Detecting miRNA Biomarkers For Alzheimer's Disease Using Next Generation Sequencing Data

- Final Report -

Imalsha Dinuwanthi
Hasini Thilakarathna
Vidwa Sripadi

Department of Computer Engineering
University of Peradeniya

Final Year Project (courses CO421 & CO425) report submitted as a requirement of the degree of
B.Sc.Eng. in Computer Engineering

April 2021

Supervisors: Dr. Damayanthi Herath (University of Peradeniya) and Prof. Roshan Ragel (University of Peradeniya)

We would like to dedicate this thesis to our loving parents, supervisors and all the lecturers in our department....

Declaration

I/We hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my/our own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgments.

Imalsha Dinuwanthi
Hasini Thilakarathna
Vidwa Sripadi
April 2021

Acknowledgements

First of all, we would like to acknowledge our supervisors Dr. Damayanthi Herath and Prof. Roshan Ragel for all the support given for us from the begin until the end of this project. You guided us in the correct path whenever we discussed the ideas in weekly meetings. So, we are grateful for your guidance and motivation without which we would not have come this far in the completion of this project.

We would also like to show our gratitude to all the lecturers in the department of computer engineering, who have been supporting us whenever we need to clarify anything regarding theoretical knowledge which indeed was a great help in the completion of this project.

Finally, we would like to acknowledge Prof. Sen-Lin Tang of Biodiversity Center in Taiwan, who directed us to the right path in the beginning of this project.

Abstract

Alzheimer's disease (AD) has been identified as one of the most common diseases found in people aged 65 or above by different researches done in past few years, which is still well known in medical field as a disease which has no efficient cure. We introduce a method for identifying AD patients using microRNAs (miRNAs). Different technologies such as microarray technology, Sanger Sequencing and Next Generation Sequencing been used by various researchers for gathering samples in previous researches. In this project, we used samples gathered from next generation sequencing technology. Out of many different sequencing techniques available in NGS like, Illumina Sequencing, Roche Sequencing and SoLiD Sequencing, we considered a data set which was collected with the Illumina Sequencing technique. To find the required samples, we used the NCBI database. The initial data set includes 70 samples and 2652 miRNAs. We used Galaxy tool for preprocessing the raw NGS data. Man Whitney test, fold change, and Receiver operating characteristics area under curve (ROC AUC) values are the statistical methods used in the obtaining the most significant miRNAs. Then we used PCA, Random Forest and correlation coefficient values for selecting the best set of biomarker miRNAs out of the features obtained from the statistical methods. Further, machine learning approaches were used for the classification purposes. Based on these we identified 25 biomarker miRNAs and we validated them using HMDD v3.2.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

AD	Alzheimer's Disease
AUC	Area Under Curve
DLB	Dementia with Lewy Bodies
HC	Healthy Controls
HMDD	Human MiRNA Disease Database
KNN	k Nearest Neighbour
MCI	Mild Cognitive Impairment
NCBI	National Centre for Biotechnology Information
NGS	Next Generation Sequencing
PCA	Principal Component Analysis
ROC	Receiver Operating Characterstics
SVM	Support Vector Machine
WMW	Wilixcon Man Whitney

Chapter 1

Introduction

1.1 What is Next-Generation Sequencing?

Next Generation Sequencing is a high throughput option introduced for gathering multiple DNA samples in parallel. Although different techniques have been introduced for next generation sequencing, all those techniques follow some common guidelines as sample preparation, sequencing machines and data output. To understand the Next Generation Sequencing, one needs to get a clear idea on how these guidelines [1] are used in the process of Next Generation Sequencing. For any sample to be sequenced, it is needed to be prepared into a sample library; which is a collection of genomes with different sizes, either from genomic DNA or total RNA. When the samples are prepared into a library, the sequencing of the genomes begins, providing a sequenced sample at the end of the process(Figure 1.1).

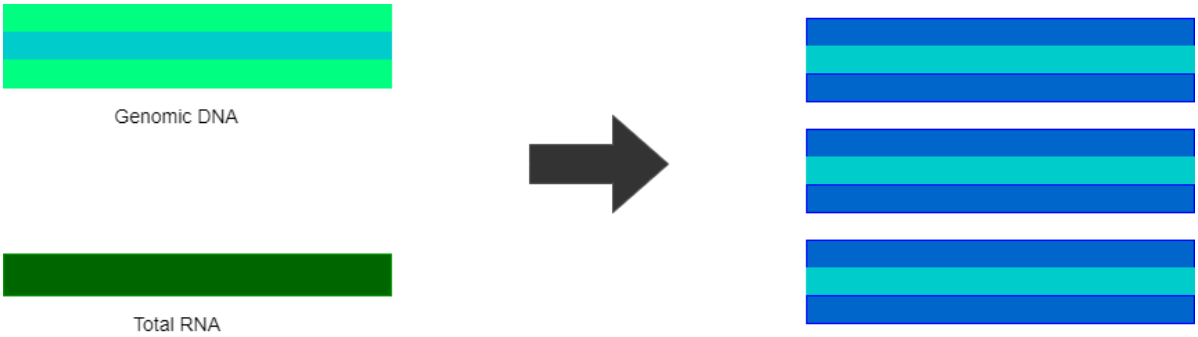

Fig. 1.1 Preparing library by genomic DNA or total RNA

The most popular technique used in diagnosis of human diseases is the Illumina technique, which is also the technique used in our project. In this technique, sequencing

by synthesis method is used for DNA Sequencing. The process of sequencing by synthesis method is illustrated in the given Figure 1.2 to provide the basic idea of how the sequencing process works.

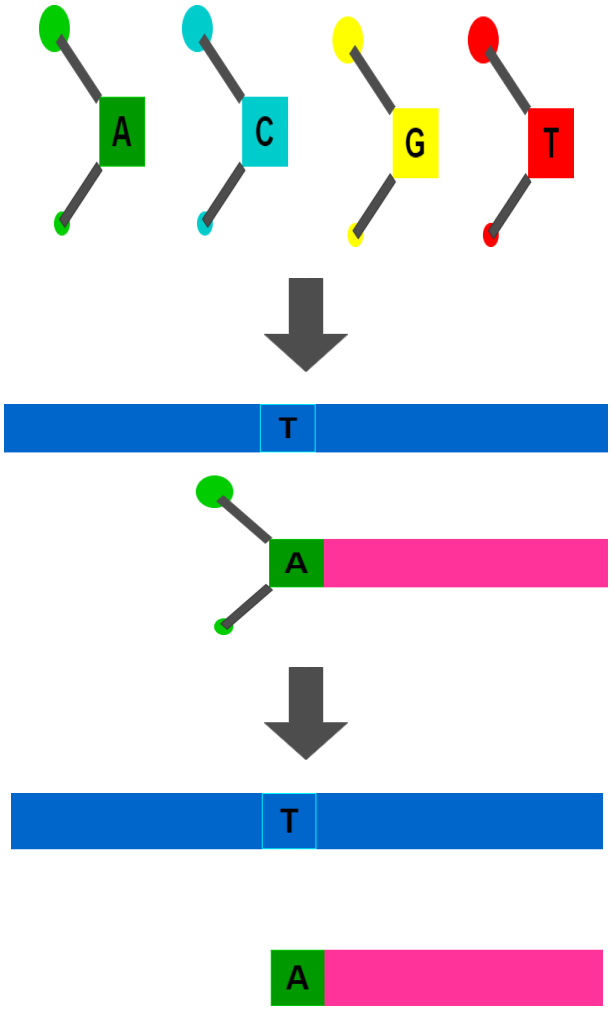

Fig. 1.2 Sequencing by synthesis process

1.2 Why detecting biomarker miRNAs for Alzheimer's disease using NGS?

1.2.1 Background

According to the Alzheimer's Association, no cure has still been found for AD. Many methods have been developed in order to diagnosis AD in past few years. Different technologies such as microarray technology, Sanger Sequencing and Next Generation Sequencing have been used by various researchers for gathering samples. Out of theses, next generation sequencing has become more common now a day, as it is a powerful platform which enables to sequence thousands or millions of DNA molecules simultaneously. Samples are used in many forms as gene expressions, MicroRNAs and RNAs. miRNAs are small on coding RNAs which mainly help for regulating gene expressions. Most of the researches were done using gene expressions to find AD biomarkers. But, the potential of miRNAs as biomarkers for disease diagnosis has been emphasized by several researchers in the past few years [2].

1.2.2 The problem

Alzheimer's disease has been identified as one of the most common diseases found in people aged 65 or above by different researches done in past few years, which is still well known in medical field as a disease which has no efficient cure. It is identified as a deadly disease which initially spreads in human body before 20 years when the symptoms are shown. Therefore, if a method to diagnosis AD at early stage can be found, then there's an opportunity for curing the disease at its early stage which will save many lives. According to United States National Institute of Aging, there is even no way to weaken the progress of spreading the AD in patient's body. It also mentions that the only available treatment in the medical field is for reducing its symptoms, like memory loss and confusion. Although many studies have carried out for diagnosis AD, none of them have been succeeded in addressing all the key features needed to be addressed when developing a method for finding biomarkers. The Alzheimer's disease spreads in the brain, which means that most significant features can be extracted from the brain samples. But, most of the researches done so far, were done using only blood sample probably because they are easily available. Only few have included brain samples in their data, which could give better results when finding candidate biomarkers for AD. In this project, we have used a data set which contains brain miRNAs in addition to blood samples. Most of the previous studies have followed statistical methods for finding

biomarkers. Only few researchers have focused on using machine learning algorithms for the classification purposes. Applying machine learning approaches in addition to statistical methods could provide better results than using only the statistical methods as features are filtered by two methods instead of one. If there's a proper diagnosis method, then there won't be increase of AD patients day by day as estimated by the Alzheimer's association. Therefore, clearly there's an opportunity for finding a more accurate method to identify AD patients.

1.2.3 The proposed solution

We are going to use a set of samples which are collected using next generation sequencing method. Preprocessing of the samples is going to be carried out using Galaxy tool by adjusting suitable settings, in order to obtain a data set with miRNA read counts for each sample. We are going to subject the prepared data set to statistical analysis to filter out the most significant miRNAs. This can be done using Wilixcon Man Whitney test which is a test which gives significance values (p values) for each miRNAs. miRNAs with p values less than the significance level (0.05) are identified as the most significant miRNAs, can be passed into the next level of analysis. For further identifying the most up regulated and the most down regulated miRNAs AUC values can be used. Machine learning approaches can be used for classification and analyzing the miRNAs filtered out by the statistical analysis step. Then we are going to use correlation values for identifying the best set of biomarker miRNAs. Following key points explains why our proposed solution could make a significant change in the field of medical when diagnosing AD patients.

- Both statistical and machine learning approaches are going to be used for finding most significant miRNAs which will provide better results.
- More accurate results since features are selected using more than one methods (using Random Forest, PCA and correlation coefficients).
- Validation will be done for the miRNAs selected from statistical analysis and machine learning approach.

1.2.4 Deliverable and milestones

Following are the milestones and deliverable we are presenting for the phase 1 of our project.

- Milestone 01: Doing background study for understanding the topic and then narrowing down the project scope for a practically achievable one. For the selected scope, come up with a methodology for finding the biomarkers. Search for datasets to be used to initiate the proposed methodology. While doing the background search write a review paper with the project related articles.

- Milestone 02: Choosing a suitable tool for preprocessing the selected dataset and use the most effective one to preprocess the dataset. Study the dataset by visualizing the its features using python. Using a suitable normalization technique normalize the dataset.

- Milestone 03: Analyze different statistical methods for finding p values and find the p values for each miRNA using the most suitable approach. Filter out the most significant features using significance level. Introduce the most dysregulated features by calculating receiver operating characteristics area under curve values. Classification of AD and control samples using machine learning approaches.

Gantt chart is given in the Figure 1.3 for showing the progress of phase 1.

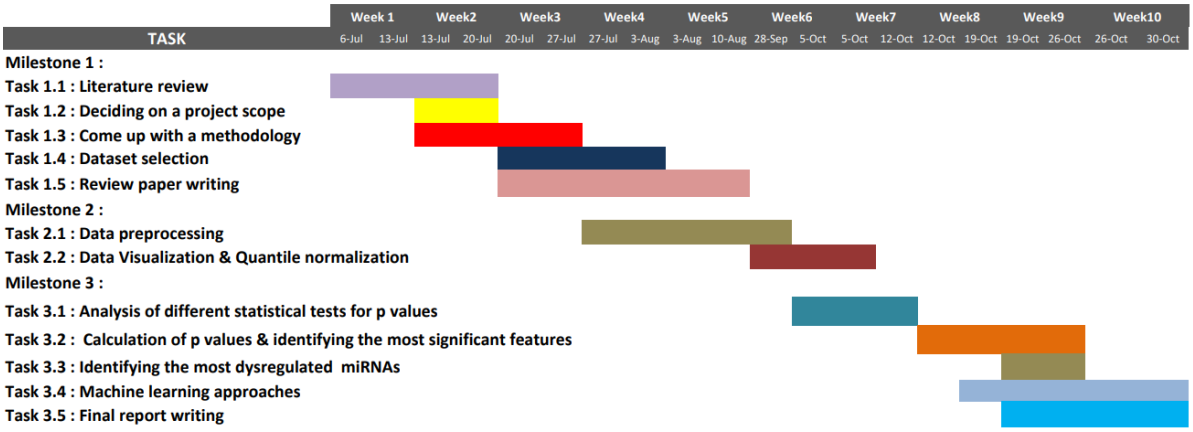

Fig. 1.3 Gantt chart for phase 1

Milestones and deliverable for the phase 2 of our project are as follows.

- Milestone 01: Identify the biomarker miRNAs using Random Forest analysis, PCA, and correlation coefficient values. Use Random Forest analysis and PCA for selecting best set of features from previously filtered data set from the statistical analysis. Use correlation coefficient values for further analyzing the less correlated features obtained from those two methods. Again out of those less correlated

features also, with the aid of the classification algorithms, obtain the set of most accurate miRNAs which are suitable as biomarkers for AD diagnosis. Obtain the final results as a combination of the results obtain from all three methods.

- Milestone 02: Validate the final results using the Human Disease Database v3.2.
- Milestone 03: Analyze the miRNA data set, which can be found under the accession number GSE147218 in NCBI database, using the same methodology. Use that data set for observing the response of the developed methodology to different data sets.

Figure 1.4 shows the Gantt chart for showing the progress in phase 2.

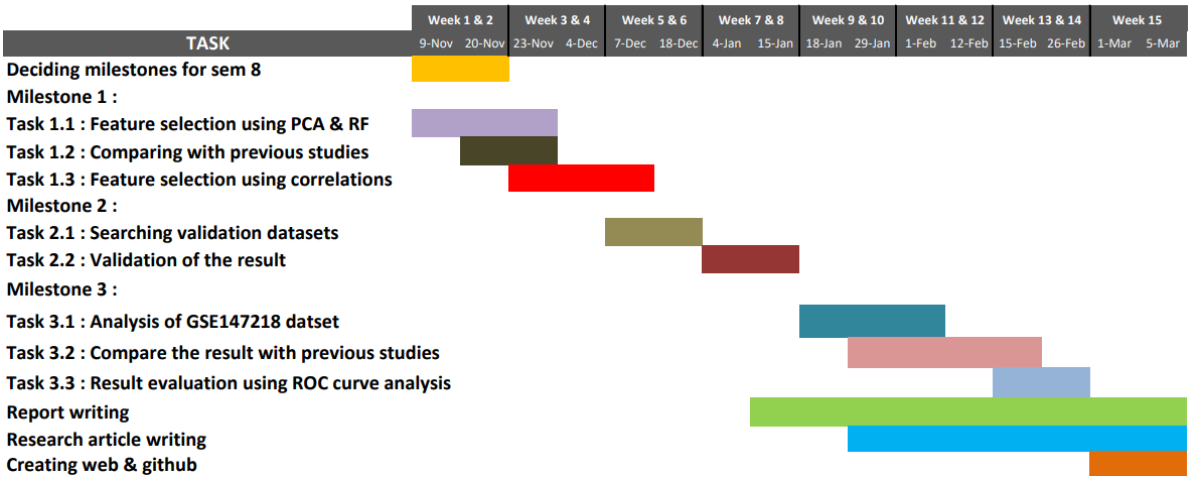

Fig. 1.4 Gantt chart for phase 2

1.2.5 Outline of the report

An introduction giving the basics of the research are discussed in this chapter. In the next chapter we are looking at the previous researches done for finding miRNA biomarkers using next generation sequencing. In chapter 3 and 4 we discuss the methodology used and how that methodology is implemented (respectively). Chapter 5 discusses the most important section in the report which is the results that we obtained. Conclusions and future works of this project are discussed in the final chapter.

Chapter 2

Related work

2.1 Introduction

The diagnosis of AD patients has now become a huge challenge which has still not been properly addressed. There are several researches which were carried out to detect biomarkers for Alzheimer's disease using different sequencing methods like microarray, Sanger Sequencing and Next Generation Sequencing [3]. In the past few years, Next Generation Sequencing technology has emerged as the most powerful sample collection method. Diverse researches have used different types of sequencing data for their investigations. miRNAs can be used to regulate AD-related proteins in the brain and it is considered as a potential novel biomarker which is mostly used in AD diagnosis [4, 5].

2.2 Sample selection

When detecting biomarkers for Alzheimer's disease, initially we have to select a sample for performing analysis. Mostly blood samples are used due to the high availability. Different types of blood samples including whole blood [6–10], serum [11–14] and plasma [15, 16] are used by many previous researchers where they have tried to find miRNA biomarkers. If we use brain samples it would give most accurate results than blood samples since AD is most prominently active in brain [17]. We would be able to give more accurate results if both blood and brain samples are used. Samples can be taken from participants generally as, AD and controls [7, 8, 11, 15] and also they can be taken considering the different stages as severe, moderate, mild AD and controls [13]. Another approach in collecting samples is taking then from participants with HC, MCI and AD [12]. The number of samples used when developing a diagnosis method can be identified as one of the main

factors which could affect the final results. Next Generation Sequencing platform is the most trending method used for gathering samples for various disease diagnosis researches. Many techniques like Illumina Sequencing technique [6, 7, 9, 11, 13, 4] are introduced for working with NGS data. Preprocessing the raw sequence counts can be done using a bioinformatic pipeline, which gives the read counts for each miRNA as the final outcome.

2.3 Normalization

Normalization of sequencing read counts can be performed using several normalization methods. Quantile normalization is one way that we can do normalization when we are having a high dimensional dataset. It excludes selected samples to minimize noise [6, 7]. Mean normalized read counts also can be used to filter out the miRNAs. Also we can follow a stepwise procedure to do normalization as below [13].

- From all the samples, find sequences which are common.
- Build a reference dataset using those common sequences.
- Apply logarithmic transformation.
- Calculate the logarithmic difference between each sample and reference dataset.
- Form a subset by taking sequences which has a difference<2.
- Perform Linear Regression.
- Calculate the mid value.

Looking at the results obtained from the study which used the above normalization method, it can conclude that this type of step wise normalization method can be used for obtaining the best set of miRNAs. Data visualization can be used for selecting which normalization method is best suit for a given dataset.

2.4 Statistical Analysis

Initial detection of miRNAs can be done by initially calculating a significance value(p value). P value is a value between 0 and 1, which shows the level of statistical significance. If a p value is less than the significance level (0.05), it is considered as a nominally significant p value and we can select those miRNA as the most impacting miRNAs. WMW test [11, 14], Wald test and Fisher's exact test [8] can be used to calculate the p

values and these p values can be adjusted for multiple testing using an approach like Benjamini-Hochberg approach [6, 7]. Other than that, t test and kruskal test can also be used to calculate significance values [15].

2.5 Validation of samples

Validation of the samples makes it easier for the next steps in the investigation and also it makes the final results more accurate. After the statistical analysis process, for validating the obtained samples, quantitative real time-polymerase chain reaction (qRT-PCR) method is used by many researchers. It analyzes the expression of single miRNAs by applying the method on previously used samples for sequencing [7, 9–14]. But in a previous study [7], they have additionally included patients with AD and also patients with other neurological disorders in the validation step, to analyze the the set of miRNAs they obtained in the previous step. After the validation is carried out, the miRNAs can be further filtered out to obtain the most significant miRNAs [13].

2.6 Receiver operating characteristic curves

Receiver operating characteristic curve analysis is used to evaluate the performance or accuracy of a classification model. ROC is a plot of sensitivity against specificity for selected samples. It is also used to initially detect the dysregulation of miRNAs and to discriminate between AD and NC sample groups. The area under the curve is the degree of separability. If the AUC is high, that means that particular miRNA is better to distinguish patients with AD and control.

2.7 Feature selection and Classification

If we use a classification model without using feature selection, it will take more run time due to the huge size with redundant features. Therefore it is required to apply some feature selection method to reduce those redundant features. Hierarchical clustering is a feature selection method which can be used to statistically analyze the dataset [6, 7, 9, 10, 12]. It will build clusters of miRNAs having similar patterns. Principal Component Analysis is another approach which can be used for the feature selection [6, 10, 12]. Machine learning classifier models are used to predict whether a sample belongs to AD or control. AdaboostM1, J48 decision tree, Random Forest and Support Vector Machines and radial basis SVM are some machine learning approaches that can

be used for building prediction models. In a previously done study, they have built a separate model by performing 7-way cross validation using 7 randomly picked partitions of 5 positive and 5 negative samples each for the feature selection.

2.8 Summary

According to the review we have done, we identified how we can use miRNAs to diagnosis AD and what are the miRNA diagnostic biomarkers which can be found in AD patients. In each study, for filtering out the candidate miRNA, step wise procedures including initial detection and statistical analysis have performed. When consider about previous studies, there are several limitations. The most common limitation of most of the research is they used a limited number of the cohort to their experiments. It is hard to find a large number of Alzheimer's disease patients to do massive experiments. But we can obtain better results if we expand the cohort size. In many studies, samples with analyzed dementia and controls have used. But not discussing about the possibility to discover pre-clinical biomarkers for Alzheimer disease is a limitation of most of the previous studies. A model which was built in a one previously done study [12], does not develop to anticipate movement from HC to MCI or MCI to AD. Also, this model was incapable of applying for late-stage AD findings. In another study [14], they have mentioned that they were unable to recognize a mechanism to identify the variation of miRNAs in serum samples. Considering all the drawbacks, limitations and also the developments found in the previous studies, in this research, we are focusing on finding a more accurate solution for detecting AD biomarkers. Figure 2.1 shows a summary of different methods used and the results obtained in previous studies. According to this diagram, only 4 studies have used machine learning algorithms and only 5 studies have used statistical methods in their studies. Out of the results obtained from above mentioned 9 studies, 7 miRNAs were identified as common for those 9 studies.

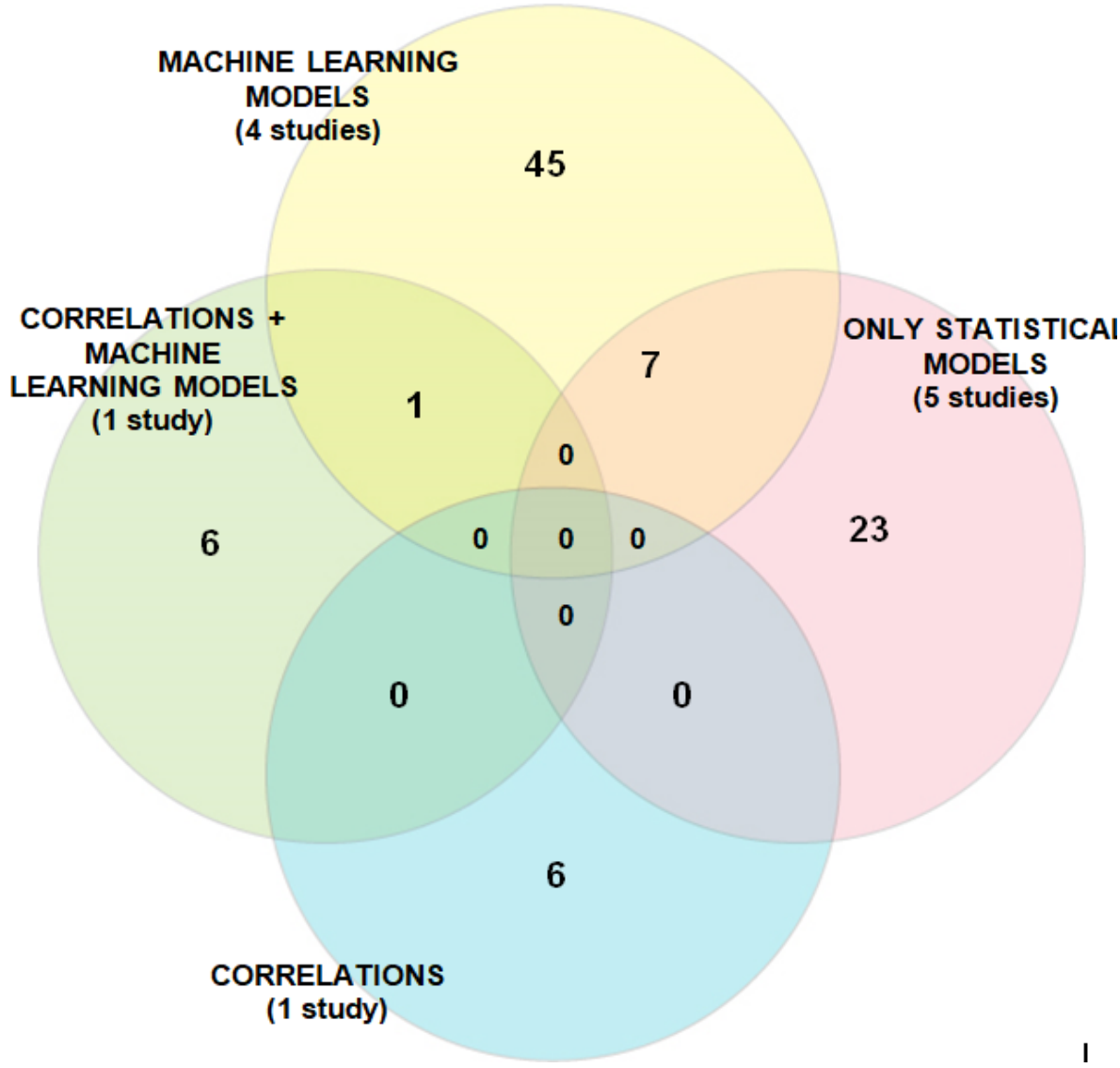

Fig. 2.1 Summary of methods used and results obtained in previous studies

Chapter 3

Methodology

3.1 Proposed Methodology

This project is about identifying the best set of biomarker miRNAs for Alzheimer's disease using NGS data. It is important to identify the biomarkers to pre-identify the disease. To achieve this task we plan to follow the following procedure.

- Finding a suitable NGS dataset from the public repository
  - The dataset should contain data from NGS platform such as Illumina MiSeq, Illumina HiSeq, SOLiD, etc.
  - The dataset should contain blood or brain miRNAs from AD patients and healthy humans
- Preprocessing sequencing data
  - Trimming adapters, indexes and low quality sequences
  - Filtering short read sequences
  - Filtering remaining low quality reads
- Creating a summarized dataset by including miRNA read counts for each miRNAs
- Removing lowly abundant data
- Data visualization and normalization
- Statistical analysis of summarized data
  - Finding the most significant miRNAs using P value (significance value)

  - Calculating AUC value for ROC curve
  - Finding dysregulated miRNAs including upregulated miRNAs and downregulated miRNAs using AUC
- Feature selection using Machine Learning models
  - PCA
  - Random Forest
- Feature reduction using correlation coefficient
- Classification using Machine Learning techniques
  - Support Vector Machine
  - Logistic Regression
  - Random Forest
- Validation of results using HMDD v3.2
- Developing a method for clinical use

3.2 Design and Methodology

3.2.1 Conceptual design

In Figure 3.1, the overall overview of the proposed methodology is shown.

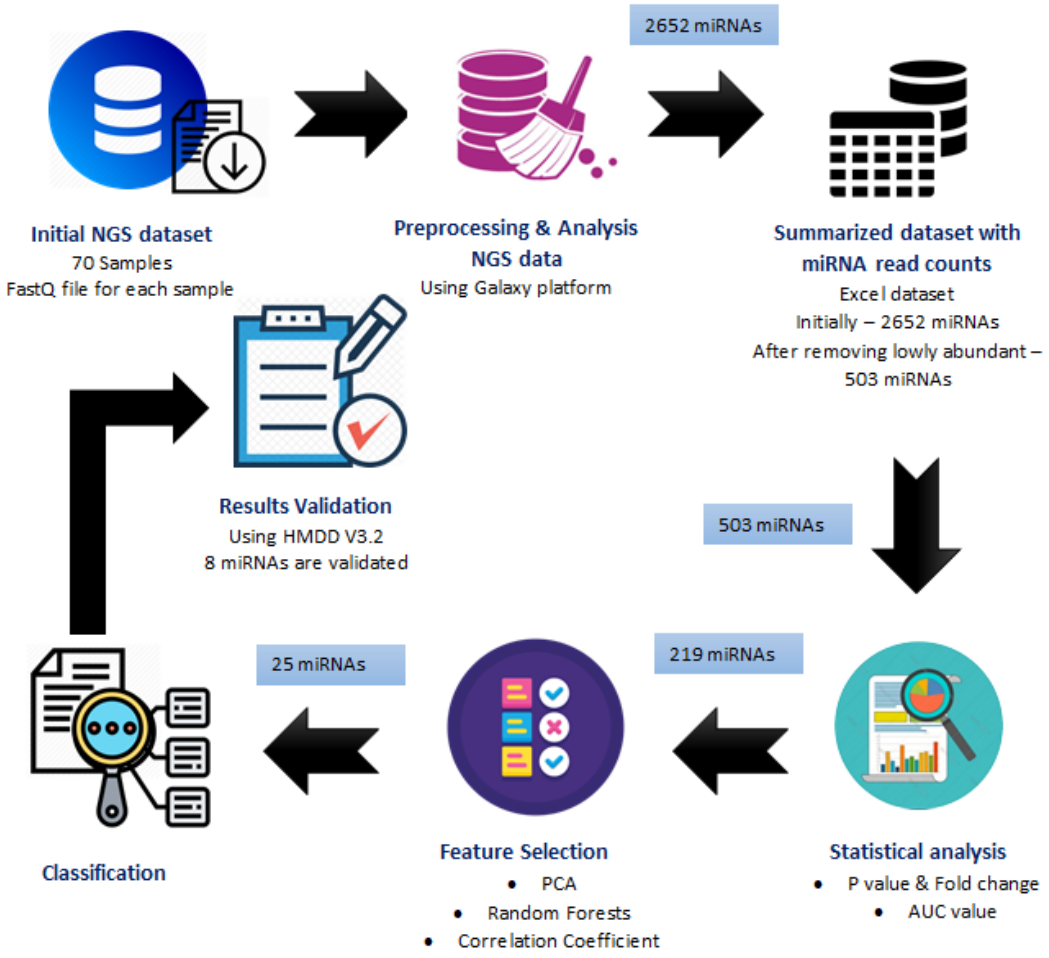

Fig. 3.1 Overview of the proposed methodology

3.2.2 Methodological approach

3.2.3 Sample Selection

We used a data set from NCBI database which can be found under the accession number GSE46579. This data set contains both blood and brain miRNAs.

3.2.4 Next generation sequencing data analysis

The raw sequencing data which was collected using the Illumina HiSeq 2000 platform was available in the NCBI GEO database under accession number GSE46579. The dataset contains samples from 48 AD patients (n = 48) and 22 healthy controls (n = 22). Sample data was downloaded and extracted to fastq type using Sratoolkit. Then

the data preprocessing and analysis were done by using the Galaxy platform. Galaxy is a web-based, open-source platform that is created for scientific data analysis [18]. It provides a lot of bioinformatics tools for data preprocessing and analyzing. First, the quality report of sequencing data was generated using the FastQC tool. FastQC tool is used to quality checking in row sequencing data. Then, using the tool Trim Galore, data trimming was performed. The package Trim Galore allows both quality trimming and adapter trimming at once [19]. Adapters, Poly A tails, and low-quality reads were trimmed in the trimming procedure. Then, the data filtering procedure was performed using the Filter FASTQ tool. Short read sequences and low-quality sequences were removed in filtering. After that, the NGS reads were mapped against a reference genome (h38) using Bowtie2. Bowtie2 is memory efficient and ultrafast tool for aligning sequencing reads to the long reference sequencing reads. Then the reads were mapped against the hsa.gff3 miRNA precursor sequences from the miRBase database (v22) [20] and the read counts of each miRNA were founded using the htseq-count tool. This preprocessing procedure was done for every sample using the galaxy platform and then the summarized data set was created by using miRNA read counts of each sample. Using the hsa-mir-200a miRNA as an example, we have illustrated in Figure 3.2 how the number of read counts are obtained in the preprocessing stage.

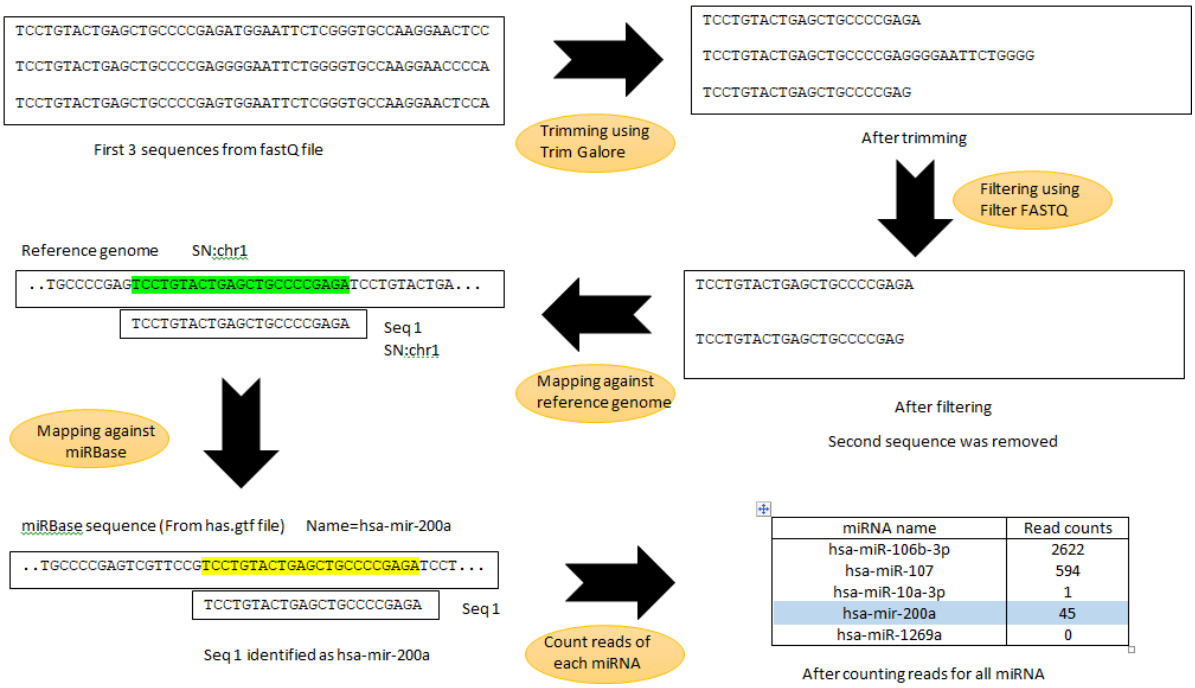

Fig. 3.2 Next Generation Sequencing Data analysis

3.2.5 Data Visualization

All the features and also some random features were visualized in order to get an idea of the data distribution. The Normalization method was chosen by studying the visualized distribution of data set. In addition to that, data visualization techniques were used for studying the correlation of selected features.

3.2.6 Normalization

As mentioned in the previous section, before normalizing the data visualization was mostly used. Considering the mean distribution of the data sets from different normalization methods, we decided to use Quantile normalization method over general min-max normalization method. The quantile normalization method removed unwanted variations from noisy data [21].

3.2.7 Statistical Analysis

The summarized data set from the NGS data analysis was used to perform statistical analysis. The statistical analysis was done by using python which has become a popular data analysis language in bioinformatics because of clean syntax, straightforward semantics, and third-party toolkits. As mentioned in the previous section, quantile normalization was done to make all distributions identical in statistical properties. We calculated the p values (significance values) for each miRNA using Wilcoxon-Mann-Whitney test and adjusted for multiple testing with the Benjamini-Hochberg adjustment technique. Fold change was calculated for each miRNA in the data set. Generally, it is a technique which is used to get an idea of how much change occurs going from one value to another. Here, we tried to get fold values for each miRNA, to check how each miRNA changes going from AD samples to control samples. P values were used to identify the most significant miRNAs. Addition to p values, we used fold change for each miRNA to get the most significant set of features. We used TAM (A tool for miRNA data analysis) on those selected features to get an idea of the most down regulated and the most up regulated features in the Alzheimer's disease. For the previously selected features using p values and fold changes, we calculated the receiver operator characteristics area under the curve values for each miRNA. The AUC values are used to identify the most upregulated and the most downregulated miRNAs.

3.2.8 Feature Selection

We used PCA and random forest for selecting miRNAs from the miRNAs filtered out from statistical analysis step. From the miRNAs obtained from those two methods, the set of miRNAs with less correlations was evaluated further with the correlation coefficient and classification accuracy. By changing the correlation coefficient, the accuracy of the data set was obtained repeatedly and the set of miRNAs which gave the highest accuracy was chosen as a set of biomarker miRNAs. The Figure 3.3 shows the process of the features selection in a flow diagram.

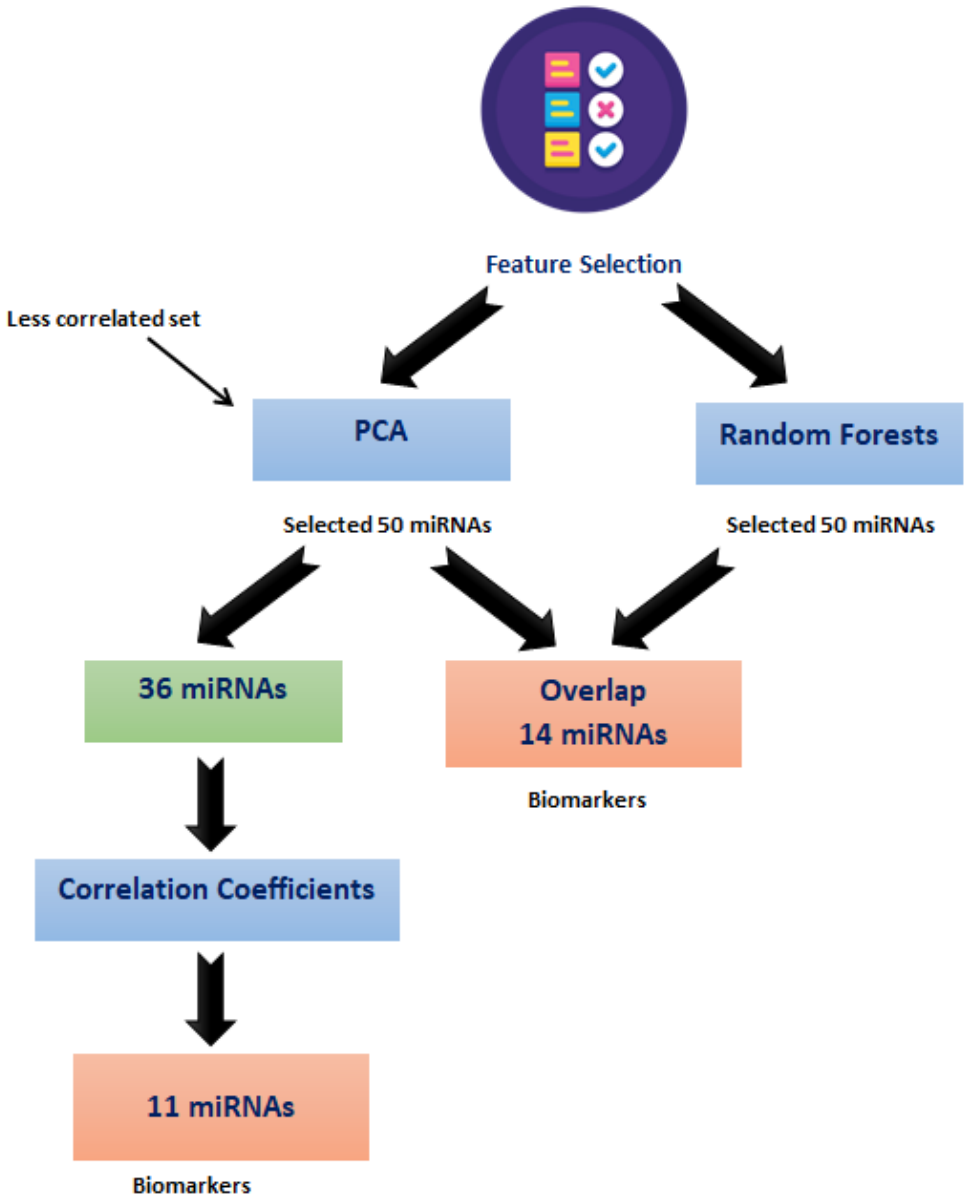

Fig. 3.3 Process of Feature Selection

3.2.9 Classification

We used machine learning approaches for the classification. Performances of different classification algorithms with our data set were analyzed for obtaining the best suited algorithms. Three classifiers with the highest accuracy were then used for obtaining the classification accuracy of the selected features filtered out from each stage. Specifically,

classification accuracy was used when obtaining the most accurate set of miRNAs using the correlation coefficients.

3.2.10 Validation of results

Human miRNA Disease Database (HMDD v3.2) was used for validating the final results obtained. The set of miRNAs which was identified as biomarker miRNAs was checked by a validation code to confirm whether they are included in this database as biomarkers of AD.

Chapter 4

Experimental Setup and Implementation

4.1 Technical Tools

- Sratoolkit : The toolkit was used to download SRA data sets from the NCBI database and convert them to fastq format.

The data preprocessing was done by using the Galaxy platform. Galaxy is a web-based, open-source platform that is created for scientific data analysis. It provides a lot of bioinformatics tools for data preprocessing and analyzing. The following tools were used to analyze data in the galaxy.

- FastQC : FastQC tool is used for quality checking in row sequencing data. It generates a quality report of sequencing data
- Trim Galore : Data trimming was performed using Trim Galore. The package Trim Galore allows both quality trimming and adapter trimming at once. Adapters, Poly A tails, and low-quality reads were trimmed using Trim Galore.
- Filter FASTQ : Using Filter FASTQ, short-read sequences and low-quality sequences were removed.
- Bowtie2 : The NGS reads were mapped against a reference genome (h38) using Bowtie2. Bowtie2 is memory efficient and ultrafast tool for aligning sequencing reads to the long reference sequencing reads.

- Htseq-count : The reads were mapped against the hsa.gff3 miRNA precursor sequences from the miRBase database (v22) [20] and the read counts of each miRNA were founded using the htseq-count tool.

The statistical analysis was done by using python programming language which is a popular data analysis language in bioinformatics because of clean syntax, straightforward semantics, and third-party toolkits. Following python libraries were used to the procedure.

- Numpy : Numpy supports large matrices and arrays analysis
- Scipy : Scipy supports mathematics, science and engineering
- scikit learn : Provide an efficient Machine Learning environment for analysis and classification of data
- Matplotlib : Matplotlib is used to plotting graphs in python

These are the tools that we used to perform NGS data analysis and statistical analysis. Also, we used Jupyter notebook as the IDE. It is an open-source web application that allows to create and share documents with code, comments, results and visualizations.

4.2 Sample Selection

A set of samples that were collected using the Illumina HiSeq 2000 platform was used in the initial data set. It was available in the NCBI GEO database under accession number GSE46579. There are 70 samples in this data set which includes 48 Alzheimer's disease patients and 22 controls. Also, another data set was used to compare the results of the method. It is also available in the NCBI GEO database under accession number GSE147218. This data set was collected using the Illumina Miseq platform. The data set contained 14 samples with 7 DLB samples and 7 control samples.

4.3 Data Manipulation and Testing

4.3.1 Data Visualization

As mentioned in the previous chapter, data visualization is mostly carried out to study the features in the data set and their distributions. We visualized density plots as in Figure 4.1 and Figure 4.2 of the normalized data for getting an idea how the distribution is after Qunatile normalization and after minmax normalization. Since the distribution

looked the same, we decided to go with the quantile normalization as in this kind of bioinformatic analysis, quantile normalization is most commonly used, mostly considering the size of the data set used. Figure 4.3 and Figure 4.4 shows the mean distribution and the box plot of selected features which we used to study the features of the data set.

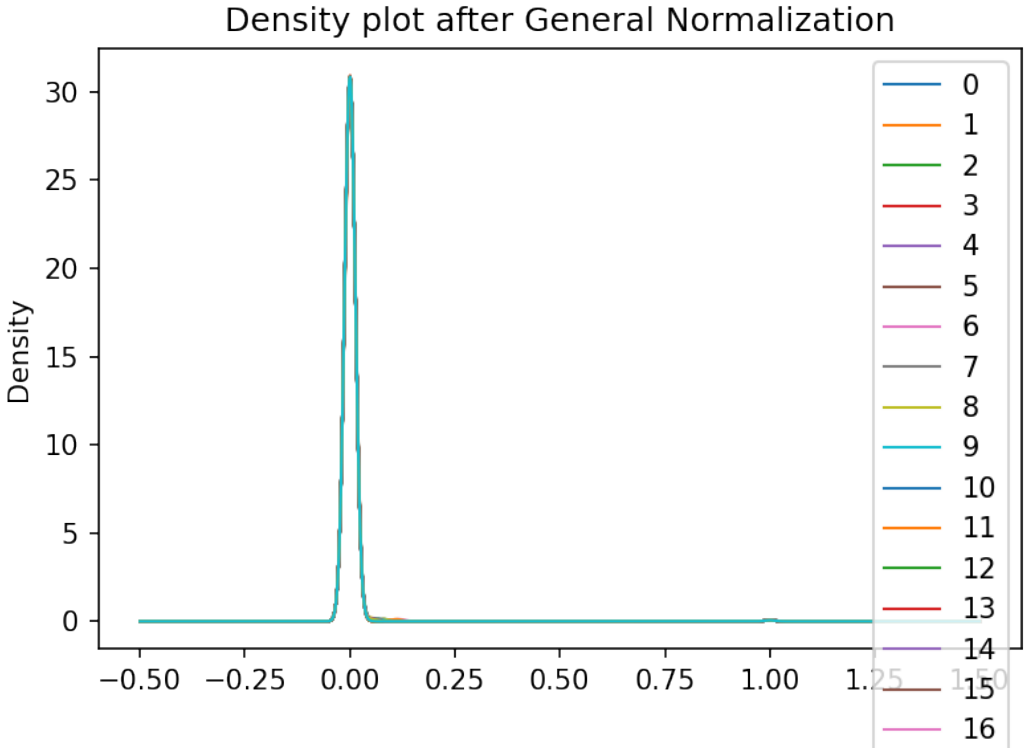

Fig. 4.1 Density plot after minmax normalization

4.3.2 Statistical analysis

After removing the lowly abundant miRNAs (miRNAs with summed up read count < 50) there were 503 miRNAs left. Statistical analysis is mainly used for finding the most dysregulated miRNAs out of 503 miRNAs in the data set. The most significant miRNA have the ability to identify strong distinction between AD and control sample. Therefore, using WMW test we calculated p values for each miRNA across AD and control samples in order to obtain those significant features. We used significance level of 0.05 to obtain 228 most significant set of miRNAs, as mentioned in the previous chapter. The null hypothesis is taken as that two distribution of AD and control samples for a particular miRNA is identical. This was rejected for those with significance value

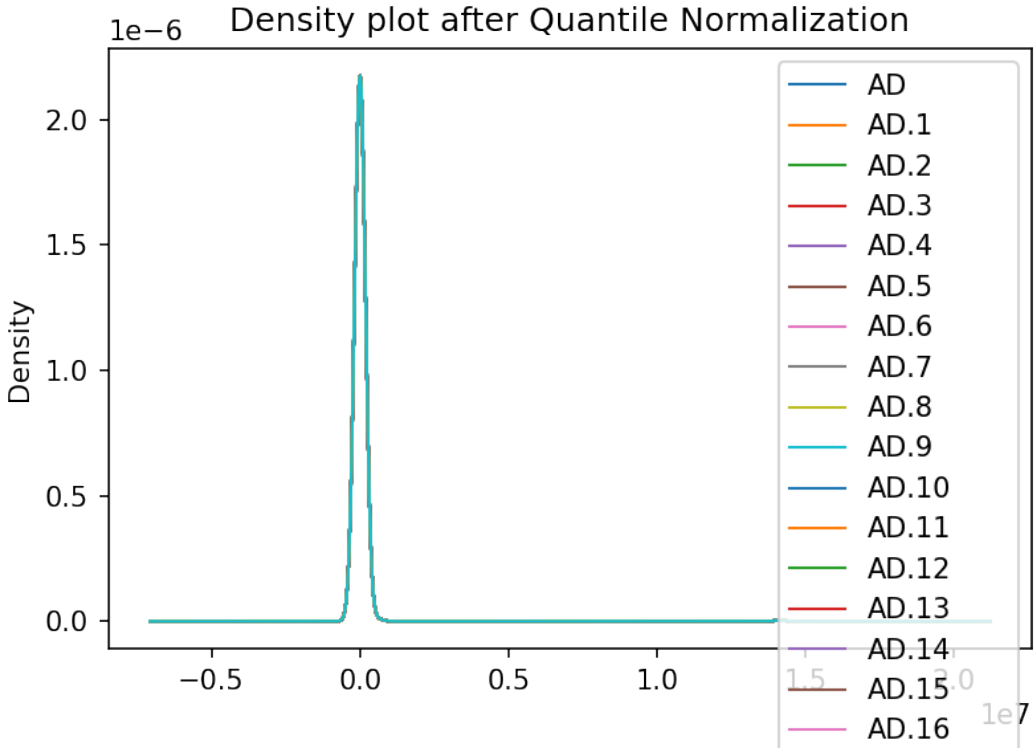

Fig. 4.2 Density plot after Quantile normalization

less than 0.05, making the alternative hypothesis, which is that two distributions for a miRNA is different, true. Using both the fold change values and the p values, we obtained 228 set of features with high significance values. For each of those significant miRNAs, AUC values were calculated for identifying the most up regulated and the most down regulated miRNAs. Using AUC values, we identified 219 down and up regulated miRNAs. As mentioned in the previous chapter, we also used the TAM tool for analyzing those selected significant features.

4.3.3 Feature Selection

After doing statistical analysis, there were 220 miRNAs left in the data set. To select biomarkers (features) out of them, three methods were used as PCA, Random Forest, and correlation coefficient analysis.

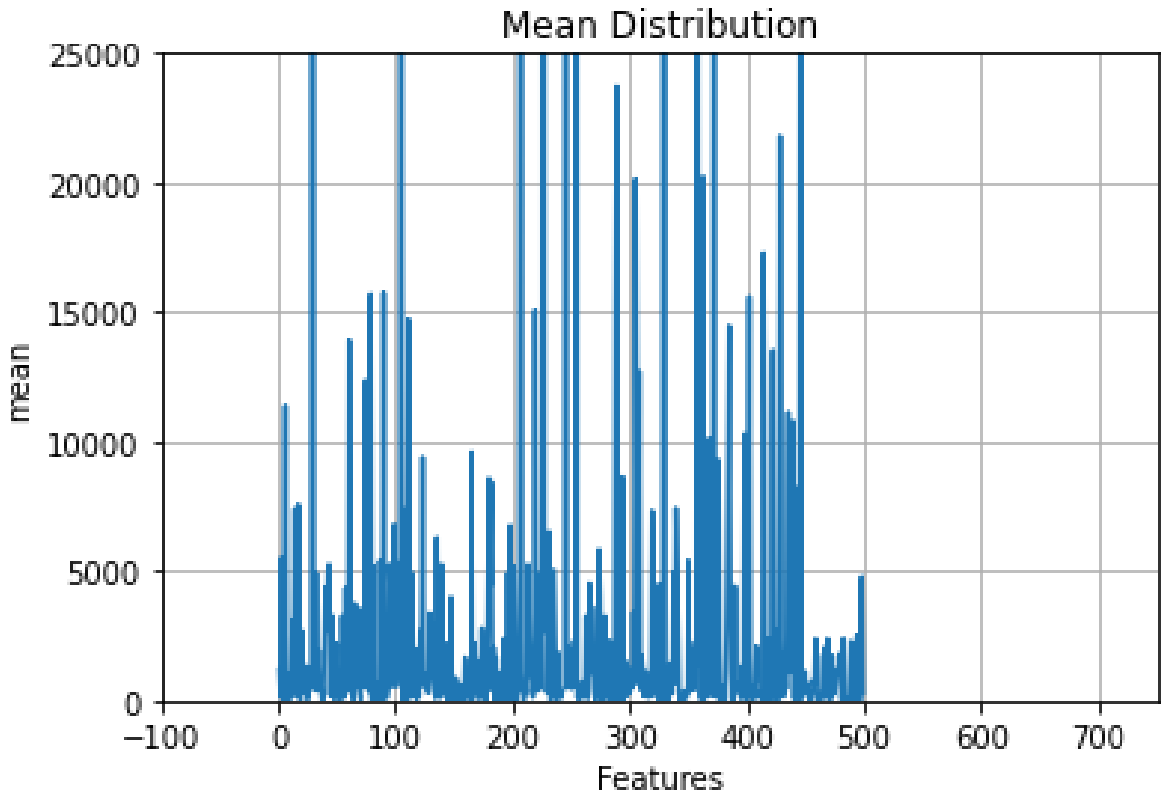

Fig. 4.3 Mean Distribution

PCA

Principal Components Analysis is used to transform high dimensional data to low dimension. PCA transforms a correlated feature set to an uncorrelated component set using the orthogonal transformation technique. In this research, we used PCA as a feature selection method. We selected high scored features (miRNAs) from each component. In PCA, each component consists of a set of correlated features. By selecting the highest scored features of each component, we founded the most effective and uncorrelated features (miRNAs).

Random Forest

Random forests are made out of multiple decision trees. Those decision trees are trained independently using random subsets of data. When training these decision trees, features are naturally ranked according to how well they improve the purity of each node. Using the measurement of decreasing impurity over all trees, the features can be ranked. The

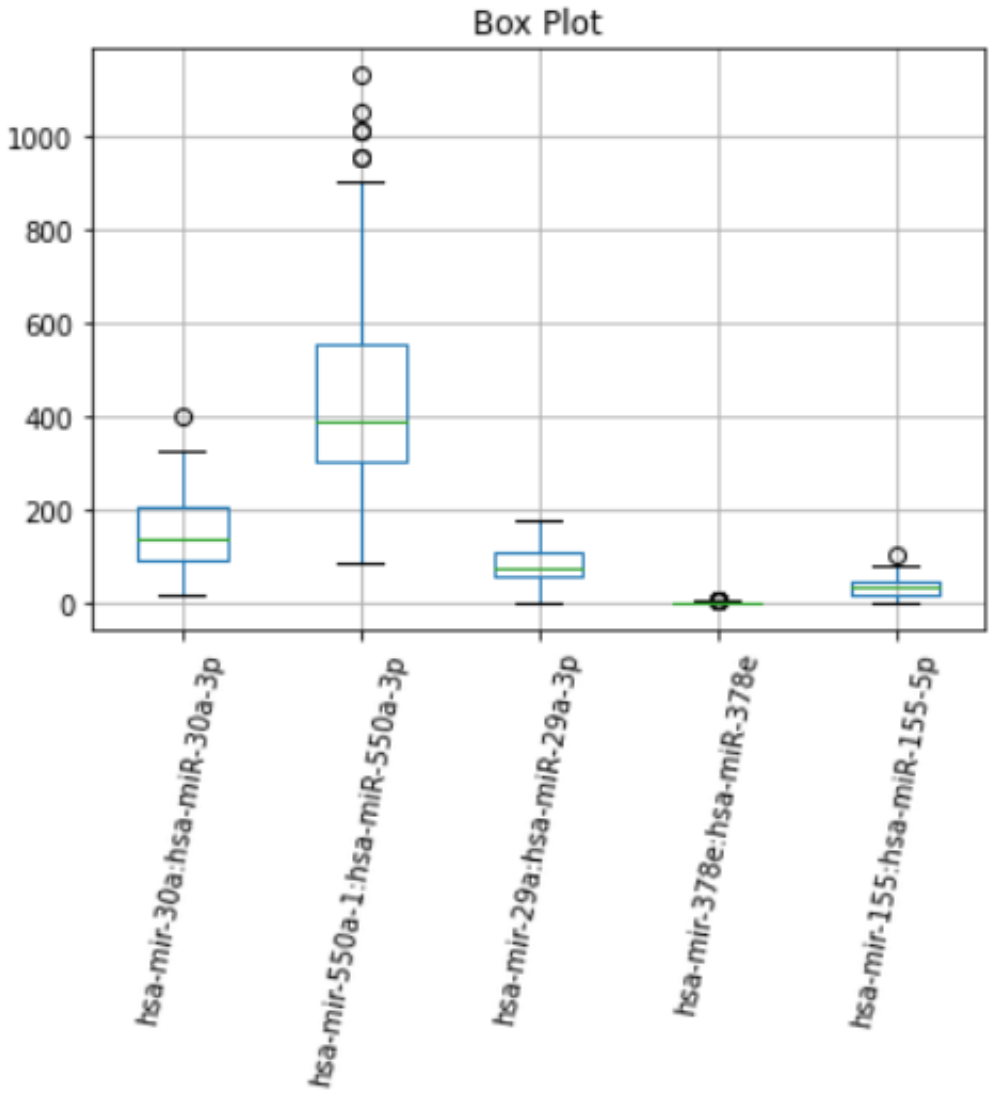

Fig. 4.4 Box plot of five features

sklearn library in python uses this measure to train decision trees. In this research, we selected features (miRNAs) with the highest scores. The highest score features are the features with the greatest decrease in impurity.

We selected 50 miRNAs from each method PCA and Random Forest. Then as our first biomarker miRNA set, we selected a set of features that overlaps from PCA and Random Forest analysis. Then, we obtained a set of less correlated features from both PCA

and Random Forest analysis. Those features were further analyzed using correlation coefficient values.

Correlations

The correlation matrix shows how features are correlated with each other. Figure 4.5 shows the correlation heatmap of selected 220 features from statistical analysis. Correlation coefficients varying from 1 to 0. 1 for the highest correlated features and 0 for the least correlated features.

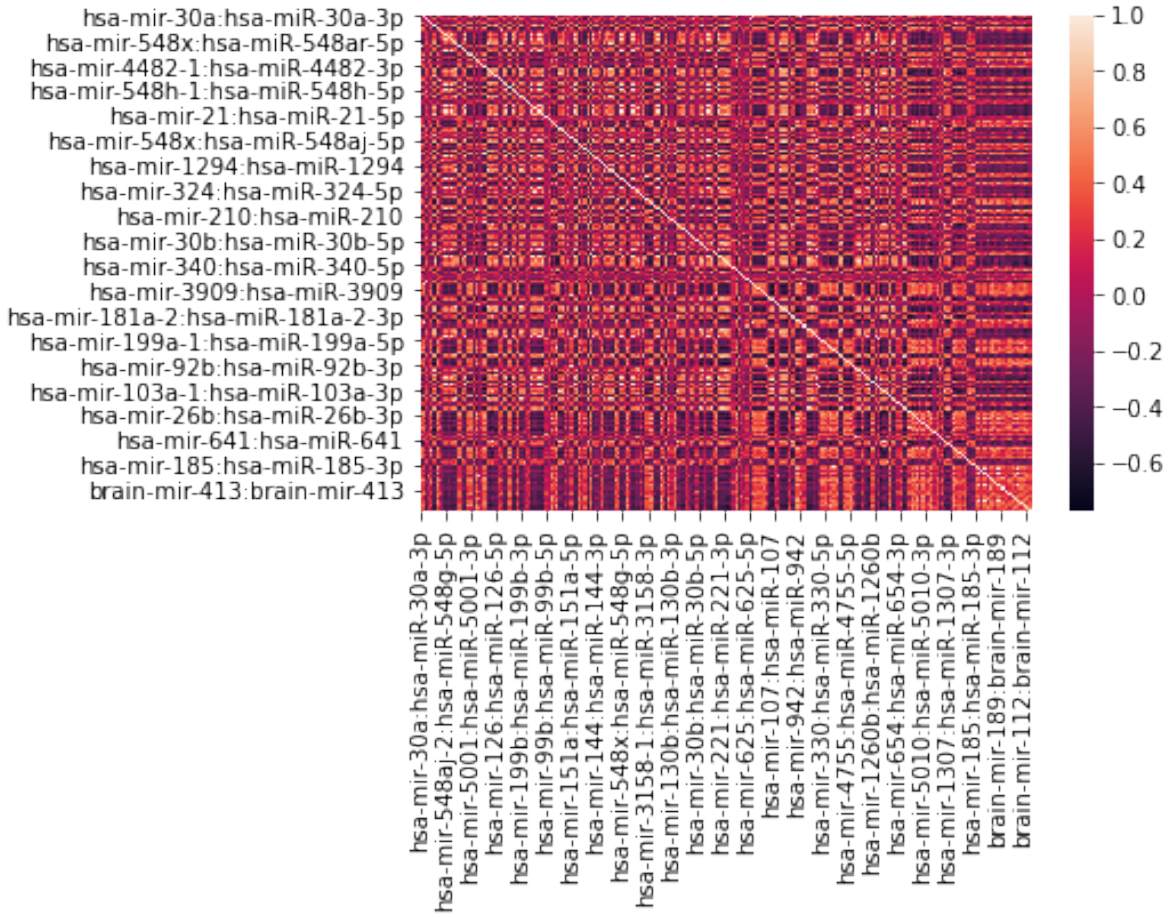

Fig. 4.5 Correlation heatmap for 220 miRNAs selected from statistical analysis

After selecting the first biomarker set, there were 36 miRNAs left in each set selected from PCA and random Forest. Using correlation heatmaps we found a less correlated set from both PCA and Random Forest. By reducing the less correlated set according to

the correlation coefficient we found the prediction accuracies. Then we selected the set with the highest accuracy as our next biomarker set.

4.3.4 Classification

As mentioned in the previous chapter, for some selected classifiers, we fit our dataset with 503 miRNA to analyze the accuracies (Table 4.1) of each classifier. We used this analysis to select the best fit model for using in the classification of AD and control samples for the selected miRNAs.

Table 4.1 Accuracies of different algorithms modelled for the dataset

Model	Training	Testing
Logistic regression	0.82	0.71
Naive Bayes	0.73	0.50
KNN	0.78	0.50
Linear SVM	0.83	0.71
Random Forest	1	0.64
Gaussian SVM	0.73	0.5

From the above table, we selected three machine learning classification algorithms for classification. Considering both train and test accuracy we selected linear Support Vector Machine (SVM), Logistic Regression and Random Forest for classification.

4.3.5 Result Validation

For validate our results we used Human MiRNA Disease Database version 3.2 (HMDD v3.2). This database contains a significant number of miRNAs and related diseases which was collected from literature. [22] The version 3.2 was released in 27th March 2019. The HMDD v3.2 is available at https://www.cuilab.cn/hmdd. It is a freely accessible database.

4.3.6 Pitfalls and workarounds

When starting this project, the main challenge that we have to face was lack of background knowledge about the bioinformatics. Initially we didn't properly understand what kind of data that generated from Next Generation Sequencing platforms. Also we hadn't knowledge to how to analyze those sequencing data. To overcome that issue we had to

do lots of background searches. Also we had to use some articles and tutorials to learn about next generation sequencing data analysis tools. Another problem that we had to face was finding a proper NGS miRNA Alzheimer's disease dataset for analysis. After searching a lot we found GSE46579 from NCBI public database. But the initial dataset was large (about 33GB) and there were fastq type datasets for each sample. So it was hard to download each dataset to a PC and perform analysis. So we used the galaxy data analysis platform for NGS data analysis. In the galaxy, there was a way to automate the analyzing procedure. Also, the data was used from the NCBI database without downloading it to the PC. Because of the lack of knowledge to find a new methodology at once we selected some articles and followed their methods to data analysis first. Because of the separated samples of data we had to create a summarized data set for further analysis after NGS analysis. When doing statistical analysis, we had to go through a lot of theories to understand the concepts since we were not familiar with most of those.

Chapter 5

Results and Analysis

5.1 Results

5.1.1 NGS analysis

There are 69 sample data files in the initial database. They contained NGS data in fastq format (Figure 5.1)

```
@SRR837486.1 HWI-ST937:130:D10R9ACXX:6:1101:1641:1978 length=50
TCCTGTACTGAGCTGCCCCGAGATGGAATTCTCGGGTGCCAAGGAACTCC
+SRR837486.1 HWI-ST937:130:D10R9ACXX:6:1101:1641:1978 length=50
CCCFFEEFHHHHHJJJJJIHHIFCCECHDHIIGIGH)8?BDH;B8CF<=F
```

Fig. 5.1 One read from fastq data file of sample SRR837486

As in Figure 5.1, the fastq sequencing data file contains the header, the sequence (+ denotes the end of the sequence), and the quality score of sequences for each read. For the sequencing data, FastQC generated a quality report about the sequences. It contains basic statistics about the file, per base, and per tile sequence quality, details about sequence length, and details about the adapter content of sequences.

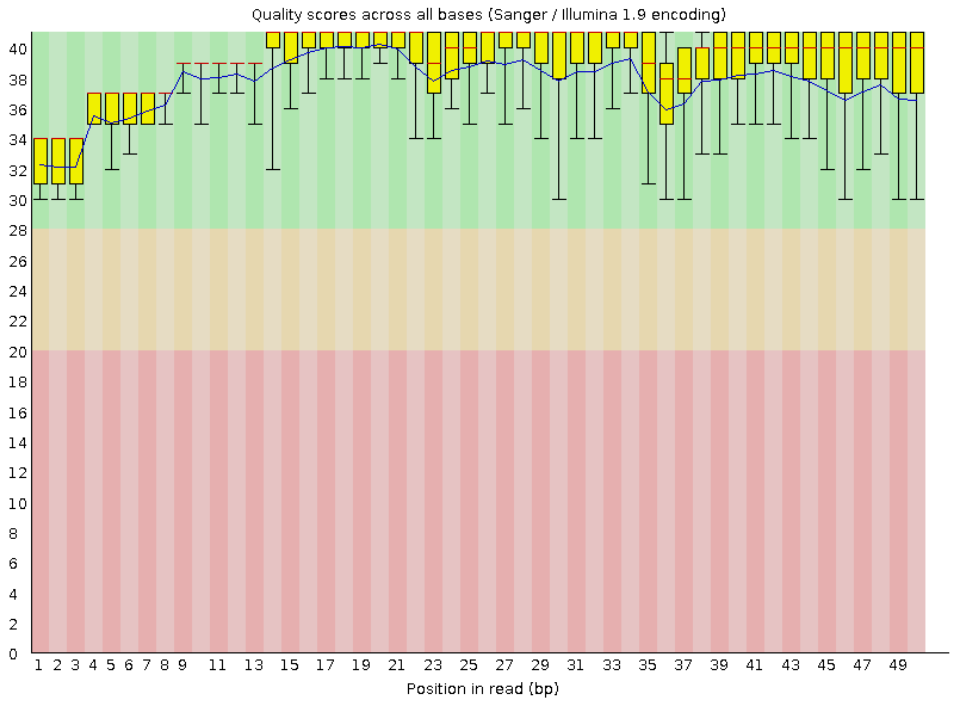

Fig. 5.2 Per base sequence quality for sample SRR837486

Figure 5.2 is a BoxWhisker type plot that shows an overview quality ranges of each sequence in fastq file of sample SRR837486. Yellow box represents 25 Adapters in sequences contain index sequences, primer binding sites, and the sites that allow fragments to connect with flow cells. Figure 5.3 shows the adapter contents of sequencing data in sample SRR837486. According to Figure 5.3, there was Illumina small RNA 3 adapter in row sequences.

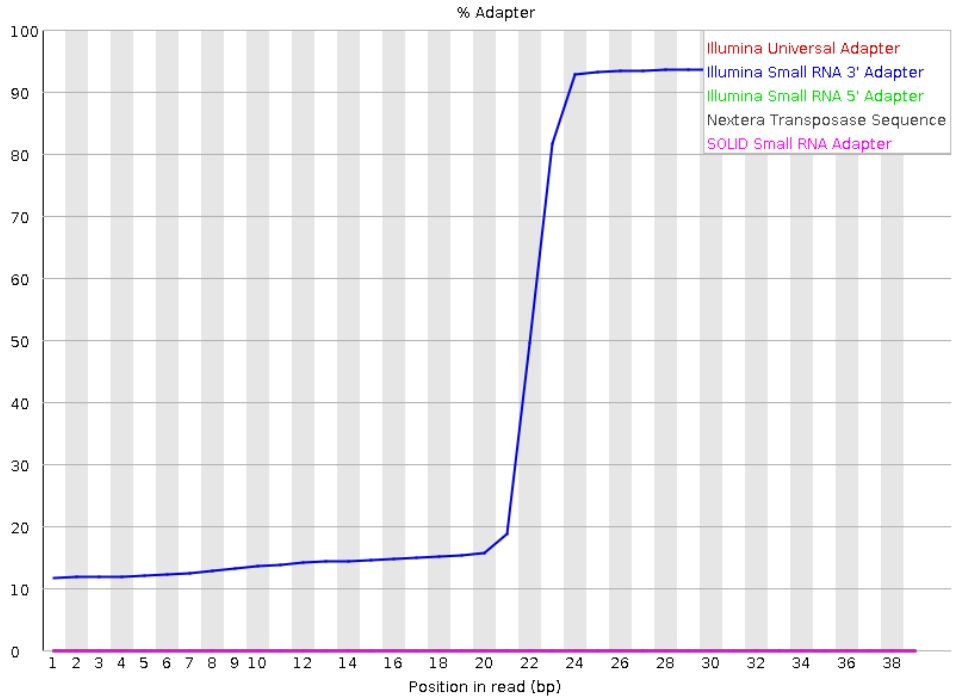

Fig. 5.3 Adapter contents of data

The NGS data preprocessing procedure contains trimming and filtering. Trimming and filtering were performed to improve the quality of data. In this procedure, the adapters and the low quality reads were removed.

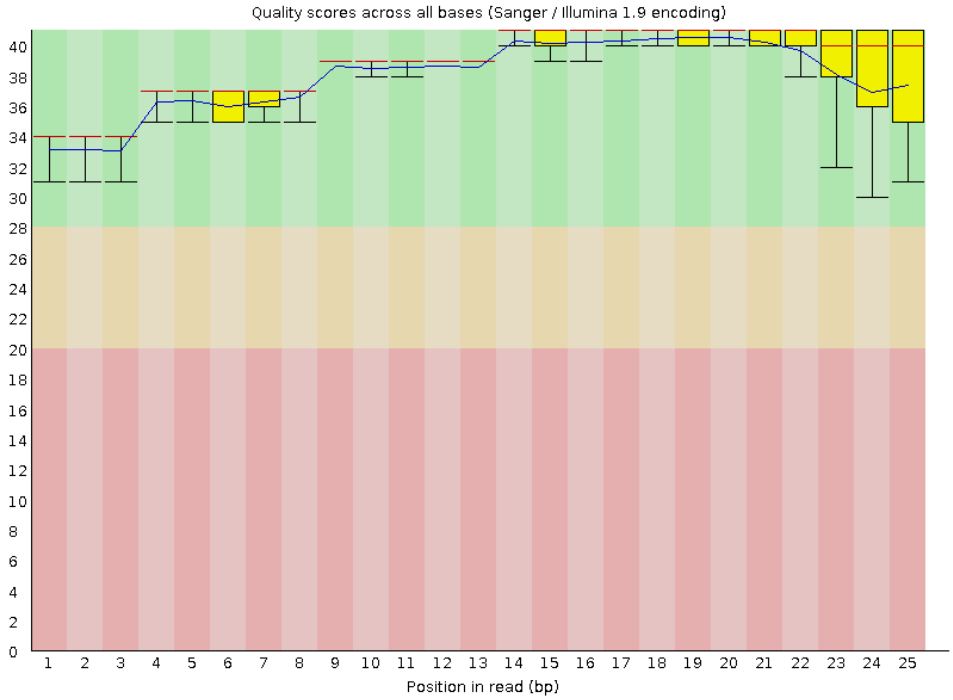

Fig. 5.4 Per base sequence quality for sample SRR837486 (After preprocessing)

Figure 5.4 shows the BoxWhisker plot for the sequence quality of sample data. As in Figure 5.4, the sequence quality was improved and the length of the sequences was reduced.

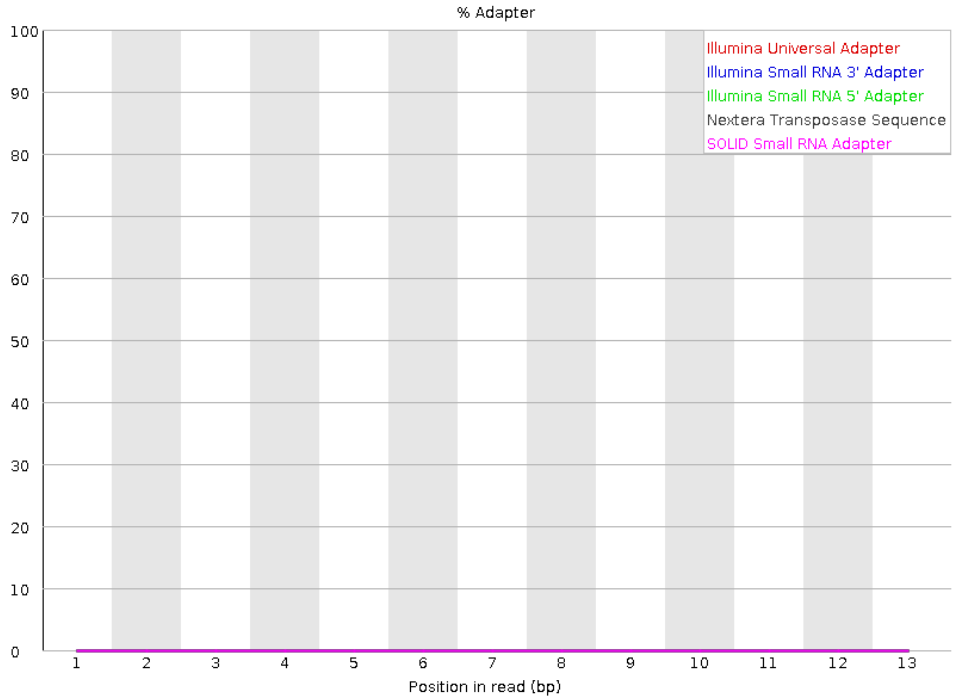

Fig. 5.5 Adapter contents of data (After preprocessing)

As show in Figure 5.5, all adapter contents were removed in the preprocessing procedure. There were about 13 million reads in sample SRR837486 at the beginning. After preprocessing, about 2 million reads were removed which are considered as low quality sequences or short read sequences. Then, the mapping to the reference genome resulted in the BAM type file and using miRBase we identified the read counts of each miRNA of each sample (Figure 5.6).

Geneid	Bowtie2 on data 11: alignments
hsa-let-7a-2-3p	0
hsa-let-7a-3p	0
hsa-let-7a-5p	9996
hsa-let-7b-3p	45
hsa-let-7b-5p	4190
hsa-let-7c-3p	0
hsa-let-7c-5p	151
hsa-let-7d-3p	345
hsa-let-7d-5p	1096
hsa-let-7e-3p	0
hsa-let-7e-5p	21
hsa-let-7f-1-3p	0
hsa-let-7f-2-3p	0
hsa-let-7f-5p	3131
hsa-let-7g-3p	0
hsa-let-7g-5p	2088
hsa-let-7i-3p	41
hsa-let-7i-5p	2986
hsa-miR-1-3p	0
hsa-miR-1-5p	0
hsa-miR-100-3p	0
hsa-miR-100-5p	1197
hsa-miR-101-2-5p	0

Fig. 5.6 miRNA read counts of SRR837486

In this procedure we identified 2652 miRNAs. By removing <50 summed up read counts we created a summarized data set for further analysis. After removing <50 read counts, there were only 503 miRNAs were left.

5.1.2 Statistical analysis

228 miRNAs were identified as the most statistically significant miRNAs using the significance level of 0.05 and the fold change of |1| (>|1|). Then, 219 miRNAs were obtained from the feature selection method, which was done based on the univariate ROC AUC classification. According to the AUC values of each miRNA, we identified the most down regulated miRNA as hsa-mir-361:hsa-miR-361-5p (AUC = 0.875) and the two most

up regulated miRNAs as hsa-mir-181a-1:hsa-miR-181a-3p and hsa-mir-98:hsa-miR-98 (AUC = 0.25). Figure 5.7 shows the distributions of those two miRNAs across AD and control samples.

As mentioned in the previous two chapters, the 228 features obtained using significance value and fold change, were also subjected to TAM tool analysis. From that we obtained 9 down regulated and 14 up regulated miRNAs which can be used for further analysis purposes.

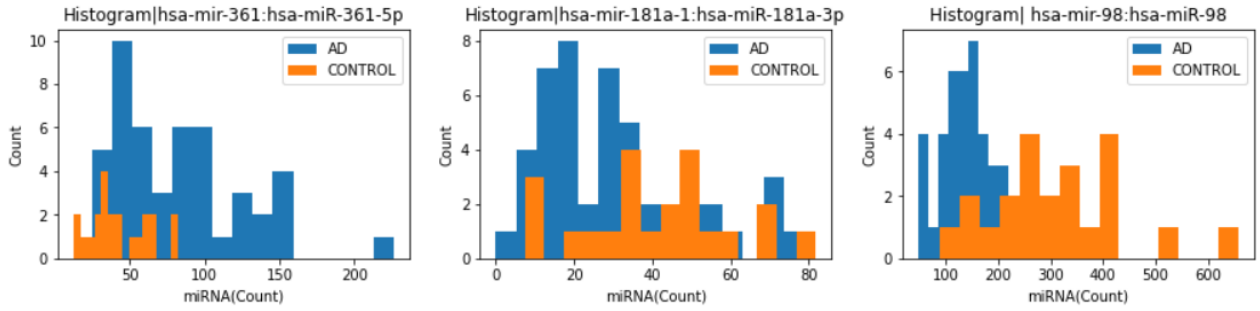

Fig. 5.7 Distributions of the most up regulated miRNA and the most down regulated miRNA

5.1.3 Feature selection

Figure 5.8 shows the plot of 219 miRNAs with univariate selection scores. It implied that first 50 miRNAs are more related to Alzheimer's disease.

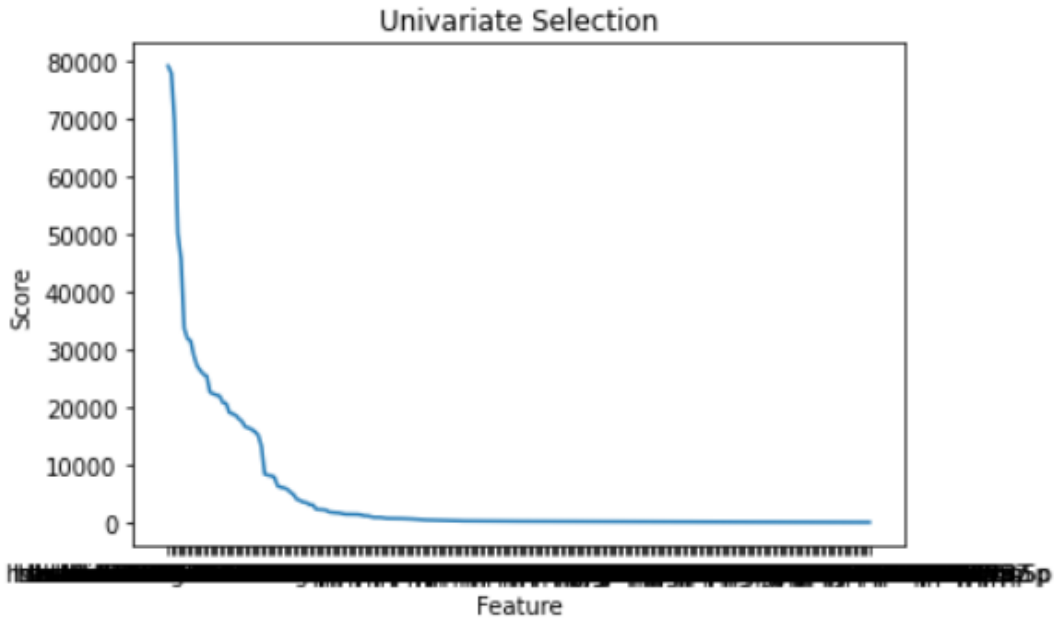

Fig. 5.8 Univariate selection for 219 miRNAs

As mentioned in the earlier chapters, PCA and Random Forest were used to select the most significant set of miRNAs. From each method, the best 50 miRNAs were selected for further analysis. The visualization of the overlap of the selected miRNAs using PCA and Random Forest shows in Figure 5.9 as a Venn diagram.

Fig. 5.9 Visualization of the overlap feature set selected from PCA and Random Forest

As shown in the Figure 5.9, 14 miRNAs are overlapped between PCA and RF. The overlapped results were more reliable and hence, the obtained 14 miRNAs were considered as biomarkers for Alzheimer's disease.

overlap
hsa-mir-186:hsa-miR-186-5p
hsa-mir-144:hsa-miR-144-3p
hsa-mir-151a:hsa-miR-151a-3p
hsa-mir-99b:hsa-miR-99b-5p
hsa-mir-98:hsa-miR-98
hsa-mir-148a:hsa-miR-148a-3p
hsa-let-7g:hsa-let-7g-5p
hsa-let-7f-2:hsa-let-7f-5p
hsa-let-7a-1:hsa-let-7a-5p
hsa-mir-30d:hsa-miR-30d-5p
hsa-mir-15a:hsa-miR-15a-5p
hsa-mir-589:hsa-miR-589-5p
hsa-mir-144:hsa-miR-144-5p
hsa-let-7f-1:hsa-let-7f-5p

Fig. 5.10 Selected miRNAs from overlap between PCA and RF

Figure 5.11 and Figure 5.12 show correlation heat maps for selected 50 miRNAs from PCA and Random Forest respectively.

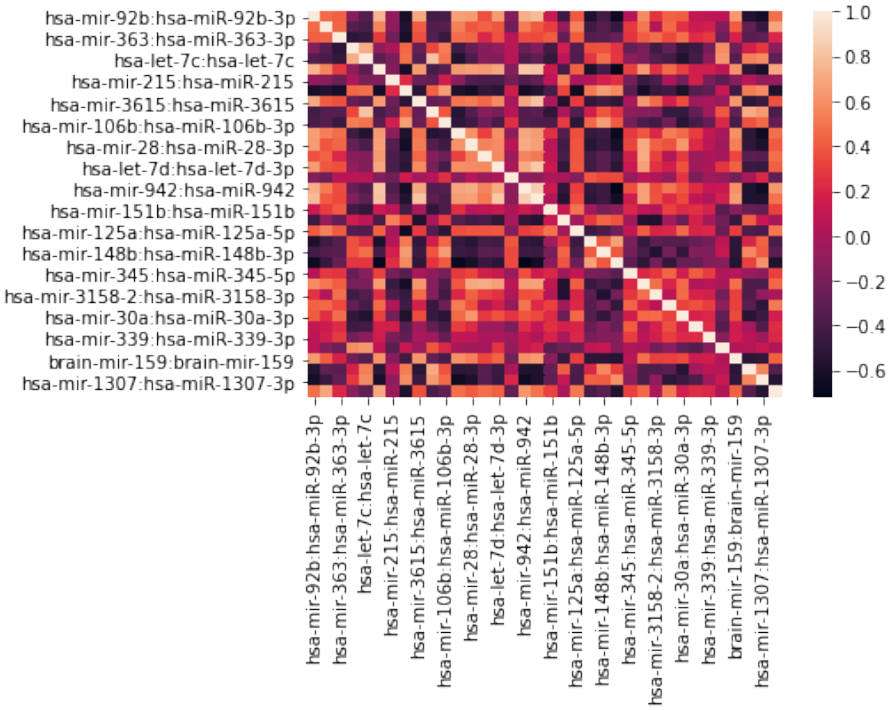

Fig. 5.11 Correlation Heatmap of selected miRNAs from PCA

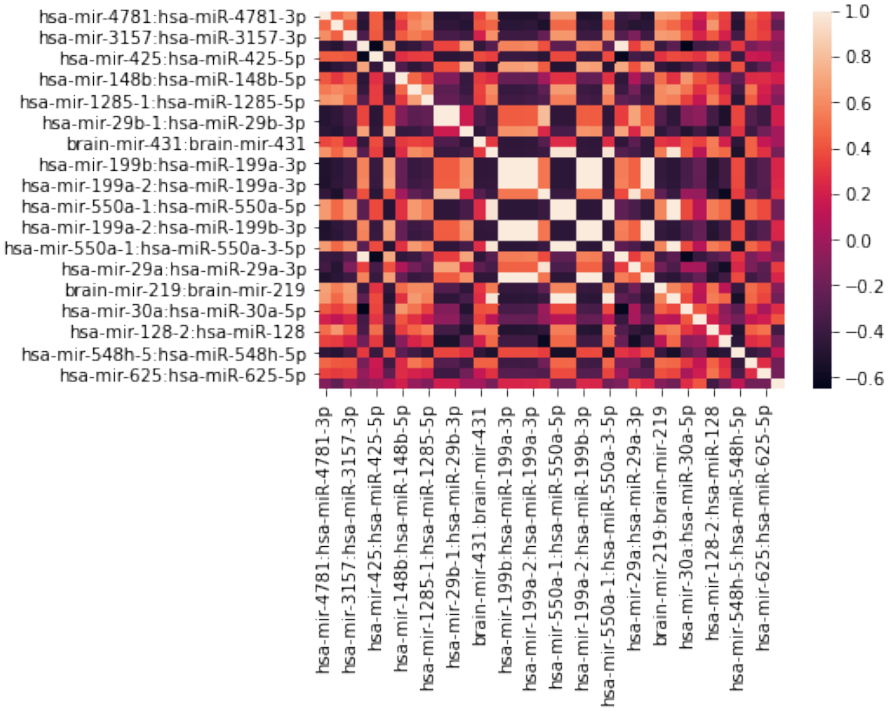

Fig. 5.12 Correlation Heatmap of selected miRNAs from Random Forest

From the above two heat maps, the one obtained from PCA was identified as the one with the less correlated miRNAs. Hence, it was further evaluated using correlation coefficient to obtain the most accurate set of miRNAs. Plot in the left hand side of the Figure 5.13 shows the classification accuracies of miRNAs obtained only from PCA evaluation with Linear SVM, Logistic Regression and Random Forest classifiers with respect to correlation coefficients. Plot shown in the right hand side of the Figure 5.13 shows how the number of features varies for different correlation coefficients. From this evaluation, 11 miRNAs were identified as the most accurate set of miRNAs. Figure 5.14 shows the set of identified 11 miRNAs.

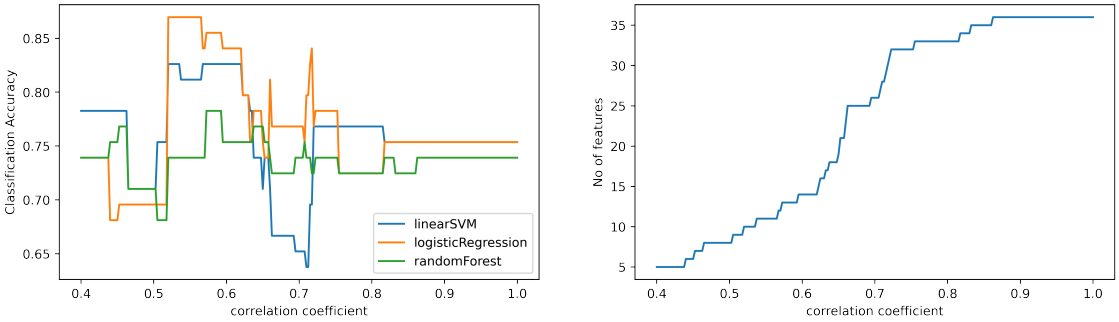

Fig. 5.13 Variation of the correlation coefficients of the miRNAs selected only by PCA with classification accuracy

Correlation
hsa-mir-4781:hsa-miR-4781-3p
brain-mir-112:brain-mir-112
hsa-let-7a-3:hsa-let-7a-5p
hsa-mir-148b:hsa-miR-148b-5p
hsa-mir-29b-2:hsa-miR-29b-3p
brain-mir-431:brain-mir-431
hsa-mir-378a:hsa-miR-378a-5p
hsa-mir-548h-5:hsa-miR-548h-5p
hsa-mir-3909:hsa-miR-3909
hsa-mir-625:hsa-miR-625-5p
hsa-mir-24-1:hsa-miR-24-3p

Fig. 5.14 Selected miRNAs from cross correlation

5.1.4 Classification

Over fitting of data happens when the machine learning algorithm fits the given data too well. Therefore to make good predictions it's important to avoid over fitting of data. In this project, we addressed that issue with cross validation. Three sub sets (3 folds) of the data set were taken and repeatedly each subset (train set) was fit on a model. As the 3-fold cross validation happens, the train and test data accuracies were calculated.

Table 5.1 Classification accuracy

ML model	Overlapped set	Overlapped + correlations set
Linear SVM	80.95%	85.71%
Logistic Regression	95.24%	95.24%
Random Forest	90.48%	95.24%

The classification accuracies were calculated for each data set obtained at different stages (Table 5.1). As in the table, classification accuracy of the set of miRNAs identified using correlation coefficients are high compared to others. Figure 5.15 shows a graphical representation of train and test data accuracies for the final set of miRNAs.

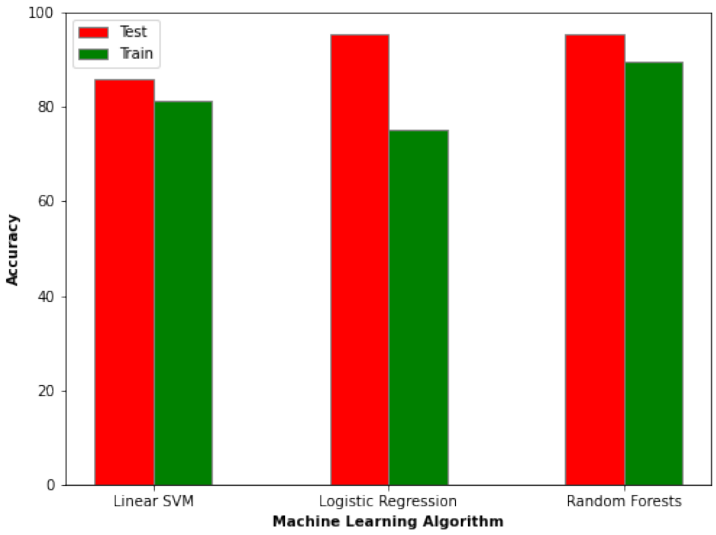

Fig. 5.15 Train and test data accuracy for the final set of miRNAs.

With the aid of this figure, we can conclude that the selected miRNAs are able to discriminate between AD and control well.

5.1.5 Validation

As the final result, 25 miRNAs were identified; 11 from less correlated set and 14 from the set of overlapped miRNAs between PCA and Random Forest. Those set of final miRNAs were validated using HMDD v3.2 as mentioned in chapter 3. Out of 25, 10 miRNAs were identified as validated miRNAs; including 2 from less correlated set and 8 from overlapped set. We identified 15 miRNAs as novel biomarkers. Figure 5.16 and Figure 5.17 shows the validated miRNAs using less correlated set and overlapped set respectively.

```
corr_validate
```

```
['hsa-mir-148b', 'hsa-mir-29b-2']
```

Fig. 5.16 Validated miRNAs from less correlated set

```
overlap_validate

['hsa-mir-186',
 'hsa-mir-144',
 'hsa-mir-151a',
 'hsa-mir-98',
 'hsa-mir-148a',
 'hsa-let-7g',
 'hsa-mir-15a',
 'hsa-mir-144']
```

Fig. 5.17 Validated miRNAs from overlapped set from both PCA and RF

5.1.6 Performance evaluation

Addition to the classification accuracy mentioned in the previous subsection, we can make a better argument for the possibility of our final miRNAs as biomarkers again, with the aid of AUC analysis. For that we used the model we developed with the Random Forest algorithm. We used true positive rate (TPR) and false positive rate (FPR) for this. They can be given as,

$$TPR = TruePositives/(TruePositives + FalseNegatives)$$

$$FPR = FalsePositives/(FalsePositive + TrueNegatives)$$

We acquired TPR or sensitivity value and FPR or the specificity value as 94.11% and 87.5% respectively for the Random Forest model which gave a test accuracy of 95.24%. Leidinger et al.[7], who have done a study with the same data set used in this project (GSE46579) have stated that their model's accuracy, sensitivity, and specificity were obtained as 93.3%, 91.5%, and 95.1% respectively.

Figure 5.19 shows the AUC analysis for the combined 25 miRNAs. Having an AUC value of 0.88 shows the power of the final 25 miRNAs to differentiate between AD and controls.

Variation of true positive rate with False positive rate of five of the most differentially expressed miRNAs are given with their AUC values in percentages to present their ability on differentiating between AD and controls (Figure 5.18).

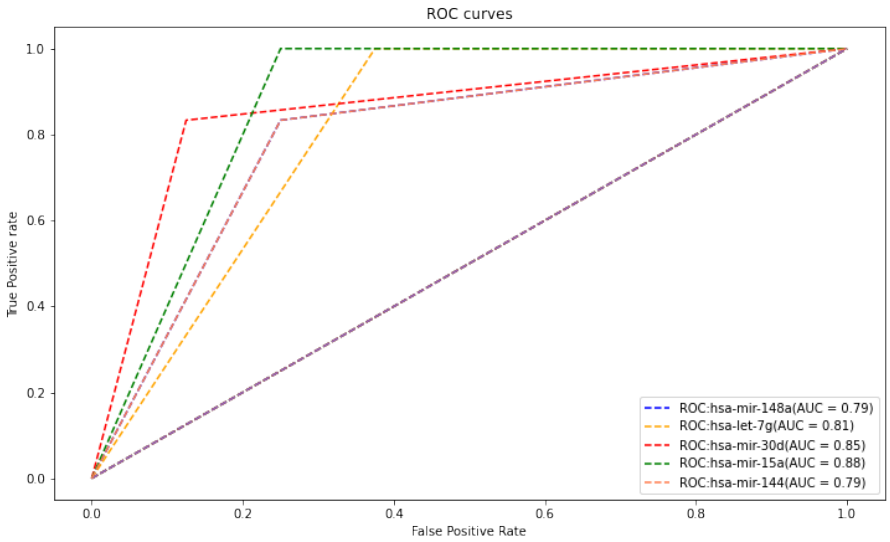

Fig. 5.18 ROC curves of the four most differentially expressed miRNAs.

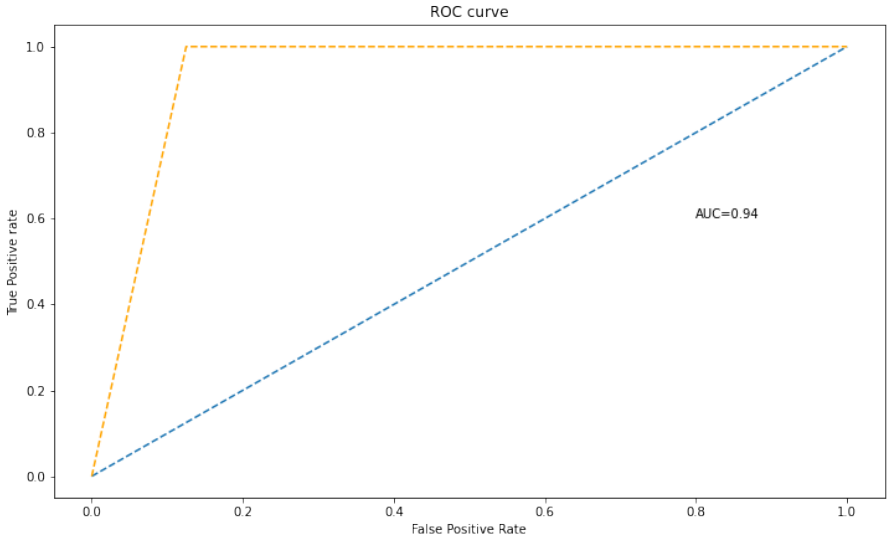

Fig. 5.19 ROC curve for the combined set of 25 miRNAs .

Out of the identified 25 miRNAs, *hsa-mir-186, hsa-mir-144, hsa-mir-151a, hsa-mir-98, hsa-mir-148a, hsa-let-7g, hsa-mir-15a, hsa-mir-144, hsa-mir-148b, and hsa-mir-29b-2* are found in the HMDD v3.2 data base. Among these 10 miRNAs, 4 miRNAs out of five which we identified previously as the most differentially expressed, are found.

5.1.7 Application Of Another Data Set To The Developed Methodology

As mentioned in the chapter 1, we applied another data set which is stored under the accession number GSE147218 in the NCBI database, to the same methodology to observe how our methodology responds to different data sets. Thus, we carried out the following steps for this data set as well.

- Reduced lowly abundant miRNAs
- Quantile normalization
- Identified most significant miRNAs using p value and fold change
- Identified dysregulated miRNAs using AUC
- Selected final set of biomarker miRNAs from PCA, Random Forest analysis and correlation coefficient values

At the end of the feature selection stage we obtained 9 miRNAs as the biomarkers for AD. Those 9 miRNAs we obtained as the final results are *hsa-miR-193b-5p, hsa-miR-1287, hsa-let-7f-1, hsa-miR-1283, hsa-miR-4703, hsa-miR-155, hsa-miR-1254, hsa-miR-3131, and hsa-miR-26b.*

5.1.8 Details of the Methods and comparison of Results with previous methods

Details of methods used and results obtained are given in the Table 5.2.

Table 5.2 Detailed summary of the methods used and results obtained in previous studies

Article	**Methods**	**Identified miRNAs**	**Overlap with our results**
[6]	P value, two-tailed t tests, Wilcoxon Mann-Whitney (WMW) tests, AUC Hierarchical clustering PCA Analysis of variance (ANOVA) SVM for classification	hsa-miR-107, hsa-miR-103a-3p, hsa-let-7d-3p, hsa-let-7f-5p, hsa- miR-151a-3p, hsa-miR-1285-5p, hsa-miR-532-5p, hsa-miR-5010-3p, hsa-miR-26a-5p, hsa-miR-26b-5p (10 miRNAs)	hsa-let-7f-5p, hsa-miR-151a-3p
[15]	P value, Fold change , two-tailed t- test, ANOVA, Kruskal-Wallis test, nonparametric Wil-coxon test, Correlation analysis, AUC J48 decision trees, SVM, adaboostM1 for classification ML analysis	hsa-miR-185-5p, hsa-miR-342-3p, hsa-miR-141-3p, hsa-miR-342-5p, hsa-miR-23b-3p, hsa-miR-338-3p, hsa-miR-3613-3p (7 miRNAs)	
[11]	P value, Fold change, AUC, Mann- Whitney test, ANOVA, Pearson's correlation coefficient	hsa-miR-98-5p, hsa-miR-885-5p, hsa-miR-483-3p, hsa-miR-342-3p, hsa-miR-191-5p, hsa-let-7d-5p (6 miRNAs)	
Continued on next page			

Table 5.2 – continued from previous page

Article	Methods	Identified miRNAs	Overlap with our results
[12]	P value, Fold change Partek Genomics Suite, ANOVA, Generalized linear modeling, Random Forest analyses for Classification and Prediction	hsa-miR-30e-5p, hsa-miR-101-3p, hsa-miR-15a-5p, hsa-miR-20a-5p, hsa-miR-93-5p, hsa-miR-106b-5p, hsa-miR-18b-5p, hsa-miR-106a-5p, hsa-miR-1306-5p, hsa-miR-582-5p, hsa-miR-143-3p, hsa-miR-335-5p, hsa-miR-361-5p, hsa-miR-424-5p, hsa-miR-342-3p, hsa-miR-15b-3p (16 miRNAs)	hsa-miR-15a-5p
[13]	ANOVA with Bonferroni's post-hoc test, Chi-square test, Student's t- test, AUC, P value	hsa-miR-26a-5p, hsa-miR-181c-3p, hsa-miR-126-5p, hsa-miR-22-3p, hsa-miR-148b-5p, hsa-miR-106b-3p, hsa-miR-6119-5p, hsa-miR- 1246, hsa-miR-660-5p (9 miRNAs)	hsa-miR-148b-5p

Continued on next page

Table 5.2 – continued from previous page

Article	Methods	Identified miRNAs	Overlap with our results
[7]	P value, Fold change, AUC, Wilcoxon-Mann-Whitney (WMW) test SVM for classification	brain-miR-112, brain-miR-161, hsa-let-7d-3p, hsa-miR-5010-3p, hsa-miR-26a-5p, hsa-miR-1285-5p, hsa-miR-151a-3p, hsa-miR-103a- 3p, hsa-miR-107, hsa-miR-532-5p, hsa-miR-26b-5p, hsa-let-7f-5p (12 miRNAs)	brain-miR-112 hsa-miR-151a-3p hsa-let-7f-5p
[8]	Fisher's exact test, P value, fold change, AUC, false discovery rate (FDR)	hsa-miR-146b-5p, hsa-miR-15b-5p (2 miRNAs)	
[16]	P value, , two-tailed t-test, Kruskal-Wallis non-parametric test, Dunn's test, AUC	hsa-miR-21-5p, hsa-miR-26a-5p, hsa-let-7i-5p, hsa-miR-126-3p, hsa-miR-451a, hsa-miR-23a-3p, hsa-let-7f-5p, hsa-miR-409-3p, hsa-miR-92a-3p, hsa-let-7b-5p, hsa-miR-151a-3p, hsa-miR-24-3p, hsa-miR-143-3p, hsa-miR-423-5p, hsa-miR-183-5p (15 miRNAs)	hsa-let-7f-5p hsa- miR-151a-3p hsa- miR-24-3p
[17]	Wald test, P value, Fold change, Fisher's exact test	hsa-miR-501-3p, hsa-let-7f-5p, hsa-miR-26b-5p (3 miRNAs)	hsa-let-7f-5p

Continued on next page

Table 5.2 – continued from previous page

Article	Methods	Identified miRNAs	Overlap with our results
[9]	P value, ROC, hierarchical clustering analysis (HCA) with omiRas	hsa-miR-26b-3p, hsa-miR-28–3p, hsa-miR-30c-5p, hsa-miR-30d-5p, hsa-miR-148b-5p, hsa-miR-151a-3p, hsa-miR-186–5p, hsa-miR-425–5p, hsa-miR-550a-5p, hsa-miR-1468, hsa-miR-4781–3p, hsa-miR-5001–3p, hsa-miR-6513–3p, hsa-let-7a-5p, hsa-let-7e-5p, hsa-let-7f-5p,hsa- let-7g-5p, hsa-miR-15a-5p, hsa-miR-17–3p, hsa-miR-29b-3p, hsa-miR-98–5p, hsa-miR-144–5p, hsa-miR-148a-3p,hsa-miR-502–3p, hsa-miR-660–5p, hsa-miR-1294, hsa-miR-3200–3p (27 miRNAs)	hsa-miR-148b-5p, hsa-miR-151a-3p, hsa-miR-186–5p, hsa-miR-4781–3p, hsa-let-7a-5p, hsa-let-7f-5p,hsa- let-7g-5p, hsa-miR-15a-5p, hsa-miR-29b-3p, hsa-miR-144–5p, hsa-miR-148a-3p
[14]	P value, Fold change, AUC Mann-Whitney U-test Student's t-test/two-sided $X^2 test$	hsa-miR-31, hsa-miR-93, hsa-miR-143, hsa-miR-146a	

Chapter 6

Conclusions and Future Works

6.1 Conclusion

In this report we have discussed about how to detect miRNA biomarkers for Alzheimer's disease using next generation sequencing. Initially we have discussed about the need of a solution to identify Alzheimer's disease in the early stage. Then we have mentioned about the literature review we have done. When we were doing the literature review, we have identified several miRNA biomarkers in different studies which used NGS. In these studies there were some limitations. In our approach so far, initially we have taken samples from participants with AD and control. Then samples were preprocessed and statistically analyzed. Significance values were calculated using Wilcoxon-Mann-Whitney (WMW) test. Also we have used ROC analysis. Using this procedure, here we have identified a set of significant miRNAs for AD. Using PCA, Random Forests and Correlation coefficient we identified 25 biomarker miRNAs for AD. In the next phase we validated the the result using HMDD v3.2. Leidinger et al.[7], who have carried out a different method to find biomarkers using the same data set, have stated that they have obtained an accuracy of 93.3% where we obtained an accuracy of 95.24%. Addition to that we evaluated the results with specificity, sensitivity and AUC values as discussed previously. In addition to diagnosis of AD patients with the final set of biomarkers, the followed methodology can be used to identify different cures for other neurological diseases including AD, by effortlessly analyzing various data sets.

6.2 Future Work

For making the best out of this methodology, we are planning to create a web based tool to visualize the distributions of datasets, preprocessing, statistical analysis and classification. It will be useful for people who are working with genomic data science and also for the clinical use.

References

[1] S. Behjati and P. S. Tarpey, "What is next generation sequencing?," *Archives of Disease in Childhood-Education and Practice*, vol. 98, no. 6, pp. 236–238, 2013.

[2] S. Gholamin, A. Pasdar, M. Sadegh Khorrami, H. Mirzaei, H. Reza Mirzaei, R. Salehi, G. A Ferns, M. Ghayour-Mobarhan, and A. Avan, "The potential for circulating micrornas in the diagnosis of myocardial infarction: a novel approach to disease diagnosis and treatment," *Current pharmaceutical design*, vol. 22, no. 3, pp. 397–403, 2016.

[3] G. Stiglic, M. Bajgot, and P. Kokol, "Gene set enrichment meta-learning analysis: next-generation sequencing versus microarrays," *BMC bioinformatics*, vol. 11, no. 1, p. 176, 2010.

[4] M. Basavaraju and A. De Lencastre, "Alzheimer's disease: presence and role of micrornas," *Biomolecular concepts*, vol. 7, no. 4, pp. 241–252, 2016.

[5] C. Nie, Y. Sun, H. Zhen, M. Guo, J. Ye, Z. Liu, Y. Yang, and X. Zhang, "Differential expression of plasma exo-mirna in neurodegenerative diseases by next-generation sequencing," *Frontiers in Neuroscience*, vol. 14, p. 438, 2020.

[6] A. Keller, C. Backes, J. Haas, P. Leidinger, W. Maetzler, C. Deuschle, D. Berg, C. Ruschil, V. Galata, K. Ruprecht, *et al.*, "Validating alzheimer's disease micro rnas using next-generation sequencing," *Alzheimer's & Dementia*, vol. 12, no. 5, pp. 565–576, 2016.

[7] P. Leidinger, C. Backes, S. Deutscher, K. Schmitt, S. C. Mueller, K. Frese, J. Haas, K. Ruprecht, F. Paul, C. Stähler, *et al.*, "A blood based 12-mirna signature of alzheimer disease patients," *Genome biology*, vol. 14, no. 7, p. R78, 2013.

[8] H. Z. Y. Wu, A. Thalamuthu, L. Cheng, C. Fowler, C. L. Masters, P. Sachdev, K. A. Mather, A. I. Biomarkers, and L. F. S. of Ageing, "Differential blood mirna expression

in brain amyloid imaging-defined alzheimer's disease and controls," *Alzheimer's Research & Therapy*, vol. 12, pp. 1–11, 2020.

[9] J.-i. Satoh, Y. Kino, and S. Niida, "Microrna-seq data analysis pipeline to identify blood biomarkers for alzheimer's disease from public data," *Biomarker insights*, vol. 10, pp. BMI–S25132, 2015.

[10] C. Backes, B. Meder, M. Hart, N. Ludwig, P. Leidinger, B. Vogel, V. Galata, P. Roth, J. Menegatti, F. Grässer, *et al.*, "Prioritizing and selecting likely novel mirnas from ngs data," *Nucleic acids research*, vol. 44, no. 6, pp. e53–e53, 2016.

[11] L. Tan, J.-T. Yu, M.-S. Tan, Q.-Y. Liu, H.-F. Wang, W. Zhang, T. Jiang, and L. Tan, "Genome-wide serum microrna expression profiling identifies serum biomarkers for alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 40, no. 4, pp. 1017–1027, 2014.

[12] L. Cheng, J. Doecke, R. Sharples, V. Villemagne, C. Fowler, A. Rembach, and B. Australian Imaging, "Lifestyle (aibl) research group.(2015). prognostic serum mirna biomarkers associated with alzheimer's disease shows concordance with neuropsychological and neuroimaging assessment," *Molecular Psychiatry*, vol. 20, no. 10, pp. 1188–1196.

[13] R. Guo, G. Fan, J. Zhang, C. Wu, Y. Du, H. Ye, Z. Li, L. Wang, Z. Zhang, L. Zhang, *et al.*, "A 9-microrna signature in serum serves as a noninvasive biomarker in early diagnosis of alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 60, no. 4, pp. 1365–1377, 2017.

[14] H. Dong, J. Li, L. Huang, X. Chen, D. Li, T. Wang, C. Hu, J. Xu, C. Zhang, K. Zen, *et al.*, "Serum microrna profiles serve as novel biomarkers for the diagnosis of alzheimer's disease," *Disease markers*, vol. 2015, 2015.

[15] G. Lugli, A. M. Cohen, D. A. Bennett, R. C. Shah, C. J. Fields, A. G. Hernandez, and N. R. Smalheiser, "Plasma exosomal mirnas in persons with and without alzheimer disease: altered expression and prospects for biomarkers," *PloS one*, vol. 10, no. 10, p. e0139233, 2015.

[16] A. Gámez-Valero, J. Campdelacreu, D. Vilas, L. Ispierto, R. Reñé, R. Álvarez, M. P. Armengol, F. E. Borràs, and K. Beyer, "Exploratory study on microrna profiles from plasma-derived extracellular vesicles in alzheimer's disease and dementia with lewy bodies," *Translational neurodegeneration*, vol. 8, no. 1, p. 31, 2019.

[17] N. Hara, M. Kikuchi, A. Miyashita, H. Hatsuta, Y. Saito, K. Kasuga, S. Murayama, T. Ikeuchi, and R. Kuwano, "Serum microrna mir-501-3p as a potential biomarker related to the progression of alzheimer's disease," *Acta neuropathologica communications*, vol. 5, no. 1, p. 10, 2017.

[18] P. Uva, G. Cuccuru, A. Bruselles, G. Marangi, and T. Pippucci, "A galaxy training platform on "data analysis and interpretation for clinical genomics" sponsored by sigu," 11 2019.

[19] S. V. Toshchakov, I. V. Kublanov, E. Messina, M. M. Yakimov, and P. N. Golyshin, "Genomic analysis of pure cultures and communities," in *Hydrocarbon and Lipid Microbiology Protocols*, pp. 5–27, Springer, 2015.

[20] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, "mirbase: from microrna sequences to function," *Nucleic acids research*, vol. 47, no. D1, pp. D155–D162, 2019.

[21] S. C. Hicks and R. A. Irizarry, "When to use quantile normalization?," *bioRxiv*, 2014.

[22] Z. Huang, J. Shi, Y. Gao, C. Cui, S. Zhang, J. Li, Y. Zhou, and Q. Cui, "HMDD v3.0: a database for experimentally supported human microRNA–disease associations," *Nucleic Acids Research*, vol. 47, pp. D1013–D1017, 10 2018.