BIBS SEMINAR

On Thursday, April 18th, 2024

NDAGIJIMANA, Frank Aimee Rodrigue
Seoul National University
Interdisciplinary Program in Bioinformatics

서울대학교 통계학과
생물정보통계연구실

BIBS

THIS WEEK'S ARTICLE

Heliyon 10 (2024) e23195

CellPress

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

PaCMAP-embedded convolutional neural network for multi-omics data integration

Hazem Qattous [a,*], Mohammad Azzeh [b], Rahmeh Ibrahim [c], Ibrahim Abed Al-Ghafer [b], Mohammad Al Sorkhy [d], Abedalrhman Alkhateeb [e,*]

TABLE OF CONTENTS

INTRODUCTION

Integrating Biomedical Data Enhances the Understanding of Diseases

- **Advancements in Biomedical Technologies:** Rapid progress in biomedical technologies has led to extensive data generation across various omics platforms.
- **Generation of Comprehensive Data:** These technologies allow for the measurement of diverse molecular features in biological systems at a large scale and high resolution.
- **Significance for Disease Understanding:** Integrating data from different sources(levels of omics) is crucial for gaining comprehensive insights into complex diseases like cancer.
- **Challenges in Data Integration:** Concatenating data from different sources may struggle to effectively combine matrices with varying scales in a biologically meaningful way.

Extracting Discriminative Features from Multi-sources of Data

Researchers have proposed a number of models to extract discriminative features from multi-sources of data:

- **Multimodal Feature Encoder** to extract features from imagery data[1]
- **Dimensionality Reduction** to extract multi-omics features from single cell data
- **DeepTraSynergy:** a model combining protein-protein interactions, cell-target interaction, and drug sequences to predict various tasks such as drug-target interactions and drug combination synergy.[2]

[1] Razzaghi, Parvin, et al. "Multimodal brain tumor detection using multimodal deep transfer learning." *Applied Soft Computing* 129 (2022): 109631.

[2] Rafiei, Fatemeh, et al. "DeepTraSynergy: drug combinations using multimodal deep learning with transformers." *Bioinformatics* 39.8 (2023): btad438.

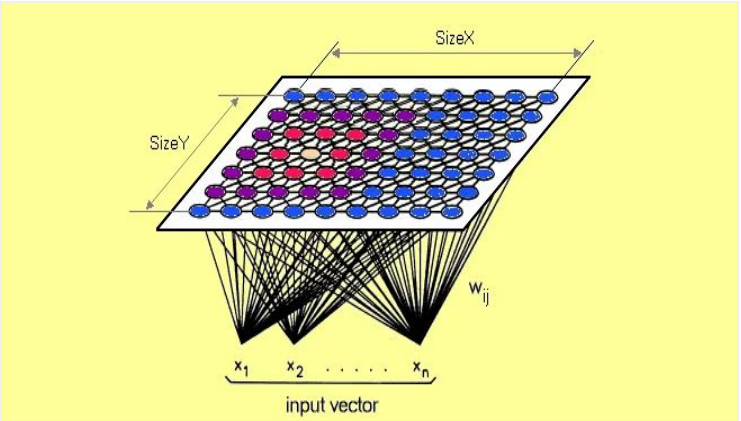
Dimension Reduction Techniques are essential in the Visualization of High Dimensional Data

- Dimension reduction techniques are essential in **data visualization** because they make it easier to understand and **analyze high dimensional data** by making it **less complicated**.
- **Techniques:**
  - Self-organizing maps (SOM)
  - Uniform manifold approximation and projection (UMAP)
  - t-distributed stochastic neighbor embedding (t-SNE)
- Each of these algorithms has its own strengths and limitations, and the optimal choice depends on the nature of the data and the problem at hand.

Dimension Reduction Techniques are essential in the Visualization of High Dimensional Data

(1) Self Organizing Maps (SOMs)

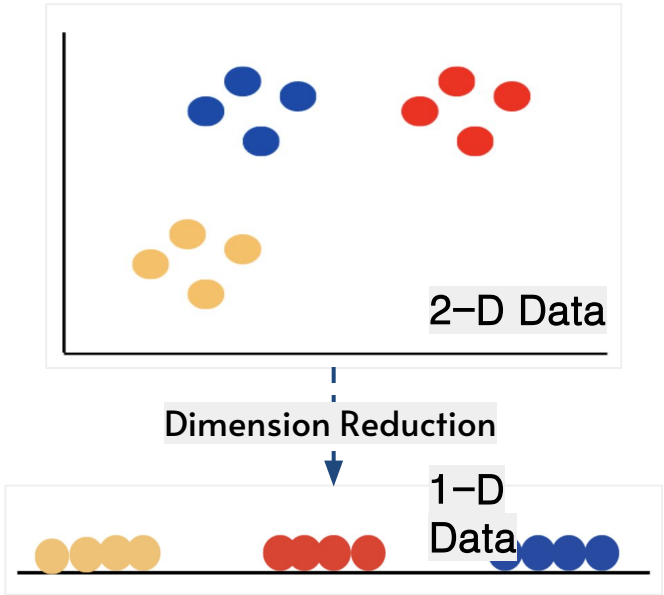
- A self-organizing map (SOM) is an artificial neural network (ANN) trained through unsupervised learning to generate a low-dimensional representation, often two-dimensional, of the input space from training data.
- SOMs adjust neuron weights through competitive learning, distinct from other neural networks that utilize backpropagation with SGD.
- Self organizing maps have two layers: the input and output layers.
- Unlike other types of artificial neural networks, SOMs do not have an activation function in their neurons.



Dimension Reduction Techniques are essential in the Visualization of High Dimensional Data

(2) t-distributed Stochastic Neighbor Embedding (t-SNE)

- t-SNE is an unsupervised non-linear dimensionality reduction technique for data exploration and visualizing high-dimensional data.
- It is often used to visualize complex datasets into two and three dimensions, while preserving the information in the high dimensional data.
- t-SNE preserves the relationships between data points in a lower-dimensional space, making it quite a good algorithm for visualizing complex high-dimensional data.



Dimension Reduction Techniques are essential in the Visualization of High Dimensional Data

(2) t-distributed Stochastic Neighbor Embedding (t-SNE)

- **Advantages of t-SNE**
  - t-SNE can give very intuitive visualizations as it preserves the local structure of the data in the lower dimensions
- **Disadvantages of t-SNE**
  - Computationally Expensive
  - Not very good at preserving global structure
  - Sensitive to hyperparameters
  - Can get stuck in local minima
  - Interpretation is challenging

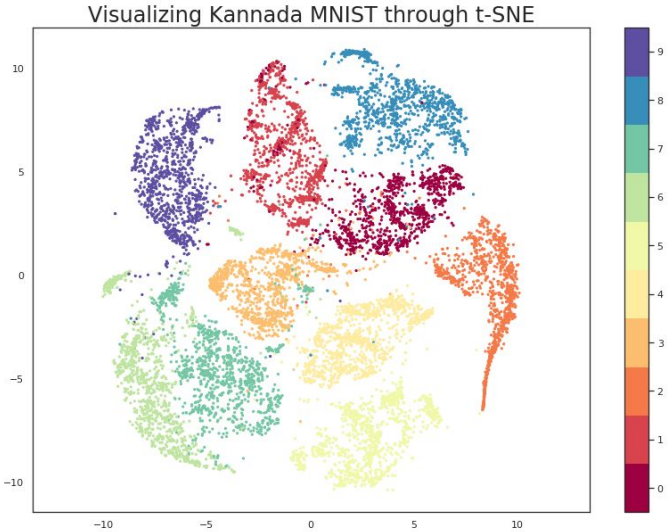
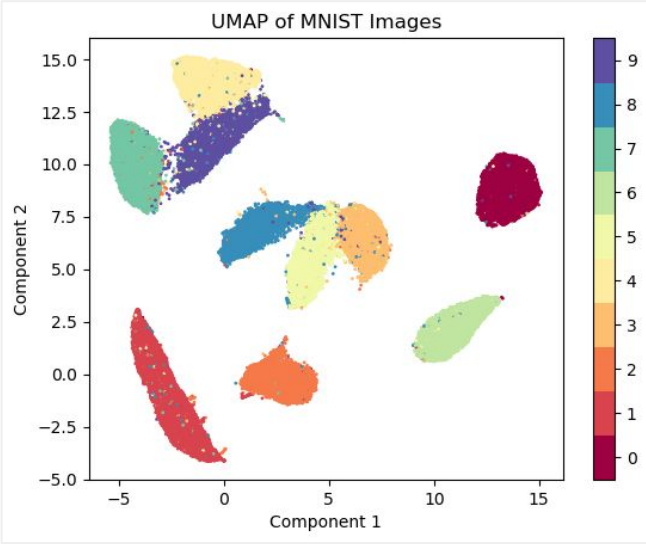


Image Source: Visualizing Kannada MNIST with t-SNE

Dimension Reduction Techniques are essential in the Visualization of High Dimensional Data

(3) Uniform Manifold Approximation and Projection (UMAP)

- UMAP, at its core, works very similarly to t-SNE - both use graph layout algorithms to **arrange data in low-dimensional space**.
- In the simplest sense, UMAP constructs a high dimensional graph representation of the data then optimizes a low-dimensional graph to be as structurally similar as possible.
- UMAP is often **better at preserving global structure** in the final projection. This means that the inter-cluster relations are potentially more meaningful than in t-SNE.



Source: Tezuka, Naoya & Ochiai, Hideya & Sun, Yuwei & Esaki, Hiroshi. (2022). Resilience of Wireless Ad Hoc Federated Learning against Model Poisoning Attacks. 10.48550/arXiv.2211.03489.

ABOUT THIS STUDY

Incorporating PaCMAP Dimension Reduction Technique

- The current study incorporates a recent DR technique known as **Pairwise Controlled Manifold Approximation (PaCMAP)**.
- PaCMAP is a **dimensionality reduction** method that optimizes low-dimensional embedding by utilizing three kinds of point pairs: neighbor pairs, mid-near pairs, and farther pairs.
- **Advantages of the Current Method**
  - Improved global vs. local trade-off.
  - Performs well without parameter tuning.
  - Substantially quicker than other algorithms.
  - Simple to use.
  - Highly intuitive hyper-parameters.

MATERIALS AND METHODS

Dataset

- **Data Composition**
  - The study analyzed the **TCGA Prostate Adenocarcinoma (PRCA) dataset**, which uses Gleason scores for prostate cancer aggressiveness classification.
  - The dataset comprises three omics: **copy number alteration (CNA), DNA methylation,** and **gene expression.**
  - It includes a total of **499 samples**, categorized into three classes based on Gleason scores: 4+3, 3+4, and a combined class of 4+5 and 5+4 due to limited samples.
  - **387 samples** with **all three omics** were selected for downstream analysis.

※ The Gleason Score(Click for More Information):

- The Gleason Score is a critical grading system for assessing prostate cancer aggressiveness.
- It ranges from 1 to 5, indicating the degree of resemblance between cancerous and healthy tissue.

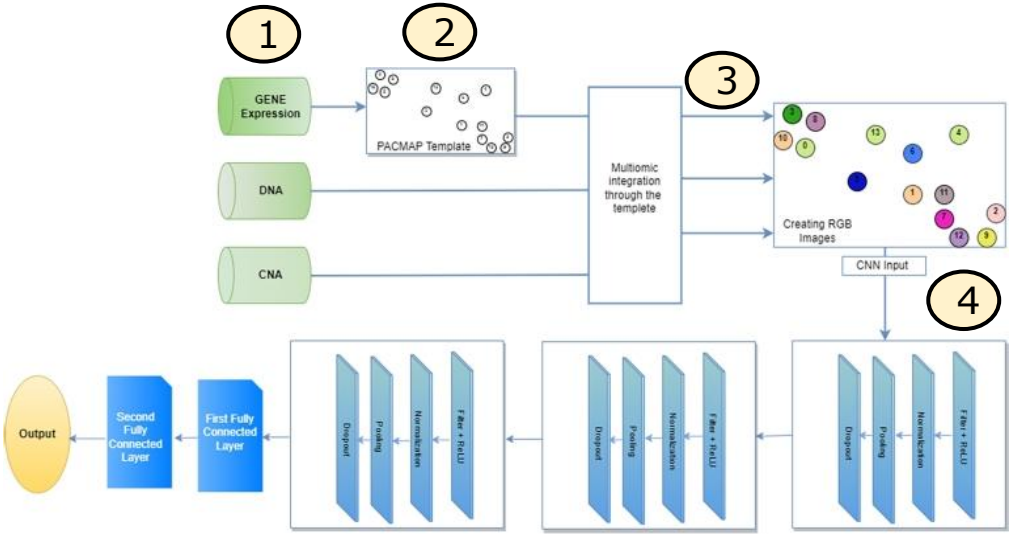
Dataset

- **Data Preprocessing**
  - Gene expression features were filtered initially to remove those with a **variance less than 0.2%**, reducing the number from around 39,000 to approximately 16,000.
  - Genes not listed in HUGO format were eliminated after normalizing all three omics data on an average scale.
  - The **MutSigCV algorithm** was employed to further refine the gene list by determining the False Discovery Rate (FDR). Genes with **FDR ≤ 0.1** were considered to have undergone significant mutations.
  - As a result, **14 mutated genes** were selected for inclusion in the study from the MutSigCV output.

※ MutSigCV algorithm identifies genes that are significantly mutated in cancer genomes, using a model with mutational covariates.(MutSigCV (v1))

Proposed Workflow

1. **Input:** Gene Expression, DNA methylation and Copy Number Alteration data
2. The model first builds a **Gene Similarity Matrix(GSN)** with **PaCMAP** to transform into a two dimensional map and create feature template
3. The template **aggregates** all the omics data and renders each sample as a **colored picture** with all of the omics data filled in.
4. The pictures are then sent to **CNN for classification.**



Pairwise Controlled Manifold Approximation Projection(PaCMAP)

- PaCMAP is a **dimensionality reduction** method that may be used for visualization and **maintains the local and global structure of the data** in the original space.
- PaCMAP uses three distinct point pairings to maximize low-dimensional embedding: neighbor, mid-near, and further pairs.
- Previous dimensionality reduction techniques like t-SNE and UMAP emphasize either local or global structure, with parameter tuning aiming to balance between the two.
- PaCMAP distinguishes itself by dynamically utilizing a specific set of **mid-near pairs** to capture global structure while enhancing local structure simultaneously.
- PaCMAP consists of three main steps: **construction**, **initialization of the solution**, and **iterative optimization** using a custom gradient descent algorithm.

Pairwise Controlled Manifold Approximation Projection(PaCMAP)

Step 1: Graph Construction:

- PaCMAP employs edges as graph building blocks.
- This DR distinguishes between neighbor pairs, mid-near pairs, and further pairs of edges.
- The first group is made up of each observation's number of closest neighbors in the high-dimensional space.
- The metric scaled distance is defined as appears in equation:

$$d_{ij}^{2,select} = \frac{\|Xi - Xj\|^2}{\sigma ij} \; and \; \sigma ij = \sigma i \, \sigma j,$$

where ***σi*** is the average separation between ***i*** and its fourth to sixth closest Euclidean neighbors.

- When choosing neighbors, in this case, the scaled distances $d_{ij}^{2,select}$ are only used; they are not used for optimization.

Pairwise Controlled Manifold Approximation Projection(PaCMAP)

- The second group comprises a number of **mid-near pairs**, which are chosen randomly.
- In addition, the third group is made up of additional points that were randomly chosen from each observation.
- For each kind of pair, PaCMAP employs three different loss functions in equation:

$$Loss_{NB} = \frac{\tilde{d}_{ij}}{10 + \tilde{d}_{ij}} \qquad Loss_{MN} = \frac{\tilde{d}_{ik}}{10000 + \tilde{d}_{ik}} \qquad Loss_{FP} = \frac{1}{1 + \tilde{d}_{il}}$$

where $\tilde{d}_{ab} = \|\mathbf{y}_a - \mathbf{y}_b\|^2 + 1$

- The coefficients $\mathbf{W_{NB}}$, $\mathbf{W_{MN}}$, and $\mathbf{W_{FP}}$, which combine to find the total loss, are added as additional weightings for the pairs.

Pairwise Controlled Manifold Approximation Projection(PaCMAP)

Step 2: Initialization of PaCMAP

- Although the initialization method has little effect on the results of PaCMAP, the Principal component analysis (PCA) DR is still used to actually reduce the running time.

Step 3: Dynamic Optimization

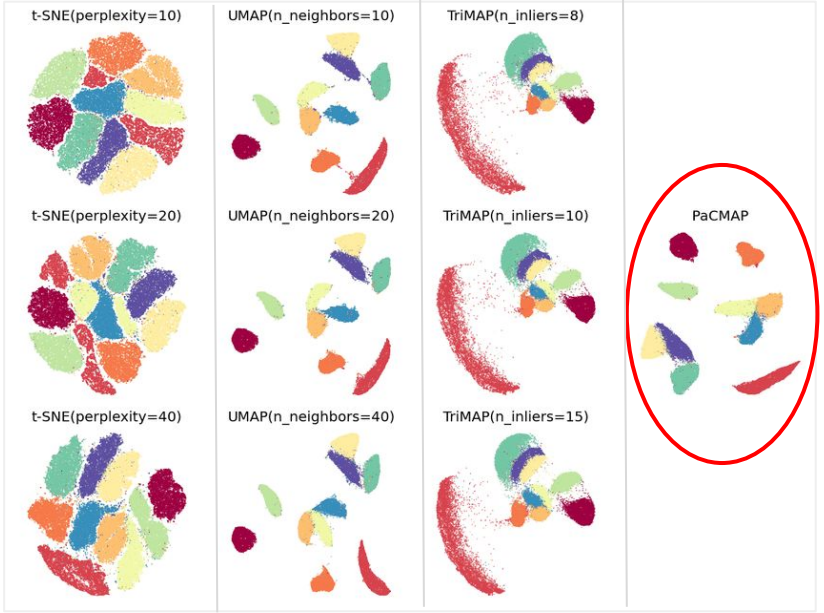
- Three phases make up the optimization process, each of which is intended to prevent local optima.
    - Initial placement emphasizes improving global structure, with mid-near pairs playing a significant role.
    - Subsequent phases refine local structure while maintaining the established global structure, gradually reducing the weight of mid-near pairs and neighbors.

More about PaCMAP: PaCMAP: An algorithm for dimension reduction

Pairwise Controlled Manifold Approximation Projection(PaCMAP)

PaCMAP's performance[4]

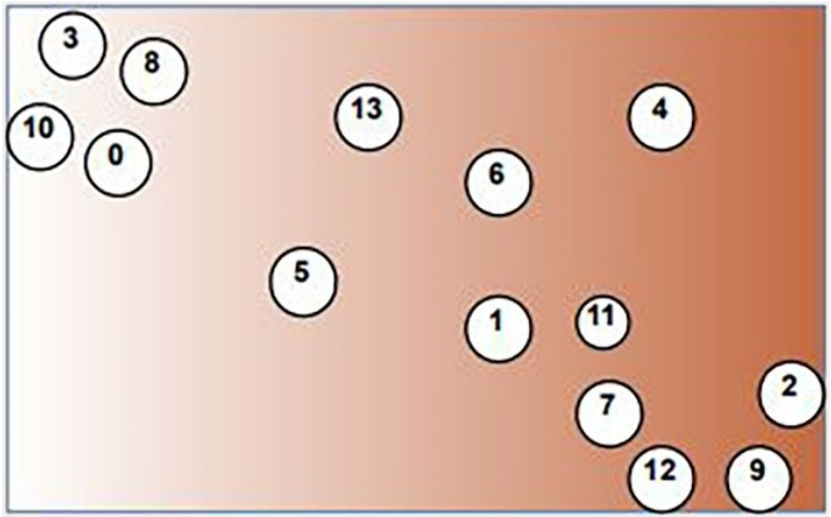
- PaCMAP's global structure preservation is similar to TriMAP.
- Local structure preservation performance is similar to UMAP and t-SNE.
- Computation is much faster than other algorithms.



[4] Wang, Yingfan, et al. "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* 22.201 (2021): 1-73.

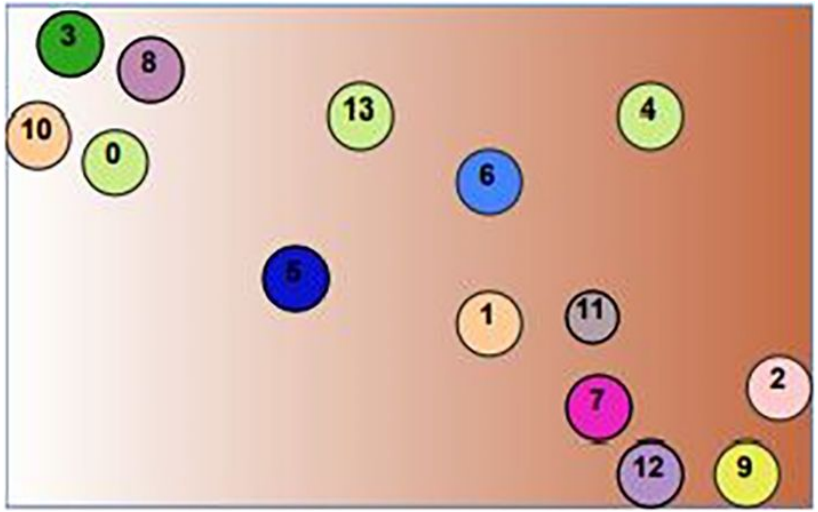
Omics Integration and Gene Similarity Networks

- PaCMAP is used for the gene expression omics, which generates the GSN and displays the genes on a two-dimensional map.
- The two-dimensional map shows the relationships between related genes as well as how similar or dissimilar the genes are arranged.
- The figure shows a GSN built by applying PaCMAP to gene expression omic.



Omics Integration and Gene Similarity Networks

- Following the design of the two-dimensional map, all omics data is integrated.
- The integration is carried out by constructing a circular zone with a predetermined radius around the gene sites and then filling those zones with various colors depending on the kind of omics.
- Gene expression is supported by the red color (R), DNA methylation by the green color (G), and CNA by the blue color (B).
- The figure shows the GSN map after coloring it by integrating the three omics values into the RGB system.



The Prediction Model

- CNNs are designed for processing grid-like data, particularly images.
- They comprise convolutional, pooling, and fully connected layers.
- Convolution and pooling layers extract features, while the fully connected layer handles classification.
- The described CNN architecture includes two identical convolutional layers with 32 filters (3×3), followed by max pooling (2×2) with a stride of 1×1, normalization, and a 20% dropout rate, as well as a third convolutional layer with similar specifications but with max pooling (2×2) having a stride of 2×2 and a 50% dropout rate.
- Rectified Linear Unit (ReLU) activation functions are applied in the three convolutional layers, accompanied by normalization and dropout layers to prevent overfitting.

EXPERIMENTS AND RESULTS

Experiment

- The authors utilized grid search to optimize the model's performance.
- The best accuracy was achieved with a learning rate of 0.05 and 80 epochs.
- The authors conducted experiments on the same dataset using default parameter values to compare their approach with the iSOM-GSN model[5].

[5] Fatima, Nazia, and Luis Rueda. "iSOM-GSN: an integrative approach for transforming multi-omic data into gene similarity networks via self-organizing maps." *Bioinformatics* 36.15 (2020): 4248-4254.

Result I: The Current Model Outperforms iSOM-GSN Model

- The model achieves an Area Under Curve (AUC) of 0.9996, compared to 0.9913 for iSOM-GSN.
- **Evaluation Metrics:**
  - Recall = $\frac{TP}{TP+FN}$
  - Precision = $\frac{TP}{TP+FP}$
  - Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
  - F1- measure = $\frac{2*(Precision * Recall)}{Precision+Recall}$

Hint:

- ❑ **TP:** Correctly predicted positive when actual positive.
- ❑ **TN:** Correctly predicted negative when actual negative.
- ❑ **FP:** Falsely predicted positive when actual negative.
- ❑ **FN:** Falsely predicted negative when actual positive

Performance assessment of the suggested model and the iSOM-GSN.

Performance measurements	The PRCA Dataset	
	Proposed model	iSOM-GSN
Accuracy	98.89%	97.89%
Precision	98.89%	98.82%
Recall	98.89%	98.72%
F1-measure	98.89%	98.71%
AUC	0.9996	0.9913

Result 2: Training and Validation Accuracy, and Loss for the Proposed Model

Fig. Training and Validation Accuracy

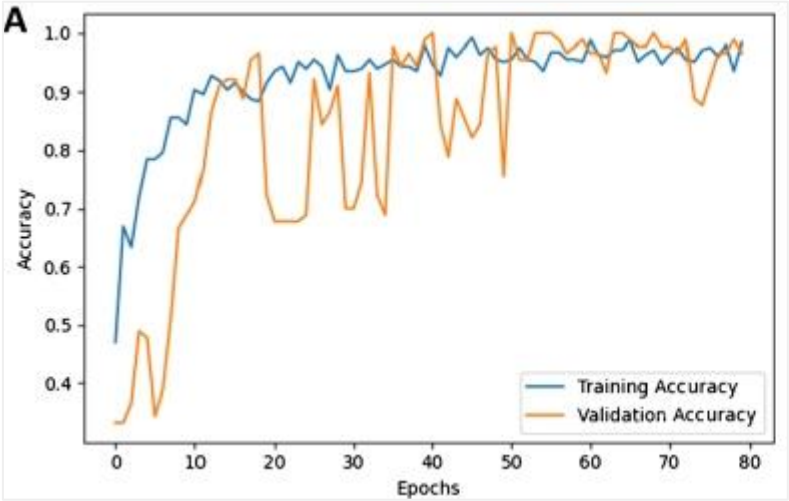
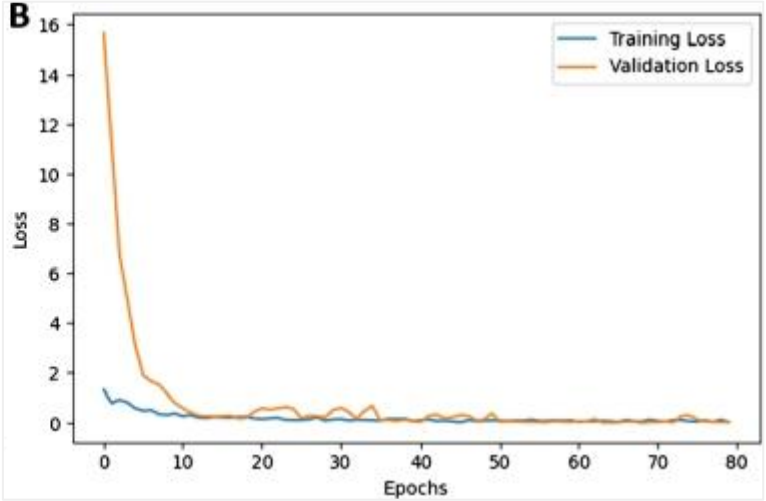
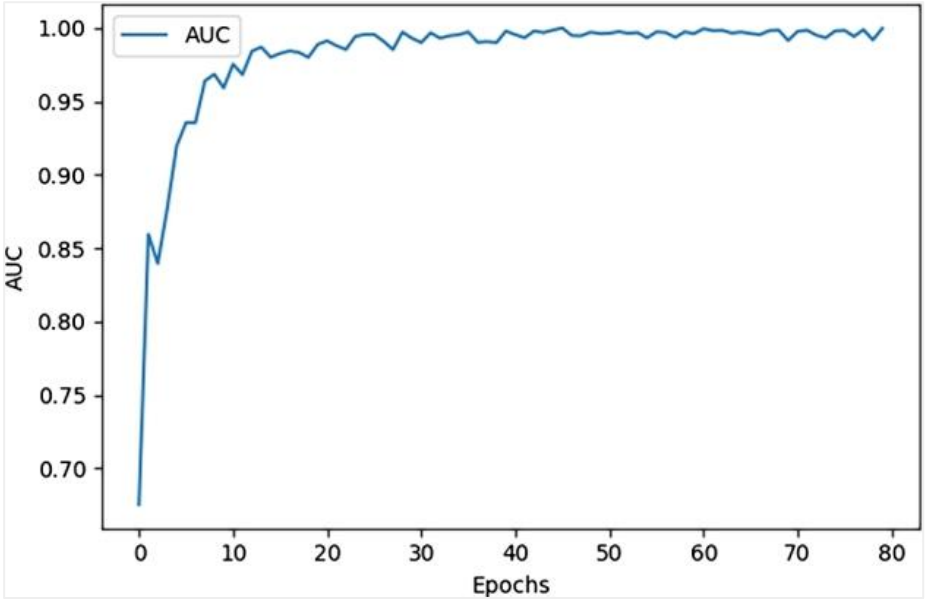


Fig. Loss Results for the Proposed Model



Result 3: AUC Results

- The figure presents the area under the curve (AUC) for running the proposed model with various numbers of epochs.
- The best accuracy was achieved at 80 epochs and a learning rate of 0.05



DISCUSSION

DISCUSSION

- The direct concatenation of different in-nature data with different scales may lead to bias or inconsistent prediction model performance.
- Earlier models in transforming multi-omics data into latent variables used PCA as an embedding technique, which implicitly assumes the linear relationships among the features.
- To address the limitations of previous approaches, this study proposes PaCMAP dimension reduction technique as an embedding method to transform the various omics into latent space before merging them into the classification model(CNN).
- The generated map reduces the 16-dimensional features into a visualized 2-dimensional picture. The visual space represents the relationships between the extracted 16 genes in the x-axis and y-axis style. The value from each omic of the sample contributes to the map by coloring a channel in the RGB system.
- The performance measurements of the proposed model surpassed those of the state-of-the-art i-SOM-GSN model by leveraging the integration of multi-omics data and the CNN architecture.

CONCLUSION

CONCLUSION

- This study presents a novel methodology that combines copy number alteration (CNA), DNA methylation, and gene expression data to predict the malignancy Gleason score for individuals with prostate cancer.
- A visually informative representation of the integrated data was generated by successfully merging the three omics data into a two-dimensional (2D) map using the PaCMAP dimensionality reduction technique and utilizing the RGB coloring scheme. This colored 2D map was fed into a CNN to predict the Gleason score class.
- The model highlights the efficacy of multi-omics data integration in predicting health outcomes.
- The proposed methodology, which combines PaCMAP for DR, RGB coloring for visualization, and CNN for prediction, provides a comprehensive framework for integrating heterogeneous omics data and improving predictive accuracy.
- The authors highlight that future work will embed sequence data and sparse data, including time series and mutation frequency, respectively, to validate the model to more data types.

Q&A

BIBS

THANK YOU FOR YOUR ATTENTION