**Blood Gene Expression as a Biomarker for Alzheimer's Disease Detection: A Review**

J. Hariharan, R. Jothi

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

**Abstract:**

**Alzheimer’s a type of neurological disorder is seeing a surge in the numbers of cases from the past decade. This paper focuses to explore and summarize the various feature extraction and classifier algorithms that can make use of blood gene expression data and find the advantages of using blood gene expression data over other sources like MRI images and gene expression extracted from other tissues of body. 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***

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.

FEATURE SELECTION TECHINIQUES:

High dimensionality poses a need for a feature selection mechanism since the number of dimensions in a gene expression dataset would be close to 40,000. Feature extraction techniques will allow us to design our classification models only based on the features that have a significant impact on exhibiting characteristics that can lead to development of AD.

Significance Analysis on Microarrays (SAM):

SAM is a statistical method used to identify genes or other features in a microarray dataset that are significantly different between two or more experimental conditions. Microarrays are a type of high-throughput technology used to measure the expression levels of thousands of genes or other biological features simultaneously. Genes with a low significance score are significantly differentially expressed, while genes with a high significance score are not considered to be significantly differentially expressed. Lee, T et al. (2020) [2]: The authors have published a study on extracting differentially expressed genes (DEG) using SAM by taking use of three publicly available datasets ADNI, ANM1, ANM2. This involved comparing SAM with other state of the art feature selection algorithms like Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest (RF). Before putting the data under the feature selection algorithms, the dataset was first filtered out to remove genes with expression value less than the median of the gene expression values in 100 samples and if a gene had multiple probes median of their values where used. This gave the authors around 21,698 unique genes to work with for feature selection. Among the various feature selection algorithm experimented Significance Analysis of Microarrays yielded the best results. Variable Auto Encoders were also used in this study but found to be not helpful in improving the performance when used along with the DEG found using SAM. Area Under the Curve (AUC) of 0.874 for ANM1 dataset was observed by the authors using ANM1 dataset.

Adaboost:

Adaptive Boosting technique is an Ensemble method using decision trees. These uses set of weak learners for fast implementation and to converge faster into the results and doesn’t require any prior domain knowledge in creating the weak learners. The goal of the weak leaners is to identify the weak hypothesis. Mahendran N et al. (2022) [4]: have conducted their study on Adaboost to find 12 differentially methylated that are most significant in identifying the AD. Out of these 6 were hypo-methylated and the other 6 were hyper-methylated. These 12 identified genes were k-fold validated using Logistic Regression (LR) model and a validation accuracy of 87.1% was observed when k = 3. The genes associated with these positions were MS4A4, MYNN, TXNIP, CORO2B, NOG, BEX2, PIGA, FAM82A1 and CDKN1C.

CLASSIFICATION TECHINIQUES:

With the extracted genes classification techniques are used to distinguish between healthy control samples (CN) from samples collected from people with AD or mild cognitive impairment (MCI). The choice of feature extraction algorithms was seen to affect the ability of the classification model due to the high dimensionality problem.

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 Logistic Regression (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.

K-Nearest Neighbors (KNN):

Kamal et al. (2021) [3]: The authors have conducted a study on KNN by designing a multi-model diagnostic system using a combination of classifiers that use MRI images as well as gene expression data. The study is focused on reducing the computational time as previous research were focused on the use of complex DNN and feature selection techniques. Using GSE174367 dataset classifier techniques such as KNN, SVM, Xboost are explored. KNN a classifier which determines the class level based on the Euclidean distance was found to not perform well due to the input having very large dimension.

Deep Recurrent Neural Network (DRNN):

This is a special kind of muti-layer neural network in which output of one layer is again fed as input to the layer. This acts as a memory which carries necessary information about the previous states. This is different from the usual neural network as in typical network the input and output neurons are completely independent of each other. Mahendran N et al. (2022) [4]: have conducted study on DRNN and have proposed an enhanced version called Enhanced DRNN (EDRNN). The selected genes from the Ada Boost feature selection technique are provided as input to these models in their study. EDRNN was implement with early stopping criteria to avoid overfitting and had considerable depth in the hidden layers of the output area for better accuracy. This model in their study was compared against state-of-the-art approaches like RNN and CNN. The accuracy of 89.4% showed it outperformed the existing systems.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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% |

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.