Hooghly Engineering & Technology College

Department of Computer Science & Engineering
B. Tech. Final Year Project

A Project on:
A Machine Learning technique in Gene Expression dataset to find influencing genes of Alzheimer’s disease(AD)

Under guidance of :
Mr. Arup Mallick

Presented By -
Rishav Das- 17600121016
Pamela Pal- 17600121023
Srija Mazumder- 17600121040
Souvik Maity- 17600121045

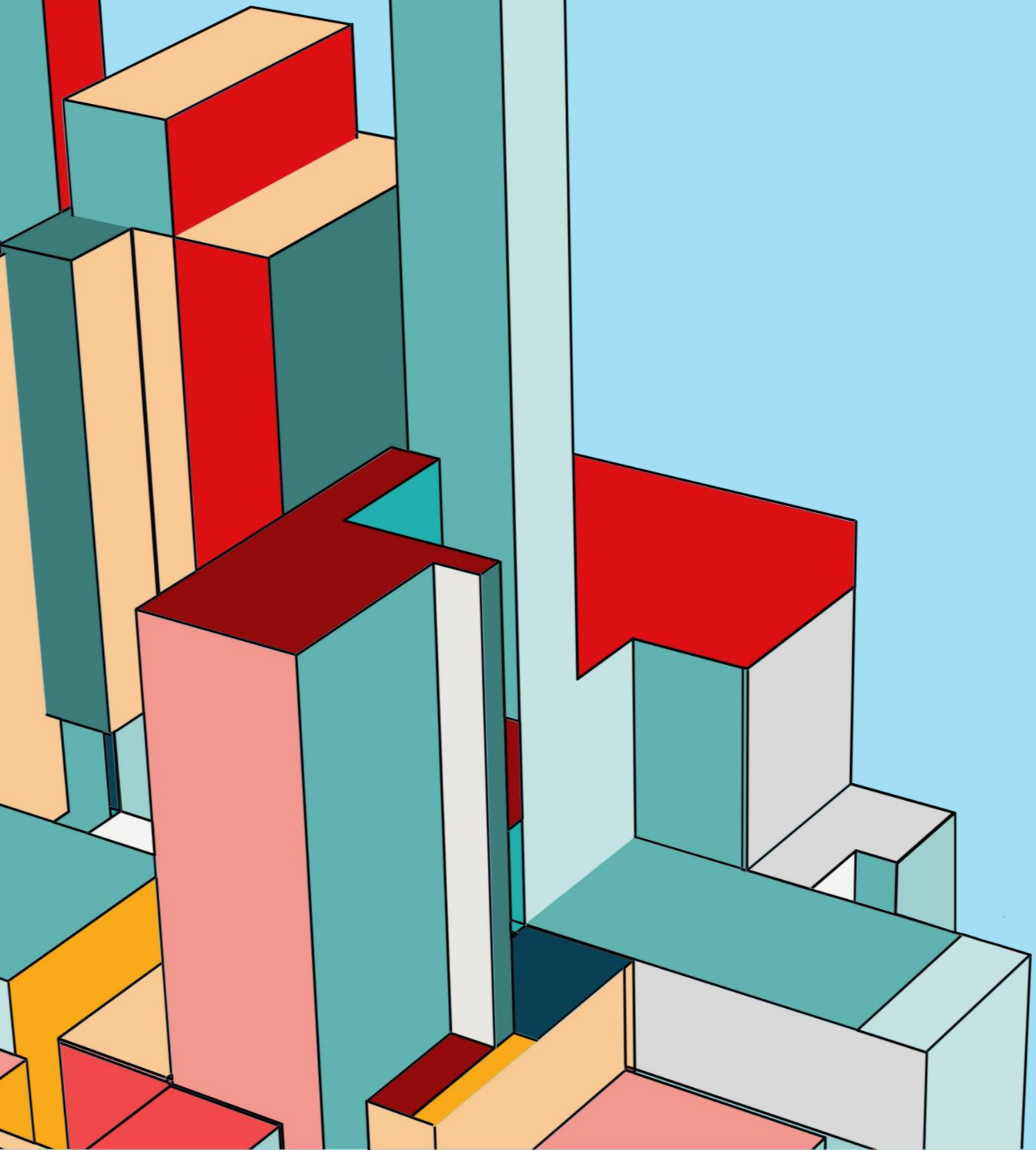


TABLE OF CONTENTS

1 INTRODUCTION

3 PROBLEM STATEMENT

4 BRIEF THEORETICAL BACKGROUND

5 METHODS INVOLVED

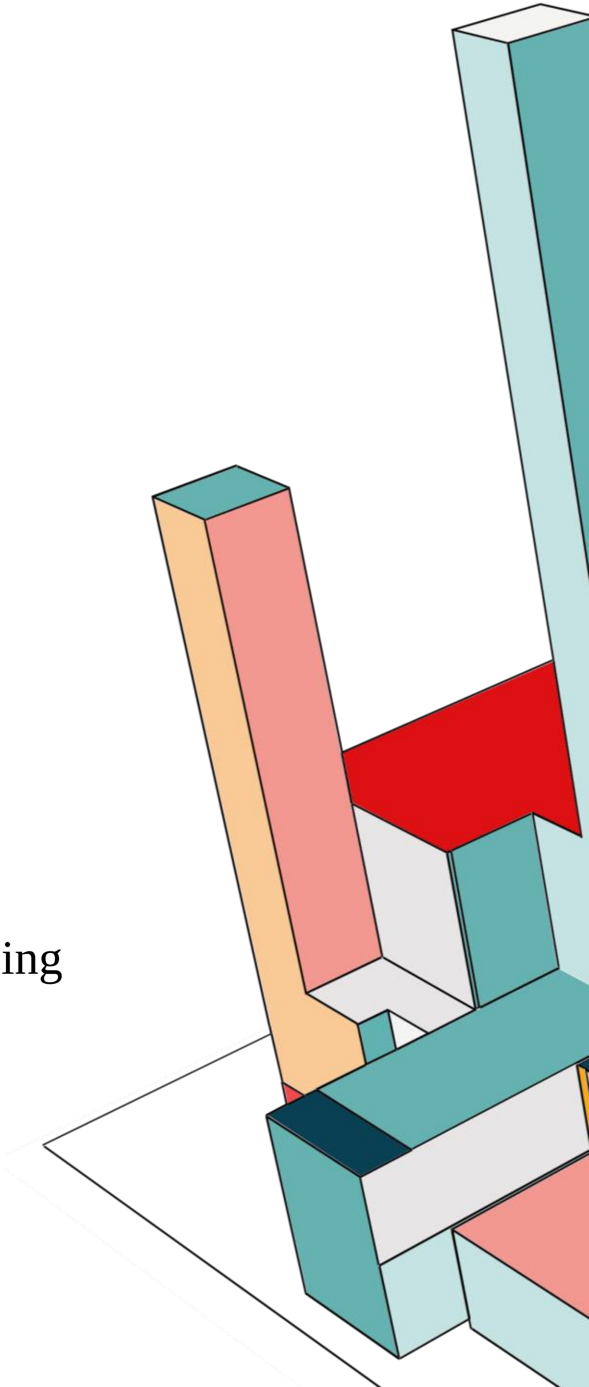
6 RESULTS & DISCUSSION

7 CONCLUSION

8 REFERNCE

INTRODUCTION:-

- Global health issue affecting millions, with a rise expected from 47 million (2015) to significantly more by 2050
- Symptoms progress from memory loss to complete cognitive decline, loss of bodily functions, and death.
- Early diagnosis is crucial for slowing progression.
- MRI and neuropsychological tests detect AD at advanced stages.
- Limited in early detection and non-invasive diagnostics.
- Gene expression data provides molecular-level detection.
- Allows earlier and more precise AD diagnosis, compared to traditional methods.
- Machine learning models (SVM, CNN, Random Forest) outperform clinicians in diagnosing AD.
- ML techniques are effective with MRI and gene expression data.
- Provides a non-invasive, scalable, and more efficient diagnostic solution.



PROBLEM STATEMENT:-

- Alzheimer’s Disease (AD) is a neurodegenerative disorder characterized by memory loss and cognitive decline.
- Despite extensive research, the precise genetic factors contributing to the onset and progression of Alzheimer’s remain poorly understood.
- Early diagnosis is challenging due to the complex nature of genetic factors associated with Alzheimer’s.
- The project aims to utilize machine learning techniques on gene expression datasets (GSE48350, GSE11882 from GEO) to identify genetic markers associated with Alzheimer’s.
- Gene expression datasets GSE48350 and GSE11882 from the GEO database are used.
- Techniques like feature selection, dimensionality reduction, and classification algorithms will be applied to the high-dimensional gene expression data.
- The goal is to pinpoint significant genes differentiating healthy individuals from those affected by Alzheimer’s.
- Results could provide insights into the genetic mechanisms behind Alzheimer’s.
- Findings may contribute to the development of precision medicine approaches for Alzheimer’s diagnosis and treatment.
- Link to the datasets used –
  i. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE48350
  ii. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11882

BRIEF BACKGROUND:-

1. Normalization (Min-Max Scaling):
   - Rescales gene expression data between 0 and 1.
   - Ensures all features contribute uniformly to machine learning models.
   - Prevents features with larger ranges from skewing results.
2. Principal Component Analysis (PCA):
   - Reduces dataset dimensionality while retaining key information.
   - Projects data onto principal components, capturing the most variance.
   - Improves computational efficiency and reduces overfitting risks.
3. K-Means Clustering:
   - Unsupervised algorithm that groups genes based on similarities in expression patterns.
   - Iteratively adjusts centroids to minimize variance within clusters.
   - Essential for identifying meaningful gene patterns related to AD

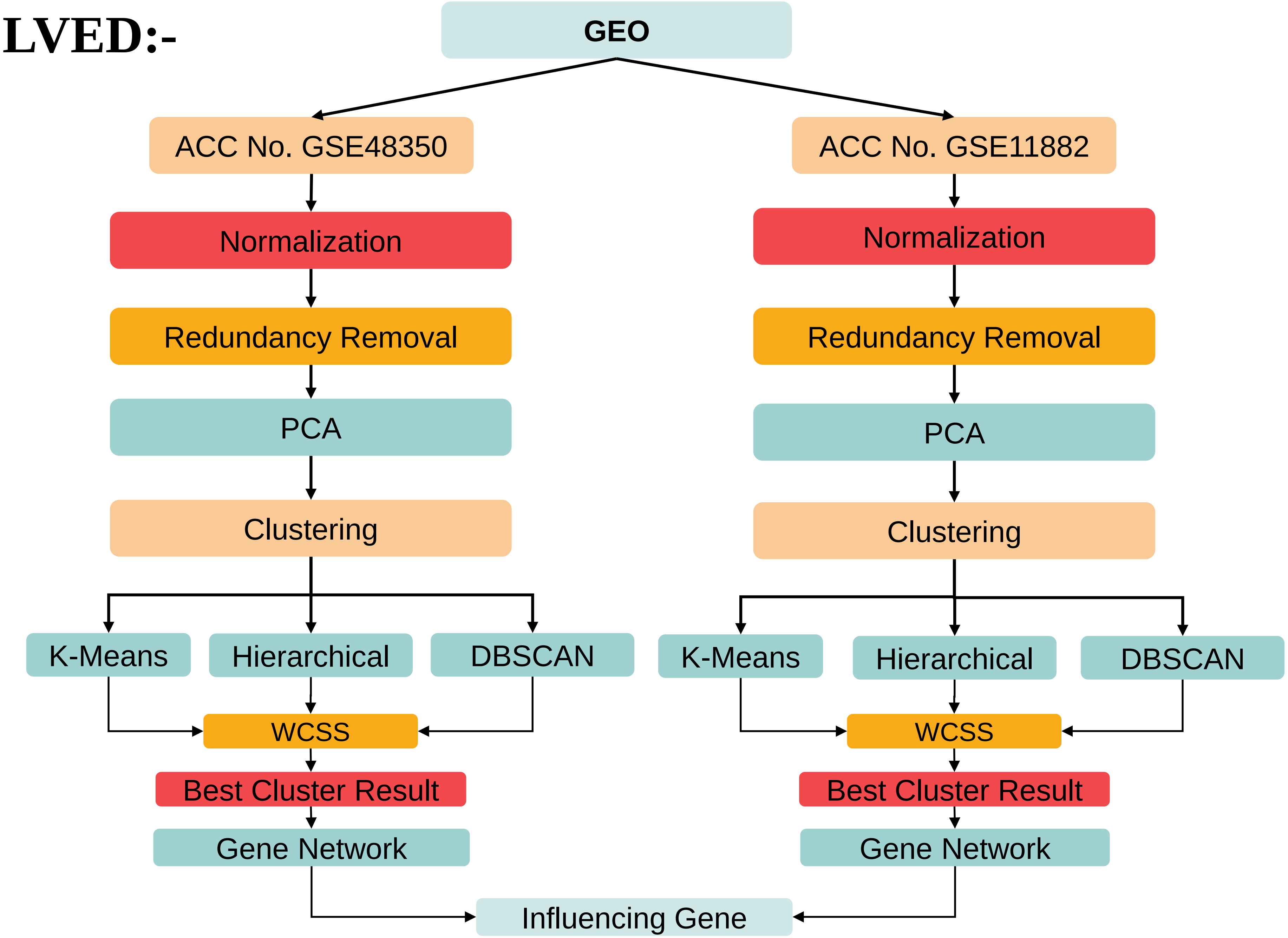
MATERIALS:-

1. We have used gene expression datasets (GSE48350, GSE11882) from the Gene Expression Omnibus (GEO) database.
2. This data set contains profiles of genes potentially associated with Alzheimer’s Disease.
3. This dataset contains microarray data from normal controls (aged 20-99 yrs) and Alzheimer's disease cases, from 4 brain regions: hippocampus, entorhinal cortex, superior frontal cortex, and post-central gyrus.
4. URLs for the two datasets

   i) GSE48350 – https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE48350

   ii) GSE11882 - http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11882

METHODS INVOLVED:-



GEO

ACC No. GSE48350

Normalization

Redundancy Removal

PCA

Clustering

K-Means

Hierarchical

DBSCAN

WCSS

Best Cluster Result

Gene Network

ACC No. GSE11882

Normalization

Redundancy Removal

PCA

Clustering

K-Means

Hierarchical

DBSCAN

WCSS

Best Cluster Result

Gene Network

Influencing Gene

1. Dataset Selection:

- Utilized gene expression datasets (GSE48350, GSE11882) from the Gene Expression Omnibus (GEO) database.
- Contains profiles of genes potentially associated with Alzheimer’s Disease.

2. Data Preprocessing:

- Normalization: Scales gene expression data between 0 and 1 for uniform analysis.
- Redundancy Removal: Removes duplicates and irrelevant features to reduce complexity.

3. Principal Component Analysis (PCA):

- Reduces dataset dimensionality, focusing on features with the highest variance.
- Simplifies complex gene expression data while preserving essential patterns.
- Helps avoid overfitting in high-dimensional datasets.

Continued….

4. Clustering Techniques:

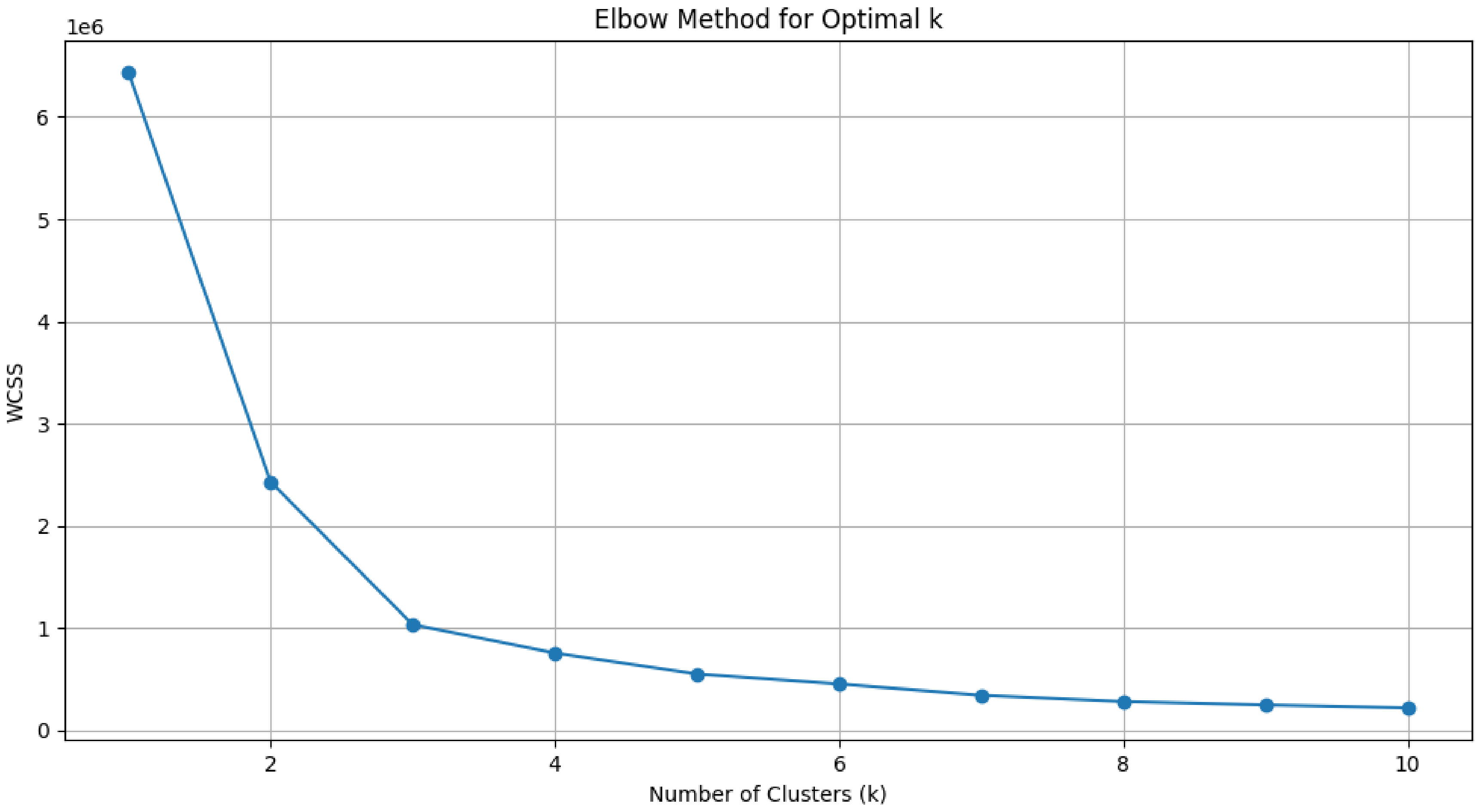
- K-Means Clustering: Groups genes into predefined clusters based on similarities in expression.
- Hierarchical Clustering: Organizes genes in a tree-like structure, showing relationships between them.
- Combination of both approaches offers robust and interpretable results.

5. Visualization:

- Visualizes gene clusters to provide a clear understanding of patterns.
- Intuitive representation of gene groupings aids in biological analysis.

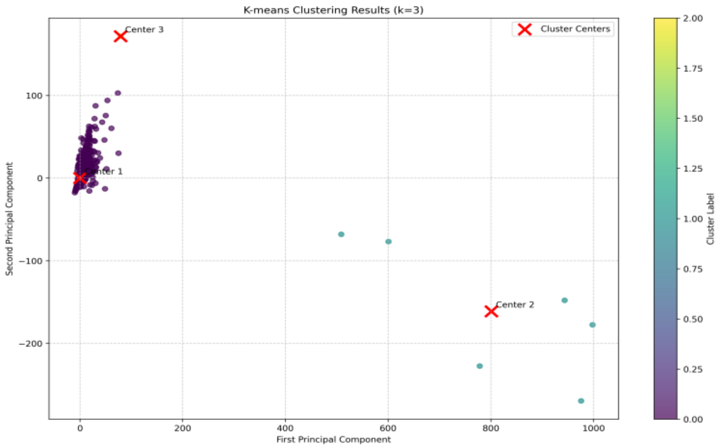
RESULTS & DISCUSSION:

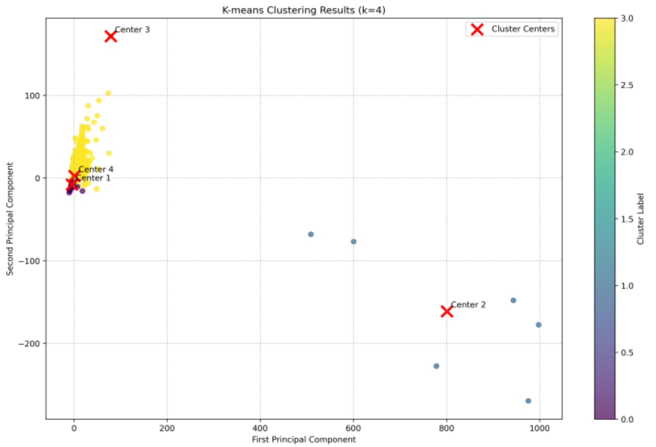
- **Elbow method for optimal no of clusters (GSE48350):**

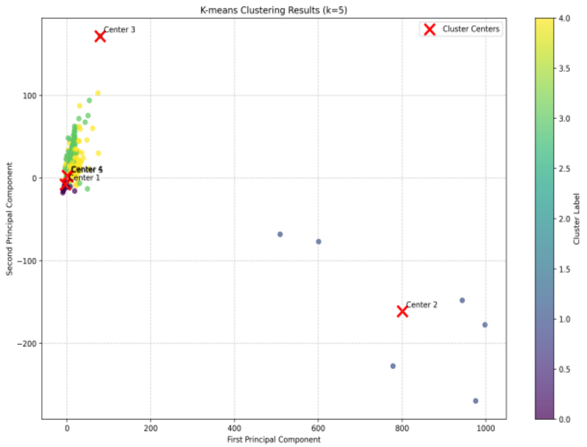


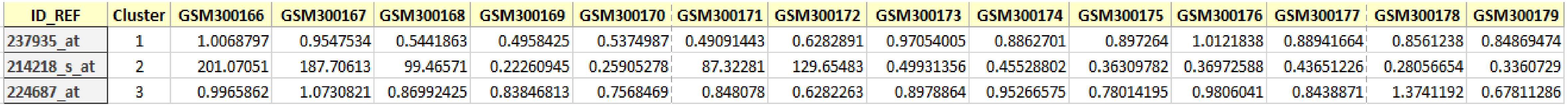
Continued …

Results of K-Means Clustering of dataset Acc no. - GSE48350

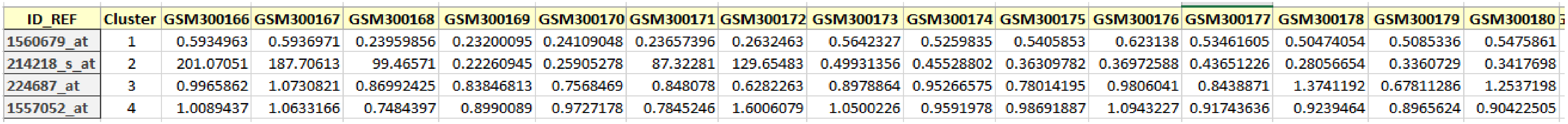




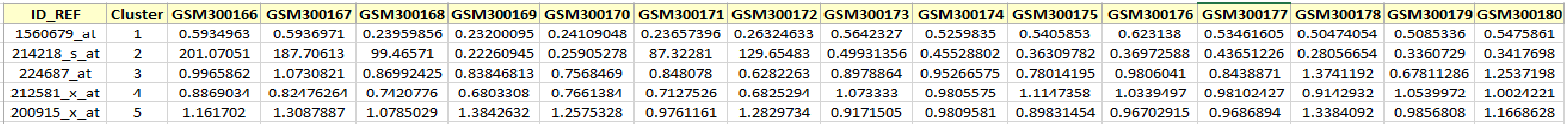




ID_REF	Cluster	GSM300166	GSM300167	GSM300168	GSM300169	GSM300170	GSM300171	GSM300172	GSM300173	GSM300174	GSM300175	GSM300176	GSM300177	GSM300178	GSM300179
237935_at	1	1.0068797	0.9547534	0.5441863	0.4958425	0.5374987	0.49091443	0.6282891	0.97054005	0.8862701	0.897264	1.0121838	0.88941664	0.8561238	0.84869474
214218_s_at	2	201.07051	187.70613	99.46571	0.22260945	0.25905278	87.32281	129.65483	0.49931356	0.45528802	0.36309782	0.36972588	0.43651226	0.28056654	0.3360729
224687_at	3	0.9965862	1.0730821	0.86992425	0.83846813	0.7568469	0.848078	0.6282263	0.8978864	0.95266575	0.78014195	0.9806041	0.8438871	1.3741192	0.67811286



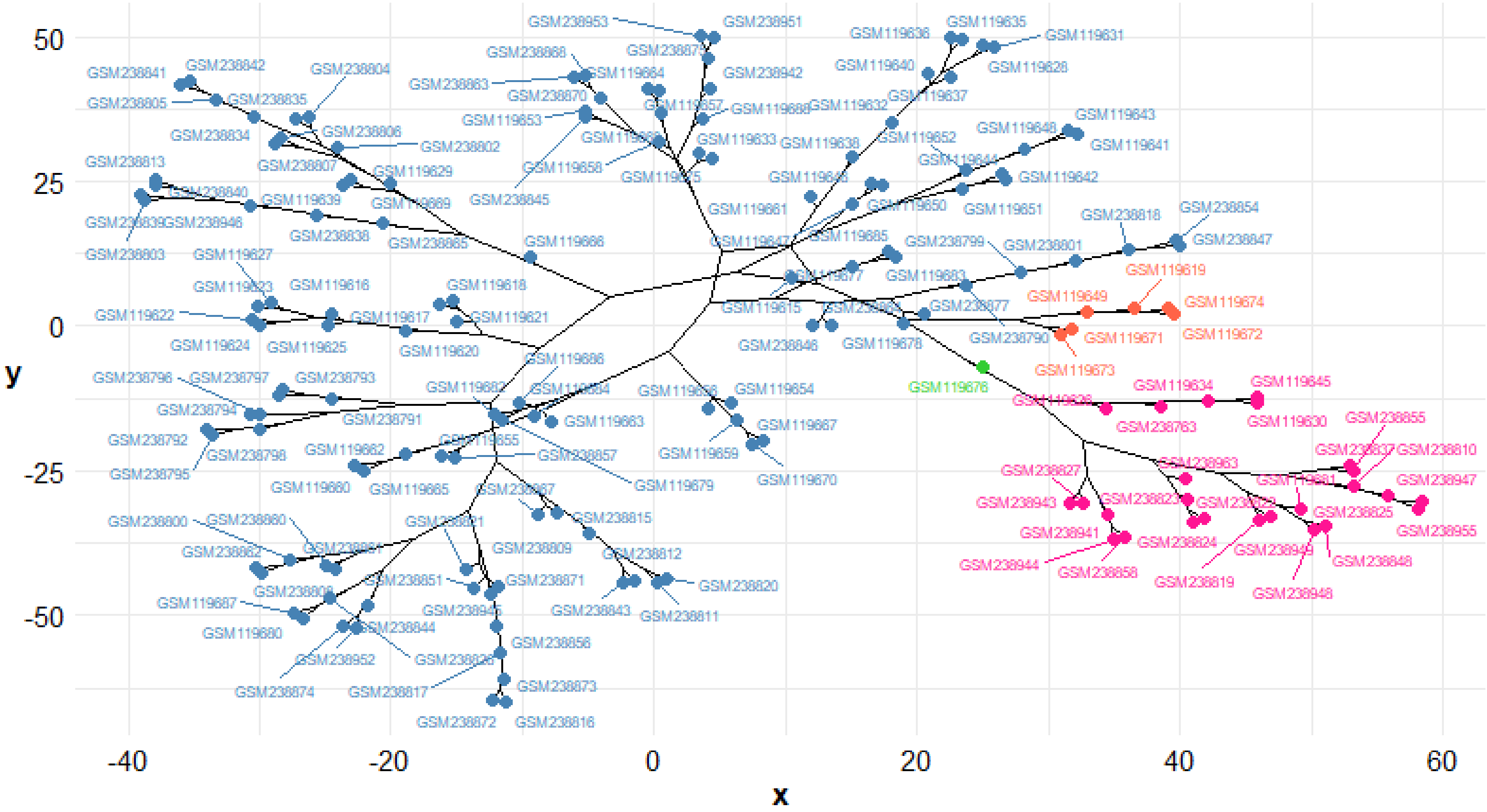
ID_REF	Cluster	GSM300166	GSM300167	GSM300168	GSM300169	GSM300170	GSM300171	GSM300172	GSM300173	GSM300174	GSM300175	GSM300176	GSM300177	GSM300178	GSM300179	GSM300180
1560679_at	1	0.5934963	0.5936971	0.23959856	0.23200095	0.24109048	0.23657396	0.2632463	0.5642327	0.5259835	0.5405853	0.623138	0.53461605	0.50474054	0.5085336	0.5475861
214218_s_at	2	201.07051	187.70613	99.46571	0.22260945	0.25905278	87.32281	129.65483	0.49931356	0.45528802	0.36309782	0.36972588	0.43651226	0.28056654	0.3360729	0.3417698
224687_at	3	0.9965862	1.0730821	0.86992425	0.83846813	0.7568469	0.848078	0.6282263	0.8978864	0.95266575	0.78014195	0.9806041	0.8438871	1.3741192	0.67811286	1.2537198
1557052_at	4	1.0089437	1.0633166	0.7484397	0.8990089	0.9727178	0.7845246	1.6006079	1.0500226	0.9591978	0.98691887	1.0943227	0.91743636	0.9239464	0.8965624	0.90422505



ID_REF	Cluster	GSM300166	GSM300167	GSM300168	GSM300169	GSM300170	GSM300171	GSM300172	GSM300173	GSM300174	GSM300175	GSM300176	GSM300177	GSM300178	GSM300179	GSM300180
1560679_at	1	0.5934963	0.5936971	0.23959856	0.23200095	0.24109048	0.23657396	0.26324633	0.5642327	0.5259835	0.5405853	0.623138	0.53461605	0.50474054	0.5085336	0.5475861
214218_s_at	2	201.07051	187.70613	99.46571	0.22260945	0.25905278	87.32281	129.65483	0.49931356	0.45528802	0.36309782	0.36972588	0.43651226	0.28056654	0.3360729	0.3417698
224687_at	3	0.9965862	1.0730821	0.86992425	0.83846813	0.7568469	0.848078	0.6282263	0.8978864	0.95266575	0.78014195	0.9806041	0.8438871	1.3741192	0.67811286	1.2537198
212581_x_at	4	0.8869034	0.82476264	0.7420776	0.6803308	0.7661384	0.7127526	0.6825294	1.073333	0.9805575	1.1147358	1.0339497	0.98102427	0.9142932	1.0539972	1.0024221
200915_x_at	5	1.161702	1.3087887	1.0785029	1.3842632	1.2575328	0.9761161	1.2829734	0.9171505	0.9809581	0.89831454	0.96702915	0.9686894	1.3384092	0.9856808	1.1668628

Continued …

SAMPLE GROUPS USING HIERARCHICAL CLUSTERING (GSE48350):



CONCLUSION:-

- Machine Learning Techniques: Normalization, PCA, and K-Means clustering were applied to analyze gene expression data for Alzheimer’s Disease.
- Normalization (Min-Max Scaling): Ensured all features contributed equally, preventing bias from varying scales.
- PCA: Reduced dimensionality, retaining key information while simplifying the dataset and minimizing overfitting.
- K-Means Clustering: Identified inherent patterns and grouped similar genes, aiding in the discovery of potential biomarkers for early AD detection.
- Outcome: These techniques provide a more efficient and insightful analysis, advancing diagnostic and research efforts in Alzheimer's Disease.

REFERENCE:-

1. Tanveer, M., Richhariya, B., Khan, R., Rashid, A., Khanna, P., Prasad, M., & Lin, C. (2020). Machine learning techniques for the diagnosis of Alzheimer’s disease: A review. *Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16, 1-35. https://doi.org/10.1145/3401634

2. Bringas, S., Salomón, S., Duque, R., Lage, C., & Montaña, J. L. (2020). Alzheimer’s disease stage identification using deep learning models. *Journal of Biomedical Informatics*, 109, 103514. https://doi.org/10.1016/j.jbi.2020.103514

3. Wang, S. H., Phillips, P., Sui, Y., Liu, B., Yang, M., & Cheng, H. (2018). Classification of Alzheimer’s disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling. *Journal of Medical Systems*, 42(5), 85. https://doi.org/10.1007/s10916-018-0926-3

4. Chen, H., He, Y., Ji, J., & Shi, Y. (2019). A machine learning method for identifying critical interactions between gene pairs in Alzheimer’s disease prediction. *Frontiers in Neurology*, 10, 1162. https://doi.org/10.3389/fneur.2019.01162

5. Li, W., Zhao, Y., Chen, X., Xiao, Y., & Qin, Y. (2018). Detecting Alzheimer’s disease on small dataset: A knowledge transfer perspective. *IEEE Journal of Biomedical and Health Informatics*, 23(4), 1234-1242. https://doi.org/10.1109/JBHI.2018.2869384

6. Bryan, R. N. (2016). Machine learning applied to Alzheimer’s disease. *Radiology*, 281(3), 665-668. https://doi.org/10.1148/radiol.2016161491

7. Neelaveni, J., & Devasana, M. G. (2020). Alzheimer disease prediction using machine learning algorithms. In *Proceedings of the 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, 6–7 March 2020.

8. Alam, S., & Kwon, G. R. (2017). Alzheimer disease classification using KPCA, LDA, and multi-kernel learning SVM. *International Journal of Imaging Systems and Technology*, 27(2), 133–143. https://doi.org/10.1002/ima.22243

THANK YOU