Early Detection of Alzheimer's using Deep Learning and Cross Validated Features from Blood Gene Expression Data

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 (AD), a type of neurodegenerative disorder, has seen an increase in cases over the past decade, necessitating the construction of a comprehensive early detection method. Existing methods are typically invasive and costly, so our research concentrates on blood gene expression as a possible biomarker. The high dimensionality of the gene expression data and the small sample size complicate blood gene expression data analysis. Our novel approach attempts to address these issues by identifying a suitable feature selection method to reduce the dimension size from 12454 to 1258 which are validated with 5-fold cross validation. GAN based synthetic data modelling is used to address the issue of a small sample size. The classification of the resulting dataset using DNN yielded an accuracy of 91% and with precision of 95% in identifying AD samples. Feature selection along with synthetic data modelling significantly enhanced the precision of early detection of AD using blood gene expression.**

Keywords: Blood Gene Expression, Feature Selection, Synthetic Data Modelling.

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.	GSE63060 GSE63061	SAM	697	SVM	AUC: 87.4%
H. M. AL-Bermany et al.	GSE63060 GSE63061	ANOVA + k-means	2500	CNN	ACC: 92.9%

Kalkan H et al.	GSE63060 GSE63061 GSE140829	LASSO	488	CNN	ACC: 84.2% AUC: 87.5%
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%

Table 1: Summary of some of the notable studies conducted on detection of AD using various feature selection and classification techniques.

The Table 1 summarizes several studies on the detection of Alzheimer's disease (AD) using different feature selection and classification techniques. These studies utilized various datasets, including GSE3060, GSE3061, GSE63060, GSE63061, GSE140829, GSE33000, GSE44770, GSE44768, GSE44771, and GSE5281. Different feature selection algorithms such as Significance Analysis of Microarrays (SAM), Analysis of Variance (ANOVA), k-means, LASSO, χ2 (Chi-square), and Mutual Information (MI) were employed to identify informative genes. Classification algorithms like Support Vector Machine (SVM), Convolutional Neural Network (CNN), and others were used to build predictive models. The performance metrics reported included Accuracy (ACC) and Area Under the Curve (AUC), with the achieved accuracies ranging from 84.2% to 97.5% and AUCs ranging from 87.4% to 97.2%. These studies highlight the diverse strategies employed to detect AD and showcase varying levels of accuracy in distinguishing between AD and non-AD cases. El-Gawady, A et al. [6] have achieved a maximum accuracy of 97.5% in their work using datasets extracted from different brain regions viz. Prefrontal cortex, Medial temporal gyrus, Hippocampus and Entorhinal cortex. As this would require autopsy-biopsy procedures, our study focuses on less invasive source by using gene expression from blood tissue. Previous work done using blood gene expression dataset viz. H. M. AL-Bermany et al. [9] have achieved a max accuracy of 92.9% but the selected genes symbols have not yet been validated yet using cross validation to eliminate the possibility of bias and overfitting. Previous studies on using blood gene expression have not yet addressed the low sample size problem which poses a challenge in coming up with a generalized model. Our proposed tries to solve this by using a variation of Generative Adversarial Network (GAN) that can be used for tabular data.

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%

Table 2: Summary of studies using biomarkers other than gene expression

The Table 2 summarizes some of the notable previous studies using MRI image-based datasets that were utilized to classify AD and CTL samples. S. Pavalarajan et al. [14] employed the OASIS dataset and utilized the RF algorithm, achieving an accuracy of 83.50%. S. S. Rajeswari et al. [15] focused on the ADNI dataset, using the VGG-19 algorithm, and achieved an impressive accuracy of 98.00%. J. Li et al. [16] also worked with the ADNI dataset, employing the 3-D U-Net CNN algorithm, and obtained an accuracy of 95.06%. Lastly, S. Basheer et al. [17] used the OASIS dataset with the CapNet algorithm and achieved an accuracy of 92.39%. The results from Table 1 and Table 2 clearly indicate the proposed method of using blood gene expression data as an alternative to MRI images proves to be not only better in detecting AD in its early stages but also with comparable accuracy.

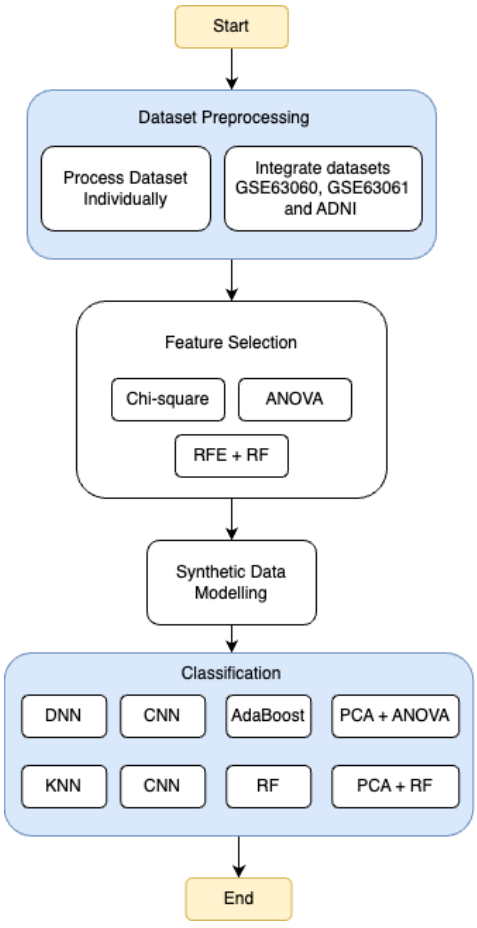

Fig 1: Overview of the proposed method

II. METHODS

A. Dataset:

For the study 3 different datasets were explored viz. GSE63060 [23], GSE63061 [24] and ADNI [25]. GSE63060 and GSE63061 gene expression samples were collected from GEO gene expression omnibus repository as SOFT formatted family files and ADNI gene expression samples were collected as part of the Alzheimer's Disease Neuroimaging Initiative (ADNI) which is a longitudinal

multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of AD. 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). GSE63061 contained samples collected from 382 individuals out of which 139 were AD samples, 134 were CTL, and 109 were samples collected from people with MCI. Each sample from GSE63060 and GSE63061 contained expression values of 38323, 32049 probes respectively, 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. ADNI contained 431 sample out of which all were AD samples. Each sample from ADNI contained gene expression values of 48548 gene symbols.

B. Preprocessing:

We performed the study using the datasets in 2 ways. One involved analyzing the classification models performance individually on each dataset and the other involved integrating the 3 datasets together into a single set. Before integrating datasets GSE63060, GSE63061 and ADNI we normalized them individually using Min-Max normalization which scaled the datasets to have gene expression values between the range 0 and 1. Individually normalized datasets were then integrated by finding the common columns or gene symbols present in all the three the datasets. This was done in python using simple intersection operation between the columns of the three datasets which resulted in 12459 common columns. Post integration the dataset was again normalized using Min-Max normalization to avoid classifier models favoring samples from a particular set since the intensity of gene expression value may vary with the apparatus and measurement procedures used by the lab. In the study it was observed the number of AD samples significantly exceeded the number of CTL samples. Under Sampling of AD samples was done to avoid models favoring AD samples. The resultant dataset contained 238 AD and 238 CTL samples which were then put for 80-20 split for generating train and test sets.

C. Feature Selection:

As there are over 12,459 dimensions in the integrated 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) Chi-square ($\chi 2$): $\chi 2$ 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.

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

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. Using $\chi 2$ test as the estimator and the number of genes to select for classification as the optimal parameter, hyperparameter tuning was performed to identify 10814 genes to be the optimal number of genes to be selected for classification model. Since the dataset contained features to be continuous gene expression values and target to be categorical taking values AD and CTL the continuous values were transformed into discrete categorical values using Binning using scikit package in python. Accuracy of the training set was found to be 93.75% but on further analysis using cross-validation splits of the dataset accuracy dropped drastically indicating the overfitting. Using deep-learning based classification model also produced similar results. Figure 2 represent the loss and accuracy curve variation with respect to the number of features selected. It clearly depicts the fact that maximum accuracy was achieved when top 10814 genes are selected for the classification task.

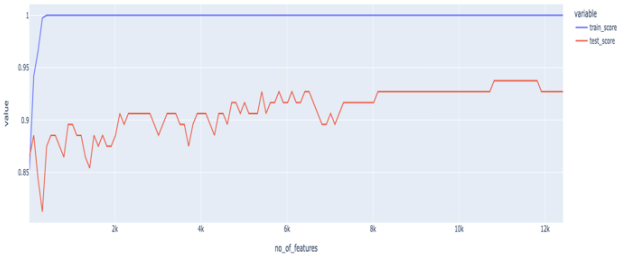

Fig 2: Variation of Accuracy of classification model with respect to the number of features selected using $\chi 2$.

b) 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.

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

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

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

Fig 3: Variation of Accuracy of classification model with respect to the number of features selected using ANOVA.

Where F represents the ANOVA coefficient, and MS_B represents Mean Square between and Groups and MS_E represents 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. Using ANOVA as the estimator and the number of genes to select

for classification as the optimal parameter, hyperparameter tuning was performed to identify 514 genes to be the optimal number of genes to be selected for classification model. Since the dataset contained features to be continuous gene expression values and target to be categorical taking values AD and CTL it required no preprocessing of the dataset unlike $\chi 2$ test. Accuracy of the training set was found to be 91.66% but on further analysis using cross-validation splits of the dataset accuracy dropped drastically indicating the overfitting. Using deep-learning based classification model also produced similar results. Figure 3 represent the loss and accuracy curve variation with respect to the number of features selected. It clearly depicts the fact that maximum accuracy was achieved when top 514 genes are selected for the classification task.

c) Recursive Feature Elimination (RFE) with Cross Validation (CV): RFE is a technique for selecting the most pertinent features for a given task that is commonly employed in machine learning. It is an iterative procedure designed to eliminate irrelevant or redundant characteristics from a dataset. The RFE algorithm begins by training a model on the entire feature set and assigning importance scores to each feature based on their contribution to the performance of the model. The least significant characteristics are then eliminated, and the procedure is repeated for the remaining characteristics. This elimination process is repeated until a predetermined number of features, or a desired level of feature importance is attained. RFE identifies a subset of features that are most informative and influential in predicting the target variable by systematically eradicating features. It considers the interactions between features and their impact on the performance of the model, which can be particularly useful when working with high-dimensional datasets. To ensure robustness and prevent overfitting, RFE is frequently paired with a cross-validation scheme. The selection of the underlying model and the metric used to evaluate the significance of a feature can vary based on the problem domain and the characteristics of the data. For this study RFE was explored with 5-fold cross-validation. Decision Tree Classifier (DTC) was used as an estimator to select the optimal number of features. Other estimators such as Logistic Regressor, SVM and DTC was found to be optimal in selecting the genes. As a result, set of 1258 genes were found to be optimal with a confidence of 90.625% when validated using the testing set. Table 3 summarizes the accuracy of various splits vs the 3 sets of gene symbols with the top mean accuracy to be 82.63% corresponding to the 1258 genes. By cross-referencing the genes selected with the gene selected by the previous studies and indicated few new set of gene symbols that might be contributing to classification of AD.

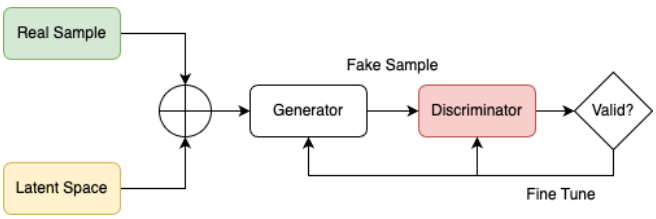

Fig 4: Tabular GAN pipeline

D. Synthetic Data Modelling:

Synthetic data modeling involves generating artificial data that resembles real-world data to simulate various scenarios for research, testing, and analysis purposes. It allows researchers and data scientists to create controlled environments, explore hypothetical situations, and evaluate the performance of algorithms or models without the need for sensitive or limited data. Synthetic data modeling can aid in data privacy protection, algorithm validation, and developing robust machine learning models that generalize well to real data.

Tabular data is typically represented in a structured format, consisting of columns and rows with categorical or numerical values. GANs are traditionally designed for continuous data, such as images. Adapting GANs to handle tabular data requires finding appropriate representations and architectures that can effectively capture the underlying data distribution. Tabular data can have a high number of features or dimensions, making it challenging for GANs to learn the complex interactions and dependencies between variables. High-dimensional data often requires specialized techniques, such as dimensionality reduction or feature selection, to improve the performance and stability of GAN models. Tabular data often includes categorical variables, which are not directly compatible with GANs that typically operate on continuous data. Converting categorical variables to continuous representations can introduce challenges in preserving the semantics and relationships between categories during the generative process. GANs are susceptible to mode collapse, where the generator fails to capture the full diversity of the data distribution and instead produces limited variations. In the context of tabular data, mode collapse can manifest as the generator generating only a subset of representative samples, ignoring less frequent patterns or outliers. Assessing the quality and performance of GAN-generated tabular data poses its own challenges. Traditional evaluation metrics used for image generation, such as Inception Score or Fréchet Inception Distance, may not be directly applicable. Developing appropriate evaluation techniques specific to tabular data is an ongoing research area. Tabular datasets may suffer from class imbalance, where certain classes or categories are underrepresented compared to others. GANs may struggle to generate balanced synthetic data that accurately represents the distribution of the original imbalanced data. Addressing these challenges often requires specialized techniques and modifications to the GAN architecture, loss functions, or training procedures. Researchers are actively exploring these challenges to enhance the applicability of GANs for generating high-quality synthetic tabular data. TGAN developed by I. Ashrapov et al. [26] model overcomes the problem discussed above by the use dynamic latent space based on data type of the column. As shown in Fig 4 the discriminator constantly tweaks the generator function to generate fake samples closer to the real samples, hence avoiding the overfitting. Post processing of the samples is also done to remove samples having noise greater than the threshold. As shown the Table 4 we can clearly see that the models that suffered overfitting have performed better when synthetic data modelling is included in the process. When the synthetic data modelling was applied on the individual dataset overfitting in TGAN model was observed.

Split 1	Split 2	Split 3	Split 4	Split 5	Mean
85.52	84.21	82.89	80.26	78.94	82.36
88.15	81.57	85.52	82.89	75.00	82.63
82.89	82.89	82.89	84.21	75.00	81.57

Table 3: Accuracy across 5-fold splits for top 3 groups of gene selected using RFE + RF

E. Classification:

Support Vector Machine (SVM): SVM is the most popular and widely used linear classifier technique. This involves classification based on supervised learning approach. An optimal hyperplane that separates data points is found to predict different classes of the target feature. In this study when SVM was applied on the dataset 98%

accuracy on the training set and 84% accuracy on the test set was observed which clearly indicated overfitting in the model. Hence other models were explored.

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. A 100% accuracy on the training set and 90.02% accuracy on the test set was observed in the study which clearly indicated overfitting. Hence other ensemble models were explored.

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. In the study conducted with the RF model similar results to the AdaBoost model was observed which indicated overfitting.

Deep Learning Classifiers:
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. Since in the study we performed feature selection with RFE and RF when the classification was done using a complex model the result clearly indicated overfitting. Hence the model was made simple with multiple dropout layers between the dense layers. Table 5 discusses the DNN model in detail and Table 8 summarizes the training and testing accuracy. A max accuracy of 91% was observed on testing set with precision of detecting AD class to be 95% which clearly indicates the model performed has performed well in classification task. This performance was greatly reduced when the synthetic data modelling was not done. Accuracy on validation set was seen to be less when the model was put for classification on individual dataset sources instead of integrated dataset.

Layer	Output Shape	Param #
Dense 1	(None, 7)	8813
Dropout 1	(Node, 7)	0
Dense 2	(None, 6)	48
Dropout 2	(None, 6)	0
Dense 3	(None, 6)	42
Dropout 3	(None, 6)	0
Dense 4	(None, 6)	42
Dense 5	(None, 5)	35
Dense 6	(None, 1)	6
Total Params	8,986	
Trainable Params	8.986	
Non-Trainable Params	0	

Table 5: DNN proposed models summary

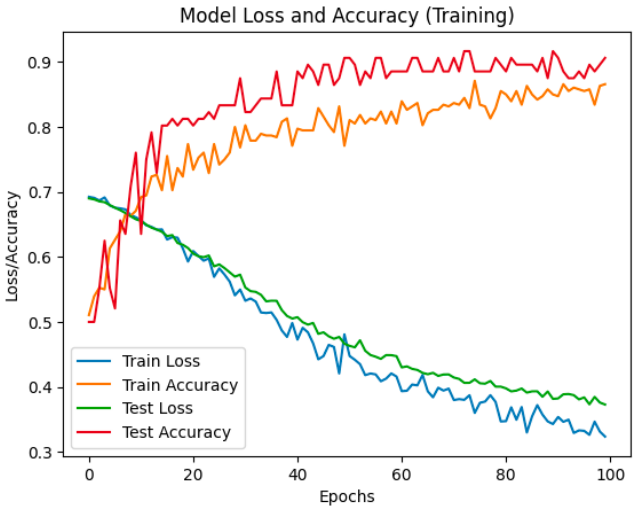

Fig 5: Train and Test set Loss/Accuracy variation curve for the DNN

Figure 5 clearly proves the fact that the DNN model developed is generalized since we the variation of Loss/Accuracy curve of Train and Test is same with the max accuracy attaining peak value at 91%. The study was conducted for 100 epochs since post 100 epochs the accuracy was not observed to improve further.

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.

Layer	Output Shape	Param #
Conv1D 1	(None, 1256, 4)	16
Max Pooling 1	(None, 628, 4)	0
Conv1D 2	(None, 626, 3)	39
Max Pooling 2	(None, 313, 3)	0
Conv1D 3	(None, 311, 3)	30
Max Pooling 3	(None, 155, 3)	0
Flatten 1	(None, 465)	0
Dense 1	(None, 4)	1864
Dense 2	(None, 2)	10
Total Param	1,959	
Trainable Params	1,959	
Non-Trainable Params	0	

Table 6: 1-D CNN proposed models summary

Gene expressions can be represented as a grid-like structure, where the spatial relationships between gene expression values are important. Hence CNNs can be used for classification tasks involving gene expression data. Table 6 summarizes the 1-D CNN model used in the study. In study it was observed increasing the complexity of the CNN resulted in overfitting. A max accuracy of only 80.67% on test set was observed even though the model performed with 98.88% in the training set which clearly showed us the characteristics of the overfitting. This helped us to conclude to go with DNN based model of classification since CNN was too complex model for the dataset with 1258 gene expression features.

Feature Selection	Feature Extraction	No of Genes Selected	Classification	Modelled Synthetic Data	Accuracy (Training Set)	Accuracy (Testing Set)
Chi Square	-	10814	SVM Gaussian Kernel	No	99.67%	77%
Chi Square	-	10814	DNN	No	96.68%	89.33%
Chi Square	PCA	30	RF	No	100%	75%
ANOVA	PCA	30	RF	No	100%	82%
ANOVA	-	514	AdaBoost	No	100%	90.20%
ANOVA	-	514	RF	No	100%	86.45%
ANOVA	-	514	SVM Gaussian Kernel	No	99.83%	87.42%
ANOVA	-	514	DNN	Yes	93.77%	90.60%
RFE + RF	-	1258	1-D CNN	Yes	98.88%	80.67%
RFE + RF	**-**	**1258**	**DNN**	**Yes**	**93.65%**	**91%**

Table 7: Summary of results observed while classifying AD from CTL samples using the dataset formed by integrating data from 3 sources.

Figure 6 clearly proves the fact that the 1-D CNN model developed is not suitable for the classification of AD since from the variation of Loss/Accuracy curve of Train and Test shows huge variation clearly proving overfitting. Also form the further analysis post 10 epochs the loss in the test set starts increasing even though it achieves close to 99% accuracy in the training set.

Fig 6: Train and Test set Loss/Accuracy variation curve for the 1-D CNN indicating worse performance compared to DNN.

III. RESULTS AND DISCUSSION:

Based on the methods discussed in the previous section, various models over two type of dataset groups were applied which are summarized in the Figure 7. RFE + RF were found to be effective in extracting useful 1,258 gene symbols from the 12,459. AARSD1, AASDHPPT, BCL9L, BDP1, C11ORF1, DUSP22, ROR1, TGFB2, ZNF33B where some of the top 1.258 genes selected. Comparing the results of the dataset groups clearly proves the fact that the training and test model on individual dataset does not help in designing a generalized model compatible with different sources of blood gene expression datasets. Even though models trained with individual dataset perform better on training set they don't carry the same performance into the testing set. Low sample size is found to be the major contributing factor for the low accuracy and overfitting in the model. From Figure 7 we can clearly understand the overfitting w.r.t to the huge difference in the testing and training accuracy when individual dataset is used. For the integrated dataset even though the training accuracy is only 93.65% the testing accuracy is of around 91% along with 5-fold cross validation of the features selected rules out the possibility of overfitting. These results also prove the proposed model to be a viable replacement of the exiting early detection methods based on MRI images briefly discussed in the Table 2.

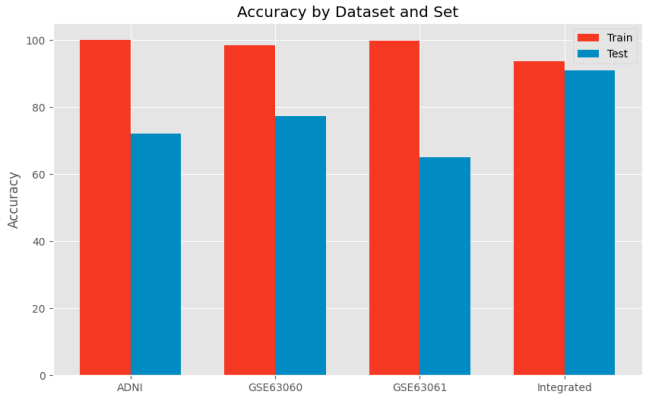

Fig 7: Comparison between the performance of Individual dataset vs Integrated dataset

Table 8 discussed the various metrics of the best performing model evaluated using the Integrated dataset. The precision for class CTL (0.87) and class AD (0.95) indicates that the classifier has a high ability to correctly identify instances of each class. A higher precision value suggests a low false positive rate. The recall for class CTL (0.96) and class AD (0.85) indicates that the classifier can capture a high percentage of true positives for each class. A higher

recall value suggests a low false negative rate. The F1-score combines precision and recall into a single metric. Both class CTL (0.91) and class AD (0.90) have high F1-scores, which suggests a good balance between precision and recall. The overall accuracy of 0.91 indicates that the classifier correctly predicts the class labels for approximately 91% of the instances. Macro Avg and Weighted Avg: Both the macro average and weighted average F1-scores are 0.91, indicating consistent performance across classes. The weighted average considers the class imbalance, giving higher weight to the class with more instances (if any). Overall, based on these evaluation metrics, the classifier appears to be performing well with high precision, recall, and F1-score for both classes.

	Precision	Recall	F1-Score	Support
CTL	0.87	0.96	0.91	48
AD	0.95	0.85	0.90	48
Accuracy			**0.91**	96
Macro avg	0.91	0.91	0.91	96
Weighted avg	0.91	0.91	0.91	96

Table 8: Result metrics of the Proposed Method on the test set.

IV. 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. Study also proved using integrated dataset from multiple sources followed by synthetic data modelling using TGAN to increase the sample proved effective in building a more generalized model. However, care should be take when integrating dataset from different sources as each source may have different sensitivity to gene expression values. 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 Deep learning models such as DNNs and 1-D CNNs classification methods, we were able to achieve high accuracy rates in distinguishing AD patients from healthy controls CTL. DNN models were observed to performed better than 1-D CNN based models.

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.

Availability: *https://github.com/f16hari/sop has the necessary supplementary files and code used for this study and experiment.*

V. 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:

[8] 10.1109/MERCon50084.2020.9185336.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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

[16] Computing Technologies and Applications (ICACTA), Coimbatore, India, 2022, pp. 1-5, doi: 10.1109/ICACTA54488.2022.9752827.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Dubey, S. (2019, December 26). Alzheimer's Dataset (4 class of Images), Kaggledatase. https://www.kaggle.com/tourist55/alzheimers-dataset-4-class-of-images?select=Alzheimer_s%2Bt.

[25] "GSE63060 - GEO DataSets - NCBI." [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63060. [Accessed: Nov 16, 2022].

[26] "GSE63061 - GEO DataSets - NCBI." [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63061. [Accessed: Feb 27, 2023].

[27] "ADNI Data Samples - Access Data." [Online]. Available: https://adni.loni.usc.edu/data-samples/access-data/. [Accessed: Mar 2, 2023.

[28] I. Ashrapov, "Tabular GANs for uneven distribution," in arXiv:2010.00638 [cs. LG], 2020.