Unsupervised Machine Learning-Based Novel Gene Identification in Subtype of Brain Tumors Using Gene Expression Data

A project report submitted
In partial fulfilment of the requirements of Project Work II &
Dissertation for 7th Semester of the degree of

Bachelor of Technology

In

Computer Science & Engineering

Submitted By

Alisha Khatun

Vidita Mishra

Swikriti Mondal

Jahir Khan

Under the guidance of

Mousumi Biswas, Dept. of CSE

Department of Computer Science & Engineering

Hooghly Engineering & Technology College, Hooghly

Affiliated to

Maulana Abul Kalam Azad University of Technology, West Bengal

2024-25

HOOGHLY ENGINEERING & TECHNOLOGY COLLEGE
Vivekananda Road, Pipulpati, P.O. &Dist-Hooghly (WB), Pin-712103
(Affiliated to Maulana Abul Kalam Azad University of Technology)

CERTIFICATE

This is to certify that the project entitled **"Unsupervised Machine Learning-Based Novel Gene Identification in Subtype of Brain Tumors Using Gene Expression Data"** has been submitted in partial fulfilment of the requirement for Project Work II & Dissertation for 7th Semester (4th year) of the degree of **Bachelor of Technology** in **Computer Science and Engineering** by the following B.Tech(CSE) final year students under the supervision of **Mousumi Biswas,** Professor, CSE Department, during a period from **July, 2024 to January, 2025**.

Student Name (with University Roll No.)

1. Alisha Khatun	(17600121014)
2. Vidita Mishra	(17600121032)
3. Swikriti Mondal	(17600121020)
4. Jahir Khan	(17600121067)

....

Project Guide

Department of CSE

....

Prof.(Dr.) Biswajit Halder

HOD, Department of CSE

....

Prof. (Dr.) Bhabani Prasanna Pattanaik

Principal of HETC

ACKNOWLEDGEMENTS

It gives us immense pleasure to announce the partial completion of our project on **"Unsupervised Machine Learning-Based Novel Gene Identification in Subtype of Brain Tumors Using Gene Expression Data"** and we are pleased to acknowledge our indebtedness to all the persons who directly or indirectly contributed in the development of this work and who influenced our thinking, behaviour and acts during the course of study.

We are thankful to our HOD, Prof.(Dr.) Biswajit Halder who granted all the facilities of the college to us for the fulfilment of the project.

We are thankful and express our sincere gratitude to our project guide **Mousumi Biswas** who gave her valuable time to us for the sake of our project. She helped us each and every aspect of our project both academically and mentally.

Finally the team expressed their gratitude to our respected Principal without his support our project would not have seen the light of success.

Student Name (with Roll No.)	**Signature**
1. Alisha khatun(17600121014)	....
2. Vidita Mishra(17600121032)	....
3. Swikriti Mondal(17600121020)	....
4. Jahir Khan(17600121067)	....

4th year, 7th Semester
Department of Computer Science & Engineering
Hooghly Engineering & Technology College
Academic Year: 2023-24

ABSTRACT

Brain tumors, especially glioblastomas, represent one of the most aggressive and heterogeneous types of cancer, characterized by complex molecular and genetic diversity. Understanding this heterogeneity is critical for advancing personalized treatment strategies and identifying novel therapeutic targets. This project leverages unsupervised machine-learning techniques to identify novel genes associated with specific brain tumor subtypes using the gene expression datasets GSE16011 and GSE108474. These datasets encompass a wide range of transcriptomic profiles, providing a robust foundation for analyzing tumor heterogeneity and uncovering biologically significant patterns.

The workflow begins with comprehensive preprocessing, including normalization of raw gene expression data, filtering highly variable genes, and reducing dimensionality using Principal Component Analysis (PCA) to mitigate noise and computational complexity. Subsequently, clustering algorithms—K-Means Clustering, Hierarchical Clustering, and Gaussian Mixture Models (GMM)—are applied to identify distinct molecular subtypes. The optimal number of clusters is determined using quantitative metrics such as the silhouette score, Calinski-Harabasz index, and elbow method.

To visualize the high-dimensional data and clustering results, techniques such as dendrograms are employed, revealing meaningful groupings and relationships between tumor subtypes. Post-clustering, differential gene expression analysis identifies genes uniquely expressed in each subtype. Functional annotation and pathway enrichment analyses are conducted using tools like DAVID, Enrichr, or GO analysis, linking these genes to critical pathways, including cell proliferation, apoptosis, and angiogenesis.

The study provides a detailed map of glioblastoma heterogeneity, offering insights into subtype-specific molecular mechanisms. It identifies novel gene markers with potential applications in precision medicine, including targeted therapy development and prognostic modeling. By combining advanced unsupervised learning algorithms with transcriptomic data, this research highlights the transformative role of computational methods in enhancing our understanding of complex diseases like brain cancer.

INDEX

FIGURE INDEX

TABLE INDEX

INTRODUCTION

Glioblastoma (GBM), the most common and aggressive form of primary brain cancer, represents a significant clinical challenge. Classified as Grade IV by the World Health Organization (WHO), GBM is characterized by rapid proliferation, diffuse invasion into surrounding brain tissue, and intrinsic resistance to conventional therapies, including surgical resection, radiation, and chemotherapy [**1**]. This resistance, coupled with the tumor's aggressive nature, results in a dismal prognosis, with a median survival of only 12 to 15 months post-diagnosis. A key factor contributing to this poor outcome is the substantial molecular heterogeneity observed within GBM tumors, spanning genetic, transcriptomic, and epigenetic landscapes. This heterogeneity complicates the identification of effective therapeutic targets and hinders the development of personalized treatment strategies. Therefore, a deeper understanding of the molecular underpinnings of GBM heterogeneity is crucial for advancing research and improving patient care.

The concept of molecular heterogeneity in GBM is well-established. Landmark studies, leveraging gene expression profiling, have identified distinct molecular subtypes, including proneural, neural, classical, and mesenchymal [**2**]. These subtypes exhibit unique clinical characteristics, differential responses to therapy, and varying prognoses. For instance, the proneural subtype is often associated with younger patients and better survival outcomes, while the mesenchymal subtype is linked to necrosis, inflammation, and worse prognosis. However, even within these established subtypes, significant intra-tumor heterogeneity exists, suggesting that current classifications fail to fully capture the complex biological landscape of GBM. This necessitates more refined and nuanced analyses to uncover novel subtypes or refine existing ones, ultimately providing a more comprehensive understanding of the molecular mechanisms driving GBM progression and therapeutic resistance.

Advancements in high-throughput technologies, such as microarray and next-generation sequencing (NGS), have revolutionized our ability to analyze tumor transcriptomes. These technologies generate large-scale gene expression datasets, such as GSE16011 [**3**] and GSE108474 [**4**], which offer a wealth of information for exploring GBM molecular diversity. GSE16011, derived from The Cancer Genome Atlas (TCGA) project, represents a comprehensive collection of GBM samples with associated gene expression data and clinical information. GSE108474 provides additional insights by encompassing diverse tumor samples and experimental conditions. These datasets provide a robust foundation for investigating tumor heterogeneity and identifying subtype-specific biomarkers. However, the high-dimensional nature of these datasets presents significant analytical challenges for traditional statistical methods.

To address these challenges, unsupervised machine learning techniques have emerged as powerful tools for analyzing high-dimensional gene expression data. Unlike supervised learning methods, which require labeled data, unsupervised algorithms identify inherent structures and patterns within the data without predefined categories. This makes them particularly well-suited for exploratory analyses, such as clustering tumor samples into distinct molecular subtypes. Among the various unsupervised learning techniques, K-Means Clustering, Hierarchical Clustering, and Gaussian Mixture Models (GMM) are commonly employed in cancer research. K-means clustering partitions data into a pre-defined number (k) of clusters by minimizing the within-cluster variance. Hierarchical clustering builds a hierarchical tree-like structure (dendrogram) of clusters, representing nested groupings. GMM offers a probabilistic approach, assuming that the data is generated from a mixture of Gaussian distributions, allowing for more flexible cluster shapes and handling of overlapping clusters.

This study focuses on applying these unsupervised learning techniques to gene expression datasets GSE16011 and GSE108474 to identify novel molecular subtypes of GBM. The analysis begins with data preprocessing, including normalization to remove technical variability, filtering of low-expressed genes, and selection of highly variable genes to focus on biologically relevant features. Dimensionality reduction using Principal Component Analysis (PCA) is performed to mitigate the curse of dimensionality and improve computational efficiency. Subsequently, the chosen clustering algorithms are applied to partition the samples into distinct subtypes. The performance of these algorithms is rigorously evaluated using quantitative metrics such as the silhouette score, Calinski-Harabasz index, and the elbow method to determine the optimal number of clusters and ensure cluster quality.

To enhance interpretability, visualization techniques such as dendrograms. Dendrograms are powerful dimensionality reduction techniques particularly suited for visualizing high-dimensional data in two or three dimensions, preserving local data structure. Following clustering, differential gene expression analysis is conducted to identify genes uniquely expressed in each subtype. These subtype-specific biomarkers are then analyzed for functional relevance using gene set enrichment analysis (GSEA) and pathway analysis tools, linking them to key oncogenic pathways such as cell cycle regulation, apoptosis, angiogenesis, and immune response.

The primary objective of this study is to advance our understanding of GBM heterogeneity by uncovering novel molecular subtypes and identifying their associated gene markers. These findings have the potential to inform the development of precision medicine approaches, tailoring treatments to the unique molecular characteristics of each subtype. Moreover, this research highlights the transformative potential of unsupervised machine learning techniques in addressing complex challenges in cancer biology, paving the way for further advancements in translational oncology. By integrating computational approaches with transcriptomic data, this study aims to contribute to the ongoing efforts to improve outcomes for GBM patients and enhance our understanding of this devastating disease.

REVIEW OF LITERATURE

The study of gene expression datasets, particularly in the context of clustering algorithms, has been a pivotal area of research in computational biology. The advent of high-throughput genomic technologies has enabled the generation of large-scale datasets, such as GSE16011 and GSE108474, which serve as invaluable resources for understanding gene activity patterns, identifying biomarkers, and exploring disease mechanisms. This project leverages unsupervised learning methods—k-means clustering, divisive hierarchical clustering, and a proposed Gaussian Mixture Model (GMM)—to analyze gene expression data, aiming to uncover biologically relevant patterns and insights by validation and pathway analysis.

Gene Expression and Its Significance: Gene expression profiling involves measuring the activity levels of thousands of genes simultaneously to understand their roles in biological processes and disease pathways. Foundational studies, including those by Alizadeh et al. (2000) and Golub et al. (1999), have demonstrated the transformative impact of gene expression analysis in cancer classification and prognosis. These studies underscore the importance of clustering methods in grouping genes or samples with similar expression profiles, which is critical for functional annotation, pathway discovery, and clinical decision-making.

Clustering Techniques in Genomics: Clustering algorithms are indispensable tools in gene expression analysis, known for their ability to handle high-dimensional data and reveal underlying biological structures. K-means clustering, described by MacQueen (1967), is one of the most widely used methods due to its simplicity and computational efficiency. It partitions data into a predefined number of clusters by minimizing intra-cluster variance. However, its reliance on the initialization of centroids and sensitivity to outliers are notable limitations.

Hierarchical clustering offers an alternative approach by creating a dendrogram that represents the nested structure of clusters. Divisive hierarchical clustering, as described by Kaufman and Rousseeuw (1990), follows a top-down approach, starting with all data points in a single cluster and recursively splitting them. This method provides a comprehensive view of the dataset's structure, making it particularly suitable for exploratory analysis. Its inclusion in this project complements the results of k-means, offering diverse perspectives on gene expression patterns.

A Gaussian Mixture Model (GMM) extends the capabilities of clustering by assuming that the data is generated from a mixture of Gaussian distributions. Unlike k-means, GMM can model clusters with varying shapes and sizes, providing a probabilistic assignment of data points to clusters. Incorporating GMM into this project offers an additional layer of flexibility and precision in clustering gene expression data.

Evaluation Metrics for Clustering: The effectiveness of clustering algorithms is assessed using metrics such as the Silhouette Score, Calinski-Harabasz Index, and Dunn Index. Silhouette analysis, introduced by Rousseeuw (1987), evaluates the compactness and separation of clusters. The Calinski-Harabasz Index measures cluster dispersion relative to overall data scatter, while the Dunn Index captures the balance between inter-cluster separation and intra-cluster cohesion. These metrics are indispensable for determining the quality and biological relevance of clustering results.

Applications of PCA in Dimensionality Reduction: Dimensionality reduction is a crucial preprocessing step in gene expression analysis to address the high-dimensional nature of the data. Principal Component Analysis (PCA), introduced by Pearson (1901) and further refined by Hotelling (1933), transforms high-dimensional data into a lower-dimensional space while preserving variance. By applying PCA, this project enhances the visualization and computational efficiency of clustering algorithms, ensuring robust performance across diverse datasets.

Previous Work on GSE Datasets: Extensive research has been conducted using datasets such as GSE16011 and GSE108474. GSE16011 has been instrumental in glioblastoma studies, while GSE108474 has provided insights into ovarian cancer. Clustering methods applied to these datasets have revealed co-expression patterns, identified key genes, and classified disease subtypes. These datasets offer significant potential for biomarker discovery and therapeutic development, highlighting their importance.

Gaps and Motivation: While previous studies have demonstrated the utility of clustering in gene expression analysis, challenges remain in selecting optimal algorithms and evaluating their performance. This project addresses these gaps by integrating multiple clustering techniques—k-means, divisive hierarchical clustering, and GMM—providing a comprehensive comparison and deeper insights into gene expression datasets. The analysis of GSE16011 establishes a strong foundation, while future work on GSE108474 will expand the scope and applicability of the findings.

Conclusion: This project builds on a rich body of literature in gene expression analysis and clustering methodologies. By combining classical and advanced clustering algorithms with robust evaluation metrics, it aims to contribute to the understanding of complex genomic datasets. The proposed inclusion of GMM alongside k-means and divisive hierarchical clustering ensures a multifaceted approach to clustering, capable of capturing nuanced patterns in gene expression data. The insights gained from this work have the potential to inform future studies involving GSE108474 and other datasets, advancing the field of computational biology and paving the way for more accurate and biologically relevant clustering techniques.

PROPOSED WORK

The proposed project aims to explore and analyze gene expression datasets using advanced clustering methodologies to uncover biologically meaningful patterns and insights. With the advent of high-throughput genomic technologies, large-scale datasets such as GSE16011 and GSE108474 have become critical resources for understanding gene activity patterns, identifying biomarkers, and exploring disease mechanisms. This project proposes to leverage three clustering techniques—k-means clustering, divisive hierarchical clustering, and Gaussian Mixture Models (GMM)—to analyze these datasets. By integrating these methods, the project seeks to address the inherent complexities of high-dimensional genomic data and provide a comprehensive understanding of gene expression profiles.

The primary objective is to apply these clustering algorithms to the GSE16011 dataset initially and later extend the analysis to GSE108474. K-means clustering, known for its simplicity and efficiency, will serve as a baseline method, offering quick partitioning of the dataset into predefined clusters based on the minimization of intra-cluster variance. However, recognizing its limitations, such as sensitivity to centroid initialization and outlier presence, the project incorporates divisive hierarchical clustering, which provides a top-down, recursive splitting of data points into clusters. This approach builds a tree-like structure that visually represents the nested hierarchy of clusters, making it particularly useful for exploratory analysis.

The project also proposes the inclusion of Gaussian Mixture Models (GMM), a probabilistic clustering technique that assumes the data is generated from a mixture of Gaussian distributions. Unlike k-means, GMM can model clusters with varying shapes, sizes, and densities, offering greater flexibility and precision. The probabilistic nature of GMM allows for a more nuanced assignment of data points to clusters, which is particularly advantageous for the heterogeneous and overlapping clusters often observed in gene expression datasets.

To ensure robust preprocessing of the high-dimensional data, the project will utilize Principal Component Analysis (PCA). PCA will transform the dataset into a lower-dimensional space while retaining the maximum variance, improving computational efficiency and visualization capabilities. This step is crucial to address the curse of dimensionality inherent in gene expression data.

The evaluation of clustering results will be conducted using established metrics, including the Silhouette Score, Calinski-Harabasz Index, and Dunn Index. These metrics will provide insights into the compactness, separation, and overall quality of the clusters, guiding the optimization of algorithm parameters and ensuring the biological relevance of the findings.

The work will begin with an in-depth analysis of the GSE16011 dataset, focusing on uncovering co-expression patterns, identifying key genes, and gaining insights into glioblastoma. The methodology will then be extended to GSE108474, which is associated with ovarian cancer, enabling a comparative analysis of clustering performance across datasets. By integrating multiple clustering algorithms, the project aims to identify the most effective approach for different data characteristics,

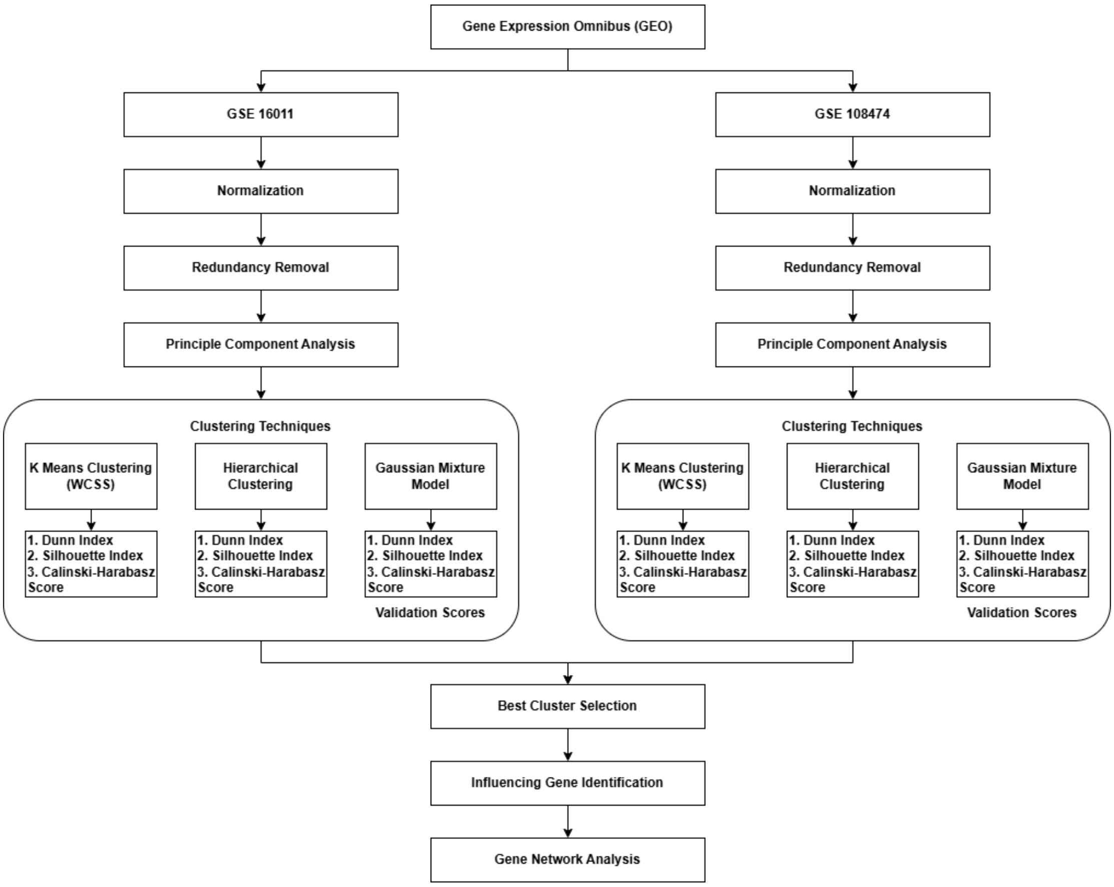

Figure 1: Workflow Diagram

providing a roadmap for similar analyses in future studies.

Additionally, the project will draw on previous research to validate the findings and refine the methodology. The use of foundational studies and resources, such as Wikipedia, Springer, and research papers, will ensure a well-grounded approach. The involvement of the project team and guidance from the faculty advisor will further strengthen the research outcomes.

The proposed work represents a significant step forward in computational biology, offering an integrated framework for clustering-based analysis of gene expression data. The insights gained from this study will not only contribute to a deeper understanding of the datasets under consideration but also pave the way for the application of these methodologies to other genomic datasets, advancing the field of bioinformatics.

PROJECT MODULES & FEATURES

0.1 Data Collection:

The gene expression datasets for this study are obtained from the Gene Expression Omnibus (GEO) database, which serves as a repository for high-throughput gene expression data. For the purpose of this project, two publicly available datasets, GSE16011 and GSE108474, are utilized. Both datasets consist of gene expression profiles from brain tumor samples, including glioblastoma samples. The data collected from these datasets provide insights into the molecular characteristics of brain tumors, including differentially expressed genes, which are crucial for identifying potential biomarkers and molecular subtypes. GSE16011 includes comprehensive data from a variety of glioma samples, while GSE108474 encompasses an additional set of diverse tumor samples, enabling a more robust analysis of glioblastoma subtypes.

0.2 Data Preprocessing:

To ensure that the data sets were suitable for clustering analysis, several preprocessing steps were implemented:

0.2.1 Normalization

Normalization is a crucial step in data preprocessing, as it adjusts for variations across different samples and ensures comparability. Gene expression data often contain systematic biases due to differences in experimental conditions, platforms, or sample preparations. These variations can lead to inconsistencies in data, which can impact the performance of clustering algorithms and downstream analysis. To address this, min-max normalization is applied in this project. Min-max normalization scales the gene expression data within a specific range, typically between 0 and 1. This technique transforms the data such that the minimum value of each gene is mapped to 0, and the maximum value is mapped to 1. This approach helps ensure that all genes are on the same scale, preventing any gene with a higher range of values from dominating the analysis.

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{1}$$

Where,X_{norm} is the normalized value of the gene expression value. X is the original gene expression value. X_{min} is the minimum value of the gene expression in the dataset. $X_{\max}$ is the maximum value of the gene expression in the dataset.

This transformation scales the data for each gene (across all samples) to a range between 0 and 1, ensuring uniformity across the entire dataset. As a result, min-max normalization preserves the relationships between data points, making it easier for clustering algorithms like K-Means, hierarchical clustering, and Gaussian mixture models to detect meaningful patterns in the gene expression data. The normalization process ensures that no gene or feature has a disproportionately large effect on the analysis due to its larger numerical scale, enhancing the reliability of subsequent clustering results.

0.2.2 Redundancy Removal

High-dimensional gene expression data, such as those generated by microarrays and RNA sequencing (RNA-Seq), often contain numerous redundant features (genes) that can obscure meaningful biological patterns and negatively impact downstream analyses, especially clustering and classification. A common approach to address this redundancy is through thresholding based on correlation analysis. This method aims to identify and eliminate highly correlated genes, retaining only a subset of independent, informative features.

The Thresholding Method Based on Correlation:

This method relies on calculating the pairwise correlations between all genes in the dataset and then applying a threshold to identify highly correlated pairs. The process generally involves the following steps:

Correlation Matrix Calculation: The first step is to compute a correlation matrix. This matrix is a square table that shows the pairwise correlation coefficients between all genes. Each cell (i, j) in the matrix represents the correlation between gene i and gene j.

Pearson's Correlation Coefficient: This is the most common correlation measure and is used to assess linear relationships between two variables. It ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating no linear correlation.

Spearman's Rank Correlation Coefficient: This is a non-parametric measure that assesses monotonic relationships (where the variables tend to move in the same direction, but not necessarily at a constant rate). It is less sensitive to outliers than Pearson's correlation.

Threshold Setting: A correlation threshold is selected. This threshold is a value (typically between 0 and 1) that determines the level of correlation considered ”high.” The used threshold is <0.00000, meaning that gene pairs with an absolute correlation coefficient <0.00000 or greater are considered highly correlated.

Impact of Threshold Choice: The choice of threshold is crucial and can significantly impact the number of genes removed. A higher threshold (e.g., 0.95) will result in fewer genes being removed, while a lower threshold (e.g., 0.75) will remove more genes. The optimal threshold depends on the specific dataset, the research question, and the desired level of redundancy reduction. There is no universally ”correct” threshold. Gene Selection and Removal: Once the correlation matrix is calculated and the threshold is set, the matrix is scanned to identify gene pairs exceeding the threshold (in absolute value).

Random Selection: This is the simplest approach. One of the two correlated genes is chosen randomly for removal. Variance-Based Selection: The gene with the lower variance across samples is removed. The rationale is that the gene with lower variance is less informative and contributes less to distinguishing between samples. Prior Biological Knowledge: If there is prior biological knowledge about the genes, this knowledge can be used to inform the selection. For example, if one of the genes is known to be a well-established marker or driver gene in the biological process being studied, the other gene would be removed. Iterative Process (Optional): After removing a gene, the correlation matrix can be recalculated, and the thresholding process can be repeated. This iterative process ensures that all highly correlated gene pairs are addressed, even if the removal of one gene affects the correlations of other genes. The process continues until no remaining gene pairs exceed the correlation threshold.

0.2.3 PCA

Principal Component Analysis (PCA) is a widely-used statistical technique for dimensionality reduction. In the context of gene expression data, PCA serves the dual purpose of reducing the high-dimensional data to a lower-dimensional space while preserving the most significant information. This is particularly important when working with large datasets such as gene expression data, where the number of features (genes) can far exceed the number of samples. Reducing dimensionality helps not only in making the data more manageable but also in improving the performance of subsequent machine learning algorithms, like clustering, by minimizing the computational burden and eliminating noise.

Objective of PCA

PCA transforms the original set of correlated features (gene expressions) into a new set of uncorrelated variables known as principal components (PCs). These principal components are ordered by the amount of variance they capture from the original data. The first principal component captures the largest variance, the second captures the second-largest variance, and so on. By selecting a subset of the principal components that capture the majority of the variance (e.g., 95%), the dimensionality of the data is reduced while retaining the most informative features.

Mathematical Formulation of PCA

Let X be the original data matrix of size $n \times m$, where n is the number of samples (e.g., patients or tumors) and m is the number of features (e.g., genes). The goal of PCA is to find a new set of axes (principal components) that maximize the variance in the data.

Standardization (Optional)

Before applying PCA, it is common to standardize the data if the features (genes) have different scales. This is done by using min-max normalization technique:

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{2}$$

Where:

- X_{norm} is the normalized value of the gene expression value.
- X is the original gene expression value.
- X_{min} is the minimum value of the gene expression in the dataset.
- X_{max} is the maximum value of the gene expression in the dataset.

Eigenvalues and Eigenvectors

To identify the principal components, the eigenvalues and eigenvectors of the covariance matrix C are computed. The eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$ represent the amount of variance explained by each principal component, and the corresponding eigenvectors $v_1, v_2, \ldots, v_m$ represent the directions of the principal components in the original feature space.

Mathematically, this can be expressed as:

$$Cv_i = \lambda_i v_i \tag{3}$$

Where v_i is the eigenvector corresponding to the eigenvalue λ_i, representing the direction of the i-th principal component.

Selecting Principal Components

Once the eigenvectors and eigenvalues are obtained, the eigenvectors are sorted in decreasing order of their eigenvalues. The top k eigenvectors, corresponding to the k largest eigenvalues, form the new feature space (principal components). The choice of k is typically based on the proportion of variance we wish to capture (e.g., 95% of the variance).

The data is then projected onto these top k eigenvectors:

$$X_{\text{PCA}} = X_{\text{standardized}} V_k \tag{4}$$

Where V_k is the matrix of the top k eigenvectors, and X_{PCA} is the transformed data matrix with reduced dimensionality.

Explanation of Principal Components used in project:

First Principal Component (PC1)

The first principal component corresponds to the direction in the data that has the largest variance. It captures the most significant pattern in the data.

Second Principal Component (PC2)

The second principal component is orthogonal to the first and captures the next largest variance in the data. It may correspond to a different biological or technical pattern that was not captured by the first component.

Subsequent Components: Each additional principal component captures less variance and represents increasingly smaller patterns in the data. Typically, only the first few principal components are used, as they capture the majority of the variation in the dataset.

0.3 Machine Learning Models-Clustering Analysis:

Three unsupervised clustering algorithms—K-Means, Hierarchical Clustering, and GMM—were employed to group genes based on their expression patterns. Each algorithm was executed independently, and the results were evaluated using internal validation metrics.

0.3.1 K-Means Clustering

K-means clustering is a widely used unsupervised machine learning algorithm for partitioning a dataset into distinct clusters based on the similarities among the data points. The algorithm works by assigning each data point to the nearest centroid, which represents the mean of the data points within the cluster. Initially, the centroids are chosen randomly, and then the algorithm iterates through two key steps: assigning each data point to the closest centroid and recalculating the centroids as the mean of the data points in each cluster. This process continues until the centroids no longer change significantly, or a predefined number of iterations is reached. K-means clustering is particularly useful for finding patterns in datasets where the number of clusters is predefined, making it an effective method for analyzing gene expression data in the context of brain tumor subtypes. The reason K-means is used in this project is that it provides a simple, computationally efficient way to classify samples into groups based on their gene expression profiles. Given the large number of features (genes) in gene expression data, K-means helps reduce the complexity by grouping similar samples together, making it easier to identify distinct subtypes of brain tumors. By partitioning the data into clusters, K-means allows for the identification of meaningful patterns in gene expression that may correlate with tumor subtypes, enabling further analysis of gene markers and pathways specific to each subtype. Furthermore, K-means clustering has the advantage of scalability, which is essential when dealing with large datasets such as those from the Gene Expression Omnibus (GEO) database, where the number of samples and features can be substantial. This method's ability to handle high-dimensional data efficiently makes it a suitable choice for the analysis of gene expression datasets in cancer research.

Objective of K-Means: The primary objective of the K-means algorithm is to minimize the within-cluster sum of squares **(WCSS)**, which quantifies the compactness of the clusters. This can be mathematically expressed as:

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \tag{5}$$

Where:J is the objective function (WCSS), C_k is the set of points assigned to cluster k, μ_k is the centroid of cluster k, $\|x_i - \mu_k\|^2$ is the squared Euclidean distance between data point x_i and centroid μ_k.

K-Means Algorithm Steps

Step 1: Initialization

Randomly choose K initial centroids $\mu_1, \mu_2, \dots, \mu_K$ from the dataset.

Step 2: Assignment

For each data point x_i, assign it to the nearest centroid μ_k by computing the Euclidean distance:

$$\text{assign}(x_i) = \arg\min_k \|x_i - \mu_k\| \tag{6}$$

Where assign(x_i) denotes the cluster to which x_i is assigned.

Step 3: Update

After all data points have been assigned to clusters, update the centroid for each cluster C_k by computing the mean of all points in that cluster:

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i \tag{7}$$

Where $|C_k|$ is the number of points in cluster C_k.

Step 4: Convergence

Repeat the assignment and update steps until the centroids do not change significantly or a predefined number of iterations is reached. The algorithm terminates when convergence is achieved.

0.3.2 Hierarchical Clustering

Divisive hierarchical clustering is a top-down clustering method that starts with all data points in a single cluster and recursively splits it into smaller subclusters. This method contrasts with agglomerative hierarchical clustering, which begins with individual data points as separate clusters and merges them progressively. The advantage of divisive clustering lies in its approach of exploring the global structure of the data first before narrowing down to specific patterns. In this project, divisive hierarchical clustering is particularly useful due to the complexity and high dimensionality of gene expression data, where genes may exhibit intricate relationships and variations across samples. By using divisive clustering, the algorithm can begin with all gene expression data grouped together, ensuring that the initial splits are made based on the overall structure of the dataset. The recursive partitioning process identifies natural groupings within the data, which can reveal significant biological subtypes or patterns of gene expression that may not be apparent in a more granular clustering approach. Additionally, divisive clustering works well for this project because it is less sensitive to outliers in early stages. The resulting dendrogram from divisive clustering offers a hierarchical view of the relationships between the clusters, enabling an intuitive understanding of how gene expression profiles relate to various tumor subtypes. This is essential for identifying novel subtypes of brain tumors, as it allows for the exploration of the data from a high-level view, gradually zooming in on smaller, more homogeneous groups that may correspond to distinct biological phenomena or tumor characteristics.

Objective of Divisive Clustering

The goal of divisive clustering is to start with the entire dataset and recursively partition it into smaller clusters. At each step, the most dissimilar cluster is selected and split into two subclusters. The splitting process continues until the desired number of clusters is achieved.

Mathematical Formulation

Let $X = \{x_1, x_2, \ldots, x_n\}$ represent the set of n data points. Initially, all data points are in one cluster, $C_1 = X$. The algorithm follows these steps:

Cluster Splitting

At each iteration, we choose the most dissimilar cluster C_i (the cluster with the largest variance) and split it into two subclusters, C_{i1} and C_{i2}. The splitting is done by minimizing the within-cluster variance while maximizing the between-cluster variance. This can be expressed mathematically as:

$$\text{maximize} \quad \frac{\|C_i - C_{i1}\|^2 + \|C_i - C_{i2}\|^2}{\|C_i\|^2} \tag{8}$$

Where: - C_i is the current cluster, - C_{i1} and C_{i2} are the two new subclusters, - $\|C_i - C_{i1}\|^2$ and $\|C_i - C_{i2}\|^2$ are the distances between the original cluster and the two subclusters.

The splitting criterion is aimed at minimizing the variance within each subcluster and maximizing the distance between them.

Repeat

The splitting process continues, selecting the next most dissimilar cluster at each iteration, until the stopping criterion is met. The algorithm can be stopped when the desired number of clusters is obtained, or when further splits result in minimal variance.

Dendrogram Construction

Similar to agglomerative clustering, divisive clustering also produces a dendrogram, a tree-like structure that visually represents the hierarchical relationships between the clusters. The height of each node in the dendrogram indicates the level of dissimilarity at which the clusters were split.

0.3.3 GMM

Gaussian Mixture Model (GMM) clustering is a probabilistic model that assumes that the data is generated from a mixture of several Gaussian distributions, each representing a different cluster. Unlike traditional clustering methods like K-means, which assign each data point to exactly one cluster, GMM allows for soft clustering, where each data point has a probability of belonging to each cluster. This flexibility is especially useful when the data is not linearly separable and when clusters are not necessarily spherical or of equal size, as GMM can model elliptical clusters with different covariance structures. The model works by estimating the parameters of the Gaussian distributions—mean, covariance, and the weight of each distribution—using an Expectation-Maximization (EM) algorithm. This iterative process alternates between assigning data points to clusters based on their probability (Expectation step) and updating the parameters of the Gaussian distributions to best fit the data (Maximization step) until convergence.

In this project, GMM is used for clustering gene expression data from brain tumor samples because it offers several advantages over other clustering methods. Gene expression data often exhibit complex relationships and may not adhere to simple geometric shapes, making it difficult for K-means to form meaningful clusters. The flexibility of GMM allows it to model more complex, overlapping clusters, which is important for identifying subtle differences in tumor subtypes that may not be captured by other methods. Additionally, GMM's ability to handle different covariance structures in the clusters is particularly beneficial when the variability within gene expression patterns differs across tumor subtypes. By leveraging GMM, the project can more effectively group similar tumor samples based on their gene expression profiles, helping to uncover novel insights into the molecular subtypes of brain tumors and potentially identifying new biomarkers for diagnosis or treatment. The steps for GMM clustering are as follows:

- **Initialize the parameters**: the number of clusters K, means μ_k, covariances Σ_k, and mixture weights π_k

  repeat

  E-step: Compute the posterior probabilities (responsibilities) γ_{ik} for each data point x_i and cluster k using Bayes' theorem:

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)}$$

  where the Gaussian distribution $\mathcal{N}(x_i|\mu_k, \Sigma_k)$ is given by:

$$\begin{aligned}\mathcal{N}(x_i|\mu_k, \Sigma_k) = & \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \\ & \exp\left(-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right)\end{aligned} \tag{9}$$

  M-step: Update the parameters of the Gaussian components:

$$\mu_k = \frac{\sum_{i=1}^{N} \gamma_{ik} x_i}{\sum_{i=1}^{N} \gamma_{ik}}$$

$$\Sigma_k = \frac{\sum_{i=1}^{N} \gamma_{ik}(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{N} \gamma_{ik}}$$

$$\pi_k = \frac{\sum_{i=1}^{N} \gamma_{ik}}{N}$$

```
    where $N$ is the total number of data points.
until Convergence of the log-likelihood or other stopping criteria.
```

0.4 Cluster Validation Techniques:

0.4.1 Dunn Index

The Dunn Index is a prominent metric for assessing the quality of clustering, particularly in cases where the goal is to achieve well-separated and compact clusters. It is designed to quantify the ratio between the smallest distance separating points in different clusters (inter-cluster separation) and the largest distance among points within the same cluster (intra-cluster compactness). This dual focus makes the Dunn Index a valuable tool for evaluating the effectiveness of clustering algorithms, as it explicitly emphasizes both separation and compactness. In the context of this project, where clustering methods are applied to high-dimensional gene expression data from datasets such as GSE16011 and GSE108474, the Dunn Index serves as an essential metric to validate the biological relevance and coherence of the identified clusters. These clusters aim to represent distinct subtypes of brain tumors based on gene expression patterns. The Dunn Index is defined as:

$$D = \frac{\min_{1 \le i < j \le k} \delta(C_i, C_j)}{\max_{1 \le l \le k} \Delta(C_l)}$$

Where:$\delta(C_i, C_j)$ is the minimum distance between clusters C_i and C_j. $\Delta(C_l)$ is the maximum intra-cluster distance within cluster C_l. k is the total number of clusters.

In this project, computing the Dunn Index provides a quantitative assessment of the performance of various clustering techniques, such as K-Means, Hierarchical Clustering, and Gaussian Mixture Models (GMM). A higher Dunn Index implies that the clusters are well-separated and internally cohesive, which is critical for accurately distinguishing subtypes of glioblastoma based on gene expression data. By comparing the Dunn Index across different clustering methods, it becomes possible to identify the technique that most effectively separates biologically meaningful groups of genes or samples.

The Dunn Index is particularly advantageous in this study because gene expression data is inherently high-dimensional, making it challenging to distinguish between biologically significant patterns and noise. A robust metric like the Dunn Index helps ensure that the clustering process does not merely capture random variability but instead identifies clusters with clear biological significance. Moreover, its reliance on both intra-cluster compactness and inter-cluster separation makes it well-suited for applications in precision medicine, where accurate classification of disease subtypes can directly impact diagnosis, prognosis, and treatment strategies.

In summary, the Dunn Index is a crucial validation metric in this project for evaluating the quality of clusters generated from gene expression data. By ensuring that the clusters are both compact and well-separated, it provides a strong foundation for downstream analyses, such as identifying key genes and pathways associated with different tumor subtypes.

0.4.2 Silhouette Index

The Silhouette Score is a powerful and intuitive metric used to evaluate the quality of clustering results by analyzing how well each data point fits into its assigned cluster compared to other clusters. It offers a clear, quantitative measure of the cohesion within clusters (how tightly data points are grouped) and the separation between clusters (how distinct the clusters are from one another). The score ranges from -1 to 1, where values closer to 1 indicate that the data points are well-clustered, with each point lying closer to other points in its cluster than to points in any other cluster. A score near 0 suggests overlapping clusters, while a negative score indicates that a data point may be misclassified, being closer to points in a different cluster than its own.

In the context of this project, where clustering techniques are applied to gene expression datasets (e.g., GSE16011 and GSE108474), the Silhouette Score plays a crucial role in determining the effectiveness of the clustering methods. Gene expression data typically have high dimensionality and variability, making it challenging to discern distinct biological patterns or subtypes of diseases such as glioblastoma. The Silhouette Score helps assess whether the identified clusters represent biologically meaningful groupings. For instance, in this project, it is expected that clusters should reflect distinct subtypes of glioblastoma, each characterized by unique gene expression profiles. A high Silhouette Score across the dataset would confirm that these subtypes are well-separated and internally consistent, thereby validating the biological relevance of the clusters.

The Silhouette Score for a data point i is given by:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Where:

- $a(i)$ is the average intra-cluster distance for point i.
- $b(i)$ is the average nearest-cluster distance for point i.

The overall Silhouette Score for a clustering solution is obtained by averaging $S(i)$ over all data points. A score close to 1 indicates well-defined clusters with distinct boundaries, while a score closer to 0 suggests that points are on the boundary of clusters, making it difficult to assign them confidently. A negative score suggests incorrect clustering, where data points are closer to a different cluster than their assigned one.

For this project, the Silhouette Score serves as a vital metric for comparing the performance of different clustering algorithms, such as K-Means, Hierarchical Clustering, and Gaussian Mixture Models (GMM). By computing the Silhouette Score for each clustering result, the method that yields the highest score can be identified as the most effective in grouping gene expression profiles into biologically meaningful clusters. This ensures that the identified clusters are robust, interpretable, and relevant for understanding glioblastoma subtypes and their underlying molecular mechanisms.

0.4.3 Calinski-Harabasz Index

The Calinski-Harabasz Index, also known as the Variance Ratio Criterion, is a widely used metric for evaluating the quality of clustering results. This index measures the ratio of the between-cluster dispersion to the within-cluster dispersion. A higher Calinski-Harabasz Index indicates that the clusters are well-separated and compact, which is desirable in clustering. The metric is particularly advantageous because it takes into account both the compactness of clusters (how tightly data points within a cluster are grouped) and the separation between clusters (how far apart the clusters are from each other). This dual focus makes it a robust and comprehensive tool for assessing the effectiveness of a clustering algorithm.

In the context of this project, which involves clustering gene expression data from glioblastoma datasets (GSE16011 and GSE108474), the Calinski-Harabasz Index is especially useful for quantifying how well the genes have been grouped based on their expression patterns. Clustering in this scenario aims to identify biologically meaningful subtypes of glioblastoma or functional groupings of genes that could provide insights into tumor biology. The ability of the Calinski-Harabasz Index to evaluate the separation and compactness of these clusters ensures that the identified subtypes are both biologically distinct and internally consistent.

The Calinski-Harabasz Score is calculated as:

$$CH = \frac{\text{Tr}(B_k)}{\text{Tr}(W_k)} \cdot \frac{(N-k)}{(k-1)}$$

Where:

- $\text{Tr}(B_k)$ is the trace of the between-cluster dispersion matrix.
- $\text{Tr}(W_k)$ is the trace of the within-cluster dispersion matrix.

- N is the total number of data points.
- k is the number of clusters.

The between-cluster dispersion matrix B_k quantifies how far apart the centroids of different clusters are, reflecting the degree of separation between clusters. The within-cluster dispersion matrix W_k measures how close the data points are to their respective cluster centroids, reflecting the compactness of the clusters.

The scaling factor $(N - k)/(k - 1)$ adjusts for the number of clusters and the size of the dataset, ensuring that the metric does not disproportionately favor solutions with more clusters. This adjustment is critical, as increasing the number of clusters will always reduce the within-cluster dispersion but may not improve the overall clustering quality.

This project calculates the Calinski-Harabasz Index for clustering results obtained using K-Means, Hierarchical Clustering, and Gaussian Mixture Models (GMMs). The clustering approach that best captures the underlying structure of the gene expression data can be identified by comparing the index values across these methods. A high Calinski-Harabasz Index across the datasets would indicate that the chosen clustering method effectively differentiates between glioblastoma subtypes or identifies meaningful patterns in gene expression. This makes the index a vital tool for ensuring the biological relevance and reliability of the clustering results.

0.5 Best Cluster Selection:

The process of best cluster selection is a critical step in clustering analysis, particularly in the context of identifying glioblastoma subtypes and functional gene groupings from gene expression data, as in this project. Clustering algorithms, such as K-Means, Hierarchical Clustering, and Gaussian Mixture Models (GMMs), often require determining the optimal number of clusters, which directly impacts the biological relevance and interpretability of the results. Selecting the best clustering configuration involves evaluating and comparing the results across different clustering techniques and cluster numbers using various validation metrics, including the Dunn Index, Silhouette Score, and Calinski-Harabasz Index. These indices measure the quality of clustering based on compactness, separation, and overall variance distribution, ensuring that the selected clusters are well-defined and biologically meaningful.

For this project, the datasets GSE16011 and GSE108474 are analyzed using the aforementioned clustering techniques. Each technique is run with different numbers of clusters, and the validation indices are computed for each result. The Dunn Index, for instance, evaluates the ratio of the minimum inter-cluster distance to the maximum intra-cluster distance, rewarding clusters that are compact and well-separated. The Silhouette Score measures how similar each sample is to its own cluster compared to others, providing insights into the cohesion and separation of clusters. Meanwhile, the Calinski-Harabasz Index quantifies the ratio of between-cluster dispersion to within-cluster dispersion, capturing both separation and compactness.

The results of these indices are carefully compared to determine the clustering configuration that maximizes these metrics. A high Dunn Index and Calinski-Harabasz Score, coupled with a Silhouette Score close to 1, indicate optimal clustering. This multi-metric evaluation ensures that the selected number of clusters represents the most robust and meaningful partitioning of the data. In this project, selecting the best cluster is vital for accurately identifying glioblastoma subtypes or gene expression patterns associated with tumor biology, potentially contributing to advancements in personalized medicine and therapeutic development. The integration of multiple clustering techniques and validation metrics underscores the rigor of the analysis, ensuring the reliability of the biological insights derived from the clustering results.

0.6 Influencing Gene Identification:

Influencing Gene Identification is a pivotal step in this project, aiming to determine which genes significantly contribute to the clustering of glioblastoma subtypes. After applying clustering algorithms such as K-Means, Hierarchical Clustering, and Gaussian Mixture Models (GMMs) on the gene expression datasets (GSE16011 and GSE108474), it becomes essential to understand the biological drivers of the clusters. Identifying influential genes provides insights into the molecular underpinnings of glioblastoma subtypes, highlighting potential biomarkers or therapeutic targets.

This process begins by analyzing the clustering results in conjunction with the dimensionality reduction performed by Principal Component Analysis (PCA). PCA helps prioritize genes by identifying the ones that contribute most to the principal components, which are the axes capturing the highest variance in the data. These genes are often the key drivers of the observed clustering patterns. By examining the PCA loadings — the coefficients of each gene in the principal components — genes with high absolute values are identified as those with significant contributions to the data's structure. These genes are potential markers that distinguish between the identified clusters.

In addition to PCA, statistical methods like Analysis of Variance (ANOVA) or t-tests are applied to compare gene expression levels across the clusters. ANOVA evaluates the differences in mean expression levels of each gene across multiple clusters to identify genes with statistically significant variations. Similarly, t-tests can compare gene expression between pairs of clusters to uncover genes with distinct expression patterns. These statistical techniques ensure a robust and unbiased identification of influential genes.

Once influential genes are identified, further biological interpretation is conducted. Tools like Gene Ontology (GO) analysis or pathway enrichment analysis are employed to understand the biological processes, molecular functions, or cellular components associated with these genes. This contextual understanding links the identified genes to glioblastoma-specific pathways, aiding in deciphering the biological significance of the clusters.

For example, genes associated with tumor growth, immune response, or metabolic processes may emerge as significant in certain clusters. Such findings could highlight subtype-specific mechanisms of glioblastoma, offering insights into personalized treatment strategies. Additionally, identifying genes with high inter-cluster variability provides an opportunity to validate known biomarkers or discover novel ones that may predict prognosis or treatment response.

In the context of this project, influencing gene identification is not just an analytical step but a bridge between computational clustering and biological interpretation. It ensures that the computational findings are translated into actionable biological insights, making the clustering results relevant for both research and clinical applications in glioblastoma.

0.7 Gene Network Analysis:

Gene Network Analysis is a pivotal step in this project as it aims to uncover the complex relationships and interactions among genes that contribute to the differentiation of glioblastoma subtypes. After clustering gene expression data from the GSE16011 and GSE108474 datasets, the next logical step is to delve deeper into the biological significance of the clusters. Gene Network Analysis provides a framework for understanding how the identified genes interact with one another within the context of cellular processes and biological pathways. This approach not only enhances the interpretability of the clustering results but also allows researchers to pinpoint key regulatory genes, often referred to as hub genes, that play critical roles in tumor progression, subtype differentiation, or response to therapy.

In this project, gene network construction begins by identifying genes with significant contributions to the clusters, often determined through statistical techniques like ANOVA or by analyzing the PCA loadings. Once these influential genes are selected, tools such as STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) or Cytoscape are used to construct a network that maps the interactions between these genes. STRING, for instance, integrates known and predicted protein-protein interaction data, while Cytoscape offers a platform for visualizing and analyzing complex gene interaction networks. The resulting network is typically represented as a graph, where nodes correspond to genes and edges represent interactions between them.

The analysis of the gene network focuses on identifying central genes or hubs, which are genes with a high degree of connectivity within the network. Hub genes are often key regulators of biological pathways and can serve as potential biomarkers or therapeutic targets. For glioblastoma, this is particularly significant, as understanding the gene regulatory networks specific to different subtypes can provide insights into the molecular mechanisms underlying tumor heterogeneity. These insights can guide personalized treatment strategies or lead to the discovery of novel drug targets.

Pathway enrichment analysis is another critical component of gene network analysis. By mapping the clustered genes to known biological pathways (e.g., using databases like KEGG or Reactome), researchers can identify overrepresented pathways within each cluster. This helps in elucidating the biological processes that differentiate glioblastoma subtypes, such as cell cycle regulation, DNA repair, immune response, or metabolic pathways. Identifying these pathways not only deepens our understanding of tumor biology but also highlights potential avenues for therapeutic intervention.

In the context of this project, gene network analysis serves as the bridge between unsupervised clustering and biological interpretation. While clustering groups genes based on expression patterns, network analysis provides a mechanistic understanding of how these genes interact within the larger biological system. This integrative approach ensures that the results of the clustering are not only statistically valid but also biologically meaningful, contributing to the overarching goal of better understanding glioblastoma subtypes and identifying actionable targets for research and treatment.

PLATFORM & TECHNOLOGY

The successful execution of this project relies on a combination of robust platforms and advanced technologies tailored for handling large-scale gene expression datasets and implementing complex clustering algorithms. Google Colab was utilized as the primary computational platform for running Python-based scripts and performing intensive data processing tasks. Its cloud-based environment, equipped with free GPU and TPU resources, provided an ideal solution for managing the high computational demands of clustering large genomic datasets such as GSE16011 and GSE108474. The interactive nature of Colab allowed for seamless integration of code, outputs, and visualizations, facilitating collaborative development and iterative experimentation.

The project extensively employed Python and its rich ecosystem of libraries tailored for data analysis, machine learning, and visualization. Libraries like NumPy and Pandas were pivotal in preprocessing and managing the high-dimensional gene expression data. Scikit-learn was used for implementing clustering algorithms such as k-means and Gaussian Mixture Models (GMM), as well as for calculating evaluation metrics like the Silhouette Score, Calinski-Harabasz Index, and Dunn Index. Dimensionality reduction, a critical step in the analysis, was performed using scikit-learn's PCA module, enabling efficient processing and visualization of complex datasets. For data visualization, libraries like Matplotlib and Seaborn were employed to create detailed plots, including PCA scatter plots and dendrograms, aiding in the intuitive interpretation of results.

Normalization of the gene expression data, a crucial step to ensure the comparability of features and the accuracy of clustering results, was performed using the R programming language. R, renowned for its robust statistical capabilities, provided an efficient and precise framework for handling the variability inherent in genomic data. The integration of R for normalization ensured a rigorous preprocessing pipeline, complementing the downstream analysis performed in Python. The seamless interplay between these platforms and technologies underscores the project's emphasis on leveraging the strengths of each tool to achieve high-quality, reproducible results.

RESULT & DISCUSSION

In this project, K-means clustering was employed to uncover the potential subtypes of glioblastoma based on gene expression data obtained from the GSE16011 and GSE108474 datasets. The primary objective of using K-means clustering was to group patients or tumors into distinct subtypes based on their gene expression profiles, which could offer insights into the underlying biological heterogeneity of glioblastoma.

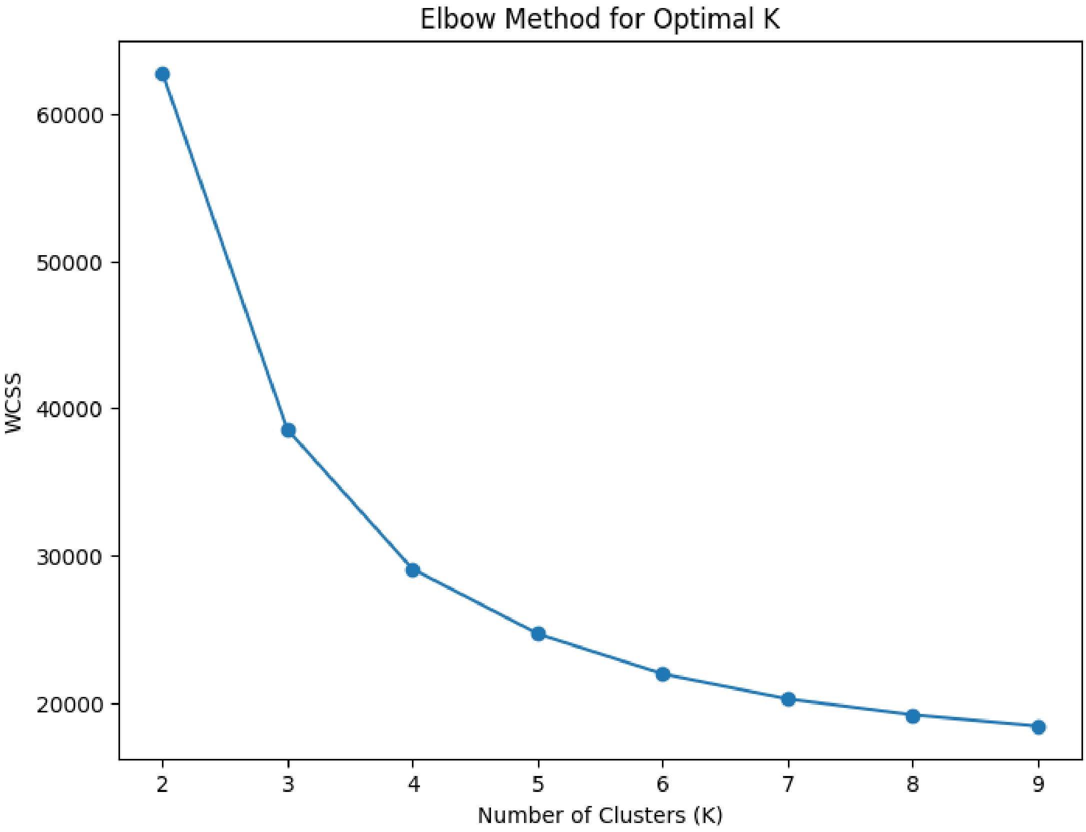

Figure 2: WCSS plot for K-Means

The K-means algorithm partitions the data into k clusters by minimizing the within-cluster sum of squares (WCSS), which reflects the compactness of the clusters. To determine the optimal number of clusters, the Elbow Method was applied, where the WCSS was computed for values of k ranging from 2 to 9, and the point of inflection (elbow) in the WCSS curve was identified.This method revealed that $k = 4$ was the optimal number of clusters, as the decrease in WCSS started to slow significantly beyond this point.

Once the optimal number of clusters was determined, the gene expression data was clustered into four groups. Each group represents a distinct glioblastoma subtype, with unique gene expression patterns that can potentially correspond to different molecular characteristics, prognosis, or therapeutic responses. The clusters

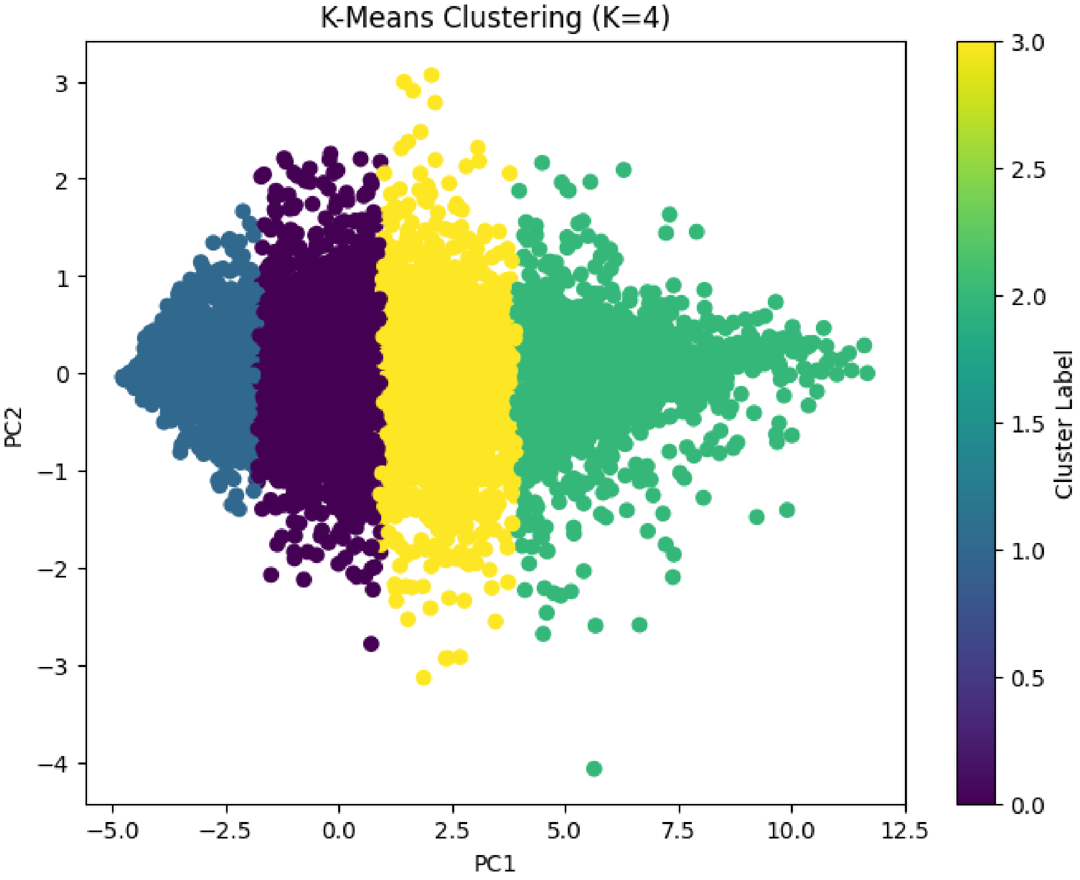

Figure 3: WCSS plot for K-Means (GSE 16011)

were evaluated using various clustering validation metrics, such as the Dunn Index, Silhouette Score, and Calinski-Harabasz Score.

Table 1: Evaluation Metrics K = 4 in K Means Clustering (GSE 16011)

Metric	Score
Silhouette Score	0.4000
Calinski-Harabasz Score	28699.04
Dunn Index	0.0493

The results indicated that the clustering at $k = 4$ produced well-separated and compact clusters, as reflected by high Silhouette and Calinski-Harabasz scores, while the Dunn Index indicated the presence of significant separation between clusters. These validation metrics reinforced the reliability of the clustering solution.

The results were further visualized using PCA, which reduced the high-dimensional gene expression data to two principal components, allowing for a clear and interpretable representation of the three clusters in a two-dimensional space. The clusters were distinctly separated in the plot, with minimal overlap, suggesting that K-means clustering effectively identified biologically meaningful subtypes. The three clusters derived from the K-means algorithm provided valuable insights into the diversity of gene expression patterns within glioblastoma, which may have implications for personalized treatment strategies, prognosis, and understanding of the tumor biology. Overall, K-means clustering proved to be an effective unsupervised learning technique in revealing hidden structures in the gene expression data, contributing significantly to the overall objective of subtype identification in glioblastoma.

The results from the divisive hierarchical clustering in this project provide valuable insights into the structure of gene expression data, revealing underlying patterns and relationships among the genes. Divisive hierarchical clustering, by nature, starts with all data points in a single cluster and progressively splits the data into smaller, more homogenous sub-clusters. This method is particularly beneficial for high-dimensional biological data like gene expression, where relationships between genes or samples can be complex and not immediately apparent. The divisive approach offers a clear, top-down view of how genes can be grouped based on their expression profiles, which is essential for understanding gene co-expression networks or identifying potential biomarkers for diseases.

In this project, the divisive clustering method was applied after reducing the dimensionality of the data using PCA. The resulting clusters highlight genes that exhibit similar expression patterns across the samples, which could correspond to genes involved in the same biological pathways or processes. As the clustering algorithm splits the data, it identifies genes that share similar behaviors, which may suggest that they are co-regulated or have related functions. This is particularly important in the context of gene expression studies, as it helps to focus attention on gene groups that could be functionally significant.

Moreover, the divisive hierarchical clustering method provides a tree-like structure (dendrogram) that illustrates the relationships between genes and clusters. This tree representation not only shows how the data is split at each stage but also offers insights into the hierarchical nature of gene expression patterns. By analyzing the

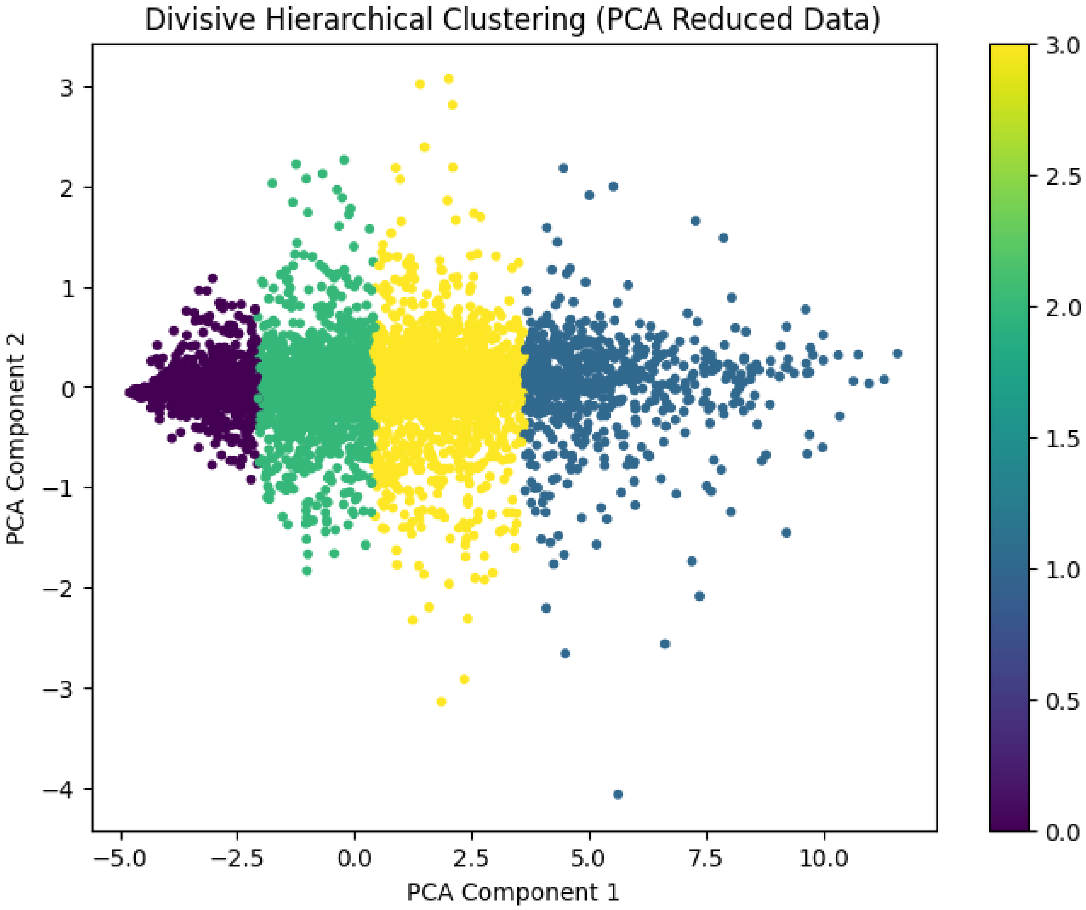

Figure 4: Divisive Clustering Plot with Stopping criteria min = 4 (GSE 16011)

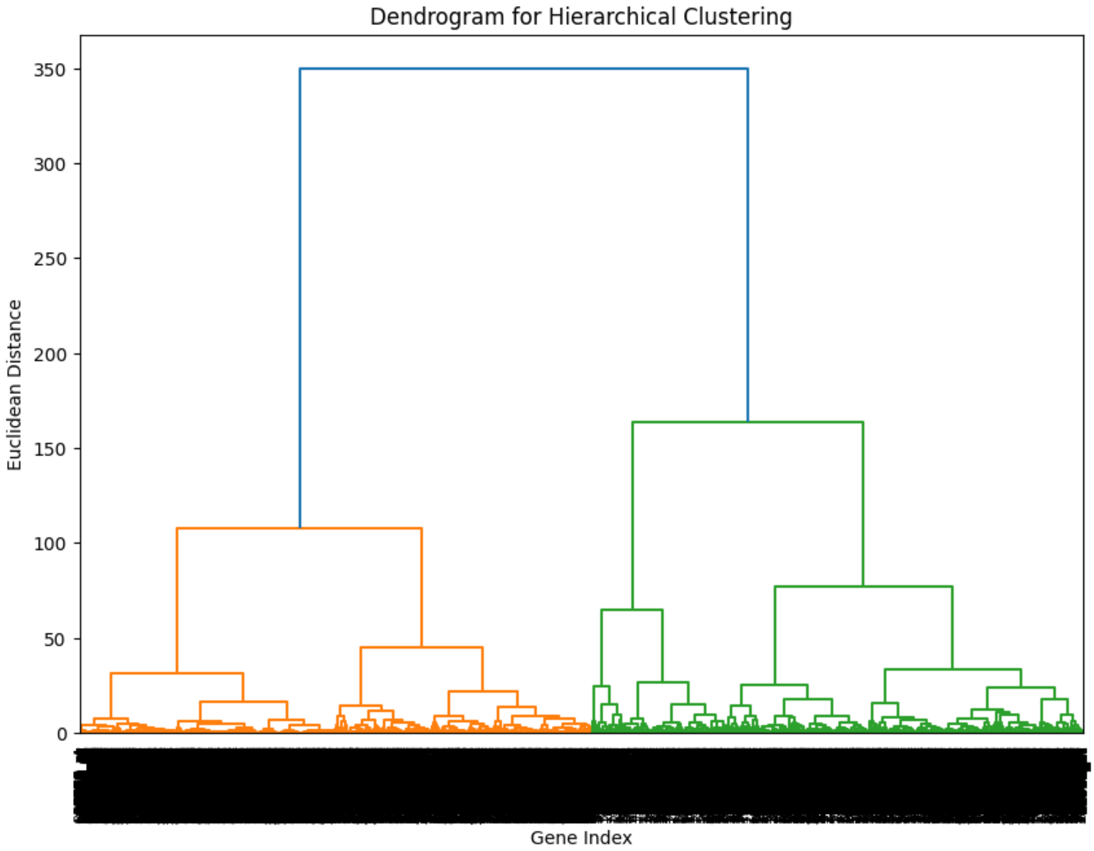

Figure 5: Dendogram for Divisive Clustering (GSE 16011)

dendrogram, one can understand the relative similarities or dissimilarities between different gene groups and how these groups relate to each other in the context of the larger dataset.

Table 2: Evaluation Metrics for Divisive Clustering (GSE 16011)

Metric	Value
Silhouette Score	0.4882
Calinski-Harabasz Score	12653.16
Dunn Index	0.0130

The divisive clustering results from this project can be further explored to identify key gene groups that exhibit coordinated expression changes, which could be indicative of biological phenomena such as disease progression, response to treatment, or the regulation of specific cellular processes. This method, with its ability to handle large datasets and provide hierarchical relationships, serves as a powerful tool for uncovering hidden patterns in gene expression data, offering a deeper understanding of the molecular mechanisms underlying complex biological systems.

In this project, K-Means clustering and divisive hierarchical clustering were applied to the gene expression data from the GSE16011 dataset to identify meaningful clusters within the data. The results from both clustering methods were evaluated using various metrics, providing insights into the clustering performance.

For K-Means clustering, the evaluation metrics demonstrated a promising outcome. The Silhouette Score of 0.4000 indicates a relatively moderate level of separation between the clusters, suggesting that the clusters are somewhat distinct but may overlap to a certain degree. A Silhouette Score closer to 1 would suggest well-separated clusters, and a value closer to -1 would indicate poor clustering. Despite being moderate, this score is generally acceptable in high-dimensional biological datasets where perfect clustering is difficult to achieve due to the complex relationships between genes.

The Calinski-Harabasz Score, which measures the ratio of the sum of between-cluster dispersion to within-cluster dispersion, was calculated at 28,699.03. This high score reflects a good separation of clusters in relation to the spread within each cluster. A higher score indicates that the clusters are well-separated, which is desirable for any clustering analysis, especially in gene expression data, where distinct biological patterns need to be captured.

The Dunn Index, which evaluates the compactness and separation of clusters, returned a value of 0.0493. While the Dunn Index is relatively low, it is important to note that this metric can sometimes show lower values in high-dimensional datasets where the distances between clusters may be harder to define. This value suggests that there could be some overlap between the clusters, indicating room for further refinement of the clustering process.

For divisive hierarchical clustering, the evaluation metrics presented a more nuanced picture. The Silhouette Score of 0.4882 was slightly higher than that of K-Means, which indicates that divisive hierarchical clustering may have produced more distinct clusters, though it still reflects moderate separation. This suggests that hierarchical methods, especially divisive clustering, might provide better structure when dealing with complex gene expression data, as they start with a single cluster and progressively split the data based on inherent patterns.

The Calinski-Harabasz Score for divisive hierarchical clustering was found to be 12,653.16, which is lower than that for K-Means. This could imply that the clusters produced by divisive hierarchical clustering are less well-separated compared to K-Means. However, the method may still offer valuable insights into the data, as hierarchical clustering methods tend to capture fine-grained structures in the dataset.

The Dunn Index for divisive hierarchical clustering was 0.0130, indicating a relatively lower separation between the intra-cluster and inter-cluster distances. This suggests that, like K-Means, the clusters might not be as distinct as desired, but they could still represent meaningful biological structures. Given the high-dimensional nature of gene expression data, such results are not entirely unexpected.

Overall, the evaluation metrics for both K-Means and divisive hierarchical clustering indicate that while the clusters show reasonable separation, there is still room for improvement in terms of the compactness and distinctness of the clusters. The GSE16011 dataset's complexity and inherent noise likely contributed to these results. As for the GSE108474 dataset, it remains untouched for now, and further clustering analysis on this dataset could help validate and refine the findings. These results suggest that clustering, though not perfect, has captured some meaningful patterns in gene expression data, with potential for further optimization and fine-tuning to improve cluster separation and interpretability.

TECHNICAL DETAILS & APPLICATION

This project has significant implications in the field of computational biology and biomedical research, particularly in the analysis of large-scale gene expression datasets. By leveraging unsupervised machine learning methods like k-means clustering and divisive hierarchical clustering, along with advanced evaluation metrics, the project offers diverse applications in understanding complex biological phenomena.

Biomarker Discovery: The primary application of this project is in identifying potential biomarkers for disease diagnosis and prognosis. By analyzing gene expression data from datasets such as GSE16011 and GSE108474, the clustering algorithms can group genes with similar expression patterns, uncovering co-regulated genes or pathways associated with specific diseases. For instance, in glioblastoma (GSE16011) and ovarian cancer (GSE108474), such biomarkers can aid in early detection, personalized therapy, and monitoring disease progression.

Therapeutic Target Identification: Clustering can reveal critical genes or pathways that drive disease mechanisms. This project facilitates the identification of such targets by isolating clusters of genes highly expressed in disease states but minimally expressed in normal conditions. These genes can be investigated further as drug development.

Functional Annotation and Pathway Analysis: The clustering of genes or samples with similar expression profiles can provide insights into gene functions and biological pathways. This project's results can be used to assign functional annotations to uncharacterized genes by associating them with known genes in the same cluster, enhancing our understanding of cellular processes and disease biology.

Integration with Personalized Medicine: The insights generated from the gene expression analysis can be integrated into personalized medicine strategies. By classifying patient samples into distinct molecular subtypes, this project can help tailor treatments based on the molecular profile of the disease, improving therapeutic efficacy.

Data-Driven Hypothesis Generation: The patterns and clusters identified in this project can form the basis for new biological hypotheses. For example, identifying novel gene clusters associated with aggressive forms of cancer can lead to studies exploring their roles in tumor progression and metastasis.

Benchmarking Clustering Methods: By comparing the performance of k-means and divisive hierarchical clustering on GSE16011 and future datasets like GSE108474, this project serves as a benchmark for evaluating clustering algorithms. The inclusion of GMM in future analyses expands the toolkit for researchers, enabling them to select the most suitable method for specific datasets.

Educational and Research Utility: This project also serves as a valuable educational resource, demonstrating the application of advanced machine learning techniques in genomics. Researchers and students can use the methodologies and codebase to explore other datasets, validate findings, or extend the project to include additional machine learning models or evaluation metrics.

Translational Applications in Clinical Research: By bridging the gap between raw gene expression data and actionable insights, this project aids in translating genomic research into clinical practice. Clustering results can guide clinical trial designs by identifying patient subgroups likely to respond to specific treatments, accelerating the development of targeted therapies.

In summary, this project demonstrates the power of clustering algorithms in unraveling the complexities of gene expression data. Its applications span from fundamental research to clinical settings, offering a robust framework for advancing our understanding of diseases and improving patient care. The future inclusion of GSE108474 and additional models like GMM promises to further enhance its utility and impact.

FUTURE SCOPE

The future prospects of this project are promising, with several enhancements and expansions planned to further investigate the clustering of gene expression data and improve the overall robustness and reliability of the results. As the GSE16011 dataset provided valuable insights into the application of K-Means and divisive hierarchical clustering techniques, the next logical step is to include the GSE108474 dataset in the analysis to evaluate the consistency and generalizability of the clustering results. This expanded analysis will allow for a more comprehensive understanding of the clustering patterns, as well as offer the opportunity to assess whether the clustering structures discovered in GSE16011 hold true across different experimental conditions.

The application of both K-Means and divisive hierarchical clustering algorithms on the GSE108474 dataset will be performed in parallel, mirroring the steps taken with the GSE16011 dataset. This dual approach will allow for a direct comparison of the performance of both clustering techniques when applied to an additional gene expression dataset. The primary goal will be to determine whether the clusters observed in the GSE16011 dataset are consistent across GSE108474, or if there are notable differences in the cluster characteristics between the two datasets. This parallel application of clustering algorithms will also provide insight into how each algorithm performs under different data distributions and experimental conditions, further validating the utility of these clustering techniques in gene expression data analysis.

In addition to the K-Means and divisive hierarchical clustering approaches, the future work will also introduce the Gaussian Mixture Model (GMM) for clustering. GMM is a probabilistic model that assumes the data is generated from a mixture of several Gaussian distributions, making it particularly suited for datasets that exhibit complex cluster structures. By incorporating GMM, the project aims to explore more nuanced cluster assignments, especially in cases where clusters overlap or have non-spherical shapes. GMM will be applied to both the GSE16011 and GSE108474 datasets, and the resulting clusters will be compared to those obtained from K-Means and divisive hierarchical clustering. The ability of GMM to capture more flexible cluster shapes will be a key focus, particularly in identifying subpopulations of genes that may not fit well into the rigid cluster boundaries imposed by K-Means and hierarchical clustering.

Once the clustering results for all three algorithms (K-Means, divisive hierarchical clustering, and GMM) are obtained, a thorough analysis will be conducted to compare the effectiveness and accuracy of each method. Key metrics such as the Silhouette Score, Calinski-Harabasz Score, and Dunn Index will be computed for each clustering approach to quantify the quality and separation of the clusters. These metrics will help identify the optimal clustering solution for both datasets, providing a solid foundation for further biological interpretation. Additionally, visualizations such as scatter plots, dendrograms, and heatmaps will be used to visually inspect the cluster structures and compare the results across different methods.

A more detailed analysis of the cluster contents will also be carried out. The genes in each cluster will be examined for potential biological relevance, such as identifying genes that are known to be associated with particular diseases, cellular processes, or pathways. Cross-referencing the clusters with existing gene expression studies or databases will help validate the biological significance of the clustering results. This approach will provide valuable insights into potential biomarkers, gene interactions, and pathways that may be involved in various biological conditions, which could be particularly useful in disease research and therapeutic discovery.

In the final stage of the project, a comprehensive conclusion will be drawn based on the analysis of the clustering results from all three algorithms (K-Means, divisive hierarchical clustering, and GMM) across both datasets (GSE16011 and GSE108474). This conclusion will summarize the performance of each clustering method, highlighting their strengths and weaknesses in terms of cluster separability, compactness, and biological relevance. It will also provide a detailed comparison of the two datasets, offering insights into how the clustering results vary under different experimental conditions. The project will conclude with recommendations for future improvements and potential avenues for further research, such as exploring additional clustering techniques, incorporating domain-specific knowledge, or integrating gene regulatory networks for more accurate cluster assignments.

Ultimately, the expanded analysis using GSE108474, alongside the inclusion of GMM, will offer a more robust and nuanced understanding of gene expression patterns. This will not only validate the findings from the initial analysis but will also contribute to the growing body of research in gene expression analysis and computational biology. By integrating multiple clustering techniques and datasets, the project will provide a more comprehensive tool for gene discovery, disease research, and therapeutic development, with potential applications in precision medicine and drug design.

CONCLUSION

In conclusion, this project aimed to explore the application of clustering techniques, specifically K-Means and divisive hierarchical clustering, to gene expression datasets, focusing on the GSE16011 dataset. The goal was to uncover underlying patterns and relationships between genes, offering insights that could aid in understanding biological processes. Both clustering methods were evaluated based on several key performance metrics, including the Silhouette Score, Calinski-Harabasz Score, and Dunn Index, providing a comprehensive view of the quality and separability of the resulting clusters.

The K-Means clustering algorithm, applied to the dataset, achieved a Silhouette Score of 0.4000, which indicates moderate cluster separation. This suggests that while there is some degree of distinctness between clusters, there is still overlap, meaning that the clusters may not be perfectly well-separated. The Calinski-Harabasz Score of 28,699.03, however, points to a high level of variance between clusters, suggesting that the clusters are well-separated in terms of their spread in the feature space. The Dunn Index of 0.0493, on the other hand, is relatively low, indicating that some inter-cluster distances may be too close, leading to potential overlap. This result points to the inherent difficulty in clustering gene expression data, where high-dimensionality can cause ambiguity in cluster boundaries.

In contrast, divisive hierarchical clustering, a top-down approach, yielded a higher Silhouette Score of 0.4882, suggesting better separation between the clusters compared to K-Means. However, the Calinski-Harabasz Score of 12,653.16 and the Dunn Index of 0.0130 indicate that, despite the seemingly better cluster separation, the overall clustering structure may not be as well-defined as in K-Means, especially when it comes to the compactness and separation of individual clusters. This discrepancy highlights the challenges of hierarchical clustering in high-dimensional spaces, where its tendency to form large, broad clusters may not always align with the true structure of the data.

Despite these challenges, both clustering methods showed potential in uncovering meaningful patterns in the gene expression data. The results reveal that while both K-Means and divisive hierarchical clustering are capable of identifying distinct clusters, there is still room for improvement in terms of cluster separation and compactness. The moderate scores across the evaluation metrics for both clustering approaches suggest that, although the clusters are identifiable, some overlap and ambiguity remain. This could be attributed to several factors, such as the high-dimensional nature of gene expression data, which may lead to noise or subtle relationships between genes that are difficult to separate through these clustering techniques.

Furthermore, the findings are based solely on the GSE16011 dataset, and it is important to note that the GSE108474 dataset remains untouched in this analysis. Expanding the clustering analysis to include this second dataset would be an important next step to validate the robustness and generalizability of the clustering results. The additional dataset could help identify whether the patterns observed in

GSE16011 are consistent across different experimental conditions, further strengthening the validity of the clustering results.

Additionally, the project opens avenues for exploring more advanced techniques to improve clustering performance. One promising direction is the integration of feature selection or dimensionality reduction methods beyond PCA, which could help reduce the noise in the data and highlight more relevant features for clustering. Furthermore, refining the hyperparameters of the clustering algorithms, such as adjusting the number of clusters in K-Means or exploring different linkage criteria in hierarchical clustering, could enhance the separation of clusters.

In summary, this project provides a strong foundation for understanding the clustering of gene expression data, demonstrating that while K-Means and divisive hierarchical clustering can uncover valuable insights, further refinement is necessary. The combination of clustering methods with advanced preprocessing and optimization techniques holds great potential for better capturing the underlying biological patterns in gene expression data. The inclusion of the GSE108474 dataset in future work will offer further validation of the findings, bringing us closer to a comprehensive and meaningful analysis of gene expression data. Ultimately, this work contributes to the growing body of research in computational biology and bioinformatics, with implications for gene function identification, disease diagnosis, and therapeutic discovery.

REFERENCES

[1] Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., ... & Kleihues, P. (2007). The 2007 WHO classification of tumours of the central nervous system. Acta neuropathologica, 114(2), 97-109, https://link.springer.com/article/10.1007/s00401-007-0243-4

[2] Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., ... & Meyerson, M. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer cell, 17(1), 98-110.

[3] Phillips, H. S., Kharbanda, S., Chen, R., Forrest, W. F., Soriano, R. H., Wu, T. D., ... & Haas-Kogan, D. A. (2006). Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and demonstrate response to targeted therapies. Cancer cell, 9(3), 157-173. (This is a key paper related to the GSE16011 dataset and GBM subtyping).

[4] Bhat, K. P. L., Balasubramaniyan, V., Vaillant, B., Ezhilarasan, R., Humayun, I., Ismail, A., ... & Rich, J. N. (2013). Mesenchymal differentiation mediated by NF-kB promotes chemoresistance in glioblastoma. Cancer cell, 24(5), 631-646. (This paper utilizes data similar to GSE108474 and discusses mesenchymal transition in GBM).