Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Identification of cancer associated biomarkers by analysing biologically enriched clusters using MSAGK_CL

Samir Kumar Sett
Techno India Salt Lake
Subir Hazra
Meghnad Saha Institute of Technology
Jadav Chandra Das
Maulana Abul Kalam Azad University of Technology
Anupam Ghosh
Netaji Subhash Engineering College
Sahabul Alam
Brainware University
Arunangshu Pal
arunangshu.pal@jaipur.manipal.edu

Manipal University Jaipur

Article

Keywords: Gustafson-Kessel Algorithm, Fuzzy Clustering, Membership Score Analysis, Validity Indices, Gene Expression Analysis, Adaptive Thresholding

Posted Date: March 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8817220/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Identification of cancer associated biomarkers by analysing biologically enriched clusters using MSA^{GK_CL}

Samir Kumar Sett[1], Subir Hazra[2], Jadav Chandra Das[3], Anupam Ghosh[4], Sahabul Alam[5], Arunangshu Pal[6]

[1] Techno India Salt Lake, WB, India
[2] Meghnad Saha Institute of Technology, Kolkata, WB, India,
[3] Department of Information Technology, Maulana Abul Kalam Azad University of Technology, West Bengal, NH-12, P.O- Simhat, Dist-Nadia, West Bengal-741249, India
[4] Department of Computer Science and Engineering, Netaji Subhash Engineering College, Techno City, Kolkata-700152, India
[5]Department of Computer Science and Engineering, Brainware University, Barasat, Kolkata, West Bengal-700124, India
[6] Department of Artificial Intelligence and Machine Learning, Manipal University Jaipur, Rajasthan-303007, India

samirkumarsett@gmail.com, hazrasubir@gmail.com, jadavchandra.das@makautwb.ac.in, anupam.ghosh@rediffmail.com, sahabul2009@gmail.com, arunangshu.pal@jaipur.manipal.edu

Abstract:
Recent clustering techniques have witnessed the rise of fuzzification systems designed for various medical applications. Computational tools, in the last decades, have considerably empowered the tasks of physicians and biologists in finding cancer-mediating biomarkers. By taking advantage of these benefits, we introduce a fuzzy clustering-based methodology to find genes associated with particular cancers. The work (MSA^{GK_CL}) applies the Gustafson-Kessel (GK) algorithm to gene expression datasets, including normal and carcinogenic states, showing its effectiveness on gene expression profile. The adaptive behaviour of the GK algorithm provides a robust treatment of non-spherical cluster shapes, which is crucial in gene expression analysis. To overcome the problem of deciding the number of clusters optimally, we used cluster validity indices like Xie-Beni (XB), Fukuyama-Sugeno (FS), and Dunn Index. These indices give a quantitative basis for the quality of clustering and help in deciding the best clustering configuration. The second approach was a dynamic threshold-based membership score analysis to identify significant genes. This analysis calculates the maximum absolute differences in membership scores between cancerous and normal datasets using the percentile of these differences to adaptively select the threshold. Its effectiveness is validated by the study's findings through precision, recall, and F1-score metrics, demonstrating its usefulness in discovering genes related to cancer. Such an integrative framework thereby opens up a promising path for future research in the area of cancer genomics, helping identify therapeutic targets for personalized medicine

Keywords: Gustafson-Kessel Algorithm, Fuzzy Clustering, Membership Score Analysis, Validity Indices, Gene Expression Analysis, Adaptive Thresholding

1. Introduction:

The genes control the cell functioning in human body which includes how quickly a cell can grow, along with how often it can be divided and how long it can live. According to researchers, a single cell contains about 30,000 unique genes, and these genes reside on chromosomes within the cell. Genes regulate cell function by producing proteins, enabling these proteins to carry out the cell's specific roles. Cancer starts when one or more genes within a cell undergo erroneous mutations, which are changes that either produce defective proteins or may block the protein production. This might result in uncontrollable growth of cells and it might become cancerous. Researchers study the expression of many genes using various technologies such as 'Microarray Technology' to detect various mutations in gene. The challenge lies in the fact that millions of genes exist in the human body, making it difficult to identify perfectly those associated with a specific cancer. However, gene expression data generated through advanced technologies like DNA sequencing and DNA microarrays have significantly impacted cancer research [1]. As DNA microarray technology advances, measuring the expression levels of thousands of genes becomes available; but, the key challenge is extracting meaningful information and understanding underlying patterns from the enormous gene expression data. [2]. Over the past few decades, various computational methods have led to the development of technologies that

improve cancer identification, significantly aiding biologists and doctors in detecting biomarkers associated with cancer [3]. But there is not much work has been done to understand the associativity of genes in different stages of cancer. Analysing microarray data and selecting the genes which are informative genes is a critical work considering the complexity related with different types of cancer. Advances in biotechnology have enabled the production of vast amounts of data by employing high-density oligonucleotide chips and cDNA arrays which helps in simultaneous measuring of data [4][5]. Despite the availability of this data, suitable algorithms to effectively gather insights and mine information are lacking. Gene expression data from high-throughput methods appears as a matrix with rows indicating gene expression levels and columns representing different samples. Gene expression data features a large number of genes, but experimental samples remain limited, which generating significant difficulties for effective analysis. The elevated dimensionality presents a considerable challenge in processing these datasets. Numerous algorithms drawing from artificial intelligence (AI) techniques, particularly those rooted in machine learning (ML) as an AI branch, have serve as powerful tools for analysing such data [6].

2. Motivation and contribution

Traditional clustering techniques are effective in grouping similar patterns but often assume distinct and well-separated cluster boundaries. Such an assumption is usually unrealistic for gene expression data, which are inherently noisy and significantly overlap across clusters. Fuzzy clustering methods address this limitation by accommodating overlapping cluster memberships and adapting to non-spherical data structures —features that are especially realized in the Gustafson-Kessel (GK) algorithm. Inherently, the GK algorithm takes advantage of cluster-specific covariance matrices, rendering it quite suitable for analysing complex datasets derived from cancer genomics. One of the significant problems in clustering analysis is to find the optimal number of clusters for a given dataset. In addressing this, cluster validity indices like Xie-Beni (XB), Fukuyama-Sugeno (FS), and Dunn Index are adopted to assess the quality of clustering so that the resultant configuration adopted maximizes compactness and separation [2]. To another aspect, dynamic thresholding methods have been demonstrated to be an effective method to identify those genes that show significant changes in the membership scores between cancerous and normal states. Then, by investigating the difference in memberships, one may identify the potential cancer-related genes with a high degree of precision. Using the algorithm (MSA^{GK_CL}) we have tried to make the identification of high-throughput cancer genes with precision and biological relevance. Demonstration of how fuzzy clustering can be highly suitable for noisy high-dimensional data-related problems. Completion contributions to personalized medicine on behalf of suggesting potential therapeutic targets and biomarkers.

3. Literature review

Fuzzy clustering has proved to be a promising microarray data analysis method since it is capable of detecting overlapping and unclear gene behaviour — a property of high value in biological systems. High dimensionality and noise in microarray data make fuzzy approaches preferable to hard clustering approaches that assign genes to one category without regard to biological overlap.

Hazra et al. presented a fuzzy clustering algorithm based on validity indices (FCVI) to differentiate between differential expressed genes related to a few forms of cancer, specifically lung, colon, and leukaemia cancer. Regarding its construction, the model combined the clustering algorithm Fuzzy C-Means (FCM) with cluster validity indices including Xie-Beni, Fukuyama-Sugeno, etc., yielding recognizable clusters as a result of these indices. Comparisons of the membership matrices of the carcinogenic and normal gene expression signatures delineated cancer mediating genes in the FCM clusters. The authors also engaged in validity tests and biological classification tests using known cancer genes in the NCBI database to confirm their validation was statistically significant. The biological validation refuted the rejected clusters producing biological significance and validated that fuzzy clustering offers a clear advantage for bioinformatics with high-dimensional data [3]. Recognizing a need for improvement in the way cluster validations are performed, Ghosh and De created the Gaussian Fuzzy Index (GFI) for assisting the quality of clusters derived from gene expression data. They applied this index in tandem with FCM method on human lung cancer microarray datasets, and reported that GFI can outperform previous indices such as Dunn and Xie-Beni with clusters that are more biologically meaningful than the prior indices. The authors confirmed a number of cancer-related

genes with pathway databases EGFR, TP53, VEGA, and noted the GFI can be one direction into exploring cancer-relevant clusters of gene information [2]. Hazra et al. additionally advanced a Gene Correlation Difference Network Analysis (GCDNA) for examining gene co-expression networks in lung cancer. This approach involved the use of Pearson correlation to calculate difference in gene-gene relationship in normal and cancer tissues. The authors calculated which genes had the largest differences in correlation and generated a co-expression network to interpret their resultant findings. This approach allowed identification of important genes (e.g., EGFR and TP53) as well as the verification of 24 genes that were already known by utilizing the NCBI database, indirectly affirming the reliability of their process in identifying the genes [5]. Continuing their work on association rule mining, Hazra et al. conducted analysis using the Apriori algorithm to identify differentially expressed genes at progressive stages of lung cancer. They analysed the associations that were lost, retained, or formed between normal and the two carcinogenic stages, I and III. The analysis identified several genes of interest such as MIG and APM, and was validated against the NCBI database. But there were limits related to the rigid filtering thresholds and possible non-linear, association-in-expression characteristics of the Apriori model [1].

In an alternative direction, Liu et al. proposed a generalized linear regression model in order to forecast gene expression patterns from the binding properties of the Oct4 transcription factor. They utilized transcriptome and ChIP-seq data from mouse cells to show, for example, that Oct4 binding features like width and distance impacted gene expression across time. While this method was promising, the linear assumptions of the model for predicting complex cellular processes would not necessarily be biological and makes the model less generalizable for applicability beyond mouse cells and/or univariate modelling of a single factor [6]. De Souto et al., performed a comparative assessment of clustering algorithms in which they evaluated the effectiveness of methods including k-means, Gaussian mixtures, and hierarchical clustering on 35 cancer gene - expression datasets. Using the corrected Rand index [8], they found that finite mixture models and k-means outperformed hierarchical methods. On the contrary, they carried out their evaluations without biological validations and did not explore the generalizability to RNA-seq datasets, therefore their findings may not have wide relevance to ubiquitous applications in contemporary genomics [7]. In a multi-cancer context, Xue et al. conducted a complete differential expression analysis for all five significant cancer types, in order to find overarching gene signatures. They subsequently identified hub genes like CDK1 and ASPM through GO and KEGG pathway enrichment, both of which had been linked to poor prognosis and immune infiltration. While the study provided useful insights into a common cancer biology, the absence of experimental validation, as well as a potentially insufficient range of datasets, limited the interpretative power of the study [8]. Borisov et al. took a clinical perspective on gene expression analysis and created a number of machine learning models to predict chemotherapy responses. By searching for information in repositories like TCGA and TARGET, they narrowed the scope of marker genes to find discriminate between responders and non-responders across a variety of cancers. This provided evidence that gene expression-based prediction models can have a clinical impact. However, limitations of data imbalance and number of genome-wide genes in the research challenges its additional use [9].

By and large, the paper implies that fuzzy clustering and membership-based analysis are useful to cancer gene identification but that the majority of current frameworks are static, heavily depend on pre-specified thresholds, or do not handle biological uncertainty. In this paper the algorithm (MSA^{GK_CL}) improves on these previous attempts by using a dynamic thresholding approach, multiple validity measures, and the Gustafson-Kessel algorithm, and thus providing a more variable, data-driven approach to the identification of cancer-relevant genes.

4. Preliminaries

4.1 Gene Expression Data

Gene expression microarray data is the quantification of mRNA levels expressed from genes in a cell or tissue under certain conditions. For this project (MSA^{GK_CL}), the data are obtained from biological samples that reflect both normal and cancerous states, and every gene forms a row and each sample is represented by a column. The values reflect gene activity levels and are continuous and high-dimensional in nature. Since the size of expression values and possible variation in samples, raw data is pre-normalized by Min-Max scaling

to make sure that data values are between 0 and 1. The normalized matrix is then feeded as an input to our clustering framework. This dataset format makes possible the examination of gene behavior within a population of samples so that genes with possibly changed activity under cancerous conditions can be identified in comparison to normal tissue.

4.2 Fuzzy Clustering

Unlike classical (hard) clustering algorithms Fuzzy clustering gives each gene a membership to all clusters, based on its partial involvement. This is very applicable in biological systems where genes frequently belong to multiple regulatory mechanisms. In fuzzy clustering, a membership matrix (membership_mat) of size (n_samples, n_clusters) is formed where every element u_{ij} represents the extent to which gene i belongs to cluster j as in equation (1). The values in every row of this matrix are normalized to add up to one so that cluster assignments are probabilistic.

Each gene x_i (a vector of expression values) has an associated membership score u_{ij} for each cluster j. The membership matrix U= [u_{ij}] satisfies the constraint:

$$\sum_{j=1}^{c} u_{ij} = 1 \text{ for all } i = 1, 2, 3 \ldots N \quad (1)$$

Where in Equation (1):

- N : the total number of genes (samples),
- c : the number of clusters.

This soft partitioning reflects biological reality more accurately and allows further interpretation using fuzzy metrics.

4.3 Gustafson-Kesssel (GK) Algorithm

The Gustafson-Kessel (GK) algorithm is a variation of fuzzy c-means clustering with an adaptive distance metric. Instead of assuming clusters to be spherical (as in Euclidean space), GK makes use of a Mahalanobis distance that considers the shape, orientation, and scale of clusters.

The objective function minimized by the GK algorithm is given by Equation (2). As:

$$J_{GK} = \sum_{i=1}^{N} \sum_{i=1}^{c} u_{ij}^{m} . D_{ij}^{2} \quad (2)$$

In Equation (2):

- m : the fuzzifier (commonly set to 2),
- u_{ij} : the degree of membership of gene x_i to cluster j,
- D_{ij}^2 : the squared distance between x_i and the cluster center v_j.

The Mahalanobis distance is defined in Equation (3) as:

$$D_{ij}^{2} = (x_i - v_j)^T A_j (x^i - v_j) \quad (3)$$

Here, $A_j = S_j^{-1}$, the inverse of the covariance matrix S_j for cluster j. This allows each cluster to adopt its own shape in the multidimensional feature space.

Cluster centers v_j are updated using as shown in Equation (4):

$$v_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \quad (4)$$

Covariance matrices are computed in Equation (5) as:

$$S_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} (x_i - v_j)(x_i - v_j)^T}{\sum_{i=1}^{N} u_{ij}^{m}} \quad (5)$$

The GK algorithm is iteratively applied until the process ends when once modification to membership values or cluster centers are less than a predetermined threshold by achieving convergence.

4.4 Membership Matrix

At the core of fuzzy clustering is the membership matrix U, where each entry $u_{ij} \in [0,1]$ represents how strongly gene x_i belongs to cluster j.

$$\forall_i \in \{1, 2, 3 \ldots N\} \qquad \sum_{j=1}^{c} u_{ij} = 1 \quad (6)$$

Each row of this matrix corresponds to a gene and its distribution across all clusters. A high membership value $u_{ij} \approx 1$ indicates a strong association of gene i to cluster j, while intermediate values indicate partial associations as shown in Equation (6). This allows biological interpretations where genes function in overlapping pathways.

4.5 Validity Indices

In order to choose the most suitable number of clusters c, we calculate clustering validity measures that measure the compactness and separation of clusters. These measures allow us to find out if a given value of c, results in well-shaped, meaningful clusters in the data.

4.5.1 Xie-Beni Index (XB)

The Xie-Beni index assesses the ratio between the average intra-cluster variance and the minimum inter-cluster distance. A good clustering has tightly packed clusters (low intra-cluster variance) and has clusters that are far apart (high inter-cluster separation). XB achieves this by computing the ratio between the overall fuzzified within-cluster distance and the smallest squared distance between cluster centres.
It is calculated in Equation (7) as:

$$XB = \frac{\sum_{i=1}^{N} \sum_{j=1}^{c} u_{ij}^{m} \cdot \|x_i - v_j\|^2}{N.min_{j \neq k} \|v_j - v_k\|^2} \qquad (7)$$

- x_i: i-th gene (as a feature vector)
- v_j: j-th cluster center
- u_{ij}: membership score of gene x_i to cluster j
- m: fuzzifier constant (typically 2)
- $\| x_i - v_j \|^2$: squared Euclidean distance between gene and cluster center
- N: total number of genes
- c: number of clusters

Numerator: Measures total intra-cluster fuzziness; how far genes are from the cluster they (partially) belong to.
Denominator: Measures the squared distance between the two closest cluster centers.
A lower value of XB indicates better clustering — with compact clusters that are well-separated from each other.

4.5.2 Fukuyama-Sugeno (FS) Index

The FS index evaluates both within-cluster compactness and dispersion of clusters relative to the global data mean. It balances the desire for small intra-cluster variance with meaningful inter-cluster distribution. A good clustering will have low intra-cluster distance (compact clusters) and cluster centres spread out away from the global mean as shown in Equation (8):

$$FS = \sum_{i=1}^{N} \sum_{j=1}^{c} u_{ij}^{m} . \|x_i - v_j\|^2 - \sum_{j=1}^{c} \|v_j - \bar{x}\|^2 \qquad (8)$$

Where in Equation (8), $\underline{x} = \frac{1}{N} \sum_{i=1}^{N}$, x_i: the global centroid of all genes
First term: total within-cluster fuzzified distance (like XB numerator)
Second term: total squared distance from each cluster centre to the global mean

Thus, low FS values suggest genes are close to their assigned centres and clusters are far from the global mean (indicating diversity).

4.5.3 Dunn Index

The Dunn index is a ratio between the minimum distance between any two clusters and the maximum diameter of any cluster. The Dunn index evaluates clustering quality based on separation vs. compactness —

in a very different way. It's a ratio of best inter-cluster separation (numerator) and worst intra-cluster spread (denominator). Numerator represents the minimum distance between any two cluster centres — we want this to be large (clusters well-separated). Denominator represents the largest intra-cluster distance among all clusters — we want this to be small (compact clusters). A high Dunn index suggests clusters are well-separated and clusters are internally cohesive.

4.6 Threshold Determination and Its Significance

Once both the normal and carcinogenic datasets have been clustered, we want to determine genes whose cluster associations significantly change. This is done using Membership Score Analysis through the comparison of membership matrices between both states.
For any given gene i, we calculate the maximum absolute difference in membership scores between all clusters for the two conditions as in Equation (9):

$$\Delta u_j = \left| u_{ij}^{(cancer)} - u_{ij}^{(normal)} \right| \quad (9)$$

This value represents how drastically a gene changes its association pattern between normal and diseased conditions. Genes with higher Δu_i are more likely to be biologically significant in disease progression. To filter out only the most significantly altered genes, we compute a dynamic threshold based on the distribution of Δu_i values across all genes.

Only genes for which: Δu_i>Threshold are selected as significant. This dynamic, data-driven method avoids arbitrary cut-offs and ensures that gene selection adapts to the natural variance in the data, improving both biological relevance and statistical robustness.

5. Proposed work

5.1 Methodology

Proposed methodology includes combination of a fuzzy clustering mechanism with dynamic thresholding and validity indices to ascertain cancer-related genes from gene expression datasets. This framework set to take care of the various complicated aspects related to gene expression data, including dimensions, noise, overlap and produce robust interpretation clearly embodying biological meaning. A few steps pertaining to the methodology are as follows:

5.2 Data Collection and Pre-processing

Gene expression datasets with cancerous and normal tissue were downloaded from online databases such as GEO or TCGA. The datasets are pre-processed by imputation of missing values through the use of the Min-Max scalers to normalizing the data such that each feature will be on a common scale. This is essential for bias correction so that the performance of clustering algorithms does not get hampered.

5.3 Clustering Using the Gustafson-Kessel Algorithm

The Gustafson-Kessel algorithm is one of the fuzzy clustering methods that can adjust for non-spherical cluster shapes. Its main steps include initialization of parameters, including fuzziness degree ("m"), maximum iterations, and error tolerance. Estimation of cluster centres and covariance matrices for each cluster. Updating of membership matrix according to adaptive distances so as to achieve optimal gene clustering.

5.4 Evaluation Using Validity Indices

To determine the preferred number of clusters, the following cluster validity indices are calculated: The Xie-

Beni Index measures compactness and separation of clusters. The Fukuyama-Sugeno Index evaluates intra-cluster variance. The Dunn Index assesses separation in comparison with compactness. The clustering setup optimizing these indices is then chosen for further attention.

5.5 Membership Score Analysis

A membership score analysis is conducted to identify significant genes. For each gene, the maximum absolute difference in membership scores between cancerous and normal datasets is calculated. A dynamic threshold is determined using the optimal percentile of values, ensuring adaptability to the data's distribution. Genes with exceeding this threshold are flagged as potential cancer-related biomarkers.

5.6 Validation of Identified Genes

The selected genes are validated against biological databases such as NCBI, Gene Ontology (GO), or KEGG pathways. Statistical metrics, including precision, recall, and F1-score, are used to evaluate the accuracy of the identified genes compared to known biomarkers.

The, results are represented in visual form to facilitate interpretation. Clusters are plotted up to highlight the separation and overlap between cancerous groups of genes versus normal gene groups. Membership differences and thresholds are represented with the help of histograms or scatter plots. The proposed technique (MSA^{GK_CL}) is designed to be highly-scalable for large datasets and very flexible for various cancers. Through these mechanisms, advanced fuzzy clustering and dynamic thresholding in this framework represent a viable and robust solution for biomarkers discovery and cancer genomics advancement.

5.7 Basic Flow Diagram of the Methodology

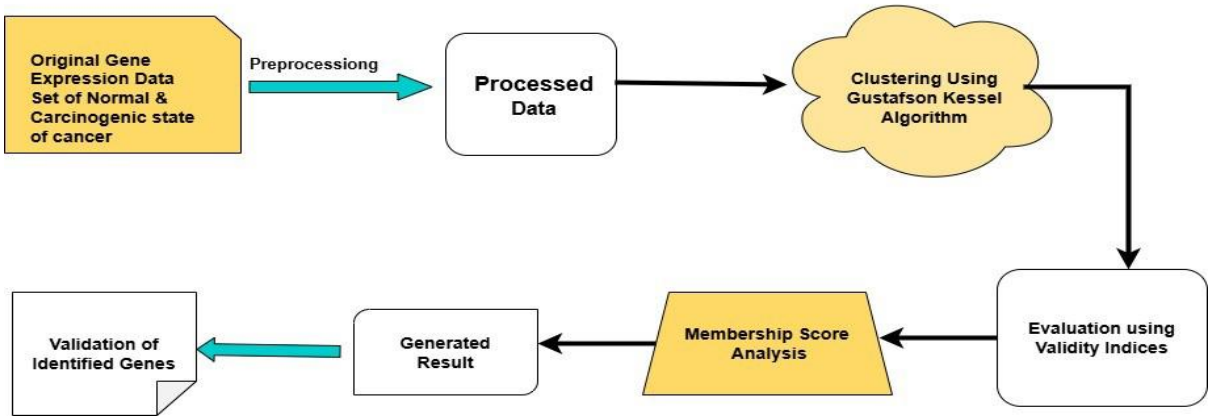


Fig 1. Flowchart of algorithm

5.8 Pseudocode of Algorithm (MSA^{GK_CL})

Step wise approach:

Step 1. Load cancerous and non-cancerous datasets.
Step 2. Pre-process and scale the datasets.
Step 3. Compute validity indices (XB, FS, Dunn) for both datasets.
Step 4. Determine the optimal number of clusters (c*) using validity scores.
Step 5. For each dataset (cancerous, normal):

a. Initialize GK algorithm parameters:
- Number of clusters (c), fuzziness (m), max_iterations.

b. Initialize membership matrix (U) randomly.

c. Repeat until convergence:
i. Compute cluster centres.
ii. Compute covariance matrices.
iii. Calculate adaptive distances.
iv. Update membership matrix.

Step 6. Perform Membership Score Analysis:
For each gene g:
i. Calculate membership differences (Delta U_g).
ii. Compute dynamic threshold.
iii. Select genes with Delta U_g > threshold.

Step 7. Validate selected genes using ground truth:
Compare identified genes with known cancer-related genes.

6. Analysis of Algorithm (MSA^{GK_CL})

The analysis and experimental phase of this project was conducted on two distinct gene expression datasets: one related to lung cancer and another pertaining to leukemia. Both datasets consisted of normalized expression values for thousands of genes across several normal and cancerous samples. The experiments were carried out in a structured sequence comprising pre-processing, clustering, validity evaluation, membership score analysis, and biological validation of the predicted genes.

6.1 Dataset Description

Human lung gene expression data is derived and taken from Affymetrix microarray experiments data. Common type of cancer i.e., lung cancer and Leukaemia Cancer was selected and downloaded. These types of cancers have high prevalence and mortality rate in the developing countries. The current study used a microarray dataset shown in Table 1. as follows: Lung: This dataset consisted of 7131 genes and 96 sample id that were divided into 2 groups of Normal state dataset consisting of 10 sample id and carcinogenic state dataset consisting of 86 sample id. Leukaemia: This dataset consisted of 22397 genes and 56 sample id that were divided into 2 groups of Normal state dataset consisting of 13 sample id and carcinogenic state dataset consisting of 43 sample id.

Table 1. Number of Genes and Samples Used in the Current Study for Lungs and Leukemia Expression Data

Data Set	Number of Genes	Number of Genes in Normal State	Number of Genes in Carcinogenic State
Lungs Expression Data	7129	10	86
Leukemia Expression Data	22397	13	43

6.2 Determination of Optimal Number of Clusters

For each dataset, the normal and cancerous gene expression matrices were subjected to separate applications of the Gustafson-Kessel (GK) fuzzy clustering algorithm. By using adaptive Mahalanobis distance, the algorithm was able to identify irregular and ellipsoidal clusters that more closely matched the biological complexity of the gene data.

For every run, membership matrices were created, giving each gene a fuzzy score across clusters. In the later phases of analysis, membership-based comparisons were based on these scores.

To identify the most meaningful number of clusters for each condition (normal and cancerous), clustering was repeated for k = 2 to 10 and 20 clusters. Three internal cluster validity indices were calculated for each run:

- **Xie-Beni Index (XB)**: minimized when clusters are compact and well-separated.
- **Fukuyama-Sugeno Index (FS)**: minimized when clusters are compact and centered far from the global mean.

Dunn Index: maximized when inter-cluster distance is large and intra-cluster spread is small.The values of these indices were plotted across the range of cluster numbers as in Fig 2, Fig 3, Fig 4 and Fig 5, and the elbow point or minima/maxima were identified as the optimal number of clusters.

For Lungs:

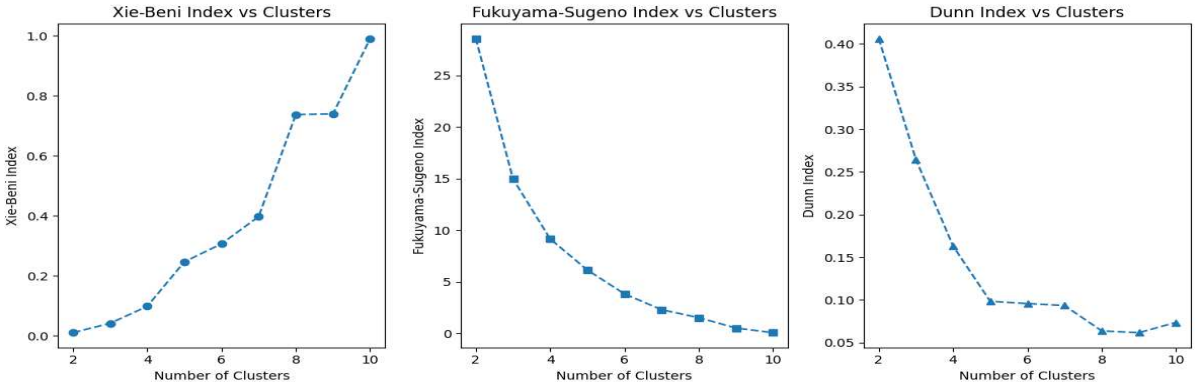


Fig 2. Figure represents the Validity indices for Lung Expression Data-Normal State

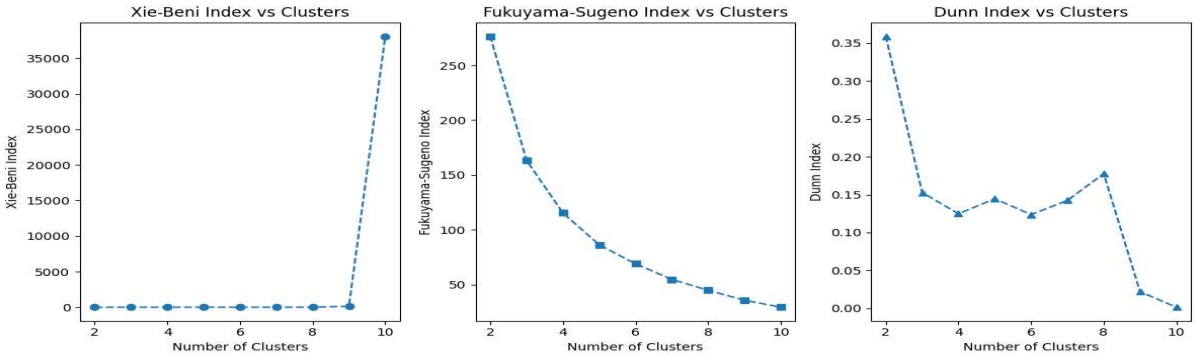


Fig 3. Figure represents the Validity indices for Lung Expression Data-Carcinogenic State

For the lung dataset as determined from Fig 2 and Fig 3:
Normal samples: Optimal clusters determined at k = 3
Carcinogenic samples: Optimal clusters determined at k = 8

For Leukaemia:

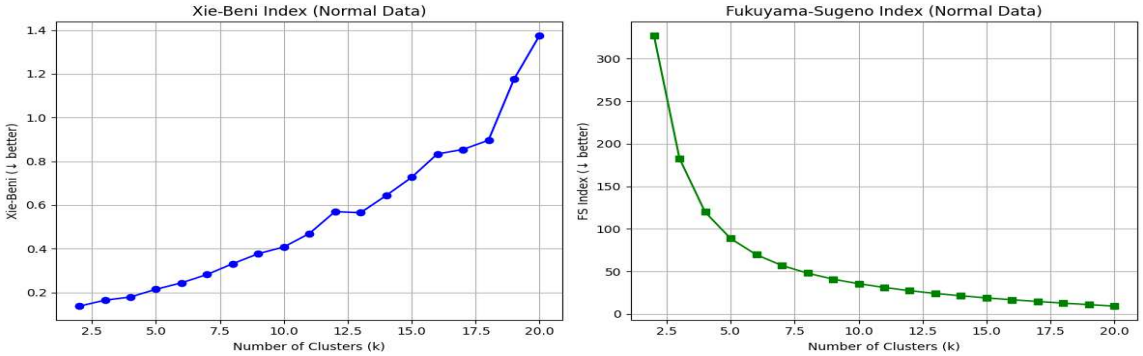


Fig 4. Figure represents the Validity indices for Leukemia Expression Data-Normal State

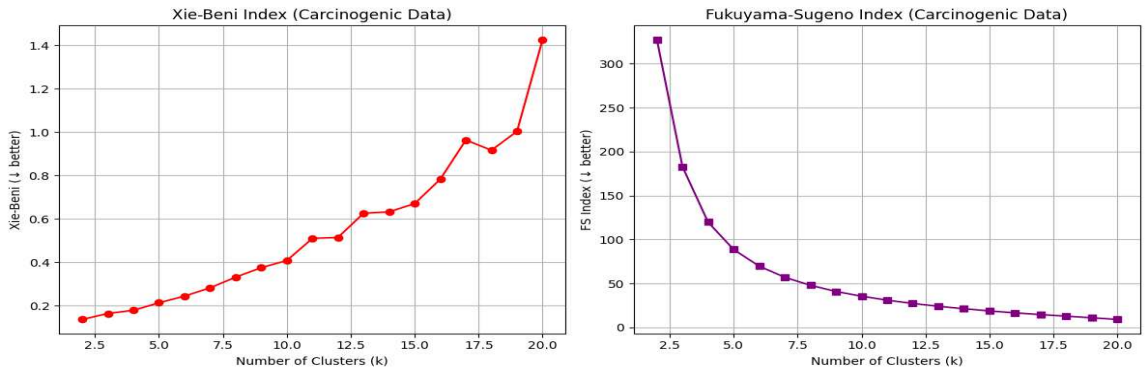


Fig 5. Figure represents the Validity indices for Leukemia Expression Data-Carcinogenic State

For the leukaemia dataset as determined from Fig 4 and Fig 5:

- Normal samples: Optimal clusters at k = 3
- Carcinogenic samples: Optimal clusters at k = 14

The MSA^{GK_CL} algorithm was applied separately to the normal and cancerous gene expression matrices as in Fig 6 and Fig 7 for each Lung and Leukaemia expression dataset.

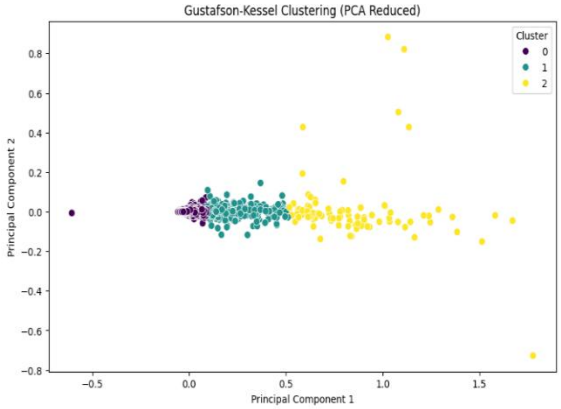


Fig. 6(a). Clustering for Lung Expression Data (Normal State)

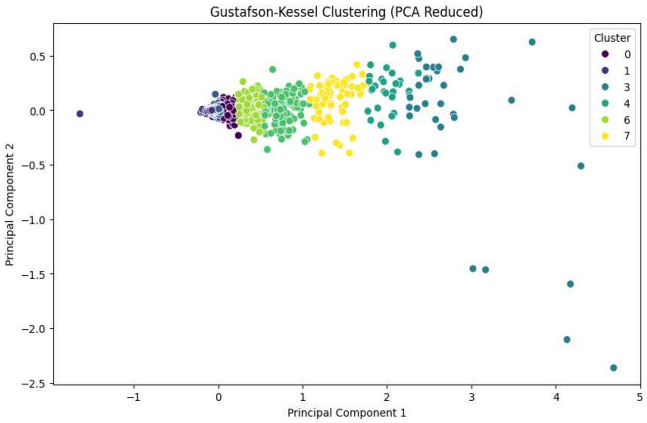


Fig. 6(b). Clustering for Lung Expression Data (Carcinogenic state)

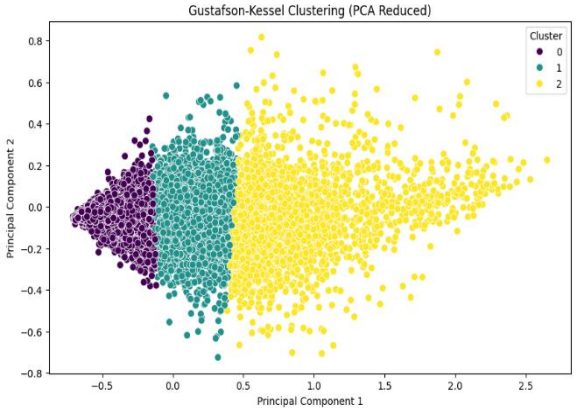


Fig. 7(a). Clustering for Leukaemia Expression Data (Normal State)

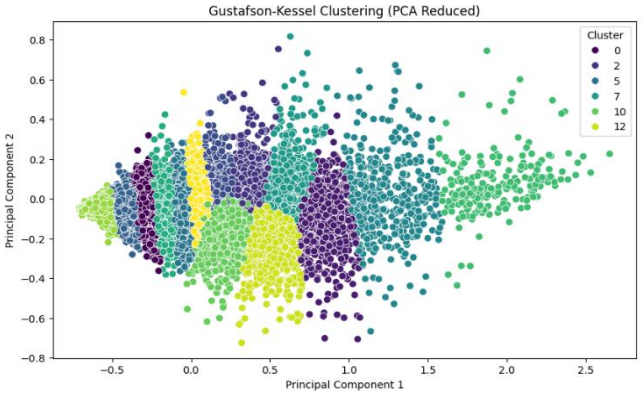


Fig. 7(b) Clustering for Leukaemia Expression Data (Carcinogenic state)

6.3 Membership Score Analysis and Gene Identification

After determining the optimal clusters, the membership matrices from normal and cancerous conditions were compared for each gene. For every gene, the maximum absolute change in membership across clusters was calculated using Equation (9):

$$\Delta u_j = \left| u_{ij}^{(cancer)} - u_{ij}^{(normal)} \right| \quad (9)$$

The dynamic threshold calculation
Let:

- $\Delta u_1, \Delta u_2, \ldots, \Delta u_N$ be the set of absolute membership score differences for each gene.
- $\mu_{\Delta u}$ be the **mean** of all Δu values.
- $\sigma_{\Delta u}$ be the **standard deviation** of the Δu values

Then, the dynamic threshold T is defined in Equation (10) as:

$$T = \mu_{\Delta u} + 2 \cdot \sigma_{\Delta u} \quad (10)$$

Where in Equation (10):

- $\Delta_u = \frac{1}{N} \sum_{i=1}^{N} \Delta u_i$
- $\sigma \Delta_u = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\Delta u_i - \mu \Delta_u)^2}$

This approach sets the threshold as in Fig 8 and Fig 9 at two standard deviations above the mean, which assumes that Δu values roughly follow a normal distribution. Genes with Δu scores greater than this threshold are considered significantly different in behaviour between normal and cancerous conditions.

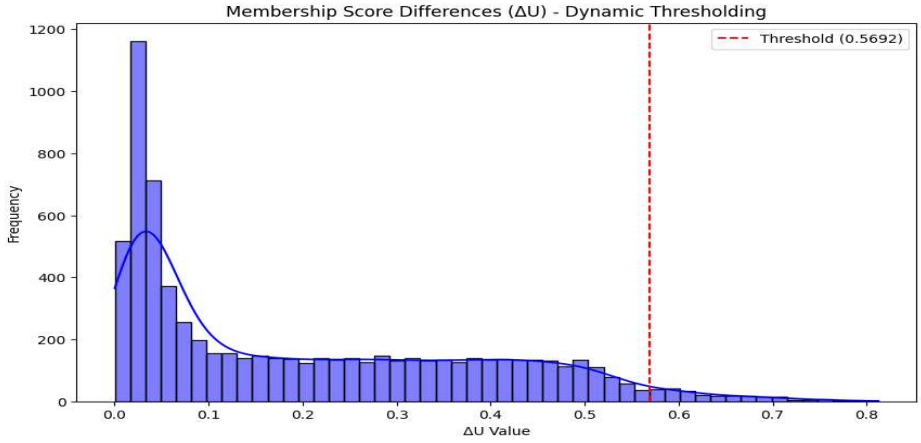


Fig 8. Figure represents the threshold for Lungs Expression Data

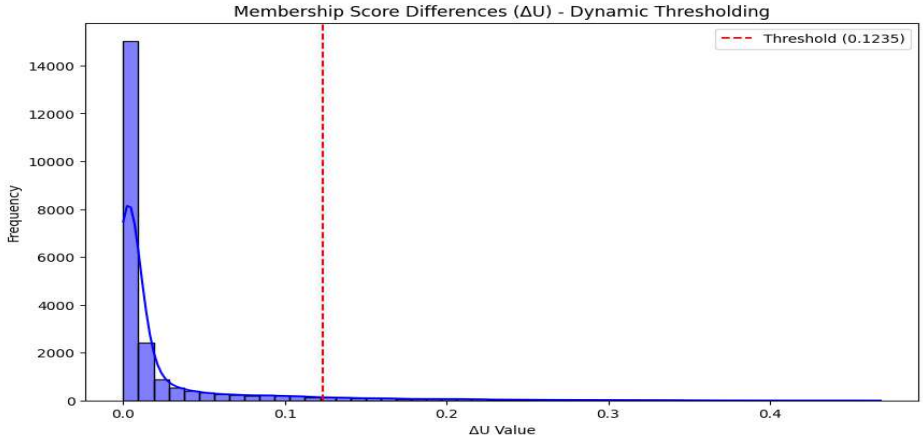


Fig 9. Figure represents the threshold for Leukaemia Expression Data

- Lung dataset: 220 genes identified
- Leukaemia dataset: 1100 genes identified

6.4 Biological Validation, Evaluation and Comparison

In the final component of this project, we completed the biological validation of the predicted genes that had been established through analysis of their membership scores. The purpose of the biological validation was to determine if the genes identified by our fuzzy clustering method were biologically relevant to cancer, thus supporting the biological validity of our method.

This validation occurred after we had determined the best number of clusters through the MSA^{GK_CL} algorithm in conjunction with internal validity indices, fully constructed the membership matrices for both the normal and the carcinogenic samples, and applied a dynamic threshold to calculate the set of differentially behaving genes based on the maximum change in cluster membership scores (Δu) for each gene from the normal to the carcinogenic states. The identified genes that demonstrated significant differences in fuzzy cluster membership were regarded to be possibly associated with the transition from normal behaviour of a cell to a more carcinogenic behaviour.

The predicted gene list used for the biological validation step was then matched against a curated reference list (known cancer mediating genes and their aliases) of cancer associated genes. This reference list comprised of genes from the gene database of National Centre for Biotechnology Information (NCBI) (as standardized gene symbol, common alias name, and some synonyms in cancer). During the matching process, we assessed for the detection of proper matches in the references list.

The biological rationale behind this validation step is grounded in the observation that genes contributing to cancer typically undergo substantial alterations in regulatory behaviour, which would manifest as notable shifts in their fuzzy membership across different conditions. Therefore, if the algorithm identifies such genes purely based on expression-based clustering without supervision, and those genes align with independently documented cancer genes, it offers strong evidence of the method's biological credibility.

The results of the validation process were highly encouraging. In the case of the lung expression dataset, 220 genes were predicted to be significantly altered. Among these, 131 genes matched known cancer-related genes or their aliases as documented in NCBI. Similarly, for the leukaemia expression dataset, 1100 genes were predicted, 200 were considered out of which 113 matched known genes, corresponding to a validation accuracy as in Fig 10.

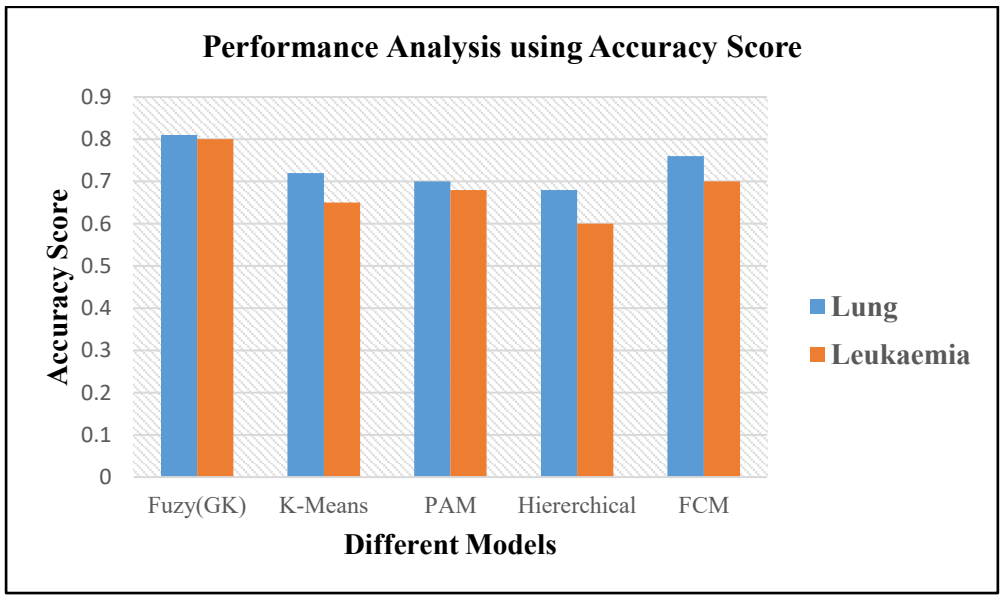


Fig. 10. Comparative study with other models using accuracy score

Comparative analysis between the proposed fuzzy clustering approach using the MSA^{GK_CL} algorithm and traditional hard clustering methods such as k-means as shown in Fig 10. The fuzzy clustering method showed a clear advantage in terms of accuracy in identifying cancer-associated genes, as validated through known biological databases. This performance gap is primarily attributed to the soft assignment property of fuzzy clustering, where each gene is assigned degrees of membership across multiple clusters. This is particularly important in gene expression data, where genes may be involved in multiple pathways or regulatory networks, a complexity that hard clustering fails to capture due to its exclusive one-cluster-per-point assignment. Moreover, fuzzy clustering facilitates membership score analysis (Δu) and dynamic thresholding, enabling more nuanced detection of gene behaviour between cancerous and normal states. In contrast, hard clustering lacks the capability for such granular analysis, which limits its sensitivity in identifying subtle yet biologically significant gene expression shifts.

Following are some of the significant identified genes as in Table 2 among the list of obtained as present in the output:

Table 2. Some significant (TP) Genes identified from Expression Datasets

Lung Expression Data Identified Significant Genes (True Positive)	**Leukaemia Expression Data Identified Significant Genes (True Positive)**
PLAB	TRIP12
AQP3	SMARCD2
LGALS4	MAPK14
AKR1C3	ABCC1
PTGS2	LPHN1
ANXA2	TPMT
TFPI2	RUNX3
TOB1	PIGF
S100A6	SPI1
ID1	IL8

The gene selection was entirely unsupervised, guided only by expression dynamics and membership shifts. No prior biological labels were used during clustering or thresholding. This biologically informed evaluation not only confirms that the clustering and thresholding process is statistically robust but also demonstrates its ability to uncover genes of functional and pathological relevance. Moreover, the unmatched genes may serve as novel biomarker candidates—genes not yet documented in known cancer databases but exhibiting significant behavior in our analysis, warranting further experimental investigation.

7. Conclusion

A systematic and optimised neutral approach to cancer identification and classification plays a vital role in new improve cancer treatment and empower drug discovery. Previous methods for cancer identification depended mostly on clinical data and were limited in their diagnostic capabilities. It is established that gene expressions are recognized as containing the essential guideline for solving primary problems in cancer detection, treatment and discovering new drugs. Recent emergence of microarray technology has helped the creation of substantial gene expression profiles, modifying the design of various clustering techniques for cancer analysis using gene expression microarray data. This project MSA^{GK_CL} has brilliantly showcased the fusion of the Gustafson-Kessel (GK) fuzzy clustering algorithm, cluster validity indices, and dynamic membership score analysis for identifying cancer-related genes from gene expression datasets. It incorporates the adaptive capabilities of the GK algorithm to deal with the complexities arising due to high-dimensional and overlapping biological data. In addition, the validity indices such as Xie-Beni (XB), Fukuyama-Sugeno (FS), Dunn Index help to obtain an optimal clustering as a guarantee towards the trustworthiness for the results. The practice of dynamic thresholding in the membership score analysis allows for the proper identification of significance genes that show alteration in their expression profiles between normal and cancerous tissues. Furthermore, the validation of through biological databases serves to further elucidate the accuracy and relevance of the biomarkers recognized.

The research findings are of great importance in accessing cancer biology through potential therapeutic targets

and would set forth a snowballing of scalable and adaptable framework for analysing complex datasets in different biological domains. In summary, this work amplifies the realm of utilization of effective fuzzy clustering methods by bioinformatics as providing a robust solution for biomarker discovery thereby calling for furthering advancements in personalized medicine and cancer diagnostics.

Funding
This research did not receive external funding.
Data Availability
The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.
Author Declarations Conflict of Interest
The authors have no conflicts to disclose.

References

[1] Subir Hazra, Rohan Sarkar, Amartya Roy, Anupam Ghosh (2022) - Identifying Differentially Expressed Genes in Different Stages of Lung Cancer – an application of ARM Model on Gene Expression Data - IEEE.

[2] Anupam Ghosh and Rajat K. De (2013) - Gaussian Fuzzy Index (GFI) for Cluster Validation: Identification of High Quality Biologically Enriched Clusters of Genes and Selection of Some Possible Genes Mediating Lung Cancer –Springer-Verlag Berlin Heidelberg.

[3] Samir Kumar Sett, Subir Hazra, Anupam Ghosh (2019) - A fuzzy clustering algorithm influenced by validity indices (FCVI) for recognizing the differentially expressed cancer mediating genes –Meta Gene – Elsevier.

[4] E. Gomede (2021) - The Gustafson-Kessel Algorithm: Enhancing Fuzzy Clustering with Adaptive Metrics – AI Mind.

[5] Subir Hazra, Amit Kumar Shaw, Palash Das, Anupam Ghosh (2022) - Gene Co expression analysis for identifying some regulatory genes in human lung cancer –IEEE.

[6] Shuai Liu, Mengye Lu, Hanshuang Li and Yongchun Zuo (2019) - Prediction of Gene Expression Patterns with Generalized Linear Regression Model

[7] Marcilio CP de Souto, Ivan G Costa, Daniel SA de Araujo, TeresaBLudermir and Alexander Schliep (2008) - Clustering cancer gene expression data: a comparative study-
BioMed Central Ltd

[8] Jin-min Xue, Yi Liu, Ling-hong Wan, Yu-xi Zhu (2020) - Comprehensive Analysis of Differential Gene Expression to Identify Common Gene Signatures in Multiple Cancers-
Med Sci Monit.

[9] Nicolas Borisov, Maxim Sorokin, Victor Tkachev, Andrew Garazha, A. Buzdin (2020) - Cancer gene expression profiles associated with clinical outcomes to chemotherapy treatments - From 11th International Young Scientists School “Systems Biology and Bioinformatics” – SBB- Novosibirsk, Russia.

[10] Özer Özdemir and Asli Kaya (2018) - GUSTAFSON-KESSEL AND FUZZY C-MEANS ALGORITHMS BY COLON CANCER DATA IN FUZZY CLUSTERING - Technological Applied Sciences(NWSATAS).

[11] Hsiang-Chuan Liu , Bai-Cheng Jeng , Jeng-Ming Yih , and Yen-Kuei Yu (2009)- Fuzzy C-Means

Algorithm Based on Standard Mahalanobis Distances - International Symposium on Information Processing (ISIP'09).

[12] P.K. Nizar Banu, H. Hannah Inbarani (2013) - AN ANALYSIS OF GENE EXPRESSION DATA USING PENALIZED FUZZY C-MEANS APPROACH

[13] H. Robert Frost(2021)- Analyzing cancer gene expression data through the lens of normal tissue-specificity-PLOSComputational Biology.

[14] Meskat Jahan, Mahmudul Hasan (2020)- A Novel Fuzzy Clustering Approach for Gene Classification-(IJACSA) International Journal of Advanced Computer Science and Applications.

[15] E. Kavitha1, R. Tamilarasan, Arunadevi Baladhandapani and M. K. Jayanthi Kannan(2021)- A Novel Soft Clustering Approach for Gene Expression Data- Tech Science Press.