Development of in silico models for predicting Alzheimer's disease from single cell genomics

A Project Report

Submitted by

AMAN SRIVASTAVA

In partial fulfilment of the requirements
for the award of the degree of

MASTER OF TECHNOLOGY

COMPUTATIONAL BIOLOGY
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

FEB, 2023

THESIS CERTIFICATE

This is to certify that the thesis titled "**Development of in silico models for predicting Alzheimer's disease from single cell genomics**", submitted by **Aman Srivastava**, to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of Master of Technology, is a bona fide record of the research work done by him undermy supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree.

Prof. Gajendra P. S. Raghava
Thesis Supervisor
Head of Department
Dept. of Computational Biology
IIIT Delhi, 110020

Place: New Delhi

Date: 11th February 2023

ACKNOWLEDGEMENTS

This work would not have been possible without the guidance and support of several individuals who in one way or the other assisted in the preparation and completion of this thesis.

I would like to express my sincere gratitude and respect towards **Prof. G. P. S Raghava** from Indraprastha Institute of Information and Technology, Delhi for being my supervisor. I would also like to thank the Ph.D. scholars from his esteemed lab- Anjali Dhall and Sumeet Patiyal for guiding me throughout the year.

My supervisor has inspired me to do good work and has supported me all along the way. Dozens of people have helped and taught me immensely through this year long journey. I would like to thank my friends and batch mates for providing such a good environment for helping me whenever I got stuck.

Lastly, I would also like to thank my parents and family for motivating from time to time throughout the course of my thesis which enabled me to pursue my research in an efficient and structured manner.

Aman Srivastava
MT20333
M.Tech(Computational Biology)
IIIT Delhi, 110020

ABSTRACT

Alzheimer's disease is progressing as the most prevalent neurological disorder world-wide. It is the most common cause of dementia in ageing society. An artificial Neural Network (ANN) is a set of neural networks which are inspired by the human brain. They are learning algorithms which can learn and make corrections as they receive input. Despite these benefits, they are not actively used in classification problems involving single-cell genomics. Many recent studies have reported the effectiveness of Machine Learning models in predicting diseases using single-cell genomics, but the sample sizes were too small. Thus, here we have compared ANN with other ML models in prediction and biomarker identification with a large dataset. In this study, ANN was compared to other machine learning models on 169,496 cells of RNA-seq data from normal human subjects and AD patients' prefrontal cortex. Of these, 90713 were AD labelled, and 78783 were NC labelled. Two different feature sets were selected, and classification accuracies were determined with ANN, LR (Logistic Regression), RF (Random Forest) and other models. As a result, ANN showed the highest performance in both the features of 100 genes and 35 genes with accuracies of 82% and 74%, respectively. Interestingly, when the feature size was decreased to 35 genes, the ANN showed a small decline (7-8%) in accuracy, but it did not change drastically to a low value. In conclusion, it indicates that these conserved 35 genes can be used to predict Alzheimer's patients and can very well act as potential biomarkers for AD diagnosis and screening. Eventually we have developed a python package named "AlzScPred" based on the above study to facilitate the scientific community. (https://webs.iiitd.edu.in/raghava/alzscpred/)

KEYWORDS: In silico models, Alzheimer's disease, Genetic Biomarkers, Deep Learning, and Machine Learning

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AD	Alzheimer's Disease
NC	Normal Controls
AUC	Area Under Curve
DT	Decision Tree
RF	Random Forest
ANN	Artificial Neural Network
IFS	Incremental Feature Selection
mRMR	Minimal Redundancy Maximal Relevance
GO	Gene Ontology
sc RNA	Single Cell Ribosomal Nucleic Acid

CHAPTER 1

INTRODUCTION

Alzheimer's disease (AD) is one of the main reasons for dementia and has become one of the biggest challenges for the medical industry [1]. About 6.2 million individuals are currently affected with AD-induced dementia globally, and this number is predicted to double every 20 years if a cure is not found [2, 3]. It is a progressive neurological disorder that results in the death of brain cells along with brain atrophy and is characterized by a steady decline in social, behavioral, and cognitive abilities, and impairs a person's capacity for independent functioning [4]. Some of the pathological biomarkers of AD include neuroinflammation, deposition of amyloid-beta peptides, and tau neurofibrillary tangles [5].Refer figure 4.3. However, their presence does not necessarily indicate AD specific dementia. It is unclear how they are directly related toneurodegeneration [6]. The pathogenesis of AD usually starts much earlier than the symptoms show up. There are several studies and clinical trials that have been designed to focus on these pathological changes but were unsuccessful in treating AD [6]. Presently, there is no known effective cure for this disease. The current treatment procedures only suppress the symptoms [7].

One of the reasons for the absence successful treatment methods for AD is the lack of knowledge of molecular underpinnings of cell-type specific responses for the pathogenesis of atrophy and neurodegeneration [8]. The bulk-tissue-level analysis may obscure the complexity of modifications between within cells, particularly for rare cell types [9]. It is one of the major challenges to understand the complex human brain which comprises a huge variety of cells which include billions of neurons of different subtypes [10]. Single-cell RNA sequencing provides an approach to analyze thousands of individual cells and study the compositional as well as activity changes within the cells to gain a clear insight of mechanism of AD [11].

In this study, we have used the single-cell RNA sequencing data to identify gene-based biomarkers for Alzheimer's. To achieve this, we have applied computational methods to a dataset [12] that contains 169,496 nuclei from normal human subjects,

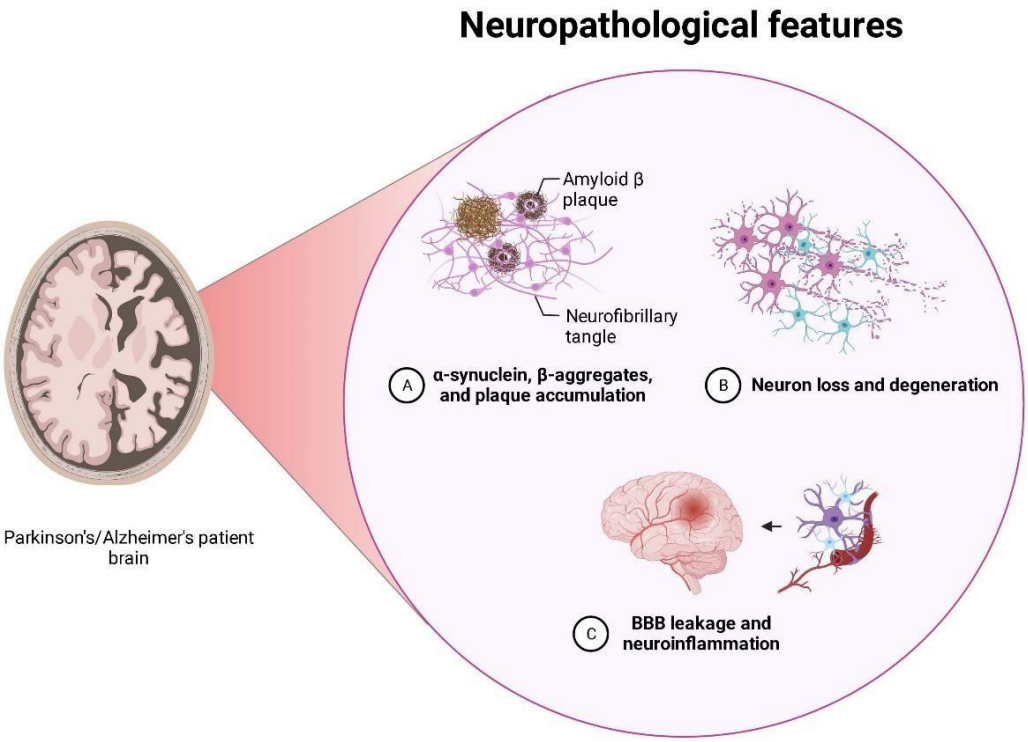

Figure 1.1: Neuro Pathological Features of Alzheimer's affected brain

and AD patients' prefrontal cortex [13]. The gene expression profiles of these samples were analyzed using the maximum relevance minimum redundancy (mRMR) method. After applying mRMR, a set of 35 genes was retrieved. The 35 genes recognized in this study can prove to be essential biomarkers for AD. Since we have identified the biomarkers using single-cell data, they could be helpful in finding more specific targets for the development of treatment methods for Alzheimer's. We believe this study will be helpful for the scientific community working on finding molecular-based therapies for Alzheimer's disease.

1.1 Single cell Genomics

Genomics is an interdisciplinary field of biology that deals with the study of the entire genome in order to understand the functioning at the genetic/ fundamental level.It increases our understanding of illnesses and diagnostic methods at the most fundamental level. Although there are many other forms of sequencing, such as single-cell DNA methylome sequencing, single-cell assay for transposase-accessible chromatin sequencing, and so on, we are focused on RNA-sequencing profiles. Traditionally,

RNA-seq profiling was done with bulk-RNA seq samples, which are made up of a variety of different cells. The bulk-RNA seq has numerous uses, such as identifying characteristic biomarkers between tissues of healthy/diseased samples or control/treated samples, discriminating between tissues by comparing transcriptomes, and locating and labelling novel genes, among others. However, bulk RNA-seq provides an estimate of the average expression level of each gene across a population of cells without taking into account cell heterogeneity. As a result, it does not provide a good picture of the individual cells in a sample and cannot be utilized to examine heterogeneous systems such as early development studies.

With the introduction of next-generation sequencing technology, transcriptome profiling became less expensive and time-consuming, paving the way for single-cell RNA sequencing. The single-cell RNA sequencing (scRNA-seq) fully overcame the constraints of bulk- RNA seq and made it feasible to estimate the distribution of expression levels of each gene throughout the population of cells. As a result, it is now able to answer fundamental biological questions such as what sort of cells are present in the tissue, what tasks these cells carry out, and how these functions differ from healthy tissues. Cell-type-specific information may be understood, which can aid in the discovery of novel or unusual cell types, the knowledge of cell differentiation throughout development, and the determination of cell composition in healthy and diseased tissues.

1.1.1 Challenges with scRNA-seq data

Because the starting material per cell is so little, scRNA-seq data presents numerous challenges. As a result, the data is quite sparse and contains a lot of zeros. The zeroes in the data may or may not be accurate. When a gene is not expressed in a cell, it is considered as a "actual" zero; however, when a gene is expressed in a cell but cannot be detected because of technological difficulties, it is counted as a "fake" zero or a "dropout." This causes unwanted variance across cells that did not arise due to biological variation but rather due to technical difficulties, such as the gene not being PCR amplified to a suitable level. This issue, however, can be resolved by doing normalization. Another concern with single-cell RNA sequencing data is batch effects. One of the most important stages in single-cell data analysis is data integration or data

harmonization. The batch effect is a phenomenon that occurs when many datasets from different laboratory settings are combined and sequenced using diverse technologies and equipment to generate a single huge reference dataset. This causes technical noise in the data, making it harder to determine the true biological variation present in the data.

CHAPTER 2

PREDICTION

2.1 Workflow

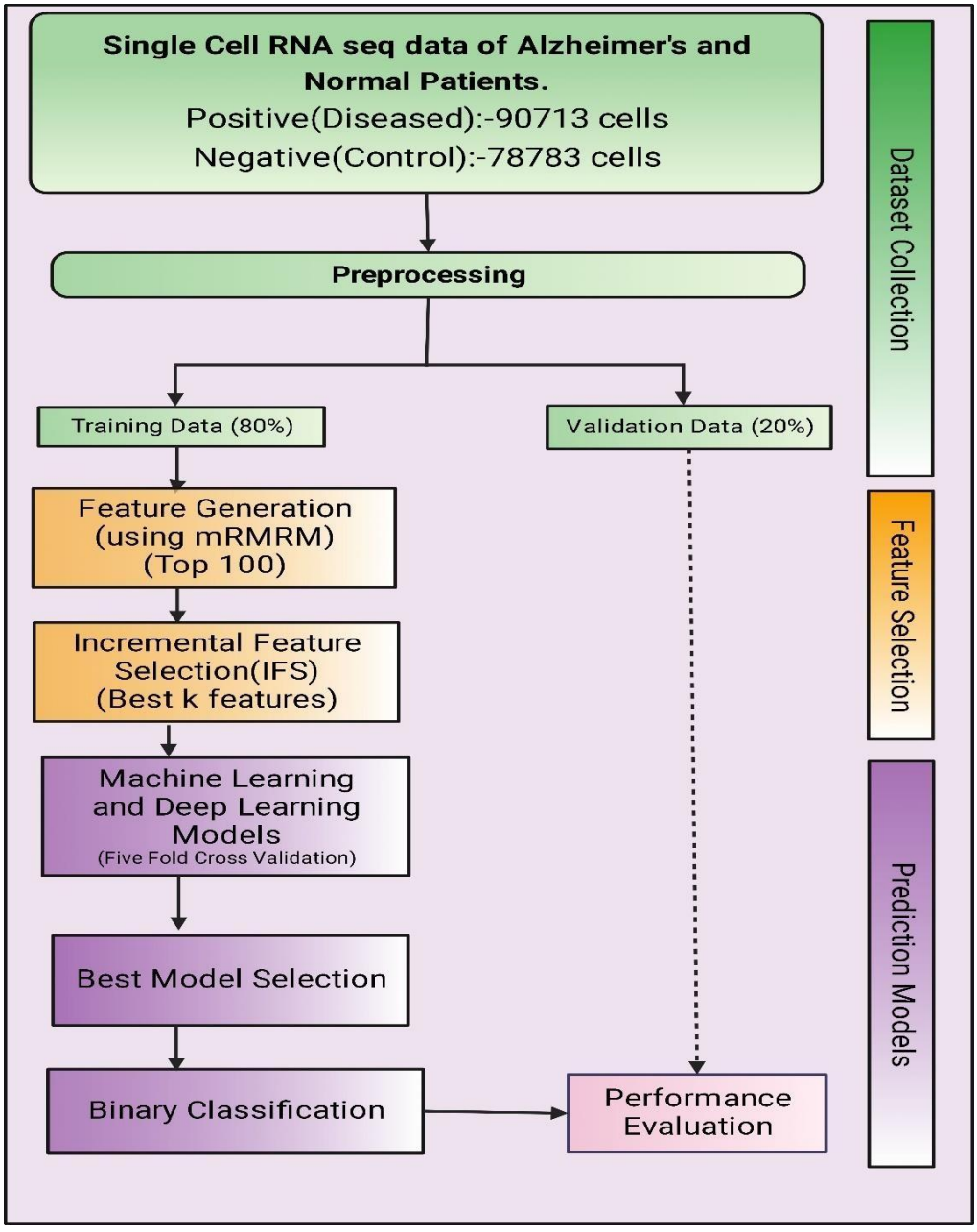

Figure 2.1: Workflow of data collection to prediction

2.2 Materials and Method

2.2.1 Data Collection

We obtained the dataset from NCBI GEO (Gene Expression Omnibus) with GEO accession number GSE157827 [13]. The keywords used for search were Alzheimer's, Single cell, RNA seq data, Control data etc. The dataset consisted of Single-cell expression profiling data of both Alzheimer's (Diseased) and Normal Patients.

2.2.2 Data description

The dataset consists of single-nucleus RNA sequencing of 169,496 nuclei from the pre-frontal cortical samples of AD patients and normal control (NC) subjects. The total Number of Patients samples was 21, out of which 12 patients were Alzheimer's affected, and 9 were normal controls. The data present was in the form of 10x single sequencing data consisting of raw counts, barcodes and gene files for each patient. The sequencing platform used for sequencing was Illumina NovaSeq 6000 (Homo sapiens), which measured a total of 33538 genes.

We determined the number of expressed genes, or the number of genes with mapped reads, in each patient sample. 5401 genes on average were expressed across all samples.

2.2.3 Data Pre-Processing

First of all, the full data processing and analysis were done in the python3 environment on a Jupyter notebook. Since the data provided by Cell Ranger Pipeline was in the sparse format, i.e. contains only the non-zero entries to minimize the file size. So using the os, csv, gzip and scipy.io modules we converted the data into a matrix in the form of a data frame.

- **Data loading and filtering**

Similarly, each patient's data was converted into a data frame, and then the dataset was pre-processed to remove insignificant columns and cells. The genes which didn't have

any mapped expression reads to more than 80% of the cells were removed and the cells were filtered with the help of scanpy's library scanpy.pp.filter_cells [14]. Then the filtered data frames were labelled with 0's and 1's. Healthy patients were labelled with 0's, and Alzheimer's disease patients were labelled with 1's.

- **Normalization**

Prior to performing any type of analysis, the count data must be normalized because the sequencing depth causes the range of values for the features to differ. In order to give the values of highly expressed and lowly expressed genes equal weight, we performed the CPM (counts per million) normalization to our data and then performed a log transformation, this was done with the help of scanpy.pp.normalize_total [14] library.

- **Data Partition**

Initially, the dataset was split into 2 parts validation and training, with 3 validation samples from both Alzheimer's and Normal for validation. Finally, all the training data frames [15] were combined to form a large data frame containing the entries of both the classes of Alzheimer's and Normal.

2.2.4 mRMR Feature selection algorithm

Many statistical methods have been developed to identify the differentially expressed genes (DEG's) in cells. But the relationship between cells was ignored in such cases. The number of DEG which came out after DEG analysis were quite large in number to be applied as biomarkers. Therefore we used the method of mRMR gene selection algorithm [16]. mRMR stands for Minimum Redundancy and Maximum Relevance which is a feature selection algorithm which helps us to select features which have a high correlation with the class (output) and have a low correlation between themselves.

The main benefit of using mRMR is that it is designed to find the smallest relevant subset of features from the total given features i.e. to find the smallest subset of features which has the maximum predictive power. Whereas on the other hand the majority of other feature selection techniques uses an all-relevant approach in which they find out all

the features which have some or other relationship with the output classes. The mRMR approach takes into account both gene redundancy and correlations between genes and samples. Only the gene that is the most significant will be chosen if multiple genes are similar [17]. This method has been widely adopted and demonstrated to be useful for a variety of biological feature selection tasks, particularly in single-cell RNA-Seq analysis.

Because the single-cell sample data was vast and sparse, any other statistical technique yielded too many significant genes, which was not a good criterion for selecting biomarkers. As a result, the mRMR technique was best suited to selecting thebest subset of a minimal number of non-redundant biomarkers for single-cell data analysis.

We applied a K value of 100 to select the top 100 ranked features identified by mRMR from the total 5401 expressed genes in the dataset. These 100 genes were then used to classify the cells based on their category of diseased and non-diseased using different Machine Learning and Deep Learning techniques.

2.2.5 Machine Learning Models

Various machine Learning models have been developed which help to classify the data into categories. In bioinformatics, a variety of strategies have been used to solve classification problems. Here, we've selected methods like Decision Trees, Random Forest, extra tree classifiers, logistic regression, K neighbors' classifier and neural networks.

a) **Artificial Neural network: -** This approach is inspired by biological neuron net- works [18]. They are made up of numerous layers, each of which contains several nodes(or neurons) that aid in decision-making. Each node is initialized with a random weightat the start, which is then fine-tuned after each iteration and set to the best-suited value as the learning process goes on [19]. The final output is the predicted label (Diseased or Normal) of the sample.

b)**Logistic Regression: -** It is similar to the linear regression algorithm, which converts the probability of prediction of each sample into a yes/no decision [20].It makes use of the idea of the sigmoid function, which is used to calculate the likelihood that a sample belongs to a specific class. [21].

c) **Decision trees**: - A decision tree is a decision-making method that presents possibilities and their results in the form of a tree [22]. Decision tree utilizes the concept of entropy and information gain to make classification choices [23].

d)**Random Forest**: - The random forest technique refers to a categorization system composed of several decision trees [24]. It employs the ideas of bagging and feature randomization in an attempt to generate an independent and identically distributed forest of trees whose prediction is more accurate than that of any individual tree [25].

e) **Extra Tree Classifiers**: - This is an ensemble-based decision tree classifier. It combinesthe output of multiple de-correlated trees. Finally, we arrive at the classification output thatis required [26]. The main ideology is similar to that of Random forests, but decision trees are constructed in a different manner.

f) **K neighbors classifier**: - The K-NN approach keeps all current data and classifies newdata points based on similarity. This means that when new data is created, it may be swiftly classified into a suitable category using the K- NN approach [27]. It is a non-parametric algorithm.

2.2.6 Evaluation Parameter

To evaluate how good our classifier is, several metrics have been chosen to show the results. Our performance is evaluated on 5-fold cross-validation. Metrics used for evaluation are:

a) **Accuracy**: It tells how many of our predictions are correctly predicted in the dataset.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

b) **Sensitivity**: It is also called the true positive rate. It is expressed as the ratio of the number of times a sample was classified as positive when it was actually positive to the total number of positive samples.

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

c) **Specificity**: The true negative rate is a different name for it. It is calculated as the ratio of the total number of negative samples to the number of times a sample was incorrectly labeled as negative.

$$Specificity = \frac{TN}{FP + TN} \tag{3}$$

d) **Precision**: It is also called the Positive Predictive value. It is expressed as the ratio of a total number of times a sample was classified as positive when it was actually positiveto the total number of times the classifier labeled a sample positive.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

e) **ROC-AUC**: By graphing the True positive rate and the False positive rate, the Receiver Operating Characteristics (Area Under Curve) visual tool helps to demonstrate the predictive ability of a classifier.

$$TruePositiveRate(TPR) = Recall = \frac{TP}{TP + FN} \tag{5}$$

$$FalsePositiveRate(FPR) = 1 - Specificity = \frac{FP}{TN + FP} \tag{6}$$

f) **F1 Score**: A classifier's precision and recall are combined into one metric by theF1-score in statistics of the classification model by calculating their harmonic means.

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{7}$$

(where TP, FN, FP, and TN stand for true positive, false negative, false positive, and truenegative, respectively.)

2.2.7 Cross- Validation

The dataset was divided in an 80:20 ratio, with training data accounting for 80% and independent validation data accounting for the remaining 20%. During the 5-fold cross-validation procedure, the training data was further separated into training and testing datasets, and the mean of the findings for each fold of the cross-validation was recorded. The total training data is split into five equivalent folds in the 5-fold cross-validation method, with four folds used for training and the fifth fold utilized for testing. The entire method is iterated five times, with each fold having a chance to be utilized as testing data. This is a common practice in many studies [28, 29].

2.2.8 Model Architecture

For this study, we have prepared a customized Artificial Neural Network Model to classify samples based on their diagnosis. The neural Network consists of an input layer, 3 hidden layers and an output layer. Also, a dropout of 0.3 is done at each step to reduce the over fitting of neural networks. This approach is inspired by biological neuron networks. An ANN is composed of a network of interconnected systems or nodes known as artificial neurons, which are generally designed like neurons in the human brain [18]. They are made up of numerous layers, each of which contains several nodes (or neurons) that aid in decision-making [19]. The final output is the predicted label (Diseased or Normal) of the sample.

The Neural Network Architecture is shown in the figure below.

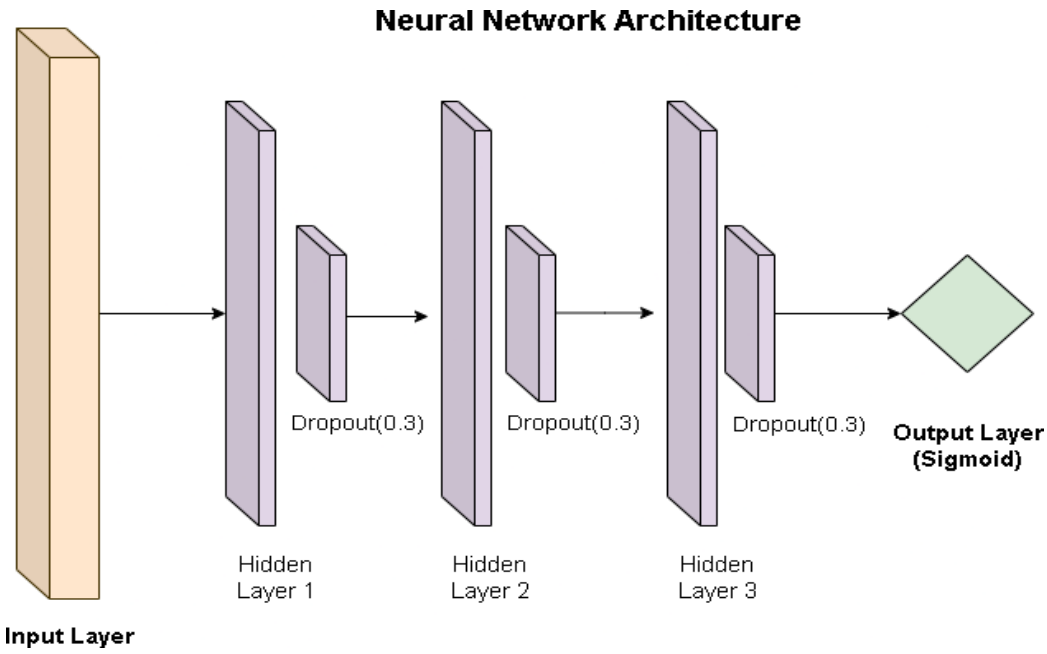

Figure 2.2: The diagram shows ANN model architecture with 1 input layer, 3 Hidden Layers, 3 dropout layers and 1 output layer

2.2.9 Packages and Tools

The entire data analysis and prediction pipeline were coded in Python 3.9.13 using 'scikit-learn' (sklearn) [30] library for machine learning and mRMR classification library [16] was used for feature selection. 'Pandas' [15] library for loading and preprocessing of data and multiple scanpy[14] libraries for preprocessing and filtering. Also, the library used to build ANN model was Tensor flow [31] and Keras [32].

2.3 Biomarkers Optimization

2.3.1 Incremental feature selection

Feature selection is typically used to decrease a large number of biological characteristics in order to establish a robust data-independent classification or regression model [33]. The disease diagnosis panel development are heavily dependenton the efficiency of the feature selection technologies.

An Incremental Feature Selection (IFS) algorithm is used to evaluate the performance of top n ranked features where n = (1, 2, 3, 4 n), where n denotes the total number of given features [34].

After the application of mRMR on ANN classifiers, we still did not know the optimal number of features to select as biomarkers. We adopted the IFS method to optimize the selected features (genes), which could act as potential biomarkers with highpredictive power.

After every iteration, a new feature was added to the previous feature set and a new set was obtained and fed to the ANN classifier. Then new ANN classifiers were built, and labels were predicted and accuracy was determined by using five-fold cross-validation. The IFS curve was plotted for all the combinations of gene set (features) on the x-axis, and the accuracy value was obtained from all sets on the y-axis as shown in plot 3.4.

CHAPTER 3

RESULTS

3.1 Analysis on top 100 genes

mRMR feature selection algorithm was applied to find out the actual discriminative effect of important features based on iteration. Our main objective was to find out the features (genes) that were directly related to the sample classes and were not redundant with all the other features present. Using the mRMR method the top 100 genes we obtained are shown in the table 3.1.

Table 3.1: Top 100 genes obtained by mRMR

Rank	Genes	Rank	Genes	Rank	Genes	Rank	Genes
01	ARL17B	26	HIBADH	51	DDX3X	76	APOD
02	NAIP	27	ZBED5	52	NSL1	77	KIF9-AS1
03	BCOR	28	PTDSS2	53	TMED10	78	TYW1
04	XIST	29	ATG4B	54	UGT8	79	GNAI2
05	TSC22D4	30	PWWP2A	55	SNX1	80	BAZ1B
06	HEPACAM	31	XRRA1	56	ALG13	81	MBTPS1
07	FGF17	32	OTUD7B	57	LINC00320	82	CDH4
08	EZH1	33	SCD	58	RAD9A	83	RAB40B
09	FOXN2	34	UBE2Z	59	RGS12	84	SPP1
10	NDUFAF6	35	PIGQ	60	ST13	85	GPBP1L1
11	CC2D1A	36	PCMTD2	61	PTN	86	FSCN1
12	MARCKSL1	37	COL4A5	62	USP8	87	CAPZA1
13	ZDHHC11B	38	ARFIP1	63	EDF1	88	SPPL2B
14	PLXNB1	39	CCND3	64	SLCO1A2	89	MED15
15	PLPPR2	40	FOXK2	65	NUP153	90	C1GALT1
16	AC090517.4	41	CPOX	66	SYNRG	91	ITGB3BP
17	CDK18	42	STXBP3	67	EIF3E	92	CCDC82
18	LGI4	43	ITPKB	68	LPCAT4	93	CDK12
19	CHD7	44	TBCB	69	ARMCX4	94	EGLN1
20	RBMX	45	SRSF10	70	PREX2	95	CCDC57
21	CDKL1	46	SPTLC2	71	APC2	96	SEL1L
22	DNAJC7	47	LYPLAL1	72	SLC38A9	97	CHORDC1
23	SLC25A13	48	FAM107B	73	UTP23	98	ATG10
24	PER1	49	PDIA2	74	RCN2	99	C4orf48
25	LPAR1	50	C1orf61	75	PRR14L	100	AC097103.2

a) **Training: -** After training Machine Learning and Deep Learning models on these 100 genes modelsperformance are shown in the table below in table 3.2.

Model Name	Accuracy	Sensitivity	Specificity	AUC-ROC	Precision	f1-Score	Mis-Classification
Decision Tree	0.9953	0.99	1.00	0.9948	1.00	0.99	0.001
Random Forest	0.98	0.96	1.00	0.9784	1.00	0.98	0.02
Logistic Regression	0.9953	0.99	1.00	0.9948	1.00	0.99	0.001
XGBClassifier	0.99	0.99	1.00	0.9948	1.00	0.99	0.00
ExtraTree Classifier	0.9953	0.99	1.00	0.9948	1.00	0.99	0.00
K Neighbors classifier	1	0.99	1.00	0.9948	1.00	0.99	0.00
Deep Learning Model	0.99	0.99	1.00	0.9948	1.00	0.99	0.00

Table 3.2: Training results on top 100 genes

All the models were trained with very high accuracy i.e. 99% but Machine learning techniques such as logistic regression, decision trees etc failed on predicting the samples with low accuracy. Whereas, compared to other Machine Learning techniques Deep learning technique performed significantly well in the prediction part as compared to other models.

b) Validation:- The results achieved were 99% accuracy on the training dataset and 82% accuracy on testing dataset. AUC-ROC achieved on the validation set was 0.84 on Deep learning ANN model. The external validation data results are shown in the table 3.3.

As it can be seen in the table that Deep Learning ANN model outperforms all other models with accuracy of 0.82, Sensitivity 0.86, Specificity 0.77, AUC-ROC 0.84 on

Models	Accuracy	Sensitivity	Specificity	AUC-ROC	Precision	f1-Score	Mis-Classification
Decision Tree	0.40	0.97	0.02	0.49	0.02	0.04	0.59
Random Forest	0.41	0.89	0.09	0.49	0.09	0.16	0.58
Logistic Regression	0.38	0.68	0.18	0.43	0.18	0.28	0.62
XGBClassifier	0.47	0.36	0.55	0.45	0.55	0.43	0.53
ExtraTree Classifier	0.43	0.85	0.13	0.50	0.13	0.23	0.57
K Neighbors classifier	0.37	0.33	0.39	0.36	0.39	0.36	0.63
Deep Learning Model	0.82	0.86	0.77	0.84	0.82	0.80	0.18

Table 3.3: Validation results on top 100 genes

the validation dataset, whereas other models fail to predict sample labels with very low accuracy. A graphical representation of the training and validation accuracy of all the models on the top 100 genes identified by mRMR is shown in the chart below 3.1.

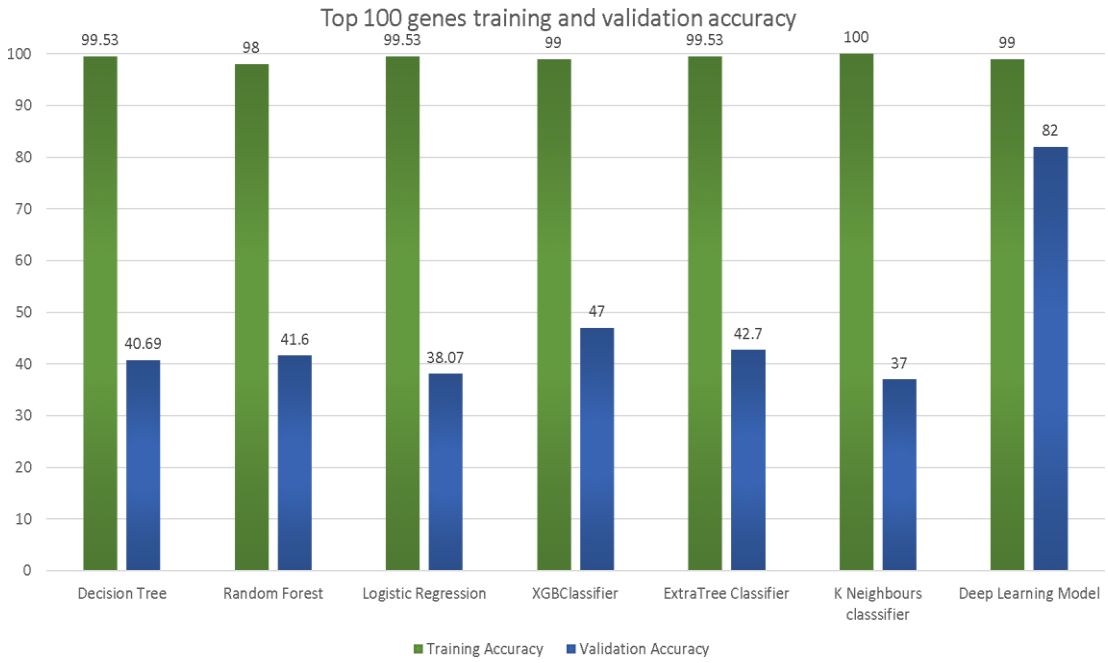

Figure 3.1: Training and validation accuracy of all models on top 100 genes

3.2 Patient wise analysis on top 100 genes

Since we had the data in a patient-wise format, we decided to analyze how accurately our best-selected Deep Learning Model (ANN) could predict the patient's diagnosis. Each patient's full single-cell RNA seq profiling was given to the model to predict the percentage of 1's and 0's in it. 1's label denotes the number of diseased cells, and 0's denotes the number of normal cells. The table 3.4 below shows patient-wise obtained results via the ANN Deep learning model.

Patient Name	Predicted Diseased Cells	Predicted Normal Cells
Alzheimer Test 1	43.00%	57.00%
Alzheimer Test 2	83.27%	16.70%
Alzheimer Test 3	83.97%	16.02%

Table 3.4: Alzheimer's Test patient's analysis results on top 100 genes

Patient Name	Predicted Diseased Cells	Predicted Normal Cells
Normal test 1	31.00%	69.00%
Normal Test 2	23.21%	76.80%
Normal Test3	22.98%	77.01%

Table 3.5: Healthy Test patient's analysis results on top 100 genes

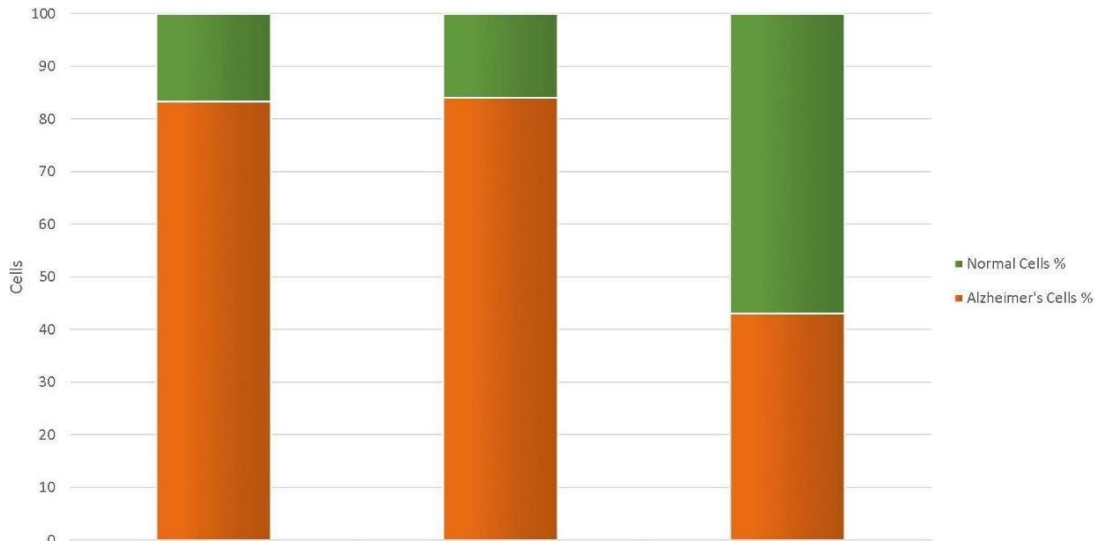

Figure 3.2: Diagrammatic representation of normal and diseased cells in Alzheimer's Patients

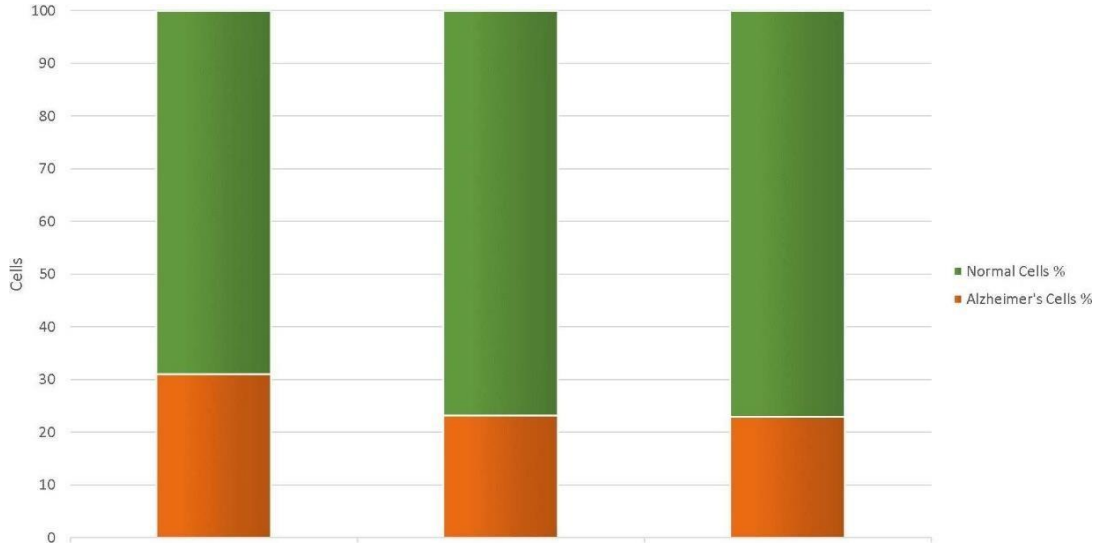

Figure 3.3: Diagrammatic representation of Normal and diseased cells in Normal Patients

A graphical representation of the amount of predicted diseased cells and normal cells for each sample of Alzheimer's test and Normal Test patient is shown in Figure

3.2 and 3.3

3.3 Biomarkers Optimization - Incremental Feature Selection

Based on the top 100 mRMR extracted genes, we prepared 100 different ANN classifiers and applied an Incremental Feature selection method to find out the best optimum number of genes which could act as biomarkers.

Initially, the top 1st gene was taken to train the model and then accuracy was determined, then top 2 genes were taken to train the model and similarly accuracy of this model was determined. Subsequently, models were trained with more subsets of top 3, top 4, top 100 genes to determine the accuracy of all the subsets. The accuracy values are shown in the table 3.6.

Geneset(TOP)	Accuracy values	Geneset(TOP)	Accuracy values	Geneset(TOP)	Accuracy values	Geneset(TOP)	Accuracy values
01	0.685286	26	0.948541	51	0.9765	76	0.9852
02	0.685286	27	0.952565	52	0.9765	77	0.9857
03	0.753898	28	0.952565	53	0.9767	78	0.9861
04	0.753898	29	0.957285	54	0.9771	79	0.9866
05	0.803021	30	0.957285	55	0.978	80	0.9876
06	0.803021	31	0.957285	56	0.978	81	0.989
07	0.838826	32	0.960116	57	0.9781	82	0.9891
08	0.838826	33	0.960116	58	0.9793	83	0.9896
09	0.848441	34	0.964502	59	0.9796	84	0.9898
10	0.859855	35	0.964502	60	0.9796	85	0.9909
11	0.871953	36	0.964502	61	0.98	86	0.9909
12	0.878375	37	0.966266	62	0.9808	87	0.9911
13	0.893185	38	0.966266	63	0.9811	88	0.9913
14	0.893185	39	0.969948	64	0.9812	89	0.9916
15	0.906203	40	0.969948	65	0.9818	90	0.9921
16	0.906203	41	0.971426	66	0.9819	91	0.9924
17	0.918308	42	0.971426	67	0.9819	92	0.9924
18	0.917066	43	0.97365	68	0.9829	93	0.9935
19	0.928251	44	0.97365	69	0.9834	94	0.9936
20	0.928251	45	0.975414	70	0.9839	95	0.9938
21	0.935705	46	0.975414	71	0.9839	96	0.9944
22	0.935705	47	0.975414	72	0.9841	97	0.9945
23	0.937469	48	0.97639	73	0.9844	98	0.9959
24	0.943465	49	0.97639	74	0.9845	99	0.996
25	0.948541	50	0.976014	75	0.9848	100	0.9969

Table 3.6: IFS accuracy values on top 100 genes

A graphical representation of Accuracy vs Gene subsets using IFS is shown in figure 3.4. Out of the total top 100 genes returned by mRMR, the optimal value of biomarkers

found out was 35. Because after 35 genes the curve starts to flatten out and no significant increase in training accuracy is observed. Thus the optimal 35 genes are shown in Table 3.7

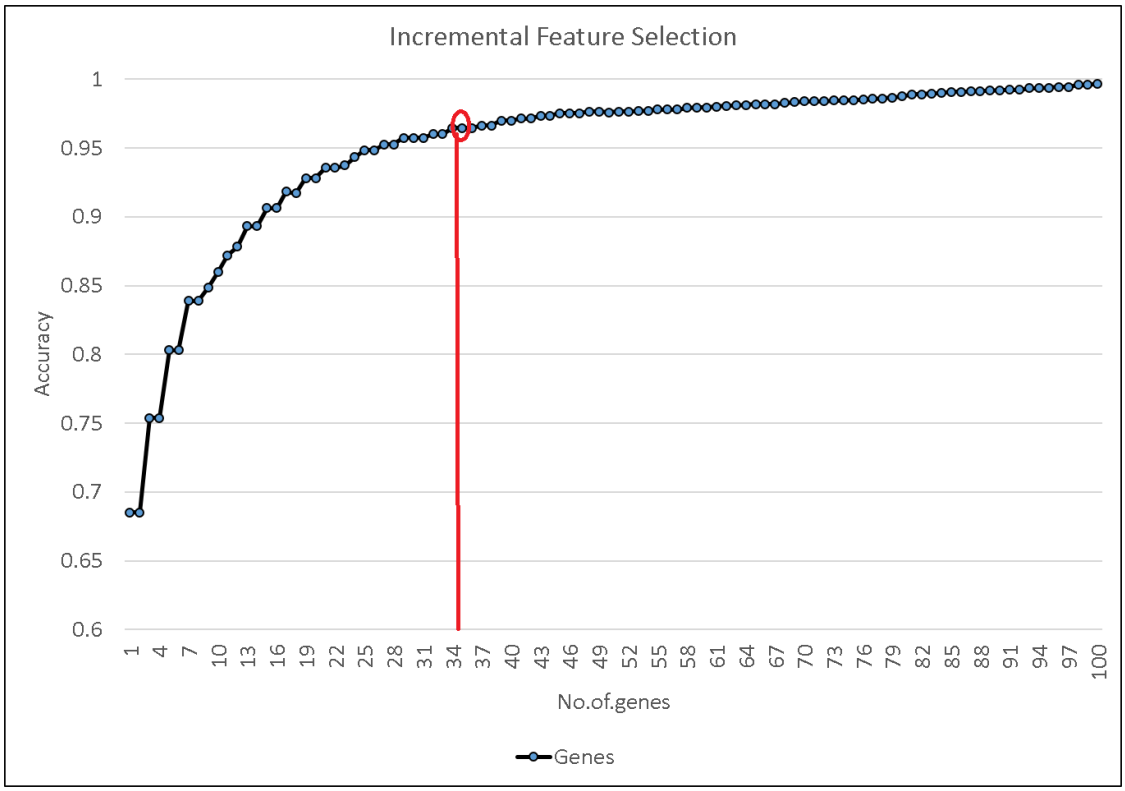

Figure 3.4: IFS accuracy graph representing accuracy vs gene subsets, out of which 35 genes is selected as optimal.

Rank	Gene	Rank	Gene	Rank	Gene
01	ARL17B	13	ZDHHC11B	25	LPAR1
02	NAIP	14	PLXNB1	26	HIBADH
03	BCOR	15	PLPPR2	27	ZBED5
04	XIST	16	AC090517.4	28	PTDSS2
05	TSC22D4	17	CDK18	29	ATG4B
06	HEPACAM	18	LGI4	30	PWWP2A
07	FGF17	19	CHD7	31	XRRA1
08	EZH1	20	RBMX	32	OTUD7B
09	FOXN2	21	CDKL1	33	SCD
10	NDUFAF6	22	DNAJC7	34	UBE2Z
11	CC2D1A	23	SLC25A13	35	PIGQ
12	MARCKSL1	24	PER1		

Table 3.7: Optimal 35 genes selected by IFS

3.3.1 Analysis on 35 selected Biomarkers

a) **Training**: - The genes mentioned in the above table 3.7 were used to train multiple Machine Learning and a Deep Learning (ANN) model. After training, the performance of training of models are shown in the table below in table 3.8.

Models	Accuracy	Sensitivity	Specificity	AUC-ROC	Precision	f1-score	Mis-Classification
Decision Tree	0.96388	1.00	0.93	0.967	0.93	0.97	0.04
Random Forest	0.9593	0.91	1.00	0.9551	1.00	0.95	0.04
Logistic Regression	0.96388	1.00	0.93	0.967	0.93	0.97	0.04
XGB Classifier	0.9638	1.00	0.93	0.967	0.93	0.97	0.04
Extra tree Classifier	0.96	0.91	1.00	0.9551	1.00	0.95	0.04
Deep Learning Model	0.964	1.00	0.93	0.967	0.93	0.97	0.04

Table 3.8: Different models training results on 35 selected genes

b) **Validation**:- All the models were trained with very high accuracy i.e. 96% but Machine learning techniques such as logistic regression, decision trees, Random Forest etc. failed to predict the samples and showed low prediction accuracy. Whereas, compared to other Machine Learning techniques, the Deep learning technique performed significantly well in the prediction part compared to other models. The results of Validation accuracy can be shown in table 3.9

Models	Accuracy	Sensitivity	Specificity	AUC-ROC	Precision	f1-score	Mis-Classification
Decision Tree	0.42	0.92	0.07	0.50	0.07	0.13	0.58
Random Forest	0.42	0.81	0.15	0.48	0.15	0.25	0.58
Logistic Regression	0.38	0.62	0.4	0.38	0.4	0.38	0.62
XGB Classifier	0.46	0.54	0.41	0.48	0.41	0.47	0.54
Extra tree Classifier	0.46	0.5	0.43	0.47	0.43	0.46	0.54
Deep Learning Model	0.74	0.75	0.66	0.75	0.66	0.71	0.34

Table 3.9: Different models validation results on 35 selected genes

As it is significant from the above-mentioned Table 3.9 that the Deep learning ANN model outperforms all other models in the validation assessment of samples. Thus ANN outperforms all other machine learning models. The ANN model displays significant accuracy of 0.74, Sensitivity 0.66, AUC- ROC 0.75 etc.

Also, A graphical representation of the comparison of training and validation accuracy of all models on the 35 selected genes by IFS is shown in the chart 3.5

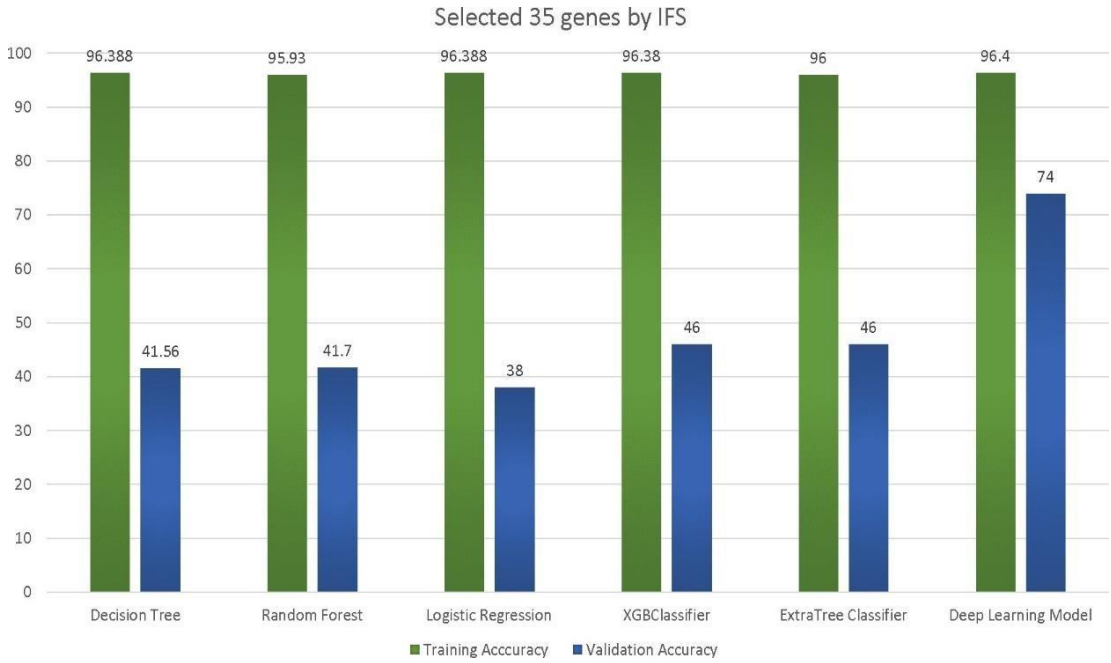

Figure 3.5: Training and validation accuracy comparison of all models on 35 genes

3.3.2 Patient-wise Analysis on 35 genes

Now, each patient's sample data was taken and the optimal 35 genes which were selected as biomarkers were extracted from each sample and fed to the model to check the percentage of diseased and healthy cells in the sample which could help us get an insight about the patient's diagnosis. The percentage of cells predicted in each test sample was calculated. The results of patient-wise cell prediction are shown in Table 3.10 below.

Patient Name	Predicted Diseased Cells	Predicted Normal Cells
Alzheimer Test 1	50.41%	49.58%
Alzheimer Test 2	83.98%	16.01%
Alzheimer Test 3	83.97%	16.02%

Table 3.10: Alzheimer's Test patient's analysis results on 35genes

A graphical representation of the amount of predicted diseased cells and normal cells for each sample of Alzheimer's test and Normal Test patient with the selected 35

Patient Name	Predicted Diseased Cells	Predicted Normal Cells
Normal test 1	43.00%	57.00%
Normal Test 2	31.00%	69.00%
Normal Test3	25.17%	74.80%

Table 3.11: Healthy Test patient's analysis results on top 35 genes

genes is shown in the Figure 3.6 and 3.7.

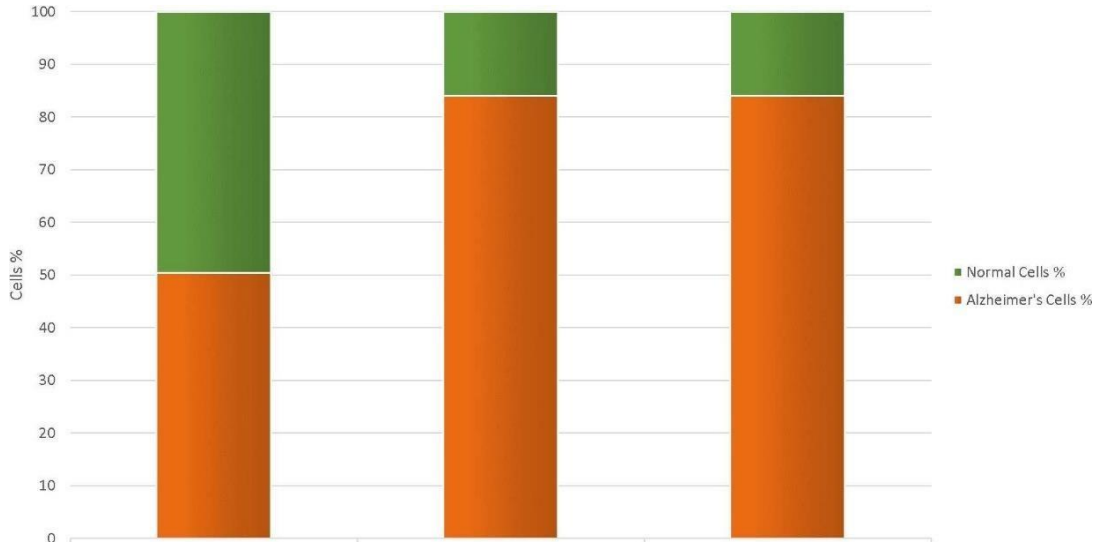

Figure 3.6: Diagrammatic representation of normal and diseased cells in Alzheimer's Patients (35 genes)

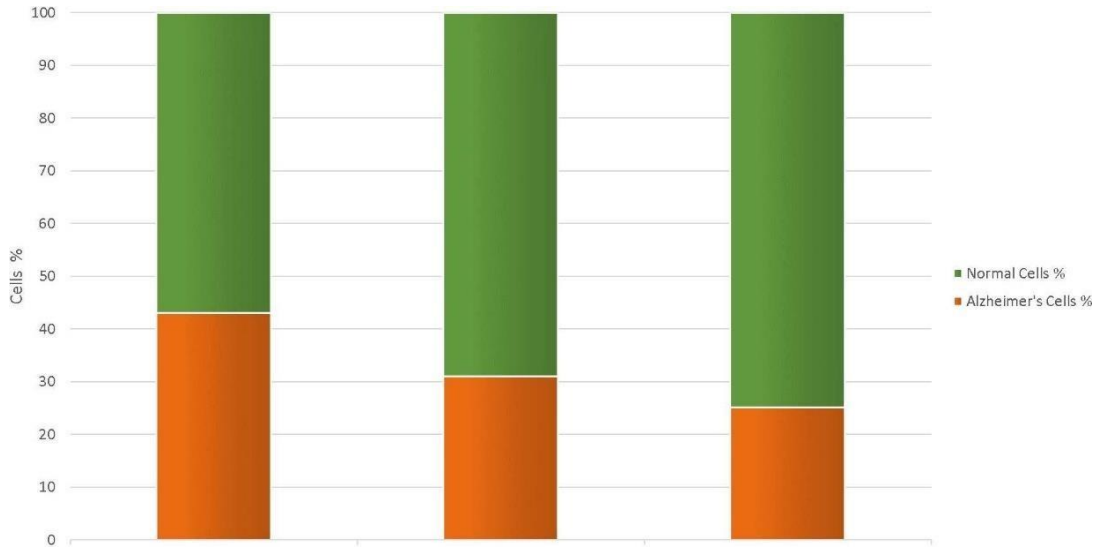

Figure 3.7: Diagrammatic representation of Normal and diseased cells in Normal Patients (35 genes)

3.4 Data Visualization of Samples

Since the data present with us is in form of a high-dimensional space, thus in order to represent it into a low-dimensional space with the goal of keeping the low-dimensional representation as near as possible to the actual dimension of the original data we have tried to visualize them using two different techniques namely tsne and umap [35]. Because sample tissues are typically a combination of disease-affected and normal cells, sample purity may result in misclassifications. So, we have selected the 35 genes from samples and plotted both tsne and Umap visualization of the data points in 2D and3D.

a) **t-SNE**:- t-Distributed Stochastic Neighbor Embedding (t-SNE) is a dimensionality reduction approach that is well known for visualizing large datasets. It represents data by giving each data point a location in a two or three-dimensional map. [36]. The t-SNE 2-Dimensional and 3-Dimensional Visualization is shown in figure 3.8 below. As we can see 2d graph shows a clear separation between both the classes of Diseased (Alzheimer's) cells and Normal (Healthy) cells. Whereas in the 3- dimensional representation we can see some overlap between the cells. As some of the data points seem to be merged with each other.

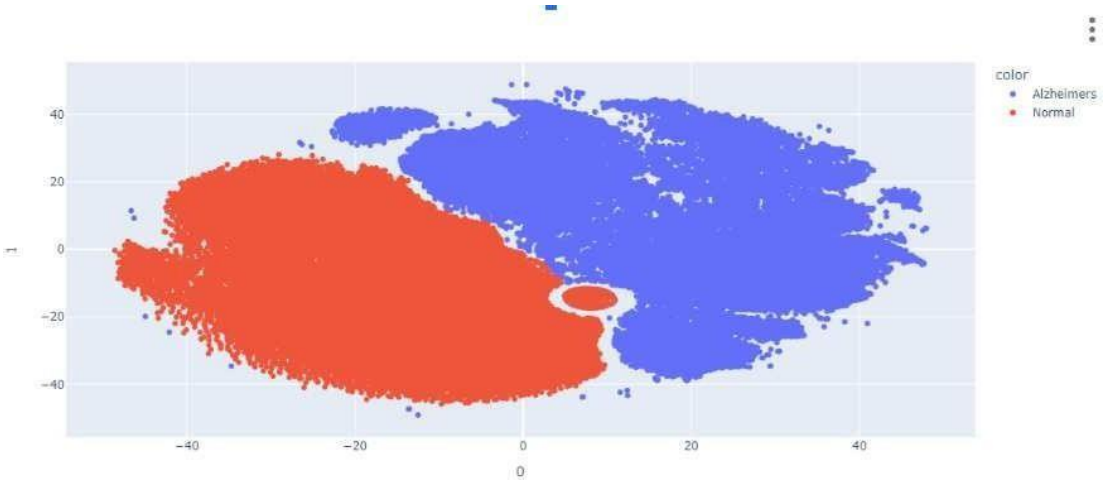

Figure 3.8: 2D visualization using tsne of both classes

b) **UMAP**: - UMAP stands for uniform manifold approximation and projection [37]. It is a dimension reduction approach that like t-SNE, may be used for visualization as well as generic non-linear dimension reduction. Despite having certain benefits over tSNE in terms of separating batch effects, recognizing pre-

defined biological groups, and exposing in-depth clusters in two-dimensional space, UMAP outperforms PCA

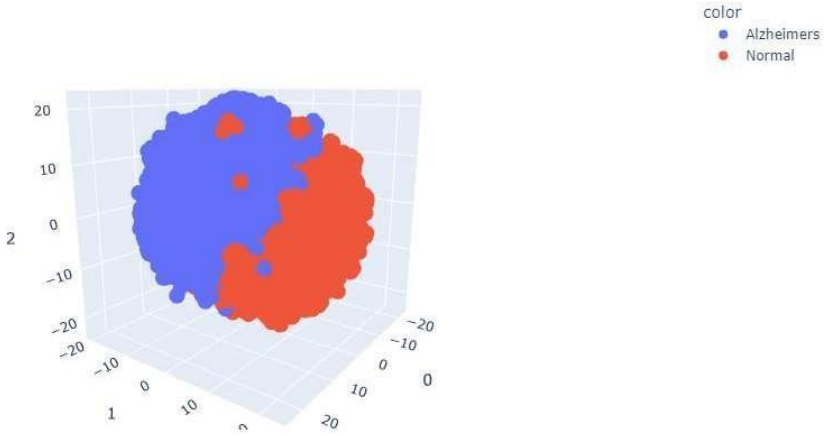

Figure 3.9: 3D visualization using tsne of both classes

and MDS. UMAP's sample clustering is important since it exposes biological traits and clinical importance [37].

UMAP 2d and 3d representations of the dataset on the selected 35 genes have been developed. We can clearly see the separation of data based on the selected 35 genes which could act as biomarkers.

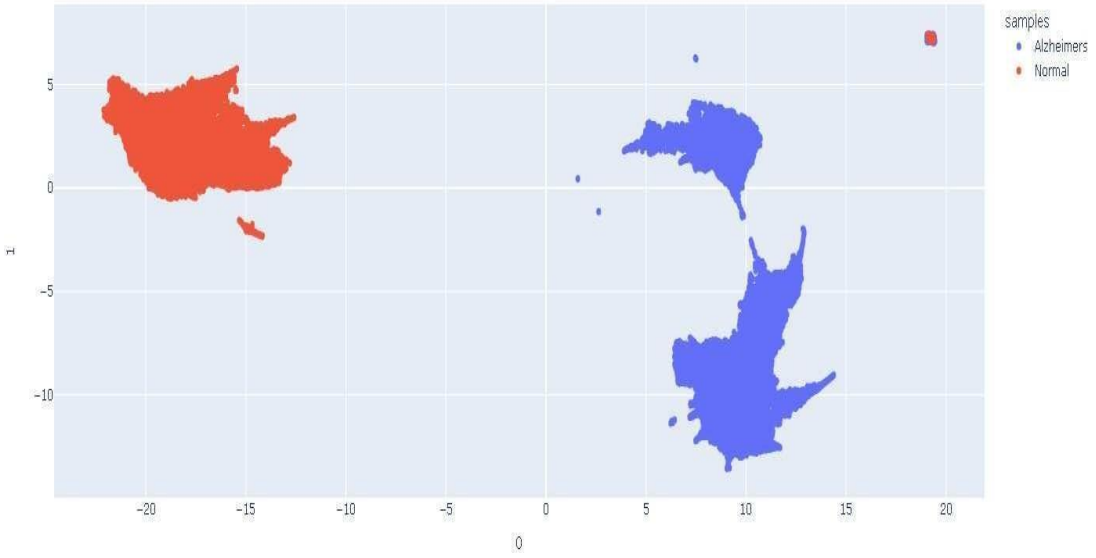

Figure 3.10: 2D visualization using umap of both classes

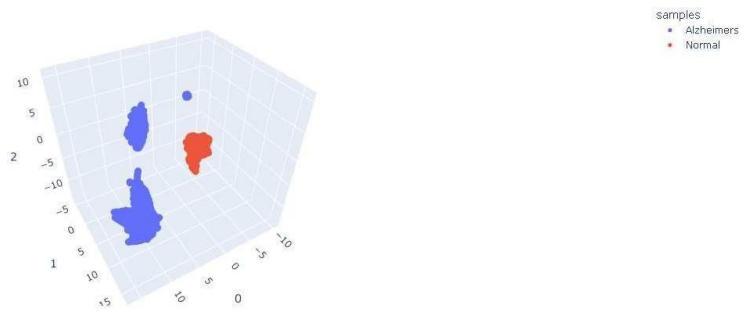

Figure 3.11: 3D visualization using umap of both classes

3.5 Biological Functions of the Selected Genes

After completion of analysis by mRMR, 35 genes were extracted which could serve as potential diagnostic biomarkers of Alzheimer's. We then performed Gene Ontology (GO) Enrichment Analysis on these extracted 35 genes to map the biological functions of the selected genes. The Go enrichment analysis results are shown in the Table 3.12.

GO term	Activity	Genes
(GO:0140657)	ATP-dependent activity	CHD7
(GO:0005488)	Binding	CHD7, ARL17B, UBE2Z, FGF17, OTUD7B, CDKL1, MARCKSL1, FOXN2, NAIP, PER1, CC2D1A, AC090517, LPAR1, PWWP2A, CDK18
(GO:0003824)	Catalytic Activity	ZDHHC11B, CHD7, UBE2Z, HIBADH, OTUD7B, CDKL1, NAIP, CDK18, PLPPR2
(GO:0098772)	Molecular Function Regulator	UBE2Z, FGF17, NAIP
(GO:0060089)	Molecular Transducer Activity	PLXNB1, FGF17, LPAR1
(GO:0140110)	Transcription Regulator Activity	BCOR, FOXN2, CC2D1A, AC090517
(GO:0005215)	Transporter Activity	SLC25A13

Table 3.12: Gene ontology Enrichment analysis results

As we can see, most of the Genes are involved in the Binding Activity and catalytic activity of various metabolic processes. Other activities associated with reported genes are ATP-dependent activity, Molecular function regulator, Molecular Transducer activity, transcription regulator activity, and transporter activity.

CHAPTER 4

PIPELINE PACKAGING

The Deep Learning model (ANN) trained on the training set was saved, and the next step of packaging the code into a proper pipeline was performed. Python packaging includes a total collection of all essential requirements and prerequisites to run the code efficiently on any system containing python. The pipeline was built so the user could input their respective data and get the output whether the patient is "Diseased" or "Normal/Healthy". The python package was uploaded on https://www.pypi.org, and the package was named "AlzScPred".

4.1 Directory Structure

The directory is the place/folder where all the prerequisites and requirements are kept. The structure of the directory is as shown in the chart below 4.1. In place of the mentioned YOUR _USERNAME _HERE, mention the name of the package you want to create. The name of the package should be unique such that it does not conflict with people that have uploaded previously with the same name. Also, make sure the directory/ folder consisting of python scripts should have the same name as the project. This simplifies the configuration and installation for new users. Also, an initialization file for every folder of the project is necessary, naming it as init .py, which can also be an empty file. It is required to import the folder as a package.

4.2 Package Requirements

Along with the python script the package directory also needs to have certain important configuration files for easy installation and processing of the package on any user system. The other required files are as follows:

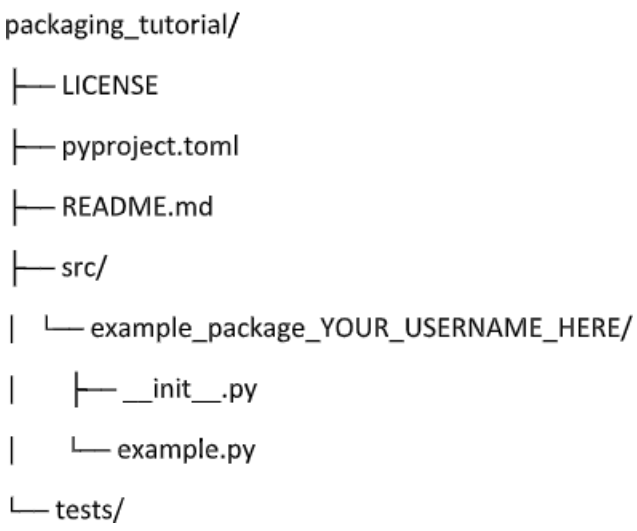

Figure 4.1: Package directory structure example

- Setup.py
- Manifest.in
- README.md
- LICENSE.txt
- Example files
- test files

The setup.py file consists of all the important configuration details and prerequisites for the software installation. It consists of data about the required packages needed before installation, version number, description of the package, author name, contact etc. Whereas, the Manifest.in the file is required to include extra important files in our source distribution because when we build a python package by default only a minimal set of files are included in the source distribution. So in order to include extra files and let them be recognized by the system a MANIFEST.in file containing all folder names and codes is required. README.md file contains information about the project and its detailed description for other users so that they it works as a guide for the project. It is a file containing information for the users. Similarly, a LICENSE.txt file is required which describes the copyrights, licenses, and restrictions which apply to that code.

4.3 Packaging process

After preparation of the package directory with all the requirements. The package needs to be converted into a .tar file and .whl (wheel) file with the below commands.

```
$ python3 setup.py sdist bdist_wheel
```

Figure 4.2: Wheel file creation code

This command will create a wheel file for the source distribution of the package and a tar file containing all the files of the package in the tar format. After preparation of the wheel file, you need to have twine installed which is required to upload the whl and tar file on www.pypi.org. The command file used is shown below. It will require the PyPI credential to log in and upload to the PyPI account.

```
$ twine upload dist/*
```

Figure 4.3: Upload on PyPI using twine

4.4 Package Details

4.4.1 AlzScPred

It is a computational approach tool to predict Alzheimer's affected patients from their single-cell RNA seq data using Deep Learning. This tool aims to use Artificial Neural Network (Deep Learning) model to classify Normal Control (NC) patients and Alzheimer's disease (AD) patients from their single-cell RNA seq data. It takes 10x single cell genomics data as input and predicts whether the patient is diseased or healthy with the help of a highly trained model.

An excellent feature selection method called mRMR (Minimum Redundancy Maximum Relevance) was used to find out the top 100 features for classification. Followed by Incremental Feature Selection (IFS) which led to the identification of 35 conserved genes which act as promising biomarkers in the classification and predictionof Normal and Diseased patients.

4.5 Installation

The user can download the package on any python environment with Python 3.0 or higher with the following command. It can also be found directly on www.pypi.org site. The screenshot of the PyPI is shown below.

```
$ pip install AlzScPred
```

Figure 4.4: pip install command to install package

Also in case the tool is previously installed, the tool can be upgraded to the latest version using the command.

```
$ pip install --upgrade AlzScPred
```

Figure 4.5: pip upgrade command to upgrade package

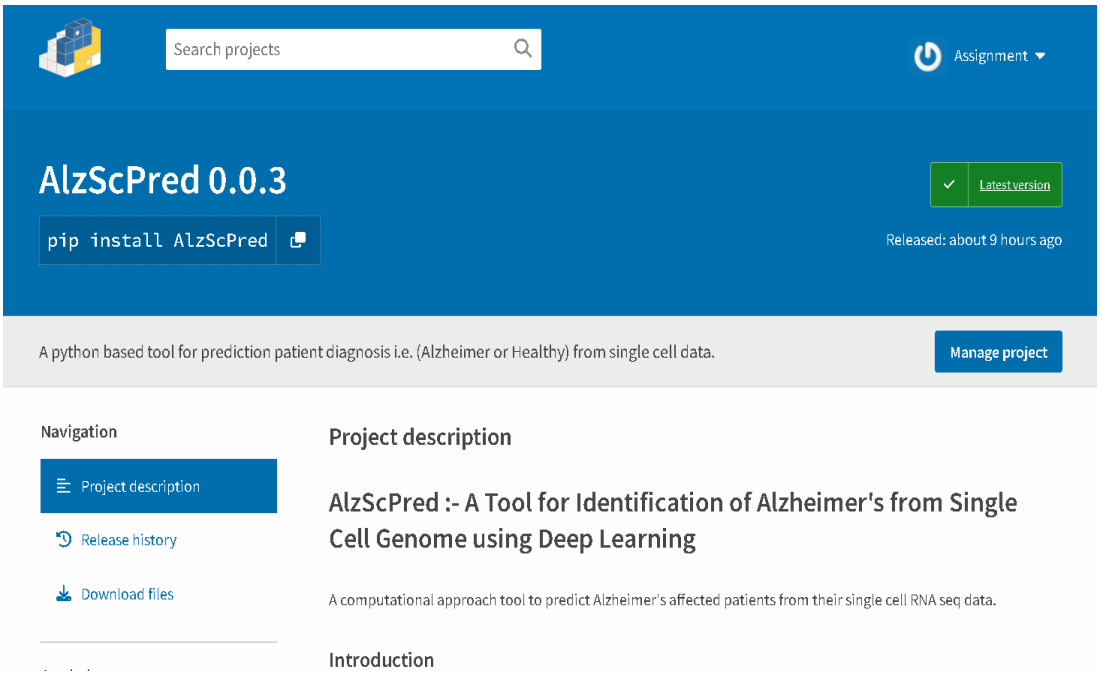

Figure 4.6: PyPI screenshot of package

4.6 Usage

After installation of the AlzScPred package in your python environment. Import the library using the below code. The AlzScPred comes with 1 inbuilt module.

- Prediction Module

Please import the modules in your python environment before executing the code below.

```
$ import AlzScPred
```

Figure 4.7: Python code to import package

```
$ from AlzScPred import Validation
```

Figure 4.8: Python code to import Validation Module

4.6.1 Input

The input file should be in the form of a data frame in which the columns should be features (genes) and the rows should be cells. The file should contain the read count data of each gene in each cell.

	Unnamed: 0	MIR1302-2HG	FAM138A	OR4F5	AL627309.1	AL627309.3	AL627309.2	AL627309.4	AL732372.1	OR4F29	...	AC007325.2	BX072566.1	AL3
0	AAACCCACATCTTAGG.1	0	0	0	0	0	0	0	0	0	...	0	0	
1	AAACGAACAACCACGC.1	0	0	0	0	0	0	0	0	0	...	0	0	
2	AAACGCTCATGACGGA.1	0	0	0	0	0	0	0	0	0	...	1	0	
3	AAACGCTGTATCTCGA.1	0	0	0	0	0	0	0	0	0	...	0	0	
4	AAACGCTGTATTCCTT.1	0	0	0	0	0	0	0	0	0	...	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
195	ACAAAGAAGGAATCGC.1	0	0	0	0	0	0	0	0	0	...	0	0	
196	ACAAAGAGTACTCGAT.1	0	0	0	0	0	0	0	0	0	...	0	0	
197	ACAAAGAGTATTGCCA.1	0	0	0	0	0	0	0	0	0	...	0	0	
198	ACAAAGAGTTGACTGT.1	0	0	0	0	0	0	0	0	0	...	0	0	
199	ACAAAGATCCGTGGCA.1	0	0	0	0	0	0	0	0	0	...	0	0	

200 rows × 33539 columns

Figure 4.9: Example of input file

4.6.2 Demo

The demo to run the python package is shown in the figure below:

```
import pandas as pd
df = pd.read_csv("Your file path here")

# Prediction:- Execute the code below to get the output. It takes 1 argument i.e the dataframe with features in

Validation.predict_patient(df)
```

Figure 4.10: Code Demo

Note: Please make sure that your single-cell data file is prepared in the above example.csv format. And the file should also contain the read count data for the selected 35 genes in the above 35_genes.txt file. Which can be found at https://webs.iiitd.edu.in/raghava/alzscpred/

4.6.3 Output

The output of the file can be obtained by the code as shown below. It will display the patient diagnosis i.e. Diseased or Healthy with the amount of Diseased or healthy cells found in the patient.

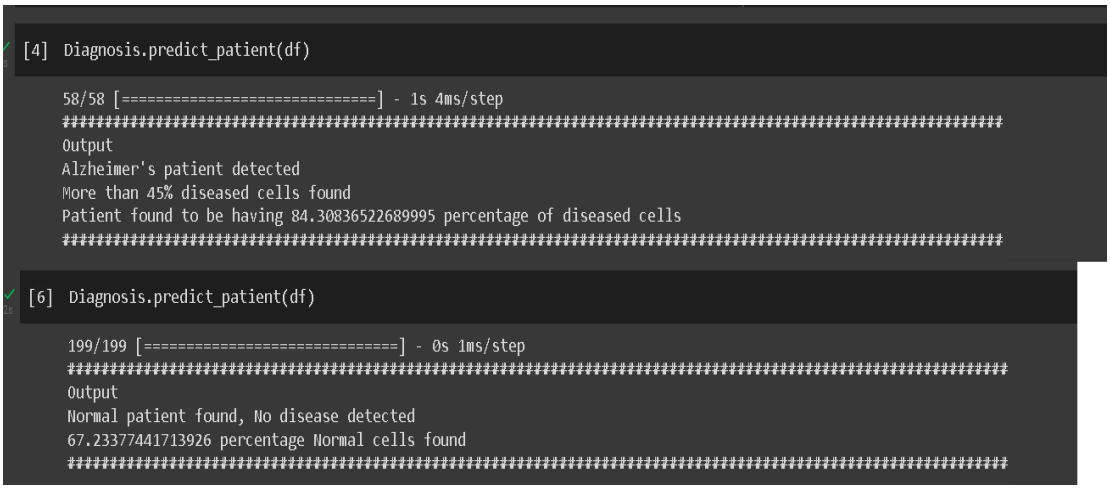

Figure 4.11: Output

CHAPTER 5

DISCUSSION

Alzheimer's has now been recognized in the category of world health concerns. It accounts for nearly 60-70% cases of dementia [38]. There are several criteria that have been proposed for the screening and diagnosis of AD, including physical symptoms, bodily fluids, and imaging studies. Despite this, there is currently no cure for AD, and the only effective therapy is symptomatic. Various drugs such as galantamine, donepezil, and rivastigmine are prescribed to increase memory power and brain alertness, but they cannot control the disease progression [39]. Numerous studies have demonstrated that altering lifestyle practices, including as food and exercise, can enhance brain health and lessen AD without requiring medical attention [40]. This is why it is recommended as a first-line prevention strategy for all AD patients.

In this study, we have obtained a subset of genes which could act as potential diagnostic biomarkers for AD. Significant neuronal loss and neuropathological abnormalities can harm several brain regions by the time it is normally identified [41]. To prevent hazardous issues like these, the study aims to highlight some specific biomarkers which could aid in early screening and diagnosis of Alzheimer's disease from the Single Cell genome. The biomarkers are shown in the table 3.7. The single-cell data is usually very large and sparse, which makes it difficult for people to analyze for multiple patients.

The principal changes in AD are the metabolic pathways for AD's neurodegenerative nature, which include extracellular amyloid plaques, intracellular neurofibrillary tangles, synaptic degeneration, and neuronal death [42].

We have obtained a set of 35 genes which could classify AD patients from Normal control patients via the Deep Learning ANN model. We have also tried to implement Machine Learning technologies on Single cell data, but they did not classify with much effect and demonstrated a low accuracy. In particular, ANN achieved 82% classification accuracy on 100 genes selected by mRMR method. However, the ANN method proved robust and showed approximately 75% accuracy on a smaller subset of 35 genes selected by the Incremental Feature Selection method.

The low performance or misclassification may be due to sample impurity and RNA sequencing dropout in the dataset. Nonetheless, we still believe it to be an interesting result to achieve decent classification accuracy from single-cell data, which is highly sparse and error-prone. A generic threshold of 45% is set to classify the patient's diagnosis, i.e. if the patient has more than 45 % of diseased cells, he/she would be diagnosed as Alzheimer's affected, and if the patient has more than 55% of normal cellspredicted, he she is classified in the category of healthy persons.

Silencing of lncRNA X-inactive specific transcript (XIST) was found to be directly connected to an increase in Alzheimer's symptoms. The XIST gene was found to be upregulated in AD models in both invitro and invivo [43]. Gene "TSC22D4" in a study in Japanese Alzheimer's Patients showed to be significantly differentially regulated as compared to normal controls [44]. "FGF17", i.e. Fibroblast growth factor dysregulation, has been reported in various brain-related (neurological and psychiatric disorders). It has been shown to be altered in epileptogenesis [45]. Transcriptional regulatory changes are prominent features of brain diseases, transcriptional changes in gene "FOXN2" highlights the convergence of genetic risk with psychiatric and neurodegenerative disorders [46].

Genes such as "CDK18" have uncharacterized mechanisms by which they may promote AD neurodegeneration, and this increases the probability that their inhibition may cause protection against AD development pathology [47]. There are also genes such as 'BCOR', 'AC090517.4', 'LGI4', 'SLC25A13' and "ZBED5', which have not yet been reported any connection with Alzheimer's disease progression and development. Thus further deep studies are required to evaluate these biomarkers, and these could act as novel findings. Genes 'SCD' and 'UBE2Z' have been reported in various studies related to cognitive impairment and diagnosis or treatment of Alzheimer's disease [48] [49]. 21 genes out of the 35 genes above in the table 3.7 have been reported with connection to Alzheimer's disease and their relation with neurodegeneration. Other genes need further study and research. We have evaluated a total of 21 samples from both Normal and Alzheimer's affected patients. The data obtained was from the prefrontal cortex of the brain. More data from areas like the hippocampus and amygdala can be obtained to add to this dataset for identifying further promising biomarkers.

CHAPTER 6

CONCLUSION

In this study, we have used various Machine Learning models and an ANN deep learning model to classify Normal Control(NC) cells and Alzheimer's Disease(AD) cells from their single-cell RNA seq data. Also, patient-wise analysis to classify the samples is also done in this study on a total of 21 samples. We compared the Deep Learning model with other Machine Learning models to find out that the deep learning model performedwell very against all the Machine learning models. Initially, the dataset was quite large with a very high number of features (>30000). The data was then preprocessed and thefeature count was reduced to a significantly low number (=5000). Since many features were co-related and redundant, so a feature selection method known as mRMR [16] was applied, to get a minimal set of features which could be helpful in classifying the samples. Out of these 5000 features, mRMR was applied to extract the top 100 featureswith minimal redundancy and maximal relevance. The IFS method was then applied toselect the optimal number of biomarkers. 35 genes were selected to be optimal. In addition, there 35 genes (features) distinguished the AD patients from NC patients withan accuracy of 74% and AUC- ROC of 0.75. ANN proved to be a powerful tool for biomarker identification and classification. However, more studies on the identified genes in clinical setups are required for in-depth analysis and their role in how they affectand cause Alzheimer's disease progression.

REFERENCES

[1] R. Mayeux and Y. Stern, “Epidemiology of alzheimer disease. cold spring harb perspect med. 2012; 2 (8).”

[2] K. G. Yiannopoulou and S. G. Papageorgiou, “Current and future treatments in alzheimer disease: an update,” *Journal of central nervous system disease*, vol. 12, p. 1179573520907397, 2020.

[3] Y. Hara and H. M. Fillit, “The dawn of a new era of alzheimer’s research and drug development,” 2022.

[4] K. Mrachek, “Cerebral amyloid angiopathy,”

[5] I. S. Padda and M. Parmar, “Aducanumab,” in *StatPearls [Internet]*, StatPearls Publishing, 2022.

[6] S. S. Ahmad, S. Khan, M. A. Kamal, and U. Wasi, “The structure and func- tion of α, β and γ-secretase as therapeutic target enzymes in the developmentof alzheimer’s disease: A review,” *CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders)*, vol. 18, no. 9, pp. 657–667, 2019.

[7] K. G. Yiannopoulou and S. G. Papageorgiou, “Current and future treatments in alzheimer disease: an update,” *Journal of central nervous system disease*, vol. 12, p. 1179573520907397, 2020.

[8] D. S. Lark and T. J. LaRocca, “Expression of exosome biogenesis genes is differentially altered by aging in the mouse and in the human brain during alzheimer’s disease,” *The Journals of Gerontology: Series A*, vol. 77, no. 4, pp. 659–663, 2022.

[9] P. J. Flannery and E. Trushina, “Mitochondrial dynamics and transport in alzheimer’s disease,” *Molecular and Cellular Neuroscience*, vol. 98, pp. 109–120, 2019.

[10] W. Chung, S. Yoon, and Y. S. Shin, “Multiple exposures of sevoflurane during pregnancy induces memory impairment in young female offspring mice,” *Korean Journal of Anesthesiology*, vol. 70, no. 6, pp. 642–647, 2017.

[11] L. Chen, Y. Li, L. Zhu, H. Jin, X. Kang, and Z. Feng, “Single-cell rna sequencing in the context of neuropathic pain: Progress, challenges, and prospects,” *Translational Research*, 2022.

[12] S.-F. Lau, H. Cao, A. K. Fu, and N. Y. Ip, “Single-nucleus transcriptome analysis reveals dysregulation of angiogenic endothelial cells and neuroprotective glia in alzheimer’s disease,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 41, pp. 25800–25809, 2020.

[13] A. X. Garcia, J. Xu, F. Cheng, E. Ruppin, and A. A. Schäffer, "Altered gene expression in excitatory neurons is associated with alzheimer's disease and its higher incidence in women," 2022.

[14] F. A. Wolf, P. Angerer, and F. J. Theis, "Scanpy: large-scale single-cell gene expression data analysis," *Genome biology*, vol. 19, no. 1, pp. 1–5, 2018.

[15] W. McKinney *et al.*, "pandas: a foundational python library for data analysis and statistics," *Python for high performance and scientific computing*, vol. 14, no. 9, pp. 1–9, 2011.

[16] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," in *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, pp. 523–528, 2003.

[17] B. Niu, G. Huang, L. Zheng, X. Wang, F. Chen, Y. Zhang, and T. Huang, "Prediction of substrate-enzyme-product interaction based on molecular descriptors and physicochemical properties," *BioMed research international*, vol. 2013, 2013.

[18] S.-C. Wang, "Artificial neural network," in *Interdisciplinary computing in java programming*, pp. 81–100, Springer, 2003.

[19] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, 1996.

[20] R. E. Wright, "Logistic regression.," 1995.

[21] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.

[22] J. R. Quinlan, "Learning decision tree classifiers," *ACM Computing Surveys (CSUR)*, vol. 28, no. 1, pp. 71–72, 1996.

[23] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 18, no. 6, pp. 275–285, 2004.

[24] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[25] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, 2017.

[26] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[27] A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, "K-nearest neighbor classification," in *Data mining in agriculture*, pp. 83–106, Springer, 2009.

[28] D. Mathur, S. Singh, A. Mehta, P. Agrawal, and G. P. Raghava, "In silico approaches for predicting the half-life of natural and modified peptides in blood," *PloS one*, vol. 13, no. 6, p. e0196829, 2018.

[29] H. Kaur, S. Bhalla, and G. P. Raghava, "Classification of early and late stage liver hepatocellular carcinoma patients from their genomics and epigenomics profiles," *PloS one*, vol. 14, no. 9, p. e0221476, 2019.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[31] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[32] A. Gulli and S. Pal, *Deep learning with Keras*. Packt Publishing Ltd, 2017.

[33] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[34] P. Guo, Y. Luo, G. Mai, M. Zhang, G. Wang, M. Zhao, L. Gao, F. Li, and F. Zhou, "Gene expression profile based classification models of psoriasis," *Genomics*, vol. 103, no. 1, pp. 48–55, 2014.

[35] L. Van Der Maaten, E. Postma, J. Van den Herik, *et al.*, "Dimensionality reduction: a comparative," *J Mach Learn Res*, vol. 10, no. 66-71, p. 13, 2009.

[36] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[37] M. W. Dorrity, L. M. Saunders, C. Queitsch, S. Fields, and C. Trapnell, "Dimensionality reduction by umap to visualize physical and genetic interactions," *Nature communications*, vol. 11, no. 1, pp. 1–6, 2020.

[38] K. Chanda and D. Mukhopadhyay, "Lncrna xist, x-chromosome instability and alzheimer's disease," *Current Alzheimer Research*, vol. 17, no. 6, pp. 499–507, 2020.

[39] Z. Breijyeh and R. Karaman, "Comprehensive review on alzheimer's disease: causes and treatment," *Molecules*, vol. 25, no. 24, p. 5789, 2020.

[40] G. K. Bhatti, A. P. Reddy, P. H. Reddy, and J. S. Bhatti, "Lifestyle modifications and nutritional interventions in aging-associated cognitive decline and alzheimer's disease," *Frontiers in Aging Neuroscience*, vol. 11, p. 369, 2020.

[41] A. Serrano-Pozo, M. P. Frosch, E. Masliah, and B. T. Hyman, "Neuropathological alterations in alzheimer disease," *Cold Spring Harbor perspectives in medicine*, vol. 1, no. 1, p. a006189, 2011.

[42] B. Hyman, "The neuropathological diagnosis of alzheimer's disease: clinical-pathological studies," *Neurobiology of aging*, vol. 18, no. 4, pp. S27–S32, 1997.

[43] D. Yue, G. Guanqun, L. Jingxin, S. Sen, L. Shuang, S. Yan, Z. Minxue, Y. Ping, L. Chong, Z. Zhuobo, *et al.*, "Silencing of long noncoding rna xist attenuated alzheimer's disease-related bace1 alteration through mir-124," *Cell Biology International*, vol. 44, no. 2, pp. 630–636, 2020.

[44] C. Humphries, M. A. Kohli, P. Whitehead, D. C. Mash, M. A. Pericak-Vance, and J. Gilbert, "Alzheimer disease (ad) specific transcription, dna methylation and splicing in twenty ad associated loci," *Molecular and Cellular Neuroscience*,vol. 67, pp. 37–45, 2015.

[45] C. A. Turner, E. Eren-Koçak, E. G. Inui, S. J. Watson, and H. Akil, "Dysregulated fibroblast growth factor (fgf) signaling in neurological and psychiatric disorders,"in *Seminars in cell & developmental biology*, vol. 53, pp. 136–143, Elsevier, 2016.

[46] J. R. Pearl, C. Colantuoni, D. E. Bergey, C. C. Funk, P. Shannon, B. Basu, A. M. Casella, R. T. Oshone, L. Hood, N. D. Price, *et al.*, "Genome-scale transcriptional regulatory network models of psychiatric and neurodegenerative disorders," *Cell systems*, vol. 8, no. 2, pp. 122–135, 2019.

[47] D. Chaput, L. Kirouac, S. M. Stevens Jr, and J. Padmanabhan, "Potential role of pctaire-2, pctaire-3 and p-histone h4 in amyloid precursor protein-dependent alzheimer pathology," *Oncotarget*, vol. 7, no. 8, p. 8481, 2016.

[48] L. K. Hamilton, G. Moquin-Beaudry, C. L. Mangahas, F. Pratesi, M. Aubin, A. Aumont, S. E. Joppé, A. Légiot, A. Vachon, M. Plourde, *et al.*, "Stearoyl-coa desaturase inhibition reverses immune, synaptic and cognitive impairments in an alzheimer's disease mouse model," *Nature communications*, vol. 13, no. 1, pp. 1–17, 2022.

[49] K.-H. Lim and J.-Y. Joo, "Predictive potential of circulating ube2h mrna as an e2 ubiquitin-conjugating enzyme for diagnosis or treatment of alzheimer's disease," *International journal of molecular sciences*, vol. 21, no. 9, p. 3398, 2020.