**Early Detection of Alzheimer’s using Blood Gene Expression Data**

J. Hariharan, R. Jothi

*School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India*

**Abstract:**

**Alzheimer’s Disease (AD) a type of neurological disorder is seeing a surge in the numbers of cases from the past decade. This paper focuses to explore the various feature extraction and classifier algorithms that can make use of blood gene expression data and to design a system using the analysis for early detection of AD. We also aim to explore explainable artificial intelligence methods (XAI) of classification for a simple human interpretation and measure its trustworthiness.**

***Keywords: Blood Gene Expression, Feature Extraction, Explainable Artificial Intelligence***

I. INTRODUCTION:

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by the gradual loss of cognitive function and memory. It is the most common cause of dementia among older adults. Especially in India the cases are expected to grow to 11,422,692 by 2050 from 3,848,118 measured in 2019 according to Lancet report as of July 2022 [1]. The disease is caused by the accumulation of amyloid plaques and tau tangles in the brain, leading to the death of nerve cells and the disruption of communication between brain cells. As the disease progresses, individuals may have trouble with everyday tasks, behavioral changes, and eventually, complete dependence on caregivers. Despite intense research efforts, there is currently no cure for AD and available treatments only offer temporary symptom relief. Early detection and diagnosis of AD is crucial for the planning of appropriate care and support for individuals and their families, as well as for the development of disease-modifying therapies. However, current diagnostic methods for AD often involve invasive and expensive procedures, such as brain imaging or lumbar punctures. In recent years, there has been increasing interest in the use of blood-based biomarkers, such as gene expression patterns, as a less invasive and more cost-effective approach for the early detection of AD. The identification of specific gene expression patterns in the blood that are associated with AD may enable the development of simple and reliable diagnostic tools that can be used in a clinical setting.

Gene expression refers to the process by which the genetic information contained in DNA is used to synthesize the various proteins and other molecules that perform specific functions within cells. This process is regulated by a complex network of signaling pathways that control which genes are turned on or off in each cell at a given time. The measurement of gene expression, or transcriptomics, allows scientists to understand how cells respond to different stimuli and how they differ from one another. By analyzing gene expression data, researchers can gain insights into the underlying mechanisms of biological processes and diseases, such as cancer or Alzheimer's disease. Gene expression data can be obtained from a variety of sources, including tissues, cells, and biofluids such as blood. The use of blood-based gene expression data has the advantage of being non-invasive and easily accessible, making it a promising tool for the diagnosis and monitoring of diseases. In recent years, there has been growing interest in the use of gene expression data for the early detection and treatment of a wide range of conditions, including cancer, cardiovascular disease, and neurological disorders. The major problem that we must address while use blood gene expression data is the High Dimensionality of the dataset, since blood tissue can be used to extract around 10,000-30,000 genes on average and each of these genes might have 1-3 gene probes. DNA probes are usually single-stranded DNA molecules that are labeled with a detectable molecule, such as a fluorescent dye or a radioactive isotope. They are designed to bind to a complementary DNA sequence and are often used to detect the presence of specific genes or to analyze DNA modifications, such as methylation. RNA probes are like DNA probes, but they are designed to bind to complementary RNA sequences. They are often used to detect the presence and abundance of specific RNA molecules, such as mRNA or non-coding RNA. Protein probes are molecules that are designed to specifically bind to and detect the presence of a particular protein. They can be antibodies, small molecules, or other types of protein-binding molecules and are often used to analyze protein expression, localization, and function.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Study** | **Dataset Used** | **Feature Selection Alg.** | **No. of Genes** | **Classification Alg.** | **Performance** |
| Lee, T et al. | ANM1 | DEG (using SAM) | 697 | SVM | AUC: 87.4% |
| M. S. Kamal et al. | GSE174367 | - | 18,234 | KNN, SVM | ACC: 64.5%  ACC: 82.4% |
| Mahendran, N et al. | GSE76105 | Adaboost | 12 | DRNN | ACC: 89.4% |
| S. Khanal et al. | ADNI | t-test + SelectFromFeature | 25 | XGBoost | ACC: 65%  AUC: 67% |
| El-Gawady, A et al. | GSE33000  GSE44770  GSE44768  GSE44771 | χ2, ANOVA, MI | 30 | SVM | ACC: 97.5%  AUC: 97.2% |
| S. Perera et al. | GSE5281 | PCA, RF, ETC | 14 | SVM | ACC: 93.9% |
| **Table 1:** Summary of studies conducted on detection of AD using various feature selection and classification techniques. | | | | | |

II. METHODS:

A. Preprocessing of Dataset:

GSE63060 and GSE63061 gene expression samples were collected from GEO gene expression omnibus repository as SOFT formatted family files. GSE63060 contained samples collected from 329 individuals out of which 145 were AD samples, 104 were healthy samples (CTL), and 80 were samples collected from people with Mild Cognitive Impairment (MCI). Each samples contained expression values of 38323 probes, which were mapped to their respective gene symbols using python GEOParse annotation package. If a gene had multiple probe values, Median of the values are taken as the expression value for the gene based on the study done by Lee, T et al. [2]. A total of 29958 unique gene expression values where this extracted and combined with other attributes such as Age, Ethnicity, Gender. Gene expression value where then normalized using Min-Max normalization method and MCI sample were ignored to avoid noise in the dataset.

B. Feature Selection Technique:

1. Chi-square (χ2): χ2 is a statistical method used to determine where there is a statistically significant relation between observed frequency and expected frequency of a particular event. If the difference between the observed and expected values are differ by a large value, then we can reject the null hypothesis and state that the variables are related. El-Gawady, A et al. (2022) [6]: have done the study on using χ2 to extract the top 30 genes. This was employed along with 2 other statistical methods (ANOVA, MI) on a group 8 gene subsets created by integrating 4 genes expression datasets (GSE33000, GSE44770, GSE44768, GSE44771) extracted from different regions of the brain. The average of these metrics was used to order the genes and select the top 30 genes. Classification done using these 30 genes yielded a maximum ACC of 97.5% and AUC of 97.2%.
2. Analysis of Variance (ANOVA): ANOVA is a statistical method used to test if two groups of variables related by checking if there is a statistical difference between the mean s of the groups. This is a powerful tool which is now widely used in many fields including biology and psychology. It can be used with gene expression dataset to extracts genes having interested properties. El-Gawady, A et al. (2022) [6]: have done their research using ANOVA to extract the top 30 genes. This was employed along with 2 other statistical methods (χ2, MI) on a group 8 gene subsets created by integrating 4 genes expression datasets (GSE33000, GSE44770, GSE44768, GSE44771) extracted from different regions of the brain. The average of these metrics was used to order the genes and select the top 30 genes. Classification done using these 30 genes yielded a maximum ACC of 97.5% and AUC of 97.2%.

C. Classification Technique:

1. Support Vector Machine (SVM): SVM is the most popular and widely used linear classifier technique. This involves classification based on supervised learning approach. Lee, T et al. (2020) [2]: The authors conducted the study on SVM along with various classification models like LR, L1 regularized LR (L1-LR), Deep Neural Network (DNN), RF. SVM was observed to be well paired with SAM feature selection algorithm. AUC of 0.874 was observed in the ANM1 dataset which was significantly higher when SVM was used with other feature extraction algorithms like VAE as authors found VAE lost critical information’s while reducing the dimensions. Kamal et al. (2021) [3]: This study also used SVM for classification for their multi-model diagnostic system and found it to outperform k-nearest and Xboost techniques. Accuracy of 82.4% with Precession of 81.8% was observed indicating the advantage of using it against a high dimensional dataset. El-Gawady, A et al. (2022) [6]: have conducted study by using SVM for classification of AD over the 30 genes extracted using the statistical methods (χ2, ANOVA, MI). SVM was used along with other classification methods like RF, LR, AdaBoost and SVM was found to outperform these techniques. Maximum ACC of 97.5% and AUC of 97.2% for the set having pairwise intersection of 4 datasets (GSE33000, GSE44770, GSE44768, GSE44771) was observed for the study.

RESULTS AND DISCUSSION:

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature Extraction** | **Classification** | **Accuracy (Training Set)** | **Accuracy (Testing Set)** |
| Chi Square | SVM Linear Kernel | 99.67% | 77% |
| Chi Square | SVM Gaussian Kernel | 98.85 | 84% |
| Chi Square | DNN (64, 128, 128, 2) | 97.13% | 77.33% |
| Chi Square | DNN (6, 4, 4, 2) | 93.68% | 89.33% |
| **Table 2:** Performance evaluation of the proposed method | | | |

|  |  |
| --- | --- |
|  |  |
| **Fig 1:** Loss and Accuracy variation chart over 100 epochs in Training and Testing set. | |

CONCLUSION:

REFERENCES:

[1] Nandi, A et al, “Global and regional projections of the economic burden of Alzheimer's disease and related dementias from 2019 to 2050: A value of statistical life approach”, EClinicalMedicine - The Lancet Discovery Science, Volume 51, 101580, 2022. https://doi.org/10.1016/j.eclinm.2022.101580.

[2] Lee, T., Lee, H. Prediction of Alzheimer’s disease using blood gene expression data. Sci Rep 10, 3485, 2020. https://doi.org/10.1038/s41598-020-60595-1.

[3] M. S. Kamal, A. Northcote, L. Chowdhury, N. Dey, R. G. Crespo and E. Herrera-Viedma, "Alzheimer’s Patient Analysis Using Image and Gene Expression Data and Explainable-AI to Present Associated Genes," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-7, 2021, Art no. 2513107, doi: 10.1109/TIM.2021.3107056.

[4] Mahendran N, P M DRV. A deep learning framework with an embedded-based feature selection approach for the early detection of the Alzheimer's disease. Comput Biol Med. 2022 Feb; 141:105056. doi: 10.1016/j.compbiomed.2021.105056. Epub 2021 Nov 22. PMID: 34839903.

[5] S. Khanal, J. Chen, N. Jacobs, and A. -L. Lin, "Alzheimer’s Disease Classification Using Genetic Data," 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021, pp. 2245-2252, doi: 10.1109/BIBM52615.2021.9669730.

[6] El-Gawady, A.; Makhlouf, M.A.; Tawfik, B.S.; Nassar, H. Machine Learning Framework for the Prediction of Alzheimer’s Disease Using Gene Expression Data Based on Efficient Gene Selection. Symmetry 2022, 14, 491. https://doi.org/10.3390/sym14030491.

[7] S. Perera, K. Hewage, C. Gunarathne, R. Navarathna, D. Herath, and R. G. Ragel, "Detection of Novel Biomarker Genes of Alzheimer’s Disease Using Gene Expression Data," 2020 Moratuwa Engineering Research Conference (MERCon), 2020, pp. 1-6, doi: 10.1109/MERCon50084.2020.9185336.