

**Identification Of Risk Genes For Neurodegenerative Diseases**

**PROJECT SUPERVISOR**

Mr. Shoaib Rauf

**PROJECT TEAM**

Robaisha Masood K20-0390

Noor Afshan K20-0266

Hira Tahir K20-0374

Submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science.

FAST SCHOOL OF COMPUTING

NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES KARACHI CAMPUS

May 2024

| Project Supervisor | Mr Shoaib Rauf |
| --- | --- |
| Project Team | Robaisha Masood K20-0390  Noor Afshan K20-0266  Hira Tahir K20-0374 |
| Submission Date | May 16, 2024 |

**Mr. Supervisor Name** Shoaib Rauf

**Supervisor**

**Dr. Muhammad Atif Tahir \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Head of Department**

FAST SCHOOL OF COMPUTING

NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES KARACHI CAMPUS

May 2024

# 

# **ANTI-PLAGIARISM DECLARATION:**

This is to declare that the above FYP report produced under the: Title: “Identification of Risk Genes for Neurodegenerative Diseases” is the sole contribution of the author(s) and no part hereof has been reproduced on as it is basis (cut and paste) which can be considered as Plagiarism. All referenced parts have been used to argue the idea and have been cited properly. We will be responsible and liable for any consequence if the violation of this declaration is determined.

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student 1

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Signature:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student 2

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Signature:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student 3

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Signature:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Supervisor (Faculty)

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Signature:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# 

# **AUTHOR’S DECLARATION:**

This states the Authors’ declaration that the work presented in the report is their own, and has not been submitted/presented previously to any other institution or organization

# 

# 

# **ABSTRACT**

The purpose of this research is to focus on Alzheimer's disease, which is a neurodegenerative condition often caused due to progressive dysfunction of neurons. Specifically, this study has unraveled Alzheimer's pathology by gaining insights from the dataset GSE118553 from NCBI GEO DATA. These genes form communities that cause malfunctioning. The Dataset contains brain tissue data from 27 control, 33 AsymAD, and 52 Alzheimer subjects. Checked affected areas (entorhinal, temporal, frontal cortex) and spared tissue (cerebellum). Derived meaningful insights from the raw data. Utilizing the advanced data mining method Weighted Gene Co-expression Network (WGCNA) is used to infer the hub genes and their role in such diseases by using pairwise correlation between genes. Furthermore, this study has exploited Machine Learning algorithms to validate our results ensuring accuracy and effectiveness.

Overall our research has uncovered novel genes that may be involved in the pathogenesis of Alzheimer's offering prospects in recognition of AD biomarkers and therapeutic targets. The study discovers the application of Machine Learning algorithms within the domain of bioinformatics. This discovery will facilitate the scientific community to produce more efficacious treatments for neurodegenerative disorders.

**ACKNOWLEDGEMENT**

We would like to express our deep appreciation and gratitude to our supervisor, Mr. Shoaib Rauf for his support and valuable insights throughout our project tenure. His unwavering guidance was the key reason behind the successful implementation of our project.

Moreover, We would like to appreciate the assistance provided by the members of the Video Surveillance Lab (VSL) for their support throughout the duration in using and performing our task on the machines installed in the Lab.

In last, we would like to thank each, and every person involved directly or indirectly in helping us to achieve our goal of successfully implementing our project titled “Identification of Risk Genes for Neurodegenerative Diseases”

# 

# 

# 

# 

Contents

[**ANTI-PLAGIARISM DECLARATION:** 2](#_2et92p0)

[**AUTHOR’S DECLARATION:** 3](#_3dy6vkm)

[**ABSTRACT** 4](#_2s8eyo1)

[**LIST OF TABLES** 8](#_44sinio)

[**LIST OF FIGURES** 8](#_z337ya)

[**CHAPTER 1. INTRODUCTION** 9](#_3j2qqm3)

[**PROBLEM DOMAIN** 10](#_1y810tw)

[**RESEARCH PROBLEM STATEMENT** 11](#_4i7ojhp)

[**CHAPTER 2. LITERATURE REVIEW** 12](#_qsh70q)

[**Related Work** 12](#_3as4poj)

[**Dataset** 21](#_49x2ik5)

[**Normalization and Pre-Processing** 24](#_32hioqz)

[**Downstream Methods** 25](#_3fwokq0)

[**Machine Learning** 27](#_28h4qwu)

[**CHAPTER 3. PROPOSED APPROACH** 29](#_111kx3o)

[**Applied Workflow** 29](#_3l18frh)

[**Chapter 4. Implementation** 31](#_1egqt2p)

[**1.**](#_3ygebqi) **Data Preprocessing-R** 31

[***Data Normalization*** 31](#_2dlolyb)

[***Data Transformation*** 31](#_sqyw64)

[***Join Operation*** 31](#_3cqmetx)

[***Joining Data with Brain Part Information*** 31](#_1rvwp1q)

[**2.**](#_4bvk7pj) **Data Preprocessing-Python** 32

[***Importing Libraries*** 32](#_2r0uhxc)

[***Command Line Arguments*** 32](#_1664s55)

[***Read Matrix File*** 32](#_3q5sasy)

[***Read Family File*** 32](#_25b2l0r)

[***Remove Duplicate Probes*** 32](#_kgcv8k)

[***Output Matrix and Trait File*** 32](#_34g0dwd)

[**3.**](#_43ky6rz) **Weighted Gene Coexpression Analysis(WGCNA)** 33

[***Loading Libraries and Data*** 33](#_2iq8gzs)

[***Preparing Dataset*** 33](#_xvir7l)

[***Detecting Outliers*** 33](#_3hv69ve)

[***Preparing Trait Data and Visualizing Clusters*** 33](#_1x0gk37)

[***Creating the Network*** 33](#_4h042r0)

[***Grouping Genes into Modules*** 34](#_1baon6m)

[***Merging Modules and Gene Dendogram*** 34](#_3vac5uf)

[***Associating Modules with Phenotype Traits*** 34](#_2afmg28)

[***Analyzing Module Membership and Gene Significance*** 34](#_pkwqa1)

[**Requirements** 35](#_48pi1tg)

[***1.***](#_2nusc19) ***Dataset Websites:*** 35

[**3.**](#_1302m92) ***Hardware***: 35

[***4.***](#_3mzq4wv) ***Process Requirements*** 36

[**Chapter 5. Validation and Testing** 37](#_haapch)

[**1.**](#_319y80a) **Machine Learning** 37

[***1.***](#_1gf8i83) ***Importing Libraries:*** 37

[***2.***](#_40ew0vw) ***Data Preparation:*** 37

[***3.***](#_2fk6b3p) ***Data Imputation*** 37

[***4.***](#_upglbi) ***Model Training*** 37

[***5.***](#_3ep43zb) ***Model Evaluation*** 37

[***6.***](#_1tuee74) ***Results*** 38

[**Chapter 6. Results and Discussion** 39](#_4du1wux)

[**Results of WGCNA** 39](#_2szc72q)

[**Chapter 7. Conclusions and Future Work** 45](#_279ka65)

[**Future Work** 45](#_meukdy)

[1.](#_36ei31r) ***Research Papers Contribution*** 45

[2.](#_45jfvxd) ***Advancement in Genetics*** 45

[**3.**](#_zu0gcz) ***Impact on Welfare and Society*** 45

[**Conclusion** 45](#_1yyy98l)

[**References** 46](#_2y3w247)

# 

# **LIST OF TABLES**

| 1. Information of GSE118553 | 20 |
| --- | --- |
| 2. Gantt Chart | 28 |
| 3. Libraries Required For Each Process | 35 |
| 4. Results Obtained After Different Machine Learning Algorithms | 37 |

# 

# **LIST OF FIGURES**

| 1. Network Analysis Outcome | 38 |
| --- | --- |
| 2. Module-trait Relationship | 39 |
| 3. Cluster Dendrogram | 40 |
| 4. Sample Clustering to Detect Outliers | 41 |
| 5. Module Membership vs Gene Significance | 42 |
| 6. Clustering of Module Eigengenes | 43 |

# **CHAPTER 1. INTRODUCTION**

Neurodegenerative diseases are chronic brain conditions when the neurons start to degenerate gradually due to abnormal accumulation and misfolding of specific proteins. Neurodegenerative diseases cause memory loss and behavioral abnormalities. The degeneration of neurons leads to neuronal damage and causes memory loss in the patient at different stages. Neurodegenerative disease syndrome includes Alzheimer's disease, Parkinson's disease, and Huntington's disease.

Neurodegenerative diseases pose significant challenges, causing a gradual decline in brain function and ultimately leading to neuronal death. Among these conditions, Alzheimer's disease (AD) stands as a prominent example, affecting millions globally.

Recent research in the domain of neuroscience has discovered some factors that potentially cause the pathogenesis of neurodegenerative disease, especially Alzheimer's. One of the research hypotheses suggests that the mental levels of deregulation in the brain may play a very vital role in the development of neurodegenerative disorders. This hypothesis is supported by the observation of amyloid aggregation and oxidative stress, both of which are pathological factors associated with the disease. To further investigate the matter, our study will focus on identifying the genes working together that govern the pathogenesis of the disease.

Our study focuses on leveraging such tools to investigate the genetic basis of Alzheimer's disease, especially by examining the [GSE118553](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118553) dataset. This dataset contains genetic information from brain samples of healthy individuals, those with early-stage AD that is AsymAD, and those diagnosed with Alzheimer's.

Our approach employs Weighted Gene Co-expression Network Analysis (WGCNA).To ensure the reliability of our findings, we also utilize machine learning algorithms. By integrating these methods, our research aims to deepen our understanding of Alzheimer's disease, identify potential biomarkers for early diagnosis, and uncover new therapeutic targets.

## **PROBLEM DOMAIN**

Neurodegenerative diseases are complex conditions characterized by the progressive degeneration of neurons in the brain, leading to a range of debilitating symptoms such as memory loss, cognitive decline, and behavioral abnormalities. Alzheimer's disease and Parkinson's disease are among the most common neurodegenerative disorders affecting millions of individuals.

The burden of neurodegenerative diseases is particularly pronounced given the shift towards an aging population. With increasing life expectancy, these disorders are on the rise, posing significant challenges for healthcare systems and society at large.

Understanding the underlying mechanism driving neurodegenerative diseases is crucial for developing effective treatments and interventions. Recent research has implicated various factors in disease pathogenesis like protein misfolding, and oxidative stress. These molecules contribute to progressive loss of neuronal integrity and function ultimately leading to neurodegeneration.

To address these challenges, approaches combining molecular biology, bioinformatics, and computational biology have emerged as tools for unraveling the complexities of neurodegenerative diseases. By integrating large-scale data and employing advanced techniques researchers aim to find genetic networks and signaling pathways underlying disease progression.

Moreover, finding key genes holds promise for early diagnosis and development of disease-modifying therapies. By computational approaches, researchers can improve the quality of life for individuals affected by neurodegenerative diseases.

## **RESEARCH PROBLEM STATEMENT**

Neurodegenerative diseases, including Alzheimer's, pose significant challenges due to their progressive nature and devastating impact on individuals and society. Understanding the molecules driving these diseases is crucial for developing effective treatments. Our study aims to address these knowledge gaps by leveraging a computational approach to analyze genomics data and identify key genes and their significance with Alzheimer's disease.

With the availability of the [GSE118553](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118553) dataset containing brain samples from control subjects, individuals with early signs of Alzheimer's, and patients diagnosed with disease there is an opportunity to extract meaningful insights into the molecular basis of Alzheimer's. By examining affected brain regions and tissues we seek to find genetic networks involved in disease progression.

Our research seeks to identify genes that play vital roles in the pathogenesis of Alzheimer's disease. It utilizes the weighted Gene Co-expression Network Analysis (WGCNA) to uncover interconnected gene molecules and their significance. Validate our findings using machine learning algorithms to ensure the reliability of results. Finally, our goal is to contribute to the development of genes for early diagnosis and targets for Alzheimer's disease to improve outcomes for affected individuals.

# 

# 

# 

# 

# **CHAPTER 2. LITERATURE REVIEW**

Neurodegenerative diseases are a significant challenge for public health. The study of neurodegenerative disorders demands a needful understanding of the genetic foundations. This literature review highlights the key findings from the recent research in neurodegenerative disease genetics.

## **Related Work**

The authors of the study introduce a novel approach called GeneEMBED, which aims to identify gene interactions associated with complex diseases such as Alzheimer's. Through the application of GeneEMBED on multiple datasets related to Alzheimer's, researchers successfully discovered previously unidentified genes (namely PLEC, UTRN, TP53, and POLD1) that play a role in this disease. Additionally, it was observed that two out of these four genes are targeted by approved pharmaceutical drugs. **[1]**

The "Network dysfunction perspective" on neurodegenerative diseases, as proposed by Jorge J. Palop, Jeannie Chin, and Lennart Mucke, focuses on understanding these diseases with the help of examination of broader neural networks and connection within the brain, rather than working with a single individual area. **[2]**

The paper "Chapter 21 - Concepts and Classification of Neurodegenerative Diseases" by Gabor G. Kovacs discusses neurodegenerative diseases and follows the classification by providing a detailed description of the neuropathology of Alzheimer's disease, alpha-synucleinopathies, tauopathies, FTLD with TDP-43 or FUS/FET proteinopathies, trinucleotide repeat disorders, and prion diseases. **[3]**

" Applications of machine learning to diagnosis and treatment of neurodegenerative diseases" by Monika A. Myszczynska, Poojitha N. Ojamies, …Laura Ferraiuolo explains how machine learning is helping in diagnosis in the early stage and interpretation of medical images as well as the discovery and development of new therapies. This also helps in multiple high-dimensional sources of data that provide different views on different diseases. **[4]**

Wim Mandemakers 1, Vanessa A Morais, and Bart De Strooper's article "A cell biological perspective on mitochondrial dysfunction in Parkinson's disease and other Neurodegenerative diseases"; focus on mutations of reasons causing Parkinson's disease.

It tells the study of these mutations and it causes other diseases that indicate mitochondrial dysfunction which becomes the main contributor to neurodegenerative processes. **[5]**

According to Stanley B. Prusiner, M.D.'s "Neurodegenerative Diseases and Prions," it is obvious that prion or neurodegenerative diseases are caused by anomalies in the course of these illnesses, which further cause the accumulation of particular neuronal proteins. Laboratory search results led to the discovery of prions, yielding findings like infectious pathogens and central nervous system degeneration. **[6]**

"Alzheimer Disease and Related Neurodegenerative Diseases in Elderly Patients With Schizophrenia" by Dushyant P. Purohit, Daniel P. Perl, Vahram Haroutunian, et al. Clinical studies suggest that severe cognitive impairment is common in older people with schizophrenia who reside mostly in psychiatric people and that its result is in conflict with Alzheimer's disease when combined with schizophrenia. **[7]**

The study, “Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease.” by Lambert, J. C., et al. (2013); was conducted on 74,046 individuals, which led to 11 discoveries associated with Alzheimer’s disease. This study served a great purpose in highlighting the importance of large-scale collaborative research in deciphering the complex genetic landscape of this condition. **[8]**

Jun et al. (2016) presented their opinion in the novel by identifying an uncharted Alzheimer's disease locus near the tau gene. This discovery essentially shows that there could be a potential link between tau protein pathology and susceptibility to Alzheimer’s disease. **[9]**

Karch, C. M., et al. (2014) presented their study, “ Alzheimer's Disease Genetics: from the Bench to the Clinic.”. The key discussion in this paper was the need to build the gap between genetic discoveries and clinical applications in Alzheimer’s disease genetics. Their work highlights the genetic factors involved in Alzheimer's disease and it also underscores the importance of applying this knowledge in clinical practice to potentially improve the diagnosis and the treatments. **[10]**

John Hardy and Dennis J. Selkoe's paper, "Alzheimer's disease: The amyloid cascade hypothesis," examines emphasis on the causal link between mitochondrial gene mutations and neurodegenerative disorders such as Parkinson’s and Alzheimer’s diseases. It reviews structural and functional studies, highlighting mitochondrial dysfunction's role. In the study, animal models provide evidence for mitochondria’s involvement in disease initiation and progression. **[11]**

“Neurodegenerative disorders associated with genes of mitochondria'' by Vaibhav S. Marde et al. (2021); emphasizes the causal link between mitochondrial gene mutations and neurodegenerative disorders such as Parkinson’s and Alzheimer’s diseases. It reviews structural and functional studies, highlighting mitochondrial dysfunction's role. In the study, animal models provide evidence for mitochondria’s involvement in disease initiation and progression. **[12]**

Common polygenic variations have a significant impact on Alzheimer's disease research, according to Escott-Price, who highlights how these genetic variations improve risk assessment, early diagnosis, and stage prediction. Microglial-Mediated Innate Immunity. **[13]**

The methodology used in the study conducted by Guan et al. is related to our project on identifying coexpression networks of genes for neurodegenerative diseases like Alzheimer's in terms of using weighted gene coexpression network analysis (WGCNA) to analyze genomic data. WGCNA is a powerful tool for identifying modules or clusters of genes that are highly correlated and may play a role in disease mechanisms. By constructing separate networks for tumor microbial data and messenger RNA (mRNA) data, the study identified the oncogene-associated microbiome module (OAM), which was found to be tumor-specific enriched and had a better prognostic value when combined with the oncogene. In our project, we aim to apply similar techniques to analyze gene expression data in neurodegenerative diseases like Alzheimer's, to identify coexpression networks that could potentially reveal novel biomarkers or therapeutic targets. **[14]**

This open-access article presents a study on Alzheimer's disease pathogenesis, in which researchers used co-expression network analysis to identify novel genes related to AD. The study involved functional enrichment analysis, protein-protein interaction network analysis, and gene set enrichment analysis for the hub genes. The results suggest that the identified genes are mainly involved in axonogenesis, synaptic transmission, and ion transport, among other processes. The study provides new insights into the molecular mechanisms underlying AD and may contribute to the development of new treatments for this disease.

In this study, the researchers used the Weighted Gene Co-expression Network Analysis (WGCNA) method to construct a co-expression network for a gene dataset with a large number of samples from the brain tissues of AD patients and normal controls. They then screened the hub genes that were related to AD pathogenesis and validated the robustness of the expression of hub genes using an independently validated cohort. They also analyzed the pathways and clinical significance of the hub genes.

Overall this study provides a useful example of how co-expression network analysis can be used to identify novel genes related to AD pathogenesis and may provide insights that can be applied to your research on identifying risk genes for AD. **[15]**

This article discusses the role of AI algorithms in medical diagnosis, specifically focusing on their contribution to classifying Alzheimer's disease. This study aims to consolidate relevant knowledge about machine learning models used in Alzheimer's identification, presenting results from different algorithms employed for diagnosis. This paper tells further about work done in the medical research field that is related to Alzheimer's disease, comparing the efficiency and error rates. The article highlights the shift from methods relying solely on individual features from sMRI images for classification, which often resulted in low accuracy. Instead, recent multimodal studies demonstrate improved classification accuracy by combining multiple features from different sMRI analysis techniques. This approach enhances the understanding and diagnosis of Alzheimer's disease and its stages. **[16]**

This paper explores the potential of artificial intelligence (AI) and machine learning (ML) algorithms to enhance Alzheimer's detection. Recognizing the limitations of conventional approaches, the study employs big data processing to gather information from diverse sources, considering the evolving nature of the disease. Previous research using Support Vector Machine (SVM) showed low accuracy, prompting the need for improved precision. The paper presents alternative algorithms to enhance efficiency, revealing that the Support Vector Machine with a linear kernel outperforms other models. The existing method utilizing SVM for prediction faces challenges such as time inefficiency, particularly in handling large, heterogeneous datasets. The paper highlights the drawbacks of traditional data mining tools and proposes a new approach using algorithms like Linear Regression, SVM, Decision Tree, Random Forest, and Naïve Bayes. The study recognizes the complexity of big data processing and emphasizes the role of supercomputing platforms in overcoming obstacles at the data, model, and system levels. Advantages of the proposed method over the conventional SVM approach include faster business value realization, early detection of Alzheimer's, identification of optimal treatment modes across different age groups, and insights driving growth and profitability. **[17]**

The brain, a remarkable organ overseeing various physical and cognitive functions, undergoes structural and functional changes in Alzheimer's Disease (AD), a chronic condition contributing to the majority of dementia cases. Early diagnosis is crucial for effective intervention, as AD worsens over time. Traditional testing methods are time-consuming, prompting the exploration of machine learning models for efficient and early detection. This paper focuses on the detection of Alzheimer's using two datasets: the Longitudinal dataset with text values and the OASIS dataset containing MRI images. Utilizing 14 machine learning algorithms, including the Random Forest Algorithm with a peak accuracy of 92.1385% and the KNN Algorithm with a baseline accuracy of 47.1910%, the OASIS Longitudinal dataset demonstrates efficient detection capabilities. For MRI images, various Transfer learning models are employed, with the InceptionV3 model and ADAM optimizer yielding the highest accuracy. The study emphasizes the potential of machine learning in achieving accurate and early Alzheimer's detection, presenting a multifaceted approach by considering diverse datasets and models. **[18]**

This study addresses the crucial need for early and accurate diagnosis of Alzheimer's disease (AD) using a multi-diagnostic and generalizable approach. Leveraging machine learning (ML) and structural MRI, the study develops classifiers for mild cognitive impairment (MCI) and AD diagnosis, demonstrating transparency, interpretability, and generalizability across datasets and MRI protocols. The classifiers, trained and tested on the AD Neuroimaging Initiative (ADNI) and Open Access Series of Imaging Studies (OASIS) databases, achieved impressive results. The "healthy controls (HC) vs. AD" classifier attains a balanced accuracy of 90.6% and Matthew's correlation coefficient (MCC) of 0.811. The "HC vs. MCI vs. AD" classifier, trained on the ADNI dataset, achieves a balanced accuracy of 62.1%, with hippocampal features identified as significant contributors. The study's diagnostic tool is unique in its multi-diagnostic approach, considering HC, MCI, and AD simultaneously, providing a more practical solution for clinicians. It also addresses the challenge of generalizability by evaluating performance across independent datasets and various MRI protocols. The study presents a promising ML-based diagnostic tool for MCI and AD, offering a comprehensive and clinically applicable solution. The multi-diagnostic approach, generalizability across datasets and protocols, and transparent reporting contribute to the robustness and potential clinical utility of the proposed classifier. **[19]**

This paper explores the application of machine learning algorithms, including Decision Trees, SVM, Logistic Regression, and Naive Bayes, to identify Alzheimer's disease at an early stage. Utilizing datasets from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Open Access Series of Imaging Investigations (OASIS), which include longitudinal MRI data and demographic information, the study aims to enhance diagnostic efforts. The datasets encompass factors such as age, gender, mini-mental status, and Clinical Dementia Rating (CDR). Precision, F1 Score, Recall, and specificity are considered in evaluating each method. The results reveal that the Decision Tree Algorithm achieves a maximum accuracy of 93.7%, showcasing its effectiveness in early Alzheimer's disease detection. Alzheimer's disease is highlighted as a progressive condition with initial signs including difficulty recalling recent events. The study emphasizes the challenges in clinical diagnosis, leading to the exploration of automated systems for medical decision support. **[20]**

This study aimed to identify feature genes associated with Alzheimer's disease (AD) by analyzing data from three Gene Expression Omnibus (GEO) databases: GSE122063, GSE15222, and GSE138260. The datasets were filtered based on AD-related keywords, Homo sapiens as the selected species, and a sample size greater than 20 for each dataset, including both normal and AD groups. GSE15222 and GSE138260 were combined as a training group to build a model, while GSE122063 served as a test group for model verification. Weighted gene co-expression network analysis (WGCNA) on the combined dataset identified AD-related module genes. The intersection of differential and AD-related module genes resulted in AD key genes. LASSO regression further filtered these genes, leading to the identification of three AD-related feature genes: SST, MLIP, and HSPB3. Differential expression analysis in the combined datasets revealed 111 common differential AD genes. Gene Ontology (GO) analysis highlighted terms related to cognition, learning, and memory. Kyoto Encyclopedia of Genes and Genome Pathways (KEGG) analysis identified enrichment in pathways such as neuroactive ligand-receptor interaction, cAMP signaling, and Calcium signaling. **[21]**

Weighted Gene Co-expression Network Analysis (WGCNA) is a widely used method in bioinformatics for exploring correlation patterns among genes in microarray samples. This approach involves identifying clusters (modules) of highly correlated genes, summarizing these clusters, and relating them to external sample traits. The WGCNA R software package, presented in this paper, offers a comprehensive set of functions for performing various aspects of weighted correlation network analysis. These functions include network construction, module detection, gene selection, topological property calculations, data simulation, visualization, and integration with external software. The package is user-friendly and can be applied not only to gene expression data but also to various biological contexts, such as cancer, genetics, and brain imaging data. The accompanying tutorials enhance the accessibility of the software, making it a valuable tool for researchers in the field. **[22]**

The study explores the shared pathophysiological mechanisms of Alzheimer's Disease (AD) and Type 2 Diabetes Mellitus (T2DM) through co-expression network analysis. Microarray data of AD and T2DM were obtained from the Gene Expression Omnibus (GEO) database, and co-expression networks were constructed using Weighted Gene Co-Expression Network Analysis (WGCNA). Gene Ontology (GO) and pathway enrichment analyses were performed on common genes related to AD and T2DM modules. The results revealed significant modules for both AD and T2DM, enriched in pathways such as circadian entrainment, phagosome, glutathione metabolism, and synaptic vesicle cycle. Protein-protein interaction network analysis identified 10 hub genes (CALM1, LRRK2, RBX1, SLC6A1, TXN, SNRPF, GJA1, VWF, LPL, AGT) shared between AD and T2DM. The findings suggest common pathogenesis and shared pathways, providing insights for further mechanistic studies and potential therapeutic targets for AD and T2DM. **[23]**

The study explores the interrelationships between pathological processes and clinical phenotypes in Alzheimer's disease (AD) using graph theory. Neuroimaging techniques reveal disruptions in brain networks, with findings pointing to a loss of highly connected areas in AD. However, variability in reported group differences suggests non-isometric brain graphs, complicating interpretation. The use of graph theory provides quantitative measurements for understanding the integrated nature of local brain activity, highlighting the significance of hubs in information processing and disease propagation within the network. Confounding factors, such as differences in graph construction methods, are identified, and recommendations for future research are provided. **[24]**

This study investigated graph theory metrics in the context of Alzheimer's disease (AD). Resting-state functional connectivity MRI was utilized to assess functional integration, functional segregation, and functional distinctness. The study found reductions in clustering coefficient and modularity, indicative of large-scale disconnection in symptomatic AD. Cognitively normal participants with preclinical AD biomarkers also exhibited similar, albeit smaller, reductions in these graph measures. The impact on modularity was influenced by age, and AD was observed to affect hub-like regions in the brain. These findings suggest significant brain changes in preclinical AD, emphasizing the relevance of large-scale disconnection even before symptomatic onset. **[25]**

In summary, this literature review offers a rich tapestry of insights into neurodegenerative diseases, with a particular emphasis on identifying risk genes. The GeneEMBED approach introduces an innovative method for uncovering genes associated with Alzheimer's, showcasing the potential for targeted gene identification. The exploration of network dysfunction, disease classification, and the integration of machine learning not only broadens our understanding but also provides practical avenues for identifying crucial risk genes. The diverse perspectives on mitochondrial dysfunction, prions, and disease co-occurrence contribute to a holistic comprehension of the genetic landscape. As our research focuses on pinpointing risk genes, this review lays a robust foundation, offering a nuanced and interdisciplinary perspective for advancing our understanding of neurodegenerative diseases at the genetic level.

## **Dataset**

Link: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118553>

Table 1. Information of GSE118553

| Sample Source | AD | Control | AsymD |
| --- | --- | --- | --- |
| Temporal Cortex(TC) | 52 | 31 | 32 |
| Frontol Cortex(FC) | 40 | 23 | 33 |
| Entorhinal Cortex(EC) | 37 | 24 | 37 |
| Cerebellum(CE) | 38 | 22 | 32 |

#### ***Summary Of Researched Items:***

The selection of an appropriate dataset lays the foundation for genomic research. The [GSE118553](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118553) dataset, chosen for its inclusion of brain samples from control subjects, individuals from asymptomatic Alzheimer's (AsymAD), and diagnosed Alzheimer's patients provides a rich resource for investigating the molecules of disease. By analyzing gene expression profiles across different disease stages and brain regions, researchers can gain insights into dynamic changes occurring in tissues that have Alzheimer's. This dataset worked closely with proposed study objectives which seek to identify key genes and pathways associated with Alzheimer's pathogenesis. Leveraging the comprehensive nature of the dataset, the proposed research aims to extract meaningful insights that can inform the development of diagnosis and targets for Alzheimer's disease.

#### 

#### **Critical Analysis of Researched Items:**

**Strengths:**

1. Variety of samples:The dataset covers a wide range of samples which also includes brain tissues from people without Alzheimer's known as controlled, those with early signs of the disease known as AsymAD, and those with Alzheimer's. The variety allows researchers or us to look closely at different stages of diseases and different parts of the brain. By studying samples from various stages and regions, scientists can understand how Alzheimer's changes over time and how it affects brain regions. This helps in finding important details about the diseases.

*2.* Access to Raw Data: Having access to raw data means researchers can find deep details in information and explore many different ideas. Raw data has a lot of details about how genes are expressed, any genetic differences, or how they are controlled. With this raw data researchers can use even advanced methods to analyze for example finding gene expression building networks or finding gene significance so this helps in uncovering new items.

**Weaknesses:**

1. Samples and Diversity: Even though there are many samples, there is a possibility that the dataset might represent everyone equally. Some groups might have more or fewer samples, which could make the results less reliable. Also, the people in the dataset might not represent all different types of people who are affected. Things like age, gender, or background can affect it and contain mostly female gender datasets so results might not apply to everyone with Alzheimer's.

2. Need for cleaning and preparing data: This dataset is complicated, it needs to be carefully checked and prepared before using it. Things like this make sure all datasets are in the same format, checking for mistakes and removing any weird results are important. If these aren't done properly, the result could be wrong or misleading and it is also important to make sure that data is of the same quality, especially with different methods so data can be trustworthy and can be used to understand Alzheimer's better.

#### ***Relationship of proposed research work***

The selection of the [GSE118553](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118553) dataset aligns with the objective of this research which aims to identify key genes associated with the pathogenesis of Alzheimer's disease. By leveraging this dataset, researchers can analyze gene expression patterns and identify potential targets for Alzheimer's disease.

## **Normalization and Pre-Processing**

#### ***Summary Of Researched Items:***

Normalization and preprocessing of this data are essential steps to ensure the accuracy and reliability of downstream analyses. In the study of Xingxing Zhao, our base paper, the author investigated various normalization methods for gene expression data, including quantile normalization and RMA. The study compares the performances of these normalization techniques in terms of preserving biological genes.

#### ***Critical Analysis of Researched Items:***

**Strengths:**

1. Comparison of normalization methods:We study multiple methods like RNA and Quantile and this provides researchers with valuable insights into their strengths and limitations. By presenting a comprehensive overview researchers can select the most appropriate normalization approach.

2. Emphasis on preprocessing importance:The researchers find the significance of pre-processing steps in mitigating batches that affect the dataset. By addressing these factors early, we can improve data quality and reduce the likelihood of factors affecting the study.

**Weaknesses:**

1. Complexity:While the study provides a comparison of normalization methods it may not fully capture the complexities associated with normalization over different platforms and then this dataset can exhibit variability due to factors like sample preparation techniques, and sequencing.

2.Limited discussion on Downstream Analyses:The researcher gives less discussion on the impact of normalization on downstream analysis like WGCNA. Understanding how normalization techniques affect this is a crucial step for interpreting study findings accurately.

#### ***Relationship of proposed research work***

Normalization and preprocessing of datasets are crucial steps in ensuring the accuracy and reliability of downstream analyses. The literature review evaluates various normalization methods and their impact on reducing technical variation and preserving signals in gene expression data. By incorporating best practices in data preprocessing, the study aims to ensure robust results. Researchers can mitigate potential biases and enhance the reproducibility of their findings.

## **Downstream Methods**

#### ***Summary Of Researched Items:***

In war to identify the significant genes associated with complex diseases like Alzheimer's researchers often turn to network-based approaches like we used Gene Regulatory Network (GRN) and Weighted Gene Co-expression Network Analysis (WGCNA). Initially, we utilize GRN analysis to explore gene interaction related to Alzheimer's disease. However, upon encountering challenges and looking at the advantages of WGCNA, we transitioned to this method. WGCNA allowed them to identify co-expressed gene modules linked to Alzheimer's revealing their significance.

#### ***Critical Analysis of Researched Items:***

**Strengths:**

1. Understanding Gene Interactions: Network-based approaches give us a big-picture view of how genes interact and control each other, which helps us understand the complex mechanism behind diseases like Alzheimer's.

2.Finding genes significance: WGCNA is especially good at finding groups of genes that work together and play a role in diseases. These gene clusters are like genes that work to find their significance and identify them.

**Weaknesses:**

1. Inconsistency in finding gene significance: Finding gene significance can vary a lot depending on the data we have, how the experiment was done, and which methods were used. This inconsistency can lead to different results making it hard to compare.

2. Need for more understanding and checking:Even though we find genes, we still need to make sure they are biologically meaningful. We have to validate our findings by checking if these gene clusters do important things. Understanding the roles of significance and how they interact within these groups is crucial for turning our discoveries into useful information for Alzheimer's research.

#### ***Relationship of proposed research work***

The transition from GRN to WGCNA reflects the iterative nature of scientific inquiry and methodology that offer greater insights and advantages. In the proposed research, the application aligns intending to identify gene significance associated with Alzheimer's disease.

### 

## **Machine Learning**

#### ***Summary Of Researched Items:***

Machine learning algorithms play a crucial role in dataset analysis, enabling researchers to uncover complex patterns. In our study, we utilized machine learning techniques such as Random Forest to validate the results of Alzheimer's disease based on gene expression profiles.

#### ***Critical Analysis of Researched Items:***

**Strengths:**

1. Accurate and scalable Analysis: Machine learning algorithms provide highly accurate prediction and can easily handle large datasets. These algorithms are good at preprocessing, making them valuable tools for analyzing complex systems like those in Alzheimer's.

2. Uncovering complex relationships: Machine learning methods can uncover non-linear relationships and interactions between genes and phenotypes which may be missed by other methods

**Weaknesses:**

1. Limited interpretability: One of the major drawbacks is their limited interpretability which can create problems in understanding disease mechanisms. While they can give accurate predictions understanding how and why they make that particular prediction can be challenging so it makes it difficult to extract meaningful insights from ML analyses

2. Dependency on Dataset characteristics: The performance of ML may vary on a dataset and its feature selection methods. Factors like data quality, sample size, and choice of feature can significantly impact predictive performance so these factors should be carefully considered.

#### ***Relationship of proposed research work***

ML algorithms offer powerful tools for analyzing complex neurodegenerative disease datasets and uncovering patterns that may not be possible in traditional methods. The literature review examines the application of machine learning techniques such as Random Forest in the validation of results of Alzheimer's disease based on their gene expressions. Based on this theory we aim to enhance the accuracy and reliability of the prediction. Through the integration of ML into the research framework seek the predictive power of computational models to advance our understanding of Alzheimer's disease.

# 

# 

# **CHAPTER 3. PROPOSED APPROACH**

Our research aims to investigate neurodegenerative diseases particularly AD by analyzing the gene expression dataset from [GSE118553](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118553) obtained from the NCBI Geo Data repository. We will use a combination of computational techniques to identify gene significance.

## **Applied Workflow**

Table 2. Gantt Chart

|  | **Sep** | **Oct** | **Nov** | **Dec** | **Jan** | **Feb** | **Mar** | **Apr** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***1. Study and Research*** | **Task 1** | **Task 1** |  |  |  |  |  |  |
| ***2.Data Collection and Data Preprocessing*** |  |  | **Task 2** |  |  |  |  |  |
| ***3. Construction of GRN*** |  |  |  | **Task 3** |  |  |  |  |
| ***4.ML Algorithm Research*** |  |  |  |  | **Task 4** |  |  |  |
| ***5. Construction WGCNA*** |  |  |  |  |  | **Task 5** | **Task 5** |  |
| ***6. Validation of results by ML*** |  |  |  |  |  |  |  | **Task 6** |

We have acquired the dataset [GSE118553](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118553) dataset which contains gene expression profiles from the brain tissues of control subjects, those with asymptomatic AD (AsymAD), and diagnosed AD patients. The dataset includes samples from affected brain regions (entorhinal, temporal, and frontal cortex) as well as spared tissue (cerebellum). Raw data will be preprocessed to remove noise and normalize gene expression levels.

We utilize the WGCNA framework to construct genes and to find gene significance in hierarchical manners associated with neurodegenerative diseases. By leveraging the WGCNA methodology we uncover those genes which are present in modules of co-expression associated with these diseases.

The constructed WGCNA and their significance will be analyzed to identify modules associated with neurodegenerative diseases. Genes in these modules will be prioritized based on their connectivity with disease phenotypes. Machine Learning models will be evaluated using cross-validation and performance metrics to assess their predictive accuracy.

# 

# 

# **Chapter 4. Implementation**

## **Data Preprocessing-R**

This step details the preprocessing steps implemented using R programming language, essential for preparing raw datasets for analysis phases. Our preprocessing aims to ensure data quality for integration and analysis necessary for identifying hub genes associated with Alzheimer's disease.

### ***Data Normalization***

Normalizing the raw data to eliminate technical variability, quantile normalization is used which adjusts the distribution of gene expression measurements across samples. This involves identifying numeric columns and applying the “normalize.quantiles” function from the preprocessingCore package. Non-numerics were excluded from normalization to maintain integrity.

### ***Data Transformation***

It is done to convert and prepare data for merging by ensuring column names and types. First extraction of the column is done which is common like Probe\_id for join and then it is loaded into the R dataframe. Then column names are aligned across the data frame to facilitate accurate merging.

### ***Join Operation***

Merge datasets to consolidate and map the gene to specific identifiers. For this base R function is used "merge" ensuring all observations are transformed in the dataset. Checking that no identifier like ID\_REF is missing was the crucial step. We have a complete normalized dataset ready for further evaluation.

### ***Joining Data with Brain Part Information***

As in the dataset, there were multiple files from which we extracted brain part information respective to a particular gene ID. For this normalized data frame headers were adjusted so we could get a common column which was "GENE\_ID" and then a join was performed. In as final view, we get to know brain parts with their type and in particular gene ID. Then Cytoscape is used to form GRN which is to visualize gene interaction.

## **Data Preprocessing-Python**

## ***Importing Libraries***

Imported necessary libraries such as sys, numpy and pandas.

### ***Command Line Arguments***

Get command line argument representing input and output file paths.

### ***Read Matrix File***

A function readMatrixFile is designed to read matrix file line by line. Parse each line to extract ID , expression list , headline , and trait information.In the end it returned a dictionary containing expression data , headline and trait information.

### ***Read Family File***

A function readFamilyFile is used to read data from the family file line by line. Parse each line to extract gene symbol and ID information,Function returned a dictionary mapping IDs to gene symbol

### ***Remove Duplicate Probes***

A function names removeDulProbes is used to remove duplicate probes and then an average value for expression values are calculated for each gene.Function iterates over the gene symbol dictionary and expression data.It returns a dictionary mapping gene symbols to expression values.

### ***Output Matrix and Trait File***

A csv file for information of trait and expression data is made.OutputTraitFile and OutputMatrixFile is generated respectively to write data to output files.

## 

## **Weighted Gene Coexpression Analysis(WGCNA)**

### ***Loading Libraries and Data***

Start with preparing the R environment with necessary WGCNA packages like WGCNA, and GO.db and configure R to handle string data and enable multithreading.

### ***Preparing Dataset***

Form clean expression data for network construction by loading the gene expression data and then transposing the data to match WGCNA requirements and removing non-numeric data.

### ***Detecting Outliers***

Identify the outliers to prevent skewing the analysis for this dendrogram is generated to inspect outliers groups.

### ***Preparing Trait Data and Visualizing Clusters***

First, import the trait data file and ensure that their sample names and ID matches expression data. Extract and process tissue type information from trait Data and then utilize a color palette to visualize different tissue types in the plot enhancing their interpretability. Generate hierarchical clustering of samples based on their expression profiles and overlay trait data as a color-coded heatmap.

### ***Creating the Network***

Select soft threshold power to use by evaluating scale-free index for key assumption in WGCNA. Then plotting mean connectivity and scale independence plot to visually assess how well different soft-threshold powers confirm to scale-free network. Then construct the network by choosing soft power to create adjacency and then transform it into a Topological Overlay Matrix (TOM) which is used to measure network connectivity of genes reflecting not just correlation but also indirect interaction.

### 

### ***Grouping Genes into Modules***

Identify modules based on TOM and analyze the relationship among eigengenes. Perform hierarchical clustering on TOM to group the genes and then apply dynamic tree cut to clustering to define gene modules as it allows to identification of distinct gene clusters. Calculate module eigengenes which are representative gene expression profiles for each module into a single profile to analyze correlated modules.

### ***Merging Modules and Gene Dendogram***

Merge closely related modules to simplify the network and reduce its complexity. This is done by assessing the dendrogram and merging modules below a certain height threshold which indicates similarity. Then construct gene modules using TOM which helps identify how genes are grouped before or after merging modules.

### ***Associating Modules with Phenotype Traits***

Analyze the relationship between gene modules and phenotype traits to uncover gene expression on specific traits so that after converting categorical values to numerical it is confirmed that trait data and expression data are consistent. Compute correlation and associated p-value between module eigengenes and numerical traits to identify significant relationships. Then generate a heatmap to visually represent the correlation between modules and traits.

### ***Analyzing Module Membership and Gene Significance***

Prepare data in numerical values for correlation analysis. Calculate the correlation between expression data and module eigengenes to establish gene module membership. Later compute p-values to assess statical significance. Determine the significance of genes concerning tissue type providing an insight into how gene expression relates to phenotype. Later plot the relationship between module membership and gene significance and convert it into a CSV file.

## 

## **Requirements**

The requirement of this project involves high-end machines with GPU due to the requirements of processing large amounts of data. Along with the machines some other requirements are mentioned below.

### ***Dataset Websites:***

* + 1. NCBI GEO Database [GSE118553](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118553)

1. ***Software:***
   * 1. R-studio version 1.4.1106
     2. Visual Studio Code
     3. Python version 3.9.x
     4. Jupyter Notebook/Google Colaboratory

### ***Hardware***:

* + 1. RAM 64 GB
    2. NVIDIA GeForce RTX 3090
    3. Driver version: 31.0.15.3713
    4. DirectX version: 12 (FL 12.1)
    5. GPU Memory 55.9 GB

### ***Process Requirements***

Table 3. Libraries Required For Each Process

| Data PreProcessing-R | 1. Affy 2. Tidyverse 3. Openxlsx 4. Dplyr 5. PreprocessCore |
| --- | --- |
| Data PreProcessing-Python | 1. Sys 2. Numpy 3. Pandas |
| WGCNA R-Packages | 1. BiocManager 2. GO.db 3. WGCNA 4. RColorBrewer 5. dynamicTreeCut |
| Machine Learning | 1. Pandas 2. sklearn.model\_selection import train\_test\_split 3. from sklearn.ensemble import GradientBoostingClassifier 4. from sklearn.impute import SimpleImputer 5. from sklearn.preprocessing import LabelEncoder 6. from sklearn.metrics import accuracy\_score, classification\_report, f1\_score, precision\_score, recall\_score 7. import xgboost as xgb 8. from sklearn.ensemble import RandomForestClassifier |

# **Chapter 5. Validation and Testing**

## **Machine Learning**

We have used three algorithms Random Forest, Gradient Boost, and XGBoost.All three algorithms follow the same approach for data preparation, model training, and prediction.

Here is the summary of the workflow:

### ***Importing Libraries:***

The necessary libraries are imported including pandas for data manipulation, Scikit Learn for Machine learning functionalities, and classifiers like GradientBoostingClassifier, xgboost, and RandomForestClassifier.

### ***Data Preparation:***

The features X are separated from the data frame, excluding columns such as identifiers and non-predictive variables. The target variable disease state is encoded using LabelEncoder to convert categorical variables into numerical values. The data is split into training and testing using train\_test\_split.

### ***Data Imputation***

Missing Values are imputed using the median value for each column.

### ***Model Training***

Classifiers are initialized and trained on imputed training data.

### ***Model Evaluation***

Predictions are made on imputed data. The accuracy of each classifier is tested using accuracy\_score. The Feature importance is obtained to find important hub genes.

### ***Results***

Model accuracy and the top 10 hub genes are printed using feature\_importance.

Additional evaluation metrics such as F1 score, Precision, and Recall are displayed in the form of classification reports are calculated and displayed.

Table 4. Results Obtained After Machine Learning Algorithms

|  | GradientBoost | XGBoost | RandomForest |
| --- | --- | --- | --- |
| Accuracy | 0.8182 | 0.8081 | 0.8184 |
| Precision | 0.8267 | 0.8166 | 0.8179 |
| Recall | 0.8182 | 0.8081 | 0.8182 |
| F1 Score | 0.8169 | 0.8078 | 0.8171 |

# **Chapter 6. Results and Discussion**

## **Results of WGCNA**

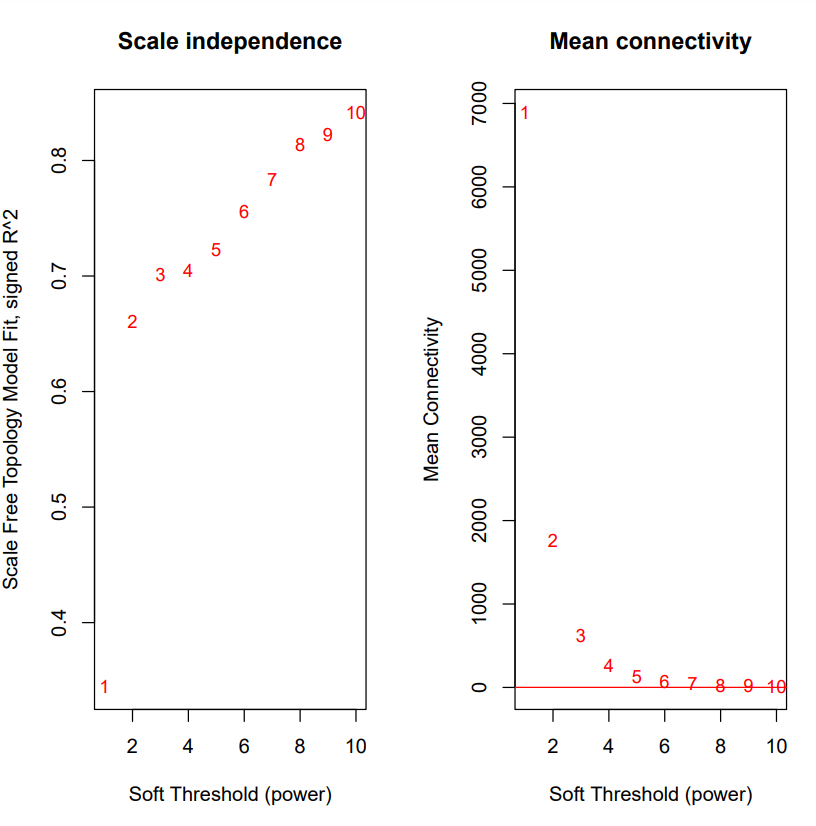


Figure 1. Network Analysis Outcomes

The graphs depict how network properties change with different soft threshold powers. The left graph assesses the scale-free fit (ideal above 0.8) and the right graph shows decreasing mean connectivity as power increases. helping select an appropriate threshold for network construction.

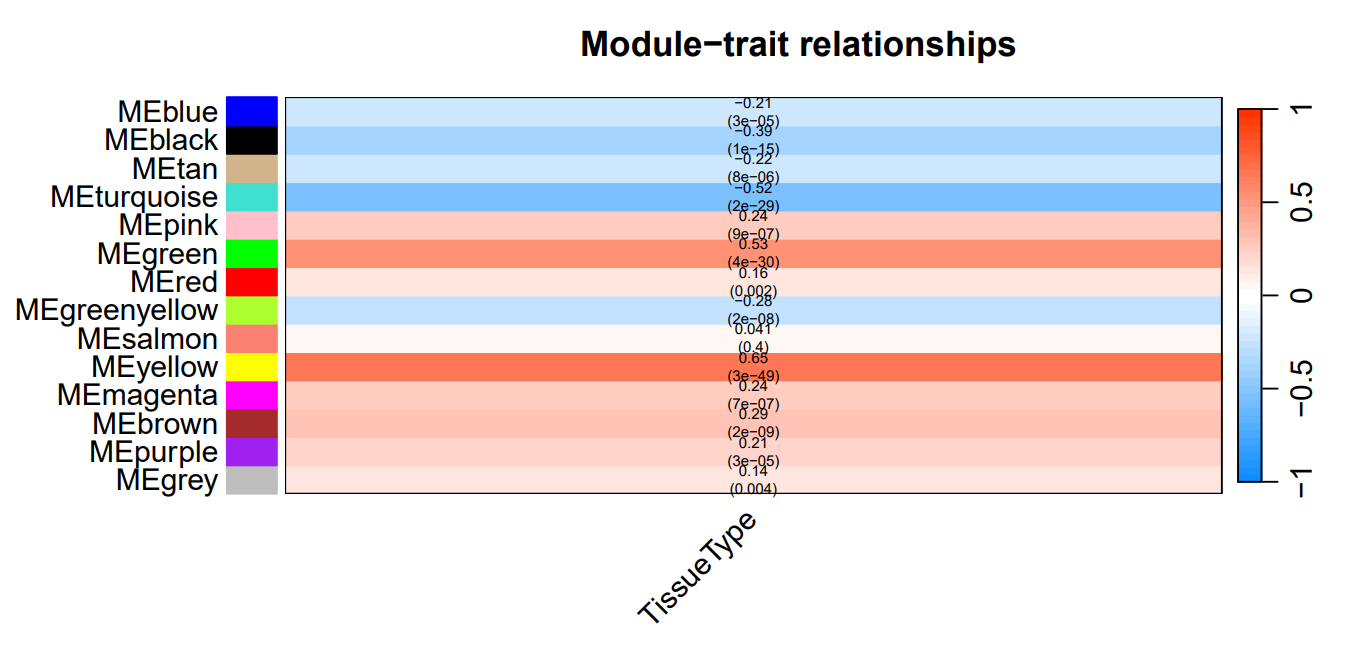


Figure 2. Module-trait Relationship

This module-trait heatmap represents the association between identified gene modules and Tissue types. It shows an understanding of how groups of genes correlate with phenotypic traits. It helps find gene significance involved in the regulation of traits. A positive correlation is in blue indicating a strong association.

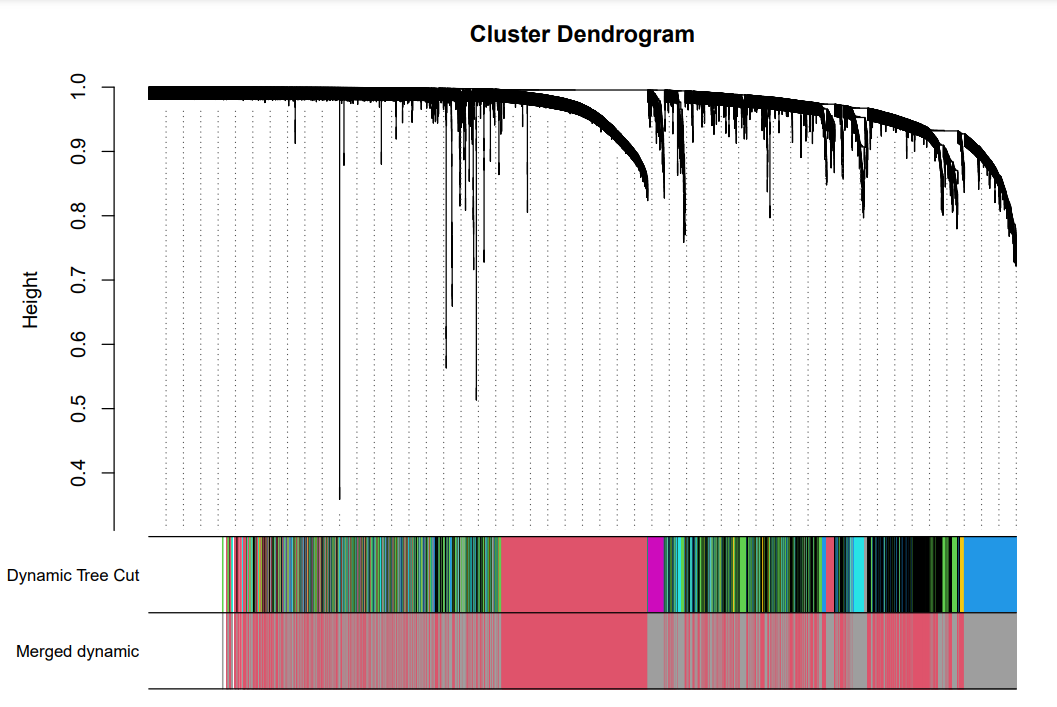


Figure 3. Cluster Dendrogram

The Dendrogram represents clustering of genes based on their expressed pattern which shows similarities or dissimilarities. The branching pattern illustrates how individual genes or smaller clusters combine into large clusters.

Dynamic Tree Cut in dendrogram identifies distinct modules which are clusters of highly correlated genes and shows different modules with genes in the same module showing similar expression profiles.

Merged Dynamic indicates results after merging some modules to further refine module distinction.

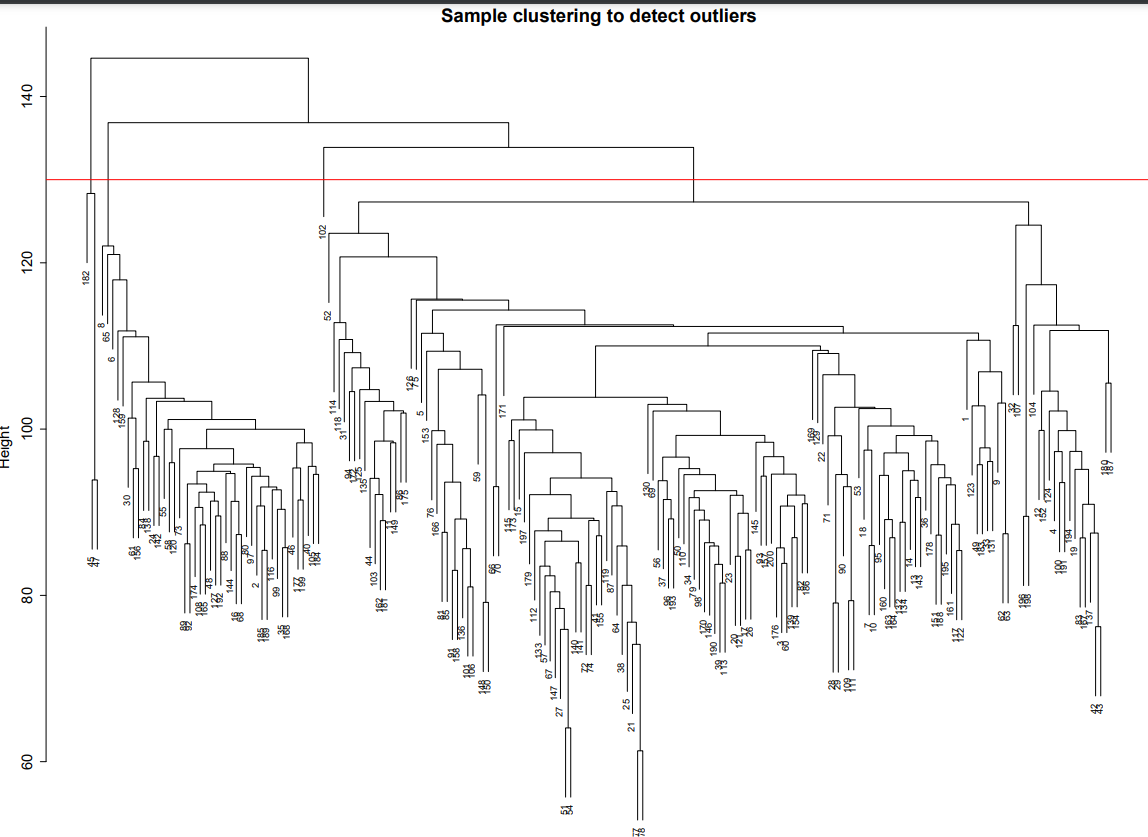


Figure 4. Sample Clustering to Detect Outliers

It is created to detect outliers among samples and it can be due to any reason like experimental errors. This clustering provides insights into relationships and similarities in samples so researchers can identify groups of samples that behave similarly under common traits. This is the branching structure of how samples are merged.

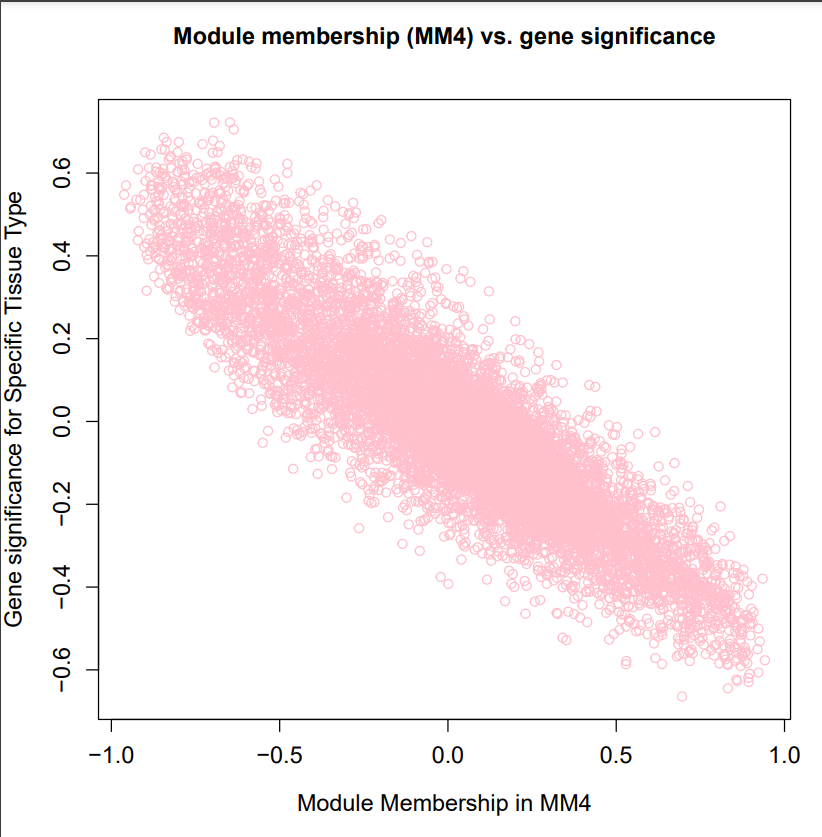


Figure 5. Module Membership vs Gene Significance

Module membership shows genes in specific modules. It quantifies how closely gene expression data is correlated with eigengenes.

Gene significance measures the association between gene expression and Tissue traits. A higher value indicates a stronger association with the trait.

The trend shows genes within modules have higher gene significance, suggesting these genes are not only tightly connected within modules but also significantly associated with traits.

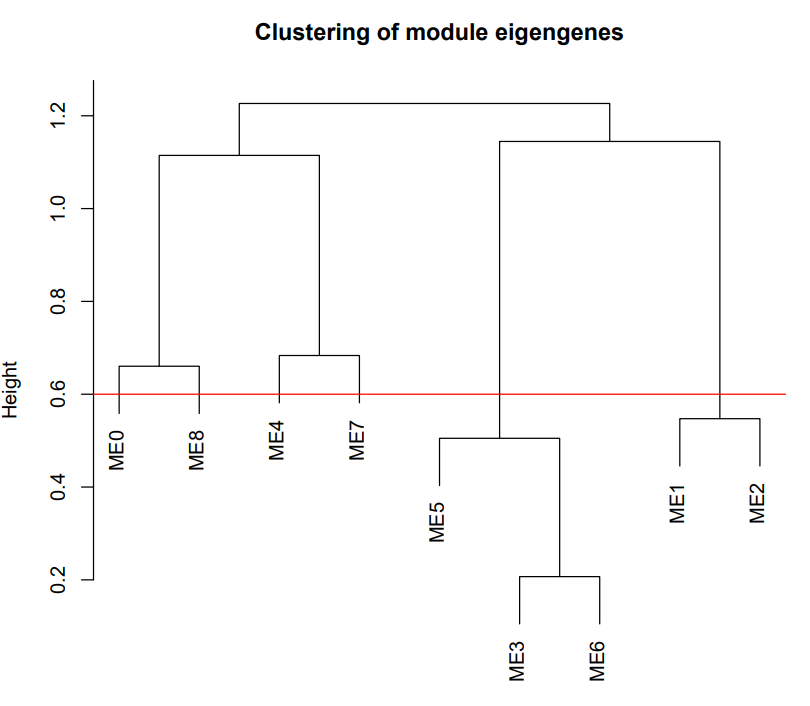


Figure 6. Clustering of Module Eigengenes

Eigengenes are used to represent summaries of gene clustering expression patterns. Each label like ME0 represents an eigengenes. It shows similarities between expression profiles. It shows how modules are grouped based on similarities of eigengenes. Thresholds for clustering determine which modules are similar to be considered to be merged in analysis.

# 

# 

# **Chapter 7. Conclusions and Future Work**

## **Future Work**

### ***Research Papers Contribution***

### We anticipate producing research papers that will contribute to the research and computational society. Our work will shed light on important causes and genes that play vital roles in neurodegenerative disease development."

### ***Advancement in Genetics***

### Identification of key genes associated with neurodegenerative diseases contributes significantly to the field of genetics. This knowledge enhances our understanding of the genetic basis of these diseases.

### ***Impact on Welfare and Society***

### Early Diagnosis Benefits: Our research aims to have a positive impact on welfare and society. Early diagnosis resulting from our work can facilitate timely interventions, as delaying treatment can significantly impact an individual's quality of life.

## **Conclusion**

In conclusion, our research is to shed light on the genetics of neurodegenerative diseases, particularly Alzheimer's disease. Through meticulous analysis of gene expression data from the GSE118553 dataset and application of advanced computing and machine learning techniques, we have embarked on a journey to uncover key genes and pathways in disease pathogenesis. By leveraging the WGCNA and Machine learning algorithms we have identified gene modules and potential biomarkers associated with neurodegenerative diseases. Our findings hold promise for advancing our understanding of these complex disorders and ultimately leading to the development of targeted therapies to improve patient outcomes.

# 

# **References**

*[1] Y. Lagisetty et al., “Identification of risk genes for Alzheimer’s disease by gene embedding,” Cell Genomics, vol. 2, no. 9, p. 100162, Sep. 2022, doi: 10.1016/j.xgen.2022.100162.*

*[2] J. J. Palop, J. Chin, and L. Mucke, “A network dysfunction perspective on neurodegenerative diseases,” Nature, vol. 443, no. 7113, pp. 768–773, Oct. 2006, doi: 10.1038/nature05289.*

*[3] G. G. Kovacs, “Concepts and classification of neurodegenerative diseases,” in Handbook of Clinical Neurology, 2018, pp. 301–307. doi: 10.1016/b978-0-12-802395-2.00021-3.*

*[4] M. A. Myszczynska et al., “Applications of machine learning to diagnosis and treatment of neurodegenerative diseases,” Nature Reviews Neurology, vol. 16, no. 8, pp. 440–456, Jul. 2020, doi: 10.1038/s41582-020-0377-8.*

*[5] W. Mandemakers, V. A. Morais, and B. De Strooper, “A cell biological perspective on mitochondrial dysfunction in Parkinson disease and other neurodegenerative diseases,” Journal of Cell Science, vol. 120, no. 10, pp. 1707–1716, May 2007, doi: 10.1242/jcs.03443.*

*[6] S. B. Prusiner, “Neurodegenerative diseases and prions,” The New England Journal of Medicine, vol. 344, no. 20, pp. 1516–1526, May 2001, doi: 10.1056/nejm200105173442006.*

*[7] D. P. Purohit, D. P. Perl, V. Haroutunian, P. Powchik, M. Davidson, and K. L. Davis, “Alzheimer disease and related neurodegenerative diseases in elderly patients with schizophrenia,” Archives of General Psychiatry, vol. 55, no. 3, p. 205, Mar. 1998, doi: 10.1001/archpsyc.55.3.205.*

*[8] J.-C. Lambert et al., “Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer’s disease,” Nature Genetics, vol. 45, no. 12, pp. 1452–1458, Oct. 2013, doi: 10.1038/ng.2802.*

*[9] G. Jun et al., “A novel Alzheimer disease locus located near the gene encoding tau protein,” Molecular Psychiatry, vol. 21, no. 1, pp. 108–117, Mar. 2015, doi: 10.1038/mp.2015.23.*

*[10] C. M. Karch, C. Cruchaga, and A. M. Goate, “Alzheimer’s Disease Genetics: From the bench to the clinic,” Neuron, vol. 83, no. 1, pp. 11–26, Jul. 2014, doi: 10.1016/j.neuron.2014.05.041.*

*[11] J. Hardy and G. A. Higgins, “Alzheimer’s Disease: The Amyloid Cascade hypothesis,” Science, vol. 256, no. 5054, pp. 184–185, Apr. 1992, doi: 10.1126/science.1566067.*

*[12] V. S. Marde et al., “Neurodegenerative disorders associated with genes of mitochondria,” Future Journal of Pharmaceutical Sciences, vol. 7, no. 1, Mar. 2021, doi: 10.1186/s43094-021-00215-5.*

*[13] V. Escott-Price et al., “Common polygenic variation enhances risk prediction for Alzheimer’s disease,” Brain, vol. 138, no. 12, pp. 3673–3684, Oct. 2015, doi: 10.1093/brain/awv268.*

*[14] Guan et al. "Weighted gene coexpression network analysis and machine learning reveal oncogenome associated microbiome plays an important role in tumor immunity and prognosis in pan-cancer." Journal of Translational Medicine 21.537 (2023): 1-21.*

*[15] Hu, Y., Yu, X., Zhou, Y., Yin, X., Hu, J., Lu, X., and Hu, C. (2020). Identification of novel genes related to Alzheimer's disease pathology using genome-wide co-expression network analysis and supervised machine learning. Frontiers in Aging Neuroscience, 12, 605961. doi: 10.3389/fnagi.2020.605961*

*[16] S. Balne and A. Elumalai, "Machine learning and deep learning algorithms used for the diagnosis of Alzheimer's" in Materials Today: Proceedings, vol. 47, no. 15, pp. 5151-5156, 2021, doi: 10.1016/j.matpr.2021.05.499.*

*[17] P. Kishore, Ch. Usha Kumari, M.N.V.S.S. Kumar, and T. Pavani, "Detection and analysis of Alzheimer’s disease using various machine learning algorithms," in Materials Today: Proceedings, vol. 45, no. 2, pp. 1502-1508, 2021, doi: 10.1016/j.matpr.2020.07.645.*

*[18] A. Amrutesh, G. B. C G, A. A, A. R. KP and G. S, "Alzheimer's Disease Prediction using Machine Learning and Transfer Learning Models," 2022 6th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Bangalore, India, 2022, pp. 1-6, doi: 10.1109/CSITSS57437.2022.10026365.*

*[19] Diogo, V.S., Ferreira, H.A., Prata, D. et al. Early diagnosis of Alzheimer’s disease using machine learning: a multi-diagnostic, generalizable approach. Alz Res Therapy 14, 107 (2022).* [*https://doi.org/10.1186/s13195-022-01047-y*](https://doi.org/10.1186/s13195-022-01047-y)

*[20] S. Harika, T. Yamini, T. Nagasaikamesh, S. H. Basha, S. Santosh Kumar, and Mrs. S. Sri DurgaKameswari, "Alzheimer's Disease Detection Using Different Machine Learning Algorithms," \*International Journal of Research in Advent Technology (IJRASET)\*, DOI: 10.22214/ijraset.2022.46937.*

*[21] H. Sun, J. Yang, X. Li, Y. Lyu, Z. Xu, H. He, X. Tong, T. Ji, S. Ding, C. Zhou, P. Han, J. Zheng, "Identification of feature genes and pathways for Alzheimer's disease via WGCNA and LASSO regression," Frontiers in Computational Neuroscience, vol. 16, p. 1001546, Sep. 2022, doi: 10.3389/fncom.2022.1001546.*

*[22] Langfelder, P., Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008),* [*https://doi.org/10.1186/1471-2105-9-559*](https://doi.org/10.1186/1471-2105-9-559)

*[23] Hong Yue, Bo Yang, Fang Yang, Xiao-Li Hu, Fan-Bin Kong, "Co-expression network-based analysis of hippocampal expression data associated with Alzheimer's disease using a novel algorithm," \*Experimental and Therapeutic Medicine\*, vol. 11, no. 5, pp. 1707-1715, 2016.*

*[24] B. M. Tijms, A. M. Wink, W. de Haan, W. M. van der Flier, C. J. Stam, P. Scheltens, and F. Barkhof, "Alzheimer's disease: connecting findings from graph theoretical studies of brain networks," Neurobiology of Aging, vol. 34, no. 8, pp. 2023-2036, Aug. 2013.* [*https://doi.org/10.1016/j.neurobiolaging.2013.02.020*](https://doi.org/10.1016/j.neurobiolaging.2013.02.020)

*[25] Brier, M. R., Thomas, J. B., Fagan, A. M., Hassenstab, J., Holtzman, D. M., Benzinger, T. L., Morris, J. C., and Ances, B. M. (2014). Functional connectivity and graph theory in preclinical Alzheimer's disease. \*Neurobiology of Aging\*, 35(4), 757-768. [*[*https://doi.org/10.1016/j.neurobiolaging.2013.10.081](https://doi.org/10.1016/j.neurobiolaging.2013.10.081)*](https://doi.org/10.1016/j.neurobiolaging.2013.10.081%5D(https://doi.org/10.1016/j.neurobiolaging.2013.10.081))

*[26]Q. Jing, H. Zhang, X. Sun, Y. Xu, S. Cao, Y. Fang, X. Zhao, and C. Li, "A Comprehensive Analysis*

*Identified Hub Genes and Associated Drugs in Alzheimer’s Disease," vol. 2021, Article ID 8893553,*

*2021,* [*https://doi.org/10.1155/2021/8893553*](https://doi.org/10.1155/2021/8893553)*.*

*[27]C. Zou, L. Su, M. Pan, L. Chen, H. Li, C. Zou, J. Xie, X. Huang, and M. Lu, "Exploration of novel*

*biomarkers in Alzheimer’s disease based on four diagnostic models," Front. Aging Neurosci., vol. 15, p.*

*1079433, Feb. 16, 2023. [Online]. Available:* [*https://doi.org/10.3389/fnagi.2023.1079433*](https://doi.org/10.3389/fnagi.2023.1079433)

*[28]Zhang, Y., Kiryu, H. Identification of oxidative stress-related genes differentially expressed in*

*Alzheimer’s disease and construction of a hub gene-based diagnostic model. Sci Rep 13, 6817 (2023).*

[*https://doi.org/10.1038/s41598-023-34021-1*](https://doi.org/10.1038/s41598-023-34021-1)

*[29]A. Sharma and P. Dey, "A machine learning approach to unmask novel gene signatures and*

*prediction of Alzheimer's disease within different brain regions," Genomics, vol. 113, no. 4, pp.*

*1778-1789, 2021, ISSN 0888-7543,* [*https://doi.org/10.1016/j.ygeno.2021.04.028*](https://doi.org/10.1016/j.ygeno.2021.04.028)*.*

*[30]Y. Chen, Z. Li, X. Ge, H. Lv, and Z. Geng, "Identification of novel hub genes for Alzheimer’s disease associated with the hippocampus using WGCNA and differential gene analysis," Frontiers in Neuroscience, vol. 18, no. 1359631, Mar. 2024.*

*[31]H. Sun, J. Yang, X. Li, Y. Lyu, Z. Xu, H. He, X. Tong, T. Ji, S. Ding, C. Zhou, and P. Han, "Identification of feature genes and pathways for Alzheimer's disease via WGCNA and LASSO regression," Frontiers in Computational Neuroscience, vol. 16, p. 1001546, 2022,* [*https://doi.org/10.3389/fncom.2022.1001546*](https://doi.org/10.3389/fncom.2022.1001546)

*[32]J. W. Liang, Z. Y. Fang, Y. Huang, Z. Y. Liuyang, X. L. Zhang, J. L. Wang, H. Wei, J. Z. Wang, X. C. Wang, J. Zeng, and R. Liu, "Application of weighted gene co-expression network analysis to explore the key genes in Alzheimer’s disease," Journal of Alzheimer's Disease, vol. 65, no. 4, pp. 1353-1364, 2018, DOI: 10.3233/JAD-180400.*

*[33]H. Wang, X. Han, and S. Gao, "Identification of potential biomarkers for pathogenesis of Alzheimer’s disease," Hereditas, vol. 158, pp. 1-8, 2021.*

*[34]J. Li, Y. Zhang, T. Lu, R. Liang, Z. Wu, M. Liu, L. Qin, H. Chen, X. Yan, S. Deng, and J. Zheng, "Identification of diagnostic genes for both Alzheimer’s disease and Metabolic syndrome by the machine learning algorithm," Frontiers in Immunology, vol. 13, p. 1037318, 2022,* [*https://doi.org/10.3389/fimmu.2022.1037318*](https://doi.org/10.3389/fimmu.2022.1037318)

*[35]J. Ren, B. Zhang, D. Wei, and Z. Zhang, "Identification of methylated gene biomarkers in patients with Alzheimer’s disease based on machine learning," BioMed Research International, 2020,* [*https://doi.org/10.1155/2020/8348147*](https://doi.org/10.1155/2020/8348147)

*[36]B. Monk, A. Rajkovic, S. Petrus, A. Rajkovic, T. Gaasterland, and R. Malinow, "A machine learning method to identify genetic variants potentially associated with Alzheimer’s disease," Frontiers in Genetics, vol. 12, p. 647436, 2021,* [*https://doi.org/10.3389/fgene.2021.647436*](https://doi.org/10.3389/fgene.2021.647436)

*[37]H. Chen, Y. He, J. Ji, and Y. Shi, "A machine learning method for identifying critical interactions between gene pairs in Alzheimer's disease prediction," Frontiers in Neurology, vol. 10, p. 490162, 2019,* [*https://doi.org/10.3389/fneur.2019.01162*](https://doi.org/10.3389/fneur.2019.01162)

*[38]H. Alamro, M. A. Thafar, S. Albaradei, T. Gojobori, M. Essack, and X. Gao, "Exploiting machine learning models to identify novel Alzheimer’s disease biomarkers and potential targets," Scientific Reports, vol. 13, no. 1, p. 4979, 2023.*

*[39]Q. Jing, H. Zhang, X. Sun, Y. Xu, S. Cao, Y. Fang, X. Zhao, and C. Li, "A comprehensive analysis identified hub genes and associated drugs in Alzheimer’s disease," BioMed Research International, 2021,* [*https://doi.org/10.1155/2021/8893553*](https://doi.org/10.1155/2021/8893553)

*[40]A. Sharma and P. Dey, "A machine learning approach to unmask novel gene signatures and prediction of Alzheimer's disease within different brain regions," Genomics, vol. 113, no. 4, pp. 1778-1789, 2021,* [*https://doi.org/10.1016/j.ygeno.2021.04.028*](https://doi.org/10.1016/j.ygeno.2021.04.028)

# **Github Repository Link**

<https://github.com/Robaisha/IdentificationOfRiskGenesForNeurodegenerativeDiseases-FinalYearProject.git>