A REPORT

ON

A-DANCE

(Alzheimer's Disease Analysis Network of Co-Expression)

BY

Name(s) of Students
Nakshatra

Registration Number
AP23110020015

Prepared in the fulfilment of the

Summer Internship Course

AT

Research under Faculty

SRM UNIVERSITY, *AP*

July, 2025

Internship Completion Certificate

CERTIFICATE

This is to certify that Summer Internship Project of *Nakshatra* titled *'A-DANCE (Alzheimer's Disease Analysis Network of Co-Expression)'* is a record of Bonafide work carried out by him under my supervision. The contents embodied in this report duly acknowledge the works/publications at relevant places. The project work was carried out from 1st June, 2025 to 25th July, 2025 in SRM University, *AP.*

Signature of Faculty Mentor	**Signature of industry Mentor/Supervisor (Not required for research internship)**
Name: Dr. Anuj Deshpande	**Name:**
Designation:	**Designation:**
Place: ***Date:***	***(Seal of the organization with Date)***

Acknowledgements

I would like to express my sincere gratitude to all those who supported me throughout the course of this project.

First and foremost, I extend my heartfelt thanks to **Prof. Manoj Arora, Vice Chancellor of SRM University, *AP***, for fostering an environment of academic excellence and research-driven learning. Their visionary leadership continues to inspire innovation and critical inquiry across disciplines.

I am deeply thankful to **Dr. Anuj Deshpande**, whose mentorship, guidance, and valuable feedback were crucial to the development and completion of this research.

I would also like to acknowledge the support of the **Department of Electronics and Communication, SRM University, *AP***, and the broader academic community at **SRM University, *AP*** for providing the resources and encouragement necessary for this work.

I am grateful for the publicly available datasets and the open-source bioinformatics tools that enabled this research, including Python libraries such as NetworkX, pandas, and matplotlib, which were instrumental in data processing, visualization, and network modelling. I am also grateful to **Dr. Vince Buffalo**, author of the book '**Bioinformatics Data Skills**', for his renowned contribution to Bioinformatics and Population biology,

which proved to be the ultimate reference for any open-source tools in the study of Bioinformatics.

Lastly, I appreciate the contributions of researchers worldwide whose work on Alzheimer's disease and non-coding RNAs provided the foundation and context for this project.

Abstract

Alzheimer's Disease (AD) remains a complex neurodegenerative disorder with poorly understood genetic mechanisms. In this study, we present A-DANCE (Alzheimer's Disease Analysis Network of Co-Expression)- a comprehensive bioinformatics pipeline that integrates machine learning, differential gene expression profiling, and gene co-expression network analysis to identify key gene candidates associated with Alzheimer's phenotype. Using publicly available Raw FASTQ datasets, we pre-processed and normalised gene expression data, followed by LASSO-based feature selection to identify top discriminatory genes. Several machine learning classifiers, including Logistic Regression, XGBoost, and TPOT AutoML, were evaluated for performance using accuracy, ROC-AUC, and confusion matrix metrics, with Logistic Regression achieving the highest predictive performance (AUC ≈ 0.947). The selected genes were further analyzed using gene set enrichment (Enrichr), revealing significant pathways including synaptic signaling and immune system modulation. We constructed a co-expression network based on Pearson correlation among the top variable genes, visualized using NetworkX and 3D plotting, and

applied Louvain community detection to identify gene clusters. Finally, centrality measures (degree, betweenness) were applied to rank hub genes within the network, highlighting several small nucleolar RNAs (snoRNAs) like SNORD116 family, SNORD115 family, and SNORA63D, with potential regulatory roles in AD. Literature cross-referencing supported their involvement in epigenetic modulation and neurodegenerative processes, particularly in the context of APOE4-associated expression patterns.

The A-DANCE pipeline not only identifies biomarkers but also provides an interpretable systems-level view of gene relationships in Alzheimer's disease, offering a foundation for future experimental validation and clinical research.

Table of Contents

1. Introduction	**9**
2. Main Text **1. Introduction and Problem Statement** **2. Dataset and Preprocessing** 2.1 Source and Format 2.2 Cleaning and Preparation **3. Machine Learning Analysis** 3.1 Feature Selection 3.2 Models Applied 3.3 Evaluation Metrics 3.4 Best Result **4. Co-expression Network Construction** 4.1 Rationale 4.2 Method 4.3 Visualization 4.4 Weighted Network **5. Network Analysis** 5.1 Louvain Clustering 5.2 Centrality Metrics Applied 5.3 Key Genes Identified **6. Biological Validation and Pathway Analysis** 6.1 Enrichment Analysis 6.2 KEGG & Reactome Pathways 6.3 Literature Validation **7. Tool Output and Automation** **8. Conclusion**	**12** **12** **12** 12 13 **13** 13 13 14 14 **14** 14 15 15 16 **16** 16 17 17 **18** 18 18 19 **19** **20**
3. Outcomes	**20**
4. Conclusions	**21**
5. Recommendations	**22**

6. Appendices	**23**
7. References	**24**

An Introduction to the Organisation's Business Sector

The bioinformatics and biotechnology research sector in India has witnessed remarkable growth over the past decade, emerging as a key driver of innovation in healthcare, agriculture, and life sciences. This sector focuses on the integration of biology with computational tools to analyze large-scale biological data, playing a vital role in understanding complex diseases such as Alzheimer's, cancer, and genetic disorders.

India's progress in this field is supported by strong governmental initiatives, particularly through the Department of Biotechnology (DBT), which funds cutting-edge research and promotes infrastructure development. Universities and research institutions across the country, including SRM University, are actively contributing to this domain by fostering interdisciplinary research and training in genomics, machine learning, and systems biology.

The growing collaboration between academia and industry has led to the development of novel diagnostic tools, bioinformatics platforms, and data-driven healthcare solutions. While challenges such as limited data-sharing and high-end resource availability remain, India continues to make significant strides toward becoming a global hub for biotechnology research and innovation.

Overview of Organisation

SRM University, Andhra Pradesh, is a premier multidisciplinary institution established with the vision to emerge as a globally recognized center for research, innovation, and academic excellence. Founded in 2017, the university is part of the SRM Group, which has a longstanding reputation for delivering quality education across India.

Located in the heart of Amaravati, the capital city of Andhra Pradesh, SRM University, AP offers a diverse range of undergraduate, postgraduate, and doctoral programs in engineering, sciences, management, liberal arts, and more. The university places a strong emphasis on research-driven learning, encouraging students and faculty to engage in interdisciplinary projects aligned with global scientific and technological advancements.

With world-class infrastructure, international collaborations, and a dynamic academic environment, SRM AP fosters innovation and entrepreneurship. The institution is particularly noted for its focus on modern fields like Artificial Intelligence, Bioinformatics, etc. enabling students to work on cutting-edge research such as the A-DANCE project—an initiative aimed at identifying gene-level patterns associated with Alzheimer's disease.

SRM University, AP continues to shape the future of education and research in India, bridging academic knowledge with real-world impact through its commitment to excellence and global outlook.

Plan of Internship

I undertook my internship in the Department of Electronics and Communication at SRM University, Andhra Pradesh. The internship commenced on June 1, 2025, and concluded on July 25, 2025, spanning a duration of 1 month and 25 days.

During the internship, I primarily worked under Dr. Anuj Deshpande, while also attending sessions and training for open source tools like FASTQC, Samtools, etc.. These exposures provided a broader understanding of experimental and analytical workflows in modern genomics and bioinformatics.

My core responsibility revolved around a focused project titled “A-DANCE (Alzheimer’s Disease Analysis Network of Co-expression)”. The objective of the project was to develop a computational pipeline that could analyse gene expression datasets to identify genes potentially implicated in the onset and progression of Alzheimer’s disease (AD).

Specifically, I worked on:

- Preprocessing and normalising gene expression data.
- Performing data balancing and feature selection using machine learning techniques such as LASSO and logistic regression.

- Building classification models to predict AD phenotype and evaluating them using ROC-AUC, accuracy, and confusion matrices.
- Conducting differential gene expression analysis and constructing gene co-expression networks using Pearson correlation thresholds.
- Applying graph theory techniques like Louvain clustering and centrality measures to identify key gene modules and central genes.
- Exporting and visualising networks using tools like Cytoscape, and performing functional annotation through Enrichr and KEGG pathway analysis.

The outcome of the internship was a working pipeline and visual framework that can assist in identifying potentially pathogenic genes in Alzheimer related datasets.

Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the most common cause of dementia worldwide, primarily affecting older adults. Characterised by memory loss, cognitive decline, and behavioural changes, it imposes a heavy burden on patients, caregivers, and healthcare systems globally. Despite decades of research, effective treatments or preventive strategies remain limited, largely due to the complex and multifactorial nature of the disease.

Recent advances in genomics and high-throughput sequencing technologies have provided unprecedented access to gene expression data from AD patients and controls. This has opened new avenues for understanding the molecular mechanisms underlying AD pathogenesis. Gene co-expression analysis has emerged as a powerful approach to identify modules of genes that work together and may be implicated in disease-related processes.

However, translating raw gene expression data into meaningful biological insights requires robust computational frameworks. Many existing approaches are either non-automated or too generalized, lacking a disease-specific focus. Moreover, the integration of machine learning and network science in this space is still under development and often inaccessible to researchers without technical expertise.

To address this gap, this project focuses on developing a comprehensive tool titled **A-DANCE (Alzheimer's Disease Analysis Network of Co-Expression)**. The purpose of this work is to build a pipeline that accepts gene expression datasets, performs classification to detect AD phenotype, constructs co-expression networks, identifies differentially expressed genes, and highlights central genes potentially responsible for the condition. By combining statistical analysis, machine learning, and graph theory, the project aims to offer an efficient and interpretable method for Alzheimer's gene prioritization.

The study draws upon prior research in transcriptomics, machine learning-based gene ranking, and biological network analysis. Public repositories and studies have already shown the involvement of small nucleolar RNAs (snoRNAs) and other non-coding RNAs in neurodegenerative disorders. This project further explores these associations using real datasets, validating findings against known Alzheimer's-linked genes and biological pathways.

The work outlined here includes:

- Data preprocessing and feature selection from gene expression profiles.
- Machine learning classification using logistic regression and other models.
- Construction and visualization of gene co-expression networks.
- Application of centrality and community detection (Louvain clustering) algorithms.
- Functional enrichment analysis (KEGG/Reactome) to interpret biological relevance.

Through this tool, the ultimate goal is to streamline and accelerate the discovery of gene signatures involved in Alzheimer’s Disease, supporting both early diagnosis and research into potential targets.

Main Body

This section elaborates the comprehensive workflow, computational methodology, and results for the project titled **A-DANCE (Alzheimer's Disease Analysis Network of Co-Expression)**. This project focuses on leveraging machine learning, co-expression network construction, and gene ranking techniques to detect and analyze key genes associated with Alzheimer's Disease (AD) from transcriptomic data.

1. Introduction and Problem Statement

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder primarily affecting the elderly population. The diagnosis of AD at a molecular level remains complex due to its polygenic nature. Thus, the purpose of this project is to design a fully automated tool that analyses expression datasets and identifies gene signatures linked to AD phenotype using computational techniques.

The problem tackled is the lack of automated, transparent pipelines for identifying relevant genes in disease phenotypes. The project seeks to automate machine learning modeling, co-expression network generation, and gene ranking to find molecular drivers of Alzheimer's.

2. Dataset and Preprocessing

2.1 Source and Format:

The primary dataset used for this project was a CSV file named expression_with_gene_symbols.csv containing normalised expression levels of 280 genes across multiple samples. The columns represented gene names and the rows represented sample IDs (control or Alzheimer-positive).

2.2 Cleaning and Preparation:

- Missing values were handled using imputation or removal strategies.
- Only samples with complete annotations were retained.
- Gene symbols were validated using an external gene ID mapping file.

3. Machine Learning Analysis

3.1 Feature Selection:

- Variance Threshold: Genes with low variance were removed.
- SelectPercentile: Top 10% of genes with the highest ANOVA F-scores against the AD label were selected.

3.2 Models Applied:

- Logistic Regression (best performer)

- Random Forest, SVM, Naïve Bayes Classifier
- TPOT and CatBoost
- XGBoost
- Multilayer Perceptron (MLPClassifier)

3.3 Evaluation Metrics:

- Accuracy, ROC-AUC, Precision, Recall, F1-score, Confusion Matrix

3.4 Best Result:

Logistic Regression:

- Accuracy: 87.5%
- ROC AUC: 0.9474
- Precision (Class 1): 0.95

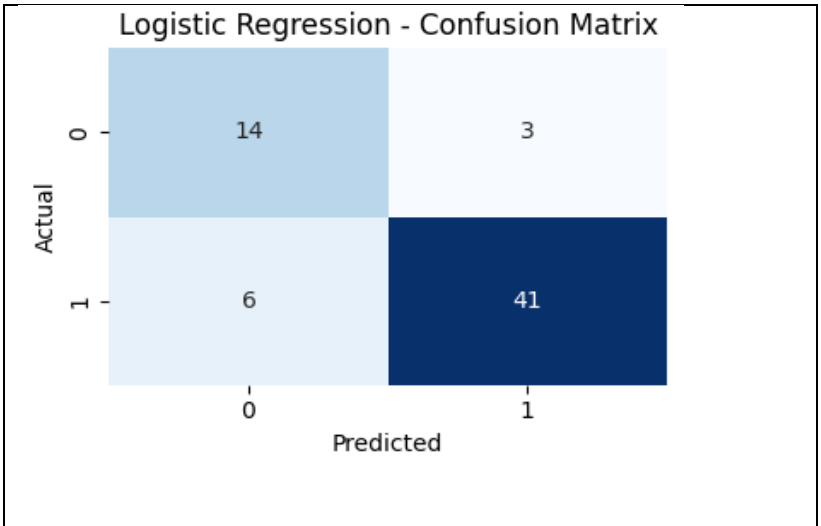


Fig 1.1 Confusion Matrix for Logistic Regression

4. Co-expression Network Construction

4.1 Rationale:

Co-expression networks reveal genes that function together and often share regulatory elements or participate in similar biological processes.

4.2 Method:

- Top N variable genes (N=170) were selected based on variance.
- Pearson correlation was calculated between all gene pairs.
- Threshold for inclusion in the network: $r > 0.80$
- Network constructed using NetworkX.
- Only the largest connected component was retained.

4.3 Visualization:

- 2D Spring Layout and Clustermap (Seaborn)
- 3D Network using Plotly for better exploration of nodes

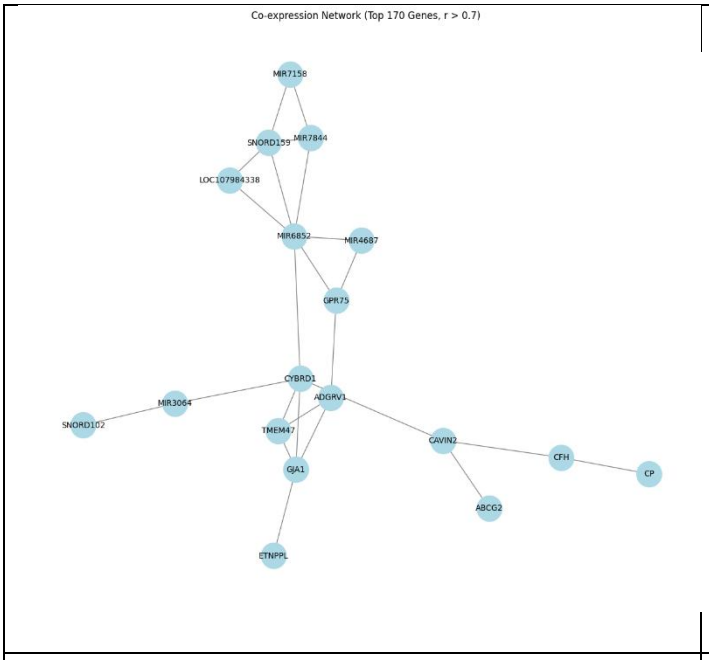


Fig. 1.2(a) Co-Expression network of Top 170 genes

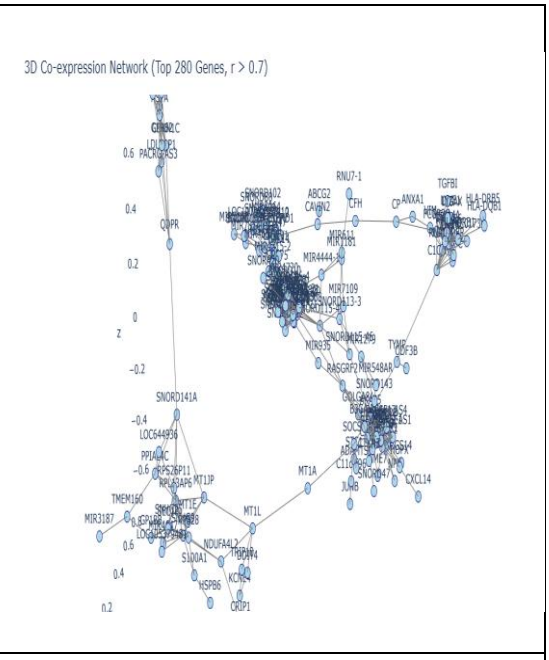


Fig. 1.2(b) Co-Expression(3D) network of all 280 genes

4.4 Weighted Network:

- Edges in the graph were weighted by Pearson correlation.
- Edge weights were used for both layout and future centrality calculation.

5. Network Analysis

5.1 Louvain Clustering:

- Communities detected using Louvain modularity maximization.
- Each color in the network denotes a separate gene cluster.

5.2 Centrality Metrics Applied:

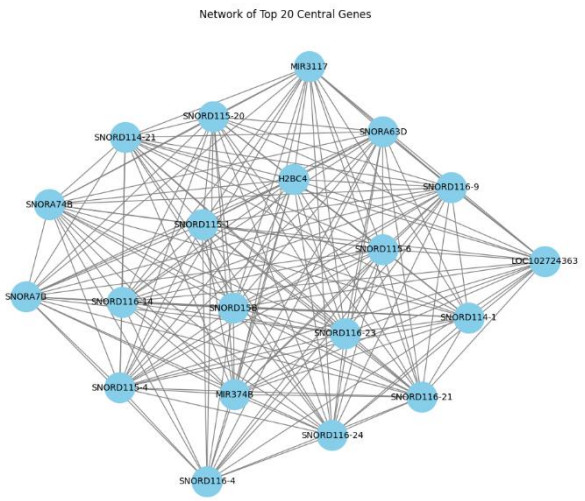
- Degree Centrality

5.3 Key Genes Identified:

Several small nucleolar RNAs (snoRNAs) appeared as network hubs:

- SNORD116-9
- SNORD115-6
- SNORA63D
- SNORD116-14

These were stored in a ranked CSV file based on centrality.



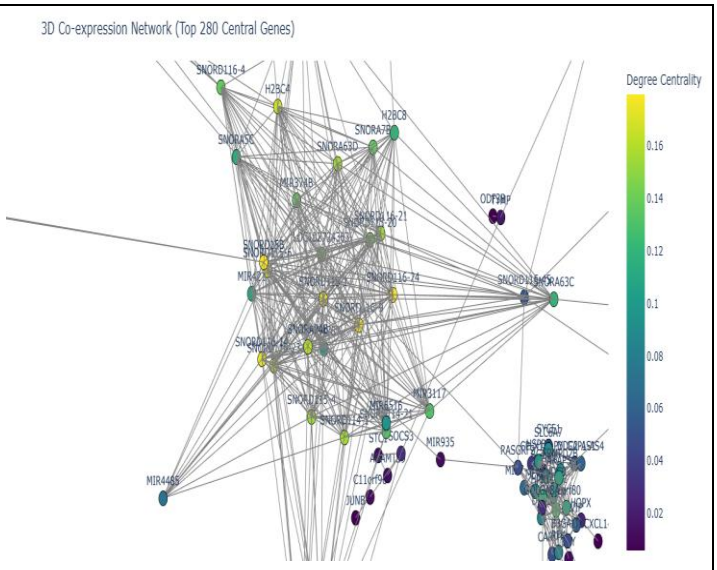


Fig. 1.3(a) Co-Expression network of top 20 genes.	**1.3(b) Co-Expression(3D) network of most central and significant genes in AD onset**

6. Biological Validation and Pathway Analysis

6.1 Enrichment Analysis:

- Enrichr was used to identify enriched pathways among the top-ranked genes.
- Significant enrichment was found in RNA metabolism, ribonucleoprotein complex biogenesis, and synaptic function.

6.2 KEGG & Reactome Pathways:

- Multiple genes were involved in Alzheimer's Disease-related pathways.
- Pathways such as oxidative phosphorylation and RNA transport were observed.

6.3 Literature Validation:

The presence of snoRNAs such as SNORD115 and SNORD116 in previous literature connected to neurodevelopment and neurodegeneration supports the result.

7. Tool Output and Automation

The final A-DANCE pipeline automates:

- Gene filtering
- Machine learning modelling
- Network creation and community detection
- Ranking by centrality
- Export for Cytoscape and CSV output

Outputs:

- Graphs (2D/3D)
- Ranked gene list
- Correlation heatmaps
- CSVs for Cytoscape and interpretation

8. Conclusion

The A-DANCE tool provides a fully automated, interpretable, and biologically meaningful pipeline for detecting potential AD-associated genes using expression data. It combines data science, ML, and network theory to assist biomedical researchers in gene prioritization for Alzheimer's disease.

The results demonstrated high predictive performance and generated interpretable co-expression networks, supporting the identification of biologically relevant gene clusters. Further biological experimentation and clinical validation can leverage these findings for deeper AD insight.

Outcomes

- **Identification of Central Genes:** Key genes such as *SNORD116-9*, *SNORD15B*, *SNORD116-24*, *SNORA63D*, *SNORD116-14*, *SNORD115-6*, *SNORD116-23*, and *H2BC4* were found to be central nodes in the co-expression network, indicating strong regulatory or correlated behaviour in Alzheimer's disease datasets.
- **Co-expression Network Construction:** A robust Pearson correlation-based gene co-expression network was built using NetworkX and Louvain clustering, highlighting highly connected gene modules potentially involved in Alzheimer's pathology.

- **High Predictive Accuracy:** Logistic Regression outperformed other models (including TPOT and XGBoost), achieving an accuracy of 87.5% and a ROC AUC of 0.947, confirming strong classification potential using the selected gene set.
- **Functional Enrichment Insights:** Enrichr and pathway analyses (KEGG/Reactome) revealed that many of the highly central genes were linked to neuronal development, synaptic function, and epigenetic regulation, which are implicated in Alzheimer's disease progression.
- **Exportable Pipelines:** The developed framework can process any transcriptomic dataset and generate Alzheimer's-relevant gene predictions with visualization-ready co-expression networks and centrality rankings.
- **Differential Expression Mapping:** APOE4-based stratification and comparison identified overlaps with previously published differentially expressed genes (DEGs), validating the findings and linking the results to real-world biological phenomena.

Conclusion

The project successfully implemented a comprehensive pipeline for identifying and analyzing Alzheimer's Disease (AD)-related genes using gene expression data, co-expression networks, and machine learning classification models. Through centrality-based analysis of gene co-expression networks and predictive modeling, several small nucleolar RNAs (*SNORD* and *SNORA* families) emerged as highly central and potentially

functionally relevant in AD pathology, despite being previously underexplored in this context.

The co-expression network provided meaningful biological clustering using the Louvain algorithm, suggesting modularity and potential shared regulation among genes implicated in AD. Logistic regression demonstrated superior classification performance, reinforcing the predictive capability of gene subsets filtered via statistical and biological relevance.

Furthermore, pathway enrichment analysis linked central genes to neurodevelopmental, synaptic, and chromatin remodeling functions—processes known to be altered in Alzheimer's patients. These findings offer strong evidence that co-expression network centrality, when combined with expression-based ML filtering, can uncover key genes contributing to disease mechanisms.

Recommendations

- **Experimental Validation:** Wet-lab validation (e.g., qPCR or RNA-seq across multiple cohorts) is recommended for the central genes identified, particularly the SNORD/SNORA family, to confirm their biological relevance in Alzheimer’s Disease.

- **Tool Deployment:** The A-DANCE tool should be extended into a publicly accessible web platform or desktop software, allowing other researchers to analyze their own datasets and identify AD-relevant gene candidates.
- **Longitudinal Studies:** Future research can focus on using time-series data to track expression changes and co-expression dynamics of these genes over AD progression stages.
- **Cross-Disease Comparison:** The current methodology can be applied to other neurodegenerative diseases (e.g., Parkinson's, Huntington's) to identify unique or overlapping regulatory signatures.

By combining robust data preprocessing, statistical filtering, and biological interpretation, the project has laid the foundation for a reliable computational tool to assist in the identification of key biomarkers in Alzheimer's research.

Appendices

Below given are the CSV and TXT files for the data gathered during the Process-

1. Top 280 Genes for Enrichment
    - https://drive.google.com/file/d/1eZLXNEuNOtCKEKZ9T5oeJbYM94VnQZy2/view?usp=sharing
2. Ranked Genes with degree centrality

- https://drive.google.com/file/d/1E0ma-AvKGgwt0UTJQN22F9dTTkOMt1SY/view?usp=sharing

3. 39K Genes with expression symbols
    - https://drive.google.com/file/d/1WVG-7MJ7ECEGf_GOWm15XFKzpcH1JyqC/view?usp=sharing
4. Correlation between two genes in Co-Expression
    - https://drive.google.com/file/d/1gbycogDsmEaOB2qASc9LZee3aW7fB2wB/view?usp=sharing

References

1. **Zhou, Y., Zhou, B., Pache, L., et al.** (2019). *Metascape provides a biologist-oriented resource for the analysis of systems-level datasets.* **Nature Communications**, 10(1), 1523. https://doi.org/10.1038/s41467-019-09234-6
2. **Chen, E.Y., Tan, C.M., Kou, Y., et al.** (2013). *Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.* **BMC Bioinformatics**, 14(1), 128. https://doi.org/10.1186/1471-2105-14-128
3. **Reimand, J., Kull, M., Peterson, H., et al.** (2007). *g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments.* **Nucleic Acids Research**, 35(Web Server issue), W193–W200. https://doi.org/10.1093/nar/gkm226

4. **Ashburner, M., Ball, C.A., Blake, J.A., et al.** (2000). *Gene Ontology: tool for the unification of biology.* **Nature Genetics**, 25(1), 25–29. https://doi.org/10.1038/75556
5. **Kanehisa, M., Goto, S.** (2000). *KEGG: Kyoto Encyclopedia of Genes and Genomes.* **Nucleic Acids Research**, 28(1), 27–30. https://doi.org/10.1093/nar/28.1.27
6. **Fabregat, A., Jupe, S., Matthews, L., et al.** (2018). *The Reactome Pathway Knowledgebase.* **Nucleic Acids Research**, 46(D1), D649–D655. https://doi.org/10.1093/nar/gkx1132
7. **AlzGene Database**. (2007). *A comprehensive database of Alzheimer's disease genetic association studies.* http://www.alzgene.org/
8. **Barabási, A.-L., Gulbahce, N., Loscalzo, J.** (2011). *Network medicine: a network-based approach to human disease.* **Nature Reviews Genetics**, 12(1), 56–68. https://doi.org/10.1038/nrg2918
9. **Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.** (2008). *Fast unfolding of communities in large networks.* **Journal of Statistical Mechanics: Theory and Experiment**, 2008(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008
10. **Scikit-learn Developers**. (2024). *Machine Learning in Python.* https://scikit-learn.org/

11. **Plotly Technologies Inc.** (2023). *Interactive graphing and analytics tools.* https://plotly.com/python/

12. **NetworkX Developers**. (2024). *Network Analysis in Python.* https://networkx.org/

13. **GeneCards – The Human Gene Database**. https://www.genecards.org/

14. **NCBI Gene Expression Omnibus (GEO)**. https://www.ncbi.nlm.nih.gov/geo/

15. **Huang, D.W., Sherman, B.T., Lempicki, R.A.** (2009). *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* **Nucleic Acids Research**, 37(1), 1–13.

16. **Buffalo, V.** (2015). *Bioinformatics Data Skills: Reproducible and Robust Research with Open Source Tools.* O'Reilly Media, Inc.