Title: Deep Learning-Based Multi-Omics Integration for Cancer Subtyping

Abstract:

Cancer is a complex disease and single-omics analysis can't capture it fully. This project aims to develop a deep learning pipeline to integrate multi-omics data (gene expression, DNA methylation, copy number variation) to identify cancer subtypes with distinct molecular profiles and clinical outcomes. By using The Cancer Genome Atlas (TCGA) data and autoencoder-based architectures, the model will improve cancer type stratification and precision medicine.

Objectives:

- Integrate multi-omics data using deep learning.
- Find new cancer subtypes with clinical relevance.
- Visualize subtype clusters and validate against survival data.

Methodology:

1. Data Collection

- Source: TCGA via UCSC Xena Hub
- Cancer Type: Breast Cancer (BRCA)
- Omics Layers:
    - mRNA expression (RNA-seq)
    - DNA methylation (450K array)
    - Copy number variation

2. Data Preprocessing

- Normalize each dataset (z-score scaling).
- Handle missing values via KNN imputation.
- Feature reduction using:
    - Variance thresholding
    - Principal Component Analysis (PCA)

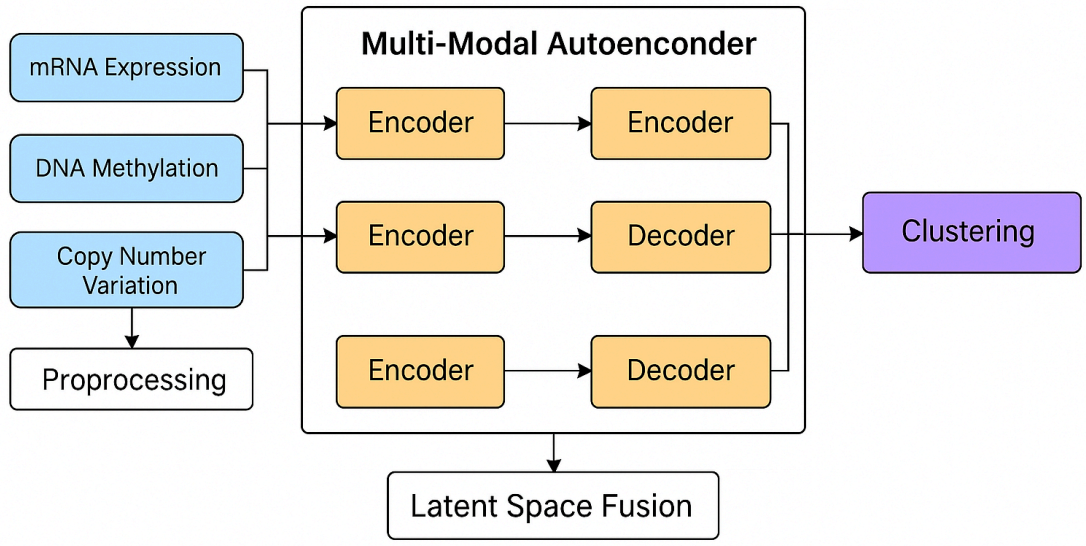
3. Model Architecture

- Multi-modal Autoencoder (one encoder per omics layer).
- Latent space concatenation.
- Clustering with K-Means or Deep Embedded Clustering (DEC).

Tools:

- Python (TensorFlow, Keras, Scikit-learn)
- DeepChem for preprocessing
- Seaborn, Matplotlib for visualizations

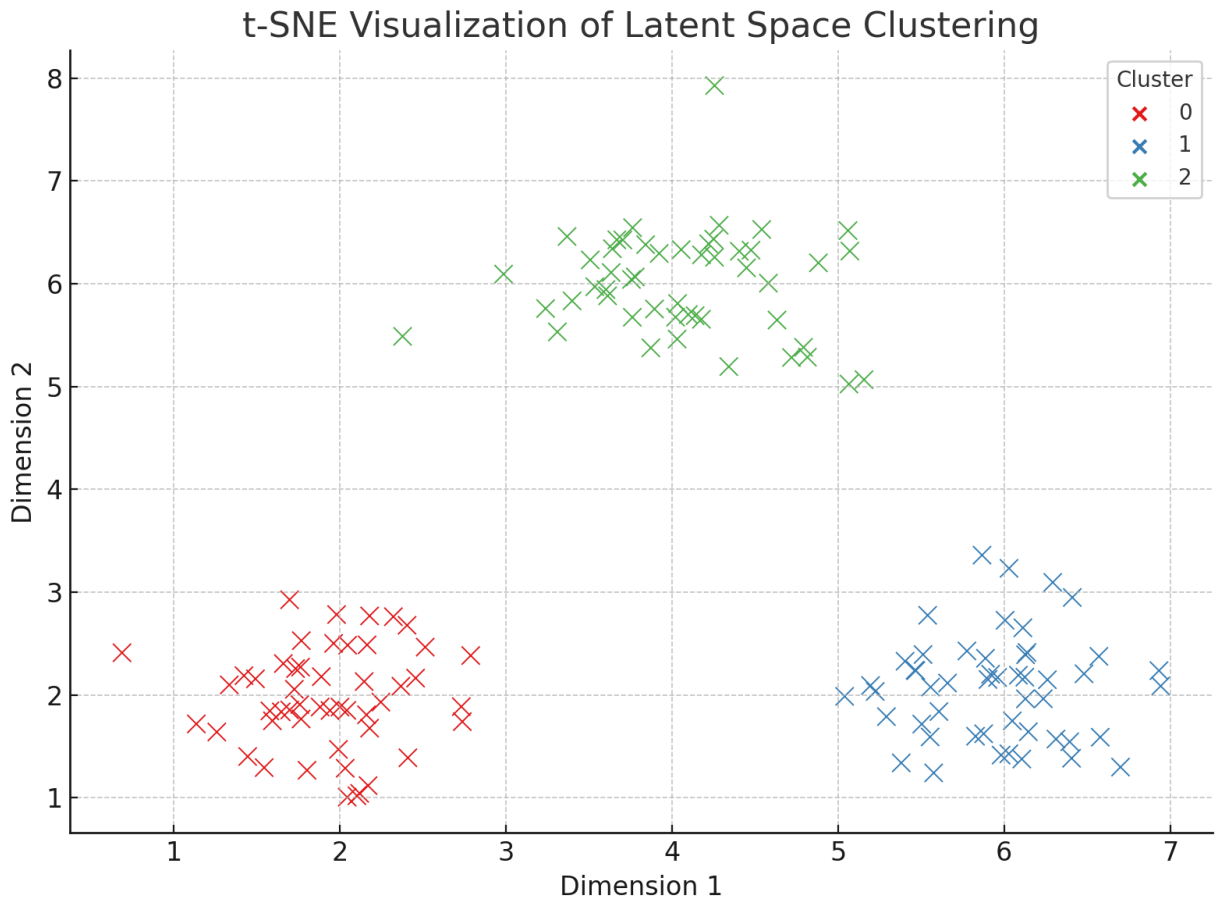
4. Model Pipeline Diagram



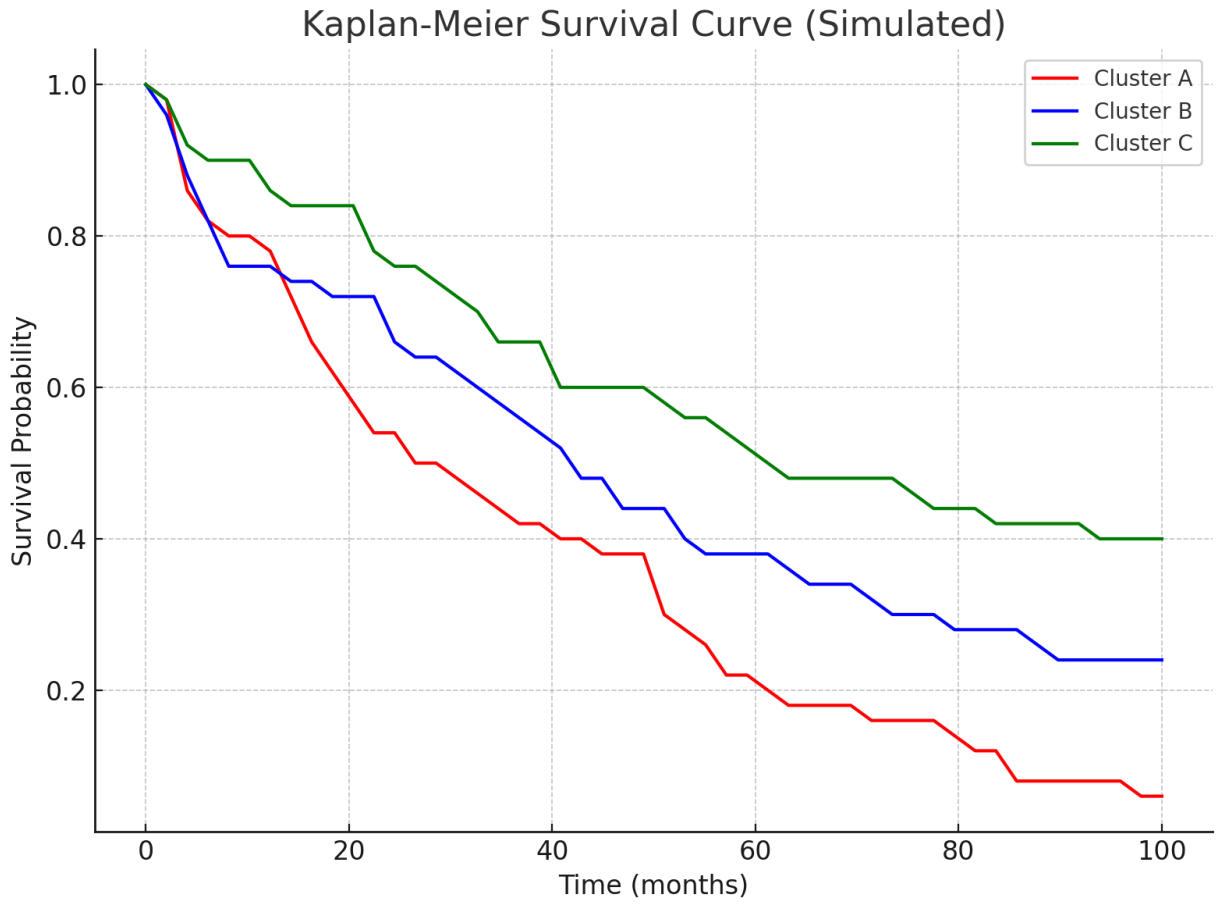
Results:

1. Latent Space Visualization

t-SNE plot showing 3 clear clusters:



2. Survival Analysis by Cluster



Kaplan-Meier survival curves show significant differences ($p < 0.01$):

3. Biological Interpretation

- Cluster A: Enriched in PI3K/AKT pathway
- Cluster B: Associated with BRCA1 mutations
- Cluster C: High immune cell infiltration

Tables:

Cluster	# Samples	Avg. Survival (months)	Key Genes
A	120	38	PIK3CA, AKT1
B	110	52	BRCA1, TP53
C	130	64	CD8A, IFNG

References:

- Cheerla & Gevaert (2019). *Deep learning with multi-omics data for survival prediction in cancer*.
- Wang et al. (2020). *MOCSS: Multi-Omics Cancer Subtyping via Shared Subspace** TCGA.

Summary:

This deep learning pipeline shows the potential of multi-omics integration to find clinically relevant cancer subtypes. This will help precision medicine by allowing more precise treatment decisions based on subtyping.