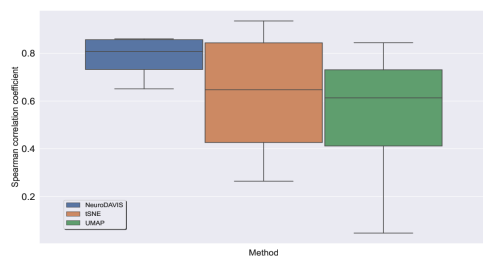
(B) Distances between Australia and others



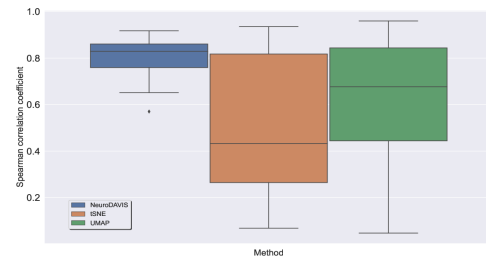
(C) Distances between North America and others



(D) Distances between South America and others



(E) Distances between Africa and others



(F) Distances between all pairs of continents

Figure 2.7: Preservation of pairwise distances between clusters (continents) in *World Map* dataset by NeuroDAVIS, t-SNE and UMAP.

2.4.2 Performance evaluation on high-dimensional datasets

Subsequently, the effectiveness of NeuroDAVIS has been demonstrated on several categories of high-dimensional datasets, including numeric, textual, image and biological data. The general steps used for evaluation are as follows.

We first reduced each dataset to two NeuroDAVIS dimensions. For each dataset, the quality of the NeuroDAVIS embedding has then been compared with that produced by each of t-SNE, UMAP, Fit-SNE, PCA and RP in two different aspects. First, we have calculated the Spearman correlation coefficient between the pairwise distances of the data points in the original space and those in the latent space, produced by NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP. Finally, for all the high-dimensional datasets excluding the biological one, the embedding produced by NeuroDAVIS has been assessed for its classification performance. A train-test split in 80 : 20 ratio has been used to generate training and test datasets from the NeuroDAVIS-generated embedding. Two classifiers, viz., k-NN and random forest (RF), have been trained, and the held-out test dataset has been used to evaluate the performance of the classifiers. The results have been compared with those obtained using the same classifiers on t-SNE, UMAP, Fit-SNE, PCA and RP-embeddings.

For the scRNA-seq (biological) dataset, we have performed cell-type clustering using k-means and hierarchical agglomerative clustering on the NeuroDAVIS-embedding, and compared the results with those obtained on t-SNE, UMAP, Fit-SNE, PCA and RP embeddings. We have further compared the results obtained with two recently benchmarked methods, viz., PHATE [42] and IVIS [50], specially developed for scRNA-seq datasets. In order to compare the performance of NeuroDAVIS with the other methods, we have used some measures used in [73]. Classification performance for the non-biological datasets has been measured using Accuracy and F1-scores, while two external indices - Adjusted Rand Index (ARI) and Fowlkes-Mallows Index (FMI) have been used to measure the quality of clusters produced in the case of scRNA-seq datasets. F1-score conveys the precision (Positive predicted value) and Recall (True positive rate) of the model since it is the harmonic mean of precision and recall, and gives a better insight into the performance of a classifier, when there is an imbalance in class labels. Thus, Accuracy and F1-scores have been reported by varying the parameters (k and $n_estimators$) of the k-NN and

RF classifiers for the imbalanced datasets considered here. The performance of t-SNE and UMAP has been evaluated by varying the parameters, which are sensitive to tiny modifications, like perplexity for t-SNE, and $n_neighbours$ and min_dist for UMAP, and the corresponding embeddings having maximum Spearman correlation have been used for comparison (Figures A.II.1 - A.II.12 in Appendices).

Numeric datasets

Two numeric datasets - *Breast Cancer* and *Wine*, originally available at [66], have been downloaded from [74] in order to evaluate the performance of NeuroDAVIS. The NeuroDAVIS-generated embedding as compared to those obtained using t-SNE, UMAP, Fit-SNE, PCA and RP for *Breast Cancer* dataset are shown in Figures 2.8(A), 2.8(B), 2.8(C), 2.8(D), 2.8(E) and 2.8(F), while Figures 2.8(G), 2.8(H), 2.8(I), 2.8(J), 2.8(K) and 2.8(L) show similar embeddings for *Wine* dataset. The correlation coefficient values, as shown in Figure 2.9, reveal that pairwise distances in the original space are more correlated with those in NeuroDAVIS (median correlation coefficient = 0.94 (*Breast Cancer*), 0.92 (*Wine*)), PCA (median correlation coefficient = 0.99 (*Breast Cancer*), 0.99 (*Wine*)) and RP embedding (median correlation coefficient = 0.98 (*Breast Cancer*), 0.99 (*Wine*)), than those in t-SNE (median correlation coefficient = 0.76 (*Breast Cancer*), 0.91 (*Wine*)), UMAP (median correlation coefficient = 0.77 (*Breast Cancer*), 0.82 (*Wine*)) and Fit-SNE embedding (median correlation coefficient = 0.77 (*Breast Cancer*), 0.92 (*Wine*)). The corresponding p - values obtained using the Mann-Whitney U test are 0.0001 (for all pairs, viz., NeuroDAVIS-t-SNE, NeuroDAVIS-UMAP, NeuroDAVIS-Fit-SNE, NeuroDAVIS-PCA and NeuroDAVIS-RP) for *Breast Cancer* dataset, and 0.0001 (NeuroDAVIS-UMAP, NeuroDAVIS-PCA and NeuroDAVIS-RP), 0.733 (NeuroDAVIS-t-SNE) and 1.000 (NeuroDAVIS-Fit-SNE) for *Wine* dataset.

The results for classification (accuracy and F1-score) using k-NN and RF, on the NeuroDAVIS-embedding, as shown in Figures A.III.1 - A.III.5 (in Appendices), also reflect superior performance of NeuroDAVIS over t-SNE, UMAP, Fit-SNE, PCA and RP for *Wine* dataset, while for *Breast Cancer* dataset, NeuroDAVIS-embedding has displayed comparable classification performance to t-SNE, UMAP, Fit-SNE, PCA and RP embeddings. Results have been recorded by repeating each experiment multiple times.

Textual dataset

Next, we have used a textual dataset to demonstrate the effectiveness of NeuroDAVIS. *Spam* dataset, originally available at [68], has been downloaded ² in this regard. Out of 5572 messages in this dataset, 747 messages are classified as spam messages. This dataset has undergone preprocessing using standard pipelines used for textual datasets, which include the following major steps:

1. URLs removal,
2. Lower case conversion,
3. Punctuations removal,
4. White-spaces removal,
5. Stopwords removal,
6. Lemmatization,
7. Tokenization, and
8. min-max scaling.

The two-dimensional embedding obtained by NeuroDAVIS on the preprocessed *Spam* dataset has been shown in Figure 2.8(M). The corresponding t-SNE, UMAP, Fit-SNE, PCA and RP embeddings on the same dataset have been shown in Figures 2.8(N), 2.8(O), 2.8(P), 2.8(Q) and 2.8(R) respectively. Although the embeddings look quite similar to each other, except RP-embedding, on close observation, it can be seen that, quite contrary to NeuroDAVIS, UMAP and PCA, the ham (not spam) cluster in both the t-SNE and Fit-SNE embedding portrays multiple sub-clusters. Figure 2.9 also reveals that the correlation coefficient values for the NeuroDAVIS embedding (median correlation coefficient = 0.77) is far better than those obtained for the t-SNE (median correlation coefficient = 0.59), UMAP (median correlation coefficient = 0.57), Fit-SNE (median correlation coefficient = 0.61) and RP (median correlation coefficient = 0.54) embeddings and only falls behind PCA (median correlation coefficient = 0.86). The corresponding p – value obtained using the Mann–Whitney U test is 0.0001 (for NeuroDAVIS-t-SNE, NeuroDAVIS-UMAP, NeuroDAVIS-Fit-SNE, NeuroDAVIS-PCA and NeuroDAVIS-RP). The classification performance on NeuroDAVIS embedding of *Spam* dataset is, however, not as good as that on the t-SNE, UMAP, Fit-SNE, PCA and RP embeddings (Figures A.III.1 - A.III.5 in Appendices).

²<https://www.kaggle.com/datasets/team-ai/spam-text-message-classification>

Image datasets

Subsequently, we have evaluated NeuroDAVIS on an image dataset, *Coil20*. *Coil20* dataset [69] contains images of 20 different objects with backgrounds discarded. Figures 2.8(S), 2.8(T), 2.8(U), 2.8(V), 2.8(W), and 2.8(X) show the NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP embeddings obtained on this dataset, respectively. We have observed that, similar to the numeric and textual datasets, NeuroDAVIS has been able to preserve the distances between objects in high dimensions, better than the other state-of-the-art methods, except for PCA. The median correlation coefficient value reported by NeuroDAVIS has been 0.61, while that for the t-SNE, UMAP, Fit-SNE, PCA and RP embeddings has been 0.53, 0.15, 0.56, 0.78 and 0.28, respectively. Figure 2.9 shows the distribution of correlation coefficient values obtained from multiple executions of this experiment. The corresponding p – values obtained using Mann–Whitney U test are 0.0007 (NeuroDAVIS-t-SNE), 0.0001 (NeuroDAVIS-UMAP, NeuroDAVIS-PCA and NeuroDAVIS-RP) and 0.002 (NeuroDAVIS-Fit-SNE).

The classification accuracy and F1-score values achieved by k-NN and RF classifiers on the NeuroDAVIS embedding for *Coil20* image dataset have been found to be better than those achieved by UMAP, PCA and RP. However, t-SNE and Fit-SNE have produced better classification results on this dataset, as depicted in Figures A.III.1 - A.III.5 (in Appendices). It may be mentioned here that methods like PCA and RP, which try to preserve global shape over local, usually show higher correlation. However, they fail to produce good classification results, which reflects their poor local shape preservation capability.

In order to demonstrate the effectiveness of NeuroDAVIS on datasets having more than 50,000 samples, we have additionally evaluated it on another image dataset, called *FMNIST* [70], containing 60,000 training images of clothing of 28×28 pixels each. Figures 2.10(A), 2.10(C), 2.10(E), and 2.10(G) show NeuroDAVIS, t-SNE, UMAP and Fit-SNE embeddings of *FMNIST* datasets, respectively. To obtain high-quality embeddings from high-dimensional datasets, researchers often use PCA as a preprocessing step [50]. For this reason, we have performed a further investigation applying NeuroDAVIS, t-SNE, UMAP and Fit-SNE on the first 50 principal components obtained from *FMNIST* data, and comparing the results with those produced by the sole usage of NeuroDAVIS, t-SNE, UMAP and Fit-SNE. Figures 2.10(B), 2.10(D), 2.10(F), and 2.10(H) show the effect

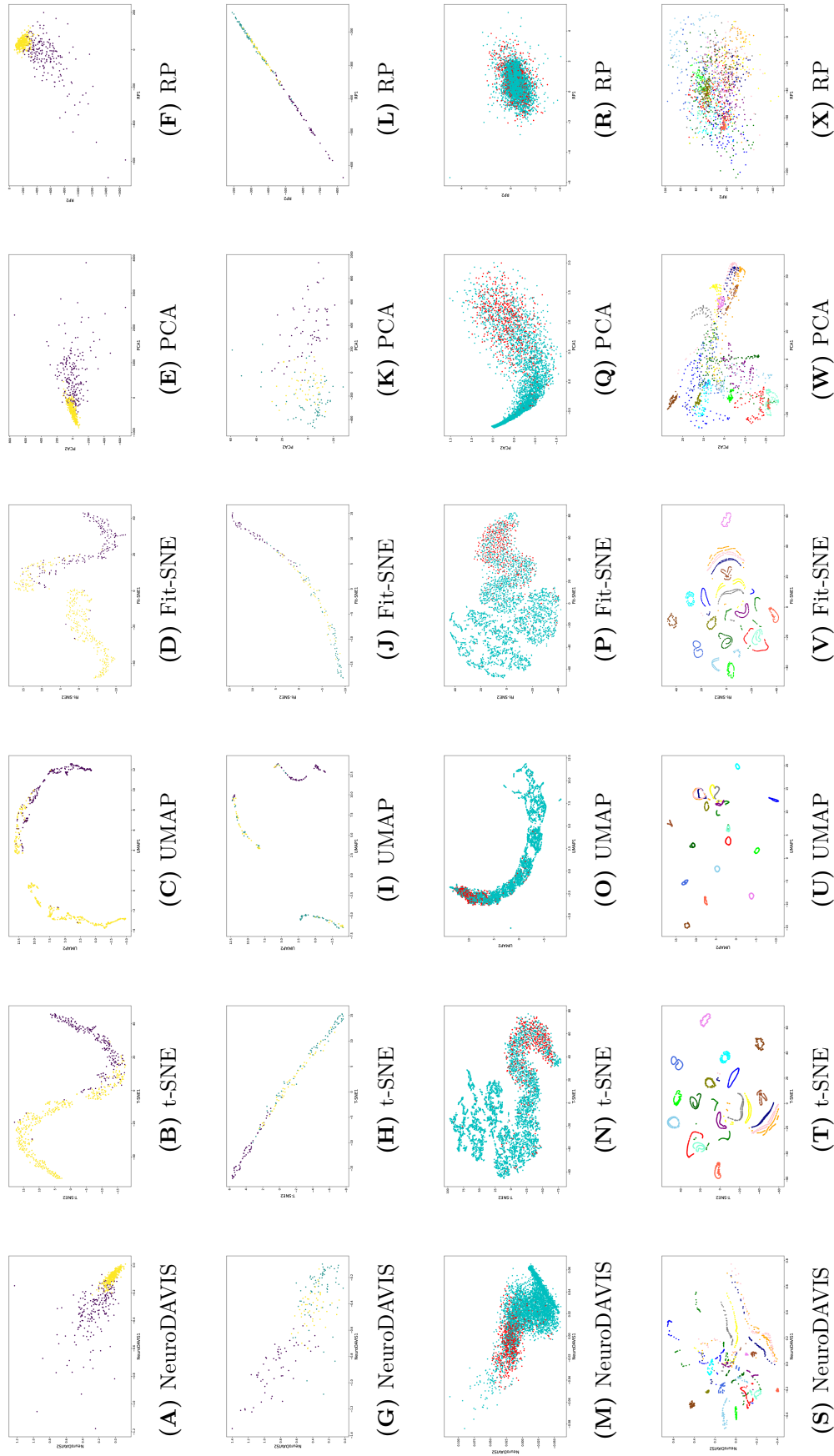


Figure 2.8: Embeddings produced by NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP for *Breast Cancer* ((A)-(F)), *Wine* ((G)-(L)), *Spam* ((M)-(R)) and *Coil20* ((S)-(X)) datasets

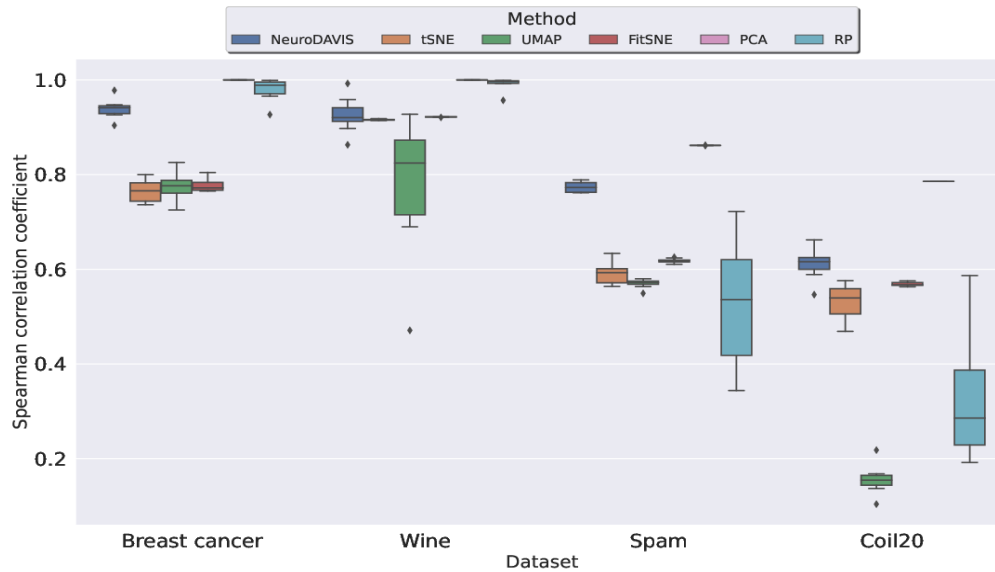
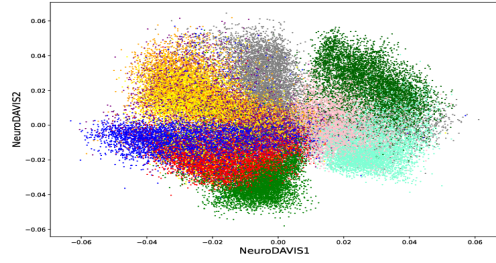


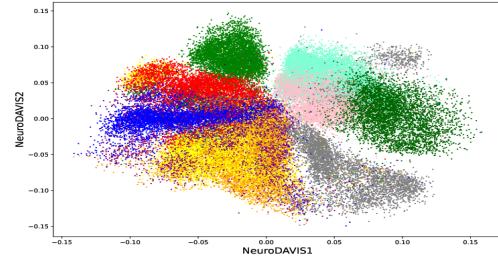
Figure 2.9: Spearman rank correlation coefficient between pairwise distances in original distribution, and pairwise distances in NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP embeddings of *Breast Cancer*, *Wine*, *Spam* and *Coil20* datasets. For *Breast Cancer* dataset, median correlation coefficient values obtained are 0.94 (NeuroDAVIS), 0.76 (t-SNE), 0.77 (UMAP), 0.77 (Fit-SNE), 0.99 (PCA) and 0.98 (RP). For *Wine* dataset, median correlation coefficient values obtained are 0.92 (NeuroDAVIS), 0.91 (t-SNE), 0.82 (UMAP), 0.92 (Fit-SNE), 0.99 (PCA) and 0.99 (RP). For *Spam* dataset, median correlation coefficient values obtained are 0.77 (NeuroDAVIS), 0.59 (t-SNE), 0.57 (UMAP), 0.61 (Fit-SNE), 0.86 (PCA) and 0.54 (RP), while for *Coil20* dataset, these values are 0.61 (NeuroDAVIS), 0.53 (t-SNE), 0.15 (UMAP), 0.56 (Fit-SNE), 0.78 (PCA) and 0.28 (RP). The corresponding p – values obtained using the Mann-Whitney U test are 0.0001 (both NeuroDAVIS-t-SNE and NeuroDAVIS-UMAP) for both *Breast Cancer* and *Spam* datasets. For *Wine* dataset, p – values obtained using Mann-Whitney U test are 0.73 (NeuroDAVIS-t-SNE), 0.002 (NeuroDAVIS-UMAP) and 1.00 (NeuroDAVIS-Fit-SNE), while for *Coil20* dataset, p – values obtained using Mann-Whitney U test are 0.0007 (NeuroDAVIS-t-SNE), 0.0001 (NeuroDAVIS-UMAP) and 0.002 (NeuroDAVIS-Fit-SNE).

of using PCA as a preprocessing step before applying NeuroDAVIS, t-SNE, UMAP or Fit-SNE.

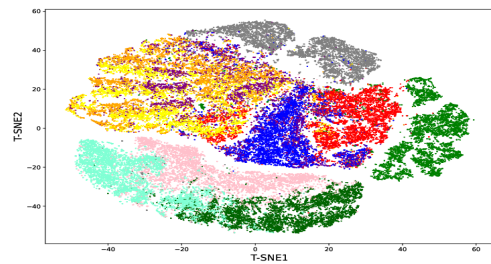
It is often difficult to assess the quality of clusters visually. Hence, in order to quantify the quality of clusters, we have obtained the Spearman rank correlation coefficients between the pairwise distances of the cluster centroids in high dimensions and the pairwise distances of the cluster centroids in the low-dimensional embedding. A high correlation coefficient signifies better preservation of inter-cluster distances. Here, we have observed that NeuroDAVIS has produced a higher correlation coefficient of 0.93, compared to that produced by t-SNE (correlation coefficient = 0.70), UMAP (correlation coefficient = 0.91) and Fit-SNE (correlation coefficient = 0.89), as shown in Figure 2.11. Interestingly, applying PCA for preprocessing has produced a lower correlation coefficient



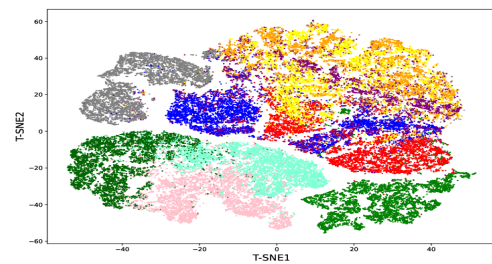
(A) NeuroDAVIS embedding of original data



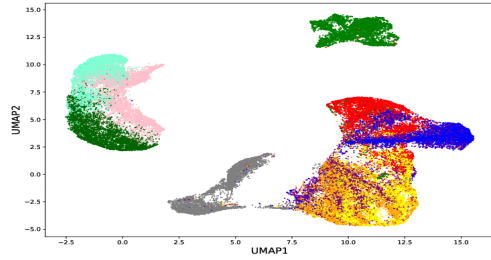
(B) PCA prior to NeuroDAVIS



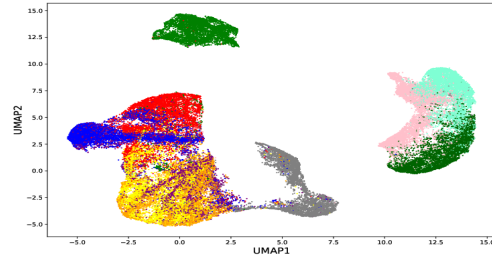
(C) t-SNE embedding of original data



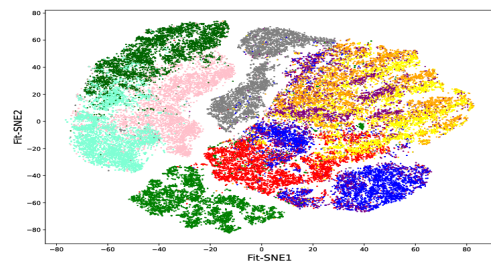
(D) PCA prior to t-SNE



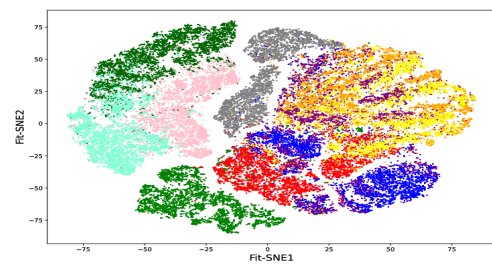
(E) UMAP embedding of original data



(F) PCA prior to UMAP



(G) Fit-SNE embedding of original data



(H) PCA prior to Fit-SNE

Figure 2.10: (A)NeuroDAVIS (C) t-SNE (E) UMAP (G) Fit-SNE embeddings obtained on *FMNIST* dataset, (B) NeuroDAVIS (D) t-SNE (F) UMAP (H) Fit-SNE embeddings obtained on the first 50 PCA components of the original *FMNIST* dataset.

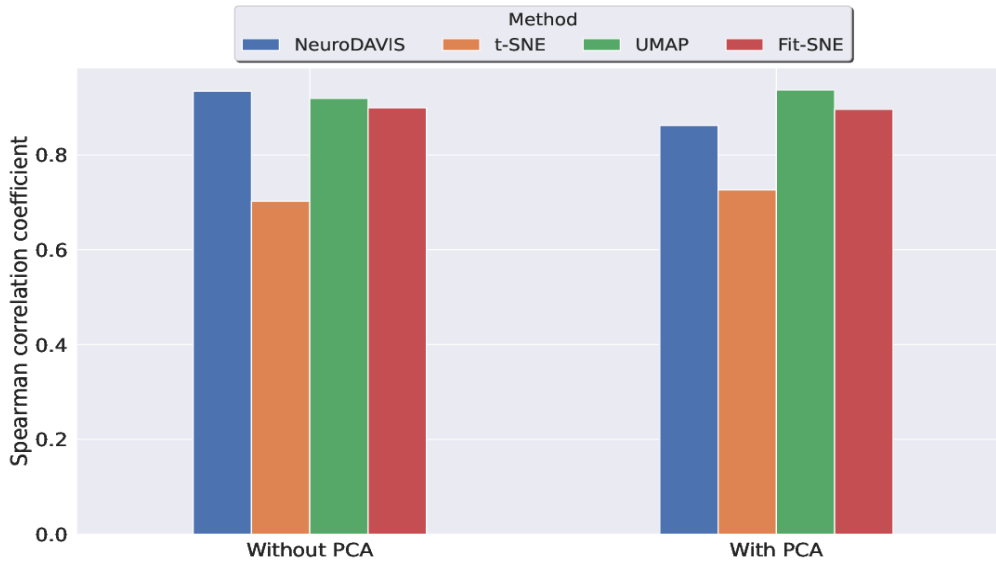


Figure 2.11: Spearman rank correlation coefficients between pairwise distances of the cluster centroids in high-dimension and the pairwise distances of the cluster centroids in the low-dimensional embedding of *FMNIST* dataset produced by NeuroDAVIS, t-SNE, UMAP, Fit-SNE and RP with and without PCA-based preprocessing.

for NeuroDAVIS (correlation coefficient = 0.86). It has improved the result for t-SNE (correlation coefficient = 0.72) and UMAP (correlation coefficient = 0.93), while FitSNE has not been affected by the usage of PCA as a preprocessing step (correlation coefficient = 0.89), as shown in Figure 2.11. These results have led us to infer that t-SNE or UMAP perform better when PCA is used as a preprocessing tool. NeuroDAVIS, on the other hand, is able to produce high-quality embeddings independently.

Since this dataset has 60,000 samples, the largest among all the datasets used in this work, we have tried to explore how NeuroDAVIS performs when a set of new observations is projected onto an existing embedding. For this reason, the dataset has been split into training and test datasets in a 80 : 20 ratio. Initially, we have trained NeuroDAVIS on the training data and then projected the test dataset onto the already-existing projection of the training dataset. As shown in Figure A.IV.1 (in Appendices), we have observed that the test dataset has precisely followed the distribution of the training dataset embedding, which fortifies the claim that NeuroDAVIS can be used for pretraining.

Biological datasets

Finally, we have explored the effectiveness of NeuroDAVIS on some biological datasets. We have used two scRNA-seq datasets for this purpose. *Usoskin* dataset [71] contains

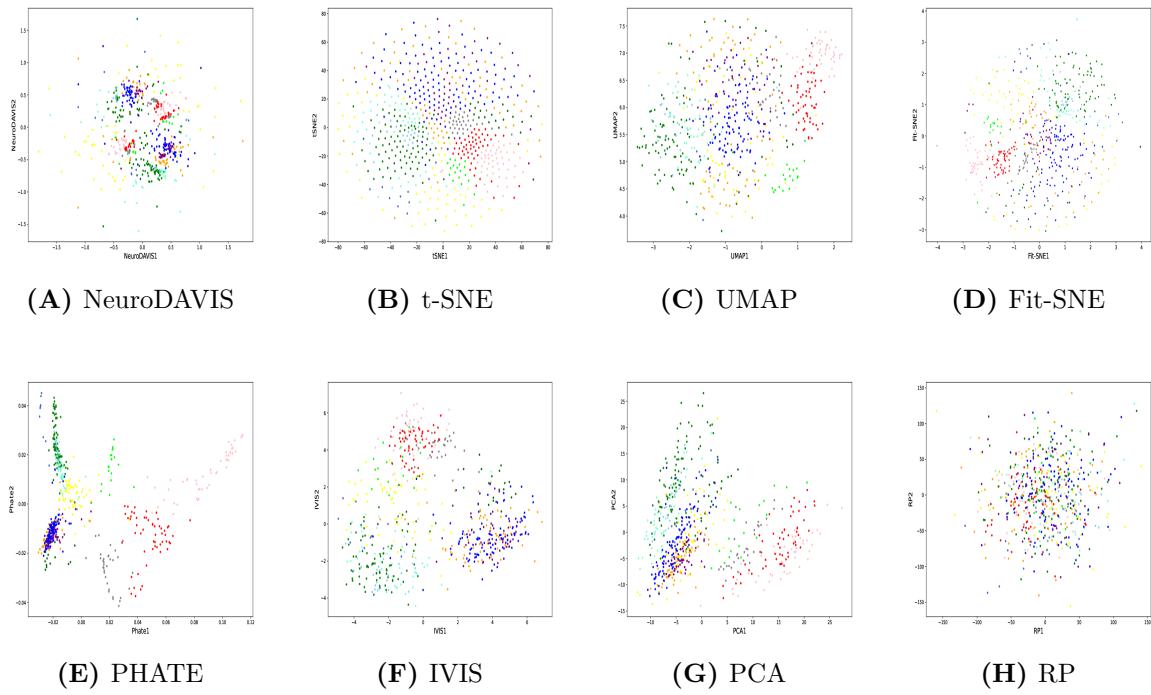


Figure 2.12: 2-dimensional embeddings of *Usoskin* data generated by NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PHATE, IVIS, PCA and RP respectively.

gene expression values for 622 cells across 25334 genes, while *Jurkat* dataset [72] contains gene expression values for 3388 cells across 32738 genes. Both these datasets have been preprocessed using Scanpy [75] following standard procedures as recommended in the tutorial. Apart from t-SNE, UMAP, Fit-SNE, PCA and RP, we have considered PHATE and IVIS for performance comparison on these datasets, which are among the current state-of-the-art algorithms for scRNA-seq datasets besides t-SNE and UMAP. It may be mentioned here that in both these datasets, the number of samples is much lower than the number of dimensions, which escalates an additional challenge in learning.

The two-dimensional embeddings obtained by NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PHATE, IVIS, PCA and RP for *Usoskin* and *Jurkat* datasets, as shown in Figures 2.12 and 2.13 respectively, have revealed that both t-SNE and UMAP have been unable to represent the clusters in the data clearly. For *Usoskin* dataset, NeuroDAVIS has been able to represent some of the clusters well, while for *Jurkat* dataset, clusters in the NeuroDAVIS projection have been much more dense and well-separated, as compared to the other methods. Overall, it can be said that NeuroDAVIS projections are better than t-SNE, UMAP, Fit-SNE and PCA projections, and are comparable to PHATE or IVIS projections only.

The correlation coefficient values measured between the pairwise distances in the original data and NeuroDAVIS-, t-SNE-, UMAP-, Fit-SNE-, PHATE-, IVIS-, PCA- and RP-generated embeddings have been reported in Figure 2.14. We have once again observed that NeuroDAVIS has shown better correlation (median correlation coefficient = 0.41) than all other methods for *Jurkat* dataset, while for *Usoskin* dataset, NeuroDAVIS has reported a median correlation coefficient of 0.23, better than all other methods except t-SNE (median correlation coefficient = 0.27) and Fit-SNE (median correlation coefficient = 0.29). On the other hand, RP-projected embedding has produced extremely poor visualization inspite of the fact that the Spearman correlation is higher than that of UMAP, PHATE and IVIS for *Usoskin* data and comparable with that of all other methods except NeuroDAVIS. The corresponding p – values obtained using the Mann–Whitney U test are 0.0001 (NeuroDAVIS-t-SNE, NeuroDAVIS-UMAP, NeuroDAVIS-Fit-SNE, NeuroDAVIS-PHATE and NeuroDAVIS-IVIS) for *Jurkat* dataset, while for *Usoskin* dataset, p – values obtained using the Mann–Whitney U test are 0.677 (NeuroDAVIS-t-SNE), 0.472 (NeuroDAVIS-Fit-SNE) and 0.0001 (NeuroDAVIS-UMAP, NeuroDAVIS-PHATE, NeuroDAVIS-IVIS and NeuroDAVIS-PCA).

As demonstrated in Figure A.III.6 (in Appendices), we have further observed that the clustering performance of k-means and agglomerative hierarchical clustering methods on NeuroDAVIS embedding of *Usoskin* data has been better than t-SNE, IVIS, PCA and RP embeddings, but poorer than PHATE and Fit-SNE embeddings, while being closely comparable to that of UMAP embedding, in terms of ARI and FMI scores. For *Jurkat* dataset, NeuroDAVIS embedding has, however, resulted in one of the best clustering performances in terms of both ARI and FMI scores, which can be challenged by IVIS and PCA embedding.

An important attribute to compare dimension reduction methods is shape preservation. While every method tries to preserve local and global structures/shapes within the data, they end up preserving one over the other. In order to analyze how NeuroDAVIS compares with the other state-of-the-art methods to preserve local and global shapes, we have compared the pairwise distances among all points in high dimensions with those in their reduced low-dimensional embedding, generated by these methods.

First, we have calculated all possible distances in the original (high-dimensional) space, and distributed them across 50 equal-width bins in sorted order. Similarly, the pairwise

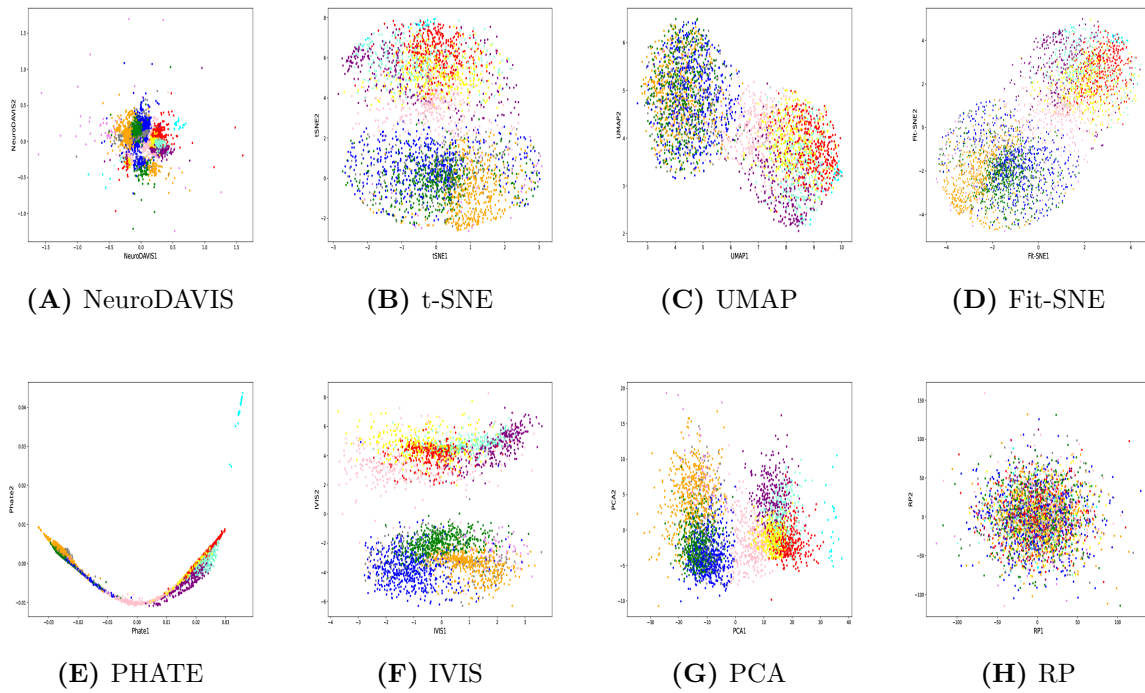


Figure 2.13: 2-dimensional embeddings of *Jurkat* data generated by NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PHATE, IVIS, PCA and RP respectively.

distances in the low-dimensional embedding have been recorded and mapped into their corresponding bins. To understand how good global distances are preserved by each embedding, we have used boxplots for displaying the distribution of the pair-wise distances in low dimension over that in high dimension, throughout the bins, for all the methods, as shown in Figures 2.15(A) and 2.15(B). The boxplots, at the beginning, represent short distances, while those at the end represent long distances. Thus, the lower the spread of the boxplot, the better the preservation of distances. Moreover, the median of the boxplots should follow an increasing order, i.e., the median for trailing boxplots representing long distances should be higher than that for the leading boxplots representing short distances. This signifies that the order of pairwise distances between observations in high dimension is similar to that between observations in low dimension.

It has been observed that for both *Usoskin* and *Jurkat*, NeuroDAVIS has preserved short distances better since the spread in the leading boxplots tends to be zero. On observing the median distances, we have found that for both the datasets, NeuroDAVIS exhibits an increasing pattern, outperforming all the other methods, which show a descent in median distances as we go towards the trailing boxplots. The Pearson correlation coefficient values between all-pair distances in high dimensions and those in low dimensions

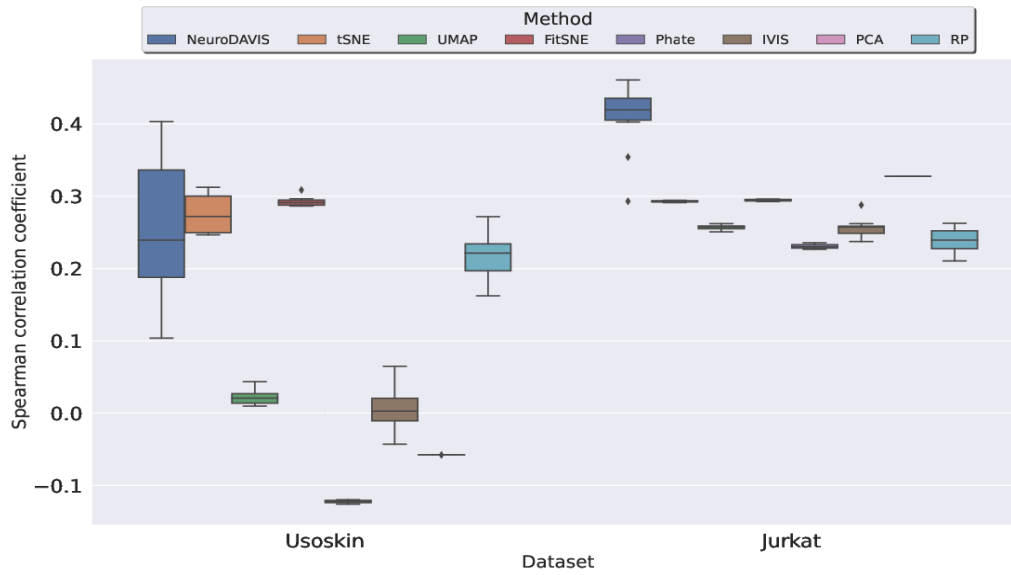


Figure 2.14: Spearman rank correlation between pairwise distances in original distribution and pairwise distances in NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PHATE, IVIS, PCA and RP-produced embeddings of *Usoskin* and *Jurkat* datasets. For *Usoskin* dataset, median correlation coefficient values obtained are 0.24 (NeuroDAVIS), 0.27 (t-SNE), 0.02 (UMAP), 0.29 (Fit-SNE), -0.12 (PHATE), 0.002 (IVIS), -0.05 (PCA) and 0.22 (RP). For *Jurkat* dataset, median correlation coefficient values obtained are 0.42 (NeuroDAVIS), 0.29 (t-SNE), 0.26 (UMAP), 0.29 (Fit-SNE), 0.23 (PHATE), 0.25 (IVIS), 0.32 (PCA) and 0.23 (RP). The corresponding p – values obtained using the Mann-Whitney U test for *Usoskin* dataset are 0.677 (NeuroDAVIS-t-SNE), 0.472 (NeuroDAVIS-Fit-SNE), 0.0001 (NeuroDAVIS-UMAP, NeuroDAVIS-PHATE, NeuroDAVIS-IVIS and NeuroDAVIS-PCA), and 0.0001 (NeuroDAVIS-t-SNE, NeuroDAVIS-UMAP, NeuroDAVIS-Fit-SNE, NeuroDAVIS-PHATE, NeuroDAVIS-IVIS, NeuroDAVIS-PCA and NeuroDAVIS-RP) for *Jurkat* dataset.

have been found to be 0.46 for both the datasets, which is higher than that obtained using the other methods. The last subfigure within Figures 2.15(A) and 2.15(B) each represents the support values, i.e., the number of pairwise distances within each bin.

Additionally, we have compared the NeuroDAVIS-generated embedding with that produced by the other methods with respect to their trustworthiness [76], a measure to quantify local neighbourhood preservation post-dimension reduction. For five different values of neighbours, the trustworthiness score has been calculated for both *Jurkat* and *Usoskin* datasets. The trustworthiness values, as shown in Figures 2.15(C) and 2.15(D), reflect that local neighbours are preserved better in the NeuroDAVIS embedding as compared to the embeddings produced by the other methods.

In order to test the robustness of the methods, we have carried out a subsample-based experiment on *Usoskin* and *Jurkat* datasets. For each of these datasets, we have initially drawn a small subsample (100 samples) randomly and applied all these methods

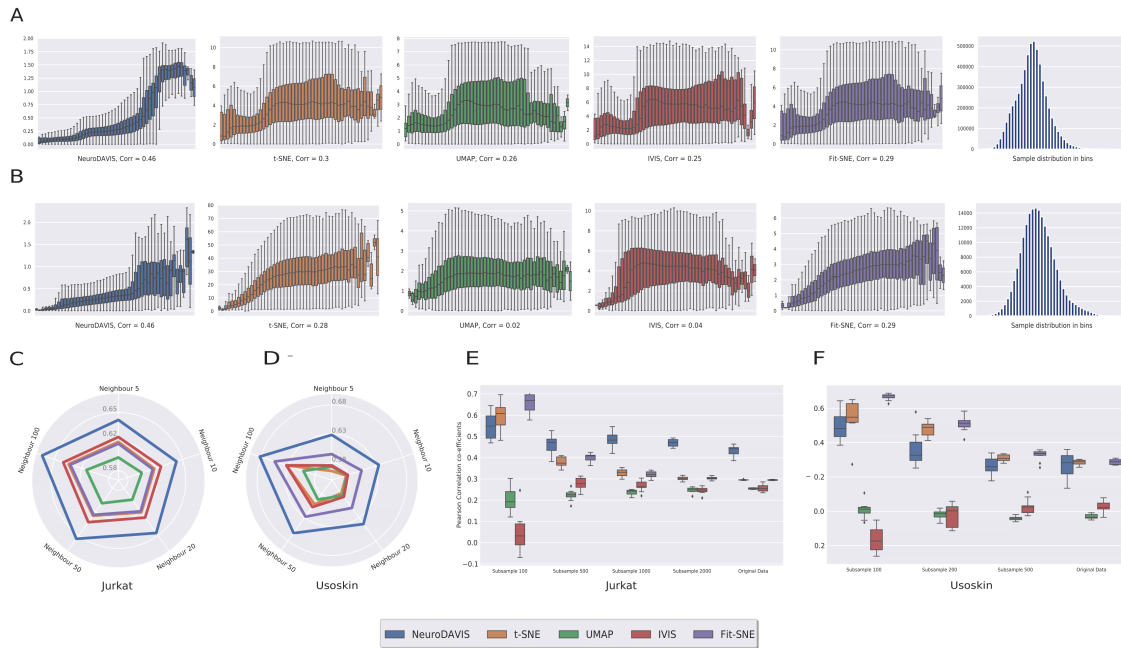


Figure 2.15: (A) and (B) show the results of performance comparison of NeuroDAVIS with that of t-SNE, UMAP, IVIS and Fit-SNE in terms of distance preservation capability and correlation between pairwise distances in high dimension to that in low dimension of *Jurkat* and *Usoskin* datasets, respectively. Distances in high dimension have been measured and distributed across 50 equal-width bins in sorted order. The corresponding distances in low dimensions have also been mapped into the same bins. The leading boxplots represent short distances, while the trailing ones represent long distances. The lower the spread of the boxplot, the better the preservation of distances. The median of the trailing boxplots representing long distances should be higher than that of the leading boxplots representing short distances. It signifies that the order of pairwise distances among observations in high dimension is similar to that among observations in low dimension. The last histogram for each dataset represents the count of distances in each bin. (C) and (D) show results of comparison of local shape preservation capability of NeuroDAVIS with that of t-SNE, UMAP, IVIS and Fit-SNE in terms of trustworthiness score on *Jurkat* and *Usoskin* datasets, respectively. (E) and (F) show results of comparison of robustness of NeuroDAVIS with that of t-SNE, UMAP, IVIS, and Fit-SNE in terms of subsampling-based correlation analysis on *Jurkat* and *Usoskin* datasets, respectively.

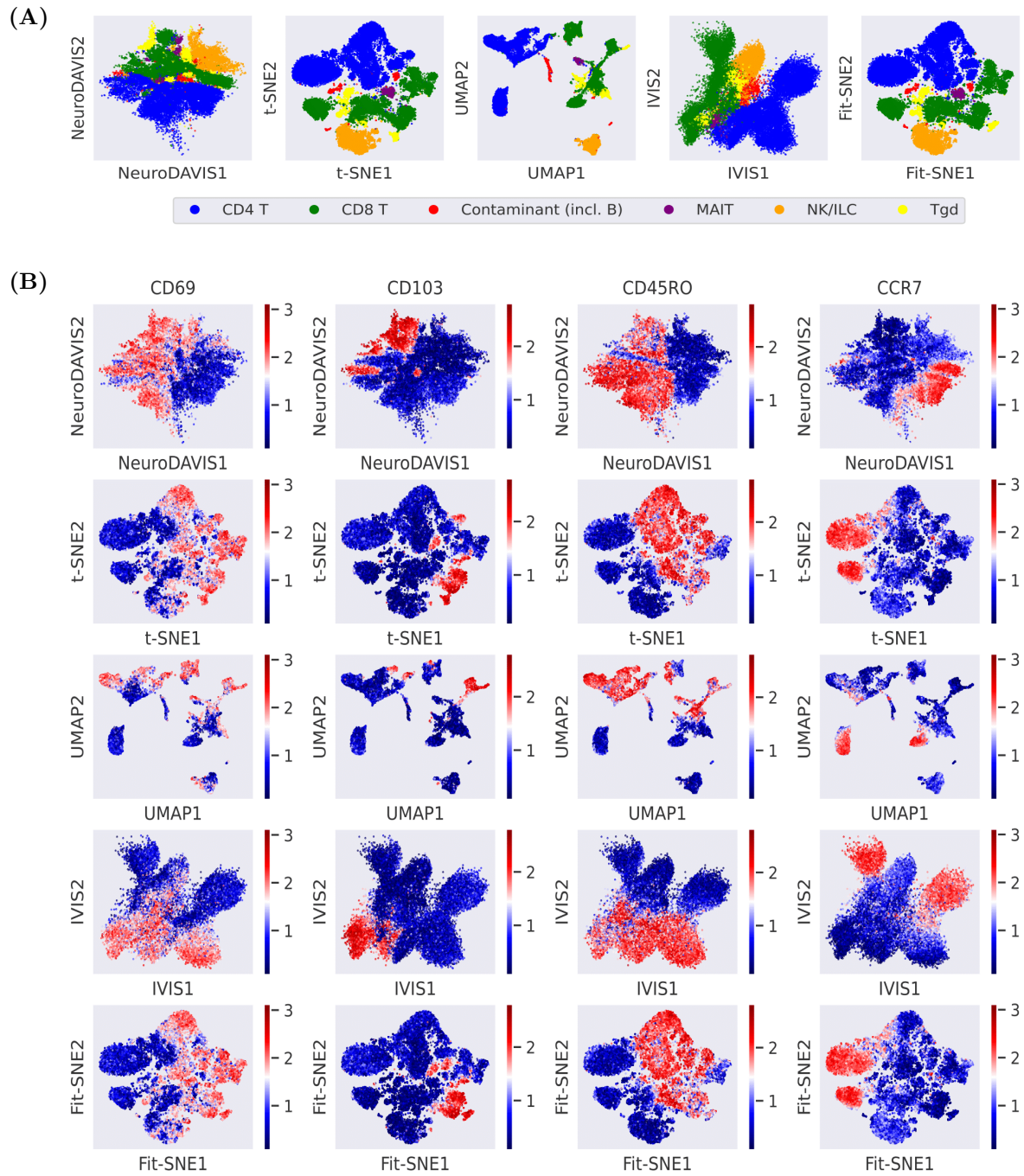


Figure 2.16: (A) 2-dimensional embeddings produced by NeuroDAVIS, t-SNE, UMAP, IVIS and Fit-SNE for the Mass Cytometry *Wong* data, coloured by cell-type; *MAIT* denotes mucosal-associated invariant *T* cells; *ILC* stands for innate lymphoid cells; and Tgd represents $\gamma\delta T$ cells. (B) Expression of *T* cell marker proteins *CD69*, *CD103*, *CD45RO*, *CCR7* on the NeuroDAVIS, t-SNE, UMAP, IVIS and Fit-SNE-generated embeddings.

for dimension reduction. We have recorded the correlation coefficient values (Pearson) between pairwise distances in the original distribution and those in the low-dimensional embedding produced by each of these methods. This experiment has then been repeated multiple times with increasing subsample sizes on both datasets. Interestingly, we have observed that the performance of NeuroDAVIS has been stable and consistent as and when we increase the subsample size (Figures 2.15(E) and 2.15(F)). However, though t-SNE and Fit-SNE have performed well on small subsample sizes, they have failed to scale for large subsample sizes. The performance of UMAP and IVIS, on the other hand, has not been good enough even for small subsample sizes. NeuroDAVIS has outperformed all the other methods for both these datasets when the subsample size reaches the maximum number of samples in the datasets. Thus, NeuroDAVIS stands out as the most robust method among all the visualization methods used in the experiment.

To further demonstrate the effectiveness of NeuroDAVIS, we have considered a Mass Cytometry data (*Wong* data), having more than 300K samples with 39 protein targets. The labels for each cell generated using Louvain clustering-based phenograph and manual annotation have been collected from [39]. Due to computational limitations, in this study, we have used a random sample of size 30k from this data for visualization using NeuroDAVIS. A similar embedding has also been produced by the other state-of-the-art methods. Comparable visualization traits have been found in the embeddings generated by all the methods (Figure 2.16(A)), where the *Tgd* ($\gamma\delta T$) cell cluster and the *Contaminant* (including *B*) cell clusters have been split into distinct clusters. However, unlike t-SNE, UMAP and Fit-SNE, NeuroDAVIS has been able to keep the *CD8T* cell cluster intact, which is only comparable to IVIS. A similar observation can be made for the *CD4T* cell cluster. Thus, NeuroDAVIS undoubtedly preserves local neighbours better than the other methods, which has also been established mathematically earlier in this chapter.

Subsequently, the quality of embeddings has been evaluated by projecting the expression profiles of a few chosen proteins onto the embeddings. As the major cell-type present in the data is *T* cell, we have tried to visualize the expression levels of different *T* cell markers, viz., *CD69*, *CD103* (a resident memory *T* cell marker), *CD45RO* (a memory *T* cell marker) and *CCR7* (naive *T* cell marker) [39]. As illustrated in Figure 2.16(B), the expression levels for all these marker proteins have revealed that NeuroDAVIS has been able to accurately reproduce the *T* cell differentiation stage (blue indicating low

expression and red indicating high expression) inside each main cluster. While all the embeddings have represented an analogous continua of clusters, they do not appear to be organized along a single, recognizable axis like that in NeuroDAVIS.

2.5 Discussion and Conclusion

In this work, we have developed a novel unsupervised deep learning model, called NeuroDAVIS, for data visualization. NeuroDAVIS can produce a low-dimensional embedding from the data, extracting relevant features useful for subsequent analyses. The effectiveness of NeuroDAVIS has been demonstrated on a wide variety of datasets, including 2D synthetic data, and high-dimensional numeric, textual, image and biological data. It has shown excellent visualization capability supported by outstanding downstream analysis (clustering/classification) capabilities, thus making it competitive against state-of-the-art methods, like t-SNE, UMAP, Fit-SNE, PCA and RP. The results obtained are also statistically significant in favour of NeuroDAVIS. The strength of NeuroDAVIS lies in the fact that it is able to preserve cluster shape, size and structure within the data well. As proved mathematically, the low-dimensional embedding produced by NeuroDAVIS has also been observed to preserve local relationships among observations in high dimensions. Furthermore, it is capable of extracting the inherent non-linearity within the data. In addition, NeuroDAVIS does not presume data distributions, which makes the method more practical. The model can be termed as a general-purpose dimension reduction and visualization system having no restrictions on embedding dimension. It is capable of producing interpretable visualizations independently without any initialization. Furthermore, NeuroDAVIS, being a neural network-based model, is parametric and thus supports pretraining.

NeuroDAVIS, being a dimension reduction and visualization method, does not come without its weaknesses. It takes as input an identity matrix of order equal to the number of observations in the input data. Thus, it consumes more space in memory than other existing methods. Moreover, hyper-parameters in NeuroDAVIS are sensitive to tiny modifications, and network tuning requires a substantial amount of time and effort by an amateur.

Nevertheless, NeuroDAVIS serves as a competing method for visualization of high-

dimensional datasets, producing a robust latent embedding better than some of the existing benchmarked methods, like t-SNE, UMAP and Fit-SNE, while also resolving the local-global shape preservation challenge. Thus, in this work, we have introduced a single solution to the long-standing problem of visualization of datasets belonging to different categories/domains/modalities. Adding a feature selection module to the existing architecture could be a future extension to NeuroDAVIS. It might also be extended to visualize high-dimensional multi-modal datasets by producing low-dimensional embeddings that can capture significant features from the data.

Visualization of high-dimensional datasets enables the discovery of latent structures by projecting the data into a lower-dimensional feature space. While such extracted features are effective for capturing complex patterns, they are often difficult to interpret in many application domains, limiting their practical utility. In scenarios where interpretability is critical, feature selection provides a viable alternative by identifying a subset of original, domain-relevant features. Motivated by this necessity, Chapter 3 presents a feature selection framework based on NeuroDAVIS model, aiming to balance pattern discovery with interpretability.

Chapter 3

NeuroDAVIS-FS: Feature selection through visualization using NeuroDAVIS

3.1 Introduction

Availability of high-dimensional datasets poses significant challenges like computational complexity, the curse of dimensionality, and model overfitting. In order to deal with such high-dimensional datasets, dimensionality reduction is widely used for data pre-processing, which in turn is effective for further data analysis [77]. Existing literature often follows two different types of dimension reduction approaches, viz., feature extraction and feature selection [9]. Feature extraction techniques are used to generate new sets of features that preserve relevant information of the original dataset [39]. Chapter 2 has presented a data visualization method, which is basically a feature extraction technique. These feature extraction techniques cannot recognize the proper weightage of the actual features, which sometimes becomes less relevant to some specific tasks [10]. At the same time, feature selection techniques can address the issues by ranking the features based on their importance, and accordingly reducing the dimension of the original datasets with respect to this ranking [10]. Existing state-of-the-art feature selection methods have some major drawbacks, e.g., huge space and time complexity, usage of the target class, and ignoring the inherent association among the features while ranking them, among others [22, 23, 21, 78, 24]. In order to address these limitations (Table 3.1), this chapter aims

to propose a novel and effective approach, called NeuroDAVIS-FS (NDFS), which is an extension of NeuroDAVIS [1], discussed in Chapter 2.

NDFS performs in an unsupervised learning setup without assuming any prior data distribution. Initially, it considers the training of the prior model NeuroDAVIS and then selects features utilizing the trained model. The performance has been demonstrated on three datasets, viz., *Breast Cancer*, *Wine*, and *Digits*, and compared against the state-of-the-art methods, viz., Variance Threshold (VT), GenericUnivariateSelect (GUS), SelectKBest (SKB), Mutual Information Classifier (MIC), Recursive Feature Elimination (RFE), and Recursive Feature Elimination with cross-validation (RFECV). Both the extracted and selected features have been compared with the state-of-the-art in terms of classification and clustering performances. In this context, Accuracy, Precision, Recall, and F1-score are calculated for classification performances; on the other hand, Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Fowlkes-Mallows index (FMI) have been reported to evaluate the clustering performance, respectively. Through this extensive study, it has been observed that NDFS is robust enough to capture the hidden patterns present in datasets from various fields. In addition, two case studies on image and biological datasets with a very low sample-feature ratio (nearly 1 : 10) have been executed.

The outline of this chapter is described as follows. Section 3.2 shows a detailed explanation of the method developed, along with a problem scenario. Section 3.3 executes the outcome and analyses of the results. Section 3.4 discusses the strengths and weaknesses of the model developed, and concludes the chapter.

3.2 Problem statement and solution approach

This section deals with the problem scenario along with the developed approach.

3.2.1 Problem scenario

Dimension reduction plays a crucial role in data science as the number of features increases rapidly. Feature extraction methods can reduce the dimension of any dataset by producing a linear or non-linear combination. However, the extracted features are not at all interpretable. In this regard, feature selection methods are one step ahead of the

Table 3.1: Comparison of existing state-of-the-art and NeuroDAVIS-FS (NDFS)

Method	Technique	Advantages	Drawbacks
VT [21]	Unsupervised	- Removes features with variance lower than certain threshold - Simple and efficient - Removes noise	- Ignores the inherent association among the features - Threshold parameter is sensitive - Does not consider Feature importance
GUS [78]	Supervised	- Considers univariate statistical tests. - Simple and efficient - Selects target specific features.	- Limited to Univariate statistical tests - Often overfits - Ignores the inherent association among the features.
SKB [23]	Supervised	- Simple and efficient - Handles both categorical and numerical data - Removes noise	- Ignores the inherent association among the features - Sensitive to threshold parameter - Ignores feature importance
MIC [22]	Supervised	- Captures non-linear hidden patterns - Handles both categorical and numerical data - Robust in nature	- Computationally expensive - Sensitive to noise - Ignores feature-feature associations
RFE [23]	Supervised	- Considers association among features - Automatically ranks features based on their importance - Reduces overfitting	- Huge computational cost - Sensitive to model - Requires careful cross-validation
RFECV [24]	Supervised	- Considers association among features - Automatically ranks features based on their importance - Cross-Validation ensures generalized feature selection	- Huge computational cost - Depends on the initial model - Iterative process and cross-validation makes it even more complex
NeuroDAVIS-FS (NDFS)	Unsupervised	- Considers association among features - Do not assume any inherent data distribution - Effective for datasets with a very low sample-feature ratio	- Depends on the initial model - Cannot handle categorical datasets directly

feature extraction method and select relevant features by utilizing the domain knowledge. Motivated by this scenario, a feature selection module, NDFS, has been developed in this chapter, which is an extension of NeuroDAVIS, a data visualization approach.

Algorithm 1 NeuroDAVIS-FS

Input: Data matrix $\mathbf{X} = [x_{ik}]_{n \times d}$, where $x_{ik} \in \mathbb{R}$

Output: Top τ features of $\mathbf{X}$, where $\tau \leq d$

```

1:  $Score \leftarrow []$ 
2: Drop columns with standard deviation  $< 0.0001$ 
3:  $\mathbf{X}_{scaled} \leftarrow \text{MinMaxScaling}(\mathbf{X})$ 
4:  $\tilde{\mathbf{X}} \leftarrow \text{NeuroDAVIS}(\mathbf{X}_{scaled})$ 
5: for  $k = 1$  to  $d$  do
6:   Compute  $L_{KLD}(x_{:k} \parallel \tilde{x}_{:k})$  using Eq. (3.1)
7:   Compute  $Score_k$  using Eq. (3.2)
8: end for
9:  $F \leftarrow \arg \min_{\tau} Score$ 
10: return  $F$ 

```

Let a data matrix $\mathbf{X}$ with n samples and d features be considered as input. The chapter aims to select the top τ features based on the prior trained NeuroDAVIS model, where $\tau \leq d$. The framework of the developed model has been described in Algorithm 1.

3.2.2 Solution approach

This section discusses the description of the datasets that have been used to demonstrate the effectiveness of the developed model, along with the detailed methodology.

Data description

In this chapter, various datasets from different domains have been used to demonstrate the effectiveness of the developed feature selection module. The detailed descriptions are provided in Table 3.2. Among the datasets, *Breast Cancer*, *Wine*, and *Digits* are utilized to evaluate the performance of NeuroDAVIS-FS in comparison with the existing state-of-the-art methods. However, the remaining datasets are used to examine the performance of NDFS when the sample-feature ratio is very low (nearly 1 : 10).

Table 3.2: Descriptions of the datasets utilized for the model evaluation

Name	Description	Samples	Features	Classes	Category & Source
<i>Breast Cancer</i>	Picture of a digitalized FNA of a breast mass is used to compute features.	569	30	2	Numeric [66]
<i>Wine</i>	A chemical examination of three distinct growers' wines, cultivated in Italy.	178	13	3	Numeric [67]
<i>Digits</i>	Grayscale images of handwritten digits	1797	64	10	Image [1]
<i>Coil20</i>	Grayscale images of objects, captured through different angles	1440	16384	20	Image [69]
<i>Olivetti Faces</i>	Grayscale images of human faces	400	4096	40	Image [74]
<i>Jurkat</i>	Sc-RNAseq data using the peripheral blood of a 14-year-old kid with acute lymphoblastic leukemia	3388	32738	10	Biological [72]

Methodology

In this section, an extended version of NeuroDAVIS, viz., NDFS, has been developed to select important features for data modeling and analysis [1]. Initially, NDFS considers the earlier data visualization module, NeuroDAVIS, for model training. In this context, NeuroDAVIS tries to reconstruct the data using random lower-dimensional points. During the process of reconstruction, it learns a suitable visualization of the dataset.

Let $\mathbf{X} = [x_{ik}]_{n \times d}$, where $x_{ik} \in \mathbb{R}$, denotes a data matrix having n samples and characterized by d features. Here, $\mathbf{x}_i$ and $\mathbf{x}_k$ denote the i^{th} sample vector (i.e., i^{th} row of $\mathbf{X}$) and k^{th} feature vector (i.e., k^{th} column of $\mathbf{X}$) respectively. For a better readability, $\mathbf{x}_i$ has been written as $\mathbf{x}_i$ through the entire chapter. Initially, NeuroDAVIS takes an identity matrix of order n as input and creates random lower-dimensional points at the *Latent layer*. These lower-dimensional points are then projected to the original high-dimensional space at the *Reconstruction layer* through a few *Hidden layers*. The reconstructed data received from *Reconstruction layer* has again been considered for calculating a loss, and accordingly, the weights and biases have been updated using the Adam optimizer [1]. After successful training, NeuroDAVIS learns to recreate the original dataset, and accordingly, 2 or 3 completely new features have been extracted at the *Latent layer*, which are useful

for visualization. It may be inferred that a feature with a lower variance contains less information, and accordingly, it can easily be reconstructed by NeuroDAVIS.

Motivated by this aspect, a novel methodology, NeuroDAVIS-FS, has been developed in this module. In order to identify important features, we have examined how well the reconstructions are compared to the original dataset and how much variation is captured by that feature. Important features that have been identified by NeuroDAVIS, have a high variance with a better reconstruction. After successful training, the reconstruction of the original features is represented by $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_{:k}]_{1 \times d}$. In order to validate the reconstruction capability, Kullback–Leibler Divergence (KLD) [79] has been used, which is defined as

$$L_{KLD}(\mathbf{x}_{:k} || \tilde{\mathbf{x}}_{:k}) = \mathbf{x}_{:k} \log \left(\frac{\mathbf{x}_{:k}}{\tilde{\mathbf{x}}_{:k}} \right) \quad (3.1)$$

For comparable results, initially, a preprocessing task, i.e., Min-max scaling, has been performed on each of the datasets and then fed to NeuroDAVIS. If NeuroDAVIS is able to reconstruct $\mathbf{x}_{:k}$ properly then $\tilde{\mathbf{x}}_{:k} \sim \mathbf{x}_{:k}$, and it indicates $L_{KLD}(\mathbf{x}_{:k} || \tilde{\mathbf{x}}_{:k}) \rightarrow 0$. Let $\sigma_{\mathbf{x}_{:k}}$ be the standard deviation of the k^{th} feature $\mathbf{x}_{:k}$. Therefore, we are looking for features that have a large standard deviation along with better reconstruction, i.e., smaller L_{KLD} loss along with a higher standard deviation. Incorporating these two aspects, the score is defined as

$$Score_j = \frac{L_{KLD}(\mathbf{x}_{:k} || \tilde{\mathbf{x}}_{:k})}{\sigma_{\mathbf{x}_{:k}}}; \text{ where } \sigma_{\mathbf{x}_{:k}} \neq 0 \quad (3.2)$$

Once the scores are obtained, the top features are selected according to their minimum value. These top τ features are then utilized for several downstream analyses, e.g., classification and clustering. In this context, k-Nearest Neighbour (k-NN) and Random Forest (RF) classifiers are utilized for classification, and the performance is quantified using Accuracy, Precision, Recall, and F1-Score. At the same time, K-Means and Agglomerative clustering are utilized, and the performance is compared using ARI, NMI and FMI scores.

3.3 Results

This section describes the comparative results with respect to the state-of-the-art methods, followed by a few case studies on datasets from different domains.

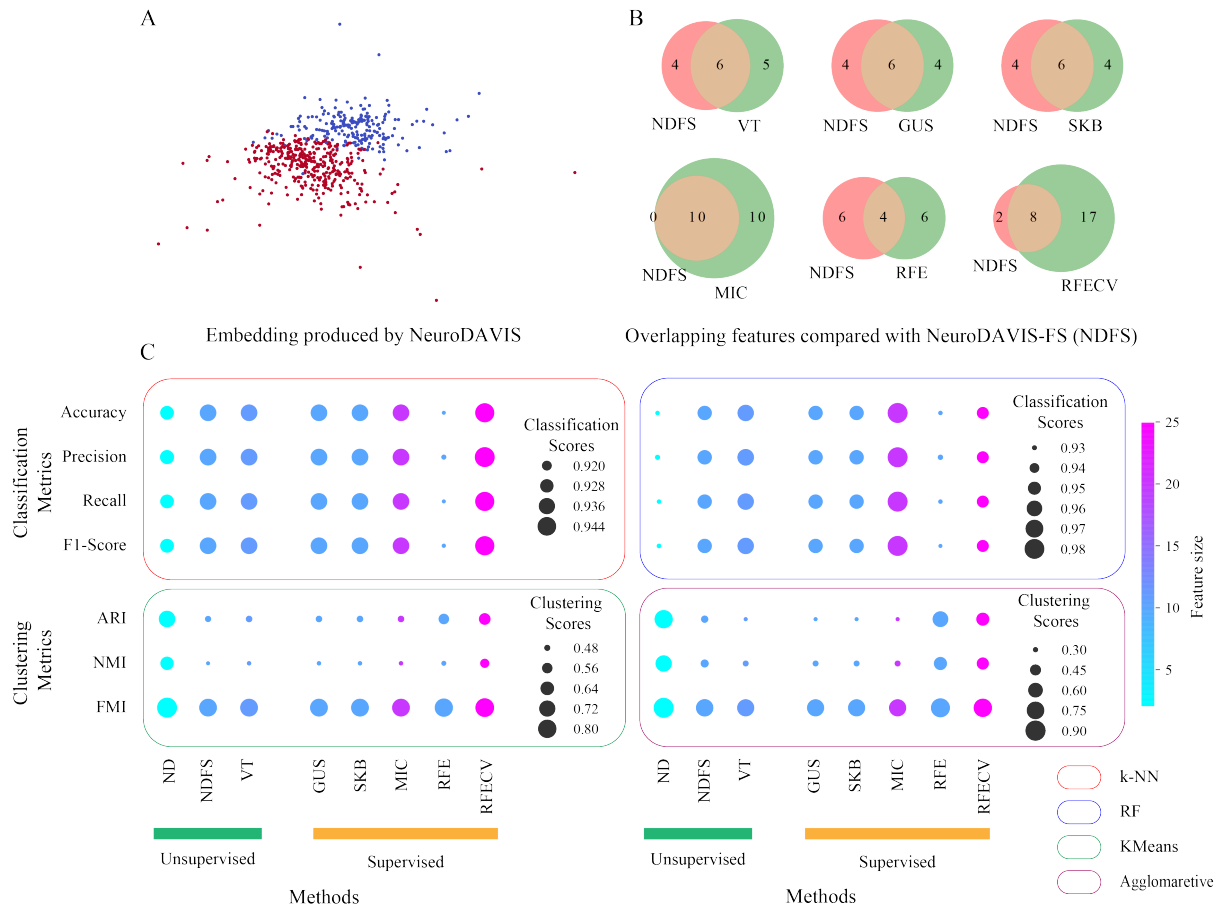


Figure 3.1: For *Breast Cancer* dataset (A) The 2-dimensional embedding produced by NeuroDAVIS (ND). (B) Overlap of features between NeuroDAVIS-FS (NDFS) and the state-of-the-art. (C) Comparison of the developed models with state-of-the-art methods based on classification and clustering performance. For classification, K-nearest neighbor (k-NN) and Random Forest (RF) models are utilized and for evaluation of clustering performance, KMeans and Agglomerative clustering techniques are used.

3.3.1 Comparison with the state-of-the-art method

In this subsection, the performance of NDFS has been demonstrated over three datasets, viz., *Breast Cancer*, *Wine*, and *Digits*.

Results on *Breast Cancer* data

Figure 3.1 describes the results related to *Breast Cancer* dataset. Figure 3.1(A) shows the embedding produced by NeuroDAVIS, which reflects those two extracted features separate the classes properly. Accordingly, the top features have been selected using NeuroDAVIS-FS, with the help of those two extracted features. After getting the efficient features, it has been compared against the state-of-the-art methods to study overlapping features (Figure 3.1(B)). 40% overlap with RFE; 60% overlap with VT, GUS, and SKB; 80% overlap

with RFECV; and a 100% overlap with MIC have been observed. In order to validate the efficiency of the features selected by NeuroDAVIS-FS, classification and clustering have been performed over the selected features. The results have again been compared against the state-of-the-art methods. It is evident from Figures 3.1(C) that both the extracted and selected features can classify the samples properly with an accuracy of 0.93. VT, GUS, and SKB shows similar classification performance. MIC and RFECV perform better in terms of accuracy, precision, recall, and F1-score. However, the dimension of the selected feature space is significantly higher compared to the other methods. Figures 3.1(C) also shows the clustering performance of the developed methods in terms of ARI, NMI, and FMI scores against the state-of-the-art methods. The extracted features have outperformed all the other methods, and the selected features have produced comparable results in terms of clustering performance. It must be mentioned that only VT and the developed approach select features in an unsupervised setup, whereas the rest use a target class vector to do the same.

Results on *Wine* data

The results on the *Wine* dataset have been demonstrated using figure 3.2. The embedding produced by NeuroDAVIS has shown effective visualization performance in Figure 3.2(A). Three different types of wines are well separated and clearly visible in the embedding. Using this embedding, NeuroDAVIS-FS selects top features. Features have also been selected using the other state-of-the-art methods, and a good amount of overlapping has been found between NDFS and the state-of-the-art methods. 30% overlap with MIC; more than 80% overlap with VT, GUS, and SKB; and a complete overlap with RFE and RFECV has been observed. For *Wine* dataset, ‘Alcohol’ came out as the top feature while NDFS was used. This shows the reliability of NDFS as a feature selection method. Apart from the overlapping feature, classification, and clustering have been performed over *Wine* dataset using two classifiers, viz., k-NN, and RF, and two clustering methods, viz., kMeans and Agglomerative clusterings, respectively, against the state-of-the-art approaches. It has been found from Figure 3.2(C), that the extracted features have outperformed all the methods in terms of classification and clustering. NDFS selected top features show a comparable result with the state-of-the-art methods by selecting comparatively fewer features and also by ignoring the class information.

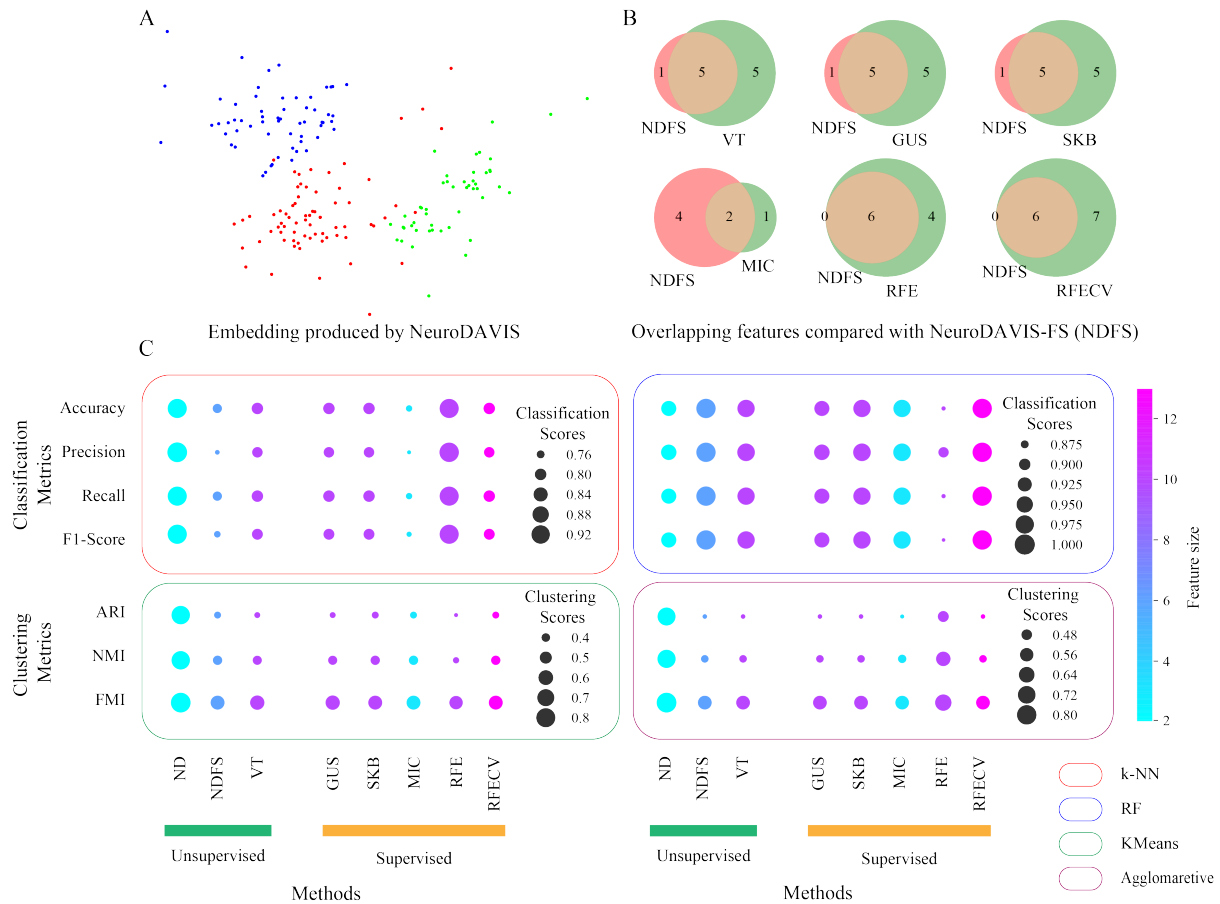


Figure 3.2: For *Wine* dataset (A) The 2-dimensional embedding produced by NeuroDAVIS (ND). (B) Overlap of features between NeuroDAVIS-FS (NDFS) and the state-of-the-art methods. (C) Comparison of the developed models with the state-of-the-art based on classification and clustering performance. For classification, K-nearest neighbor (k-NN) and Random Forest (RF) models are utilized and for evaluation of clustering performance, KMeans and Agglomerative clustering techniques are used.

Results on *Digits* data

Finally, the effectiveness of NDFS has been demonstrated on an image dataset, viz., *Digits* dataset. NeuroDAVIS projected embedding has been shown in Figure 3.3(A). All ten clusters are not very separated, whereas a few compact clusters are clearly visible. Only two or three clusters overlap with each other (Figure 3.3(A)). Moreover, NDFS has been applied using the embedding produced by NeuroDAVIS. It has been observed from Figure 3.3(B) that there exists a good amount of overlap between the features selected by NDFS and the state-of-the-art methods. 50% features are common with GUS, SKB, and RFE; nearly 80% features are common with MIC; almost a complete overlap with VT and RFECV has been observed. NDFS has again been compared with the state-of-the-art in terms of classification and clustering. Figure 3.3(C) shows that the extracted features are

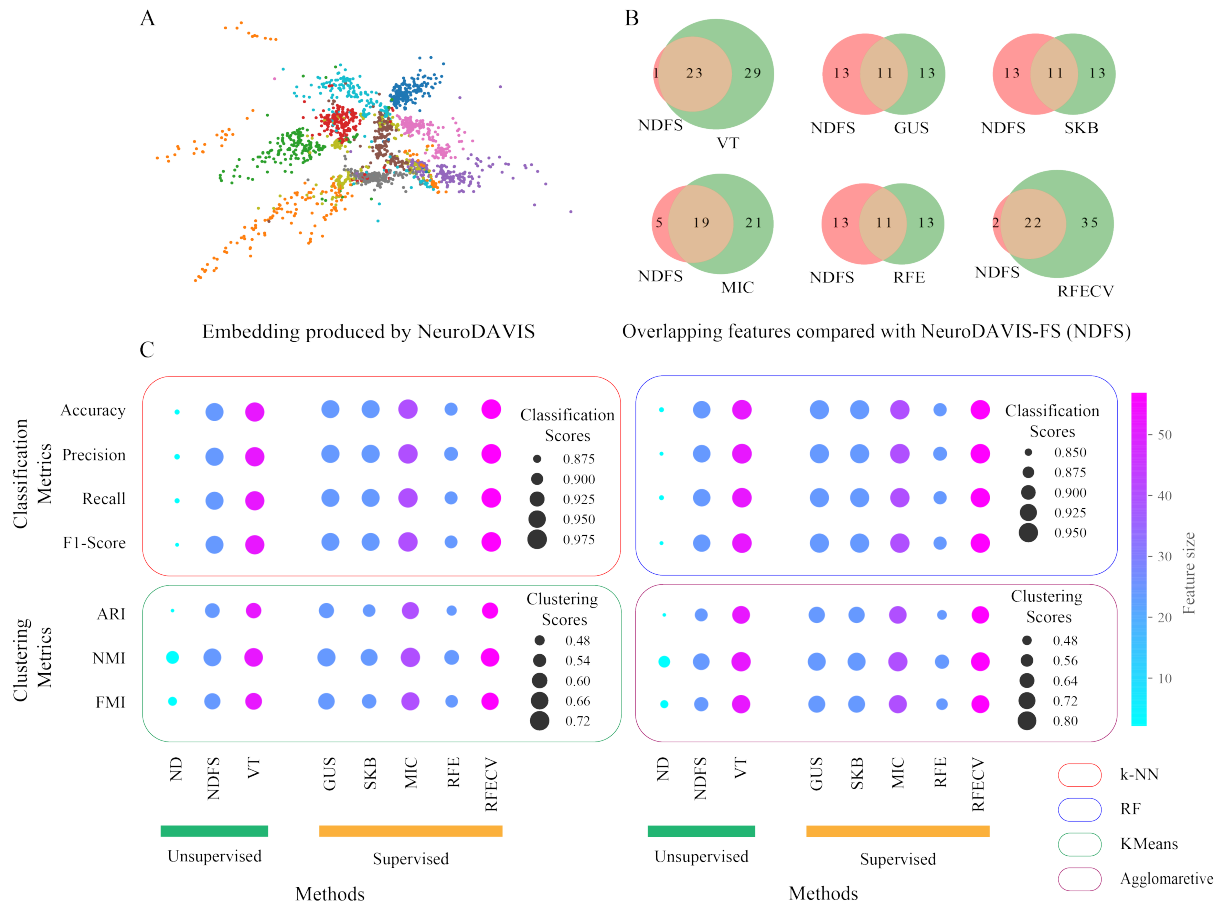


Figure 3.3: For *Digits* dataset (A) The 2-dimensional embedding produced by NeuroDAVIS (ND). (B) Overlap of features between NeuroDAVIS-FS (NDFS) and the state-of-the-art methods. (C) Comparison of the developed models with the state-of-the-art based on classification and clustering performance. For classification, K-nearest neighbor (k-NN) and Random Forest (RF) models are utilized and for evaluation of clustering performance, KMeans and Agglomerative clustering techniques are used.

not that efficient in terms of visualization, classification, and clustering. On the other hand, the selected features are performing well in terms of classification and clustering. NDFS selects only the top 35% features and shows comparable classification and clustering performance; however, methods like MIC, VT, and RFECV select a lot more features to achieve the same.

3.3.2 Case studies

In this section, two case studies have been demonstrated on the remaining two image datasets, viz., *Coil20* and *Olivetti Faces*, and on one biological dataset. These datasets consist of a comparatively larger number of features. Moreover, the smaller sample size makes the task of feature selection even more complex.

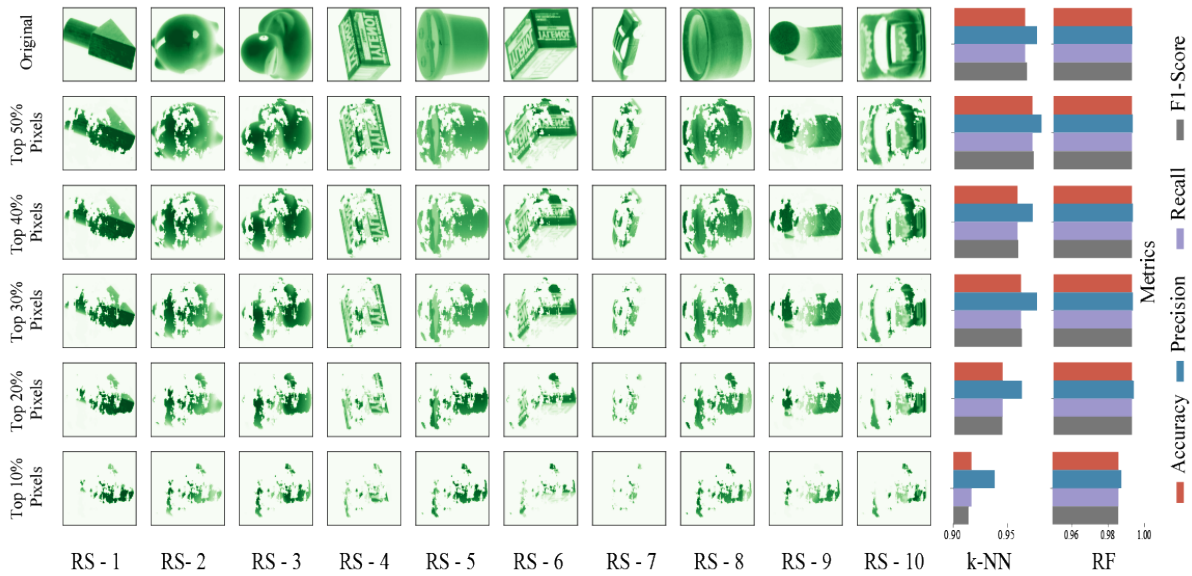


Figure 3.4: Visual representation of randomly selected samples (RS) through the top features selected by NDFS for *Coil20* dataset. The classification performance of k-NN and RF has been shown over the selected feature space.

Results on Image datasets

This Section discusses a case study on two image datasets. Initially, NDFS has been examined on different sizes of top features. In other words, this analysis will help us to understand how informative the top features are. In this context, 5 sets of top features have been considered, viz., top 10%, top 20%, top 30%, top 40%, and top 50%. 10 samples have been drawn from the datasets randomly and visualized through these top features or pixels. More precisely, the randomly selected images have been visualized only through the top pixels, and the rest of the pixels have been set to 0. It has been observed in Figure 3.4 that objects are clearly identifiable through the naked eye with only 30% top features. In this regard, a classification performance has been carried out on these selected features using k-NN and RF classifiers, and it has been found that even with 10% features both the classifiers achieve an accuracy, precision, recall, and F1-score over 0.9, i.e., even those top 10% features are efficient enough to distinguish different objects almost certainly (Figure 3.4).

A similar experiment has been carried out on the *Olivetti Faces* dataset. As shown in Figure 3.5, unlike *Coil20* dataset, here the random samples are not identifiable to the naked eye. However, both the classifiers, k-NN and RF, are able to do the same with an accuracy of at least 0.85 and 0.65 respectively. Both classifiers, with top 30% features,

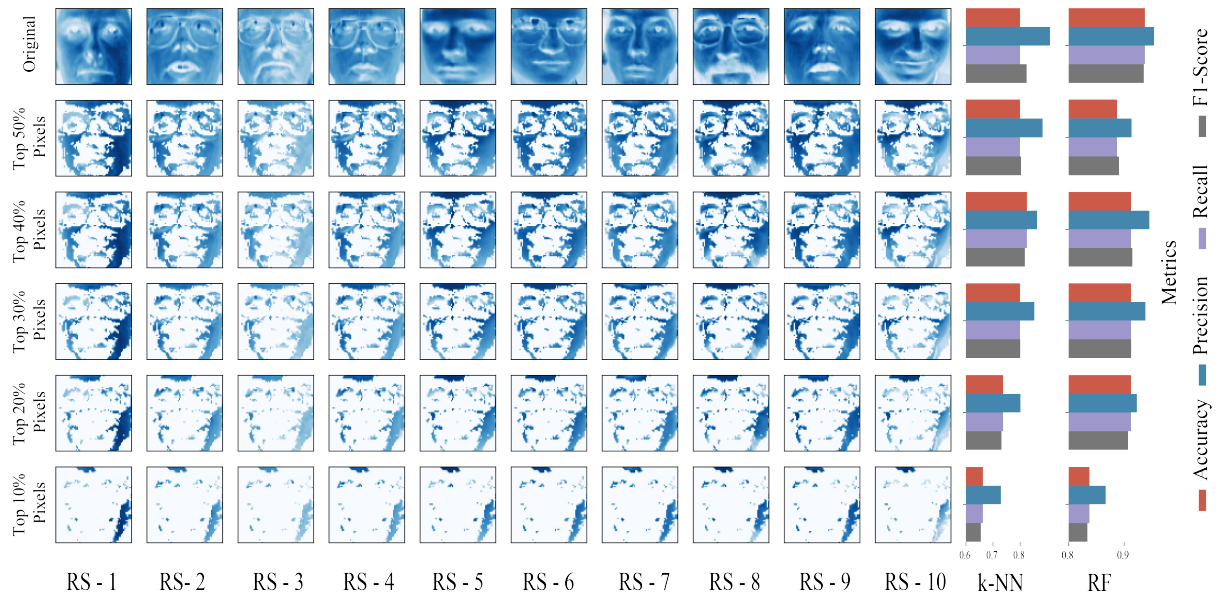


Figure 3.5: Visual representation of randomly selected samples (RS) through the top features selected by NDFS for *Olivetti Faces* dataset. The classification performance of k-NN and RF has been shown over the selected feature space.

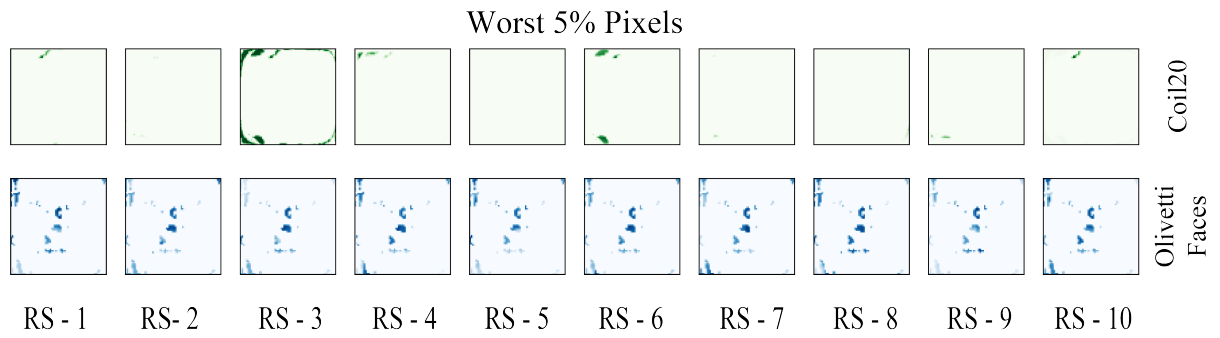


Figure 3.6: Visual representation of randomly selected samples through the worst 5% features selected by NDFS for *Coil20* (upper) and *Olivetti Faces* (lower) datasets.

have shown superior classification results.

In the previous experiments, it has been found that, for both the image datasets, the top features are efficient enough to perform downstream tasks with a higher accuracy. A similar experiment has been carried out on both datasets, considering the top 5% worst features. It has been observed from Figure 3.6 that for both the datasets, the random samples look very similar, while observed through the worst features. The k-NN classifiers also reflect poor results with an accuracy of 0.5 for *Coil20* dataset, and 0.42 for *Olivetti Faces* dataset, while applied on this worst feature space.

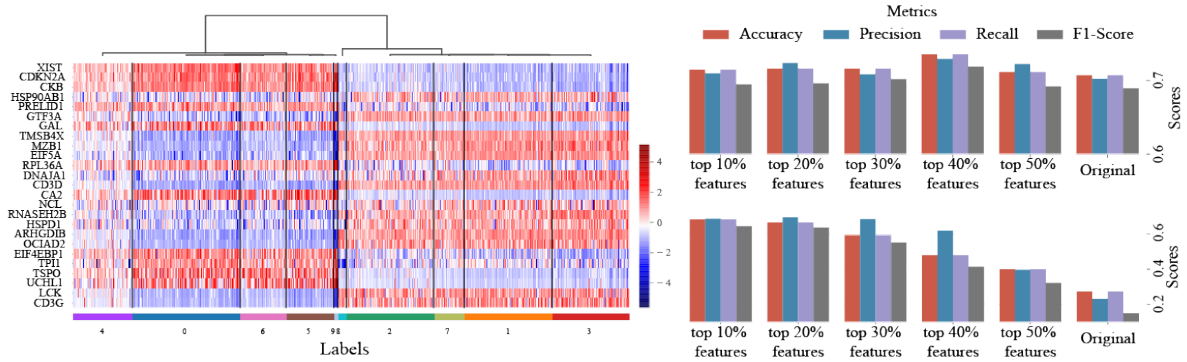


Figure 3.7: Visual representation of top 25 features selected by NDFS for *Jurkat* dataset (left). The classification performance of k-NN and RF has been shown over the selected feature spaces (right).

Results on Biological dataset

Finally, the performance of NeuroDAVIS-FS has been evaluated on a biological dataset, viz., *Jurkat* dataset. Initially, the dataset was pre-processed using the standard pipeline proposed in Scanpy¹. The dataset contains gene expressions for 3388 cells across 32738 genes. Feature extraction has been performed on *Jurkat* dataset, followed by a feature selection. As shown in Figure 3.7 (left), the top 25 features selected by NDFS have been shown using a heatmap. Cluster information in the heatmap suggests that the top features can separate the major cell clusters. Moreover, a classification performance has been performed on the different sizes of top features. It has been observed from Figure 3.7(right) that both k-NN and RF classifiers' performance improved as the feature size decreased. k-NN classifier using only top 10% of the features produces an accuracy 0.65, and on the other hand RF classifier produces an accuracy 0.73 using top 40% features. This analysis reflects the robustness of the model, developed for feature selection.

3.4 Discussion and Conclusion

Dealing with high-dimensional datasets is challenging due to computational complexity, the curse of dimensionality, and model overfitting. It necessitates reducing the dimension of the dataset for a better understanding of inherent information, and feature selection techniques are well-suited to do the same. In this chapter, a neural network-based model, NeuroDAVIS-FS (NDFS), has been developed, which is an extension of the earlier data

¹<https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html>

visualization model NeuroDAVIS. The feature selection method, NDFS, performs in an unsupervised learning setup without assuming any prior data distribution. Initially, it considers training using the model NeuroDAVIS and selecting features according to the trained model. The efficacy of NDFS has been demonstrated on various datasets from different domains, viz., Numeric: *Breast Cancer*, and *Wine*, Image: *Digits*, *Coil20*, and *Olivetti Faces* and Biological: *Jurkat*. A comparative analysis has been executed for several downstream tasks like classification and clustering with the state-of-the-art models, and it has been observed that the top features selected by NDFS are efficient for the same. Moreover, comparable results are obtained in terms of predefined metrics for the state-of-the-art methods. Being a neural network-based approach, the model is non-parametric and performs in an unsupervised setup, which makes the model novel in comparison with traditional cases. In addition to that, a good amount of overlapping features have been found in NDFS feature selection model against the other classical models. Finally, two case studies on image and biological datasets with a very low sample-feature ratio (nearly 1 : 10) have been executed and found to be effective for relevant feature selection. Even with 10% features, both the classifiers classify the classes of *Coil20* dataset with an accuracy of 0.9 approximately. However, for *Olivetti Faces* dataset, the accuracy rate of k-NN and RF are approximately 0.7 and 0.85 respectively. Whereas, for *Jurkat* dataset, the accuracy rate increases for the classification tasks with the top 10% selected features.

Though NDFS outperforms the state-of-the-art methods, it still has some limitations. As the performance of NDFS depends on the prior training of NeuroDAVIS, the model may not be efficient in the case of biased training. Moreover, the optimal number of features has not been considered in this study, which might be an extension of a future study. Though one can plot the feature scores in a descending order and select the top features by observing the elbow. The location of a bend (knee) in the plot is generally considered an optimal number of features.

In Chapters 2 and 3, we have explored a data visualization approach and a feature selection method, respectively, and demonstrated their effectiveness across multiple datasets from diverse application domains. However, when data comes from multiple modalities, looking at each one separately often gives an incomplete picture. Joint visualization brings these different views together into a shared space, making common patterns, relationships, and differences across modalities visible in an intuitive and interpretable way. Motivated

by this need, Chapter 4 introduces a neural network-based model, designed to generate a joint visualization of multi-modal data. Recent technological advancements have enabled joint sequencing techniques that allow the simultaneous measurement of multiple modalities, thereby providing complementary views of the same cells of interest. Therefore, we have evaluated the model's performance over several such multi-omics (CITE-seq and ATAC-seq) datasets.

Chapter 4

NeuroMDAVIS: Visualization of single-cell multi-omics data under deep learning framework

4.1 Introduction

Multi-modal datasets originate from different sources, and a large number of dimensions in each modality makes the data incredibly complex. Thus, reduction of dimension while preserving the inherent structures within the data, is a basic requirement for easy exploration and interpretation of data. Visualization helps in identifying patterns, clusters, outliers, and correlations within these multi-modal datasets.

Besides, with technological advancements, the realm of molecular biology has advanced into a plethora of possibilities. Single-cell technology has opened up newer dimensions for omics data analyses that lead to unravelling of the disease or developmental processes at the cellular level. Single-cell RNA-sequencing (scRNA-seq) allows measurement of mRNA expressions at a cellular resolution. Current scRNA-seq protocols include DROP-seq [80], SMART-seq2 [81], 10x Genomics [82, 83, 72, 84], 10x Chromium [85], CEL-seq2 [86] and MARS-seq [87]. ATAC-seq [88], on the other hand, allows sequencing of open chromatin regions within the genome. More recently, joint sequencing technologies have come up, which allow simultaneous measurement of more than one modalities, thereby offering multiple views of the same cells of interest. These technologies include CITE-seq [89], REAP-seq [90] (joint measurement of proteins and transcriptome), scM&T-seq [91],

scMT-seq [92], scTrio-seq [93] (simultaneous measurement of methylome and transcriptome), and sci-CAR [94], SNARE-seq [95], and SHARE-seq [96] (joint measurement of chromatin accessibility and transcriptome in single-cells or nuclei). All these technologies facilitate deeper understanding of the cellular identity and processes in organs and tissues.

Interestingly, when it comes to visualizing observations from multi-omics experiments, there is no joint data visualization method that can combine views from several omics layers. Such a joint visualization can assist researchers in explaining the underlying biology of a system/process in a better way. MultiMAP [59], a recently developed method for dimensionality reduction and multi-omics integration (a more generalized version of UMAP), enables shared visualization of multi-omics datasets. It creates a non-linear manifold representing different omics modalities, allowing the identification of shared features across cells in a co-embedded space. However, the shared visualization produced by MultiMAP generates separate embeddings for each of the omics modalities.

Thus, motivated by the wide benefits of using neural networks, this chapter introduces a multi-modal data visualization model, called NeuroMDAVIS, developed under a deep learning framework to address this challenge. NeuroMDAVIS is capable of extracting crucial features from the data and producing an effective joint visualization. It is a feed-forward deep learning model that does not assume any kind of data distribution to visualize high-dimensional data. Most of the state-of-the-art methods have been observed to perform well on complex datasets when conjugated together with an initialization step, like that of applying PCA, to reduce the number of dimensions to something less than or equal to 50 [41]. NeuroMDAVIS does not require any such initialization, which makes it more effective and efficient. Our primary aim being multi-omics data visualization, in this study, we have limited our experiments to multi-omics datasets only. NeuroMDAVIS is the first of its kind to generate a joint visualization of different omics modalities. It facilitates visualizing multi-omics modalities jointly, thereby providing an integrated view that captures correlations and interactions among modalities directly in the same space. However, it can be applied to other domains as well. Additionally, we have also explored how NeuroDAVIS, a precursor of NeuroMDAVIS [1], performs on these multi-omics datasets when used on each omics modality separately.

The organization of the remaining part of this chapter is summarized as follows. Section 4.2 provides detailed methodology of NeuroMDAVIS. Section 4.3 describes the

experimental results on multi-omics datasets, along with comparisons against NeuroDAVIS and MultiMAP, projection of test samples, and sensitivity of the model hyper-parameters. Section 4.4 provides a brief discussion of the strengths and weaknesses of NeuroMDAVIS, and concludes the chapter.

4.2 Methods

The unsupervised deep learning model, called NeuroMDAVIS, developed in this chapter, allows joint visualization of high-dimensional multi-omics datasets. To the best of our knowledge, it may be considered as the first model that allows joint visualization of multi-omics data. NeuroMDAVIS is a generalized version of NeuroDAVIS [1] introduced earlier to provide visualization of a single data modality. Nevertheless, NeuroMDAVIS incorporates significant changes to the network architecture over NeuroDAVIS. Figures 4.1(A) describes the overall workflow for this chapter, while Figures 4.1(B) and 4.2 depict the detailed architecture of the model developed.

4.2.1 Architecture

NeuroMDAVIS architecture represents a novel structure as shown in Figure 4.1(B). It consists of four different types of layers, viz., an *Input layer*, a *Latent layer*, one or more *Hidden layer(s)* and a *Reconstruction layer*. As shown in Figure 4.1(B), Hidden layer(s) are of two types, viz., *Shared hidden layer(s)* and *Modality-specific hidden layer(s)*. *Input layer*, *Latent layer*, and *Shared hidden layer(s)* are densely connected, i.e., each node in any of these layers is connected to each node of its adjacent layer(s). However, *Modality-specific hidden layer(s)* does not share a completely dense connection, instead, it shares a dense connection modality-wise only.

Let $\mathbf{X} = \{\mathbf{X}^{(r)} : \mathbf{X}^{(r)} \in \mathbb{R}^{n \times d_r}\}_{r=1}^m$ be a dataset containing m different omics modalities. In each modality, there are equal number (say n) of samples/observations, where each sample is described by data in m omics modalities. An r^{th} modality is characterized by d_r features, i.e., $\mathbf{X}^{(r)} = [x_{ik}^{(r)}]_{n \times d_r}$, where $x_{ik}^{(r)} \in \mathbb{R}$, and $r \in \{1, 2, \dots, m\}$. The *Input layer* of NeuroMDAVIS has n neurons, while the *Latent layer* contains κ ($= 2$ or 3) neurons, where κ is the desired number of dimensions to be used for visualization. Both the *Modality-specific hidden layer(s)* and *Reconstruction layer* have m modality-specific

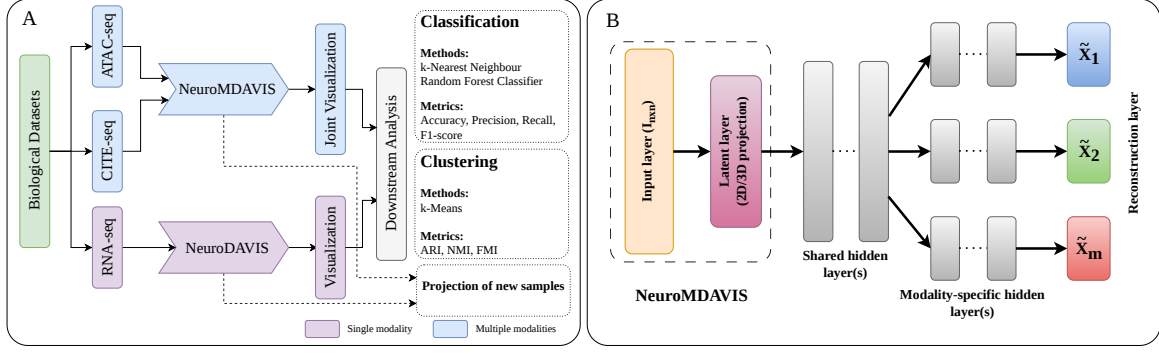


Figure 4.1: (A) A graphical abstract reflecting the overall workflow of this chapter. (B) The architecture of NeuroMDAVIS, that have been developed for multi-omics data visualization.

sub-modules for m omics modalities. In each sub-module of the *Reconstruction layer*, there are d_r neurons. However, the number of neurons in the *Hidden layer(s)* are decided empirically.

Input layer of NeuroMDAVIS takes an identity matrix of order $n \times n$ as input. Similar to NeuroDAVIS, the *Input layer* creates a random latent embedding of n samples at the *Latent layer*. *Latent layer* then tries to regress all the data modalities simultaneously through the aforesaid *Hidden layers*. The purpose of using a *Shared hidden layer* is to capture the common information across different modalities for a particular sample/observation, whereas, a *Modality-specific hidden layer* extracts the modality-specific information for that sample/observation. In this work, only one *Shared hidden layer* and one *Modality-specific hidden layer* has been used. One can use multiple such layers based on the problem requirement. When $m = 1$, the architecture resembles that of NeuroDAVIS with general *Hidden layer(s)* instead of distinct *Shared hidden layer(s)* and *Modality-specific hidden layer(s)*. Finally, the *Reconstruction layer* reconstructs the data in hand, i.e., the original individual omics modalities.

4.2.2 Forward propagation

We have considered a data matrix $\mathbf{X}$ with m omics modalities, each having n paired samples, i.e., $\mathbf{X} = \{\mathbf{X}^{(r)} : \mathbf{X}^{(r)} \in \mathbb{R}^{n \times d_r}\}_{r=1}^m$. NeuroMDAVIS takes an identity matrix $\mathbf{I}$ of order $n \times n$ as input. An i^{th} row vector $\mathbf{e}_i$ of $\mathbf{I}$ is considered for regressing the i^{th} paired sample (i.e., i^{th} sample from all the omics modality). More precisely, $\mathbf{e}_i$ propagates through the layers of NeuroMDAVIS and reconstructs an approximate version $[\tilde{\mathbf{x}}_i^{(1)}, \tilde{\mathbf{x}}_i^{(2)}, \dots, \tilde{\mathbf{x}}_i^{(m)}]$ of $[\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \dots, \mathbf{x}_i^{(m)}]$ at the *Reconstruction layer*. Let $\mathbf{a}_i^{(Inp)}$ and $\mathbf{h}_i^{(Inp)}$ correspond to

the input to and output from the *Input layer*, and $\mathbf{a}_i^{(Lat)}$ and $\mathbf{h}_i^{(Lat)}$ represent the input to and output from the *Latent layer*, respectively, on presentation of an i^{th} sample. Let $\mathbf{W}^{(Lat)}$ and $\mathbf{b}^{(Lat)}$ be the weight matrix between the *Input layer* and *Latent layer*, and the bias of the nodes in the *Latent layer* respectively. Thus, for the *Input layer*, we have

$$\begin{cases} \mathbf{a}_i^{(Inp)} = \mathbf{e}_i, \\ \mathbf{h}_i^{(Inp)} = \mathbf{a}_i^{(Inp)}, \end{cases} \quad \forall i \in \{1, 2, \dots, n\} \quad (4.1)$$

For the *Latent layer*, we have

$$\begin{cases} \mathbf{a}_i^{(Lat)} = \mathbf{W}^{(Lat)} \mathbf{h}_i^{(Inp)} + \mathbf{b}^{(Lat)}, \\ \mathbf{h}_i^{(Lat)} = \mathbf{a}_i^{(Lat)}, \end{cases} \quad \forall i \in \{1, 2, \dots, n\} \quad (4.2)$$

Here, the weight parameters, controlled by the input $\mathbf{e}_i$, create an independent low-dimensional representation of the i^{th} sample at the *Latent layer*. This is required to ensure that only the links connected to the i^{th} neuron of the *Input layer* activate neurons in the *Latent layer* on presentation of the i^{th} sample.

Let the total number of *Shared hidden layer* and *Modality-specific hidden layer* present in the network be p and q respectively. Let $\mathbf{a}_i^{(SHid_j)}$ and $\mathbf{h}_i^{(SHid_j)}$ correspond to the input to and output from a j^{th} *Shared hidden layer* on presentation of an i^{th} sample. Also let $\mathbf{W}^{(SHid_j)}$ and $\mathbf{b}^{(SHid_j)}$ be the weight matrix and the bias of the nodes in a j^{th} *Shared hidden layer* respectively. Thus, for the first *Shared hidden layer*, we have

$$\begin{cases} \mathbf{a}_i^{(SHid_1)} = \mathbf{W}^{(SHid_1)} \mathbf{h}_i^{(Lat)} + \mathbf{b}^{(SHid_1)} \\ \mathbf{h}_i^{(SHid_1)} = \text{ReLU}(\mathbf{a}_i^{(SHid_1)}), \end{cases} \quad \forall i \in \{1, 2, \dots, n\} \quad (4.3)$$

For $(p - 1)$ *Shared hidden layer(s)*, we have

$$\begin{cases} \mathbf{a}_i^{(SHid_j)} = \mathbf{W}^{(SHid_j)} \mathbf{h}_i^{(SHid_{j-1})} + \mathbf{b}^{(SHid_j)} \\ \mathbf{h}_i^{(SHid_j)} = \text{ReLU}(\mathbf{a}_i^{(SHid_j)}), \end{cases} \quad \forall i \in \{1, 2, \dots, n\}; \forall j \in \{2, 3, \dots, p\} \quad (4.4)$$

where $\text{ReLU}(\mathbf{y}) = \max(\mathbf{0}, \mathbf{y})$; $\max$ (maximum) being an element-wise operation. The output from the last *Shared hidden layer* will flow to the end of the network through different *Modality-specific hidden layer(s)*, and reconstruct every modality at the *Recon-*

struction layer. Let us assume that $\mathbf{a}_i^{(MHid_j^{(r)})}$ and $\mathbf{h}_i^{(MHid_j^{(r)})}$ correspond to the input to and output from an r^{th} module of a j^{th} *Modality-specific hidden layer* on presentation of an i^{th} sample. Also let $\mathbf{W}^{(MHid_j^{(r)})}$ and $\mathbf{b}^{(MHid_j^{(r)})}$, respectively, be the weight matrix and the bias of the nodes in an r^{th} module of a j^{th} *Modality-specific hidden layer*. Thus, for the first *Modality-specific hidden layer*, we have

$$\begin{cases} \mathbf{a}_i^{(MHid_1^{(r)})} = \mathbf{W}^{(MHid_1^{(r)})} \mathbf{h}_i^{(SHid_p)} + \mathbf{b}^{(MHid_1^{(r)})} \\ \mathbf{h}_i^{(MHid_1^{(r)})} = \text{ReLU}(\mathbf{a}_i^{(MHid_1^{(r)})}), \end{cases} \quad \forall i \in \{1, 2, \dots, n\}; \forall r \in \{1, 2, \dots, m\} \quad (4.5)$$

For the remaining $(q - 1)$ *Modality-specific hidden layer(s)*, we have

$$\begin{cases} \mathbf{a}_i^{(MHid_j^{(r)})} = \mathbf{W}^{(MHid_j^{(r)})} \mathbf{h}_i^{(MHid_{j-1}^{(r)})} + \mathbf{b}^{(MHid_j^{(r)})} \\ \mathbf{h}_i^{(MHid_j^{(r)})} = \text{ReLU}(\mathbf{a}_i^{(MHid_j^{(r)})}), \end{cases} \quad \forall i \in \{1, 2, \dots, n\}; \forall r \in \{1, 2, \dots, m\}; \quad (4.6)$$

$$\forall j \in \{2, 3, \dots, q\}$$

Finally, at the *Reconstruction layer*, a reconstruction (lossy) of the original data is formed. Let us assume that $\mathbf{a}_i^{(Rec^{(r)})}$ and $\mathbf{h}_i^{(Rec^{(r)})}$ correspond to the input to and output from an r^{th} module of the *Reconstruction layer* on presentation of an i^{th} sample. Also let $\mathbf{W}^{(Rec^{(r)})}$ and $\mathbf{b}^{(Rec^{(r)})}$ be the weight matrix and the bias of the nodes, respectively, in an r^{th} module of the *Reconstruction layer*. Thus, for the *Reconstruction layer*, we have

$$\begin{cases} \mathbf{a}_i^{(Rec^{(r)})} = \mathbf{W}^{(Rec^{(r)})} \mathbf{h}_i^{(MHid_q^{(r)})} + \mathbf{b}^{(Rec^{(r)})} \\ \mathbf{h}_i^{(Rec^{(r)})} = \mathbf{a}_i^{(Rec^{(r)})} \\ \tilde{\mathbf{x}}_i^{(r)} = \mathbf{h}_i^{(Rec^{(r)})}, \end{cases} \quad \forall i \in \{1, 2, \dots, n\}; \forall r \in \{1, 2, \dots, m\} \quad (4.7)$$

NeuroMDAVIS projects the latent embedding for a sample, obtained at the *Latent layer*, to different d_r dimensional spaces, corresponding to each of r^{th} omics modalities, through the *Hidden layer(s)*, so that the sample gets reconstructed at the *Reconstruction layer*. The term $\mathbf{h}_i^{(Rec^{(r)})}$ represents lossy reconstruction $\tilde{\mathbf{x}}_i^{(r)}$ of $\mathbf{x}_i^{(r)}$, which is an i^{th} sample in the r^{th} omics modality.

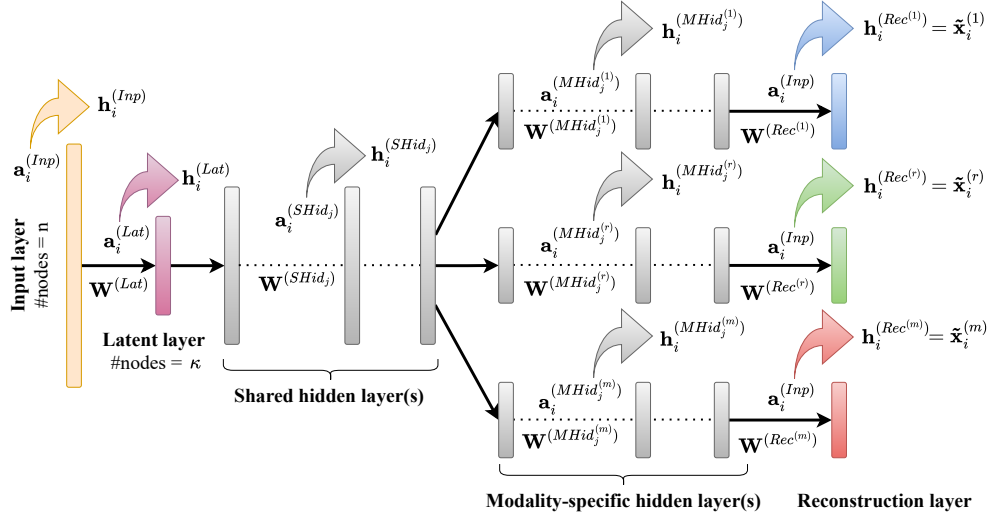


Figure 4.2: The network architecture of NeuroMDAVIS, along with the mathematical notations.

4.2.3 Learning

NeuroMDAVIS enables dimension reduction and visualization of high-dimensional multi-omics data. Similar to NeuroDAVIS, NeuroMDAVIS also tries to reconstruct the data to be visualized. For i^{th} sample from the r^{th} omics modality, $\mathbf{x}_i^{(r)}$, NeuroMDAVIS tries to minimize the reconstruction error $\|\mathbf{x}_i^{(r)} - \tilde{\mathbf{x}}_i^{(r)}\|$ in order to find an optimal reconstruction $\tilde{\mathbf{x}}_i^{(r)}$ of the sample $\mathbf{x}_i^{(r)}$. Since different omics modalities may have different number of dimensions, to avoid disparity in learning, a balancing parameter γ_r has been introduced into the objective function for NeuroMDAVIS. The value of γ_r lies in $(0, 1]$, where higher value of γ_r implies higher weightage to the r^{th} data modality. γ_r is useful when knowledge about the data modalities are available *a priori*. In absence of prior knowledge, one can simply choose $\gamma_r = 1, \forall r$. The objective function thus becomes

$$\mathcal{L}_{\text{NeuroMDAVIS}} = \frac{1}{n} \sum_{r=1}^m \sum_{i=1}^n \gamma_r \|\mathbf{x}_i^{(r)} - \tilde{\mathbf{x}}_i^{(r)}\|_2^2 \quad (4.8)$$

In order to avoid overfitting and minimization of model complexity, $L2$ regularization, involving activities of nodes and weights, has been considered. This stops both weights

and layer outputs from exploding. The objective function thus becomes

$$\begin{aligned} \mathcal{L}_{NeuroMDAVIS} = & \frac{1}{n} \sum_{r=1}^m \sum_{i=1}^n \gamma_r \|\mathbf{x}_i^{(r)} - \tilde{\mathbf{x}}_i^{(r)}\|_2^2 + \\ & \lambda_1 \sum_{i=1}^n \left[\sum_{j=1}^p \|\mathbf{h}_i^{(SHid_j)}\|_2^2 + \sum_{j=1}^q \sum_{r=1}^m \|\mathbf{h}_i^{(MHid_j^{(r)})}\|_2^2 \right] + \\ & \lambda_2 \left[\sum_{j=1}^p \|\mathbf{W}^{(SHid_j)}\|_F^2 + \sum_{j=1}^q \sum_{r=1}^m \|\mathbf{W}^{(MHid_j^{(r)})}\|_F^2 \right], \quad (4.9) \end{aligned}$$

where λ_1 and λ_2 are regularization parameters set empirically.

We have used the Adam optimizer for training NeuroMDAVIS. The number of epochs needed for convergence has been set empirically by monitoring the convergence behaviour of the training loss. Section 4.3.4 further explains how other hyper-parameters have been tuned. On convergence, NeuroMDAVIS learns a parametric function that can efficiently produce the omics modalities reconstructed from the low-dimensional latent embedding. This latent embedding has then been extracted to produce a joint visualization of the multi-omics data.

4.2.4 Projection of new observations

The first two layers of NeuroMDAVIS, viz., *Input layer* and *Latent layer*, are used to learn suitable regressors. In other words, they control the points in the low-dimensional embedding in each iteration, while the remaining part of the network tries to optimize a function that can project the low-dimensional data into high dimension. Once NeuroMDAVIS is trained on a training dataset, we can visualize the data at the *Latent layer*. The weights from the *Latent layer* through the *Reconstruction layer* have been learned in such a way that it can project any low-dimensional point to a high-dimensional space.

In order to visualize new observations (not present during training but having a similar distribution), the *Input layer* of NeuroMDAVIS has to be presented with an identity matrix which will be of the order of the number of observations in this unseen data (test dataset). This test dataset must contain the same number of omics modalities as the training data. Thus, the *Input layer* dimension will be equal to the size of the test dataset. The weights between the *Input layer* and the *Latent layer* need to be initialized again, and only the sub-network comprising the *Input layer* and the *Latent layer* needs to be re-trained, allowing the weights connecting these two layers to be updated, while the

Table 4.1: Summary of datasets used in this chapter

Dataset	Technology	Description	Cells	RNAs	ADTs/ Peaks	Batches present	Source
<i>bmcite30k</i>	CITE-seq	scRNA-seq profiles measured alongside a panel of antibodies from bone marrow	30672	17009	25	Yes	[97]
<i>kotliarov50k</i>	CITE-seq	CITE-seq profiling of 82 surface proteins and transcriptomes of 53,201 single cells from healthy high and low influenza-vaccination responders	58654	32738	87	Yes	[98, 56]
<i>pbmc10k</i>	Multome	Single-cell multiome ATAC and gene expression data from cryopreserved human peripheral blood mononuclear cells (PBMCs) of a healthy female donor	11909	36601	108377	No	[99]

others are kept frozen to the already learned values. These weights are then used to visualize new samples.

Let there be n_1 observations in the test data. In order to visualize the test data, an identity matrix of size $n_1 \times n_1$ is fed to the *Input layer* of the new network. Similar to the training phase, a loss is calculated at the *Reconstruction layer*. However, unlike training, during visualization of new samples, only the weights connecting the *Input layer* to the *Latent layer* are updated via standard back-propagation. Upon convergence, the final latent embedding for these unseen samples can be extracted from the *Latent layer*.

4.3 Results

Superior performance of NeuroMDAVIS has been demonstrated on multi-omics datasets (Table 4.1), including CITE-seq (Section 4.3.1) and multiome data (Section 4.3.1), over some state-of-the-art visualization methods, like t-SNE, UMAP, IVIS and Fit-SNE. Moreover, it has also been compared against the shared embedding produced by MultiMAP in Section 4.3.2. Table 4.2 presents a comparison of basic features of NeuroMDAVIS with those of the above existing methods. Finally, the sensitivity analysis of hyper-parameters has been discussed in Section 4.3.4.

Table 4.2: Comparison of basic features of the visualization methods used in this chapter

Method	Type	Supports multi-modal visualization (Yes/No)	Makes assumption about data distribution (Yes/No)
NeuroMDAVIS	Neural network-based	Yes (joint visualization)	No
NeuroDAVIS [1]	Neural network-based	No	No
t-SNE [29]	Distribution-based	No	Yes
UMAP [39]	Manifold-based	No	No
Fit-SNE [37]	Distribution-based	No	Yes
IVIS [50]	Neural network-based	No	No
MultiMAP [59]	Manifold-based	Yes (shared visualization)	No

4.3.1 Visualizing multi-omics data using NeuroMDAVIS

To evaluate the performance of NeuroMDAVIS with respect to multi-omics data visualization, we have used two multi-omics datasets, including two CITE-seq (joint profiling of RNA and surface protein measurements) and one multiome (paired RNA-seq and ATAC-seq) dataset. Since the existing visualization methods used for comparison do not support multi-modal visualization, for each dataset, we have concatenated the omics modalities available against the set of paired cells and used them as input to these methods for visualization. The following subsections describe the experiments carried out on these two datasets and their corresponding results.

Visualizing CITE-seq data

First, we have considered two CITE-seq datasets, viz., *bmcite30k* and *kotliarov50k*. These datasets have been downloaded from and preprocessed following [55]. CITE-Seq allows concurrent measurement of mRNA and cell surface proteins. In order to trace cell-types, most of the existing visualization methods either use separate projections for ADTs and RNA molecules, or project the protein expressions on the transcriptomic landscape. However, NeuroMDAVIS supports joint visualization of multiple modalities. As shown in Figure 4.3, for these two datasets, NeuroMDAVIS has been able to keep distinct cell types into distinct clusters, the clustering quality being qualitatively comparable only to IVIS. However, t-SNE, UMAP and Fit-SNE have been unable to keep clusters well

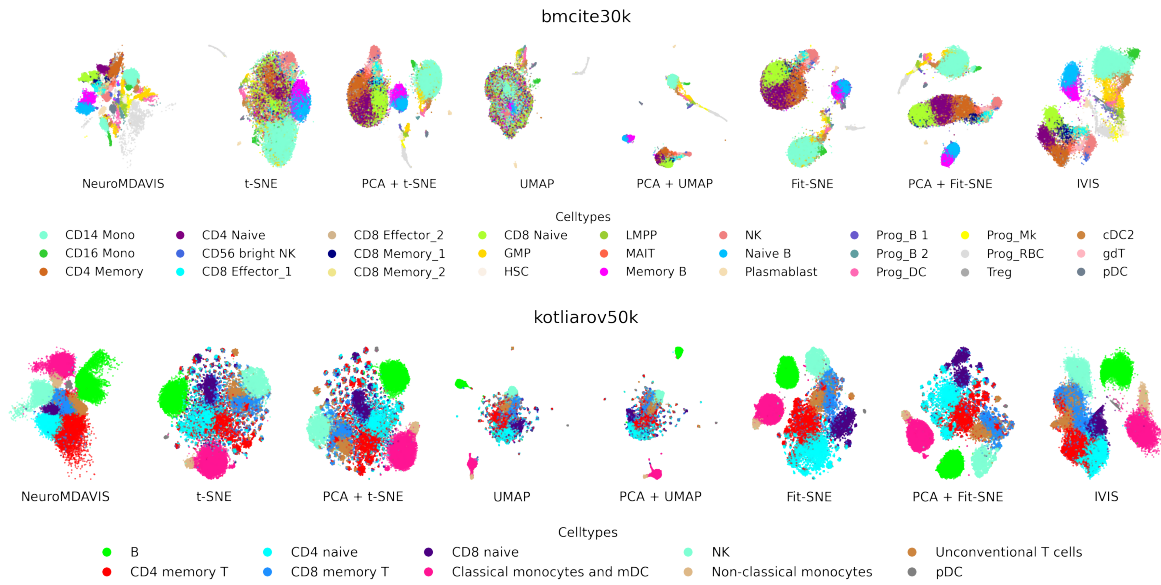


Figure 4.3: 2-dimensional embeddings produced by NeuroMDAVIS, t-SNE, PCA + t-SNE, UMAP, PCA + UMAP, Fit-SNE, PCA + Fit-SNE, and IVIS for *bmcite30k* (first row), and *kotliarov50k* (Second row) datasets.

separated. We had initially apprehended that this might be due to the absence of a prior initialization step, like the usage of PCA, which sometimes serves as a good initialization for the state-of-the-art methods, like t-SNE and UMAP [100]. Hence, we have carried out further experiments by using PCA as a prior initialization step for t-SNE, UMAP, and Fit-SNE, as shown in Figure 4.3. Though the results have improved quite a bit, NeuroMDAVIS and IVIS have still remained the best-performing models in terms of visualization.

Thereafter, we have used k -NN and Random Forest classifiers to identify cell-types based on the NeuroMDAVIS-generated projection, and compare results with that performed on projections generated by other existing methods. Training and test datasets have been prepared in a 80 : 20 ratio. The parameters k , representing the number of neighbours in k -NN, and n , being the number of estimators in Random Forest classification, have been varied between 5 and 45, and 20 and 100 respectively to ensure consistency of results. As can be seen in Figures 4.4(A)-(D), NeuroMDAVIS has outperformed all other methods with respect to Accuracy, Precision, Recall and F1-score values for both the datasets, consistently over all values of k (for k -NN) and n (for Random Forest classifiers), on the 20% held out test dataset. The prior implementation of PCA has improved the performance of t-SNE, UMAP, and Fit-SNE, but has still fallen short of achieving the

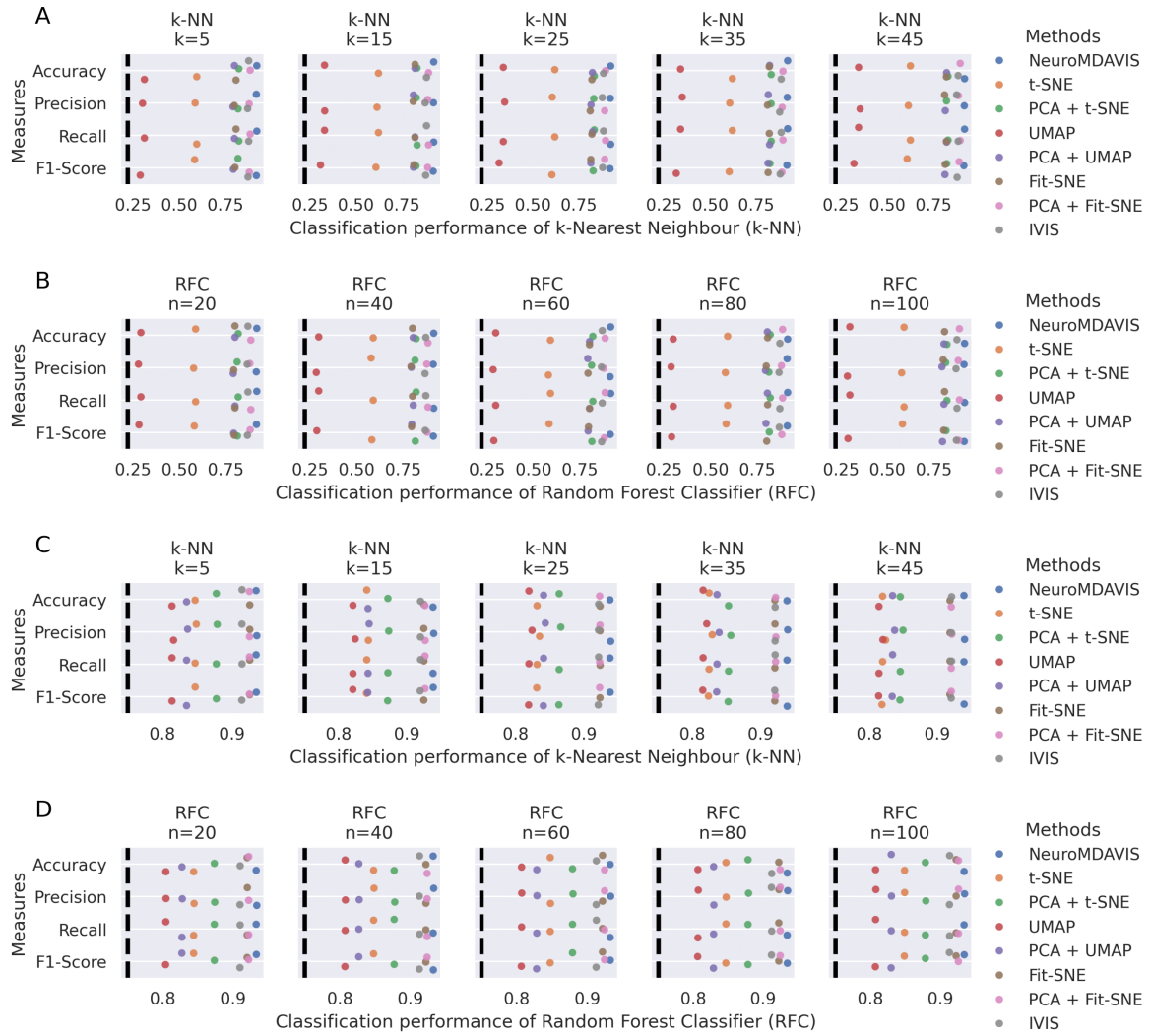


Figure 4.4: (A) and (B) show classification performance on low-dimensional embeddings of *bmcite30k* dataset generated using NeuroMDAVIS, t-SNE, PCA + t-SNE, UMAP, PCA + UMAP, Fit-SNE, PCA + Fit-SNE, and IVIS, in terms of accuracy, precision, recall and F1-score using k -NN and Random Forest classifiers respectively. (C) and (D) show the same on low-dimensional embeddings of *kotliarov50k* dataset.

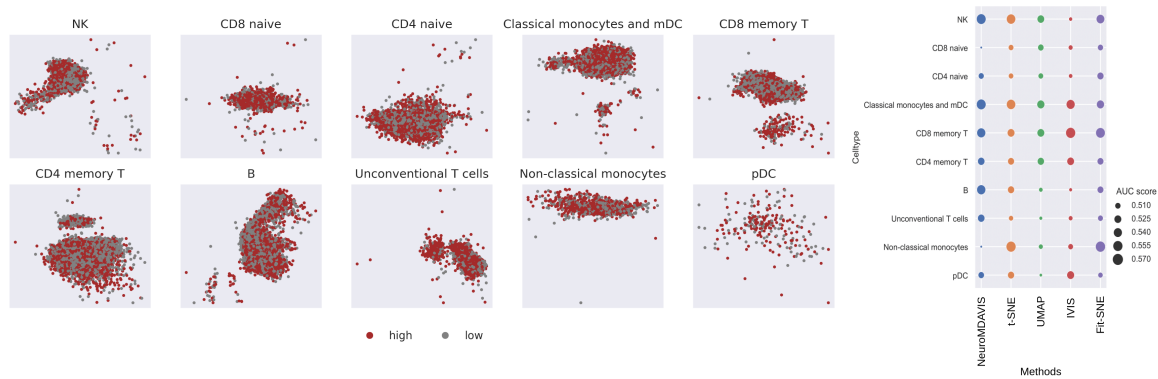


Figure 4.5: NeuroMDAVIS-produced sub embeddings (based on cell types), colored by high/low influenza vaccine responder (left). Area-Under-the-Curve (AUC)-score for high/low influenza vaccine responder classification obtained using k -NN classifier (with $k = 50$) on the NeuroMDAVIS, t-SNE, UMAP, IVIS and Fit-SNE projections of *kotliarov50k* dataset (right).

accuracy of NeuroMDAVIS.

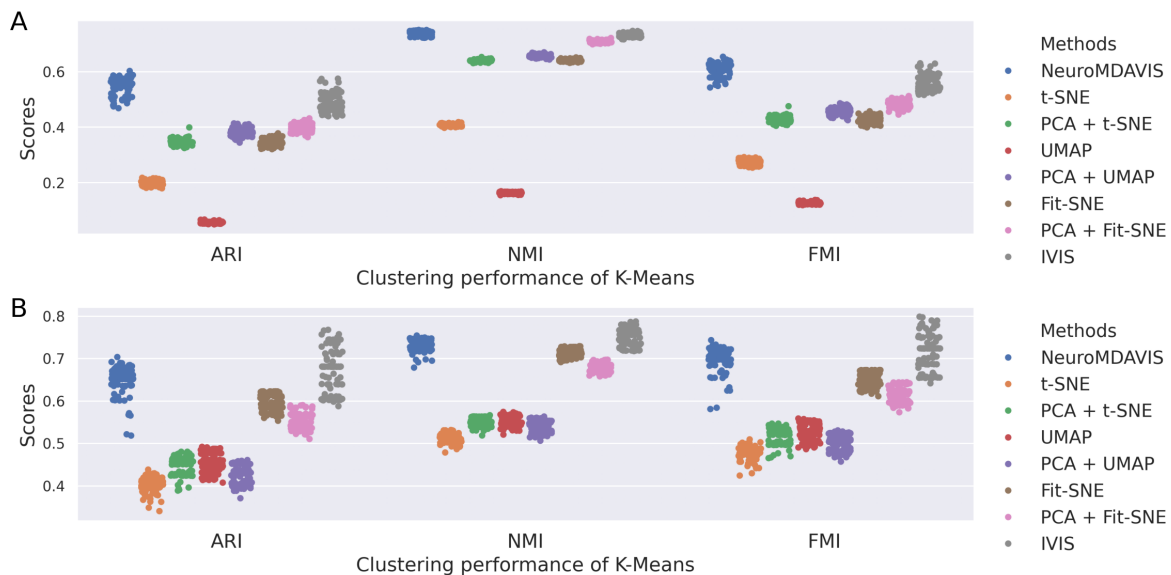


Figure 4.6: Clustering performance on low-dimensional embeddings generated using NeuroMDAVIS, t-SNE, PCA + t-SNE, UMAP, PCA + UMAP, Fit-SNE, PCA + Fit-SNE, and IVIS, with respect to ARI, NMI and FMI scores of (A) *bmcite30k*, and (B) *kotliarov50k* datasets, using k -means clustering algorithm.

Further, k -means clustering on the embedding generated by NeuroMDAVIS, has shown highly competitive performance, if not better than the other existing methods, in terms of Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Fowlkes Mallows Index (FMI) scores. As depicted in Figures 4.6(A)-(B), for *bmcite30k* dataset, NeuroMDAVIS has outperformed all the other methods while for *kotliarov50k* dataset, the performance of NeuroMDAVIS has surpassed t-SNE, UMAP, and Fit-SNE by a high

margin. In order to capture robustness, results from 100 distinct runs have been reported using different train-test partitions.

For further validation, we have performed a clinical classification on the *kotliarov50k* embedding generated by NeuroMDAVIS. The dataset *kotliarov50k* deals with single-cell CITE-seq data gathered from high and low-influenza vaccine responders. We have tried to explore whether the NeuroMDAVIS-generated projection is capable of classifying cells into low and high responder categories. For each unique cell cluster, we have performed a train-test split of 80 : 20 ratio. k -NN classifier with $k = 45$ has been used for this binary classification. Figure 4.5 (right) demonstrates that the Area-Under-the-Curve (AUC)-score obtained using NeuroMDAVIS on each of the cell-types has been highly competitive with those obtained using the other state-of-the-art methods, viz., t-SNE, UMAP, IVIS and Fit-SNE. It may be noted here that none of the methods could achieve beyond 58% AUC, which suggests that the cell type labels obtained for this dataset may not be optimal (Figure 4.5 (left)).

Visualizing multiome data

A commonly used method for assessing chromatin accessibility across the genome is the Assay for Transposase-Accessible Chromatin with sequencing (ATAC-seq). It may be mentioned here that one can learn about how chromatin packaging and other variables impact gene expression by using ATAC-seq to sequence open chromatin regions.

In this work, a multiome dataset *pbmc10k* containing paired RNA-seq and ATAC-seq data, has been further used to demonstrate the effectiveness of NeuroMDAVIS for multi-omics data visualization. This *pbmc10k* dataset has been first preprocessed using MUON [99], reduced to highly variable features only and then further processed to match cells in both the RNA and ATAC modalities, as done in [55]. NeuroMDAVIS has then been applied on the multi-omics data to generate a joint embedding of the paired assays. When compared to t-SNE, UMAP and Fit-SNE, NeuroMDAVIS has produced better visualization of *pbmc10k* data, mapping each cell-type to a distinct cluster (Figure 4.7). Prior application of PCA has improved the visualization for t-SNE, UMAP, and Fit-SNE, but still has some overlap among cell types. NeuroMDAVIS-generated embedding is only comparable to that produced by IVIS (Figure 4.7).

Additionally, to evaluate the quality of the embedding quantitatively, we have per-

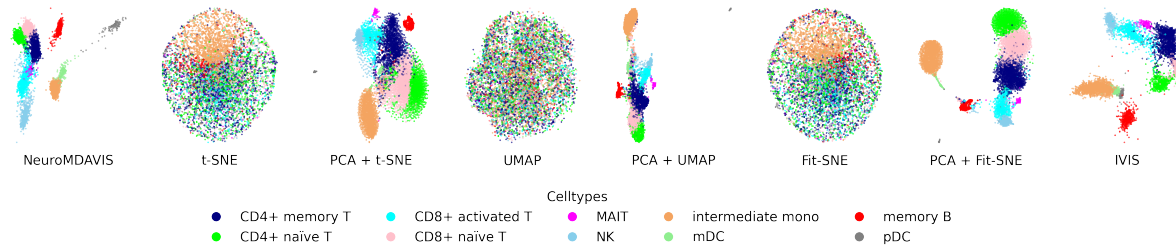


Figure 4.7: 2-dimensional embeddings produced by NeuroMDAVIS, t-SNE, PCA + t-SNE, UMAP, PCA + UMAP, Fit-SNE, PCA + Fit-SNE, and IVIS for *pbmc10k* dataset.

formed further downstream analysis, i.e., classification and clustering of cell-types on the NeuroMDAVIS-generated embedding. A similar train-test split has been used for classification in the case of CITE-seq data. Classification results on the held out test dataset, as shown in Figures 4.8(A) and 4.8(B), demonstrate NeuroMDAVIS as the best performing model for dimension reduction and visualization across all measures like accuracy, precision, recall, and F1-score. On a similar note, *k*-means clustering on the NeuroMDAVIS projection has produced the highest ARI, FMI, and NMI scores compared to all other state-of-the-art methods, like t-SNE, PCA + t-SNE, UMAP, PCA + UMAP, Fit-SNE, PCA + Fit-SNE, and IVIS, as shown in Figure 4.8(C).

4.3.2 Comparison against separate and shared visualization

In order to demonstrate effectiveness of joint-embedding from multi-omics dataset, we have compared that generated by NeuroMDAVIS with separate embeddings produced by NeuroDAVIS. We have further compared NeuroMDAVIS against another state-of-the-art dimensionality reduction and multi-omics integration method, viz., MultiMAP. MultiMAP, however, produces a shared visualization of multiple omics modalities, which projects different omics layers onto separate latent spaces while observing the combined effect of all modalities. In contrast, a joint visualization, as produced by NeuroMDAVIS, refers to generating a single and unified latent embedding by integrating multiple omics modalities.

Figure 4.9 illustrates the joint embedding generated by NeuroMDAVIS, and separate embeddings produced by NeuroDAVIS (a precursor of NeuroMDAVIS), followed by shared embeddings produced by MultiMAP on the *bmcite30k* (CITE-seq) and *pbmc10k* (Multiome) datasets. As discussed earlier, MultiMAP has generated separate embeddings for the RNA and protein modalities. NeuroMDAVIS-generated joint embedding, as observed by Figure 4.9, has resulted in better inter-cell cluster separation than both MultiMAP

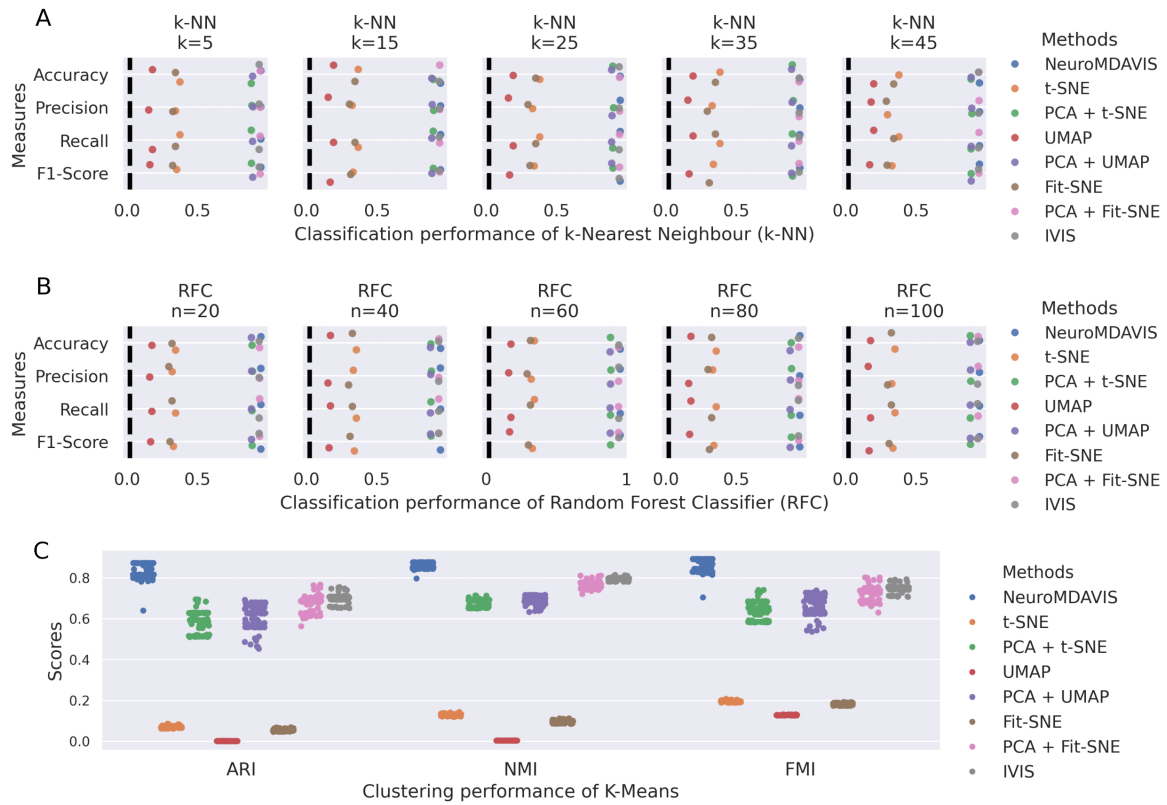


Figure 4.8: (A) and (B) show classification performance on projections of *pbmc10k* dataset generated by NeuroMDAVIS, t-SNE, PCA + t-SNE, UMAP, PCA + UMAP, Fit-SNE, PCA + Fit-SNE, and IVIS in terms of accuracy, precision, recall and F1-score using *k*-NN and Random Forest classifiers.

(C) *k*-means clustering performance on embeddings of *pbmc10k* dataset produced by NeuroMDAVIS, t-SNE, PCA + t-SNE, UMAP, PCA + UMAP, Fit-SNE, PCA + Fit-SNE, and IVIS, in terms of ARI, NMI and FMI scores.

and NeuroDAVIS-generated RNA and protein-embeddings. In addition, the comparative performance of these three methods has been evaluated for the classification and clustering tasks. Classification performance on the NeuroMDAVIS-embedding, RNA and protein-embedding generated by NeuroDAVIS and MultiMAP, has been assessed using metrics such as accuracy, precision, recall, and F1-score, while clustering performance on the same, has been evaluated using Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Fowlkes-Mallows Index (FMI). Figure 4.10 reflects NeuroMDAVIS-generated embeddings have performed better than their counterparts generated by NeuroDAVIS and MultiMAP, in all kinds of downstream tasks for both the datasets.

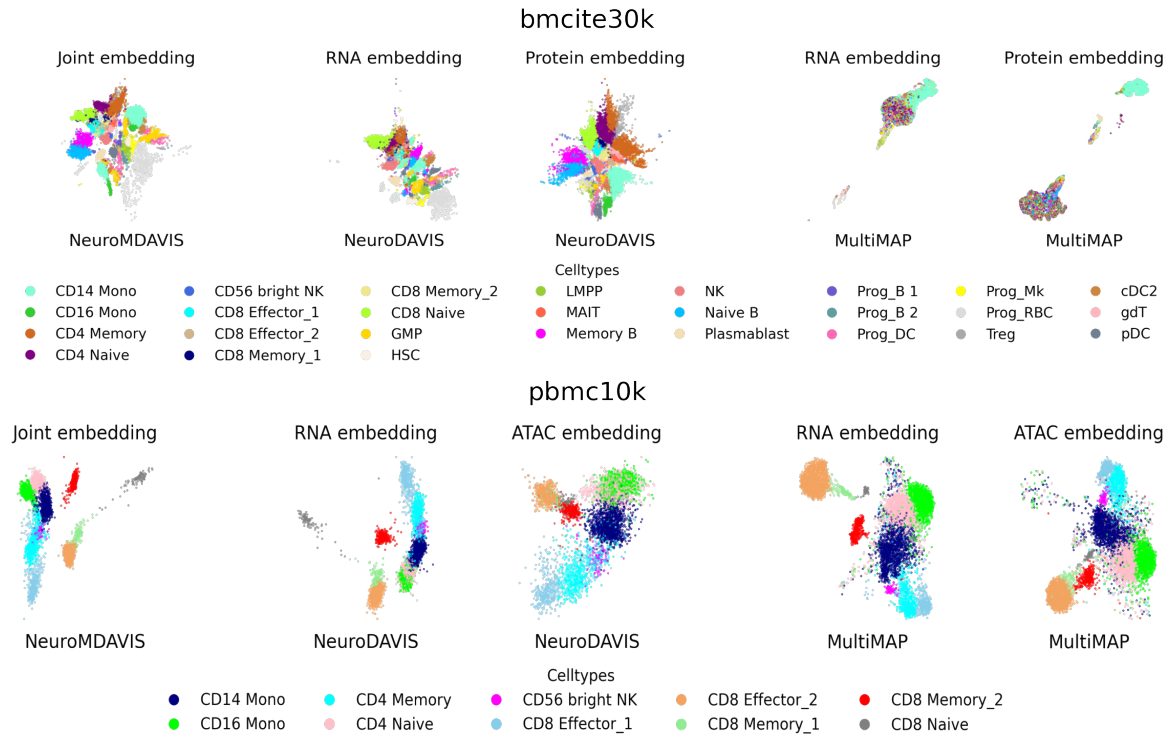


Figure 4.9: 2-dimensional joint embedding produced by NeuroMDAVIS, separate embeddings produced by NeuroDAVIS, and shared embeddings produced by MultiMAP for *bmcite30k* (first row), and *pbmc10k* (Second row) datasets.

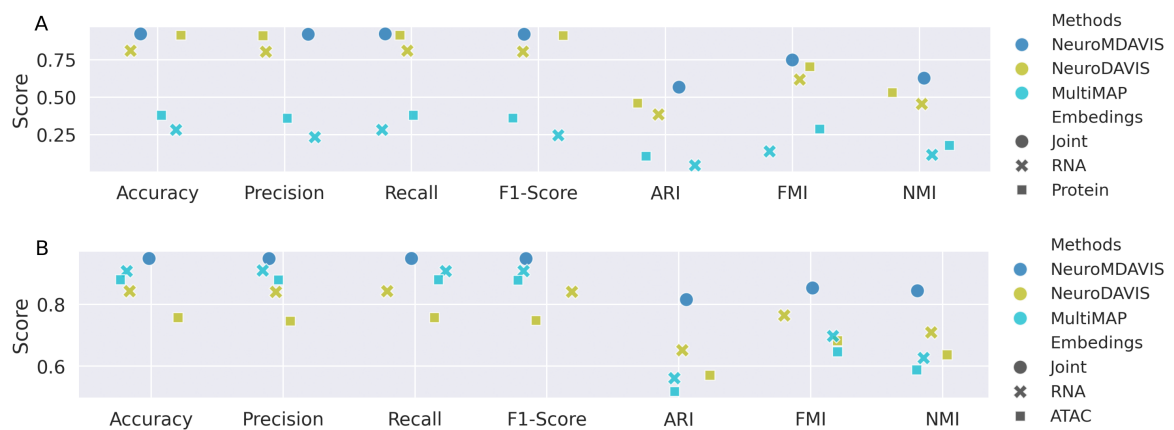


Figure 4.10: Classification and clustering performance on projections generated by NeuroMDAVIS, NeuroDAVIS, and MultiMAP, in terms of accuracy, precision, recall and F1-score, and ARI, NMI, and FMI using k -NN classifiers and k -means clustering, respectively, of (A) *bmcite30k* and (B) *pbmc10k* datasets.

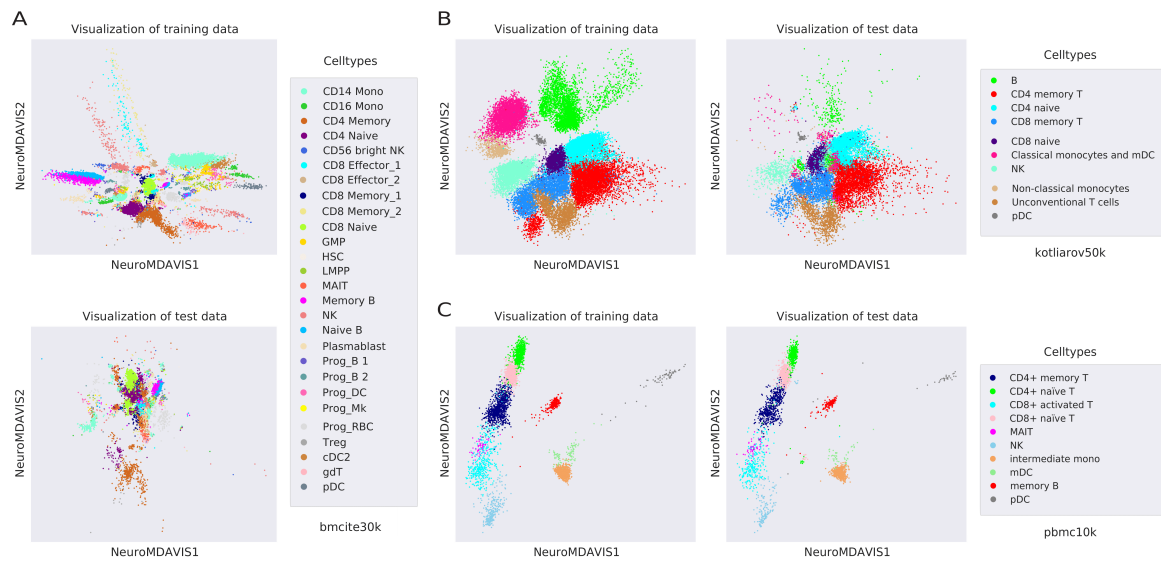


Figure 4.11: (A), (B) and (C) show 2D embeddings produced by NeuroMDAVIS on both the training and test datasets for *bmcite30k*, *kotliarov50k* and *pbmc10k* respectively.

4.3.3 Projection of new observations

NeuroMDAVIS can be used as a pre-trained model to visualize data that are not present during the training process. In order to evaluate NeuroMDAVIS for its capability to visualize new observations not present during training, we have used both CITE-seq and multiome datasets. For each of these datasets, following a 60 : 40 split of training:test datasets, NeuroMDAVIS has been trained on the 60% training data and the held-out test data of 40% samples has been used for visualization of new observations. As shown in Figure 4.11(A), for the *bmcite30k* dataset, the embeddings on training samples and on test samples are very similar. In both the embeddings, *CD8*, *CD4*, *CD14*, *NK* and *B* celltypes are clearly visible. Similar results have been observed on *kotliarov50k* dataset too (Figure 4.11(B)). All the cell-types are identifiable in both the training and test embeddings, with their distributions being quite close to each other.

For the multiome dataset, as depicted in Figure 4.11(C), the major cell types, viz., *T*, *NK* and *B*, are clearly identifiable in both the training and test embeddings. All sub-types of *T* cells are also well separated. Thus, even when the entire data is not available at the time of training, NeuroMDAVIS supports visualization of newer observations added to an existing embedding at runtime.

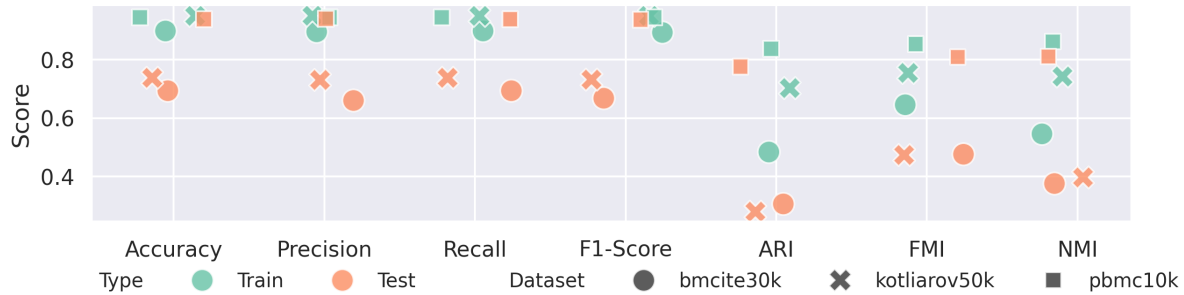


Figure 4.12: Classification and clustering performance on projections generated by NeuroMDAVIS over the train and test parts of *bmcite30k*, *kotliarov50k*, and *pbmc10k* datasets, in terms of accuracy, precision, recall and F1-score, and ARI, NMI, and FMI using k -NN classifiers and k -means clustering, respectively.

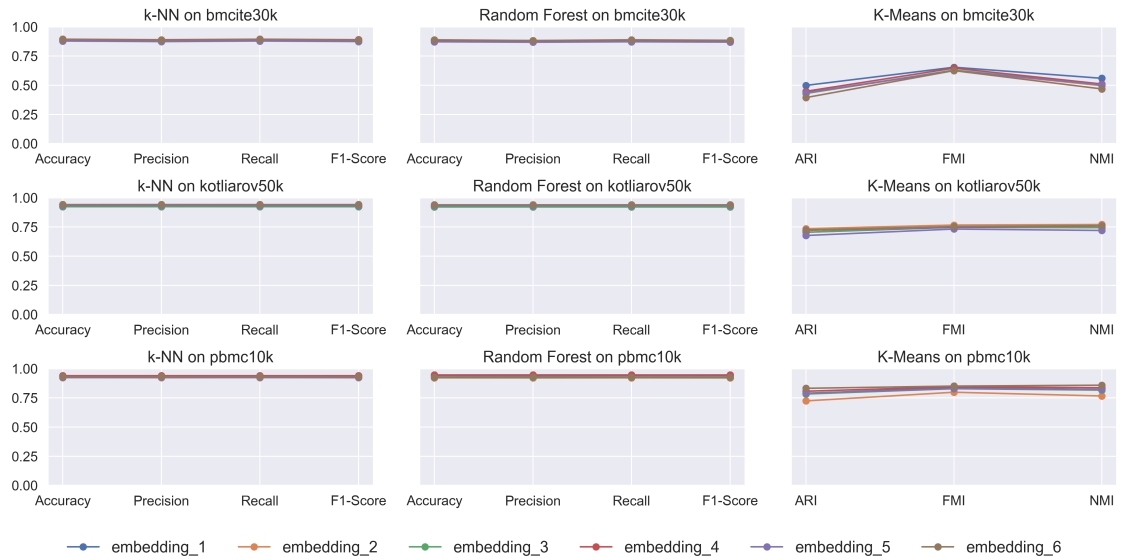
Finally, classification and clustering performance have been compared between the train and test embeddings produced by NeuroMDAVIS. Based on accuracy, precision, recall, and F1-Score, classification performance is comparable for the multi-omics dataset (*pbmc10k*) and the CITE-seq datasets (*bmcite30k* and *kotliarov50k*), as shown in Figure 4.12. Similar results have been observed for clustering in terms of ARI, NMI, and FMI. The drop in performance for CITE-seq datasets is due to poor sample sizes in some of the clusters; for example, the ‘Classical monocytes and mDC’ and ‘Non-classical monocytes’ classes could not be identified properly in their test embeddings due to their smaller sample sizes, as shown in Figure 4.11(B).

4.3.4 Sensitivity analysis

NeuroMDAVIS, being a neural network model, involves several hyper-parameters. In order to evaluate the robustness of NeuroMDAVIS, we have performed a sensitivity analysis on key hyper-parameters. We have systematically varied the key hyper-parameters around their original values. For each specific hyper-parameter value, we have generated a latent embedding for a multi-omics dataset. The embedding qualities have then been quantitatively assessed by evaluating their performance in both classification and clustering tasks. Classification performance has been measured in terms of accuracy, precision, recall, and F1-score to identify the model’s ability to discriminate the classes, while clustering performance has been assessed using ARI, NMI, and FMI to evaluate how well the intrinsic data structure is preserved in the learned latent space.

Figure 4.13 indicates that the performance of NeuroMDAVIS has been robust to

(A)



(B)

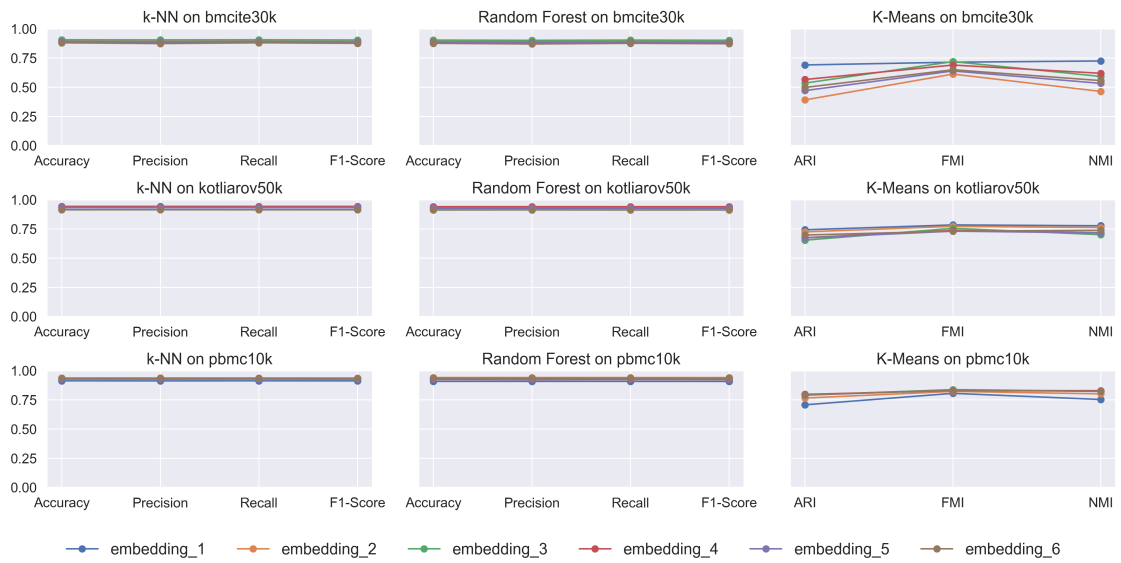


Figure 4.13: Sensitivity analysis for the hyper-parameters (A) ‘batch size’ and (B) ‘hidden layer neurons’ on *bmcite30k*, *kotliarov50k*, and *pbmc10k* datasets. (A) The batch sizes for the six embeddings for *bmcite30k* and *kotliarov50k* have been considered to be 80, 96, 112, 128, 144, and 160; and 32, 48, 64, 80, 96, and 112 for *pbmc10k*. (B) The number of hidden layer neurons for the six embeddings for *bmcite30k*, and *kotliarov50k* have been considered to be [32, [100, 64]], [32, [128, 64]], [16, [100, 64]], [16, [128, 64]], [64, [100, 64]], and [64, [128, 64]]; and [32, [100, 100]], [32, [128, 128]], [16, [100, 100]], [16, [128, 128]], [48, [100, 100]], and [48, [128, 128]] for *pbmc10k*, respectively. Here, the term $[a, [b_1, b_2]]$ conveys the structure of the hidden layers, where a is the number of nodes in the shared hidden layer, and b_1 and b_2 are the same in modality-specific hidden layer modules, respectively.

variations with respect to both batch size and number of hidden layer neurons. The figure also convey that across all three datasets (*bmcite30k*, *kotliarov50k*, and *pbmc10k*), the classification and clustering metrics remain remarkably stable. For classification, k-NN and Random Forest models have consistently achieved high scores for Accuracy, Precision, Recall, and F1-Score, with minimal variance among different embeddings. On a similar note, k-means clustering performance, as measured by ARI, FMI, and NMI, has shown little to no change across the various embeddings. This consistent performance suggests that the model’s ability to generate high-quality latent representations is not significantly affected by changes in batch size or number of neurons in the hidden layers, thereby, demonstrating its robustness and stability.

4.4 Discussion and Conclusion

In this chapter, we have developed a neural network model, called NeuroMDAVIS, for multi-omics data visualization. NeuroMDAVIS provides joint visualization combining all the omics modalities together, and is the first of its kind in this regard. The model is a generalization of NeuroDAVIS [1] recently developed by the authors for the purpose of visualizing single data modality. NeuroMDAVIS enables visualization of multiple omics modalities as a joint embedding which gives rise to a unified perspective depicting correlations and interactions among the modalities in the same space. The performance of NeuroMDAVIS has been demonstrated on CITE-seq and multiome datasets, and results have been compared with several state-of-the-art methods, like t-SNE, UMAP, IVIS, Fit-SNE and MultiMAP. Both NeuroDAVIS and NeuroMDAVIS are feed forward neural network architecture that can be used to produce a latent 2-dimensional embedding from high-dimensional single-omics and multi-omics data, respectively. This latent embedding/projection captures significant data characteristics, which can be useful for various downstream tasks, including classification and clustering.

In the context of multi-omics data visualization, NeuroMDAVIS has been able to produce qualitatively excellent visualization compared to the other state-of-the-art methods. The latent embedding, when used for cell-type classification or clustering, has surpassed those obtained by most of the other methods. NeuroMDAVIS does not assume any prior data distribution, and thereby, is non-parametric. It stands out as one of its key

strengths. Furthermore, to our knowledge, there has been no single method to this date, which supports joint visualization of multiple modalities together. In literature there exist methods for visualizing one modality overlaid on the other(s) or simultaneously distinct embeddings for each modality. NeuroMDAVIS fits into this gap perfectly. Besides, NeuroMDAVIS does not need PCA for prior initialization, contrasting to some other existing methods. Moreover, unlike major existing visualization methods like t-SNE, UMAP, both NeuroDAVIS and NeuroMDAVIS can be used as pretrained models to visualize streaming samples, which implies that they support iterative continuous integration and deployment. This attribute makes them highly suitable for large scale projects like Human Cell Atlas (HCA) or Human Tumor Atlas Network (HTAN) initiatives. Thus, NeuroDAVIS and NeuroMDAVIS together can be claimed to be the new state-of-the-art methods for omics data visualization.

NeuroMDAVIS requires an identity matrix of size equal to the number of samples as input, which may lead to high memory consumption. It is a polynomial-time algorithm. The sparse identity matrix makes the memory overhead manageable, and can be further utilized to handle the issue of scalability in case of large datasets. In the present study, NeuroMDAVIS has been successfully applied to datasets of substantial sizes without encountering computational bottlenecks. It might be due to the usage of small batch sizes during model training, i.e., during any training epoch, only a mini-batch of samples is used instead of the entire training set. For extremely large-scale scenarios (e.g., $> 100k$ cells), an additional optimization, such as block-wise computations or approximate strategies, may be incorporated to further improve scalability, which remains a scope for improvement in future.

Joint visualization can be viewed as a form of feature extraction, and thus, often suffers from limited interpretability. In order to address this limitation, Chapter 5 introduces a feature selection framework built upon the NeuroMDAVIS model, aimed at identifying the most informative features from each omics modality. Using the CPTAC cohort, genomic, transcriptomic, and proteomic data have been integrated to extract modality-specific molecular features. The effectiveness of the selected features is evaluated through Kaplan–Meier survival analysis and univariate CoxPH regression. Furthermore, the concordance index has been used to assess the predictive capability of the identified molecular features compared to conventional clinical features.

Chapter 5

NeuroMDAVIS-FS Identifies Prognostic Molecular Signatures in Lung Cancer towards Survival Analysis and Risk Stratification

5.1 Introduction

Lung cancer remains one of the leading causes of cancer-related mortality worldwide, with a high degree of heterogeneity at the molecular level and a complex tumour microenvironment. Despite surgical treatment, up to 30% of Stage I lung cancer patients experience recurrence within five years of diagnosis [101]. Lung cancer, specifically lung squamous cell carcinoma (LSCC) and lung adenocarcinoma (LUAD), exhibits distinct molecular profiles and clinical behaviors, contributing to their complexity. Both subtypes share some common genetic alterations, such as mutations in the *TP53* gene, but differ significantly in mutational landscapes and histological characteristics, affecting treatment strategies and prognosis. According to the Indian Council of Medical Research–NCDir National Cancer Registry Programme (2020), lung cancer ranks among the leading cancers in India, with tobacco-associated cancers accounting for nearly 27.1% of the overall cancer burden, and a marked rise in lung cancer incidence reported across the majority of registries¹. In the United States, lung cancers are currently the second leading cause of new cancer cases

¹https://www.ncdirindia.org/All_Reports/Report_2020/default.aspx

for both males and females². Therefore, identifying reliable biomarkers for early-stage lung cancer remains an urgent medical priority.

Biomarker prediction has long been supported by survival analysis as a cornerstone of clinical translational research, particularly in oncology, where time-to-event outcomes, such as overall survival and progression-free survival, are considered primary endpoints. Classical approaches based on the Cox proportional hazards model [102] were widely employed to identify prognostic biomarkers by associating molecular features with patient survival while accounting for censoring. As omics data are inherently high-dimensional and multicollinear, regularized extensions, such as LASSO-Cox and elastic-net Cox models, were proposed [103], enabling simultaneous feature selection and risk estimation. More recently, there has been an increasing interest in deep learning-assisted survival analysis for robust biomarker discovery driven by the exponential growth in the availability of high-dimensional omics data and by recognizing limitations of traditional Cox proportional hazard (CoxPH) models in mimicking complex, nonlinear relationships. These methods are at the core of developing multivariate prognostic signatures and risk scores that outcompete single-marker analyses and support patient stratification into clinically meaningful risk groups. The first extensions of survival analysis using neural networks, for example, Cox-nnet [104] or DeepSurv [105], have illustrated that deep feed-forward architectures can flexibly model nonlinear hazard functions while retaining the partial-likelihood framework of Cox regression, thereby enabling simultaneous feature selection (i.e., data-driven identification of survival-associated molecular features) and risk estimation. With advancements in large-scale molecular profiling, subsequent studies have [106, 107] extended this paradigm towards the integration of multi-omics data along with feature selection embedded in survival-aware neural networks to allow simultaneous risk prediction and prioritization of biologically meaningful biomarkers across genomic, transcriptomic, and proteomic layers.

Advances in multi-omics technologies have enabled researchers to integrate multiple layers of biological data along the central dogma, providing deeper insights into disease mechanisms. As these technologies have evolved, data volumes have expanded exponentially, making efficient analysis and interpretation of such high-dimensional data, especially when combined with clinical features, a significant challenge. Machine Learning (ML) and Deep Learning (DL) approaches have become increasingly popular and mainstream

²<https://seer.cancer.gov/statfacts/html/lungb.html>

research for addressing these complex problems. Recent multi-omics integration methods, such as GLUE [108], scJoint [109], TotalVI [54], UMINT [55], MUON [99], and MATILDA [110], among others, have addressed the issue of dimensionality by generating low-dimensional latent embeddings useful for downstream analyses. However, these DL methods that have been used in multi-omics data analysis are often unable to generate biologically significant information. Researchers have recently highlighted the growing relevance of biologically informed neural networks (BINNs) [111, 112] as a potential solution to this challenge [113]. Thus, feature selection and dimensionality reduction techniques play a crucial role in identifying the most informative biomarkers to aid patient prognosis, disease understanding, and future treatment strategies.

With the recent developments in computational pathology and deep learning methods, the scope of survival analysis has further extended beyond the conventional covariates by incorporating prognostic information from histology images [114, 115, 116]. Beside genomic and transcriptomic biomarkers, it has included histopathological biomarkers [117], thus enabling more comprehensive modeling of lung cancer, along with the others. Survival-aware risk scores have also been incorporated to enable the possibility of data-driven biomarker discovery for lung squamous cell carcinoma and other subtypes. The recognition of the heterogeneity of lung cancer has also underscored the need for robust survival modeling, which can enable the generalizability of lung cancer models. Survival analysis has thus emerged as a tool that not only enables prediction but also bridges the gap between lung cancer, molecular biomarkers, and treatment strategies. However, despite enhanced predictive accuracy, challenges persist in terms of model interpretability, cohort generalizability, and biological validation of predicted biomarkers, underscoring the need for integrative strategies that combine survival modeling with pathway-level and functional analyses.

In this study, we have developed NeuroMDAVIS-FS, a feature selection method to analyze two subtypes of lung cancer, viz., LSCC and LUAD. The model has built upon NeuroMDAVIS [3], which comprises a latent layer, a shared hidden layer, and a modality-specific hidden layer, to project multi-omics data into a common latent space while minimizing reconstruction loss. NeuroMDAVIS-FS examines the weights and biases of the trained NeuroMDAVIS model to rank features according to their importance. The goal of this work is to identify clinically relevant features contributing to lung cancer progression.

In order to demonstrate the effectiveness of NeuroMDAVIS-FS, we have performed a series of experiments using lung cancer datasets from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [118], having Copy Number Variation (CNV), RNA expression, and protein expression data. First, we have trained NeuroMDAVIS to obtain a low-dimensional latent representation of the multi-omics data. We have then applied NeuroMDAVIS-FS to identify the most significant features in each omics layer. Later, the selected features have been assessed using survival analysis through Kaplan–Meier (KM) statistics and univariate CoxPH, segregating patients into low-expression and high-expression groups. Thereupon, a multivariate CoxPH model has been fitted to stratify patients into high-risk and low-risk groups, which display significant differences in survival. Eventually a baseline model has been built using clinical features and compared against the features selected by NeuroMDAVIS-FS. Adding the selected features to the baseline model has improved the concordance index by a significant margin. Thus, this study has offered an advanced framework for identifying clinically relevant features in lung cancer. The overall framework is represented as Figure 5.1.

The organization of the remaining part of this chapter is summarized as follows. Section 5.2 provides detailed methodology of NeuroMDAVIS-FS and the experimental design. Section 5.3 describes the effectiveness of the top molecular features on survival prediction, risk stratification, and shows their importance over clinical features. Section 5.4 provides a brief discussion of the strengths of molecular features biologically, and finally, Section 5.5 concludes the chapter.

5.2 Methodology

In this section, we have discussed the datasets used in the experiments and the proposed methodology for selecting important features from multi-omics data. We then outline the experimental design used to evaluate the effectiveness of the selected features through survival analysis, risk stratification, and comparison with clinical features.

5.2.1 Data acquisition and pre-processing

Proteogenomic datasets for Lung Squamous Cell Carcinoma (LSCC) and Lung Adenocarcinoma (LUAD) have been obtained from CPTAC data portal. Data cohorts include

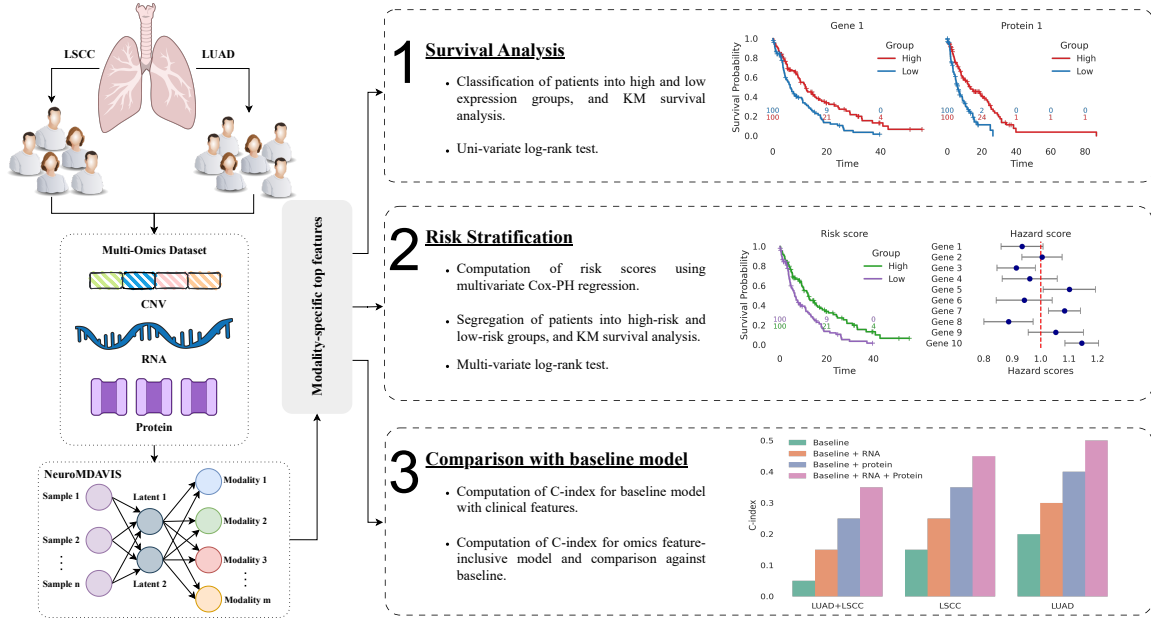


Figure 5.1: Graphical overview of the multi-modal survival analysis framework using modality-specific top features. (1) Survival analysis has been performed by stratifying patients into high and low-expression groups based on individual modality-specific features, followed by KM survival estimation and univariate log-rank testing. (2) Risk stratification has been conducted using multivariate Cox proportional hazards (CoxPH) regression to compute patient-specific risk scores, enabling separation into high and low-risk groups. (3) Model performance has been evaluated by comparing the concordance index (C-index) of a baseline clinical model with omics-augmented models across LSCC, LUAD, and the combined cohort, demonstrating improved prognostic performance with multi-modal integration.

clinical information, genomics, transcriptomics, and proteomics profiles of 205 patients (103 LSCC and 102 LUAD). For each cancer type, raw data files have been downloaded and parsed using the CPTAC Python package (version 1.5.14) [119]. Samples with missing clinical annotations or low-quality omics data have been excluded. Omics data have been log-transformed, normalized, and preprocessed using Python 3. The distribution of demographic and clinical features has been shown in Figure 5.2.

5.2.2 NeuroMDAVIS-FS

NeuroMDAVIS-FS is a feature selection model built upon the NeuroMDAVIS framework, specifically designed to identify the most informative features in multi-modal datasets. Here we have considered multi-omics data to demonstrate its effectiveness. NeuroMDAVIS itself is an unsupervised neural network framework designed for multi-omics data visualization, where the objective is to reconstruct the input omics modalities [3]. We have



Figure 5.2: Event-wise distribution of demographic and clinical variables. Kernel density plots show differences in Age and BMI between Event = 0 (censored) and Event = 1 (death), while bar plots depict event-stratified counts for Sex and Smoking status. Donut charts summarize the overall categorical composition for each variable (age groups, BMI classes, sex, and smoking categories). The left and right panels show the details of LSCC and LUAD cohorts, respectively. ‘Censored’ refers to an event that has not yet occurred within the study’s timeframe.

demonstrated earlier that the feature embedding generated by NeuroMDAVIS, solely obtained through reconstruction, is highly informative. Based on the assumption that highly informative features are reconstructed with greater precision than non-informative ones, NeuroMDAVIS-FS has been developed. Additionally, we know that feature variance strongly influences the reconstruction process, i.e., features with little to no variability are reconstructed more accurately since their values remain consistently aligned across samples. However, these features may not be informative enough. Hence, the objective function has been designed to reinforce the claim that a feature is said to be informative only if it demonstrates sufficient variability and is also effectively reconstructed by NeuroMDAVIS after training. In order to formalize this notion, we have introduced a scoring scheme as explained below.

Let $\mathbf{X} = \{\mathbf{X}^{(r)} : \mathbf{X}^{(r)} \in \mathbb{R}^{n \times d_r}\}_{r=1}^m$ be a dataset containing m different omics modalities. The r^{th} omics modality $\mathbf{X}^{(r)}$ contains n samples and is characterized by d_r features. That is,

$$\mathbf{X}^{(r)} = [x_{ik}^{(r)}]_{n \times d_r}, x_{ik}^{(r)} \in \mathbb{R}$$

where $x_{ik}^{(r)}$ denotes k^{th} feature for i^{th} sample from the r^{th} modality. The *input layer* of NeuroMDAVIS consists of n neurons, which takes an identity matrix as input and generate lossy reconstructions ($\tilde{\mathbf{X}}$) of each omics modality, where

$$\tilde{\mathbf{X}} = \{\tilde{\mathbf{X}}^{(r)} : \tilde{\mathbf{X}}^{(r)} \in \mathbb{R}^{n \times d_r}\}_{r=1}^m$$

Without loss of generality, let us assume a few more notations as follows:

$$\mathbf{x}_i^{(r)} : i^{\text{th}} \text{ sample from the } r^{\text{th}} \text{ omics modality.}$$

$$\tilde{\mathbf{x}}_i^{(r)} : \text{reconstruction of } \mathbf{x}_i^{(r)}$$

$$\mathbf{x}_{:,k}^{(r)} : k^{\text{th}} \text{ feature from the } r^{\text{th}} \text{ omics modality.}$$

$$\tilde{\mathbf{x}}_{:,k}^{(r)} : \text{reconstruction of } \mathbf{x}_{:,k}^{(r)}$$

Initially, the reconstructed features are compared with the original ones for all the omics modalities with a Kullback-Leibler Divergence (KLD) loss as shown in equation (5.1).

$$L_{KLD}(\mathbf{x}_{:k}^{(r)} || \tilde{\mathbf{x}}_{:k}^{(r)}) = \mathbf{x}_{:k}^{(r)} \log \left(\frac{\mathbf{x}_{:k}^{(r)}}{\tilde{\mathbf{x}}_{:k}^{(r)}} \right) \quad (5.1)$$

A close reconstruction of the original feature results in the loss approaching zero. Conversely, as the deviation between the reconstructed and original feature increases, the loss also increases. Formally,

$$\tilde{\mathbf{x}}_{:k}^{(r)} \sim \mathbf{x}_{:k}^{(r)} \quad \Rightarrow \quad L_{KLD}(\mathbf{x}_{:k}^{(r)} || \tilde{\mathbf{x}}_{:k}^{(r)}) \longrightarrow 0,$$

Let $\sigma_{\mathbf{x}_{:k}^{(r)}}$ represent the standard deviation of the k^{th} feature in the r^{th} omics modality. It is well-known that a feature with little to no variability can be reconstructed easily, since these values are close to a constant direction, which makes the feature less informative for downstream tasks. On the other hand, features with higher variance are more likely to capture meaningful biological or statistical differences across samples. Therefore, an informative feature should satisfy two criteria simultaneously:

- It should be reconstructed well by NeuroMDAVIS, indicated by a smaller $L_{KLD}(\mathbf{x}_{:k}^{(r)} || \tilde{\mathbf{x}}_{:k}^{(r)})$, and
- It should exhibit sufficient variability, reflected by a larger $\sigma_{\mathbf{x}_{:k}^{(r)}}$.

Based on these two criteria, we define a feature score that balances the trade-off between reconstruction quality and variability. Mathematically, the combined score function for k^{th} feature in the r^{th} modality is expressed as follows:

$$Score_k^{(r)} = \frac{L_{KLD}(\mathbf{x}_{:k}^{(r)} || \tilde{\mathbf{x}}_{:k}^{(r)})}{\sigma_{\mathbf{x}_{:k}^{(r)}}}; \text{ where } \sigma_{\mathbf{x}_{:k}^{(r)}} \neq 0 \quad (5.2)$$

Equation (5.2) suggests that the lower the combined score, the better is the importance of the feature. Finally, the set of top τ informative features ($IF^{(r)}$) from each omics modality has been selected based on these scores. That is,

$$IF^{(r)} = \arg \min_{P \subseteq \{1, 2, \dots, d_r\}, |P|=\tau} \sum_{k \in P} Score_k^{(r)} \quad (5.3)$$

These top τ features were then further utilized for several downstream analyses.

Table 5.1: Top 15 features selected by NeuroMDAVIS-FS from each of the omics modalities, viz., Protein, RNA, and CNV. Here both LSCC and LUAD samples have been utilized to train NeuroMDAVIS.

Rank	Protein	RNA	CNV
1	SLC19A3	CDY2B	LIMD1
2	TAL1	DAZ4	CXCR6
3	KRT13	DCAF4L2	SLC6A20
4	WSB2	XKR4	CCR9
5	GMCL1	PIWIL2	CCR1
6	SULT1C4	PAGE2	CCR3
7	C22orf39	PAGE5	LTF
8	MCTS2P	PNMA6E	FYCO1
9	FBXL7	PAGE2B	CCRL2
10	NEK11	SLC5A5	XCR1
11	KRT6A	HSD17B13	LZTFL1
12	PPFIA2	GP2	CCR2
13	SMAD6	OBP2B	SACM1L
14	EFCAB1	SLC28A2	LARS2
15	ADPRM	PPP4R3C	CCR5

5.2.3 Experimental Design

In order to evaluate the prognostic significance of molecular features in lung squamous cell carcinoma (LSCC) and lung adenocarcinoma (LUAD), we have developed a multi-step experimental pipeline that has integrated feature selection, survival analysis, risk stratification, and model validation. High-dimensional omics data, including CNV, RNA, and protein expression values, have been utilized for selecting top informative features using NeuroMDAVIS-FS. These top τ features have then been further utilized for these downstream analyses.

Survival Analysis

In this study, KM survival analysis, using time-to-event data, has been considered to estimate survival probabilities, model risk rates, and identify prognostic factors. The selected set of top features, including the ones chosen from each modality, has been used to predict patient survival. Patients have then been stratified into high and low-expression groups based on the median value of each feature, and log-rank tests have been conducted to assess statistical significance in survival differences.

Risk Stratification

Risk stratification is the process of classifying individuals into different risk groups (e.g., high-risk vs. low-risk) based on their predicted likelihood of experiencing an event (e.g., death). We have applied CoxPH, a semi-parametric regression model, on the top-selected features across omics modalities. It has helped us to find the relationship between covariates (e.g., clinical features) and the time until an event occurs. Subsequently, a risk score for each patient has been computed based on the coefficients of the multivariate CoxPH model. The patients have then been categorized into high-risk and low-risk groups using median risk scores, and survival outcomes between these two groups have been studied using KM curves. The hazard ratios and the hazard indices have then been critically assessed to ensure the reliability of the CoxPH model. In order to account for potential covariate interactions and validate robustness of findings across several experiments, we have evaluated statistical significance of survival differences using both univariate and multivariate log-rank tests.

Comparison with baseline model

In order to quantify the incremental prognostic value of the omics-derived signatures, we have established a baseline CoxPH model incorporating standard clinical covariates, including age, sex, BMI, and smoking history. The predictive accuracy of this baseline has then been compared against the multi-omics integrated model by evaluating the improvement in the concordance index (C-index).

5.3 Results

Initially, three distinct modalities of omics data, namely CNV, RNA, and protein expressions, have been fed as input to NeuroMDAVIS which has integrated them to generate a low-dimensional latent embedding. This latent representation has effectively condensed the complex relationships among different omics layers while preserving essential information. Thereafter, to gain deeper biological insights and improve interpretability, NeuroMDAVIS-FS (Feature Selection module) has been applied. Instead of relying solely on the latent embeddings, NeuroMDAVIS-FS has analyzed the weights and biases of the trained NeuroMDAVIS model to identify the most informative features from the

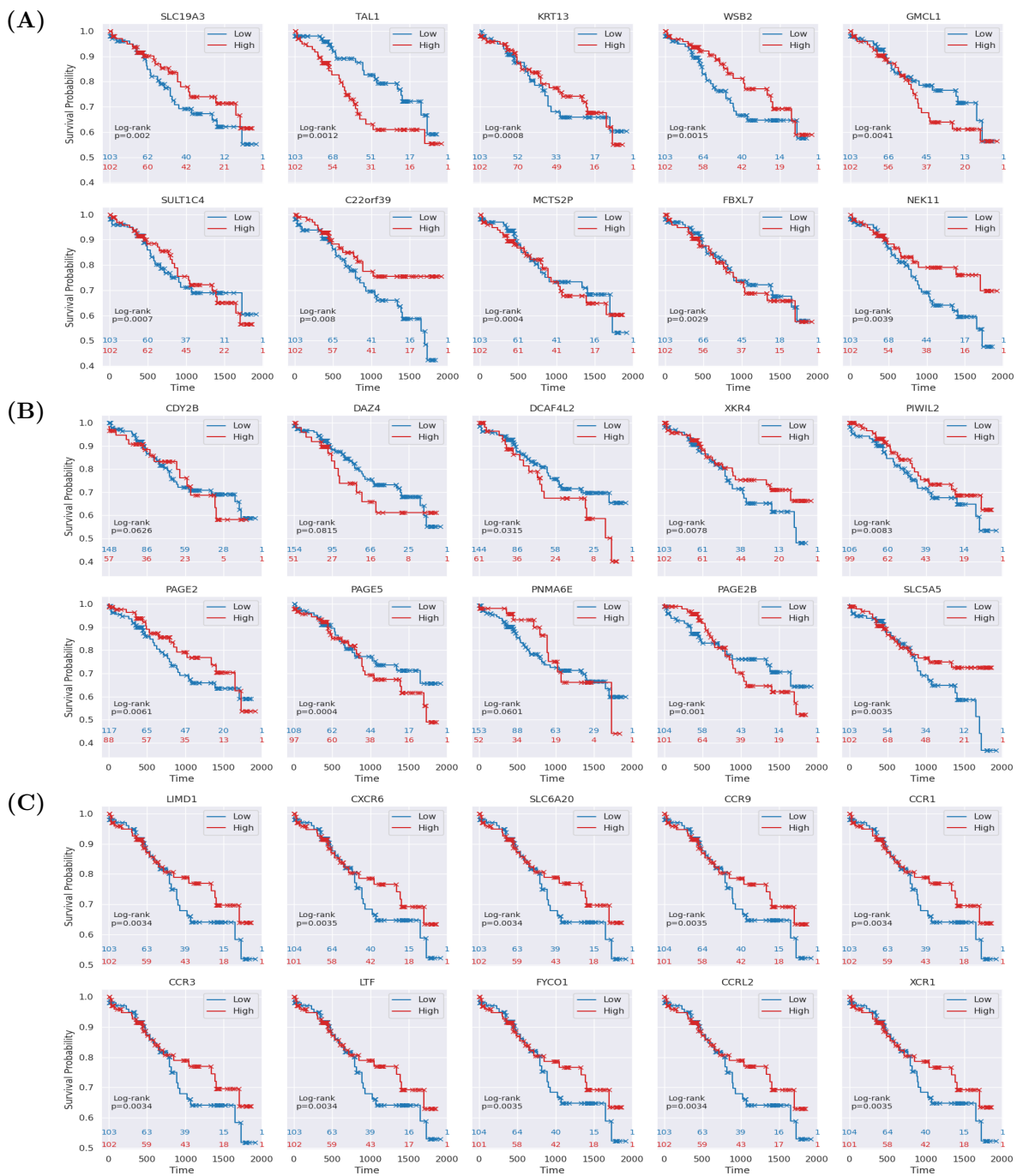


Figure 5.3: Kaplan-Meier curves based on high and low-expression values of each of the top features selected by NeuroMDAVIS-FS from (A) Protein, (B) RNA, and (C) CNV omics modalities, while applied on both the LSCC and LUAD cohort together. Patients have been stratified into high and low-expression groups based on the median value of each feature. Log-rank tests have been conducted to assess statistical significance in survival differences.

original omics data. These selected features have been expected to play a crucial role in characterizing the molecular features under study. In order to assess the impact of the selected features on patient prognosis, KM survival analysis has been carried out. The KM estimator has been utilized to evaluate the effectiveness of the selected features in distinguishing different survival groups. Further analysis has involved risk stratification using CoxPH regression and comparison against a baseline model built using clinical features only.

5.3.1 Feature Selection

To highlight cancer-specific top features, we have considered three scenarios. First, features have been selected from the combined dataset containing both cancer types, as presented in Table 5.1. Next, we have examined each cancer subtype individually, with the top features for LSCC (Table B.1 in Appendices) and those for LUAD (Table B.2 in Appendices). The features ranked by NeuroMDAVIS-FS have been found to be highly informative since they direct major biochemical pathways associated with cancer in general.

5.3.2 Survival Analysis

The KM survival plots (Figure 5.3) illustrate the prognostic value of the top-selected features (genes, CNVs, or proteins) by comparing patients with high and low expression levels. This high-low split has been considered based on the median expression value of that feature.

First, as shown in Figure 5.3(A), some of the top-ranked proteins, such as *SLC19A3*, *TAL1*, *KRT13*, *WSB2*, and *GMCL1*, have shown significant survival differences between the two groups, with log-rank p -value consistently below 0.01. Notably, higher expression levels of these proteins are generally associated with either reduced or improved survival probability, suggesting their potential role in disease progression.

Second, the top features from RNA have shown mixed trends in survival, as shown in Figure 5.3(B). While genes, such as *DCAF4L2*, *XKR4*, *PIWIL2*, *PAGE2*, *PAGE5*, *PAGE2B*, and *SLC5A5*, have demonstrated significant differences in survival (p -value < 0.05), others, like *CDY2B*, *DAZ4*, and *PNMA6E*, do not show significant associations (p -value > 0.05).

Finally, the top CNVs have also followed a similar trend (Figure 5.3(C)), focusing on

immune-related and transporter genes, like *LIMD1*, *CXCR6*, *SLC6A20*, *CCR9*, *CCR1*, *CCR3*, *LTF*, and *FYCO1*, among others. Here too, most of the comparisons have yielded statistically significant results, with log-rank p -values around 0.003. In these cases, patients who have been stratified into the high-expression group often displayed survival advantages compared to the low-expression group, reinforcing the potential importance of these genes in influencing clinical outcomes, possibly through immune regulation or cellular transport pathways.

Overall, we have observed that certain genes (*TAL1*, *GMCL1*, *PAGE5*) are linked to patient survival, with high expression values, predicting worse outcomes, whereas others, like *SLC19A3*, *LIMD1*, and *CXCR6*, have predicted better outcomes.

The above experiment has been conducted considering all samples from both LSCC and LUAD cohorts. The top molecular features from each cohort have been analyzed individually for predicting survival (Figures B.I.1 and B.II.1 in Appendices).

5.3.3 Risk Stratification

In order to evaluate the prognostic potential of the selected RNA features, a multivariate CoxPH model has been utilized. This model has estimated the hazard function by relating gene expression levels to patient survival times. Each feature has contributed a regression coefficient that reflects its impact on the risk of an event. By combining these weighted contributions across all selected features, the CoxPH model has generated a composite risk score for each patient, which has served as the basis for stratifying individuals into high and low-risk groups. KM analysis has revealed a strong separation between high and low-risk groups (Figures 5.4(A) and 5.4(B)). Patients classified as a high-risk group have demonstrated severely reduced survival probabilities over time, whereas those in the low-risk group have consistently exhibited more favorable outcomes. This observation has emphasized the prognostic value of the selected features for predicting patient survival and guiding risk-based clinical decision-making.

The hazard ratio (HR) analysis of these top RNA features (Figure 5.4(C)) has further supported this finding. Several genes, including *CDY2B*, *DAZ4*, and *DCAF4L2*, have displayed HR values above 1.0, suggesting an association with increased risk, while others, such as *XKR4* and *PIWIL2*, have HR values slightly below 1.0, uncorrelated with risk. The majority of RNA features have demonstrated wide confidence intervals, reflecting

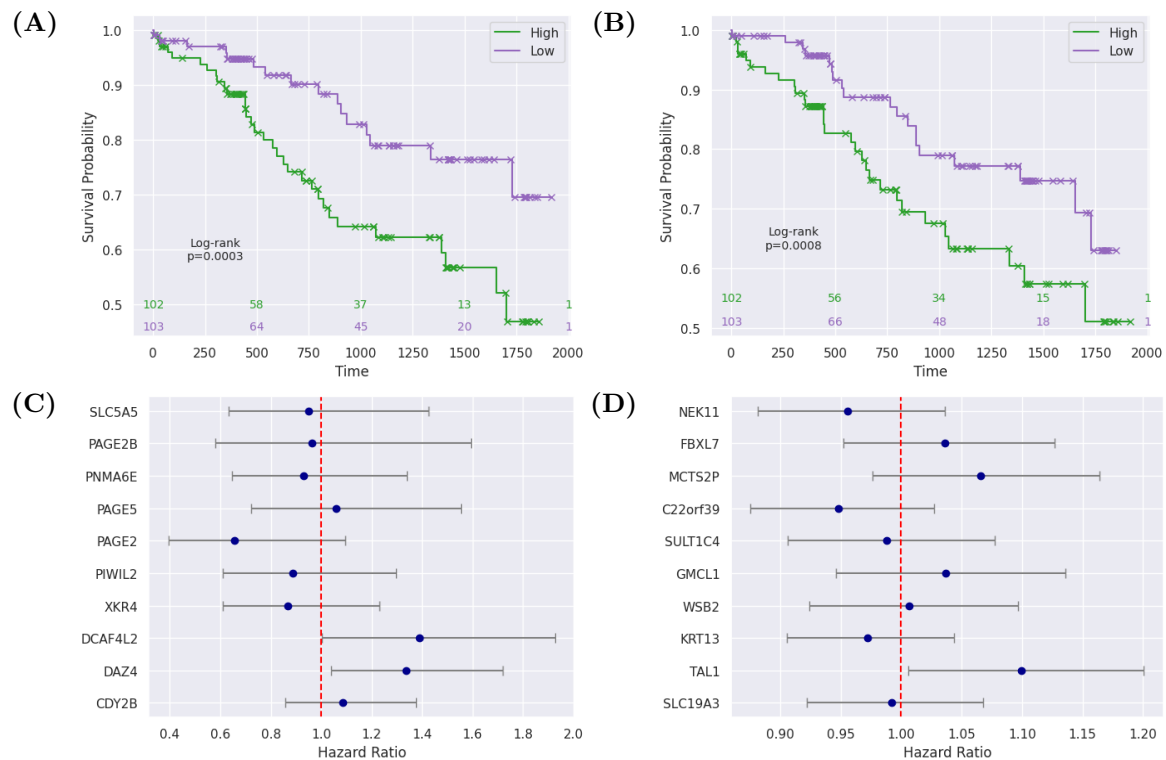


Figure 5.4: KM curves based on high and low risk scores have been obtained from a CoxPH regression applied using (A) top RNAs and (B) top proteins, selected by NeuroMDAVIS-FS. Hazard ratios of (C) top RNAs and (D) top proteins, selected by NeuroMDAVIS-FS. These results have been obtained by utilizing both the LSCC and LUAD cohorts together.

some variability, but the overall trend indicates that these features can distinguish between higher and lower risks of poor survival outcomes.

Similarly, when evaluating the top protein features (Figure 5.4D), high-risk patients have again exhibited a steeper decline in survival compared to low-risk patients. The hazard ratio plot has showed that most of the protein features clustered close to 1.0, with *TAL1*, *FBXL7*, and *MCTS2P* slightly above 1.0, indicating potential risk associations, while *NEK11* and *C22orf39* have HR values below 1.0.

These findings together have demonstrated that both RNA and protein features selected by NeuroMDAVIS-FS are potentially good survival predictors. The combination of Kaplan–Meier survival analysis and hazard ratio evaluation has provided strong evidence of their prognostic power, supporting their potential use for patient stratification and biomarker design. Moreover, the effect of top molecular features, obtained from individual analysis of LSCC and LUAD, has also shown a significant effect on patient survival (Figures B.I.2 and B.II.2 in Appendices, respectively).

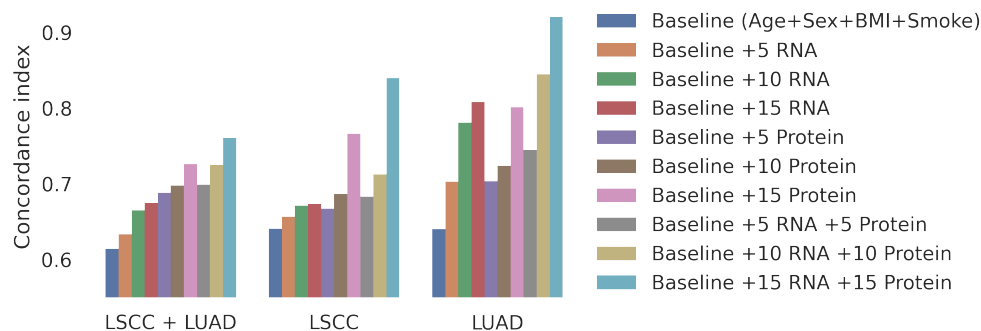


Figure 5.5: Comparison of predictive performance has been measured by the concordance index (C-index) across different feature configurations and cohorts. Results have shown for a baseline clinical model involving age, sex, BMI, and smoking status, as well as for models incrementally augmented with RNA features, protein features, and their joint combinations. Evaluations have been reported for the combined LSCC-LUAD cohort, and for the individual LSCC and LUAD cohorts.

5.3.4 Comparison with baseline

Figure 5.5 illustrates the performance of different survival prediction models across three groups of the dataset. These groups include both LSCC and LUAD combined, as well as LSCC and LUAD individually. Performance has been measured using the concordance index (C-index). The baseline model, which has incorporated only clinical variables (age, sex, BMI, and smoking history), has consistently shown the lowest predictive performance across all three groups, with C-index values around 0.62. Let us now see how predictive performance has improved gradually with incorporation of significant RNA and Protein features.

As stated above, enhancing the baseline model with RNA features has improved the C-index in a stepwise manner. Adding the top 5 RNA features has resulted in a modest improvement, while incorporating the top 10 and 15 RNA features has led to progressively higher concordance indices. A similar pattern has been observed with protein features, although the improvements have appeared slightly more moderate compared to RNA features. Notably, in the LUAD group, RNA features alone have substantially boosted the model's predictive ability, with the C-index value surpassing 0.8 when 15 features have been included. It has highlighted the strong prognostic relevance of RNA features in this cancer subtype.

The combined model integrating both RNA and protein features alongside clinical data has achieved the highest predictive performance across all groups. In particular, for

LUAD, the inclusion of 15 combined molecular features has resulted in a C-index above 0.9, demonstrating excellent survival prediction accuracy. This trend has indicated that integrating multi-omics features on top of clinical features has provided complementary information which have synergistically enhanced the model performance significantly. In contrast, LSCC and the combined LSCC-LUAD group have shown more moderate improvements, suggesting potential differences in molecular heterogeneity between these cancer types. Thus, we may conclude that while clinical features offer some predictive power, they are insufficient for accurately stratifying patient outcomes on their own. Overall, these results have emphasized the importance of incorporating molecular biomarkers, especially in combination, to improve survival prediction beyond traditional clinical factors.

5.4 Discussion

The identification of the aforesaid molecular features in the NeuroMDAVIS-FS framework underscores the model’s ability to pinpoint biologically significant markers from high-dimensional multi-omics data. By integrating genomic (CNV), transcriptomic (RNA), and proteomic (protein) data, the model has captured a holistic view of the tumor landscape that traditional clinical models overlook. In this study, the clinical utility of the NeuroMDAVIS-FS framework has been validated through its robust ability to stratify patient cohorts into distinct risk groups. By employing a multivariate CoxPH model, we have successfully categorized patients into high and low-risk subpopulations characterized by significant survival disparities (Kaplan Meier p -value < 0.001). Central to the value of this deep learning-based approach is its additive prognostic power. When the identified multi-omics features have been integrated with a clinical baseline model, we have observed a substantial increase in the concordance index—an improvement by 43.79% in LUAD, 31.05% in LSCC, and 23.76% across the combined lung cancer cohort. These results suggest that the features selected by NeuroMDAVIS-FS have captured critical molecular drivers of mortality, which are not accounted for by standard clinical variables, highlighting the limitations of relying solely on traditional staging and the necessity of multi-omics integration in precision oncology. Interestingly, we have also observed that some of the molecular hits obtained by the model are established tumor subtype-specific

drivers, tumor suppressors associated with metastasis, immune microenvironment, and therapeutic resistance, while others can be considered novel or emerging. These novel hits will need future orthogonal validation studies to establish their mechanistic role as potential biomarkers or drug targets given their strong association with prognosis in the present study. Further investigations have been performed to assess the functional and clinical associations of the top molecular features obtained from each analysis, as shown in Table B.3 in Appendices.

5.4.1 Subtype-Specific Drivers and Tumor Suppressors

In the LUAD and the combined analysis, the selection of *LIMD1* and *VIPR1* in the CNV modality aligns with their established role as critical tumor suppressors [120, 121]. *LIMD1*, located at the 3p21.3 locus, is a well-documented suppressor whose loss is a hallmark of lung adenocarcinoma progression [121]. Similarly, the model has identified *PTPRG* (protein tyrosine phosphatase gamma) in LSCC, a known tumor suppressor on 3p14.2, frequently inactivated in squamous cell carcinomas via loss of heterozygosity (LOH) [122]. These findings have demonstrated that the model has effectively prioritized genomic regions with high functional impact on survival.

5.4.2 Metastasis and the Immune Microenvironment

The model's ability to select chemokine receptors, including *CCR9*, *CCR1*, *CCR3*, and *CXCR6* from combined analysis, have suggested that it has captured essential components of the tumor microenvironment (TME) [123, 124, 125, 126]. *CCR9* is a known driver of *PI3K*/Akt-mediated metastasis [123], while *CXCR6* and LTF (Lactotransferrin) are linked to immunotherapy response and immune cell infiltration [126, 127]. The identification of *ACKR2* from the LUAD analysis, an atypical chemokine receptor that scavenges inflammatory ligands to regulate TME, further supports the model's ability to identify factors governing the interface between the tumor and the host immune system [128].

5.4.3 Therapeutic Resistance and Prognostic Gain

Beyond tumor biology, NeuroMDAVIS-FS has identified markers of therapeutic resistance, such as *ABCC2* (*MRP2*), in the LUAD analysis [129]. Overexpression of this multi-drug resistance transporter is a verified predictor of cisplatin resistance in NSCLC patients, offering a biological explanation for the poor prognosis in certain identified clusters. Many of these genes or proteins, such as the chemokine receptors *CCR1* and *CCR3*, are currently being explored as therapeutic targets [124, 125]. It has suggested that the high-risk subpopulations, identified by NeuroMDAVIS-FS, could potentially be stratified for targeted chemokine receptor inhibitors or personalized immunotherapy regimens in future clinical trials.

5.5 Conclusion

This study has demonstrated the utility of a deep learning framework for multi-omics integration for feature selection using the newly developed model, called NeuroMDAVIS-FS, which has been designed to decode the molecular complexities of lung cancer. By leveraging synchronized encoding and modality-specific decoding layers, the model maps disparate data streams, including copy number variations, transcriptomes, and proteomes, into a common latent projection that preserves critical biological signals.

This approach aligns with the emerging paradigm of Biologically Informed Neural Networks (BINNs), rather than treating the tumor as a black box. NeuroMDAVIS-FS bridges the gap between raw high-dimensional data and biological interpretability. Validation using the CPTAC lung cancer dataset confirms that the framework effectively prioritizes high-impact features, such as *LIMD1*, *ABCC2*, and *CCR9*, which are deeply rooted in lung cancer pathogenesis, therapeutic resistance, and metastatic potential. By focusing on these biologically significant nodes, the model provides a transparent mechanism for precision oncology, identifying ‘vulnerability hubs’ that enable improved patient stratification compared with traditional clinical models. This is evidenced by a concordance index improvement of up to 31.05% in LSCC and 43.79% in LUAD. While NeuroMDAVIS-FS has effectively captured key ‘vulnerability hubs’, the current architecture is yet to embed explicitly higher-order biological priors, such as pathway hierarchies or protein-protein interaction networks, which could further refine the model’s

accuracy and biological interpretability in a precision oncology context.

Future work will focus on expanding this framework to larger and more diverse cohorts, validating the findings across multiple datasets to ensure robustness and generalizability. Such studies will focus on explicitly embedding biological priors, such as pathway hierarchies and protein-protein interaction networks, directly into the architecture to further refine the model’s accuracy. Additionally, incorporating attention mechanisms or graph-based learning may further enhance the interpretability and accuracy of multi-omics integration.

By transforming complex multi-omics landscapes into actionable prognostic insights, this framework has provided a scalable foundation for precision medicine. It enables the discovery of robust biomarkers that may steer development of targeted therapy and the design of personalized immunotherapy regimens. Ultimately, NeuroMDAVIS-FS offers an advanced, interpretable deep learning paradigm that paves the way for more precise prognostic tools and improved clinical outcomes in lung cancer research.

So far in Chapter 2, a visualization method has been discussed. This model has been enhanced to introduce a feature selection model in Chapter 3. A joint visualization method has been described in Chapter 4, and finally, this chapter reflects how the selected top multi-modal molecular features have outperformed the clinical features to predict patient survival. One major drawback of embeddings produced by the earlier methods is their lack of generalization, which manifests as a shortfall in sample generation from those embeddings. The generalization capability of an embedding refers to compact, meaningful, and lower-dimensional representations of data, which extends beyond the training set and facilitates generating meaningful diverse samples. Motivated by this issue, we have developed G-NeuroDAVIS in Chapter 6, which can visualize high-dimensional datasets through generalized embedding. It can also produce samples in both conditional (supervised) and unconditional (unsupervised) scenarios and can handle high-dimensional complex datasets.

Chapter 6

G-NeuroDAVIS: A Generative Model for Data Visualization through a Generalized Embedding

6.1 Introduction

The earlier chapters have presented single and multi modal data visualization methods and have also discussed the feature selection models, that works on the visualizations achieved. This chapter develops a feed-forward neural network model, called G-NeuroDAVIS, for visualization of high-dimensional datasets through generalized embedding. The model not only reveals the hidden patterns but is also able to generate realistic high-dimensional samples. G-NeuroDAVIS takes an identity matrix as input and reconstructs the original data through a few hidden layers. During this reconstruction, it learns a close approximation of the actual data distribution using suitable parameters in the latent space. Once the latent distribution is learned, several samples in the high-dimensional space are generated by sampling from this distribution. The effectiveness of this generation process demonstrates the generalizability of the latent space, implying that it can distinguish samples based on some specific attributes. A generalized latent space allows a smooth transition from one cluster to another, thus reflecting an effective interpolation capability of the same. Although G-NeuroDAVIS can generate high-dimensional samples successfully, it lacks control over the output as each generated sample is independent and not influenced by any specific conditions or prior generations. As a result, the model cannot be controlled

to produce samples with desired properties or characteristics to a particular class, limiting its applicability in scenarios where conditional generation is required. In order to tackle this issue, we have modified the input using the label information and learned a conditional distribution in the latent space in a supervised fashion. For incorporating the label information, instead of using an Identity matrix, a one-hot encoder matrix (order being dependent on the number of samples and classes) is considered. This input ensures that all the samples belonging to a particular class are projected to the latent space through specific weight parameters. This supervised model is efficient enough to generate samples based on a specific condition. Thus, G-NeuroDAVIS is a model that can visualize data through a generalized embedding, produce samples in both conditional (supervised) and unconditional (unsupervised) scenarios, and can handle high-dimensional complex datasets.

For performance evaluation, the embedding quality, generalization capability, and classification performance of G-NeuroDAVIS have been compared against VAE, with similar capabilities. The interpolation ability of the latent space, generated by G-NeuroDAVIS, has been demonstrated via different interpolating paths. Performance of G-NeuroDAVIS has also been benchmarked against DCGAN, DDPM, and AE-guided RealNVP. The evaluation, based on improved precision and improved recall [130], has demonstrated that the developed approach has produced more diverse and authentic samples, highlighting its effectiveness in high-quality generative tasks. In addition, a computational complexity analysis, in terms of training time, inference speed, and memory usage, has been compared against the aforesaid methods. Finally, the performance of conditional data generation has been described. Here, G-NeuroDAVIS has shown comparable performance in both the tasks of data visualization and sample generation against the state-of-the-art methods.

The organization of the remaining part of this chapter is summarized as follows. Section 6.2 provides a detailed design of G-NeuroDAVIS. Section 6.3 describes the experimental results, along with comparisons on several image datasets in terms of generalized embedding, interpolation capability, downstream analyses, sample generation capability, sub-sample based analyses, complexity analyses, and finally, conditional sample generation performance. Section 6.4 provides a brief discussion of the strengths and weaknesses of G-NeuroDAVIS, and concludes the chapter.

6.2 Methodology

In this study, we have developed a generative neural network model, called G-NeuroDAVIS, which is not only useful for visualizing high-dimensional datasets but also effective for generating new samples. This section describes the problem scenario and motivation behind the model, followed by its architecture, forward propagation, and learning.

6.2.1 Problem scenario and solution approach

Here, we narrate the problem scenario and a possible solution approach.

Problem scenario

The availability of high-dimensional datasets is growing rapidly due to the advancement of technology, but it becomes challenging to visualize and extract information effectively [131]. If the number of features increases, the volume of the feature space grows exponentially, which leads to the problem of data sparsity, and the distances among the samples become less informative. This is one of the major issues of the curse of dimensionality, which increases the computational complexity and model overfitting. In this context, managing high-dimensional datasets requires specialized techniques for dimensionality reduction and visualization to reveal meaningful patterns, relationships, and insights.

Analogously, the other important aspect of machine learning is sample generation, which involves creating synthetic data for effective training of the model, especially in scenarios where data is sparse, imbalanced, or missing [132, 133]. For instance, in anomaly detection, generating realistic anomalous samples helps improve a model's ability to identify rare or outlier events. Similarly, in data imputation, generated samples fill in missing data points, enabling models to perform better in cases of incomplete data. In class-imbalance problems, generating synthetic samples from underrepresented classes helps prevent models from being biased toward the dominant class [132, 133]. However, generating such samples is particularly challenging due to the complex and non-linear associations in high-dimensional feature spaces. High-dimensional data often exhibits intricate patterns, where features interact in non-obvious ways, making it difficult to replicate these relationships in synthetic samples.

In this context, learning a generalized latent distribution is crucial in generative

models because it allows for a structured, compact representation of complex data. This approach reduces the dimensionality of the data, facilitating efficient computation and meaningful sample generation. By mapping high-dimensional data into a lower-dimensional latent space, the model captures essential features and underlying structures, making it possible to generate new, diverse, and coherent samples by sampling from this latent distribution. Additionally, a generalized latent space supports smooth interpolation among data points, provides a regularizing effect to prevent overfitting, and offers interpretability for controlling and manipulating generated outputs. Overall, learning a generalized latent distribution enhances the model's ability to generate high-quality, varied, and realistic data while maintaining generalization and adaptability for various applications.

Moreover, an additional control on the sample generation process makes the task of generation even more useful. Sampling from the latent embedding can produce a random sample that is close enough to a real one. However, conditional data generation significantly enhances the sample generation process by incorporating additional contextual information or con-

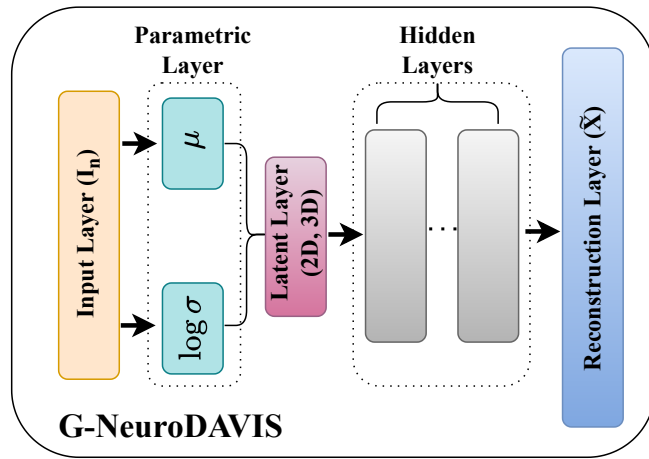


Figure 6.1: The network architectures of G-NeuroDAVIS for visualization of high-dimensional datasets.

straints into the model. This approach enables the generation of samples that meet specific requirements or attributes, leading to more targeted, relevant, and high-quality outputs.

The main objective of the chapter is to develop a novel model that not only visualizes the data in lower dimensions but also generates samples from the original high-dimensional space simultaneously by learning the inherent distribution. In addition to that, the model can generate customized samples according to the targeted scenarios, which is one of the key goals of this study.

Solution approach

Motivated by the earlier model, viz., NeuroDAVIS [1], the developed architecture of G-NeuroDAVIS can be effectively utilized for both data visualization in lower dimensions, along with the generation of samples in high dimensions. In the structure of the developed model, assuming that the latent space follows a multivariate Gaussian, and the corresponding parameters, i.e., mean and variance, have been learned accordingly. In this model, an identity matrix of order n is fed as input, which controls the parameters of the Gaussian. Using these parameters, a sample has been drawn using the reparameterization trick [60]. Once a sample is generated in the low-dimensional space, it is projected into the original space through a few hidden layers to create a replica of the original data sample. In order to achieve two major targets, i.e., sample generation and generalization of the latent space, two loss functions have been considered here. The first one is a reconstruction loss or mean squared error loss, which ensures that the projections or generations become realistic as well as close to the accurate value. On the other hand, KL-Divergence (KLD) loss has been calculated to ensure the generalization capability of the latent space. The KLD loss makes the latent distribution close to a standard normal distribution, so that the latent space may not remain sparse. Using these two loss functions, G-NeuroDAVIS has been trained on different datasets. Moreover, instead of feeding an identity matrix as input, a one-hot class matrix has been considered to control the sample generation process, so that one can customize the generated samples based on some pre-defined conditions. The detailed methodology for the development of G-NeuroDAVIS is provided in the following section.

6.2.2 G-NeuroDAVIS: Architecture, Forward Propagation and Learning

This subsection deals with the development of the architecture, forward propagation, and learning of G-NeuroDAVIS model.

Architecture

The architecture (Figure 6.1 and 6.2) of G-NeuroDAVIS is a non-recurrent feed-forward neural network. The model can effectively extract a low-dimensional embedding that can

efficiently capture the informative features for visualizing the high-dimensional datasets. G-NeuroDAVIS architecture contains five different types of layers, viz., an *Input layer*, a *Parametric layer*, a *Latent layer*, one or more *Hidden layer(s)*, and a *Reconstruction layer*.

Let $\mathbf{X} = [x_{ik}]_{n \times d}$, $x_{ik} \in \mathbb{R}$ be a data matrix comprising n samples characterized by d features each. The *Input layer* of G-NeuroDAVIS consists of n neurons/nodes. The *Parametric layer* has two sub-parts, each having κ nodes, where κ is the number of dimensions of the projected space to be used for visualization. For the purpose of visualization, one should use $\kappa = 2$ or 3 . The *Latent layer* also contains κ nodes. The number of nodes in the *Hidden layer(s)* depends on the dataset that needs to be visualized, and in this work, it has been set empirically. Moreover, the *Reconstruction layer* has d nodes, i.e., the same as the dimension of the original feature space. It has been noted that all the layers are densely connected, i.e., each node in any of these layers is connected to all the nodes of its adjacent layer(s).

The *Input layer* has been utilized to embed the original n samples/observations independently into the lower-dimensional space at the *Latent layer*, and in order to achieve the same, a *Parametric layer* has been considered. The *Parametric layer* produces the mean and variance as parameters, which have been used further to define the distribution in the latent space. In the next stage, one (or more) *Hidden layer(s)* has (have) been used to project the lower-dimensional embedding at the *Latent layer* onto the *Reconstruction layer*. Here, only two hidden layers have been taken into consideration. A *Hidden layer* has been used to capture the essence of the non-linearity present in the dataset and carry the information to its next adjacent layer. In this context, multiple *Hidden layers* can be used based on the level of complexity embedded in the dataset. However, the number of *Hidden layers* should not be too large as it tends to overfit the data. Consequently, the *Reconstruction layer* tries to recreate the given n samples in the d -dimensional space using a random low-dimensional embedding. It may be mentioned here that it is the *Parametric layer* which makes the architecture stand out from the earlier visualization model, viz., NeuroDAVIS [1], developed by the authors. Usage of a *Parametric layer* ensures a latent space with an impressive generalization capability.

Forward propagation

As stated above, the input of G-NeuroDAVIS is an identity matrix $\mathbf{I}_n$ for a dataset consisting of n samples. An i^{th} column vector $\mathbf{e}_i$ of $\mathbf{I}_n$ is fed to the input layer to create an approximate version of the i^{th} sample $\mathbf{x}_i$ as $\tilde{\mathbf{x}}_i$, at the *Reconstruction layer*. Let $\mathbf{a}_i^{(Inp)}$ and $\mathbf{h}_i^{(Inp)}$ represent the input and output of the *Input layer*, respectively, for i^{th} sample. Then

$$\begin{cases} \mathbf{a}_i^{(Inp)} = \mathbf{e}_i, \\ \mathbf{h}_i^{(Inp)} = \mathbf{a}_i^{(Inp)}, \end{cases} \quad \forall i \in \{1, 2, \dots, n\} \quad (6.1)$$

Here, $\mathbf{e}_i$ controls the weight parameters for the lower-dimensional projection of the i^{th} observation at the *Latent layer*. Thus, the links connected to the i^{th} neuron of the *Input layer* can only activate the neurons in *Parametric layer* on presentation of the i^{th} sample.

As mentioned earlier, the *Input layer* is densely connected with the *Parametric layer*, which has two sub-parts. The first part calculates the mean (μ) and the second part calculates the log-variance ($\log \sigma^2$) of the distribution assumed at the latent space (specifically a multivariate Gaussian). Let $\mathbf{W}^{(\mu)}$ and $\mathbf{b}^{(\mu)}$ represent the weights and bias, respectively, between the *Input layer* and the first part of the *Parametric layer*. Similarly, $\mathbf{W}^{(\sigma)}$ and $\mathbf{b}^{(\sigma)}$ represent the same between the *Input layer* and the second part of the *Parametric layer*. Assuming that for an i^{th} sample, $\mathbf{a}_i^{(\mu)}$ and $\mathbf{a}_i^{(\sigma)}$, respectively, represent the input to the sub-parts of the *parametric layer* before applying activation, whereas $\mathbf{h}_i^{(\mu)}$ and $\mathbf{h}_i^{(\sigma)}$, respectively, stand for the output of the same after activation. The linear activation has been used in both the sub-parts of the *Parametric layer*. Hence, on presentation of an i^{th} sample in the *Parametric layer*, we have,

$$\begin{cases} \mathbf{a}_i^{(\mu)} = \mathbf{W}^{(\mu)} \mathbf{e}_i + \mathbf{b}^{(\mu)}, & \mathbf{a}_i^{(\sigma)} = \mathbf{W}^{(\sigma)} \mathbf{e}_i + \mathbf{b}^{(\sigma)}, \\ \mathbf{h}_i^{(\mu)} = \mathbf{a}_i^{(\mu)} & \mathbf{h}_i^{(\sigma)} = \mathbf{a}_i^{(\sigma)}, \end{cases} \quad \forall i \in \{1, 2, \dots, n\} \quad (6.2)$$

As a result, the *Parametric layer* takes the parameters of the multivariate Gaussian and produces a sample from it at the *Latent layer*. However, the samples drawn directly from a distribution disrupt the training via gradient-based optimization. For this reason, the reparametrization trick has been used, which can separate the sampling process from the parameterization of the distribution [60]. Instead of directly sampling from the distribution, sample selection from a standard Gaussian distribution was done at

first, and then transformed the same to match the desired distribution parameters. This transformation is differentiable and allows the gradients to flow through the sampling process. Let $\mathbf{a}_i^{(Lat)}$ and $\mathbf{h}_i^{(Lat)}$ correspond to the input to and output from the *Latent layer* on presentation of an i^{th} sample. Hence, for the *Latent layer*, we have

$$\begin{cases} \mathbf{a}_i^{(Lat)} = \mathbf{h}_i^{(\mu)} + \exp\left(\frac{\mathbf{h}_i^{(\sigma)}}{2}\right) \odot \epsilon_i, \\ \mathbf{h}_i^{(Lat)} = \mathbf{a}_i^{(Lat)} \quad \forall i \in \{1, 2, \dots, n\} \end{cases} \quad (6.3)$$

Here, $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\odot$ represents element-wise multiplication. This sample $\mathbf{h}_i^{(Lat)} \left(\sim \mathcal{N}\left(\mathbf{h}_i^{(\mu)}, \exp\left(\frac{\mathbf{h}_i^{(\sigma)}}{2}\right)\right) \right)$, produced at the latent layer, is used to reconstruct the i^{th} sample $\mathbf{x}_i$. For this purpose, we have used *Hidden layer(s)* and *Reconstruction layer*. The sample generated from the *Latent layer* is fed to the *Hidden layer(s)* with a non-linear activation. Let $\mathbf{a}_i^{(Hid_j)}$ and $\mathbf{h}_i^{(Hid_j)}$, respectively, represent the input to and output from a j^{th} *Hidden layer* for the i^{th} sample input. Thus, for the first *Hidden layer*, we have

$$\begin{cases} \mathbf{a}_i^{(Hid_1)} = \mathbf{W}^{(Hid_1)} \mathbf{h}_i^{(Lat)} + \mathbf{b}^{(Hid_1)} \\ \mathbf{h}_i^{(Hid_1)} = \text{ReLU}\left(\mathbf{a}_i^{(Hid_1)}\right), \quad \forall i \in \{1, 2, \dots, n\} \end{cases} \quad (6.4)$$

and for the $(l - 1)$ *Hidden layer(s)*, we have

$$\begin{cases} \mathbf{a}_i^{(Hid_j)} = \mathbf{W}^{(Hid_j)} \mathbf{h}_i^{(Hid_{j-1})} + \mathbf{b}^{(Hid_j)} \\ \mathbf{h}_i^{(Hid_j)} = \text{ReLU}\left(\mathbf{a}_i^{(Hid_j)}\right), \quad \forall i \in \{1, 2, \dots, n\}; \forall j \in \{2, 3, \dots, l\} \end{cases} \quad (6.5)$$

where $\text{ReLU}(\mathbf{y}) = \max(\mathbf{0}, \mathbf{y})$; element wise $\max$ (maximum) value has been considered. $\mathbf{W}^{(Hid_j)}$ and $\mathbf{b}^{(Hid_j)}$ correspond to the weight and bias connected to the j^{th} *Hidden layer(s)*. In this work, only two hidden layers have been considered.

A reconstruction of the original data is performed at the final layer, called *Reconstruction layer*. For the *Reconstruction layer*, we have

$$\begin{cases} \mathbf{a}_i^{(Rec)} = \mathbf{W}^{(Rec)} \mathbf{h}_i^{(Hid_l)} + \mathbf{b}^{(Rec)}, \\ \mathbf{h}_i^{(Rec)} = \mathbf{a}_i^{(Rec)}, \quad \forall i \in \{1, 2, \dots, n\} \end{cases} \quad (6.6)$$

Here $\mathbf{W}^{(Rec)}$ and $\mathbf{b}^{(Rec)}$ correspond to the weight and bias associated with the *Reconstruction layer*. Thus, G-NeuroDAVIS projects the latent embedding for a sample obtained at

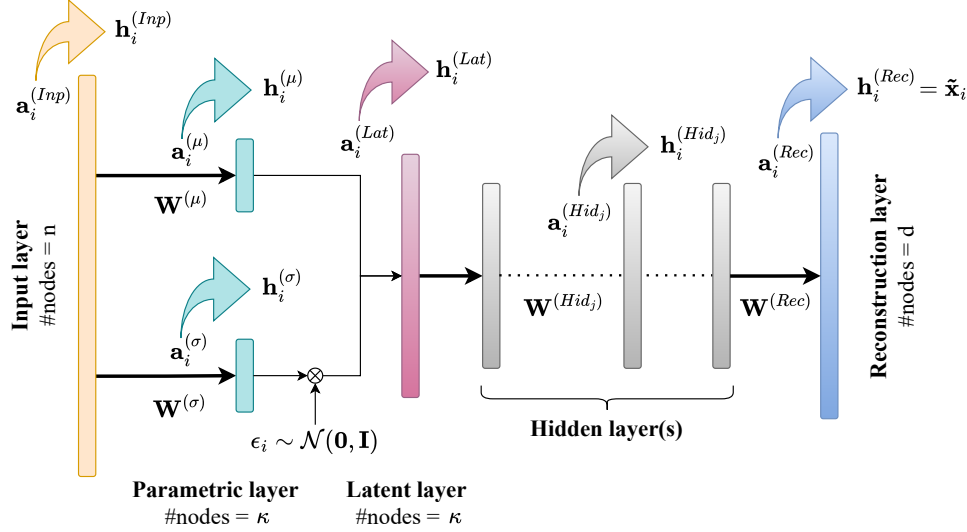


Figure 6.2: The network architecture of G-NeuroDAVIS, along with the mathematical notations.

the *Latent layer* into a d dimensional space through the *Hidden layer(s)*.

Learning

As mentioned earlier, G-NeuroDAVIS has been developed for dimensionality reduction as well as visualization of high-dimensional data. For each sample $\mathbf{x}_i$, ($i \in \{1, 2, \dots, n\}$), G-NeuroDAVIS tries to find an optimal reconstruction $\tilde{\mathbf{x}}_i$ of the input data by minimizing the reconstruction error $\|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|$. In addition, a KL Divergence (KLD) loss has been computed at the latent layer. KLD loss is defined as $p(x) \log \left(\frac{p(x)}{q(x)} \right)$, where $p(x)$ and $q(x)$ are two distributions, and this loss gets minimized when $p(x)$ and $q(x)$ are identical. This regularizes the approximate posterior distribution to match a prior distribution (assumed to be standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$ in this case) over the latent space. It helps to follow a simple distribution, aiding generalization and facilitating interpolation in the learned latent space. This Gaussian assumption of both the prior and posterior makes the loss function tractable, and simplifies the process of backpropagation. Moreover, an $L2$ regularization may be used on the nodes' activities and edge weights to avoid over-fitting. The use of $L2$ regularization ensures the minimization of model complexity as well as the

prevention of the weights from exploding. Hence, the complete objective function is

$$\mathcal{L}_{G-NeuroDAVIS} = \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|_2^2 + \beta \sum_{i=1}^n KLD\left(\mathcal{N}(\mathbf{h}_i^{(\mu)}, \exp(\frac{\mathbf{h}_i^{(\sigma)}}{2})) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})\right) + \lambda_1 \sum_{j=1}^l \|\mathbf{W}^{(Hid_j)}\|_F^2 + \lambda_2 \sum_{j=1}^l \|\mathbf{h}_i^{(Hid_j)}\|_2^2, \quad (6.7)$$

where β is a balancing parameter, and λ_1 and λ_2 are regularization parameters; $\|\cdot\|_2$ denotes the Euclidean norm of a vector, and $\|\cdot\|_F$ indicates the Frobenius norm of a matrix. The balancing parameter β helps to control the loss in a way such that both the objectives (i.e., proper reconstruction of samples and keeping the latent distribution close to a standard normal distribution) can be optimized simultaneously. Here, Adam optimizer has been used to minimize $\mathcal{L}_{G-NeuroDAVIS}$ in equation (6.7) [65]. Moreover, the number of epochs has been set empirically by observing saturation level of $\mathcal{L}_{G-NeuroDAVIS}$. It may be mentioned here that incorporation of the KLD loss prevents the latent space from exploding, and makes it close to a generalized embedding, which is a limitation of NeuroDAVIS [1].

At the initial stage of forward propagation, the weight values between the *Input* and *Parametric layer* have been fixed randomly. The learning of G-NeuroDAVIS is controlled by an n^{th} order identity matrix ($\mathbf{I}_n$), which has been fed as input to the *Input layer*. The advantage of using the identity matrix confirms the necessary modifications in the weight values that are exclusively associated with the samples in the present batch. In the beginning, random parameters of the lower-dimensional Gaussian for n observations have been set at the *Parametric layer*. Using these parameters in the *Latent layer*, a latent embedding is produced and projected onto the *Reconstruction layer* by G-NeuroDAVIS through a few *Hidden layers*. A loss is then calculated using Equation (6.7) and backpropagated up to the *Input layer*. Correspondingly, the weight and bias values are updated for a better reconstruction of the samples of the present batch for the next forward pass. The process has been continued until the loss saturates. On completion of the training phase, the transformed feature set, extracted from the *Latent layer*, is considered as the lower-dimensional embedding of the data and is used for visualization for $k = 2$ or 3 .

Once the model is trained, it generates new samples by leveraging the learned param-

eters, mean ($\boldsymbol{\mu}$) and log standard deviation ($\log \boldsymbol{\sigma}$), of the latent distribution. A latent vector $\mathbf{z}$ is sampled using the reparameterization trick as

$$\mathbf{z} = \boldsymbol{\mu} + \exp(\log \boldsymbol{\sigma}) \odot \boldsymbol{\epsilon}; \text{ where } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

This sampled point $\mathbf{z}$ is then passed through the learned transformations of the *Hidden layers*, and finally projected into the high-dimensional space via the *Reconstruction layer*. The resulting output represents a new sample that closely resembles the original data but is not a replica, preserving authenticity. The structured latent space, guided by regularization during training, ensures that generated samples remain realistic while allowing variation (using $\log \boldsymbol{\sigma}$), thereby maintaining both the diversity and authenticity of the output. However, the latent space in NeuroDAVIS has a lot of blank spaces, and thus falls short to generate realistic samples.

6.2.3 Conditional data generation in a supervised setting

Conditional data generation is important because it allows generating data conditioned on specific attributes, contexts, or conditions. This capability adds flexibility and control to generative models, enabling them to produce outputs tailored to particular requirements or constraints. In this work, G-NeuroDAVIS has been trained in an unsupervised manner and thus falls short of conditional data generation. However, additional supervision about the sample types or class (condition) overcomes this situation.

Earlier, in order to train G-NeuroDAVIS, an identity matrix of order n has been fed as an input. For reconstruction of an i^{th} sample, only the weights connected to the i^{th} input node get activated. It ensures that the task of reconstructing all the samples has been performed independently. Thus, an overall distribution of the samples has been learned, ignoring the class information. However, when it comes to generate a new sample, the developed model has no control over any prior conditions. In order to generate samples based on some pre-defined conditions, this input has now been changed to another sparse matrix, which contains the conditional information about the samples. Let n samples be present in the dataset and be distributed in c distinct classes. Now a matrix $\mathcal{I}_{n \times c}$ has been constructed, which is also sparse in nature but contains the conditional information. Each row of this matrix has exactly one non-zero entry based on the class it belongs

to. Unlike the previous case, samples belonging to the α^{th} ($\alpha \in \{1, 2, \dots, c\}$) class have been reconstructed using weights connected to the α^{th} input node. In other words, now the samples are no longer independent; they are coming from the same distribution if they belong to the same class. The number of nodes in the *Input layer* has been changed accordingly, and the other layers have been kept the same as before. This model is again trained on the training data using equation (6.7) as an objective function for a sufficient number of epochs.

After successful training, G-NeuroDAVIS can be used for data generation based on a given condition (class). Suppose one needs to generate a sample that belongs to the α^{th} class, then a vector of dimension c is fed to the *Input layer*, whose α^{th} component is 1, and the rest are zero. This input will control the weights accordingly and create a sample at the *Latent layer*, which belongs to the α^{th} class. Moreover, in order to achieve smooth variations in the latent embedding, one may include one or more hidden layers between the *Input layer* and *Parametric layer*. In this work, no such hidden layers have been used.

From the above discussions, it is evident that G-NeuroDAVIS introduces key differences from NeuroDAVIS in architecture, learning, and utilization. The parametric layer, introduced in G-NeuroDAVIS, tries to learn the latent distribution, which was absent in NeuroDAVIS. In order to keep the visualization compact and to enable accurate data generation, the objective function has also been modified. Moreover, its ability to support both independent and conditional data generation broadens the scope and applicability of G-NeuroDAVIS beyond that of NeuroDAVIS.

6.3 Results

This section describes the details of the datasets used in this study, along with the evaluation and result analyses. Here we have considered VAE, DCGAN, DDPM, and AE-guided RealNVP for comparison with G-NeuroDAVIS on *MNIST*, *FMNIST*, *Coil20*, and *Olivetti Faces* datasets. However, due to limited computational resources, we have restricted ourselves to *MNIST* and *FMNIST* datasets for comparison with DCGAN and DDPM.

Table 6.1: Descriptions of the datasets used for evaluation of G-NeuroDAVIS

Name	Description	#Samples	#Features	#Classes	Source
<i>MNIST</i>	grayscale images of handwritten digits	60,000	784	10	[134]
<i>FMNIST</i>	grayscale images of clothing items	60,000	784	10	[70]
<i>Olivetti Faces</i>	grayscale images of human faces	400	4096	40	[74]
<i>Coil20</i>	grayscale images of objects, captured from various angles	1440	16384	20	[69]

6.3.1 Data description

In this work, four image datasets have been used to demonstrate the effectiveness of G-NeuroDAVIS model. These image datasets are *MNIST* [134], *FMNIST* [70], *Olivetti Faces* [74], and *COIL20* [69]. The details of the datasets have been described in Table 6.1. *MNIST* contains 60k grayscale images of handwritten digits of dimension 28×28 each. *FMNIST* contains 60k grayscale images of clothing of dimension 28×28 each. In both cases, the number of samples is larger than the number of features, which may make the task easier. On the other hand, *Olivetti Faces* contains only 400 samples of 40 different human faces of dimension 64×64 each, and *COIL20* comprises 1440 samples of 20 different objects of dimension 128×128 each. In the latter two cases, the dimension is much larger than the number of samples, which makes the task of visualization even more complex.

Next, a few problem-specific preprocessing steps have been done for different datasets. At first, standard MinMax scaling has been done to normalize all the pixel values and then all the sample images have been flattened to a vector to fit in the model.

6.3.2 Result Analyses

Initially, the embeddings produced by G-NeuroDAVIS, in an unsupervised manner, have been compared against those obtained by VAE, and then, the generalization capability has been demonstrated. The preservation of the underlying data structure in the latent space has been validated using four different interpolating paths through the embedding.

In order to evaluate the efficacy of the embeddings, a classification task has been carried out for both the embeddings. A sub-sample-based experiment has been conducted on both the datasets, viz., *MNIST* and *FMNIST* with 60k samples each. Besides, new samples have been generated by G-NeuroDAVIS, VAE, DCGAN, DDPM, and AE-guided RealNVP, and the corresponding diversity and authenticity have been compared using improved precision and improved recall scores. Moreover, a complexity analysis has been performed against the aforesaid state-of-the-art methods. In addition, the performance of G-NeuroDAVIS has been observed in a supervised manner, in which it has generated samples based on a given condition (class). Finally, we present a case study on *Celeb* dataset and conclude the section.

G-NeuroDAVIS in an unsupervised setting

Here the performance of G-NeuroDAVIS in the context of unsupervised learning has been described. In other words, the results described in this section have been obtained by training G-NeuroDAVIS without having prior class information.

Visualization through generalized embedding

Generalized embedding offers significant advantages in data analysis by simplifying complex, high-dimensional data into more manageable and informative lower-dimensional representations through dimension reduction. The task of dimension reduction is crucial over the image datasets due to data sparsity, complex structure, and presence of noise. Data visualization is a more precise task of dimension reduction where the latent dimension is fixed to 2 or 3 so that the latent embedding can be visualized. G-NeuroDAVIS has been trained on the above-mentioned image datasets and a comparative analysis has been shown.

In Figures 6.3(A) and 6.3(B), 2-dimensional embeddings, using G-NeuroDAVIS (left) and VAE (right) for *MNIST* and *FMNIST* datasets, respectively, have been provided. The samples are colored with respect to different class labels. It has been revealed from the figures that in both cases the classes are well separated. In the embeddings generated by G-NeuroDAVIS, different sub-clusters are observed, which is reasonable due to the intra-class variety present in both datasets.

Likewise, 2-dimensional embeddings obtained on *Coil20* and *Olivetti Faces* dataset using G-NeuroDAVIS (left) and VAE (right) have been shown in Figures 6.4(A) and 6.4(B),

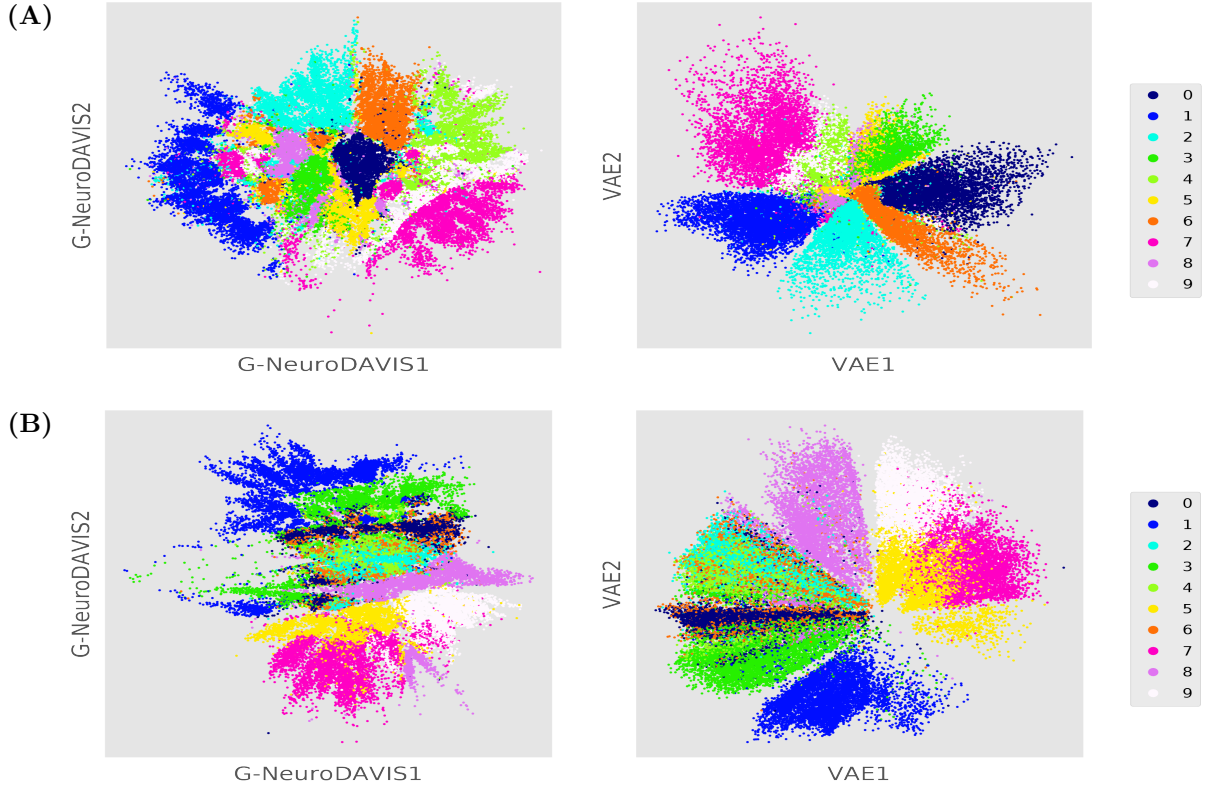


Figure 6.3: 2-dimensional embeddings produced by G-NeuroDAVIS (left) and VAE (right) for (A) *MNIST* and (B) *FMNIST* datasets respectively, in the context of unsupervised learning (Different class samples are represented by different colors).

respectively. In *Coil20* dataset, there are 72 pictures of 20 different objects, each taken from different angles, which exhibit circular patterns in the original space. It has been revealed from Figure 6.4(A) that the embedding generated from G-NeuroDAVIS produces circular clusters, whereas VAE generates compact clusters instead. It has been noted that in both the embeddings, similar points are closer. For the *Olivetti Faces* dataset, the classes are well-separated, and similar points remain closer in G-NeuroDAVIS-generated embedding. In contrast, VAE-produced embedding represents a mixture of different classes. This may happen due to the higher dimension of the dataset itself.

Moreover, G-NeuroDAVIS-generated embeddings are generalized in nature. It reveals that the samples are densely positioned and each location in the embedding can be mapped to a realistic high-dimensional sample for a dataset. G-NeuroDAVIS is designed to enhance this generalization by learning compact and informative representations that maintain the inherent semantic relationships of the original data. It outperforms VAE in terms of generalized embedding.

The key distinction between G-NeuroDAVIS and VAE lies in their encoding mechanisms.

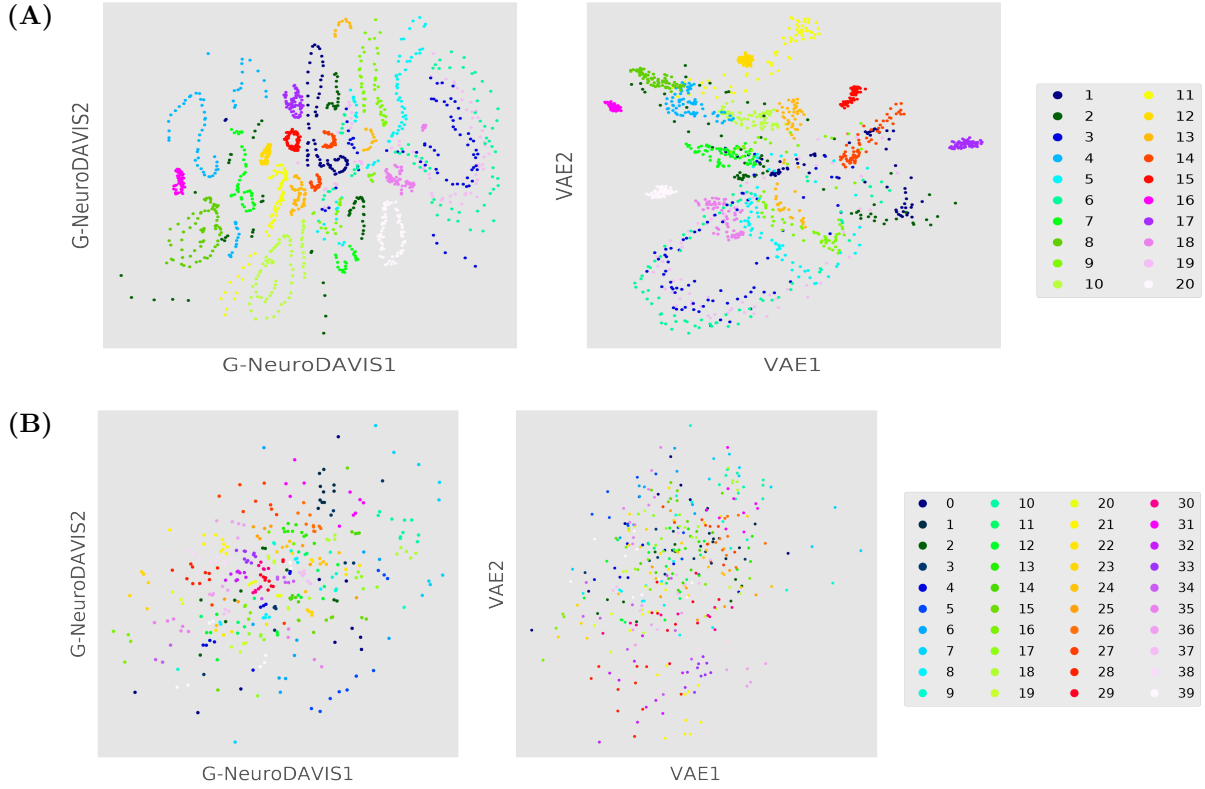


Figure 6.4: 2-dimensional embeddings produced by G-NeuroDAVIS (left) and VAE (right) for (A) *Coil20* and (B) *Olivetti Faces* datasets respectively, in the context of unsupervised learning (Different class samples are represented by different colors).

In a VAE, the encoder processes the input data to produce the mean and variance vectors, which are then used to sample from the latent embedding. In contrast, G-NeuroDAVIS employs an identity matrix as input, which is independent of the original data distribution, and introduces an additional degree of freedom over the latent space. This decoupling from the input data distribution allows the model to explore the latent space more freely and may contribute to its superior performance over VAE.

Embedding mapping

In order to evaluate the effectiveness of G-NeuroDAVIS, an embedding mapping has been visualized, and a comparison has been made against the same by using VAE. An Embedding mapping reflects the interpolation capability of the learned latent embedding. In this context, a 30×30 grid space has been considered on the entire embeddings produced by both of them. Each mesh point of the grid is also known as an interpolating point, which is a possible lower-dimensional point in the embeddings. These interpolating points have been projected to the high-dimensional space using the trained networks. These projections indicate the generation capability and diversity of the embeddings.

These projections or reconstructions have then been fitted to the grid space to generate a pictorial representation of the learned mapping as well as capture the diversity.

This experiment has been performed on all the above-mentioned datasets, and the corresponding embedding mappings have been shown in Figures C.I.1-C.I.4 in Appendices. As stated earlier, small sub-clusters have been observed in the embedding generated by G-NeuroDAVIS in the *MNIST* dataset, which has also been reflected in the embedding mapping (Figure C.I.1).

From Figure C.I.1 (left), it has been observed that the clusters, related to the number ‘6’, occurred in two places; one at the top right corner and another at the bottom left corner. Though the numbers are the same in both places, the patterns are different. It has been observed that all different classes of data are present in the embedding mapping produced by G-NeuroDAVIS, whereas VAE has failed to do so (Figure C.I.1 (right)). It may be mentioned here that a smooth transition between clusters has been observed in Figure C.I.1 (left). In this Figure, the transition from 7 to 9 (bottom right) and from 0 to 6 (top right) reflects a smooth transition, showing superior interpolation abilities of G-NeuroDAVIS.

For *FMNIST* dataset, there are no significant changes in the embedding mapping produced by the above-mentioned models (Figure C.I.2). It has been found from Figure C.I.3 that most of the clusters are well visible in G-NeuroDAVIS-generated embedding, whereas VAE-generated embedding only reflects some major clusters for the mappings for *Coil20* dataset.

Finally, for the *Olivetti Faces* dataset, the embedding mapping produced by VAE is not at all capable of reconstructing the distinct faces. Instead, it learns only the common patterns present in the data (Figure C.I.4 (right) in Appendices). On the other hand, G-NeuroDAVIS is able to produce faces with unique features, such as glasses, beards, and different facial expressions. Moreover, each face cluster is clearly visible in the compact embedding mapping (Figure C.I.4 (left) in Appendices) produced by G-NeuroDAVIS.

Interpolation through the latent space

Here, we investigate the interpolation capabilities of the latent space generated by G-NeuroDAVIS, using *MNIST* and *FMNIST* datasets. In order to perform the analysis, we begin by selecting two representative points within the latent space — one indicated as the start point and the other as the end point. The interpolation between these two points

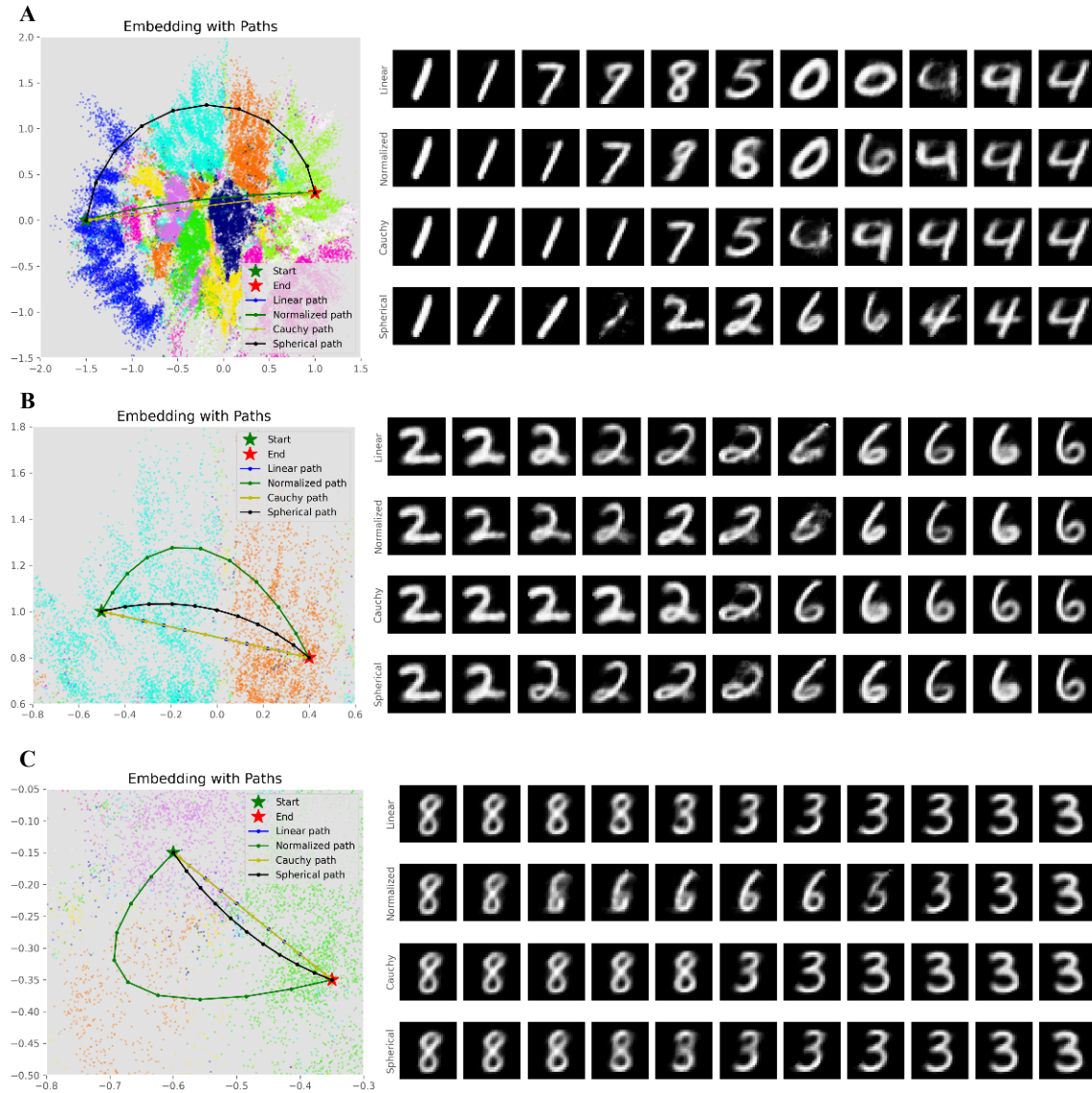


Figure 6.5: Visualization of interpolation paths in the embedding space, and their corresponding decoded outputs on *MNIST* dataset. (Left) Scatter plot showing the latent embedding colored by class labels, with interpolation paths between the start (green star) and end (red star) points using linear, normalized, Cauchy, and spherical interpolation methods. (Right) Decoded images along each interpolation path (rows) from start to end (columns), demonstrating how different interpolation strategies traverse the latent space and influence the generations. (A) Inter-cluster interpolation with embedding traversal, (B) Isolated inter-cluster interpolation, and (C) Cluster-intersecting interpolation.

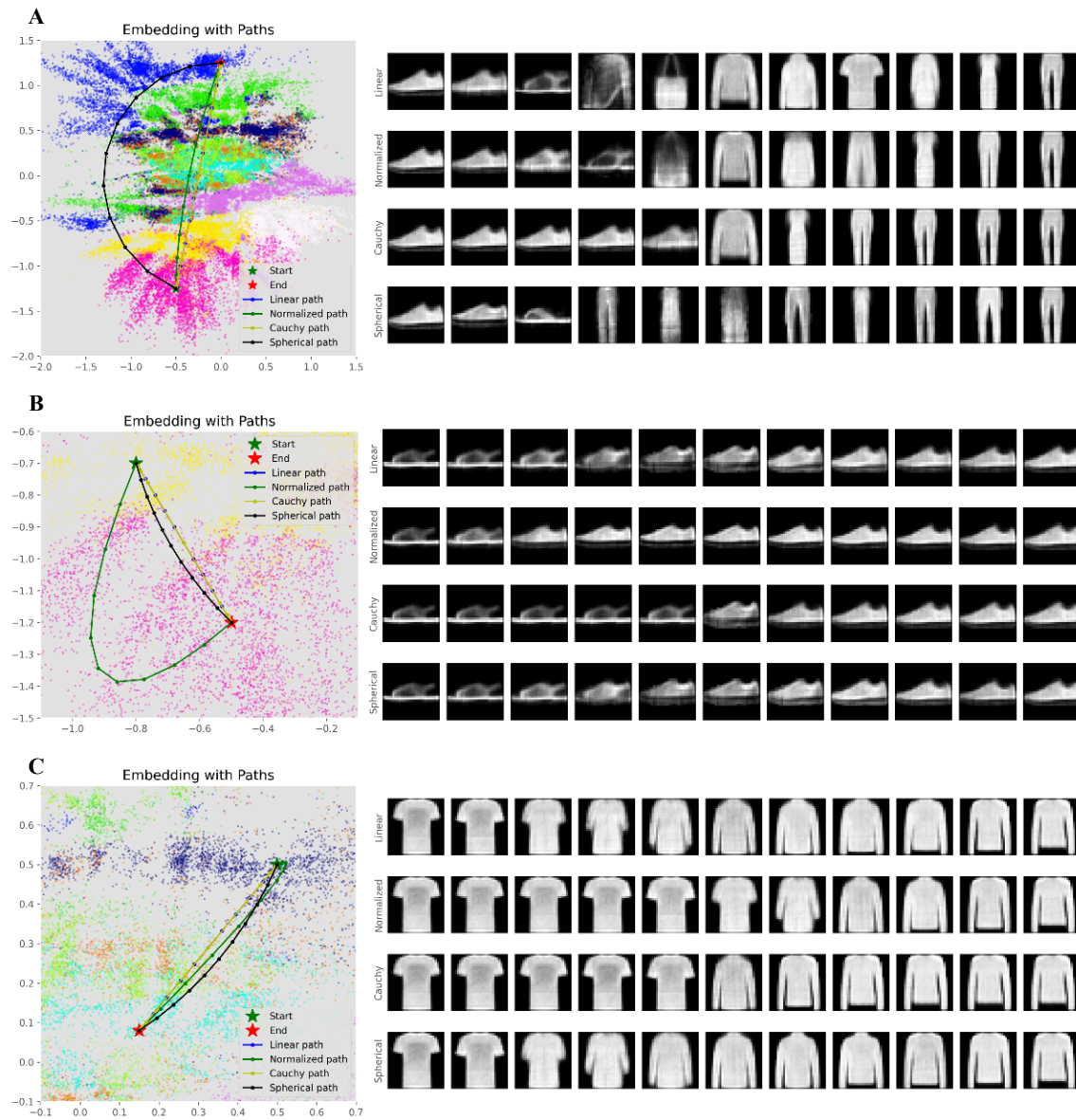


Figure 6.6: Visualization of interpolation paths in the embedding space, and their corresponding decoded outputs on *FMNIST* dataset. (Left) Scatter plot showing the latent embedding colored by class labels, with interpolation paths between the start (green star) and end (red star) points using linear, normalized, Cauchy, and spherical interpolation methods. (Right) Decoded images along each interpolation path (rows) from start to end (columns), demonstrating how different interpolation strategies traverse the latent space and influence the generations. (A) Inter-cluster interpolation with embedding traversal, (B) Isolated inter-cluster interpolation, and (C) Cluster-intersecting interpolation.

has been carried out using four distinct paths, viz., Linear, Normalized linear, Cauchy, and Spherical [135]. These paths provide diverse perspectives on how the latent space can be traversed smoothly and meaningfully. In order to ensure a comprehensive and unbiased evaluation, three specific scenarios for these interpolation paths have been considered:

- Inter-cluster interpolation with embedding traversal – where the interpolation path spans from one cluster to another, traversing a large portion of the latent space and potentially interacting with multiple intermediate clusters.
- Isolated inter-cluster interpolation – where the path connects two clusters such that no other cluster lies along the path, offering a clean view of the transition.
- Cluster-intersecting interpolation – where the interpolation path crosses exactly one additional, distinct cluster, allowing us to observe how the latent structure responds when traversing through meaningful intermediate representations.

Figures 6.5(A) - 6.5(C) illustrates the inter-cluster interpolation with embedding traversal, isolated inter-cluster interpolation, and cluster-intersecting interpolation, respectively, for *MNIST* dataset. Correspondingly, Figures 6.6(A) - 6.6(C) presents the same scenarios for *FMNIST* dataset. Figure 6.5(A) showcases interpolation over a longer trajectory, navigating through multiple intermediate clusters. For instance, the spherical path begins at the cluster of ‘1’s, passes through clusters of ‘2’s and ‘6’s, and finally reaches the cluster of ‘4’s — this progression is distinctly mirrored in the decoded images. In Figure 6.5(B), a smooth transition is observed across all the interpolation paths from the cluster of ‘2’s to the cluster of ‘6’s, demonstrating the model’s ability to preserve semantic consistency. Figure 6.5(C) reveals that the linear, Spherical and Cauchy interpolation paths traverse directly from the cluster of ‘8’s to ‘3’s, and it is reflected in the generated samples. Interestingly, the normalized path deviates slightly by passing through the cluster of ‘6’s, resulting in a two-step semantic transition — from ‘8’ to ‘6’, and then ‘6’ to ‘3’ — which is visible in the reconstructions. Figure 6.6(A) also exhibits a smooth variation of patterns through all the paths for *FMNIST* dataset. In Figure 6.6(B), a coherent transition from sandals to sneakers is evident across all interpolation strategies. Lastly, Figure 6.6(C) demonstrates a clear transformation from T-shirts to pullovers, passing through coats. Notably, the gradual increase in sleeve length from start to end exemplifies the smooth and semantically meaningful latent space learned during training.

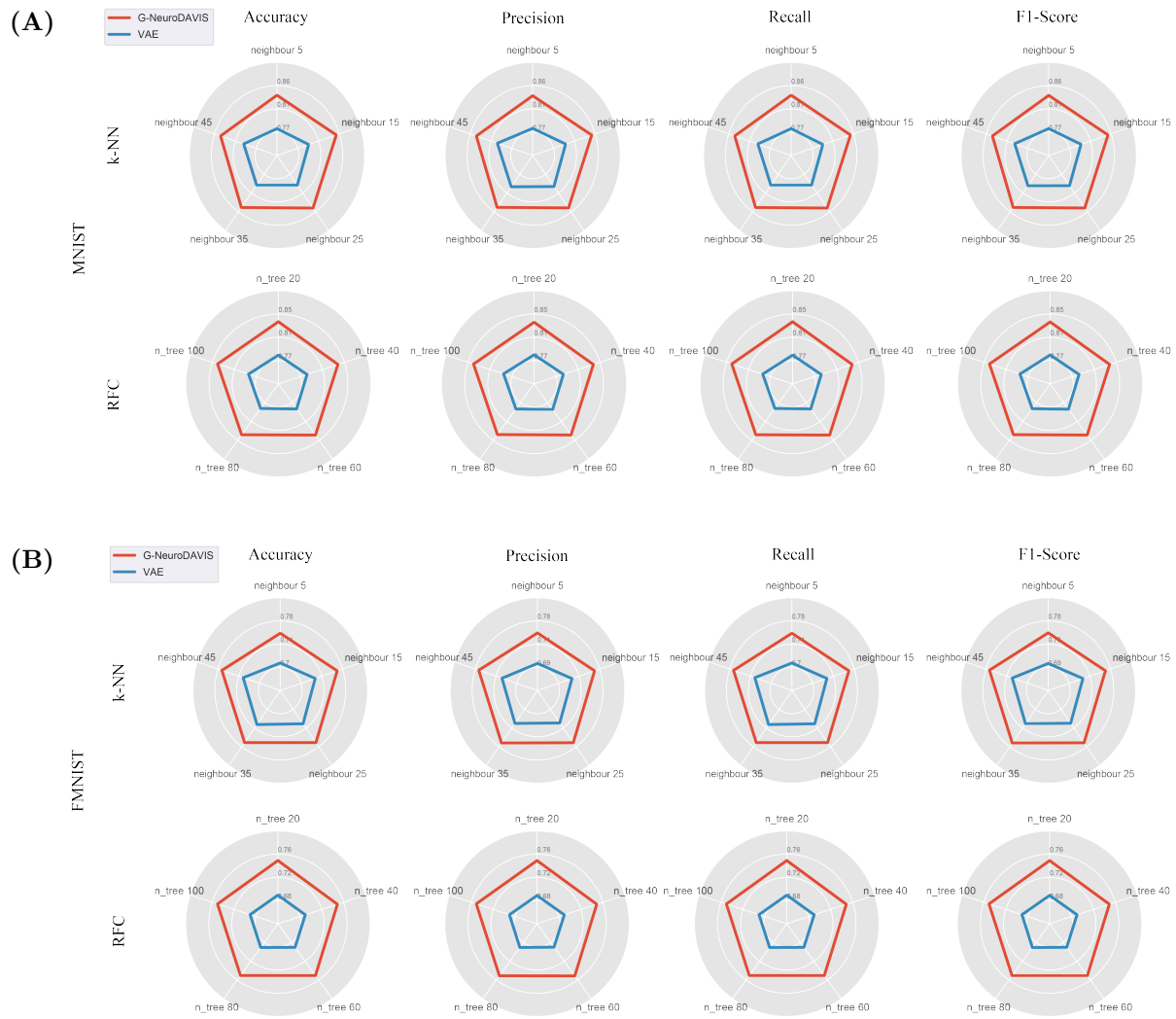


Figure 6.7: Classification performance of G-NeuroDAVIS against VAE on (A) *MNIST* and (B) *FMNIST* dataset using k-NN and RF classifiers in terms of Accuracy, precision, Recall, and F1-Score. Results have been obtained by varying parameters of k-NN (*neighbours*) and RF (*estimators* / *trees*) respectively.

Downstream Analysis

After successful training of both G-NeuroDAVIS and VAE, embeddings have been obtained from the latent layer of G-NeuroDAVIS and the bottleneck layer of the VAE respectively. These embeddings have again been compared with respect to several downstream analyses. Initially, the classification performance of both the models has been compared over all the datasets. In this study, k-NN and Random Forest (RF) classifiers have been utilized. For this comparison, all these datasets have been split into a 80 : 20 ratio in which the larger part is used to train the classifiers, and the remaining part is then used to test the classifiers. Accuracy, Precision, Recall, and F1-scores have been reported

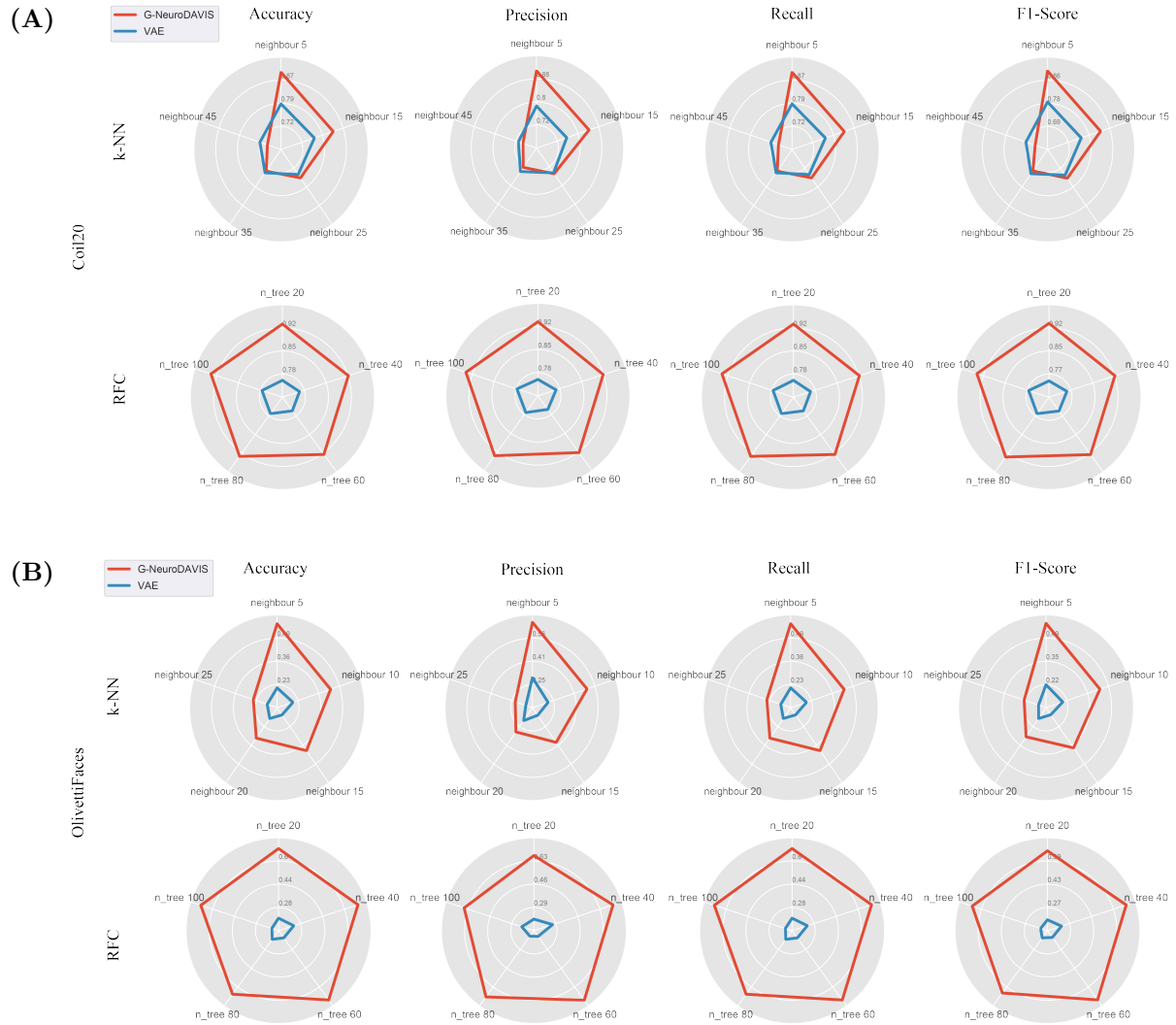


Figure 6.8: Classification performance of G-NeuroDAVIS against VAE on (A) *Coil20* and (B) *Olivetti Faces* dataset using k-NN and RF classifiers in terms of Accuracy, precision, Recall and F1-Score. Results have been obtained by varying parameter of k-NN (*neighbours*) and RF (*estimators / trees*) respectively.

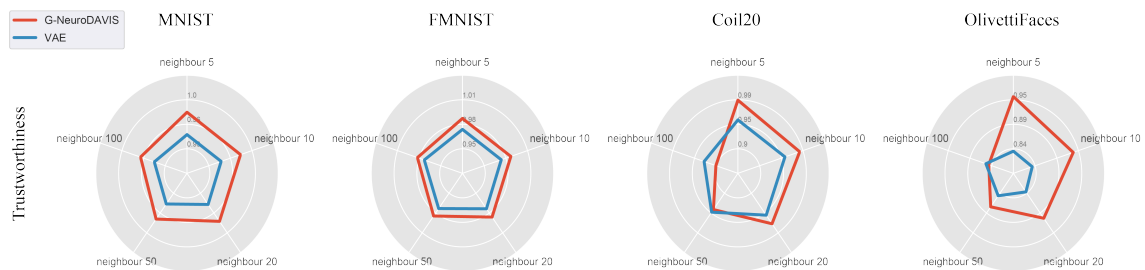


Figure 6.9: Trustworthiness scores obtained on G-NeuroDAVIS-generated embedding of *MNIST*, *FMNIST*, *Coil20*, and *Olivetti Faces* datasets and compared against that obtained by using VAE. Parameter $n_neighbour$ has been varied for robust results.

to quantify the classification performance. For both *MNIST* and *FMNIST* datasets, G-NeuroDAVIS has outperformed VAE in terms of the above-mentioned metrics (Figure 6.7).

For *Coil20* dataset, RF shows superior performance of G-NeuroDAVIS over VAE, but k-NN shows comparable performance for a higher value of the parameter k (Figure 6.8(A)). On the contrary, both k-NN and RF show superior performance of G-NeuroDAVIS-generated embedding over VAE-generated embedding for *Olivetti Faces* dataset in terms of Accuracy, Precision, Recall, and F1-score (Figure 6.8(B)). This experiment shows that G-NeuroDAVIS generated latent space has been able to preserve the class information though it has been trained in an unsupervised setting. The classifiers have been trained by varying their parameters to check their robustness.

Later, we have evaluated the embedding quality using another well known metric, called Trustworthiness. It takes both the embedding and the original high-dimensional dataset as input and calculates how close the embedding is in terms of local distances. Locality has been measured by a parameter, called $n_neighbour$. In this study, we have used five different values for $n_neighbour$ and observed superior performance of G-NeuroDAVIS over VAE for *MNIST* and *FMNIST* datasets. For *Coil20*, the results obtained are quite comparable, and for *Olivetti Faces*, a higher score has been observed for G-NeuroDAVIS for lower values of $n_neighbour$, as the value of $n_neighbour$ increases similar result has been found 6.9.

Sub-sample based analyses

Another sub-sample-based experiment has been carried out to validate the robustness of G-NeuroDAVIS against VAE, by altering the sample amount. For the same, different sizes of random subsamples (10%, 20%, $\dots$, 90%, entire data) of *MNIST* and *FMNIST* datasets with a sample size of 60k have been taken into consideration. It has been found that all the sub-samples are from a similar distribution (Figures C.II.1 and C.II.2 in Appendices).

Apart from visualization, a classification performance has been obtained on the embedding using k-NN and RF classifiers by altering their parameters. Figures C.II.3 and C.II.4 (in Appendices) demonstrate that for all embeddings of varying sizes, G-NeuroDAVIS-produced embeddings have outperformed VAE in terms of accuracy, precision, recall, and F1-Score.

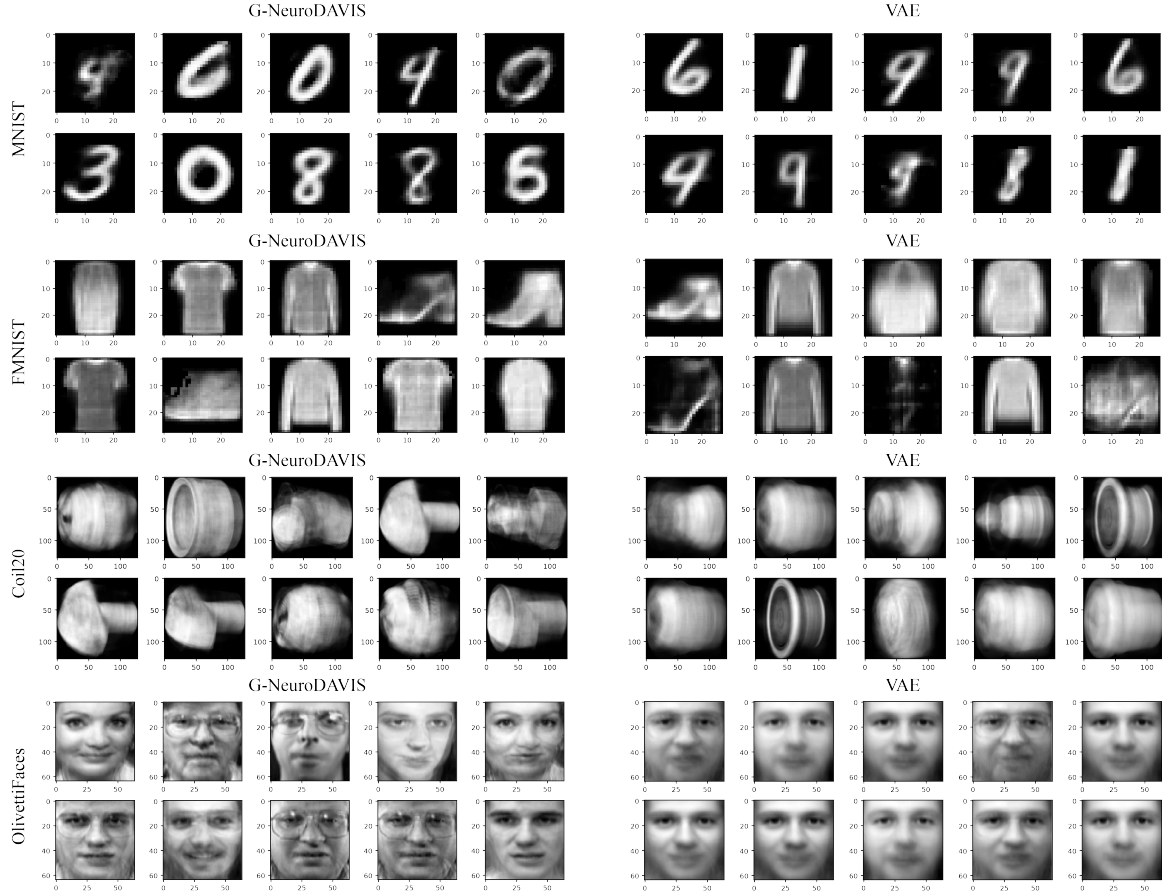


Figure 6.10: Samples generated using G-NeuroDAVIS (left) and VAE (right) in an unsupervised fashion on *MNIST*, *FMNIST*, *Coil20*, and *Olivetti Faces* dataset.

Sample Generation

This section presents a comparison of G-NeuroDAVIS with VAE, DCGAN, DDPM, and AE-guided RealNVP, in terms of their sample generation capabilities. For validation purposes, 10 samples have been generated from each of these methods and then compared against each other. G-NeuroDAVIS first learns a latent embedding, then random points are sampled from the same, and finally, these samples are projected into the high-dimensional space of the original dataset by utilizing the *Hidden layers* and *Reconstruction layer*. VAE uses its decoder to project random samples to high dimensions to generate samples. DCGAN uses its generator sub-module to generate samples. For DDPM, sampling starts from a Gaussian noise, followed by iteratively denoising it using a trained U-NET architecture [62]. At each step, the model predicts and removes noise to gradually reconstruct a clean data sample. Finally, the task of image generation with RealNVP has been considered, which is computationally expensive. For this reason, we have simplified the model with an auxiliary autoencoder (AE). The bottleneck of the AE has been used



Figure 6.11: Samples generated using AE-guided RealNVP on (A) *MNIST*, (B) *FMNIST*, (C) *Coil20*, and (D) *Olivetti Faces* datasets. Samples generated using DCGAN on (E) *MNIST* and (F) *FMNIST* datasets. Samples generated using DDPM on (G) *MNIST* and (H) *FMNIST* datasets.

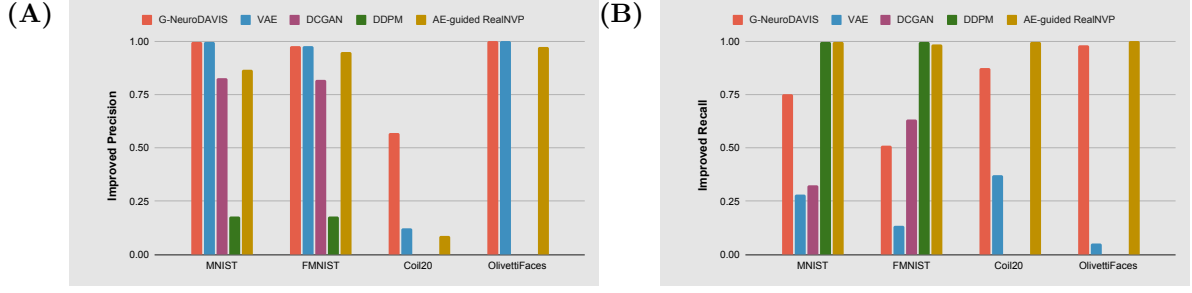


Figure 6.12: Comparison of G-NeuroDAVIS in terms of (A) improved precision and (B) improved recall scores against VAE, DCGAN, DDPM, and AE-guided RealNVP on *MNIST*, *FMNIST*, *Coil20*, and *Olivetti Faces* datasets.

for sample generation.

As observed in Figure 6.10, G-NeuroDAVIS-generated samples for *MNIST* dataset are realistic and readable; for *FMNIST* dataset, the generated samples by G-NeuroDAVIS show distinct clothing items; while for *Coil20* dataset, various objects have been found clear in the G-NeuroDAVIS-generated samples; and for *Olivetti Faces* dataset, G-NeuroDAVIS has been found to generate distinct faces. On the other hand, the generation from VAE model is not at all comparable for datasets with very high dimensions, like *Coil20* and *Olivetti Faces*. For *Olivetti Faces*, all generated samples of facial expression contain only common features but could not capture the unique features present in the data; and for *Coil20*, the samples are blurred in nature. However, comparable samples have been observed for *MNIST* and *FMNIST* datasets.

Figure 6.11((A)-(D)) shows that AE-guided RealNVP has produced comparable results for all four datasets. However, DCGAN and DDPM-produced samples are comparable for *MNIST* data, but not that clear for *FMNIST* (Figure 6.11((E)-(F)) (DCGAN) and Figure 6.11((G)-(H)) (DDPM)). For the larger two datasets, we are not able to train DCGAN and DDPM due to limited computational resources.

Quality of sample generation

All the above comparisons of generated samples have primarily relied on the visual appearances of the samples. In this section, we present a quantitative study that evaluates the quality of the sample generation from two major aspects, i.e., diversity and authenticity. In this context, we have used two different metrics, viz., improved precision and improved recall [130]. Here, diversity refers to the existence of variability in the generated samples, whereas authenticity shows how realistic the generated samples are with respect to real

data. Improved precision quantifies how the generated samples lie within the manifold of the real sample distribution, and thereby measures authenticity. On the other hand, improved recall refers to the presence of variability in the generation process, which indicates diversity. However, the key challenge is balancing both measurements.

Figure 6.12 presents a comparative evaluation of G-NeuroDAVIS against VAE, DCGAN, DDPM, and AE-Guided RealNVP across four datasets, viz., *MNIST*, *FMNIST*, *Coil20*, and *Olivetti Faces*, using improved precision and improved recall. Figure 6.12(A) illustrates the authenticity of the generated samples, where G-NeuroDAVIS consistently achieves the highest or near-highest precision across all datasets, indicating superior realism. On the other hand, Figure 6.12(B) highlights the diversity of generated outputs, and again, G-NeuroDAVIS shows strong performance, particularly on complex datasets like *Olivetti Faces* and *Coil20*, signifying a broad coverage of the data distribution. Overall, these figures demonstrate that G-NeuroDAVIS effectively balances both authenticity and diversity, outperforming or matching other models across varied datasets.

G-NeuroDAVIS in a supervised setting

Unlike other generative models, G-NeuroDAVIS can also be trained in a supervised setting and has been used for the above-mentioned datasets. G-NeuroDAVIS learns a distribution specifically for each class based on the label information provided by the data itself. This learning technique enables conditional data generation. Initially, an embedding has been produced after the model training (Figure C.III.1 in Appendices), and samples have been generated accordingly. In the context of unsupervised learning, no information is given about the classes, and as a result, some overlapping of classes has been observed in Figures 6.3 and 6.4. However, after having the class information, G-NeuroDAVIS has produced well-separated distinct classes across all the datasets (Figure C.III.1 in Appendices).

In addition, its conditional data generation capability has been examined, and accordingly, the samples have been generated from all the classes present in the dataset. For each class one sample has been generated using the trained model and all the conditional samples have been depicted in Figure 6.13.

Conditional data generation validation involves ensuring that the generated data meets specific conditions or constraints. It has been observed from Figure 6.13, clear readable images for *MNIST*; all the clothing items, including bag and shoes for *FMNIST*; for

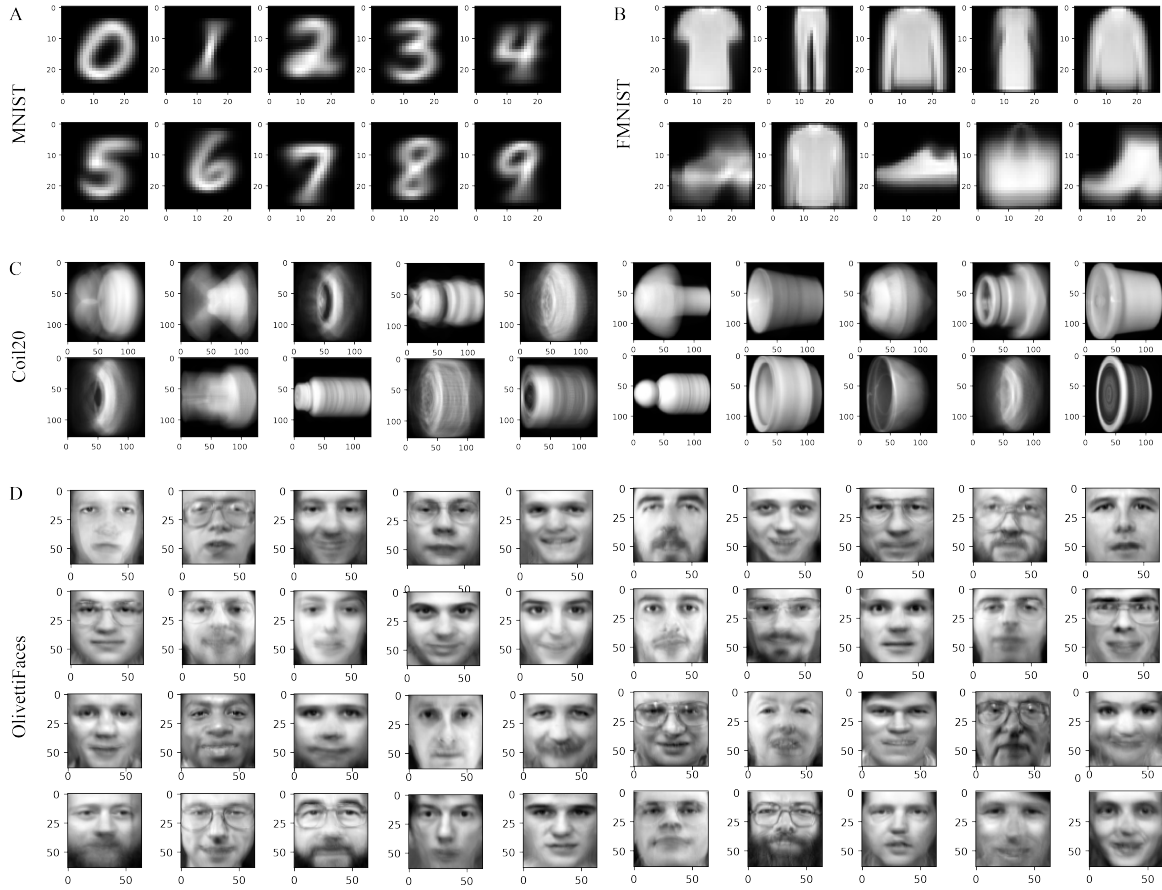


Figure 6.13: Samples generated using G-NeuroDAVIS in an supervised fashion on (A) *MNIST*, (B) *FMNIST*, (C) *Coil20*, and (D) *Olivetti Faces* dataset.

Coil20 dataset, all 20 distinct items; and for *Olivetti Faces* dataset, all 40 different faces have been generated. Each of these generations matches the specific condition very well. G-NeuroDAVIS has been able to generate those class-specific samples successfully, not only for datasets having a smaller number of classes but also for datasets having more than 20 different classes.

This supervised learning adds a control to the generation capability of G-NeuroDAVIS and makes it even more useful than the state-of-the-art methods.

Complexity analyses

In this section, we present a computational complexity analysis of G-NeuroDAVIS against the state-of-the-art methods. It refers to understand how the complexity of the models affect their performance in terms of using trainable parameters, runtime per epoch, and inference time of the models.

Figure 6.14 shows the comparative analysis on the number of trainable parameters considered for training the developed model and the state-of-the-art methods. It has been observed that G-NeuroDAVIS uses a significantly lower number of parameters compared to VAE, DCGAN, DDPM, and AE-guided RealNVP. However, the performance in terms of generalized embedding, downstream tasks like classification, and sample generation is closely comparable.

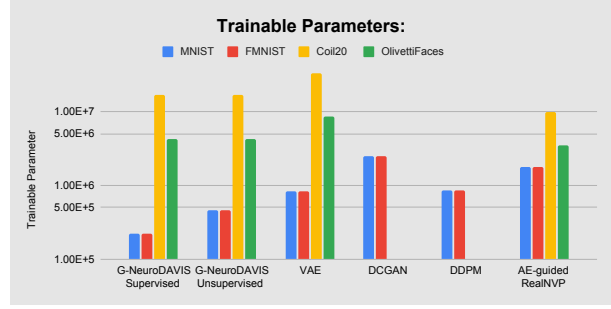


Figure 6.14: Comparison based on trainable parameters

Likewise, a time complexity analysis was also carried out with respect to runtime and inference time. As each model takes a different number of epochs to converge, the time taken for each epoch is calculated and utilized for a fair comparison. Figure 6.15(A) shows the time taken by each method on a log scale, where G-NeuroDAVIS shows comparable runtime with VAE; however, DCGAN, DDPM, and AE-guided RealNVP take much longer to train. Figure 6.15(B) shows a comparable inference time of the developed model with VAE. Here, inference time indicates the time (in log seconds) taken by any method to generate 100 samples. It is observed that both DCGAN and DDPM take much more time than G-NeuroDAVIS, but AE-guided RealNVP is the fastest to achieve the same, which might be due to the simultaneous usage of an autoencoder.

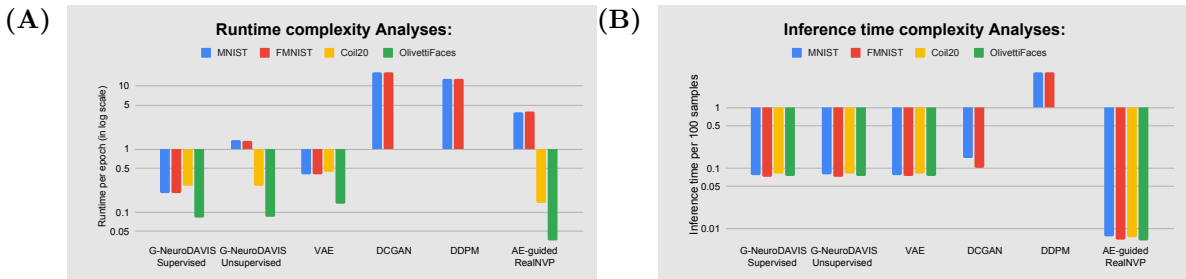


Figure 6.15: (A) Run time and (B) Inference time comparison of G-NeuroDAVIS against VAE, DCGAN, DDPM, and AE-guided RealNVP on *MNIST*, *FMNIST*, *Coil20*, and *Olivetti Faces* dataset.

Sensitivity analyses

There are several hyper-parameters involved in the design and implementation of G-NeuroDAVIS. These hyper-parameters play a crucial role in the model's performance, and

thus, sensitivity analyses of these hyper-parameters become essential. We have varied two such hyper-parameters, viz., batch size and number of neurons in the hidden layers, across all the datasets, and reported the variation in downstream analyses like classification. Performance has been evaluated using accuracy, precision, recall, and F1-scores. Figure 6.16 shows that the classification performance remains stable for *MNIST*, *FMNIST*, and *Coil20* dataset. On the other hand, a little change has been observed for *Olivetti Faces* dataset. Results are similar for both k-NN and RF classifiers. According to this experiment, G-NeuroDAVIS's performance becomes more stable for datasets with larger sample sizes.

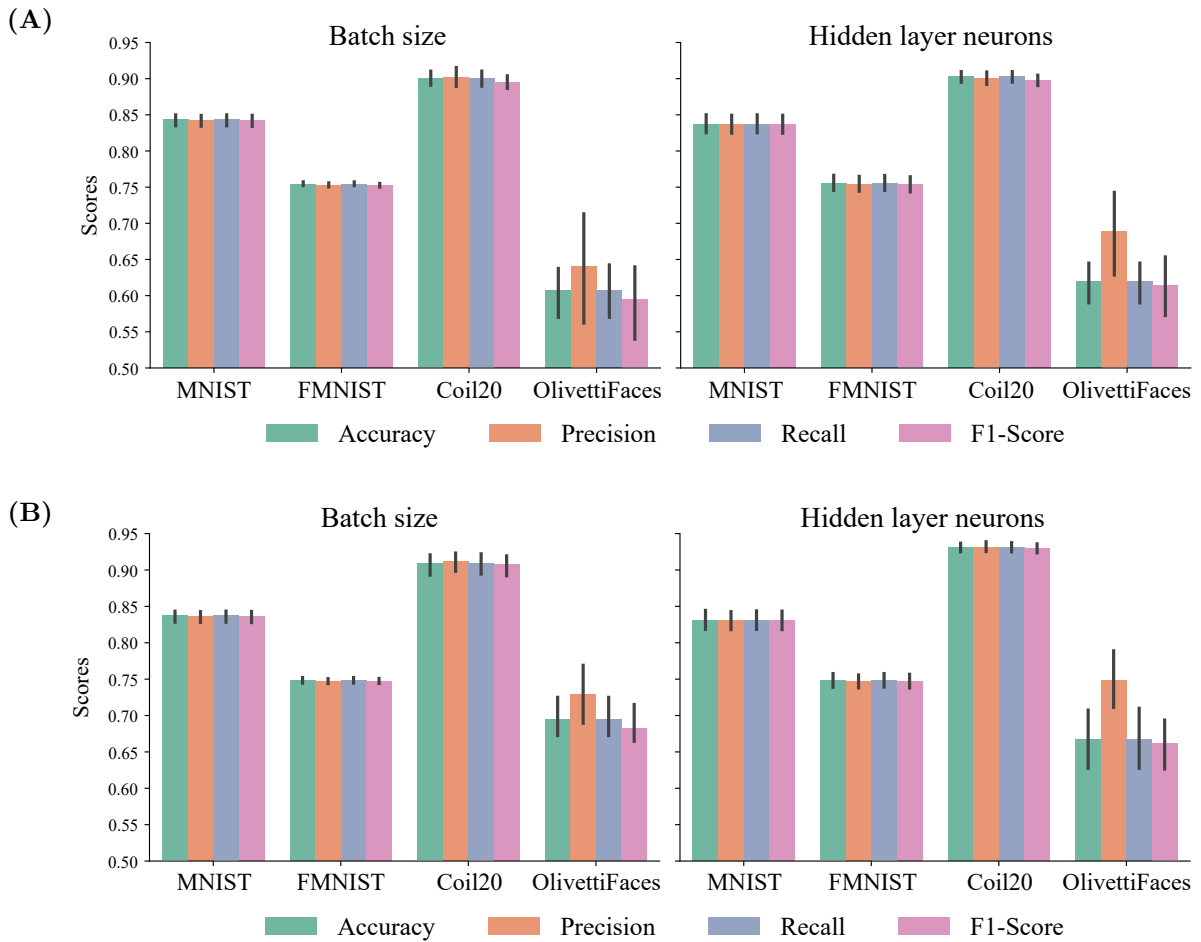


Figure 6.16: (A) k-NN and (B) RF classifiers performance with varying hyper-parameters of G-NeuroDAVIS on *MNIST*, *FMNIST*, *Coil20*, and *Olivetti Faces* dataset.

A case study on Celeb dataset

Finally, we present a case study on the *Celeb* dataset available at <https://www.kaggle.com/datasets/vishesh1412/celebrity-face-image-dataset>. This dataset contains

images of 18 Hollywood celebrities with 100 images of each celebrity. The images are of different sizes, due to which we have cropped the data into a 150×150 images to make them fit into G-NeuroDAVIS. In order to cover the faces, we first used a Haar-cascade face detection algorithm [136] before cropping. Initially, an experiment has been carried out based on different genders. Figure 6.17(A) shows the G-NeuroDAVIS-produced embedding, where each point represents a celebrity image. Colors indicate gender labels, which have not been used during the embedding process but are visualized for interpretation. Despite being unsupervised, the embedding reveals a separation between the two gender groups, suggesting that the learned features naturally capture gender-relevant visual traits. After training, the sample generation capability has been examined, and it has been observed from Figure 6.17(B) that both male and female faces have shown up during inference, which indicates diversity. However, the generated figures are a little blurred, indicating poor authenticity, which can be improved using a deep architecture of G-NeuroDAVIS.

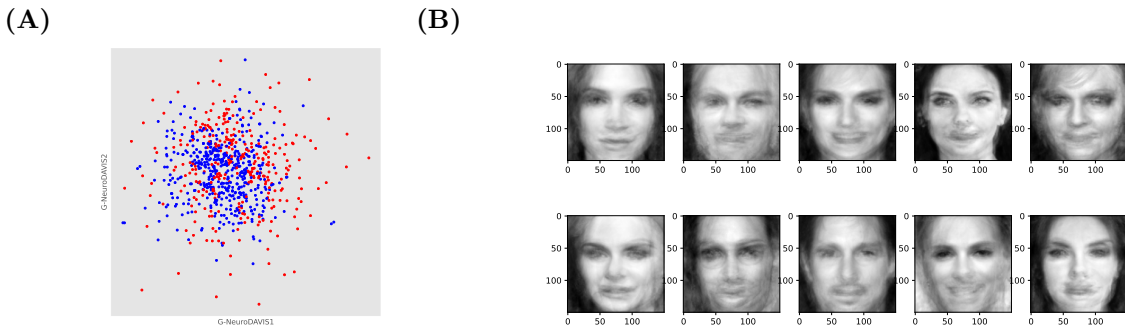


Figure 6.17: (A) 2-dimensional embeddings produced by G-NeuroDAVIS on *Celeb* dataset in the context of unsupervised learning (Blue dots indicate male faces, while Red dots represent female faces). (B) Samples generated using G-NeuroDAVIS in an unsupervised fashion on *Celeb* dataset.

6.4 Discussion and Conclusion

In this chapter, we have developed a feed-forward neural network model, called G-NeuroDAVIS, for data visualization through generalized embedding and sample generation. G-NeuroDAVIS has been able to produce meaningful embeddings by extracting features from high-dimensional datasets. The latent embedding produced by it can generate samples that are realistic and diverse as well. The generalization capability of any latent representation is crucial for sample generation. G-NeuroDAVIS encodes patterns that can

be extended beyond the training set for sample generation. The interpolation experiment has shown how well G-NeuroDAVIS preserves the underlying data structure in the latent space. It can be trained in both supervised and unsupervised settings for handling high-dimensional complex datasets. Training in a supervised setting allows it to generate samples based on some pre-defined condition, like classes, and in this way, it controls the process of generation.

The performance of G-NeuroDAVIS has been compared with VAE, which has a similar capability, on four publicly available image datasets, viz., *MNIST*, *FMNIST*, *Olivetti Faces*, and *Coil20*. Unlike VAE, G-NeuroDAVIS can generate samples based on conditions. It has been observed that though the embedding produced by G-NeuroDAVIS is similar to VAE in terms of visualization, the classification performance reflects its superiority over VAE. A smooth transition has also been observed in the latent space, which reflects a superior interpolation capability. The sub-sample-based experiment also shows the robustness of G-NeuroDAVIS over VAE. G-NeuroDAVIS has also shown superior generation capability over VAE, DCGAN and DDPM. However, AE-guided RealNVP produces comparable samples. These results are validated using improved precision and improved recall measures. Moreover, G-NeuroDAVIS is trained using a significantly smaller number of parameters compared to the other state-of-the-art methods.

G-NeuroDAVIS has performed well on various datasets but has a few limitations. First, the input it takes to visualize the data is of the order of the square of the sample size, which makes a huge space complexity. Second, being a neural network model, it has several hyper-parameters that need to be set wisely. Sometimes, a little modification can significantly impact the results. Finally, G-NeuroDAVIS assumes the embedding to follow a Gaussian distribution, which may not always be the optimal case. In addition, G-NeuroDAVIS can be used for performing conditional data generation, having a smoother transition within the latent embedding.

Chapter 7

Conclusions and Scope for Further Research

This chapter concludes the thesis by discussing the key research contributions, notable results, strengths, weaknesses, and a possible direction of future scope.

7.1 Research Contribution

The thesis makes two principal contributions. The first and most substantial contribution pertains to the novel architectures of the developed models of neural networks, and the second contribution is demonstrated through the statistically significant and insightful results achieved.

7.1.1 Architecture

In this thesis, Chapter 2 has developed a novel unsupervised deep learning model, called NeuroDAVIS, for data visualization. It is capable of producing a low-dimensional embedding from the data, extracting relevant features useful for subsequent analyses. The main novelty of this architecture lies in the utilization of the input layer. It controls the low-dimensional points in the latent space. After training, it helps the latent layer to produce the final visualization.

Chapter 3 presents a neural network-based feature selection model, NeuroDAVIS-FS, an extension of the earlier data visualization model NeuroDAVIS. Initially, it considers training using the model NeuroDAVIS and selecting features according to the trained

model. It functions in an unsupervised learning setup and does not assume any prior data distribution.

In Chapter 4, another neural network model, viz., NeuroMDAVIS, has been developed for visualizing multi-omics (multi-modal) data in a joint latent space. NeuroMDAVIS provides joint visualization combining all the omics modalities together, and is the first of its kind in this regard. The model is a generalization of NeuroDAVIS. NeuroMDAVIS enables visualization of multiple omics modalities as a joint embedding, which gives rise to a unified perspective depicting correlations and interactions among the modalities in the same space. In the architecture of NeuroMDAVIS, there are several notable modifications over NeuroDAVIS. Similar to NeuroDAVIS, NeuroMDAVIS also contains an input layer that controls the low-dimensional points in the latent space. Besides, it consists of two different types of hidden layers, viz., shared hidden layer and modality-specific hidden layer. The former captures information aligned in all modalities together, while the latter captures information which are possibly conflicting in some modalities. Moreover, here the loss function introduces a balancing parameter that allows additional control over different modalities.

Chapter 5 presents a multi-modal feature selection method, called NeuroMDAVIS-FS, which is built upon NeuroMDAVIS. NeuroMDAVIS-FS is designed to decode the molecular complexities of lung cancer. By applying synchronized encoding and modality-specific decoding layers, the model maps disparate data streams into a common latent projection that preserves critical biological signals. Later, NeuroMDAVIS-FS selected top molecular features have been found to be effective enough to classify patients based on their survival outcomes.

Both NeuroDAVIS and NeuroMDAVIS are feed-forward neural network architectures that can be used to produce a 2-dimensional latent embedding from high-dimensional single-omics and multi-omics data, respectively. This latent embedding/projection captures significant data characteristics, which can be useful for various downstream tasks, including classification and clustering.

Finally, in Chapter 6, another feed-forward neural network model, called G-NeuroDAVIS, has been developed for data visualization through generalized embedding, and thereby, sample generation. The architecture of G-NeuroDAVIS introduces a parametric layer that helps to produce a generalized embedding. It has been able to produce

meaningful embeddings by extracting features from high-dimensional datasets. The latent embedding produced by it can generate samples that are realistic and diverse as well. The generalization capability of any latent representation is crucial for sample generation. G-NeuroDAVIS encodes patterns that can be extended beyond the training set for sample generation. The interpolation experiment has shown how well G-NeuroDAVIS preserves the underlying data structure in the latent space. It can be trained in both supervised and unsupervised settings for handling high-dimensional complex datasets. Training in a supervised setting allows it to generate samples based on some pre-defined condition, like classes, and in this way, it controls the process of generation.

Overall, the thesis presents a comprehensive investigation into neural network architectural design, culminating in the development of three novel architectures, viz., NeuroDAVIS, NeuroMDAVIS, and G-NeuroDAVIS. NeuroDAVIS introduces a novel input layer that departs from conventional architectures. Building upon this foundation, NeuroMDAVIS incorporates both shared and modality-specific hidden layers, enabling the model to effectively capture complementary information across heterogeneous data modalities. Finally, G-NeuroDAVIS extends this framework by introducing a parametric layer that learns the underlying data distribution and generates a generalized embedding, thereby, enhancing the representational capacity and robustness of the architecture.

7.1.2 Results

The effectiveness of NeuroDAVIS has been demonstrated on a wide variety of datasets, including 2D synthetic data, and high-dimensional numeric, textual, image and biological data (Chapter 2). It has shown excellent visualization capability supported by outstanding downstream analysis (clustering/classification) capabilities. The results obtained are also statistically significant.

Later, the efficacy of NeuroDAVIS-FS has been demonstrated on various datasets from different domains, i.e., numeric, image, and biological (Chapter 3). A comparative analysis has been performed for several downstream tasks, like classification and clustering, with the state-of-the-art models, and it has been observed that the top features selected by the proposed model are efficient for the same. Moreover, comparable results have been obtained in terms of predefined metrics for the state-of-the-art methods. In addition, a good amount of overlapping features have been found in the proposed feature selection

model against the other classical models. Finally, two case studies on image and biological datasets with a very low sample-feature ratio (nearly 1 : 10) have been performed and found to be effective for feature selection. Even with 10% features, both the classifiers have been able to classify the classes of *Coil20* dataset with an accuracy of 0.9 approximately. However, for *Olivetti Faces* dataset, the accuracy rate of k-NN and RF are approximately 0.7 and 0.85 respectively. Whereas, for *Jurkat* dataset, the accuracy rate increases for the classification tasks with the top 10% selected features.

The performance of NeuroMDAVIS has been demonstrated on CITE-seq and multiome datasets, and results have been compared with several state-of-the-art methods, like t-SNE, UMAP, IVIS, Fit-SNE and MultiMAP (Chapter 4). In the context of multi-omics data visualization, NeuroMDAVIS has been able to produce qualitatively excellent visualization compared to the other state-of-the-art methods. It has also shown that the joint embedding produced by NeuroMDAVIS are more efficient than individual embeddings produced by NeuroDAVIS. In addition, the latent embedding, when used for cell-type classification or clustering, has surpassed those obtained by most of the other state-of-the-art methods.

Besides, NeuroMDAVIS-FS bridges the gap between raw high-dimensional data and biological interpretability (Chapter 5). Validation using the CPTAC lung cancer dataset confirms that the framework effectively prioritizes high-impact features. A significant improvement has been observed in C-index with the help of top molecular features over clinical ones. Furthermore, the selected top features are efficient enough to predict survival outcomes.

Finally, the performance of G-NeuroDAVIS has been compared with VAE, which has a similar capability, on four publicly available image datasets, viz., *MNIST*, *FMNIST*, *Olivetti Faces*, and *Coil20* (Chapter 6). Unlike VAE, G-NeuroDAVIS can generate samples based on conditions. It has been observed that though the embedding produced by G-NeuroDAVIS is similar to VAE in terms of visualization, the classification performance reflects its superiority over VAE. A smooth transition has also been observed in the latent space, which reflects a superior interpolation capability. The sub-sample-based experiment also shows the robustness of G-NeuroDAVIS over VAE. G-NeuroDAVIS has also shown superior generation capability over VAE, DCGAN and DDPM. However, AE-guided RealNVP produces comparable samples. These results have been validated using improved precision and improved recall measures. Moreover, G-NeuroDAVIS is trained

using significantly smaller number of parameters compared to the other state-of-the-art methods.

7.2 Strengths

The strength of NeuroDAVIS lies in the fact that it is able to preserve cluster shape, size and structure within the data well. As proved mathematically, the low-dimensional embedding produced by NeuroDAVIS has also been observed to preserve local relationships among observations in high dimensions. Furthermore, it is capable of extracting the inherent non-linearity within the data.

To the best of our knowledge, there has been no single method to date that supports joint visualization of multiple modalities together. In literature, there exist methods for visualizing one modality overlaid on the other(s) or simultaneously distinct embeddings for each modality. NeuroMDAVIS fits into this gap perfectly.

Both NeuroDAVIS and NeuroMDAVIS do not assume any prior data distribution, and thereby, are non-parametric. It stands out as one of their key strengths. They perform in an unsupervised setup, which makes them novel in comparison with traditional approaches. Besides, neither NeuroDAVIS nor NeuroMDAVIS, requires a prior initialization using PCA or other linear DR methods, contrasting to other existing methods. Moreover, unlike major existing visualization methods, like t-SNE and UMAP, both NeuroDAVIS and NeuroMDAVIS can be used as pretrained models to visualize streaming samples, which implies that they support iterative continuous integration and deployment. This attribute makes them highly suitable for large-scale projects like Human Cell Atlas (HCA) or the Human Tumour Atlas Network (HTAN) initiatives.

On the other hand, G-NeuroDAVIS can simultaneously visualize high-dimensional data in lower dimensions and generate new, realistic samples. The embedding it produces has a generalization capability. It produces more diverse and authentic samples compared to VAE, DCGAN, DDPM, and AE-guided RealNVP, particularly in high-dimensional datasets. Moreover, it can be trained in both unsupervised and supervised settings, which makes it stand out from other generative models.

Thus, in this thesis, a single solution to the long-standing problem of visualization of datasets belonging to different categories/domains/modalities has been systematically

introduced.

7.3 Limitations

Despite their demonstrated strengths and contributions, the models presented in this thesis are not without limitations. This section provides a critical discussion on the inherent weaknesses associated with each proposed model. While several limitations identified in earlier architectures have been addressed in subsequent model developments, certain challenges remain unsolved. These weaknesses highlight opportunities for further methodological refinement and extension.

The most common problem of these methods lies in their input mechanism. All of them take an identity matrix of order equal to the number of observations as input in the input layer. Thus, it consumes more space in memory than other existing methods for data comprising many samples. However, the sparse identity matrix makes the memory overhead manageable, and can be further utilized to handle the issue of scalability in case of large datasets. In this thesis, all these methods have been successfully applied to datasets of substantial sizes without encountering computational bottlenecks. It might be due to the usage of small batch sizes during model training, i.e., during any training epoch, only a mini-batch of samples has been used instead of the entire training set. For large-scale scenarios (e.g., data containing more than 100k samples), an additional optimization, such as block-wise computations or approximate strategies, may be incorporated to further improve scalability, which remains a scope for improvement in the future.

NeuroDAVIS could not handle multiple data modalities, which have been overcome in NeuroMDAVIS. NeuroMDAVIS can produce a joint embedding from multi modal dataset. Both NeuroDAVIS and NeuroMDAVIS generated latent spaces are not generalized in nature and thus fall short for data generation, which has been achieved by G-NeuroDAVIS. G-NeuroDAVIS assumes that the embedding follows a Gaussian distribution, which may not always be the optimal case. In the current scope, the assumption of Gaussian distribution helps to achieve a closed form of the KLD loss and thus produces tractable gradients for backpropagation. Any other choice of this assumption can be seen as a research gap to address in the future. Moreover, being neural network models, hyper-parameters in all these methods are sensitive to tiny modifications, and network tuning

requires a substantial amount of time and effort for an amateur.

On the other hand, the feature selection modules have relied completely on the training of a corresponding base model. If the base model has not been trained properly, the feature selection module also produces inefficient and biased features. The optimal number of features has not been considered in this study, which might be an extension of a future study. Though one can plot the feature scores in a descending order and select the top features by observing the elbow, the location of a bend (knee) in the plot is generally considered an optimal number of features.

7.4 Future Scope

Addressing the above-mentioned limitations in future studies may improve the robustness, scalability, and generalizability of the models developed in the thesis, and thereby paving the way for more comprehensive and adaptable solutions. However, there are several directions we can explore given the current state of research.

- G-NeuroDAVIS, as a generative model, can be extended to several benchmark tasks, including data augmentation, data imputation, class balancing, and more.
- Another important direction is modality generation. Multi modal datasets are the new state-of-the-art source to work on. However, technical faults often lead to missing modalities. G-NeuroDAVIS can be extended in a way so that it can generate new modalities given some other.
- In the current scope, we have worked with numeric, text, image, and biological datasets. In the future, we will try to analyze video datasets.
- Finally, all the current models do not capture any temporal association present in the data. Incorporating a methodology that can assess such information can be a scope for future direction.

Appendices

Here we present additional results that complement the analyses discussed in Chapter 2, Chapter 5, and Chapter 6 of this thesis. It includes supplementary experiments, extended quantitative evaluations, ablation studies, and supporting visualizations that were omitted from the main text for clarity and conciseness. These results provide further validation of the developed methodologies and offer deeper insights into their behaviour across different datasets, parameter settings, and experimental conditions. While not central to the primary narrative, the materials included here enhance the reproducibility, transparency, and completeness of the research findings.

A. NeuroDAVIS

A.I. Results on synthetic 2D datasets

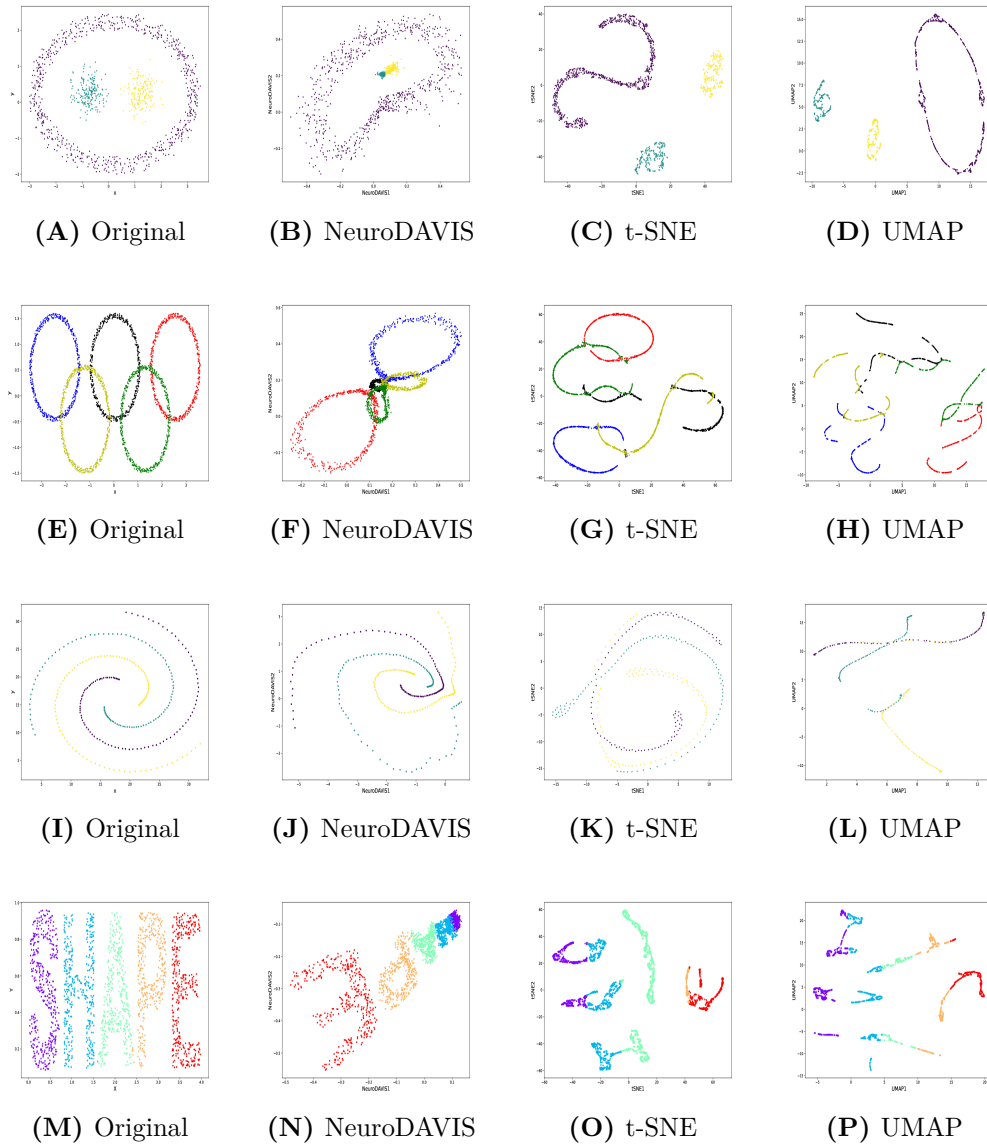


Figure A.I.1: The original distributions and the embeddings produced by NeuroDAVIS, t-SNE and UMAP after a nine-dimensional transformation $(x + y, x - y, xy, x^2, y^2, x^2y, xy^2, x^3, y^3)$ of *Elliptic Ring* ((A)-(D)), *Olympic* ((E)-(H)), *Spiral* ((I)-(L)) and *Shape* ((M)-(P)) datasets.

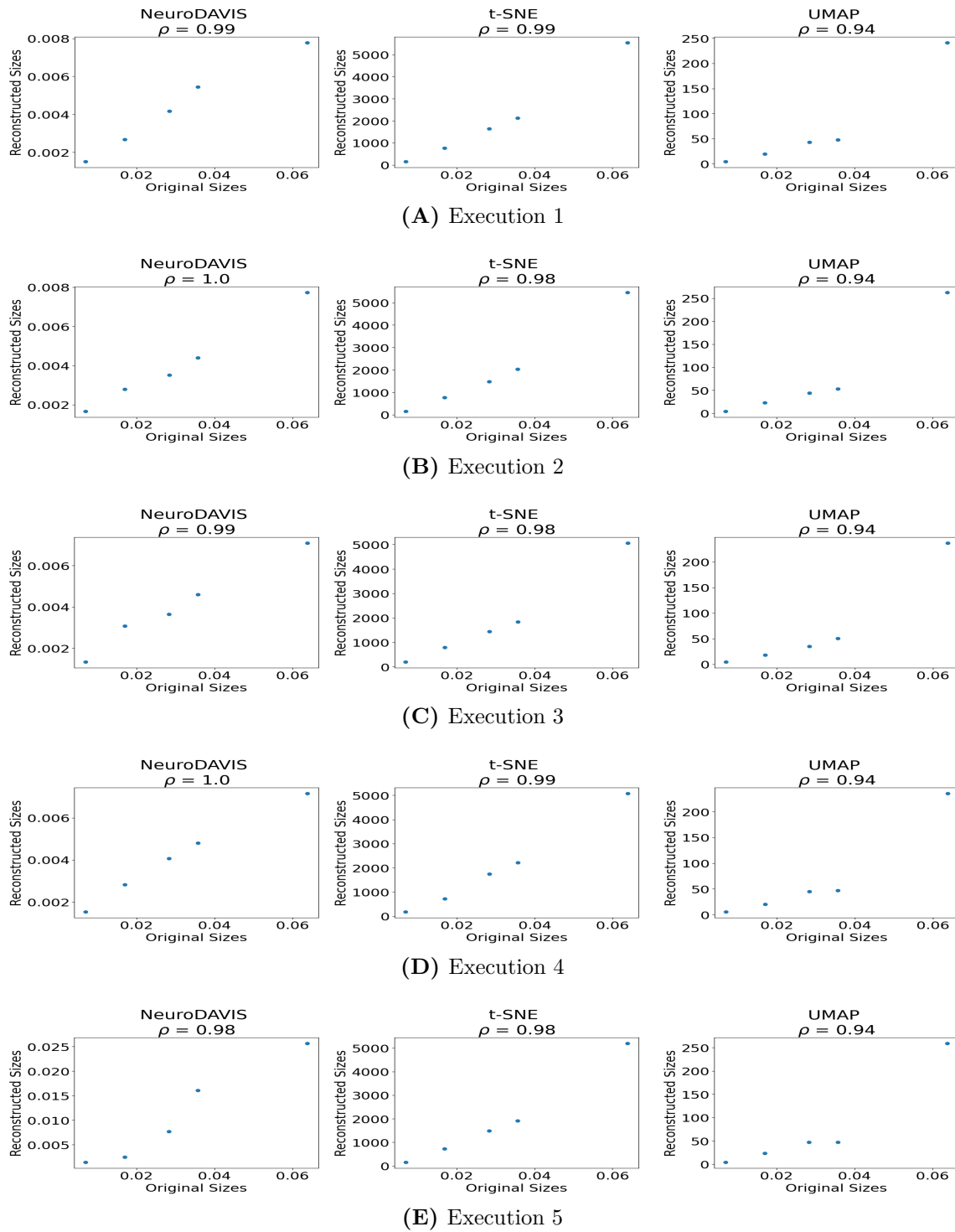


Figure A.I.2: Preservation of cluster sizes by NeuroDAVIS, t-SNE and UMAP in *World Map* dataset. ρ is the Pearson correlation coefficient. The higher the ρ value, the better the preservation.

A.II. Results on high-dimensional datasets

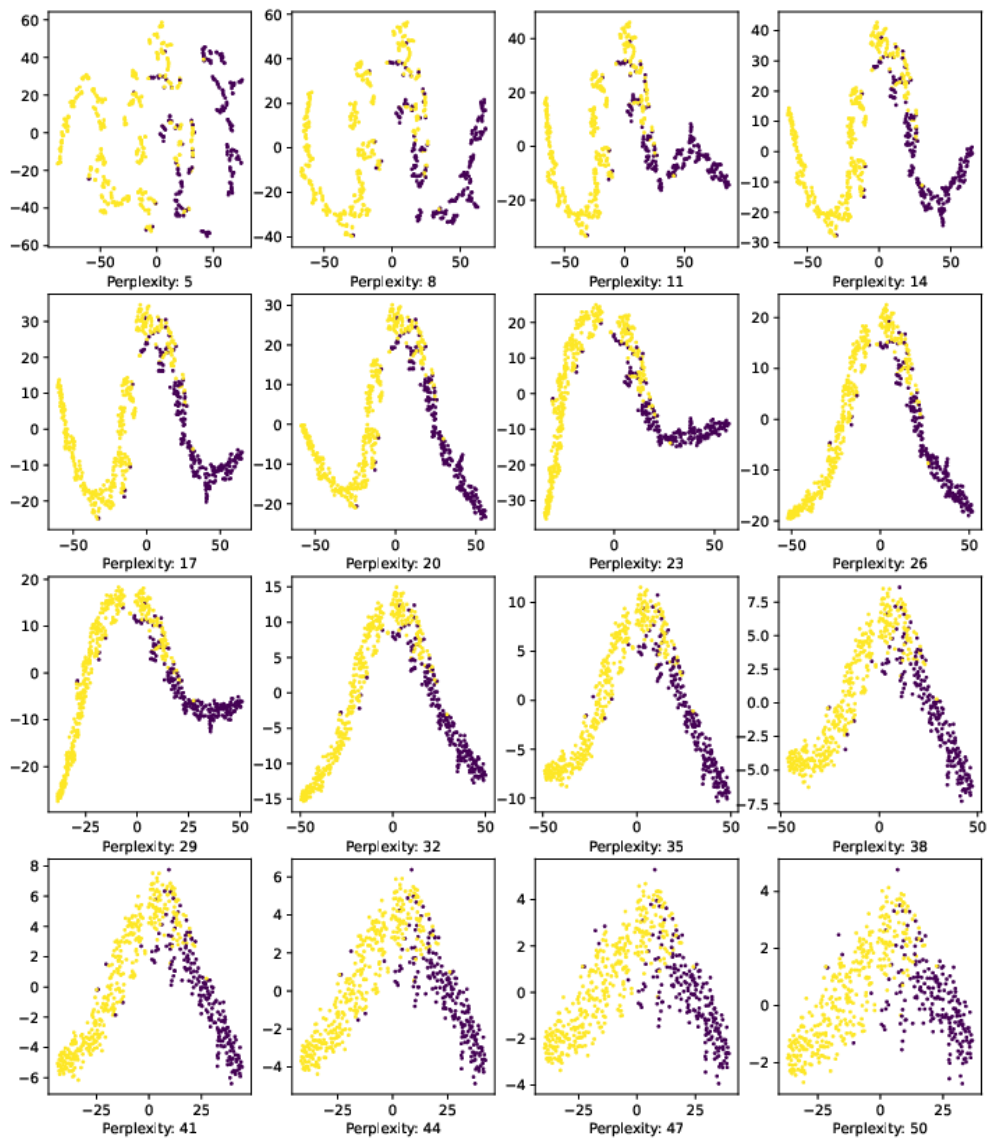


Figure A.II.1: Embeddings produced by t-SNE with varying perplexity parameter of *Breast Cancer* dataset.

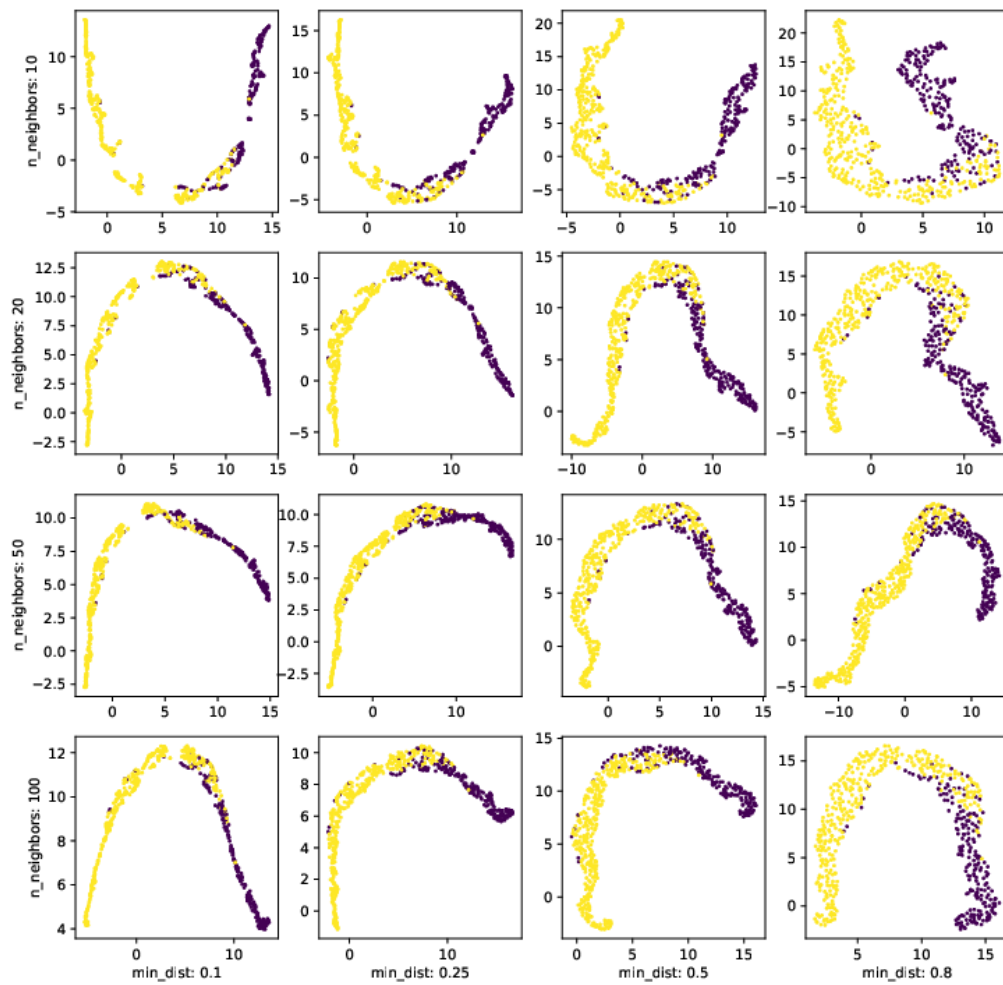


Figure A.II.2: Embeddings produced by UMAP with varying parameters $n_neighbours$ and min_dist of *Breast Cancer* dataset.

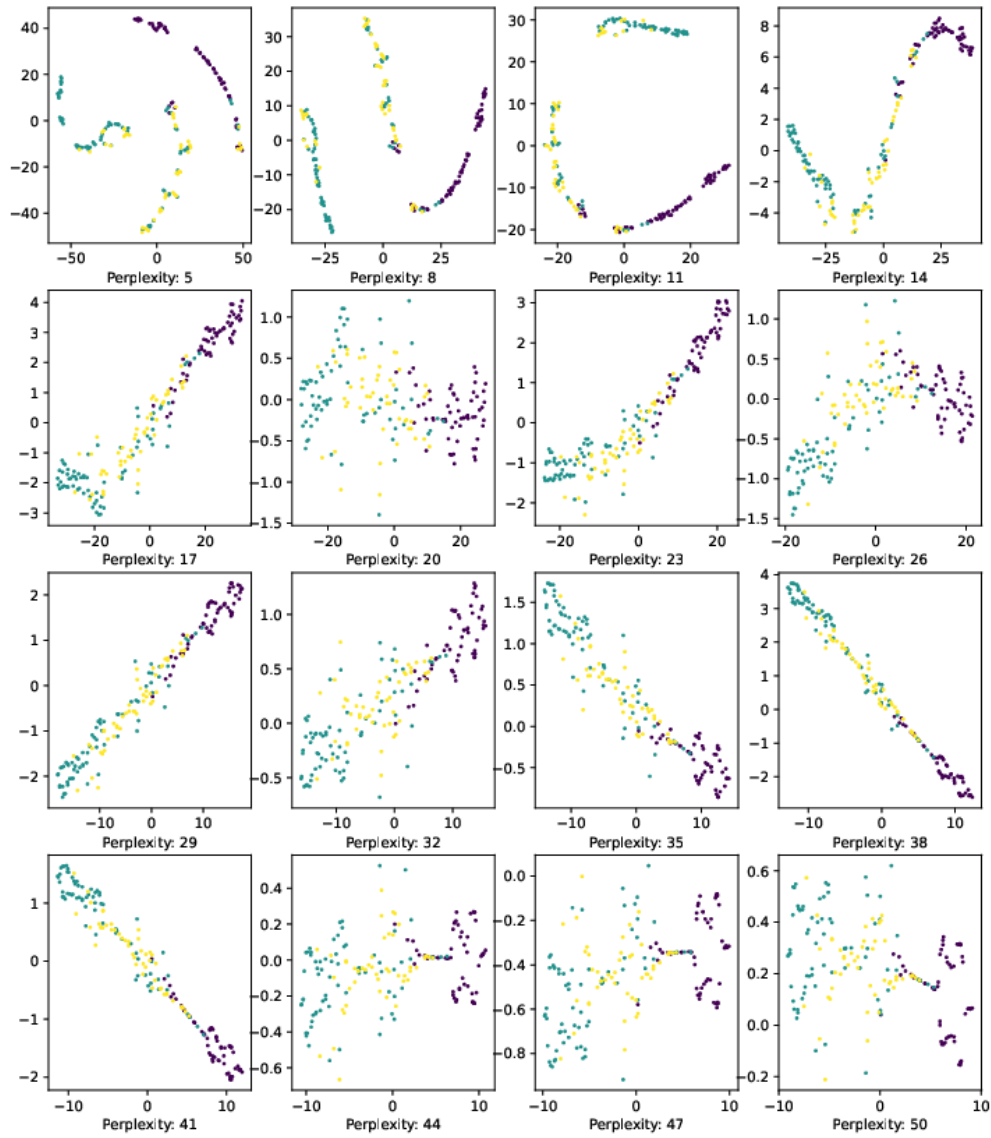


Figure A.II.3: Embeddings produced by t-SNE with varying perplexity parameter of *Wine* dataset.

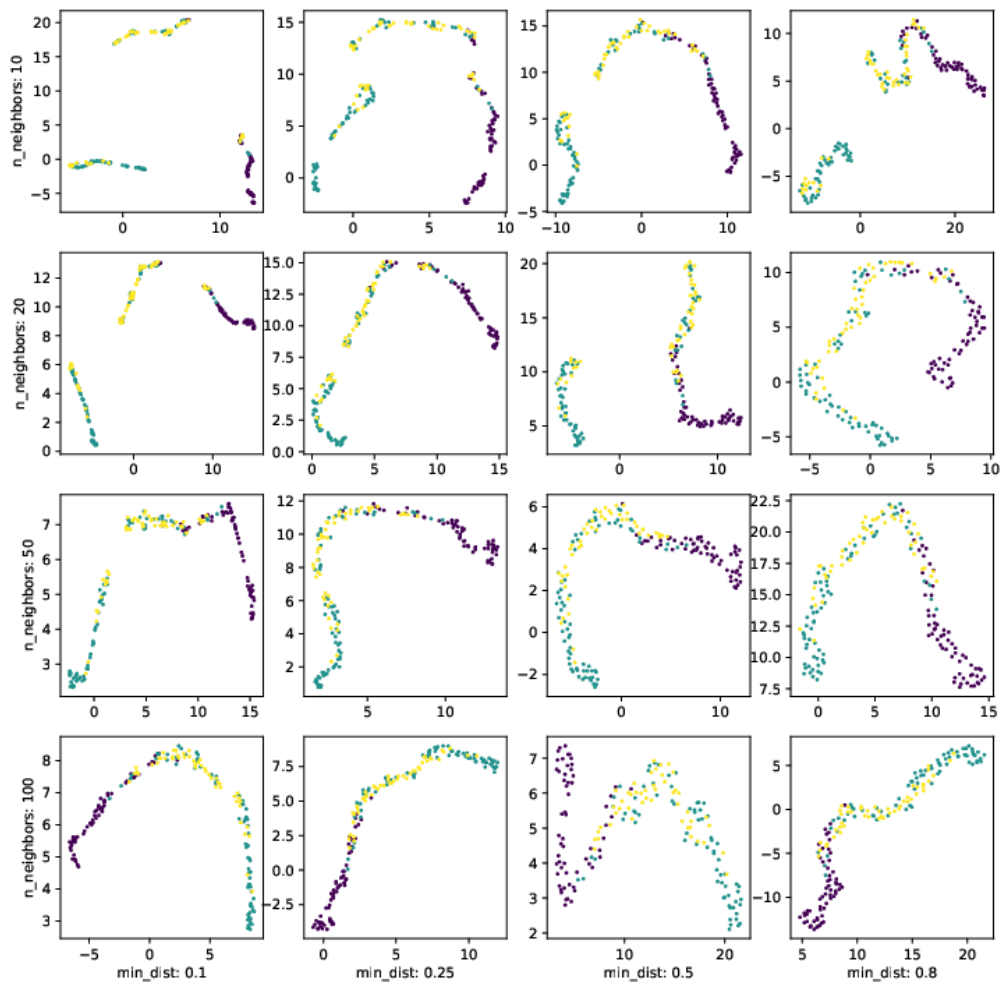


Figure A.II.4: Embeddings produced by UMAP with varying parameters `n_neighbours` and `min_dist` of *Wine* dataset.

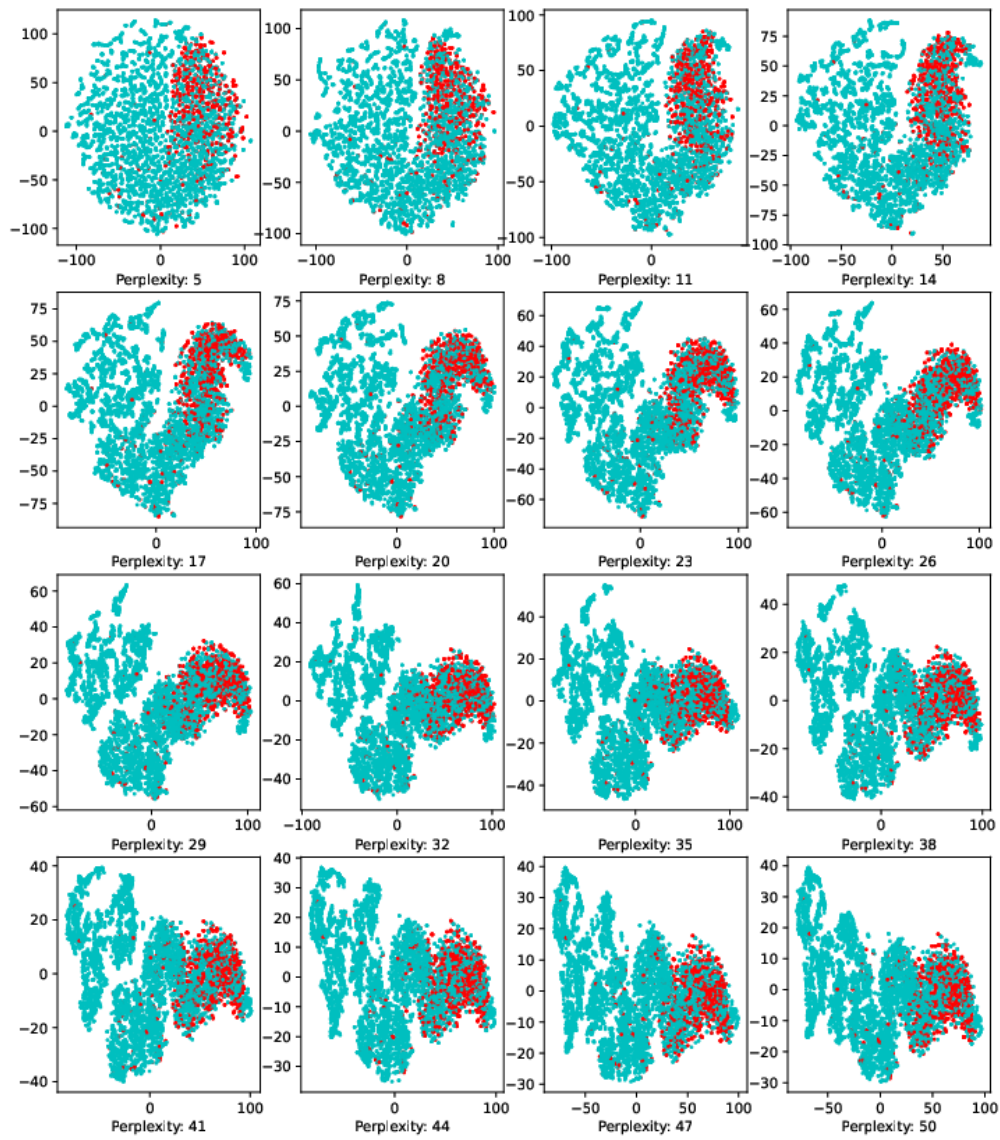


Figure A.II.5: Embeddings produced by t-SNE with varying perplexity parameter of *Spam* dataset.

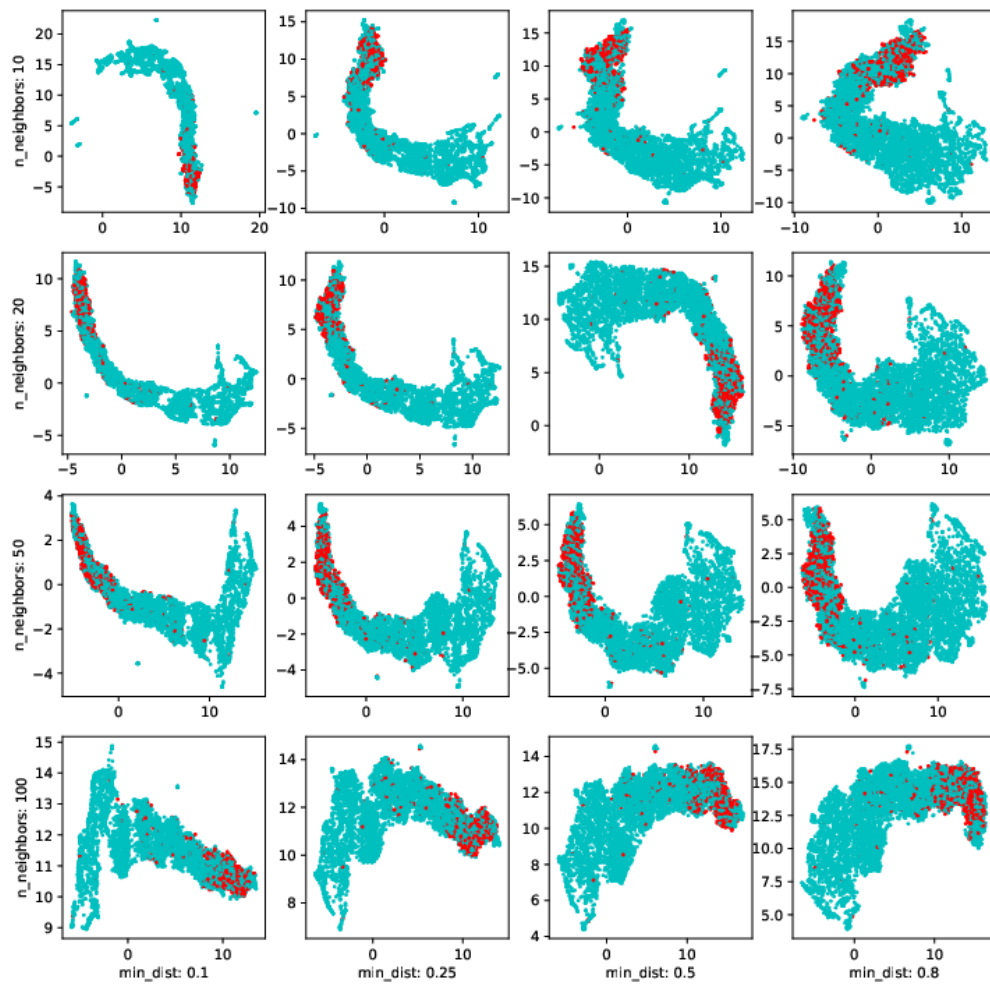


Figure A.II.6: Embeddings produced by UMAP with varying parameters `n_neighbours` and `min_dist` of *Spam* dataset.

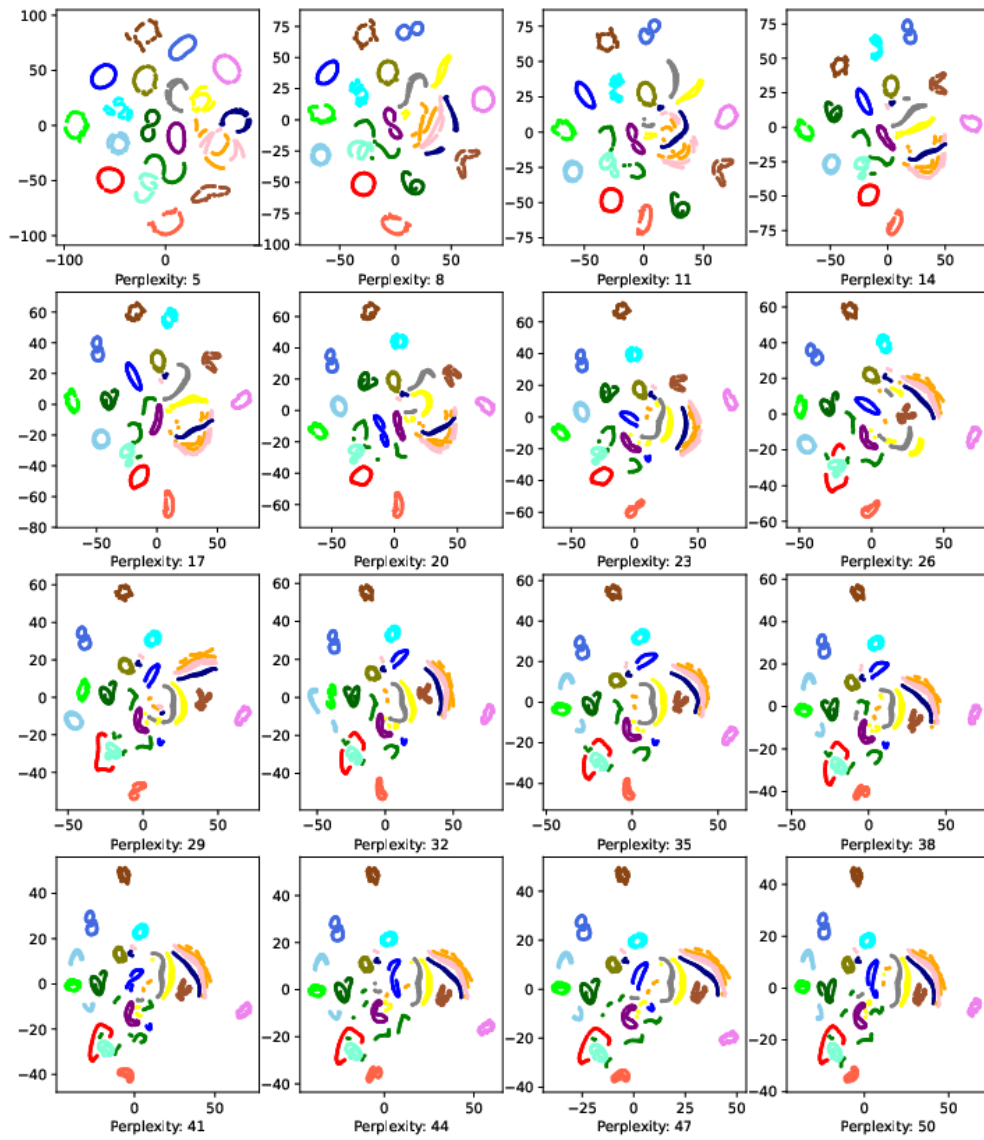


Figure A.II.7: Embeddings produced by t-SNE with varying perplexity parameter of *Coil20* dataset.

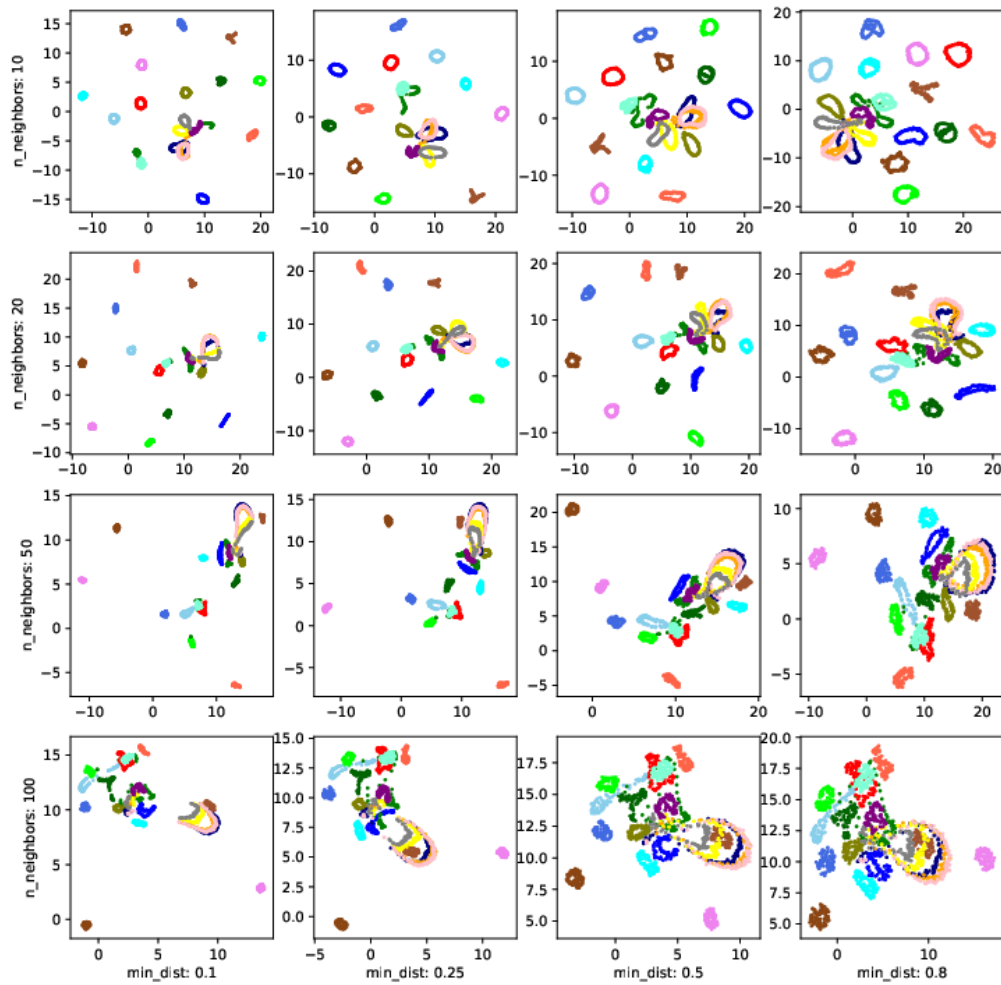


Figure A.II.8: Embeddings produced by UMAP with varying parameters `n_neighbours` and `min_dist` of *Coil20* dataset.

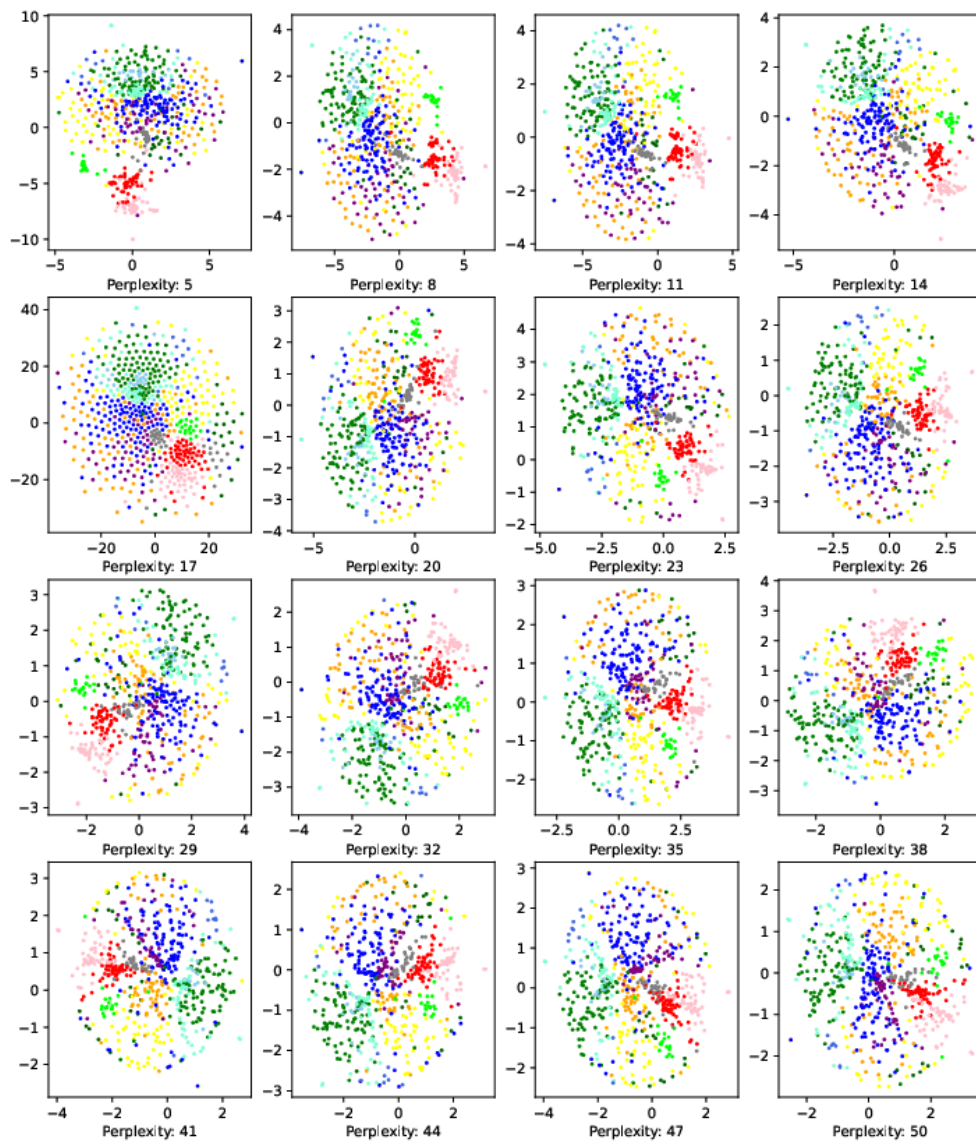


Figure A.II.9: Embeddings produced by t-SNE with varying perplexity parameter of *Usoskin* dataset.

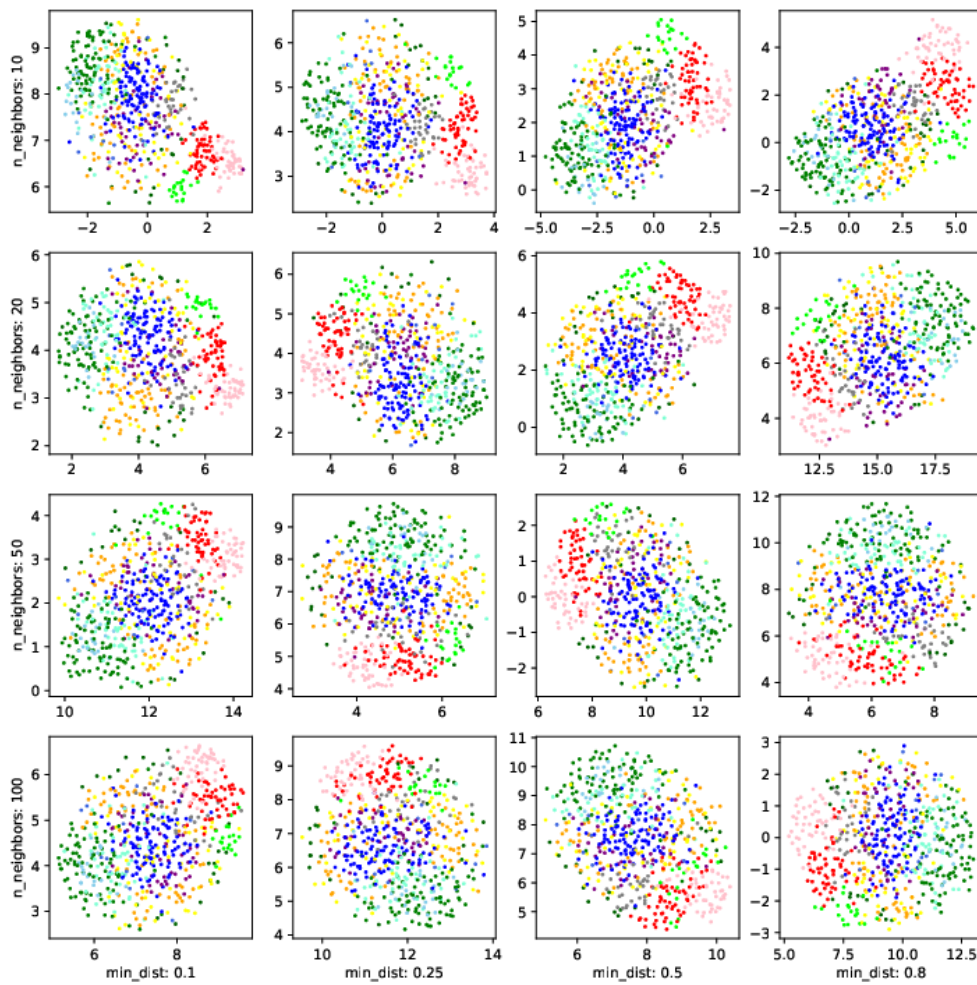


Figure A.II.10: Embeddings produced by UMAP with varying parameters `n_neighbours` and `min_dist` of *Usoskin* dataset.

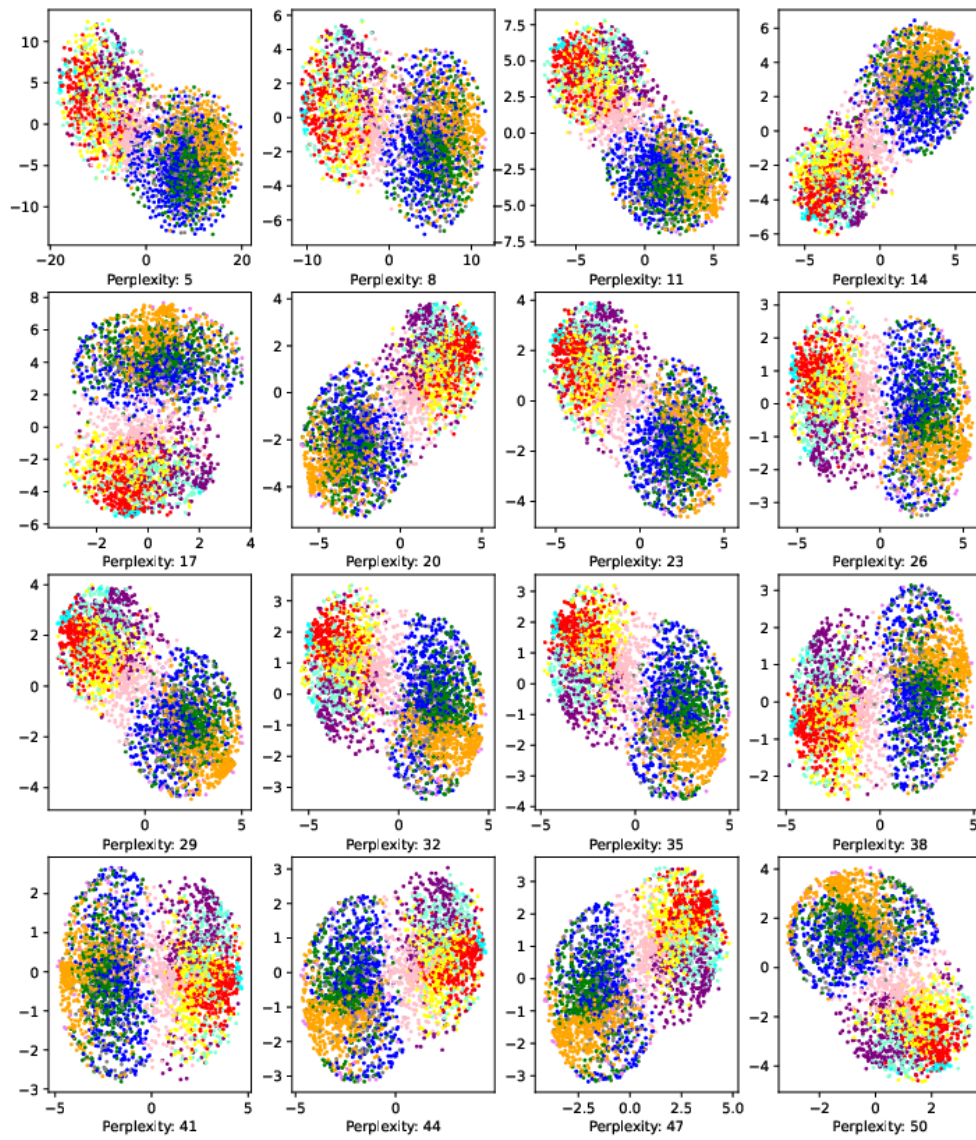


Figure A.II.11: Embeddings produced by t-SNE with varying perplexity parameter of *Jurkat* dataset.

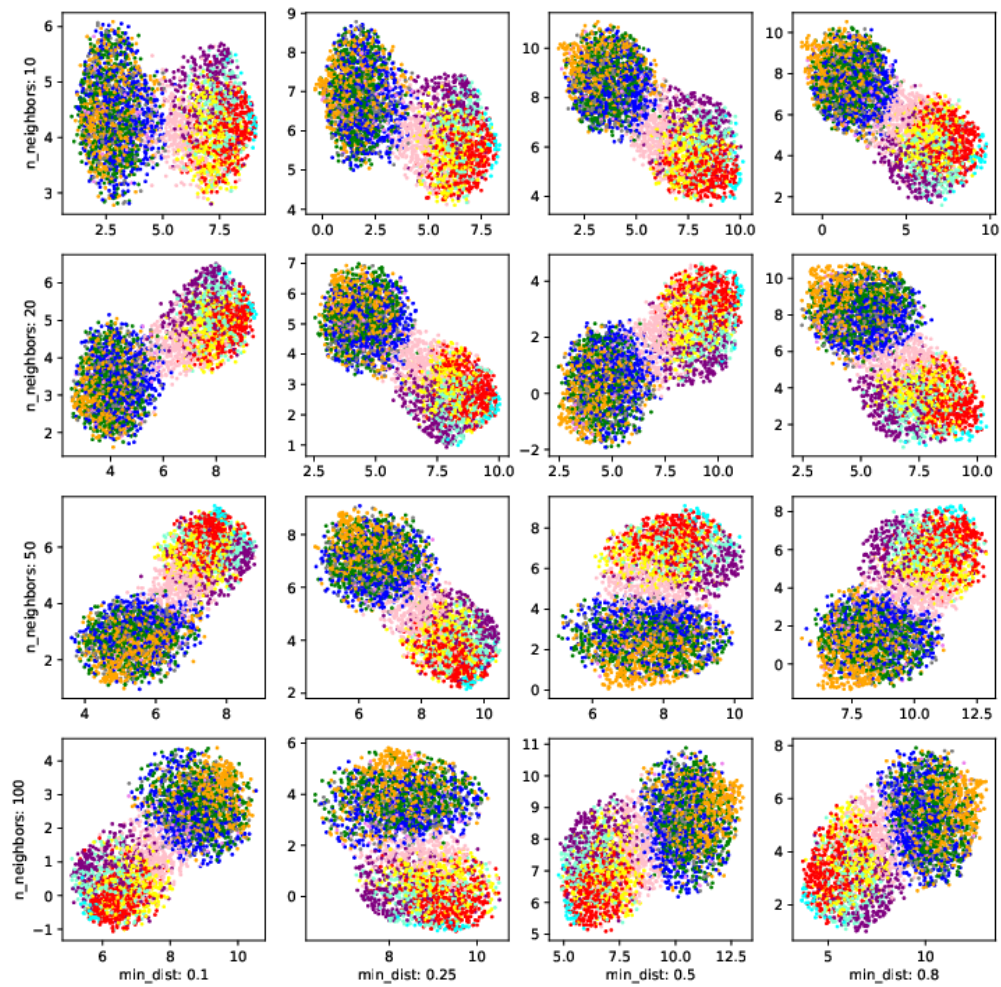
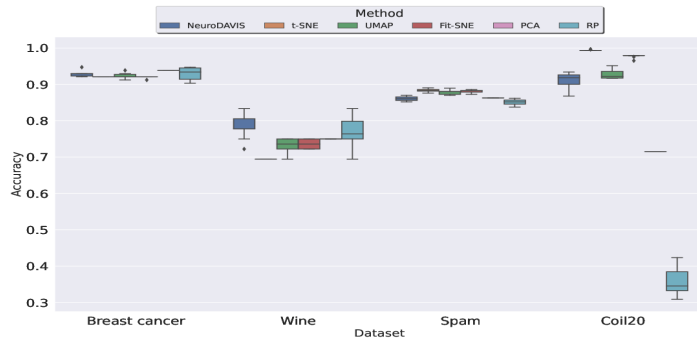
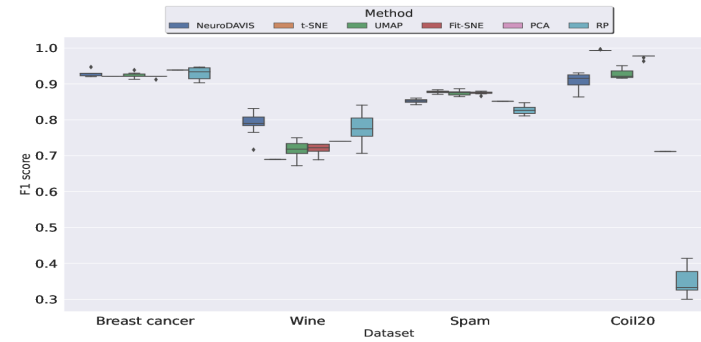


Figure A.II.12: Embeddings produced by UMAP with varying parameters *n_neighbours* and *min_dist* of *Jurkat* dataset.

A.III. Downstream Analysis



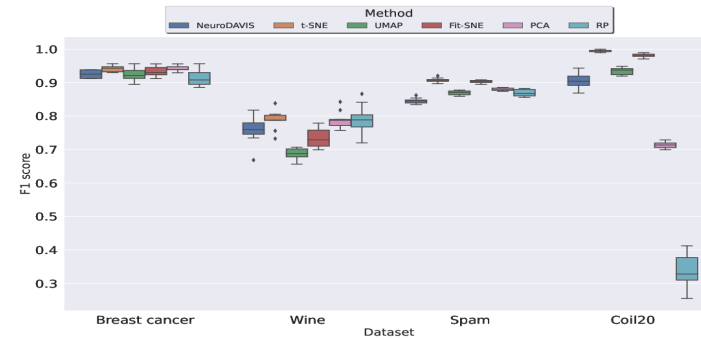
(A) Accuracy (k-NN)



(B) F1-score (k-NN)

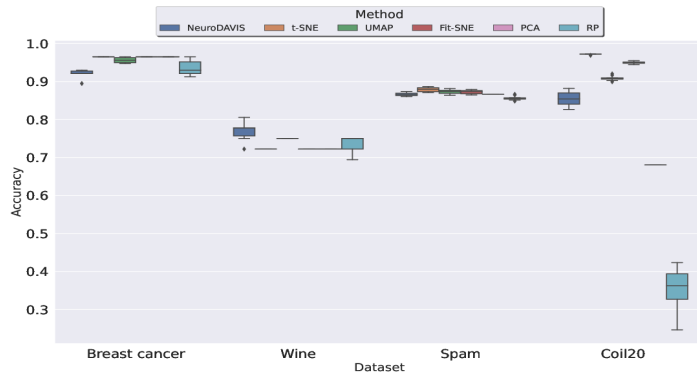


(C) Accuracy (RF)



(D) F1-score (RF)

Figure A.III.1: (A) and (B) show performance of k-NN classifier with $k = 5$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of *Breast Cancer*, *Wine*, *Spam* and *Coil20* datasets, in terms of Accuracy and F1-scores respectively, (C) and (D) show performance of RF classifier with $n_estimators = 20$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of *Breast Cancer*, *Wine*, *Spam* and *Coil20* datasets, in terms of Accuracy and F1-scores respectively.



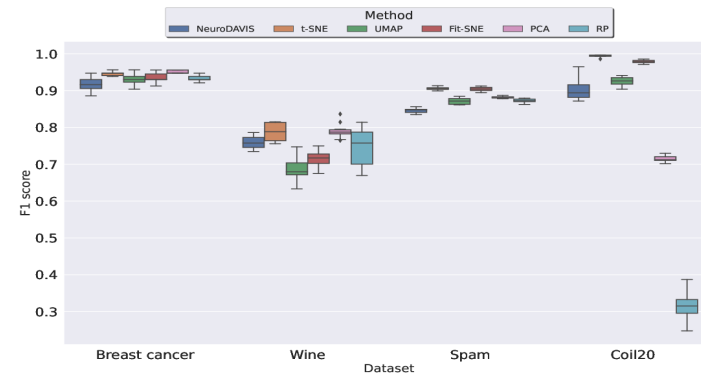
(A) Accuracy (k-NN)



(B) F1-score (k-NN)

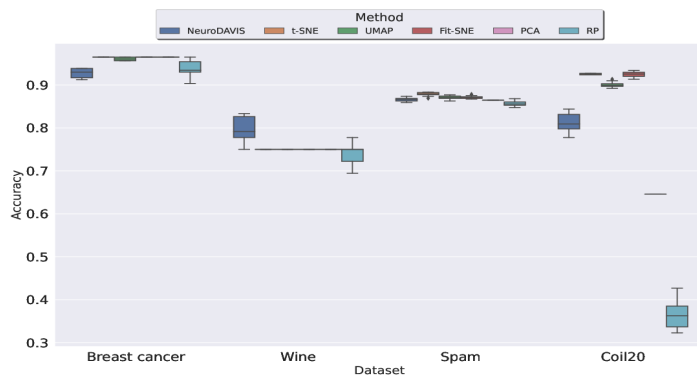


(C) Accuracy (RF)

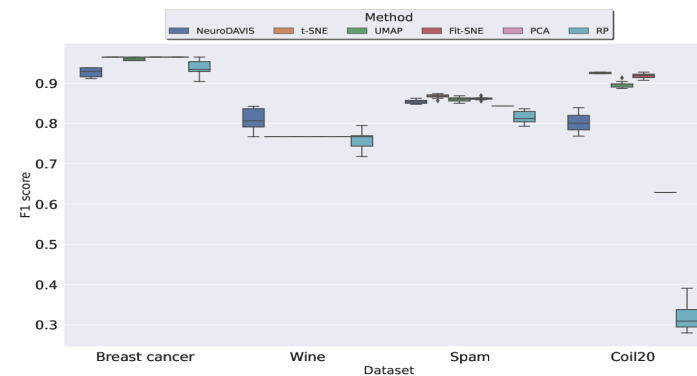


(D) F1-score (RF)

Figure A.III.2: (A) and (B) show performance of k-NN classifier with $k = 15$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of *Breast Cancer*, *Wine*, *Spam* and *Coil20* datasets, in terms of Accuracy and F1-scores respectively, (C) and (D) show performance of RF classifier with $n_estimators = 40$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of *Breast Cancer*, *Wine*, *Spam* and *Coil20* datasets, in terms of Accuracy and F1-scores respectively.



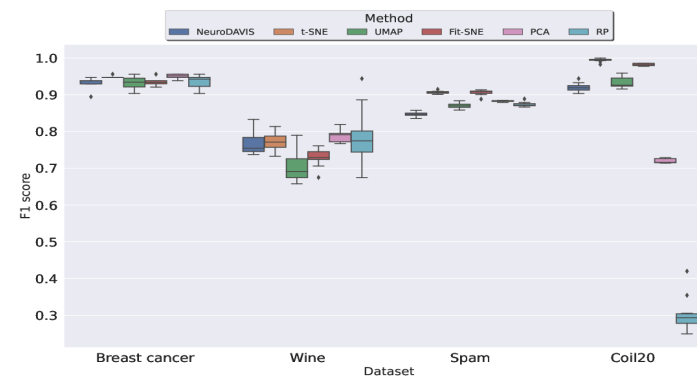
(A) Accuracy (k-NN)



(B) F1-score (k-NN)

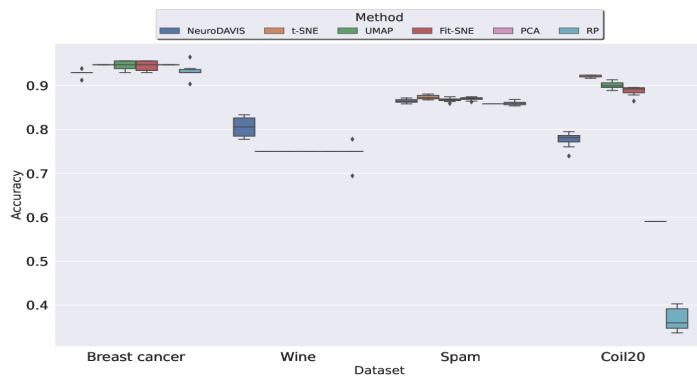


(C) Accuracy (RF)

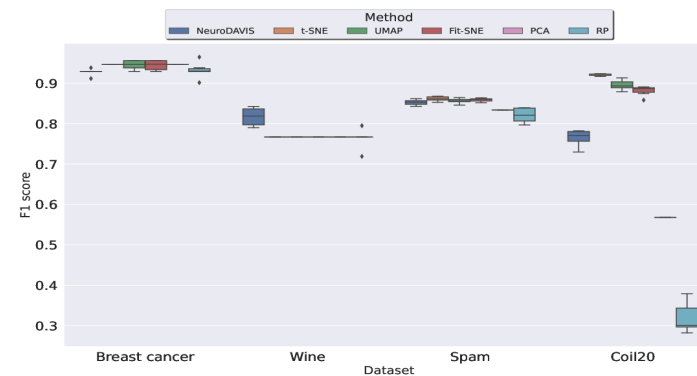


(D) F1-score (RF)

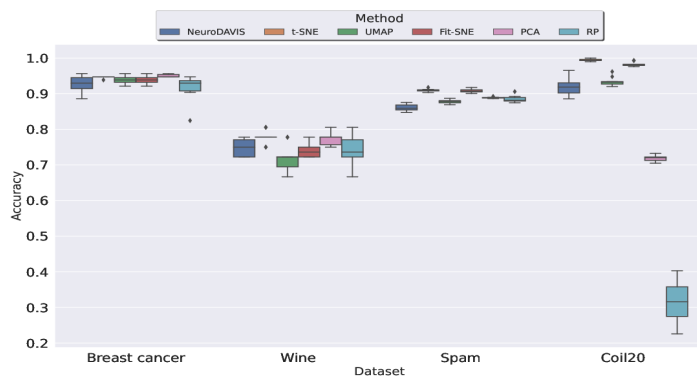
Figure A.III.3: (A) and (B) show performance of k-NN classifier with $k = 25$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of *Breast Cancer*, *Wine*, *Spam* and *Coil20* datasets, in terms of Accuracy and F1-scores respectively, (C) and (D) show performance of RF classifier with $n_estimators = 60$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of *Breast Cancer*, *Wine*, *Spam* and *Coil20* datasets, in terms of Accuracy and F1-scores respectively.



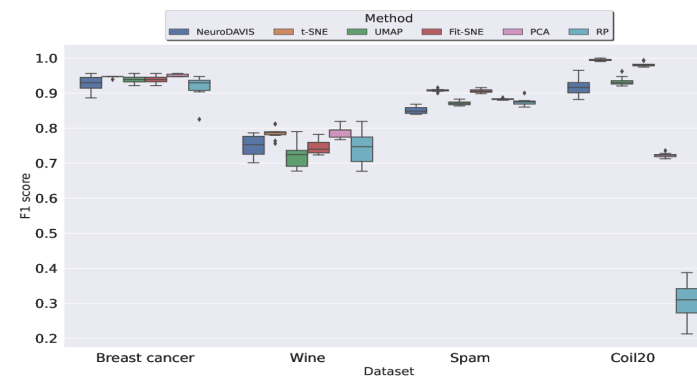
(A) Accuracy (k-NN)



(B) F1-score (k-NN)

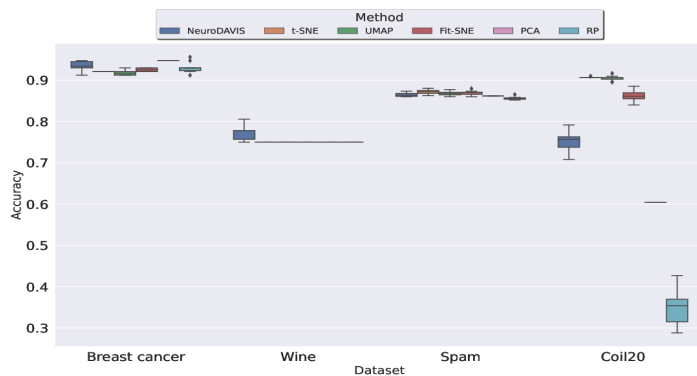


(C) Accuracy (RF)

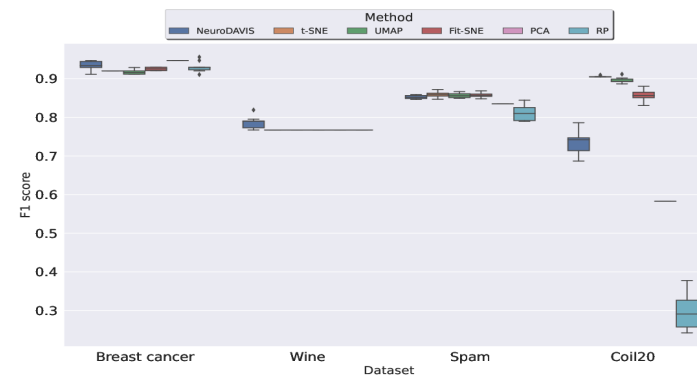


(D) F1-score (RF)

Figure A.III.4: (A) and (B) show performance of k-NN classifier with $k = 35$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of *Breast Cancer*, *Wine*, *Spam* and *Coil20* datasets, in terms of Accuracy and F1-scores respectively, (C) and (D) show performance of RF classifier with $n_estimators = 80$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of *Breast Cancer*, *Wine*, *Spam* and *Coil20* datasets, in terms of Accuracy and F1-scores respectively.



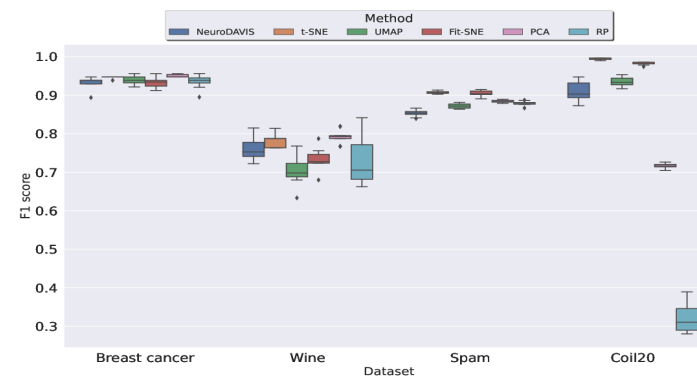
(A) Accuracy (k-NN)



(B) F1-score (k-NN)

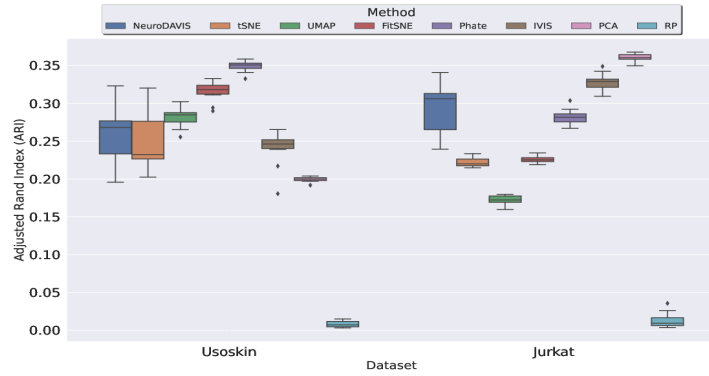


(C) Accuracy (RF)

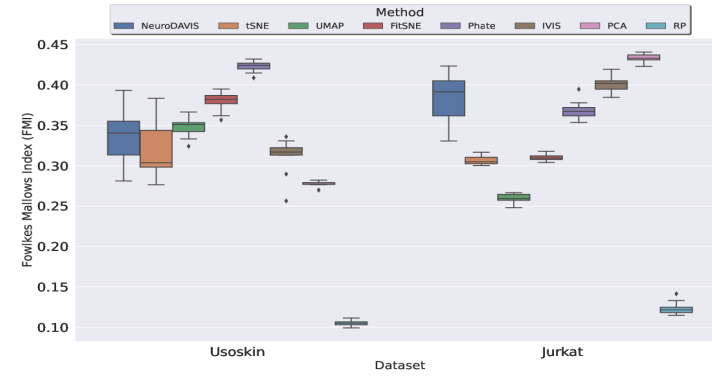


(D) F1-score (RF)

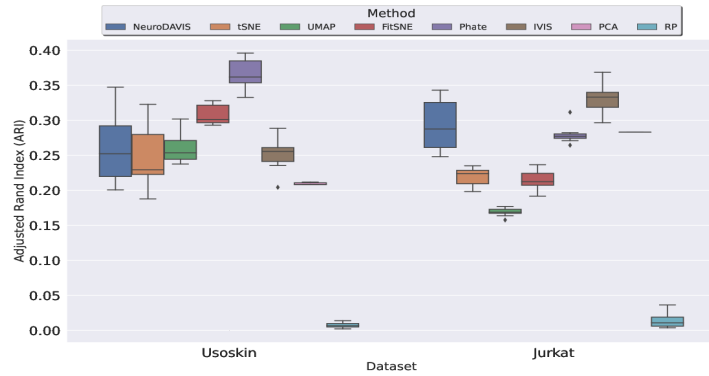
Figure A.III.5: (A) and (B) show performance of k-NN classifier with $k = 45$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of *Breast Cancer*, *Wine*, *Spam* and *Coil20* datasets, in terms of Accuracy and F1-scores respectively, (C) and (D) show performance of RF classifier with $n_estimators = 100$ on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PCA and RP-generated embeddings of *Breast Cancer*, *Wine*, *Spam* and *Coil20* datasets, in terms of Accuracy and F1-scores respectively.



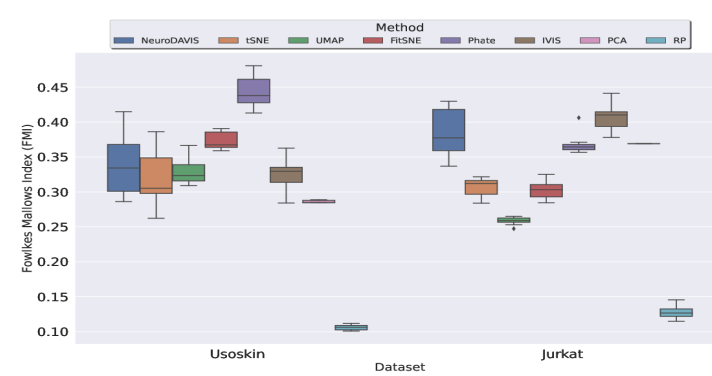
(A) Adjusted Rand Index (ARI) (k-means)



(B) Fowlkes Mallows Index (FMI) (k-means)



(C) Adjusted Rand Index (ARI) (Agglomerative clustering)



(D) Fowlkes Mallows Index (FMI) (Agglomerative clustering)

Figure A.III.6: (A) and (B) show ARI and FMI scores for cell-type clustering using k-means on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PHATE, IVIS, PCA and RP-generated embeddings of *Usoskin* and *Jurkat* datasets, (C) and (D) show ARI and FMI scores for cell-type clustering using agglomerative hierarchical clustering on NeuroDAVIS, t-SNE, UMAP, Fit-SNE, PHATE, IVIS, PCA and RP-generated embeddings of *Usoskin* and *Jurkat* datasets.

A.IV. Projection of test samples

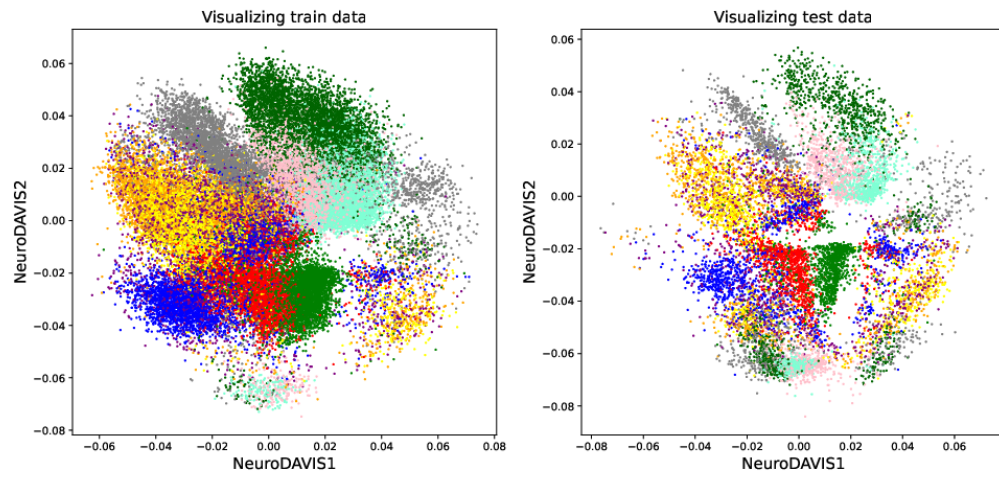


Figure A.IV.1: Embeddings produced by NeuroDAVIS on both the training and test parts of *FMNIST* dataset.

B. NeuroMDAVIS-FS

B.I. Results on LSCC cohort

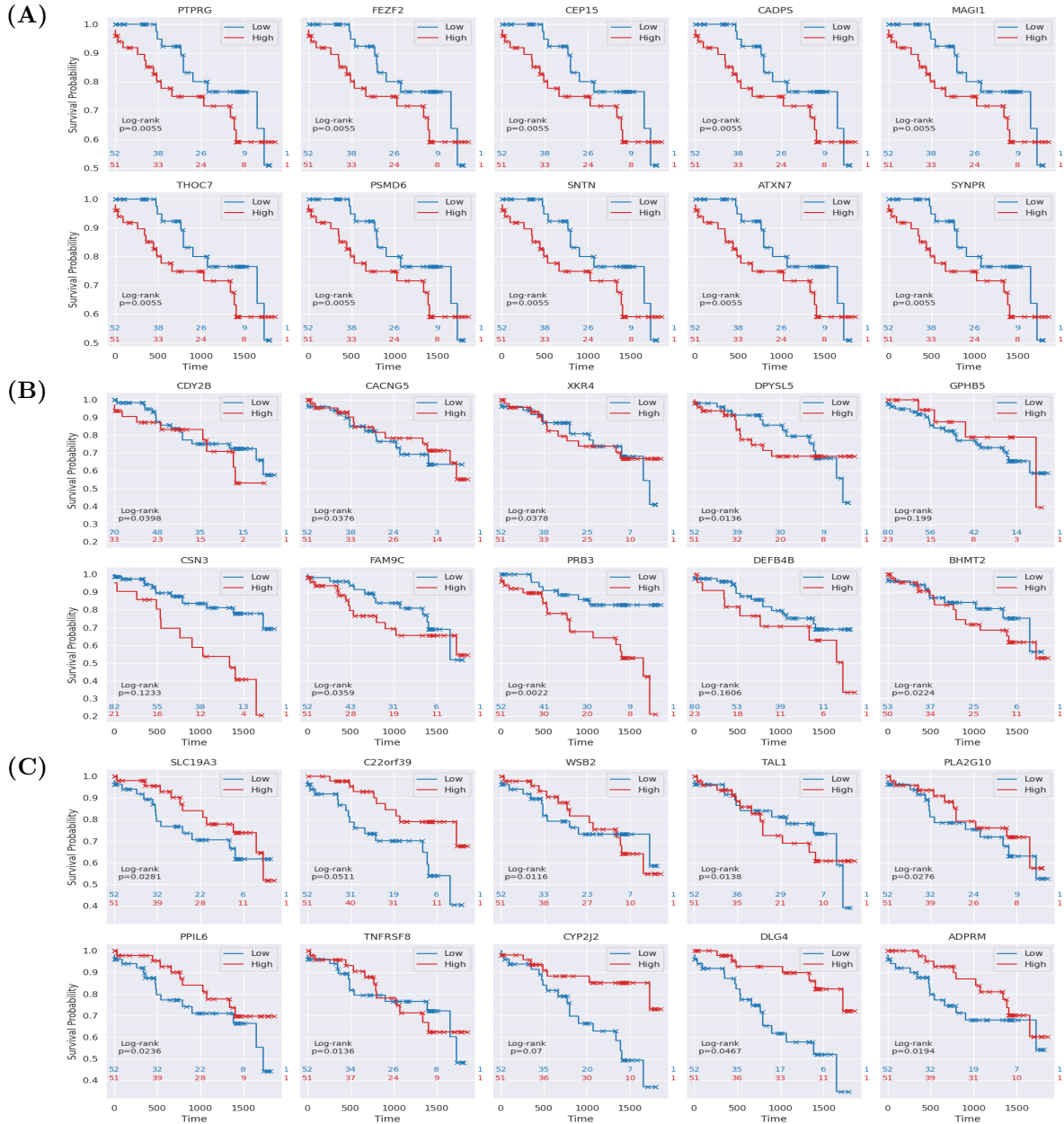


Figure B.I.1: Kaplan-Meier curves based on high- and low-expression values of each of the top features selected by NeuroMDAVIS-FS from (A) Protein, (B) RNA, and (C) CNV omics modalities, while applied on the LSCC cohort. Patients have been stratified into high and low-expression groups based on the median value of each feature. Log-rank tests have been used to assess statistical significance in survival differences.

Table B.1: Top 15 features selected by NeuroMDAVIS-FS from each of the omics modalities, viz., Protein, RNA, and CNV. Here only LSCC samples utilized to train NeuroMDAVIS.

Rank	Protein	RNA	CNV
1	SLC19A3	CDY2B	PTPRG
2	C22orf39	CACNG5	FEZF2
3	WSB2	XKR4	CEP15
4	TAL1	DPYSL5	CADPS
5	PLA2G10	GPHB5	MAGI1
6	PPIL6	CSN3	THOC7
7	TNFRSF8	FAM9C	PSMD6
8	CYP2J2	PRB3	SNTN
9	DLG4	DEFB4B	ATXN7
10	ADPRM	BHMT2	SYNPR
11	SLC16A4	DCAF4L2	PRICKLE2
12	DLL4	HSD17B13	C3orf49
13	GMCL1	TUBA3C	ADAMTS9
14	ZNF611	NOBOX	SLC25A26
15	CALCA	SST	FRG2

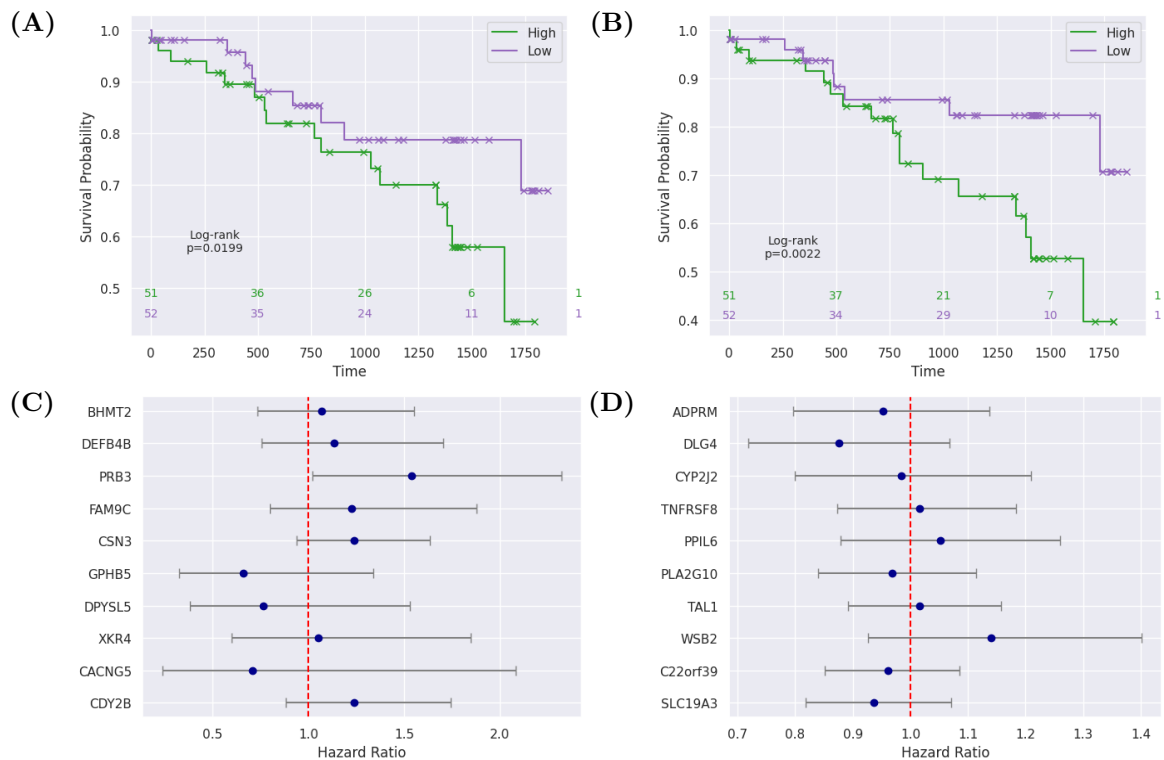


Figure B.I.2: Kaplan-Meier curve based on high and low risk scores obtained from a CoxPH regression applied using (A) top RNAs, and (B) top proteins selected by NeuroMDAVIS-FS. Hazard scores of (C) top RNAs, and (D) top proteins selected by NeuroMDAVIS-FS

B.II. Results on LUAD cohort

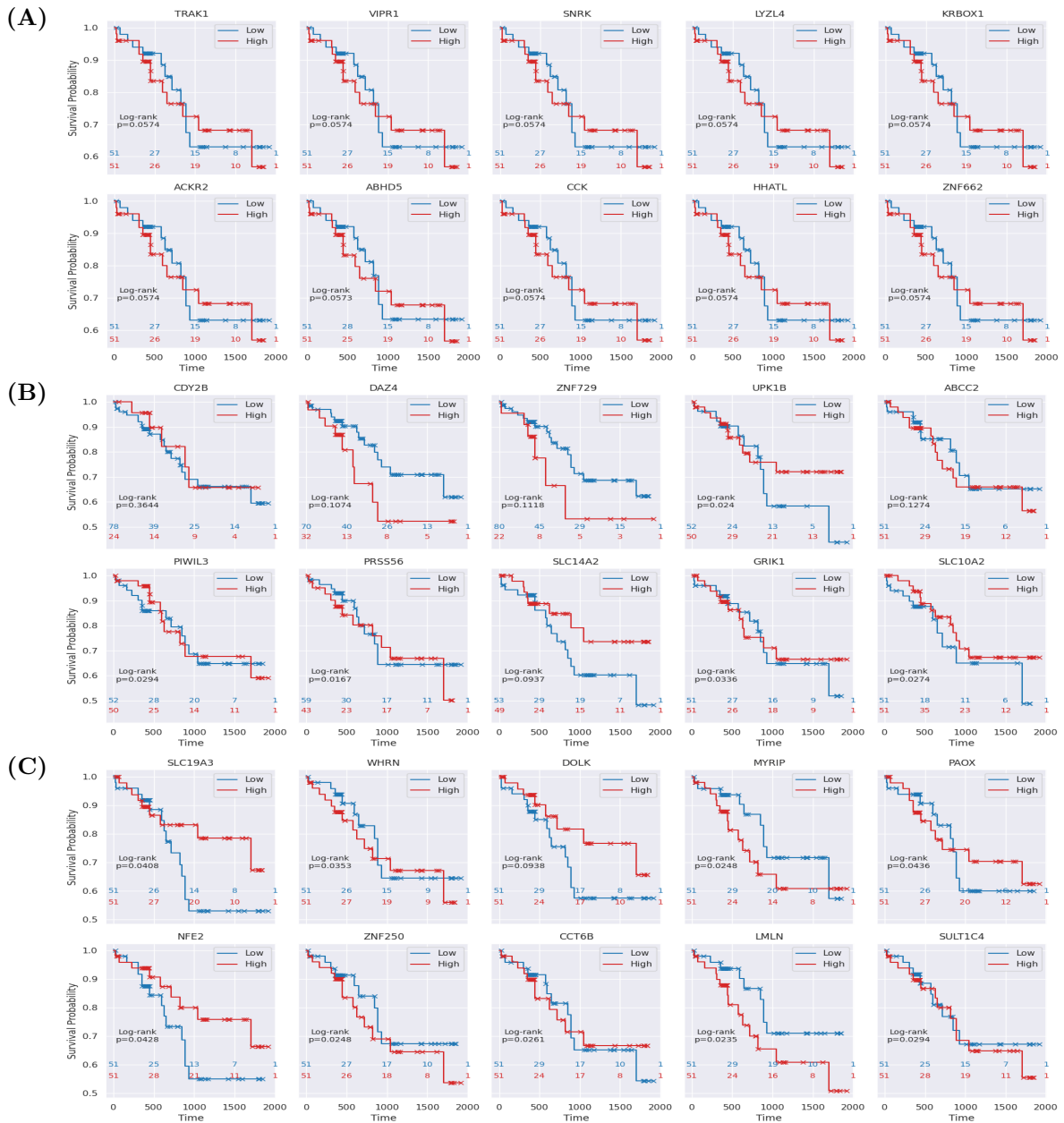


Figure B.II.1: Kaplan-Meier curves based on high- and low-expression values of each of the top features selected by NeuroMDAVIS-FS from (A) Protein, (B) RNA, and (C) CNV omics modalities, while applied on the LUAD cohort. Patients have been stratified into high and low-expression groups based on the median value of each feature. Log-rank tests have been used to assess statistical significance in survival differences.

Table B.2: Top 15 features selected by NeuroMDAVIS-FS from each of the omics modalities, viz., Protein, RNA, and CNV. Here only LUAD samples utilized to train NeuroMDAVIS.

Rank	Protein	RNA	CNV
1	SLC19A3	CDY2B	TRAK1
2	WHRN	DAZ4	VIPR1
3	DOLK	ZNF729	SNRK
4	MYRIP	UPK1B	LYZL4
5	PAOX	ABCC2	KRBOX1
6	NFE2	PIWIL3	ACKR2
7	ZNF250	PRSS56	ABHD5
8	CCT6B	SLC14A2	CCK
9	LMLN	GRIK1	HHATL
10	SULT1C4	SLC10A2	ZNF662
11	WSB2	ZFP42	POMGNT2
12	CXCR6	DSG3	HIGD1A
13	TAL1	SLC5A5	ZBTB47
14	ATXN7L2	NR0B1	ANO10
15	GRAMD1C	MAGEA10	CYP8B1

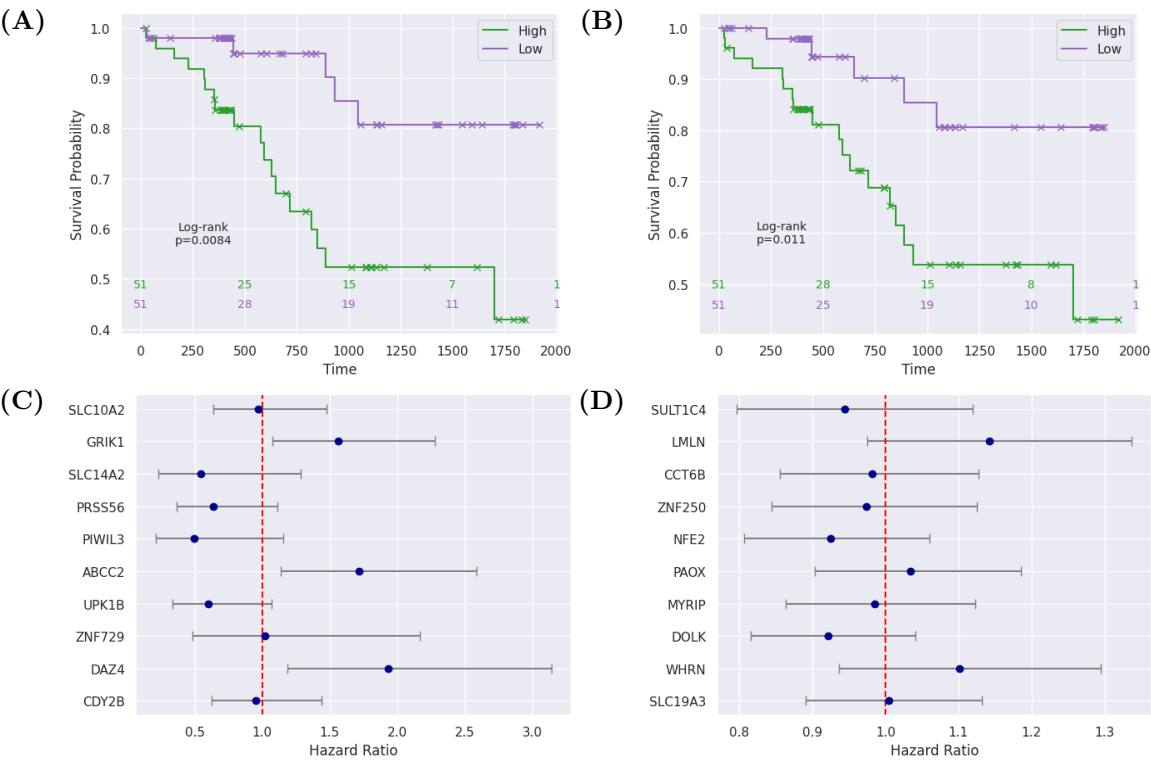


Figure B.II.2: Kaplan-Meier curve based on high and low risk scores obtained from a CoxPH regression applied using (A) top RNAs, and (B) top proteins selected by NeuroMDAVIS-FS. Hazard scores of (C) top RNAs, and (D) top proteins selected by NeuroMDAVIS-FS

Table B.3: Multi-omics features and their functional and clinical relevance in lung cancer

Modality	Gene	Lung Cancer Context & Functional Explanation	Clinical Association (Prognosis / Target)	Source
CNV	VIPR1	Vasoactive intestinal peptide receptor 1; functions as a tumor suppressor in lung adenocarcinoma (LUAD).	Low expression correlates with poor overall survival; inhibits tumor migration and invasion.	[120]
CNV	PTPRG	Protein tyrosine phosphatase gamma; tumor suppressor located at chromosome 3p14.2.	Loss of expression due to LOH or promoter methylation is common in NSCLC.	[122]
CNV	ACKR2	Atypical chemokine receptor that regulates the tumor microenvironment by scavenging inflammatory chemokines.	Associated with immune cell infiltration and patient survival across solid tumors.	[128]
CNV	LIMD1	Established tumor suppressor located at chromosome 3p21.3.	Frequent copy number loss in LUAD; modulates sensitivity to kinase inhibitors.	[121]
RNA	ABCC2	ATP-binding cassette transporter (MRP2) involved in drug efflux and multidrug resistance.	Overexpression predicts resistance to cisplatin in NSCLC.	[129]
RNA	PIWIL2/3	PIWI-like proteins involved in germline-associated stemness and RNA silencing mechanisms.	Associated with increased proliferation and poor prognosis in LUAD.	[137]
RNA	LTF	Lactotransferrin; tumor suppressor with immunomodulatory functions.	Predictive biomarker for benefit from chemo-immunotherapy in SCLC.	[127]
Protein	TNFRSF8	CD30; member of the TNF receptor superfamily involved in immune signaling.	Signature component for predicting prognosis and chemotherapy benefit in SCLC.	[138]
Protein	KRT13	Keratin 13; marker of squamous epithelial differentiation.	Specifically dysregulated in LSCC; used as a histological subtype marker.	[139]
Protein	TAL1	Transcription factor involved in cell fate determination and transcriptional regulation.	Dysregulation linked to aggressive clinical behavior in multiple cancers.	[140]
CNV	CCR9	A Promotes metastasis and aggressive disease progression via the PI3K/Akt signaling pathway.	High expression correlates with advanced tumor stage.	[123]
CNV	CXCR6	Facilitates tumor cell migration and invasion through matrix metalloproteinase activation.	Serves as a biomarker for response to anti-PD-(L)1 immunotherapy.	[126]
CNV	CCR1	High expression associated with metastatic phenotype and increased invasiveness.	Identified as a potential target for novel small-molecule antagonists.	[124]
CNV	CCR3	High expression associated with metastatic phenotype and increased invasiveness.	Identified as a potential target for novel small-molecule antagonists.	[125]

C. G-NeuroDAVIS

C.I. Embedding mapping

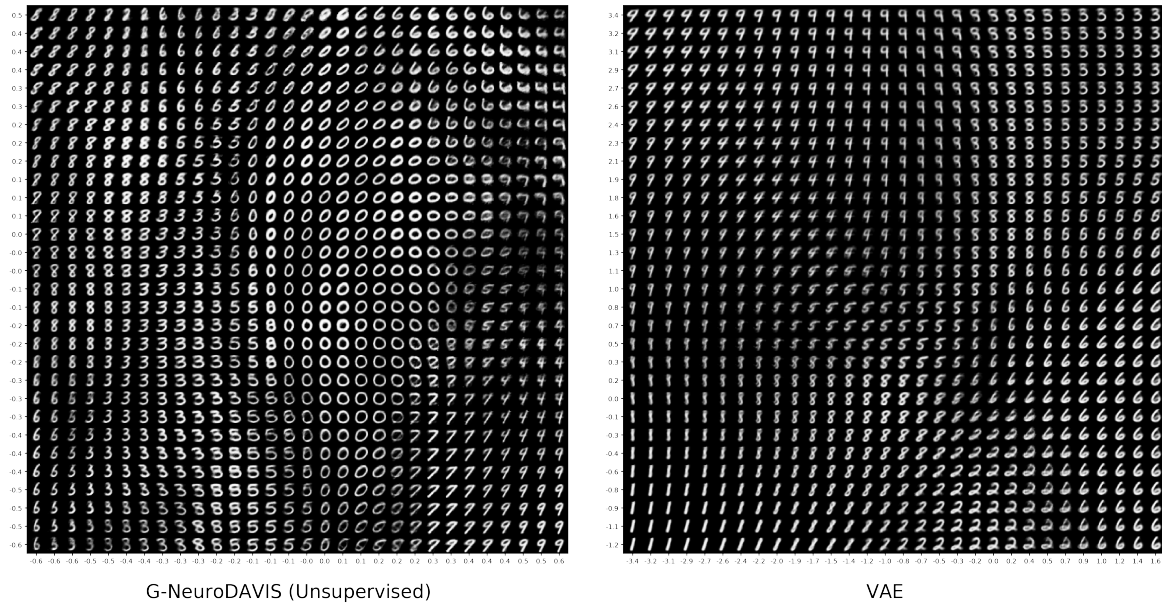


Figure C.I.1: Mapping from 2-dimensional embedding, produced by G-NeuroDAVIS and VAE, in an unsupervised fashion, to the original high-dimensional space for *MNIST* dataset.

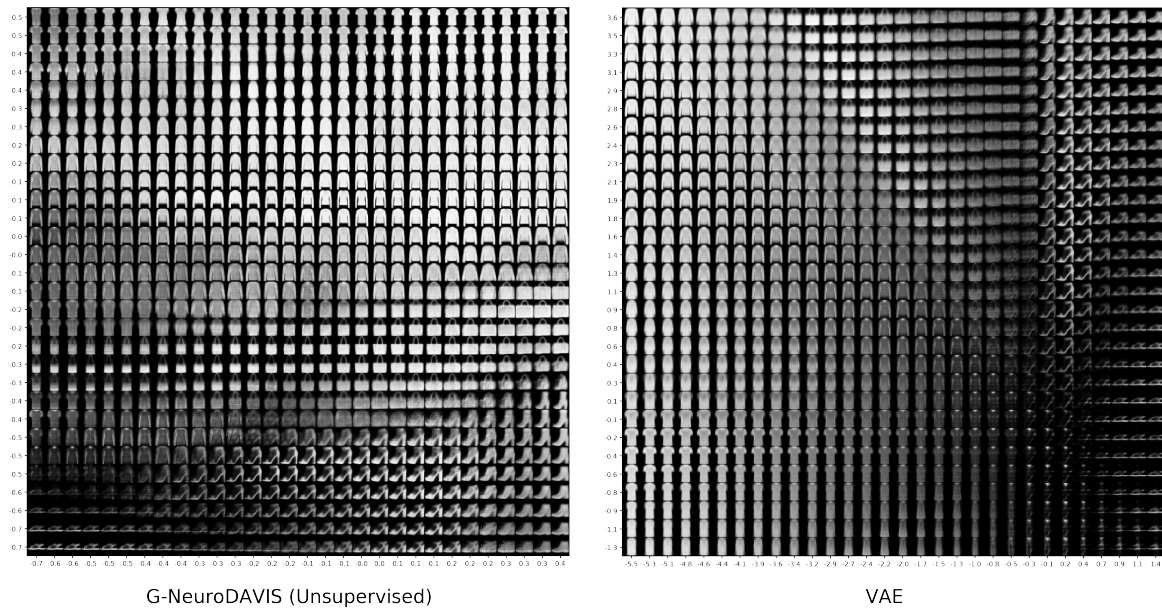


Figure C.I.2: Mapping from 2-dimensional embedding, produced by VAE, to the original high-dimensional space for *FMNIST* dataset.

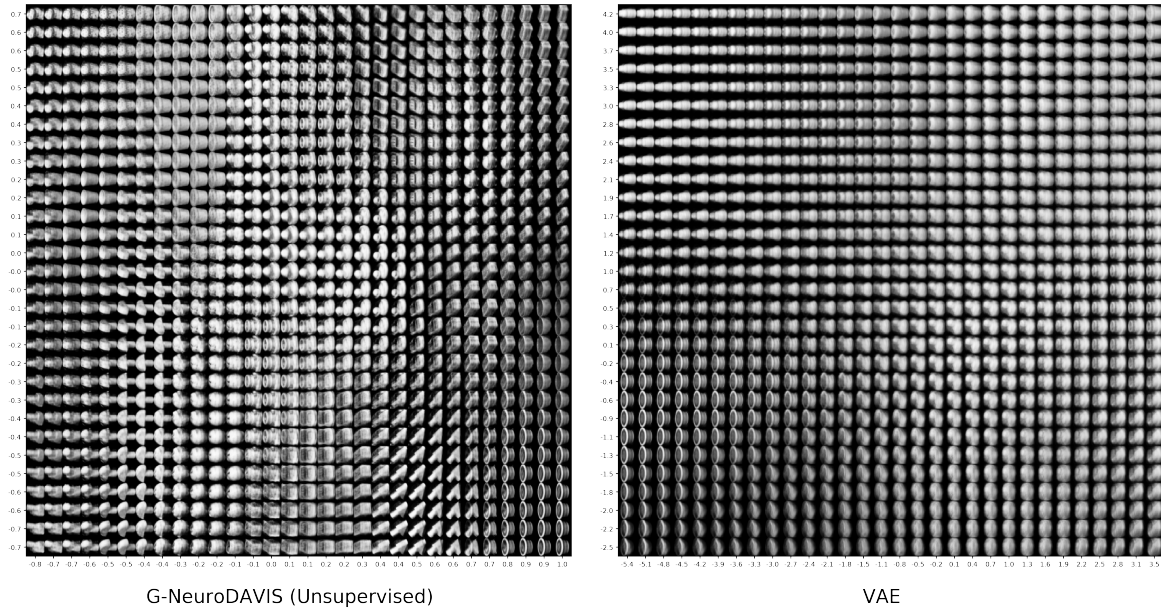


Figure C.I.3: Mapping from 2-dimensional embedding, produced by G-NeuroDAVIS, in an unsupervised fashion, to the original high-dimensional space for *Coil20* dataset.

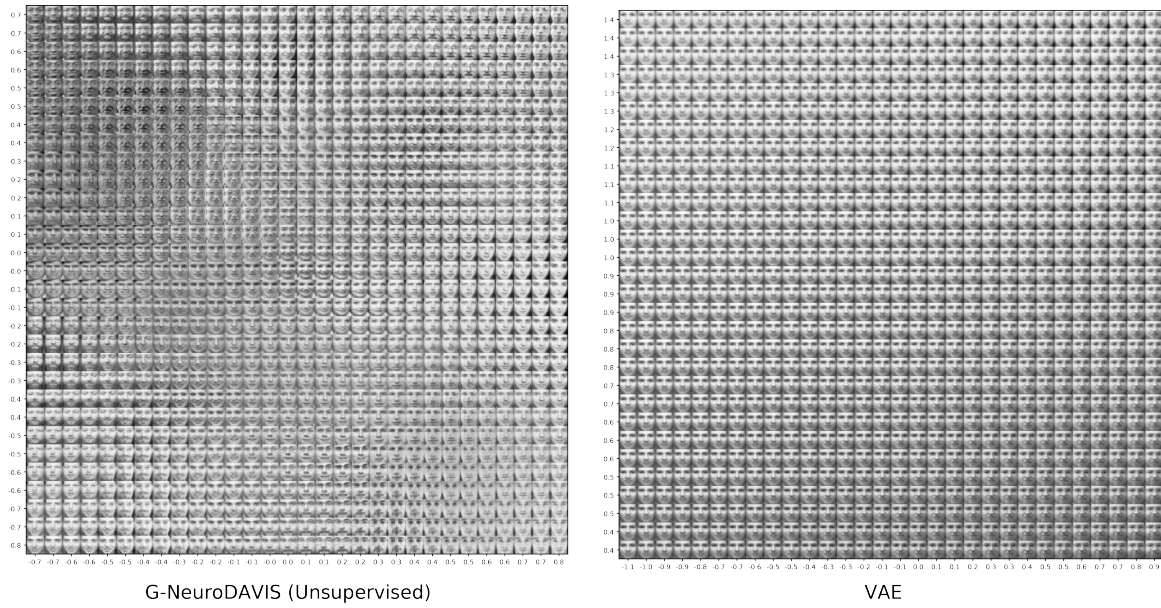


Figure C.I.4: Mapping from 2-dimensional embedding, produced by G-NeuroDAVIS, in an unsupervised fashion, to the original high-dimensional space for *Olivetti Faces* dataset.

C.II. Sub-sample based analyses

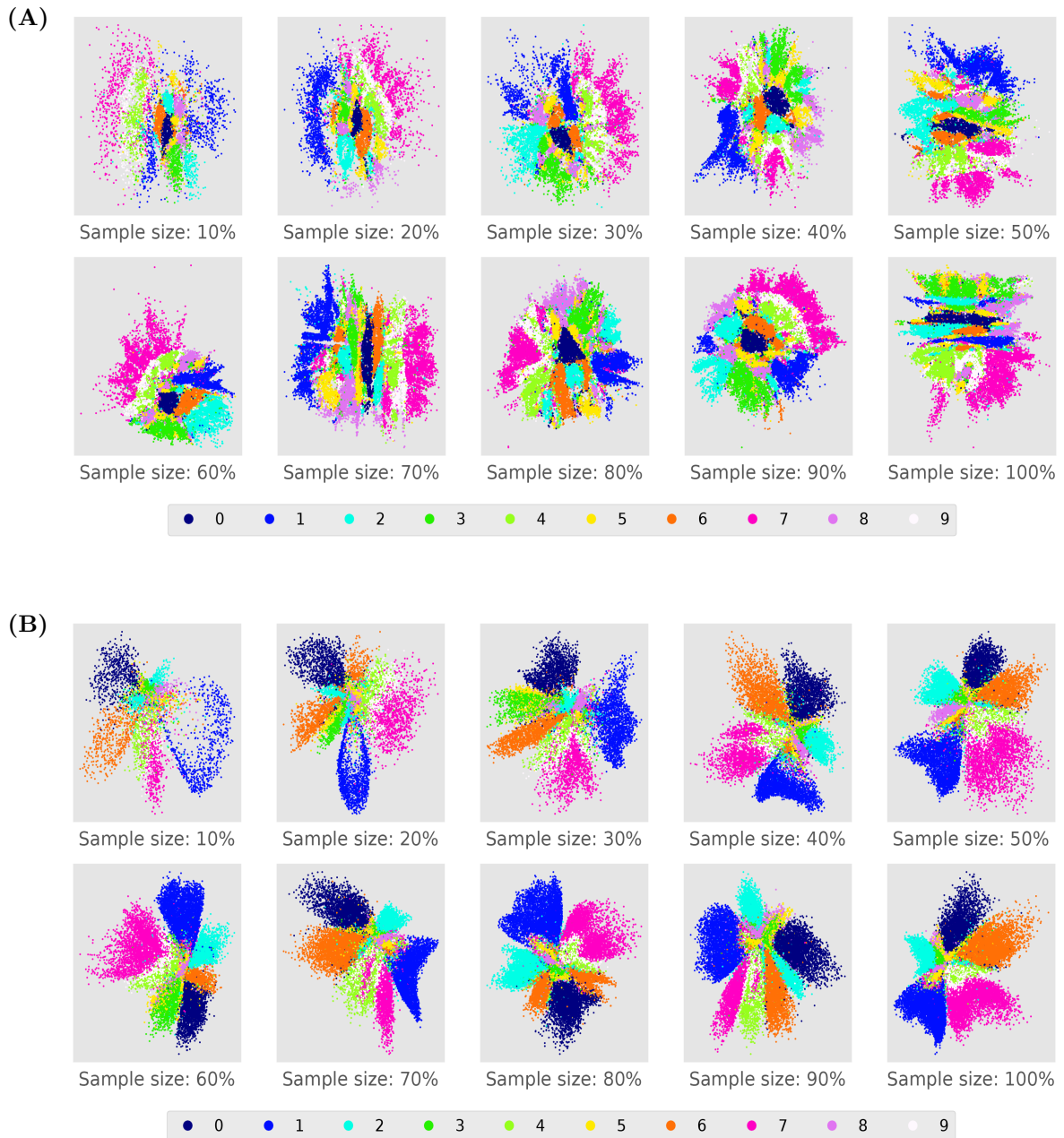


Figure C.II.1: 2-dimensional embeddings produced by (A) G-NeuroDAVIS and (B) VAE, in an unsupervised fashion, for different sub-samples with varying sizes of *MNIST* dataset, coloured by different digits.

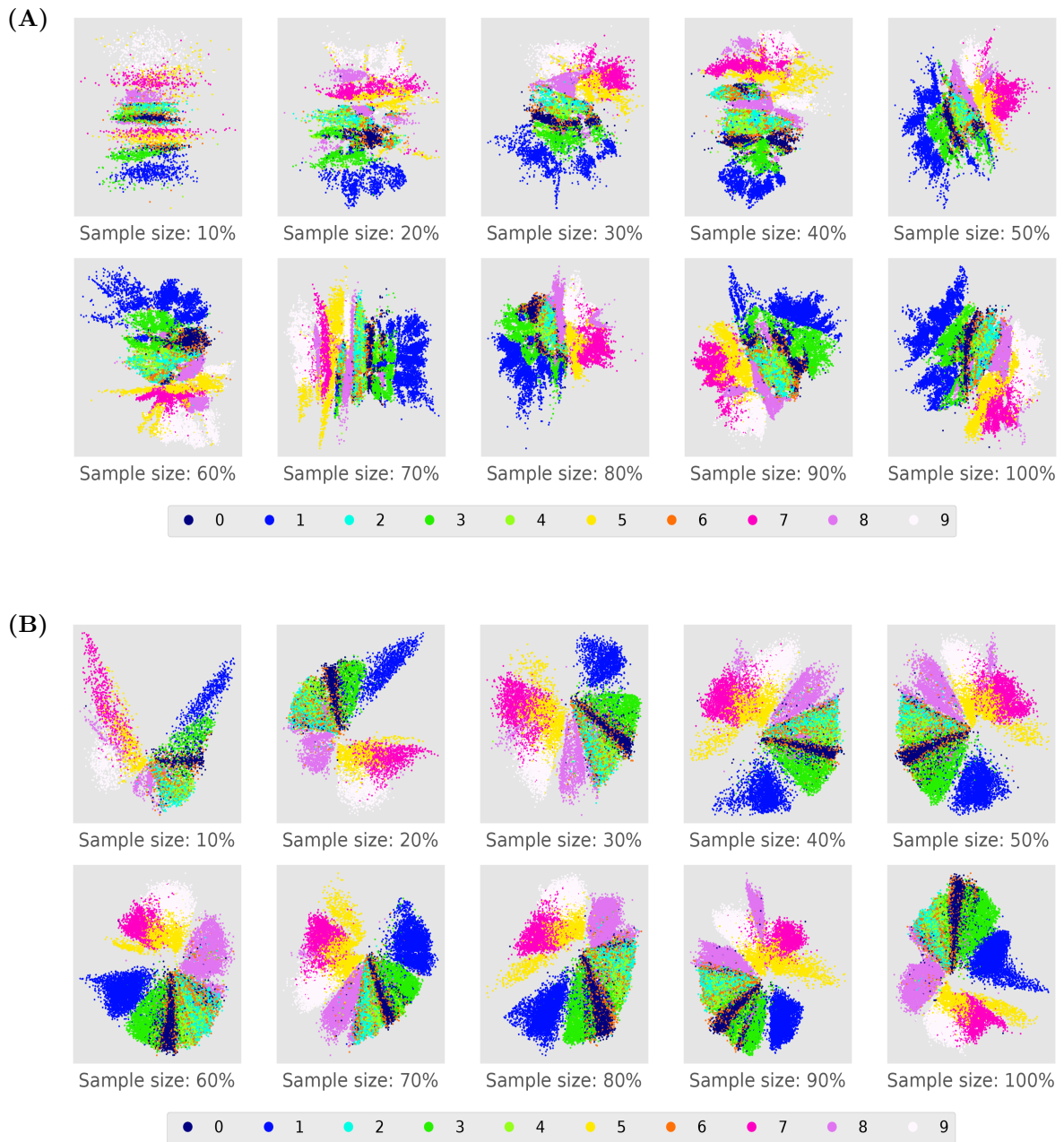


Figure C.II.2: 2-dimensional embeddings produced by (A) G-NeuroDAVIS and (B) VAE, in an unsupervised fashion, for different sub-samples with varying sizes of *FMNIST* dataset, coloured by clothing's label.

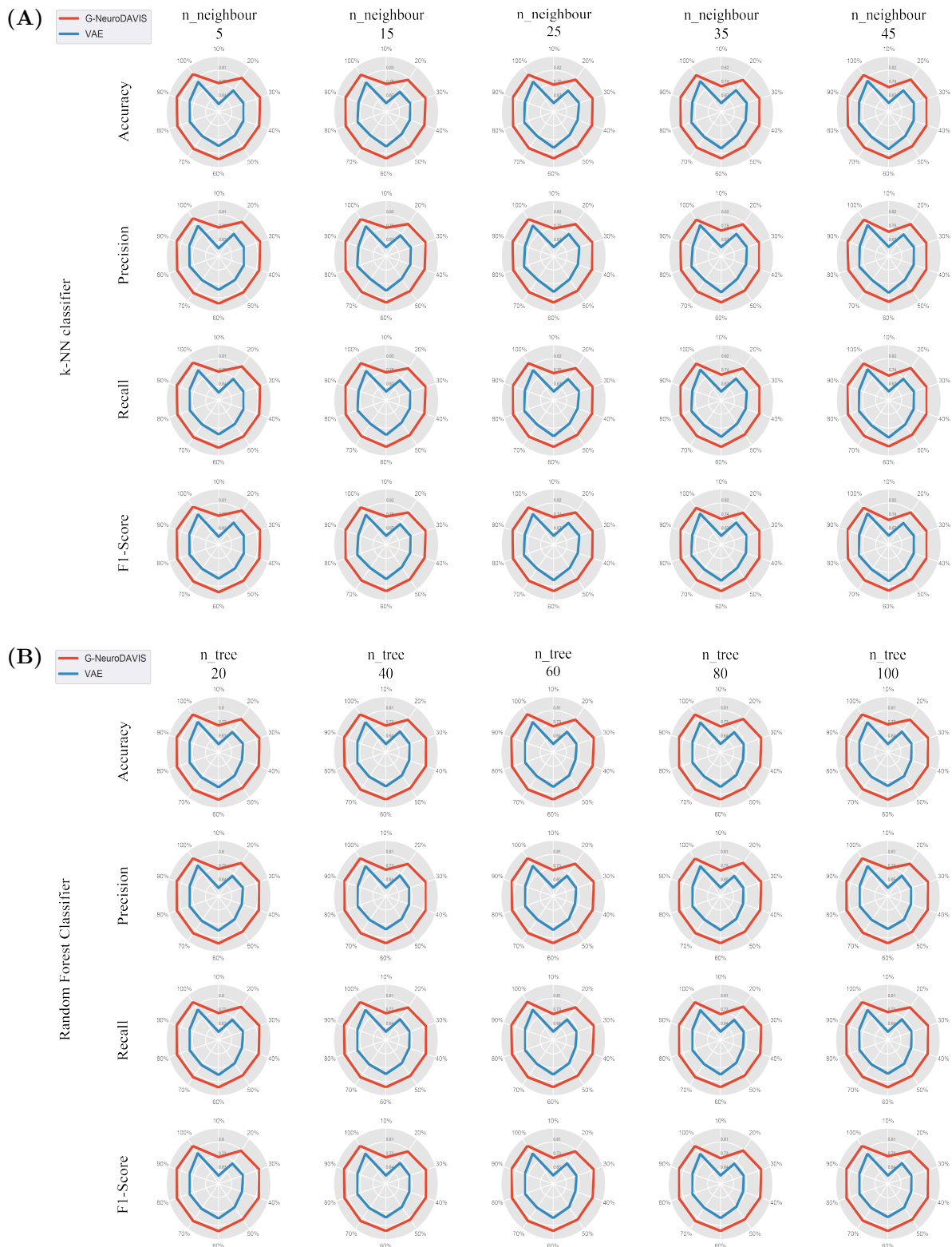


Figure C.II.3: Classification performance of (A) k-NN and (B) RF on G-NeuroDAVIS-generated embedding against that obtained using VAE for different sub-samples with varying sizes of *MNIST* dataset in terms of Accuracy, Precision, Recall and F1-Score.

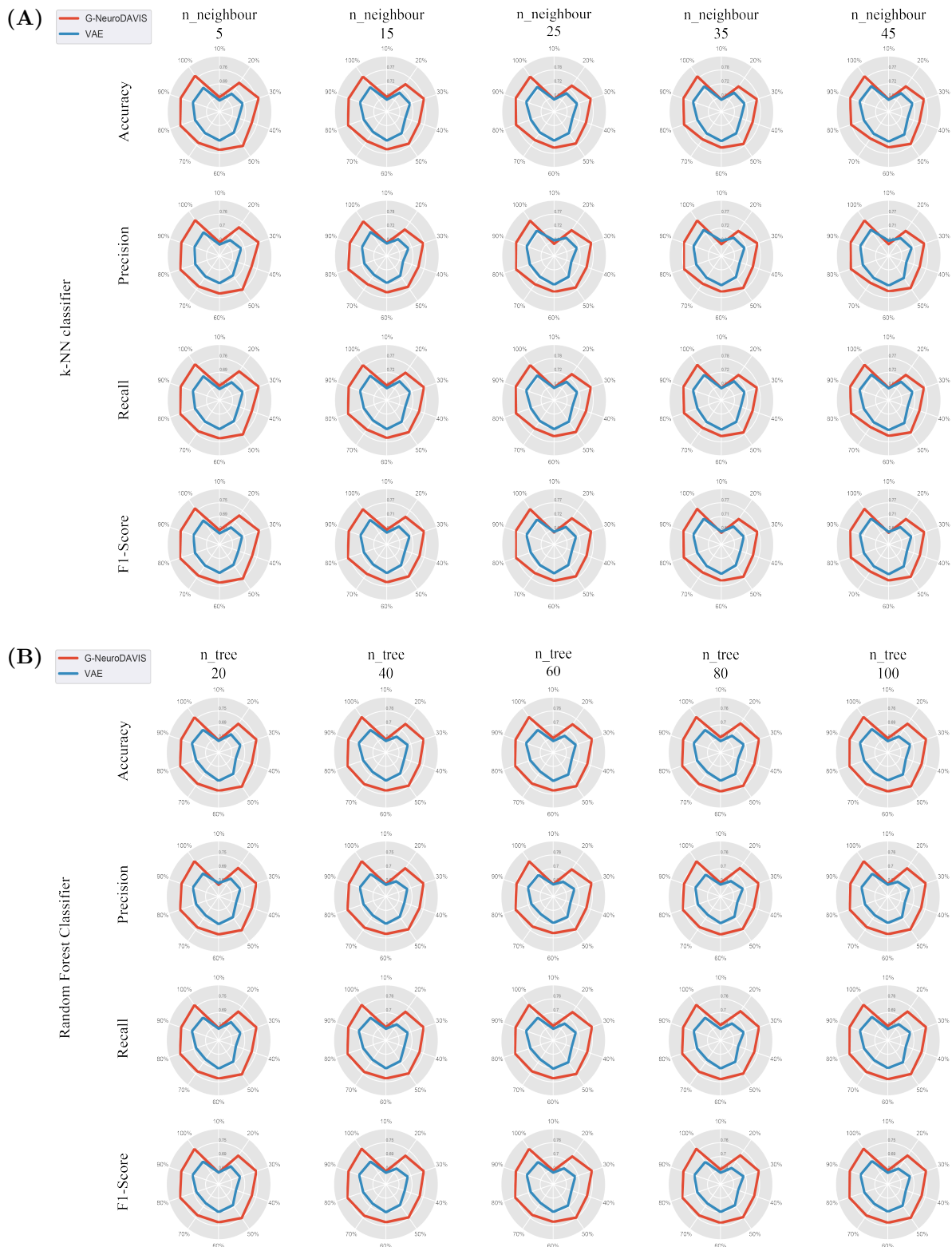


Figure C.II.4: Classification performance of (A) k-NN and (B) RF on G-NeuroDAVIS-generated embedding against that obtained using VAE for different sub-samples with varying sizes of *FMNIST* dataset in terms of Accuracy, Precision, Recall and F1-Score.

C.III. G-NeuroDAVIS in a supervised setting

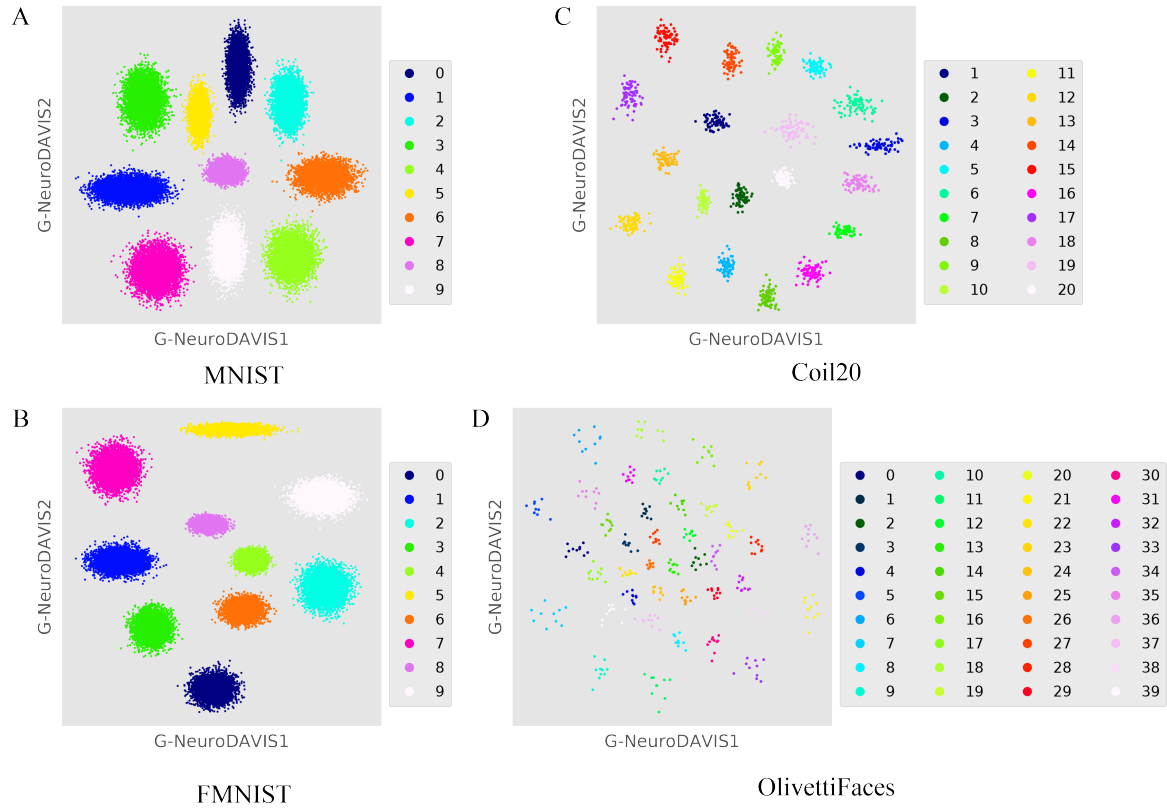


Figure C.III.1: Embeddings produced by G-NeuroDAVIS, in an supervised setting, for (A) *MNIST*, (B) *FMNIST*, (C) *Coil20* and (D) *Olivetti Faces* dataset

Bibliography

- [1] Chayan Maitra, Dibyendu B. Seal, and Rajat K. De. NeuroDAVIS: A neural network model for data visualization. *Neurocomputing*, 573:127182, 2024.
- [2] Chayan Maitra, Anwesha Sengupta, and Rajat K. De. NeuroDAVIS-FS: Feature Selection Through Visualization Using NeuroDAVIS. In *Proceedings of the International Conference on Pattern Recognition*, pages 284–298. Springer Nature Switzerland, 2025.
- [3] Chayan Maitra, Dibyendu B. Seal, Vivek Das, and Rajat K. De. NeuroMDAVIS: Visualization of Single-Cell Multi-Omics Data under Deep Learning Framework. *Journal of Computational Biology*, 0(0):15578666261423965, 2026. PMID: 41744270.
- [4] Chayan Maitra, Vivek Das, Dibyendu B. Seal, and Rajat K. De. Multi-Omics Integration for Identification of Prognostic Molecular Signatures for Survival Stratification in Lung Cancer. *MedRxiv*, 2026.
- [5] Chayan Maitra and Rajat K. De. G-NeuroDAVIS: A generative model for data visualization through a generalized embedding. *Neural Networks*, 193:107948, 2026.
- [6] Nicoleta Tantalaki, Stavros Souravlas, and Manos Roumeliotis. Data-driven decision making in precision agriculture: The rise of big data in agricultural systems. *Journal of agricultural & food information*, 20(4):344–380, 2019.
- [7] Iain M Johnstone and D Michael Titterton. Statistical challenges of high-dimensional data, 2009.
- [8] Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th international conference on world wide web*, pages 287–297, 2016.
- [9] G Manikandan and S Abirami. A survey on feature selection and extraction techniques for high-dimensional microarray datasets. *Knowledge Computing and Its Applications: Knowledge Computing in Specific Domains: Volume II*, pages 311–333, 2018.
- [10] Jundong Li, Kewei Cheng, Suhan Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6):1–45, 2017.
- [11] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417, 1933.

- [12] Christian Jutten and Jeanny Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.
- [13] Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.
- [14] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
- [15] Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 274–281, 2001.
- [16] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.
- [17] Gilbert W Stewart. On the early history of the singular value decomposition. *SIAM Review*, 35(4):551–566, 1993.
- [18] Alan Julian Izenman. Linear discriminant analysis. In *Modern Multivariate Statistical Techniques*, pages 237–280. Springer, 2013.
- [19] Chih-Wen Chen, Yi-Hong Tsai, Fang-Rong Chang, and Wei-Chao Lin. Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Systems*, 37(5):e12553, 2020.
- [20] B Venkatesh and J Anuradha. A review of feature selection and its methods. *Cybernetics and Information Technologies*, 19(1):3–26, 2019.
- [21] Muhammad Al Fatih Abil Fida, Tohari Ahmad, and Maurice Ntahobari. Variance threshold as early screening to Boruta feature selection for intrusion detection system. In *Proceedings of the 2021 13th International Conference on Information & Communication Technology and System (ICTS)*, pages 46–50. IEEE, 2021.
- [22] Gauthier Doquire and Michel Verleysen. Mutual information-based feature selection for multilabel classification. *Neurocomputing*, 122:148–155, 2013.
- [23] Yamen El Touati, Jihane Ben Slimane, and Taoufik Saidani. Adaptive Method for Feature Selection in the Machine Learning Context. *Engineering, Technology & Applied Science Research*, 14(3):14295–14300, 2024.
- [24] Adi Zaenul Mustaqim, Sumarni Adi, Yoga Pristyanto, and Yuli Astuti. The effect of recursive feature elimination with cross-validation (RFECV) feature selection algorithm toward classifier performance on credit card fraud detection. In *Proceedings of the 2021 International Conference on Artificial Intelligence and Computer Science Technology (ICAICST)*, pages 270–275. IEEE, 2021.
- [25] Jonas Ranstam and Jonathan A Cook. LASSO regression. *Journal of British Surgery*, 105(10):1348–1348, 2018.
- [26] Steven J Rigatti. Random forest. *Journal of Insurance Medicine*, 47(1):31–39, 2017.

- [27] Laurens Van Der Maaten, Eric O Postma, H Jaap van den Herik, et al. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(66-71):13, 2009.
- [28] Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. *Journal of Machine Learning Research*, 22(1):9129–9201, 2021.
- [29] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
- [30] Jiarui Ding, Anne Condon, and Sohrab P Shah. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nature Communications*, 9(1):1–13, 2018.
- [31] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to Use t-SNE Effectively. *Distill*, 2016.
- [32] Erich Schubert and Michael Gertz. Intrinsic t-stochastic neighbor embedding for visualization and outlier detection. In *Proceedings of the International Conference on Similarity Search and Applications*, pages 188–203. Springer, 2017.
- [33] Ehsan Amid and Manfred K Warmuth. A more globally accurate dimensionality reduction method using triplets. *arXiv preprint arXiv:1803.00854*, 2018.
- [34] Laurens Van Der Maaten. Learning a parametric embedding by preserving local structure. In *Proceedings of the Artificial Intelligence and Statistics*, pages 384–391. PMLR, 2009.
- [35] Francesco Crecchi, Cyril de Bodt, Michel Verleysen, John A. Lee, and Davide Bacciu. Perplexity-free Parametric t-SNE, 2020.
- [36] Yoshiaki Kawase, Kosuke Mitarai, and Keisuke Fujii. Parametric t-Stochastic Neighbor Embedding With Quantum Neural Network, 2022.
- [37] George C Linderman, Manas Rachh, Jeremy G Hoskins, Stefan Steinerberger, and Yuval Kluger. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature Methods*, 16(3):243–245, 2019.
- [38] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 2016.
- [39] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1):38–44, 2019.
- [40] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

- [41] Dmitry Kobak and George C Linderman. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology*, 39(2):156–157, 2021.
- [42] Kevin R Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B Burkhardt, William S Chen, Kristina Yim, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, et al. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12):1482–1492, 2019.
- [43] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.
- [44] Chieh Lin, Siddhartha Jain, Hannah Kim, and Ziv Bar-Joseph. Using neural networks for reducing the dimensions of single-cell RNA-Seq data. *Nucleic Acids Research*, 45(17):e156–e156, 2017.
- [45] Wei Wang, Yan Huang, Yizhou Wang, and Liang Wang. Generalized Autoencoder: A Neural Network Framework for Dimensionality Reduction. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 496–503, 2014.
- [46] Teuvo Kohonen. The self-organizing map. In *Proceedings of the IEEE*, volume 78, pages 1464–1480. IEEE, 1990.
- [47] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.
- [48] Divyanshu Talwar, Aanchal Mongia, Debarka Sengupta, and Angshul Majumdar. AutoImpute: Autoencoder based imputation of single-cell RNA-seq data. *Scientific Reports*, 8(1):1–11, 2018.
- [49] Rui Xie, Jia Wen, Andrew Quitadamo, Jianlin Cheng, and Xinghua Shi. A deep auto-encoder model for gene expression prediction. *BMC Genomics*, 18(9):39–49, 2017.
- [50] Benjamin Szubert, Jennifer E Cole, Claudia Monaco, and Ignat Drozdov. Structure-preserving visualisation of high dimensional single-cell datasets. *Scientific Reports*, 9(1):1–10, 2019.
- [51] Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124, 2018.
- [52] Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, and Oliver Stegle. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1):111, 2020.

- [53] Amrit Singh, Casey P Shannon, Benoît Gautier, Florian Rohart, Michaël Vacher, Scott J Tebbutt, and Kim-Anh Lê Cao. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17):3055–3062, 2019.
- [54] Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L Nazon, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18(3):272–282, 2021.
- [55] Chayan Maitra, Dibyendu B Seal, Vivek Das, and Rajat K De. Unsupervised neural network for single cell Multi-omics INTEgration (UMINT): an application to health and disease. *Frontiers in Molecular Biosciences*, 10:1184748, 2023.
- [56] Mohammad Lotfollahi, Anastasia Litinetskaya, and Fabian J Theis. Multigrade: single-cell multi-omic data integration. *BioRxiv*, pages 2022–03, 2022.
- [57] Yun-Hao Yuan, Quan-Sen Sun, Qiang Zhou, and De-Shen Xia. A novel multiset integrated canonical correlation analysis framework and its application in feature fusion. *Pattern Recognition*, 44(5):1031–1040, 2011.
- [58] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333–337, 2014.
- [59] Mika Sarkin Jain, Krzysztof Polanski, Cecilia Dominguez Conde, Xi Chen, Jongeun Park, Lira Mamanova, Andrew Knights, Rachel A Botting, Emily Stephenson, Muzlifah Haniffa, et al. MultiMAP: dimensionality reduction and integration of multimodal data. *Genome Biology*, 22(1):346, 2021.
- [60] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [61] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [62] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.
- [63] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.
- [64] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 31, 2018.
- [65] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [66] W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Proceedings of the Biomedical Image Processing and Biomedical visualization*, volume 1905, pages 861–870. SPIE, 1993.

- [67] Stefan Aeberhard, Danny Coomans, and Olivier de Vel. The classification performance of RDA. *Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland, Tech. Rep*, pages 92–01, 1992.
- [68] Tiago A Almeida, José María G Hidalgo, and Akebo Yamakami. Contributions to the study of SMS spam filtering: new collection and results. In *Proceedings of the 11th ACM Symposium on Document Engineering*, pages 259–262, 2011.
- [69] SA Nene, SK Nayar, and H Murase. Columbia University Image Library (COIL-20). *Department of Computer Science, Columbia University, Tech. Rep. CUCS-005-96*, 1996.
- [70] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *CoRR*, abs/1708.07747, 2017.
- [71] Dmitry Usoskin, Alessandro Furlan, Saiful Islam, Hind Abdo, Peter Lönnerberg, Daohua Lou, Jens Hjerling-Leffler, Jesper Haeggström, Olga Kharchenko, Peter V Kharchenko, et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nature Neuroscience*, 18(1):145–153, 2015.
- [72] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):1–12, 2017.
- [73] Dibyendu Bikash Seal, Vivek Das, and Rajat K De. CASSL: A cell-type annotation method for single cell transcriptomics data using semi-supervised learning. *Applied Intelligence*, pages 1–19, 2022.
- [74] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [75] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):1–5, 2018.
- [76] Jarkko Venna and Samuel Kaski. Local multidimensional scaling. *Neural Networks*, 19(6-7):889–899, 2006.
- [77] Samina Khalid, Tehmina Khalil, and Shamila Nasreen. A survey of feature selection and feature extraction techniques in machine learning. In *Proceedings of the 2014 Science and Information Conference*, pages 372–378. IEEE, 2014.
- [78] Usama Manzoor, Zahid Halim, et al. Protein encoder: An autoencoder-based ensemble feature selection scheme to predict protein secondary structure. *Expert Systems with Applications*, 213:119081, 2023.
- [79] John R Hershey and Peder A Olsen. Approximating the Kullback Leibler divergence between Gaussian mixture models. In *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP’07*, volume 4, pages IV–317. IEEE, 2007.

- [80] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.
- [81] Simone Picelli, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, 10(11):1096–1098, 2013.
- [82] Neil I Weisenfeld, Vijay Kumar, Preyas Shah, Deanna M Church, and David B Jaffe. Direct determination of diploid genome sequences. *Genome Research*, 27(5):757–767, 2017.
- [83] Patrick Marks, Sarah Garcia, Alvaro Martinez Barrio, Kamila Belhocine, Jorge Bernate, Rajiv Bharadwaj, Keith Bjornson, Claudia Catalanotti, Josh Delaney, Adrian Fehr, et al. Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Research*, 29(4):635–645, 2019.
- [84] Ansuman T Satpathy, Jeffrey M Granja, Kathryn E Yost, Yanyan Qi, Francesca Meschi, Geoffrey P McDermott, Brett N Olsen, Maxwell R Mumbach, Sarah E Pierce, M Ryan Corces, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology*, 37(8):925–936, 2019.
- [85] Nona Farbehi, Ralph Patrick, Aude Dorison, Munira Xaymardan, Vaibhao Janbandhu, Katharina Wystub-Lis, Joshua WK Ho, Robert E Nordon, and Richard P Harvey. Single-cell expression profiling reveals dynamic flux of cardiac stromal, vascular and immune cells in health and injury. *eLife*, 8:e43882, 2019.
- [86] Tamar Hashimshony, Naftalie Senderovich, Gal Avital, Agnes Klochandler, Yaron De Leeuw, Leon Anavy, Dave Gennert, Shuqiang Li, Kenneth J Livak, Orit Rozenblatt-Rosen, et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology*, 17(1):1–7, 2016.
- [87] Asif Adil, Vijay Kumar, Arif Tasleem Jan, and Mohammed Asger. Single-cell transcriptomics: current methods and challenges in data acquisition and analysis. *Frontiers in Neuroscience*, 15:591122, 2021.
- [88] Jason D Buenrostro, Beijing Wu, Howard Y Chang, and William J Greenleaf. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, 109(1):21–29, 2015.
- [89] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, 2017.
- [90] Vanessa M Peterson, Kelvin Xi Zhang, Namit Kumar, Jerelyn Wong, Lixia Li, Douglas C Wilson, Renee Moore, Terrill K McClanahan, Svetlana Sadekova, and Joel A Klappenbach. Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology*, 35(10):936–939, 2017.

- [91] Christof Angermueller, Stephen J Clark, Heather J Lee, Iain C Macaulay, Mabel J Teng, Tim Xiaoming Hu, Felix Krueger, Sébastien A Smallwood, Chris P Ponting, Thierry Voet, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*, 13(3):229–232, 2016.
- [92] Youjin Hu, Kevin Huang, Qin An, Guizhen Du, Ganlu Hu, Jinfeng Xue, Xianmin Zhu, Cun-Yu Wang, Zhigang Xue, and Guoping Fan. Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biology*, 17(1):1–11, 2016.
- [93] Yu Hou, Huahu Guo, Chen Cao, Xianlong Li, Boqiang Hu, Ping Zhu, Xinglong Wu, Lu Wen, Fuchou Tang, Yanyi Huang, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Research*, 26(3):304–319, 2016.
- [94] Junyue Cao, Darren A Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A Pliner, Andrew J Hill, Riza M Daza, Jose L McFaline-Figueroa, Jonathan S Packer, Lena Christiansen, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385, 2018.
- [95] Song Chen, Blue B Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, 37(12):1452–1457, 2019.
- [96] Dorothy Clyde. SHARE-seq reveals chromatin potential. *Nature Reviews Genetics*, 22(1):2–2, 2021.
- [97] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhao Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.
- [98] Yuri Kotliarov, Rachel Sparks, Andrew J Martins, Matthew P Mulè, Yong Lu, Meghali Goswami, Lela Kardava, Romain Banchereau, Virginia Pascual, Angélique Biancotto, et al. Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus. *Nature Medicine*, 26(4):618–629, 2020.
- [99] Danila Bredikhin, Ilia Kats, and Oliver Stegle. Muon: multimodal omics analysis framework. *Genome Biology*, 23(1):1–12, 2022.
- [100] Tara Chari and Lior Pachter. The specious art of single-cell genomics. *PLoS Computational Biology*, 19(8):1–20, 08 2023.
- [101] Ana I Robles and Curtis C Harris. Integration of multiple “OMIC” biomarkers: A precision medicine strategy for lung cancer. *Lung Cancer*, 107:50–58, 2017.
- [102] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.
- [103] Robert Tibshirani. The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395, 1997.

- [104] Travers Ching, Xun Zhu, and Lana X Garmire. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Computational Biology*, 14(4):e1006076, 2018.
- [105] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018.
- [106] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [107] Lindong Jiang, Chao Xu, Yuntong Bai, Anqi Liu, Yun Gong, Yu-Ping Wang, and Hong-Wen Deng. Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data. *npj Precision Oncology*, 8(1):4, 2024.
- [108] Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, 2022.
- [109] Yingxin Lin, Tung-Yu Wu, Sheng Wan, Jean YH Yang, Wing H Wong, and YX Rachel Wang. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. *Nature Biotechnology*, 40(5):703–710, 2022.
- [110] Chunlei Liu, Hao Huang, and Pengyi Yang. Multi-task learning from multimodal single-cell omics with Matilda. *Nucleic Acids Research*, 51(8):e45–e45, 2023.
- [111] Magdalena Wysocka, Oskar Wysocki, Marie Zufferey, Dónal Landers, and André Freitas. A systematic review of biologically-informed deep learning models for cancer: fundamental trends for encoding and interpreting oncology data. *BMC Bioinformatics*, 24(1):198, 2023.
- [112] David A Selby, Maximilian Sprang, Jan Ewald, and Sebastian J Vollmer. Beyond the black box with biologically informed neural networks. *Nature Reviews Genetics*, pages 1–2, 2025.
- [113] Erik Hartman, Aaron M Scott, Christofer Karlsson, Tirthankar Mohanty, Suvi T Vaara, Adam Linder, Lars Malmström, and Johan Malmström. Interpreting biologically informed neural networks for enhanced proteomic biomarker discovery and pathway analysis. *Nature Communications*, 14(1):5359, 2023.
- [114] Xiaofeng Jiang, Michael Hoffmeister, Hermann Brenner, Hannah Sophie Muti, Tanwei Yuan, Sebastian Foersch, Nicholas P West, Alexander Brobeil, Jitendra Jonnagaddala, Nicholas Hawkins, et al. End-to-end prognostication in colorectal cancer by deep learning: a retrospective, multicentre study. *The Lancet Digital Health*, 6(1):e33–e43, 2024.
- [115] Yibo Zhang, Zijian Yang, Ruanqi Chen, Yanli Zhu, Li Liu, Jiyan Dong, Zicheng Zhang, Xujie Sun, Jianming Ying, Dongmei Lin, et al. Histopathology images-based deep learning prediction of prognosis and therapeutic response in small cell lung cancer. *npj Digital Medicine*, 7(1):15, 2024.

- [116] Youyu Wang, Xueming Ju, Rong Hua, Ji Chen, Xiaoqin Dai, Lunxu Liu, Guifang Wang, Yifeng Bai, Honglin Hu, and Xiaohua Li. Deep learning analysis of histopathological images predicts immunotherapy prognosis and reveals tumour microenvironment features in non-small cell lung cancer. *British Journal of Cancer*, 131(11):1833–1845, 2024.
- [117] Verena Bitto, Xiaofeng Jiang, Michael Baumann, Jakob Nikolas Kather, and Ina Kurth. Deep Learning Predicts Survival Across Squamous Tumor Entities From Routine Pathology: Insights from Head and Neck, Esophagus, Lung and Cervical Cancer. *Modern Pathology*, page 100845, 2025.
- [118] Nathan J Edwards, Mauricio Oberti, Ratna R Thangudu, Shuang Cai, Peter B McGarvey, Shine Jacob, Subha Madhavan, and Karen A Ketchum. The CPTAC data portal: a resource for cancer proteomics research. *Journal of Proteome Research*, 14(6):2707–2713, 2015.
- [119] Caleb M Lindgren, David W Adams, Benjamin Kimball, Hannah Boekweg, Sadie Tayler, Samuel L Pugh, and Samuel H Payne. Simplified and unified access to cancer proteogenomic data. *Journal of Proteome Research*, 20(4):1902–1910, 2021.
- [120] Lufeng Zhao, Zipu Yu, and Baiqin Zhao. Mechanism of VIPR1 gene regulating human lung adenocarcinoma H1299 cells. *Medical Oncology*, 36(11):91, 2019.
- [121] Tyson V Sharp, Fernando Munoz, Dimitra Bourboulia, Nadege Presneau, Eva Darai, Hsei-Wei Wang, Mark Cannon, David N Butcher, Andrew G Nicholson, George Klein, et al. LIM domains-containing protein 1 (LIMD1), a tumor suppressor encoded at chromosome 3p21. 3, binds pRB and represses E2F-driven transcription. In *Proceedings of the National Academy of Sciences*, volume 101, pages 16531–16536. National Academy of Sciences, 2004.
- [122] Christian Boni and Claudio Sorio. The role of the tumor suppressor gene protein tyrosine phosphatase gamma in cancer. *Frontiers in Cell and Developmental Biology*, 9:768969, 2022.
- [123] Pranav Gupta, Praveen K Sharma, Hina Mir, Rajesh Singh, Nalinaksha Singh, Goetz H Kloecker, James W Lillard Jr, and Shailesh Singh. CCR9/CCL25 expression in non-small cell lung cancer correlates with aggressive disease and mediates key steps of metastasis. *Oncotarget*, 5(20):10170, 2014.
- [124] Chang-Li Wang, Bing-Sheng Sun, Yong Tang, Hong-Qing Zhuang, and Wei-Zhi Cao. CCR1 knockdown suppresses human non-small cell lung cancer cell invasion. *Journal of Cancer Research and Clinical Oncology*, 135(5):695–701, 2009.
- [125] Xiaomeng Wang, Juan Wang, Qiao Fu, Jing Luo, Mao Shu, and Zhihua Lin. Identification and interaction mechanism of novel small molecule antagonists targeting CC chemokine receptor 1/3/5 for treatment of non-small cell lung cancer. *Molecular Diversity*, 29(5):4159–4176, 2025.
- [126] Hina Mir, Rajesh Singh, Goetz H Kloecker, James W Lillard Jr, and Shailesh Singh. CXCR6 expression in non-small cell lung carcinoma supports metastatic process via modulating metalloproteinases. *Oncotarget*, 6(12):9985, 2015.

- [127] Shimo Shen, Yili Wu, Zhuowei Shao, You Li, Di Peng, Bing Li, Zhou Zhang, and Shibo Wu. LTF as a Potential Predictive Biomarker for Durable Benefit From First-Line Chemo-Immunotherapy in Small Cell Lung Cancer. *Cancer Science*, 116(6):1522–1536, 2025.
- [128] Robert J Torphy, Elliott J Yee, Richard D Schulick, and Yuwen Zhu. Atypical chemokine receptors: emerging therapeutic targets in cancer. *Trends in Pharmacological Sciences*, 43(12):1085–1097, 2022.
- [129] Yun Chen, Hongying Zhou, Sifu Yang, and Dan Su. Increased ABCC2 expression predicts cisplatin resistance in non-small cell lung cancer. *Cell Biochemistry and Function*, 39(2):277–286, 2021.
- [130] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved Precision and Recall Metric for Assessing Generative Models. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [131] Amy Genender-Feltheimer. Visualizing high dimensional and big data. *Procedia Computer Science*, 140:112–121, 2018.
- [132] Shuanlong Niu, Bin Li, Xinggang Wang, and Hui Lin. Defect image sample generation with GAN for improving defect recognition. *IEEE Transactions on Automation Science and Engineering*, 17(3):1611–1622, 2020.
- [133] Christine Dewi. Generative Adversarial Network for Synthetic Image Generation Method: Review, Analysis, and Perspective. *Applications of Generative AI*, pages 91–116, 2024.
- [134] Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database. *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist>, 2, 2010.
- [135] Damian Leśniak, Igor Sieradzki, and Igor Podolak. Distribution-interpolation trade off in generative models. In *Proceedings of the International Conference on Learning Representations*, 2019.
- [136] Ali Sharifara, Mohd Shafry Mohd Rahim, and Yasaman Anisi. A general review of human face detection including a study of neural networks and Haar feature-based cascade classifier in face detection. In *Proceedings of the 2014 International Symposium on Biometrics And Security Technologies (ISBAST)*, pages 73–78. IEEE, 2014.
- [137] Michelle E Pewarchuk, Vanessa GP Souza, David E Cohn, Nikita Telkar, Greg L Stewart, Katya H Bénard, Patricia P Reis, Victor D Martinez, Wendy P Robinson, and Wan L Lam. Identification of oncofetal PIWI-interacting RNAs as potential prognostic biomarkers in non-small cell lung cancer. *Frontiers in Genetics*, 16:1611805, 2025.
- [138] Michael Croft, Shahram Salek-Ardakani, and Carl F Ware. Targeting the TNF and TNFR superfamilies in autoimmune disease and cancer. *Nature Reviews Drug Discovery*, 23(12):939–961, 2024.

- [139] Hiroyuki Goda, Koh-ichi Nakashiro, Yoshifumi Sano, Tomoko Adachi, Norihiko Tokuzen, Nobuyuki Kuribayashi, Satoshi Hino, and Daisuke Uchida. KRT13 and UPK1B for differential diagnosis between metastatic lung carcinoma from oral squamous cell carcinoma and lung squamous cell carcinoma. *Scientific Reports*, 13(1):22626, 2023.
- [140] Yucheng Guo, Bin Xu, Weimin Mao, Zhao Zhao, Zhao Li, Zhengzhou Qiu, and Changying Guo. DNMT3B blocks TAL1-mediated PKM2 transcriptional repression to promote non-small cell lung cancer progression through inducing glycolysis. *Cell Division*, 20(1):25, 2025.