Objective

To discover novel subtypes of brain tumors using unsupervised machine learning, leveraging genomic or proteomic data. This can provide insights into tumor biology and aid in developing personalized treatments.

Key Steps

1. Data Collection

Datasets:

Gene Expression Omnibus (GEO):

GSE16011: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16011

GSE108474: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE108474

Normalize (min-max or z-score) and standardize gene expression values.

2. Dimension Reduction

Dimensionality Reduction:

Use PCA to reduce noise and retain the most significant features.

Gene Selection:

Identify highly variable genes or features using variance thresholds or correlation matrices.

3. Clustering Techniques

Algorithms:

K-Means: Simple and efficient for clustering high-dimensional data. Optimize the number of clusters using the elbow method or silhouette scores.

Hierarchical Clustering: Useful for understanding relationships between clusters.

Gaussian Mixture Models (GMM): Probabilistic clustering that accounts for data variance within clusters.