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

J. Hariharan
School of Computer Science and Engineering,
Vellore Institute of Technology,
Chennai, India

R. Jothi
School of Computer Science and Engineering,
Vellore Institute of Technology,
Chennai, India

***Abstract*--Alzheimer's disease, a type of neurological condition, has seen an increase in the number of cases over the last decade, necessitating the development of a comprehensive method for early detection. Existing techniques are mostly invasive and expensive, so our research focuses on blood gene expression as a potential biomarker. The main challenge in analyzing blood gene expression data is the high dimensionality of the gene expression data. Consequently, this study investigates and summarizes the numerous feature selection and classifier techniques that can utilize blood gene expression data, as well as identifies the advantages of using blood gene expression data over other sources, such as MRI images and gene expression collected from other organs.**

Keywords---Blood Gene Expression, Feature Selection

I. INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disorder marked by the gradual deterioration of cognitive function and memory. Alzheimer's disease is the leading cause of dementia in older adults. Specifically in India, the number of cases is anticipated to increase to 11,422,692 by 2050 from 3,848,118 in 2019, according to a paper published by Lancet in July 2022 [1]. The accumulation of amyloid plaques and tau tangles in the brain leads to the disease, which results in the loss of nerve cells and the breakdown of communication between brain cells. As the condition develops, patients may have trouble with daily activities, behavioral changes, and, eventually, total dependence on care takers. There is presently no cure for AD and available therapies only provide momentary symptom relief, despite significant research efforts. Early detection and diagnosis of AD are essential for the planning of appropriate treatment and support for individuals and their families as well as the development of disease-modifying medicines. However, existing approaches for diagnosing AD frequently involve invasive and costly procedures, such as brain imaging or lumbar punctures. In recent years, there has been growing interest in the use of blood-based biomarkers, such as gene expression patterns, as a less invasive and more cost-effective method for the early identification of AD.

Gene expression refers to the process through which the genetic information stored in DNA is utilized to generate proteins and other molecules with specified functions within cells. This process is controlled by a complicated network of signaling pathways that determine which genes in each cell are active at any given time. Transcriptomics, the measurement of gene expression, enables researchers to better comprehend how cells respond to various stimuli and how they vary from one another. By studying gene expression data, scientists can get insight into the underlying mechanisms of biological processes and disorders like cancer and Alzheimer's. Data on gene expression can be extracted from numerous sources, including tissues, cells, and biofluids like blood. Blood-based gene expression data collection is a potential method for disease diagnosis and monitoring because it is non-invasive and convenient. Gene expression data is increasingly being used to detect and treat cancer, cardiovascular illness, and neurological problems. Blood tissue may typically be utilized to extract between 10,000 and 30,000 genes, and each of these genes may contain between one and three gene probes. As a result, the High Dimensionality of the dataset poses the greatest challenge when analyzing blood gene expression information. The purpose of this research is to investigate the potential of using gene expression as a biomarker by investigating the use of feature selection approaches to address the high dimensionality problem and then investigating the use of classification techniques to classify AD samples.

GENE	ID_REF	VALUE	Totals
A1BG	ILMN_2055271	7.600951059	0.00
A1CF	ILMN_1779670	7.439317571	0.00
	ILMN_1806310	7.650042454	0.00
	ILMN_2383229	7.401990858	0.00
A2BP1	ILMN_1731507	7.343481699	0.00
	ILMN_1787689	7.539511288	0.00
	ILMN_2359168	7.508456385	0.00
A2M	ILMN_1745607	7.4335366	0.00
A3GALT2	ILMN_1668111	7.407560121	0.00
A4GALT	ILMN_1735045	7.47683958	0.00
A4GNT	ILMN_1680754	7.499282035	0.00
A26A1	ILMN_1671474	7.397004647	0.00
	ILMN_2321282	7.538709975	0.00
A26B1	ILMN_1772582	7.466492328	0.00
A26C3	ILMN_1653355	7.650277251	0.00
	ILMN_1705025	7.409008953	0.00
	ILMN_1717783	7.370770179	0.00
AAA1	ILMN_1659452	7.543085586	0.00
	ILMN_1767388	7.733080405	0.00
AAAS	ILMN_1755321	7.585403532	0.00
AACS	ILMN_1698554	8.148445821	0.00
AACSL	ILMN_1814092	7.594693783	0.00
AADACL1	ILMN_1676336	7.723275813	0.00
	ILMN_2061446	8.226969726	0.00

Fig 1: Summary of blood gene expression values of top 15 Gene symbols measured against different probe identifiers taken from sample GSM1539080 which was measured as part of dataset GSE63060.

II. LITERATURE REVIEW

Feature Selection Methods: As there are over 40,000 dimensions in a gene expression dataset, high dimensionality necessitates an effective approach for selecting features. Feature selection strategies will allow us to create our classification models based solely on the features that have a substantial impact on presenting characteristics that can contribute to development of AD.

A. Statistical Methods:

a. Significance Analysis on Microarrays (SAM): SAM is a statistical technique used to discover genes or other properties in a microarray dataset that differ substantially between two or more experimental conditions. Microarrays are a form of high-throughput technology used to detect the expression levels of thousands of genes or other biological properties concurrently. Genes with a low significance score are thought to have significantly different expression, whereas genes with a high significance value are not thought to have such a difference. (2020) [2]: Lee, T. et al. The authors have published a work on extracting differentially expressed genes (DEG) using SAM in conjunction with three publicly available datasets: ADNI, ANM1, and ANM2. This involved comparing SAM to various cutting-edge feature selection algorithms such as Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest, and Least Absolute Shrinkage and Selection Operator (LASSO) (RF). Before applying feature selection algorithms to the data, the dataset was filtered to eliminate genes with expression values less than the median of the gene expression values in 100 samples, and if a gene had multiple probes, the median of their values was utilized. This provided the authors with around 21,698 distinct genes to use for feature selection. Among the many experimental feature selection algorithms, Significance Analysis of Microarrays produced the best outcomes. Variable Auto Encoders were also utilized in this investigation, but it was shown that they did not improve performance when combined with the DEG determined by SAM. The authors observed an Area Under the Curve (AUC) value of 0.874.

b. t-test: t-tests are utilized to assess if there are statistically significant differences between the means of two groups. They are commonly used to evaluate the efficacy of a novel medicine or strategy. The t-test compares the magnitude of the mean difference between two groups to the variance within each group. The null hypothesis that there is no statistically significant difference between the means can be rejected if the difference between the means is large relative to the variance within the groups. In this situation, it is probable that the discrepancy is not due to random chance. There are several types of T-tests, including paired, independent, and one-sample. Its use can be expanded to Feature Selection to identify the essential features from a big feature set. If the difference between the means of two groups (feature variable and output variable) is statistically significant, it indicates the significance of the feature in determining the output variable's value. S. Khanal et al. (2021) [5]: did the research using a t-test to compare groups of interest. The groups are ranked according to their p-value values to pick the top N genes. Compared to the other groups, the group containing EMCI performed the best, with an accuracy (ACC) of 65% and an area under the curve (AUC) of 67%. C. Park et al. (2020) [8]: did a study utilizing the Limma software to extract DEG based on the t-test. Because conventional statistical approaches were incapable of reflecting the biological process, they were disregarded. GSE30000 and GSE44770 prefrontal brain gene expression data were merged, and z-scores were standardized prior to t-testing to identify DEG. Since methylation played a significant part in gene DEG regulation, gene expression datasets were combined with DNA methylation dataset GSE80970 to find Differentially Methylated Positions (DMP) utilizing intersection operation performed between the gene expression sources. In k-fold cross validation, the classification of 35 genes retrieved with the t-test yielded an average accuracy of 82.3%.

$$t = \frac{\overline{x_i} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where $\overline{x_i}$ and $\overline{x_2}$ are the means of RNA expression values of a particular gene under consideration and output variable having value AD or CTL (healthy sample) respectively. s_i represents the variance and n_i represents the size of groups respectively.

c. Chi-square (χ^2): χ^2 is a statistical tool used to examine if there is a statistically significant relationship between the observed frequency and predicted frequency of an occurrence. If the observed and predicted values differ by a substantial amount, we can reject the null hypothesis and conclude that the variables are connected. El-Gawady, A et al. (2022) [6]: have conducted the study on extracting the top 30 genes utilizing χ^2. This was used in conjunction with two other statistical methods (ANOVA and MI) on a set of eight gene subsets constructed by combining four gene expression datasets (GSE33000, GSE44770, GSE44768, and GSE44771) retrieved from distinct areas of the brain. The average of these parameters was utilized to rank the genes and determine the top 30. Using these 30 genes for classification provided a maximum ACC of 97.5% and AUC of 97.2%.

$$\chi = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Where O_{ij} is the observed frequency values E_{ij} is the expected values in the ith cell of contingency matrix which is plotted for each gene symbol. For each gene

symbol the contingency matrix has rows r corresponding to unique expression values and columns c corresponding to target output which is AD and CTL.

d. Analysis of Variance (ANOVA): ANOVA is a statistical procedure used to determine whether two groups of variables are related by determining whether there is a statistically significant difference between the groups' means. This is a potent instrument that is frequently employed in numerous domains, including biology and psychology. It can be applied to gene expression datasets to extract genes with desirable features. El-Gawady, A., et al. (2022) [6]: have extracted the top 30 genes using ANOVA in their research. This was used in conjunction with two other statistical approaches (2, MI) on a group of eight gene subsets produced by merging four gene expression datasets (GSE33000, GSE44770, GSE44768, and GSE44771) retrieved from various brain regions. The average of these parameters was utilized to rank the genes and determine the top 30. Using these 30 genes for classification provided a maximum ACC of 97.5% and AUC of 97.2%. H. M. AL-Bermany et al. (2021) [9]: conducted a study utilizing ANOVA for feature selection in addition to other statistical techniques. The study's p-value threshold was set at 0.05, and genes with values less than this were considered statistically significant and added to the gene subset for further processing using clustering algorithms. A maximum accuracy of 92.9% was achieved with ANOVA outperforming all other statistical methods examined in the study.

$$F = \frac{MS_B}{MS_E}$$

$$MS_B = \frac{\sum_{i=1}^{k} n_i(\bar{X}i - \bar{X}..)^2}{k-1}$$

$$MS_E = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_\iota\right)^2}{N-k}$$

Where F represents the ANOVA coefficient, and MS_B is the mean square between the groups and MS_E is the mean square of errors. Here the groups (X) represent the unique gene symbols expression values and target output variable values of the samples. k represents the number of groups and N represents the sample size.

e. Mutual Information (MI): MI is a statistical technique for determining the interdependence of two random variables. It can alternatively be expressed as the "Shared Information" metric between two random variables. Greater the value of MI, the greater the dependence between random variables. This is typically used to express the superior utility of one variable over another. Therefore, this can be used with a gene expression dataset to extract genes with the desired characteristics. El-Gawady, A., et al. (2022) [6]: have extracted the top 30 genes using MI in their research. This was used in conjunction with two other statistical approaches (2, ANOVA) on a collection of eight gene subsets constructed by combining four gene expression datasets (GSE33000, GSE44770, GSE44768, GSE44771) retrieved from distinct areas of the brain. The average of these parameters was utilized to rank the genes and determine the top 30. Using these 30 genes for classification provided a maximum ACC of 97.5% and AUC of 97.2%.

$$I(X;Y) = \sum_x \sum_y p(x,y) log\big(p(x,y)/p(x)p(y)\big)$$

Here X, Y represent the random variables, which in our case would be the observed gene expression values of a gene symbol and value of target output variable having values AD and CTL. To notes is that $p(x,y)$ represents the joint probability. $I(X;Y)$ higher values of this represents relation between X and Y.

f. Principal Component Analysis (PCA): PCA is a statistical method that involves linear transformation of the given dataset into a new system where the sample can be represented with few dimensions. This technique aims to find the set principal components that exhibit maximum variance. Due to this PCA is used widely in genetic studies to find important gene sets from the given set of gene in the sample. S. Perera et al. (2020) [7]: conducted study in which PCA was used to select top 50 gene from 24,438 unique genes of samples from the data set GSE5281. PCA is found to be the most efficient way of feature selection among 2 other ensemble methods used in the study yielding a maximum ACC of 93.9% when 14 gene out of 50 top gene selected was used.

$$cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Where x_i and y_i would be the variables having gene expression values of unique gene symbols and target output variable having values AD and CTL respectively. Based on the $cov(x,y)$ values covariance matrix is plotted. From this matrix eigen values are computed and top K features based on their eigen values is collected.

g. Least Absolute Shrinkage and Selection Operator (LASSO): LASSO is a variation of linear regression model that is used to select important features and shrinking the less important features. The shrinkage factor is controlled by the penalty that is assigned to each of the features. As the value of the penalty increase the value of parameter is shrunken more towards zero. A. Sharma et al. (2021) [10]: conducted study in which LASSO was used along with other Random Forest based feature selection techniques. Data from 4 regions of brain (Prefrontal Cortex, Medial temporal gyrus, Hippocampus, Entorhinal cortex) were integrated and batch normalized using z-

test. In the study R package "caret" was used to implement LASSO. Study results showed standalone LASSO feature selected performed well in classification of AD in 3 brain regions Prefrontal Cortex, Medial temporal gyrus, Entorhinal cortex with ACC of 100%, 99% and 98% respectively. Kalkan H et al. (2022) [11]: proposed a method to LASSO for feature selection from gene expression datasets GSE63060, GSE63061 and GSE140829. A total of 488 unique gene symbols were identified which were converted into a 2D image using Linear Discriminant Analysis (LDA). A maximum ACC of 84.2% was observed when classification was done using this 2D image representation.

$$minimize(1/2n)||y - Xw||_2^2 + \lambda||w||_1$$

λ is the regularization parameter, larger value of λ results in more shrinkage of irrelevant features. X represents independent variable which would be gene expression values of gene symbol and y represents the target output variable. w is vector of coefficients. $|\dots|_1$ represents the L1 norm of a vector, also known as the Manhattan or Taxicab norm and $|\dots|_2$ is the L2 norm of a vector, also known as the Euclidean norm

B. Ensemble Methods:

a. Adaptive Boosting (Adaboost): Adaboost 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.

b. Random Forest (RF): RF is a most popular ensemble technique which is used for task such as classification and regression. From the give sample of dataset random sub samples are generated for each of which a classifier is modelled. The classifiers are then ensembled together to make the final classification. This technique can be used to extract the import feature by using a metric called as feature importance. For a feature this metric is found by identifying how much role it plays in its respective decisions trees classification accuracy. S. Perera et al. (2020) [7]: The authors conducted the study using RF on dataset GSE5281. This dataset contained a total on 161 samples with 24,438 unique gene symbols. RF was studied along with 2 other feature selection technique to extract the top 50 features based on their feature importance scores and PCA was found to outperform the RF feature selection technique. D. Sun et al. (2022) [12]: conducted study with dataset from 4 sources viz. GSE5281, GSE44771, GSE109887 and GSE132903 were GSE5281, GSE44771 were combined to increase sample size and GSE109887, GSE132903 were used for validation. Using Gene Ontology and Pathway Enrichment top 120 genes where extracted. These 120 genes were reduced further to 6 gene symbols using RF. A maximum average ACC of 92.3% was observed when model validated using 5-fold cross validation.

c. Variable Selection using Random Forest (varSelRF): varSelRF is a variation of Random Forest technique which is an ensemble method using decision tree. This technique is well suited for regression, classification and feature selection task. varSelRF involves building of decision trees and selection the nodes or features that are widely available across multiple trees. Nodes present in the top of the tree are the most important features. varSelRF is robust in identifying both linear and non-linear relationships between features. A. Sharma et al. (2021) [10]: conducted study using varSelRF along with LASSO to select important genes for classification of AD using gene expression dataset extracted form 4 different brain regions (Prefrontal Cortex, Medial temporal gyrus, Hippocampus, Entorhinal cortex). Genes selected from only varSelRF performed well in 3 regions Prefrontal Cortex, Hippocampus, Entorhinal cortex with ACC of 95%, 94% and 95% respectively. However, a max ACC of 98% was observed when the features selected using LASSO were integrated with features selected using varSelRF.

d. Extra Tree Classifier (ETC): ETC is a similar technique to RF. The way in which ETC uses a random threshold for each feature in the dataset when constructing the nodes of the decision tree is what makes it different from RF. This level of randomization helps the ETC to be more reliable in reducing the noise and to avoid overfitting. S. Perera et al. (2020) [7]: The authors conducted the study on ETC using dataset GSE5281 which contained a total on 161 samples with 24,438 unique gene symbols. ETC was compared with 2 other feature selection technique to extract the top 50 features based on their feature importance score and PCA was found to outperform the ETC.

e. Adaptive Boosting for miRNA Disease Association (ABMDA): Ensemble based technique specifically designed to locate differentially expression miRNA (DEmiRNAs). Uses the augmentation of meta-analysis to identify specific RNA sites responsible for a disease under consideration. Yuen, S.C (2021) [13]: used ABMDA to extract 28 DEmiRNA which were the potential biomarkers for classification of AD.

However, these 28 DEmiRNAs identified in the study were not put of classification using conventional classification methods instead were cross validated using biological pathway analysis.

C. Others:

a. Variational Autoencoders (VAEs): It is a generative model used to recognize the features, trends and distribution in the given sample. VAEs can be used thereby to find a compacted or an encoded version of the given dataset which helps in reducing the dimension of the given dataset while preserving the important features. VAEs have 2 components called Encoder and Decoder. Encoder phase of the model is used to find a compacted representation of the input sample and Decoder phase is used to reconstruct the actual sample from the compacted input representation. The loss function is the difference between the reconstructed sample and the actual sample which is then used in backpropagation to fine tune the encoder phase. Lee, T et al. (2020) [2]: have conducted study on using VAEs to reduce the dimension of the samples collected as part of datasets ADNI, ANM1, ANM2. VAEs was used to construct a compacted version of DEG which was then used for classification phase. Study found SAM feature selection technique outperformed DEG+VAEs technique as it was found VAEs seemed to lose important features in encoding gene expression values responsible for AD.

III. 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 selection algorithms was seen to affect the ability of the classification model due to the high dimensionality problem.

A. 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 selection 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 ($\chi 2$, 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. As the datasets uses tissues collected from brain this would be only possible with autopsy or biopsy which is less feasible when compared to gene expression collected from other tissues like blood. S. Perera et al. (2020) [7]: conducted their study involving SVM as their classification technique and found to be yielding a maximum ACC of 93.9% when evaluated using top 14 genes collected using PCA technique. SVM with Linear Kernel in the study was found to outperform other techniques such as SVM with Gaussian Kernel and RF used in the study.

$$minimize \frac{1}{2}|w|^2 \; subject\ to\ y_i(w^T\phi(x_i) + b) \geq 1, i = 1,2,\dots,n$$

Where x_i is the feature vector i.e., gene expression values for gene symbols and y_i represents the feature label for the i[th] vector. w represents the normal vector to the hyperplane that is used for classification. b is the bias and n is the number of samples used for training. $\phi(x_i)$ is the function to obtain the feature vector by applying kernel function over x_i.

B. 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. H. M. AL-Bermany et al. (2021) [9]: used KNN along with Deep Neural Network for performing classification task on the gene subset extracted using ANOVA and MI. In this study KNN was used to organize genes as clusters having same data points. From the given D genes in the data K genes were extracted, where K < D is strictly maintained. ANOVA was found to work to pair well with ANOVA than MI were a maximum accuracy of 92.9% was observed.

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

The above is the formula used to compute the Euclidean distance between the feature vector x and group y. point x is associated to the group y having the minimum

distance. Similarly, the Manhattan distance is computed using the below equation.

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

C. Convolution Neural Network (CNN): A common deep learning model predominantly used for classification tasks, such as image classification, image segmentation and object detection. Has special layers called convolution layers which perform mathematical operation called convolution on the input data based on a kernel with static size. CNN models also use polling layers to down sample or to reduce dimension using min, max or average of the values over a small region. Though CNN are mainly designed for image classification this can be extended to classification of AD using gene expression as biomarker. H. M. AL-Bermany et al. (2021) [9]: used CNN for classification of AD using the gene clusters extracted using KNN. CNN's performance was evaluated without using KNN as well and found CNN performed better when gene clustering with KNN was carried out. ACC of 92.9% was observed when ANOVA for feature selection, KNN for feature clustering and CNN for classification was used. Kalkan H et al. (2022) [11]: study used CNN for classifying dataset formed by integration 3 datasets GSE63060, GSE63061 and GSE140829. The integrated datasets where Min-Max normalized to rescale the RNA expression values between 0 and 1. Gene other the common genes across all the three datasets were ignored resulting in 11,618 unique genes. A total of 488 important genes were selected using feature selection technique involving LASSO. LDA was employed to map the 488 into a 2D space which is a supervised learning model separating the dataset into groups or classes to improve the accuracy of the classification model. The extracted image was then passed to the CNN model with 6 convolution layers, 3 max pooling layers and 2 Dense layers. A maximum ACC of 84.2% was observed.

D. Deep Neural Network (DNN): DNNs with output layers having SoftMax activation functions are generally used for designing classification models. The number of layers, nodes at each layer, activation and loss functions are each layer depend on the domain over which classification is applied. DNN models with classification done using gene expression values with large dimensions require model to be complex with large number of hidden layers. If number of unique gene are fewer complex models lead to overfitting. C. Park et al. (2020) [8]: conducted study using DNN to perform classification with the 35 genes extracted using the t-test. The study proposed the DNN model with 8 layers, 306 nodes with dropout 0.85 and Learning rate of 0.02 and observed in k-fold cross validation an avg. accuracy of 82.3%. When compared to other classification techniques such as RF, SVM and Naïve Bayesian the proposed DNN model performed better in the study. D. Sun et al. (2022) [12]: used DNN in their study to classify AD using 6 genes extracted using RF based feature selection technique. The model was built using 6 input layers, 5 hidden layers and 2 output layers. An average accuracy of 92.3% was observed when model was put on 5-fold cross validation.

E. 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.

F. eXtreme Gradient Boosting (XGBoost): XGBoost is an ensemble based boosting algorithm in which multiple weak decision trees are ensembled together into a single strong model. Advantage of XGBoost is that it allows to dynamically adapt the hyperparameters like maximum depth of trees and learning rate. S. Khanal et al. [2021]: have conducted the study on using XGBoost on dataset after the feature set are filtered using 2-step feature selection techniques t-test and SelectFromFeature. A maximum ACC of 65% and AUC of 67% for the gene group EMCI vs AD was observed.

G. Elastic Net (ElasticN): ElasticN model is obtained when L1 (Lasso) and L2 (Ridge) regularized linear regression models are combined. The dominance of L1 and L2 is controlled by a parameter called alpha, where alpha = 1 then it corresponds to Lasso and when alpha = 0 it is Ridge. This model has known to have a great immunity against overfitting by shrinking the less important features close to 0 as much as possible. ElasticN has the capability to handle large and correlated features better than other classification models. A. Sharma et al. (2021) [10]: conducted the study using the ElasticN with top 208 genes selected using LASSO and varSelRF from the gene expression collected from 4 difference brain regions (Prefrontal Cortex, Medial temporal gyrus, Hippocampus, Entorhinal cortex). ElasticN was found to perform well in all the 4 brain regions compared to other classification models such as SVM and RF. An average ACC of 98% was observed in the study.

IV. DISCUSSION

In this study, we aimed to identify the most effective feature selection and classification methods for the classification of AD using gene expression data. High dimensionality is a

common problem in gene expression data, and feature selection techniques are needed to effectively analyze the data and identify potential biomarkers for AD. We evaluated several feature selection methods on multiple categories, including Statistical methods, Ensemble Learning methods and other miscellaneous methods. Table 1 summarizes the study on various feature selection and classification techniques that have been explored. Study was also aimed to explore the advantages of using blood gene expression as biomarker instead of currently widely used methods using MRI images.

Gene expression data from different brain regions was observed to perform better than the gene expression extracted from blood tissues. A max ACC of 98% and 97.2% was observed as per the study A. Sharma et al. [10] and El-Gawady, A et al. [6] respectively. However, extracting gene expression from brain tissues would involve autopsy or biopsy which are invasive and less feasible when compared to extracting gene expression from blood tissues. A max ACC of 92.9% was observed in the study conducted by H. M. AL-Bermany et al. [9] indicating gene expression is feasible and accurate for AD classification.

Study	Dataset Used	Feature Selection Alg.	No. of Genes	Classification Alg.	Performance
Lee, T et al.	GSE3060 GSE3061	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	χ^2, ANOVA, MI	30	SVM	ACC: 97.5% AUC: 97.2%
S. Perera et al.	GSE5281	PCA, RF, ETC	14	SVM	ACC: 93.9%
C. Park et al.	GSE33000 GSE44770 GSE80970	t-test	35	DNN	ACC: 82.3%
H. M. AL-Bermany et al.	GSE3060 GSE3061	ANOVA + k-means	2500	CNN	ACC: 92.9%
A. Sharma et al.	GSE33000 GSE44770 GSE118553 GSE1132903 GSE5281 GSE48350 GSE28146 GSE5281 GSE48350 GSE4757	varSelRF + LASSO	208	ElasticN	ACC: 0.98
Kalkan H et al.	GSE63060 GSE63061 GSE140829	LASSO	488	CNN	ACC: 84.2% AUC: 87.5%
D. Sun et al.	GSE5281 GSE44771 GSE109887 GSE132903	RF	6	DNN	ACC: 92.3%

Table 1: Summary of studies conducted on detection of AD using various feature selection and classification techniques with gene expression.

Statistical methods such as SAM, t-test, χ^2, ANOVA, MI, PCA and LASSO methods were observed to be simpler to implement and most effective in understanding the correlation between features. Out of all the statistical methods studied χ^2 was observed to perform well when used as a standalone feature selection method with max ACC of 97.2%

with gene expression data extracted from bran tissues extracted from different region of the brain. Ensemble methods such as RF, ETC, Adaboost, and varSelRF were also explored and were also found to be effective in the removing the dominance of irrelevant features from the input feature set. Multiple decision trees were ensemble together in these approaches to rank nodes based on their importance in the classification of AD. As per existing study maximum ACC of 95% was observed when ensemble method varSelRF. However, when Ensemble and Statistical methods were combined as per done in study A. Sharma et al. [10] a significant improvement in the accuracy was observed. Since the important features left out by one method is caught by the other the combination proved to be the most suitable way for feature selection. Combination of varSelRF and LASSO yielded a max average accuracy of 98% on the datasets extracted from different brain regions.

Multiple state-of-the-art classification techniques such as DNN, CNN, SVM, KNN, ElasticN, DRNN were also evaluated. With reference to the study done by M. S. Kamal et al. [3] from the results it was observed ACC of classification model was heavily influenced by the efficiency and accuracy of the feature selection technique. Use of high dimensional genes directly without reducing the dimensions for classification yielded ACC of just 82.4% from which we can concluded that feature selection plays an important role in reducing the problem on overfitting by the classification models. SVM was found to be the most popular and widely used linear classifier technique for the classification of AD. However, ElasticN was observed to be the top performing model with max ACC of 98% when used for classification with genes selected from varSelRF and LASSO.

In this study AD classification techniques that does not depend on gene expression data were also explored to prove the advantages of using blood gene expression as biomarker. Table 2 summarizes the studies that rely on other sources for AD classification such as MRI images. S. Pavalarajan et al. (2022): conducted the study to classify AD patients using their MRI images. Proposed model in the study was trained using MRI images from OASIS dataset. Preprocessing task such as handling missing data, feature scaling and test train split was carried out before classifying the image set using classification techniques such as RF, LR, SVM and Decision Tree. RF performed better compared to other models with a max ACC of 83.5%. S. S. Rajeswari et al. (2021) [15]: performed study using MRI images collected as part public repository ADNI. Transfer Learning was used in the study to achieve high accuracy and to overcome the large dataset constraint since to improve the accuracy of the classification model image with high resolution was used. For classification of AD state-of-the-art models such as VGG-16, VGG-19 and Resnet-50 were used. Compared to other models VGG-19 yielded better accuracy of 98%. VGG-19 model was modelled with 16 3x3 convolution layers, 5 2x2 max-pooling layers and 3 fully connected layers. J. Li et al. (2022) [16]: performed the study using 3-D brain T1-weighted structure MRI images from ADNI public source. Data preprocessing methods such as resampling, skull stripping, intensity correction and clipping were explored. U-Net CNN is a variation of vanilla CNN which was found to perform better with the 3-D MRI images. S. Basheer et al. (2021) [17]: performed study on OASIS dataset of MRI images where classification was done using CapNet. CapNet was implemented using MCapNet package in python. A max ACC of 92.39% was observed and the proposed model was observed to outperform other conventional classification methods such as CNN. Hence, as the accuracy of models using gene expression as biomarkers have accuracy closer or sometimes better compared to existing systems using MRI images, blood gene expression can be seen as an effective biomarker for AD classification.

Study	Dataset	Classification Alg.	Accuracy
S. Pavalarajan et al.	OASIS	RF	ACC: 83.50%
S. S. Rajeswari et al.	ADNI	VGG-19	ACC: 98.00%
J. Li et al.	ADNI	3-D U-Net CNN	ACC: 95.06%
S. Basheer et al.	OASIS	CapNet	ACC: 92.39%
G. Chutani et al.	Kaggle	VGG-16 + SVM	ACC: 99.33%
D. Chaihtra et al.	Kaggle	DenseNet121	ACC: 96.00%
U. R. K et al.	Kaggle	GLCM + SVM	ACC: 84.00%
S. Buyrukoğlu et al.	ADNI	Stacked Ensemble	ACC: 91.20%

Table 2: Summary of studies using biomarkers than gene expression

V. CONCLUSION

In conclusion, our study has shown that using a combination of feature selection and classification techniques can effectively classify individuals with Alzheimer's disease (AD) using blood gene expression as a biomarker. By utilizing Recursive Feature Elimination and Random Forest feature importance for feature selection, we were able to identify a subset of genes that were most informative for AD classification. Additionally, by using Support Vector Machines and Random Forest as classification methods, we were able to achieve high accuracy rates in distinguishing AD patients from healthy controls.

However, it is important to note that these results should be validated in a larger, more diverse population before any potential diagnostic use. Additionally, further research should also be conducted to evaluate the performance of these methods in different stages of AD and in comparison, with other AD diagnostic methods. Overall, our study suggests that blood gene expression has potential as a biomarker for AD classification and that utilizing advanced feature selection and classification techniques

could aid in the development of a diagnostic tool for AD in the future.

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.

[8] C. Park, J. Ha, and S. Park, "Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset," Expert Systems with Applications, vol. 140, pp. 112873, 2020, doi: 10.1016/j.eswa.2019.112873.

[9] H. M. AL-Bermany and S. Z. AL-Rashid, "Microarray Gene Expression Data for Detection Alzheimer’s Disease Using k-means and Deep Learning," 2021 7th International Engineering Conference “Research & Innovation amid Global Pandemic" (IEC), 2021, pp. 13-19, doi: 10.1109/IEC52205.2021.9476128.

[10] A. Sharma, 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, Apr. 2021, doi: 10.1016/j.ygeno.2021.04.028.

[11] Kalkan H, Akkaya UM, Inal-Gültekin G, Sanchez-Perez AM. Prediction of Alzheimer's Disease by a Novel Image-Based Representation of Gene Expression. Genes (Basel). 2022 Aug 8;13(8):1406. doi: 10.3390/genes13081406. PMID: 36011317; PMCID: PMC9407775.

[12] D. Sun, H. Peng and Z. Wu, "Establishment and Analysis of a Combined Diagnostic Model of Alzheimer's Disease with Random Forest and Artificial Neural Network," Frontiers in Aging Neuroscience, 2022. Available: https://www.proquest.com/scholarly-journals/establishment-analysis-combined-diagnostic-model/docview/2682564611/se-2. DOI: https://doi.org/10.3389/fnagi.2022.921906.

[13] Yuen, S.C., Liang, X., Zhu, H., Jia, Y., and Leung, S.W. "Prediction of differentially expressed microRNAs in blood as potential biomarkers for Alzheimer's disease by meta-analysis and adaptive boosting ensemble learning." Alzheimer's Research & Therapy, vol. 13, no. 1, 2021, p. 126. doi: 10.1186/s13195-021-00862-z.

[14] S. Pavalarajan, B. A. Kumar, S. S. Hammed, K. Haripriya, C. Preethi and T. Mohanraj, "Detection of Alzheimer's disease at Early Stage using Machine Learning," 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA), Coimbatore, India, 2022, pp. 1-5, doi: 10.1109/ICACTA54488.2022.9752827.

[15] S. S. Rajeswari and M. Nair, "A Transfer Learning Approach for Predicting Alzheimer's Disease," 2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE), NaviMumbai, India, 2021, pp. 1-5, doi: 10.1109/ICNTE51185.2021.9487746.

[16] J. Li, Y. Wei, C. Wang, Q. Hu, Y. Liu and L. Xu, "3-D CNN-Based Multichannel Contrastive Learning for Alzheimer’s Disease Automatic Diagnosis," in IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1-11, 2022, Art no. 5008411, doi: 10.1109/TIM.2022.3162265.

[17] S. Basheer, S. Bhatia and S. B. Sakri, "Computational Modeling of Dementia Prediction Using Deep Neural Network: Analysis on OASIS Dataset," in IEEE Access, vol. 9, pp. 42449-42462, 2021, doi: 10.1109/ACCESS.2021.3066213.

[18] G. Chutani, H. Bohra, D. Diwan and N. Garg, "Improved Alzheimer Detection using Image Enhancement Techniques and Transfer Learning," 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India, 2022, pp. 1-6, doi: 10.1109/INCET54531.2022.9824008.

[19] D. Chaihtra and S. Vijaya Shetty, "Alzheimer’s Disease Detection from Brain MRI Data using Deep Learning Techniques," 2021 2nd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2021, pp. 1-5, doi: 10.1109/GCAT52182.2021.9587756.

[20] U. R. K, S. S. S, U. M. G and V. B. C, "Binary Classification of Alzheimer's disease using MRI images and Support Vector Machine," 2021 IEEE Mysore Sub Section International Conference (MysuruCon), Hassan, India, 2021, pp. 423-426, doi: 10.1109/MysuruCon52639.2021.9641661.

[21] S. Buyrukoğlu, "Improvement of Machine Learning Models’ Performances based on Ensemble Learning for the detection of Alzheimer Disease," 2021 6th International Conference on Computer Science and Engineering (UBMK), Ankara, Turkey, 2021, pp. 102-106, doi: 10.1109/UBMK52708.2021.9558994.