Using Machine Learning Approaches To Create The Most Significant Alzheimer's Disease (AD) Prediction.

Ulster University Magee Campus, Northern Ireland.
MSc Smart Manufacturing Systems.
B00879178
Adebomojo Moyosore
Adebomojo-m@ulster.ac.uk

1. Abstract

The condition of dementia is still prevalent in older individuals and is a significant contributor to reliance and incapacity. Most dementia cases are caused by Alzheimer's disease (AD), a complicated and severe neurodegenerative condition without reliable diagnostic tools. The most significant risk factor for AD is age, therefore, as lifespan increases, the prevalence of AD also increases, and the diagnosis becomes complicated. An estimated 55.2 million individuals worldwide have dementia. While persons in their 30s, 40s, and 50s may get this illness, adults 65 and older make up most of those diagnosed with AD. To stop AD from worsening and spreading irreversibly, it is critical to correctly and quickly identify it. Deep learning (DL) and machine learning (ML) have seen increased success in medical imaging in recent years, and it has replaced previous techniques for analyzing medical pictures and raised awareness of AD. Deep and machine learning models are more precise and efficient in detecting AD than traditional machine learning techniques. This research aims to identify the most accurate prediction model for labeling patients as healthy or unhealthy. K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM) were three models that were taken into consideration through the RStudio software package with the utilization of a secondary dataset obtained from the Australian Imaging, Biomarker & Lifestyle (AIBL). Using the R software, proper data pretreatment steps were conducted to clean the raw data, and standard processes were used to evaluate the cleaned data. According to the results, DL and ML technologies can successfully detect AD with the best model prediction in just eight (8) features out of the thirty-one (31) features in the baseline dataset were fitted by Random Forest (RF), which had a recall score of 98.88%.

Keywords: Alzheimer's Disease Neuroimaging Initiative (ADNI), AIBL(Australian imaging Biomarkers and Lifestyle Flagship Study of Ageing), HC(Healthy Control), Mild Cognitive Impairment(MCI), Deep learning, Machine Learning, neural network, support vector machine, Alzheimer's disease(AD), and dementia.

2. Introduction

Most people agree that the brain is one of the body's most essential organs. All actions and reactions that allow us to think and believe are controlled by and supported by the brain. It also aids in preserving good memories and emotions. With advancing years come a variety of diseases. One of them, dementia, affects 60 to 80 percent of older people [1]. *Age is not the only factor associated with a neurodegenerative condition; other risk factors include gender, a history of neurological disorders, and biomarkers like amyloid (APOE 4) and tau (APOE 3)* [2]. *AD is a form of dementia that is highly severe. AD is a progressive and fatal brain condition. When AD is identified, it slowly worsens and destroys memory cells, which impairs a person's ability to think. It is a degenerative neurological condition that causes neurons to stop functioning or even die* [3]. *There are various stages of AD or dementia, such as the preclinical stage, during which the patient feels no change in daily activities but an early stage of AD and during which the brain begins to shrink. This stage is controllable. The next stage is mild dementia or mild cognitive impairment, which involves minor changes in the brain but has no impact on the patient's daily activities. The next dimension of mild AD is when the brain begins to shrink along with the mild symptoms of AD, which slightly impair daily activities. The next stage is dementia with moderate AD, in which symptoms significantly impact most aspects of daily life. The final stage is dementia with severe AD, in which patients experience various issues, including memory loss, frequent confusion, an inability to learn new things, a change in personality, and difficulty speaking and expressing the appropriate emotions. These then are AD's various stages. The spectrum of AD is the time it takes for symptoms to appear; it takes moderate cognitive impairment around 20 years to progress to AD. Early identification is crucial since AD has already altered the brain's structure. This alteration is brought on by the accommodation of amygdala protein in brain cells and the growth of the actual ventricle. Recent research found compelling evidence that individuals who survive COVID-19 have a higher chance of later getting AD* [4]. *Various efforts have been made in clinical practice to create a reliable method for detecting early-stage AD, which would provide patients with dementias diagnosed with access to critical knowledge, resources, and support. One of these procedures is the measurement of structural brain atrophy using cutting-edge brain image capture technologies like Magnetic Resonance Imaging (MRI)* [5]. *Due to the difficulties in generalizing biomarker changes, this technology (Positron Emission Tomography (PET) and Functional Magnetic Resonance Imaging (fMRI)), among many others, failed to categorize and predict AD patients accurately. Machine learning is the ideal option for early identification and evaluating collections of records and dataset sets with millions of identifiers from thousands to hundreds of people* [6]. *Machine learning is one of the most important artificial intelligence (AI) ideas to emerge from academia. The machine learning approach has a baseline of its experience and is sometimes known as "training data" or "preliminary assessment data"* [7]. *It focuses on fostering the skills necessary for using and accessing the material in the program. It is normal to practice analyzing and understanding data using machine learning. Machine learning algorithms have been extensively used in the medical picture and data extraction with various applications, including detecting brain diseases. The range of available alternatives is expanded. Using machine learning to detect and categorize AD's many kinds and stages will be crucial. Random Forest (RF), Support Vector Machine (SVM), and k-Nearest Neighbors (KNN) are three supervised machine learning classifiers that this research will use to construct a model that best predicts whether a person is a healthy (i.e. HC=1) or a non-healthy(i.e., Non-HC=0). The study will focus on the AIBL(BL) dataset, and every MCI and AD classes will be combined into one class, non-healthy control. This will provide light on the early detection and treatment of AD in elderly patients since it will reduce brain power and, over time, reduce thinking speed.*

Machine learning

It is a method for gathering and processing data that automates the creation of analytical answers. It is predicated on robots examining data, looking for patterns, and making judgments with minimal human assistance [8]. *Two classes of machine learning—supervised and unsupervised—can be used to detect AD early before advancing to the critical stage.*

Data with labels are used in supervised learning. Supervised learning requires monitoring to train the model, which is akin to a student who learns things in the presence of a teacher. Supervised learning in medical imaging may be used to detect the condition and subsequently classify it based on the results. Then, the model must train using the picture's texture, shape, and size. After training, the picture is fed into the supervising model to recognize the image and forecast the outcome according to the algorithm.

Another kind of learning algorithm is unsupervised learning. The structure and patterns can be discovered in the data through unsupervised learning. There are no norms to follow in unsupervised learning. By itself, the computer finds patterns in the data.

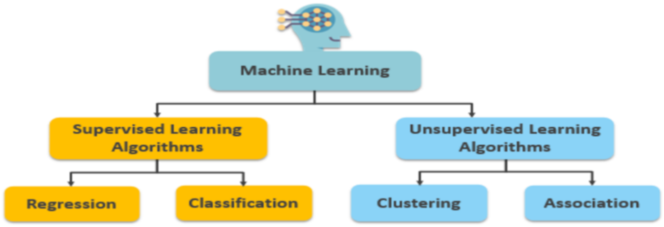

Machine Learning
Supervised Learning Algorithms
Unsupervised Learning Algorithms
Regression
Classification
Clustering
Association

Fig 2.1 Machine Learning (Rashmi Karan et al.,2022).

3. Dataset Description

The Australian Imaging, Biomarker & Lifestyle (AIBL) website is an excellent source for the dataset used for this study. On November 14th, 2006, the Flagship Study of Aging was officially established to undertake a four-and-a-half-year long-term cognitive study. The first cohort comprised 1,100 participants from two specific sites in Australia who were at least 60 years old. Each volunteer underwent evaluation, along with their cognitive traits, biomarkers, lifestyle, and health variables were noted to uncover potential precursors of the development of symptomatic AD in the future. The sample included people who were either in good health (HC), had mild cognitive impairment (MCI), or had been diagnosed with AD. The AIBL dataset was read and analyzed using RStudio software tools. R is a computer language mainly used for data mining tasks [9]. *The unique record Identifiers combined pertinent files after importing all necessary libraries.*

The 862 records (participants) and 36 attributes comprise the AIBL non-imaging baseline dataset utilized in this research. The characteristics include details on the subjects' demographics, medical histories, neuropsychological testing, apolipoprotein E genotype, blood analysis, clinical diagnoses, and five additional redundant factors connecting the several data sets. These redundant variables are subsequently removed from the dataset during the data preparation steps. Below is a quick summary of the characteristics that were utilized to analyze the AIBL dataset:

Demographics: Baseline patient gender and age (in years).

Medical Conditions: Neurologic (MH NEURL), smoking (MH SMOK), hepatic (MH HEPAT), musculoskeletal (MH MUSCL), psychiatric (MH PSYCH), malignancy (MH MALI), gastrointestinal (MH GAST), endocrine-metabolic (MH ENDO), renal-genitourinary (MH RENA), and cardiovascular (MH CARD),.

Genotypes: 2-Alleles genotypes of apoE (APGEN1/APGEN2). Among the three genotypes, one is held by each allele.

Blood tests: Mean corpuscular hemoglobin concentration (HMT102), Mean corpuscular hemoglobin (HMT100), AXT117 (Thyroid Stimulating Hormone), HMT3 (Red Blood Cell), Vitamin 12 (BAT126), HMT7 (White Blood Cell), Cholesterol (RCT120) (RCT329), Platelets (HMT13), Urea Nitrogen (RCT6), Serum Glucose (RCT11), Hemoglobin (HMT40).

Neuropsychological tests: MMSCORE (The Mini-Mental State Examination), CDGLOBAL(the Clinical Dementia Rating Global), LIMMTOTAL(the Logical Memory Immediate Recall), and the total number of story units recalled —the partial score of the LM test—are all neuropsychological tests (LDELTOTAL).

4. Methodology

The AIBL dataset was analyzed with a standard mining approach read through RStudio software with the data selection done with the study's objective, as shown in fig 4.2.1.

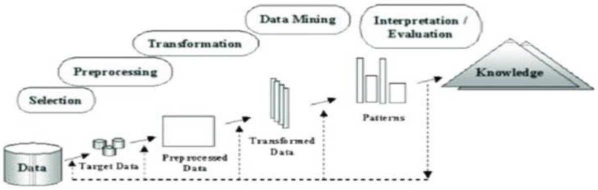
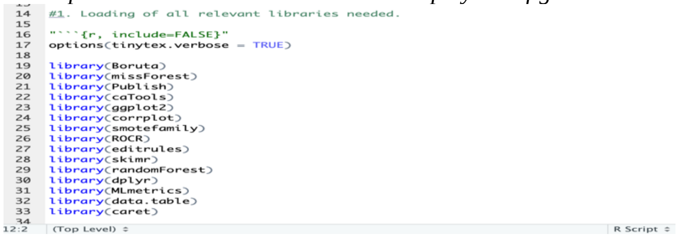


Fig 4.1. Standard Data Mining Steps (Mohotti et al., 2017) [10].

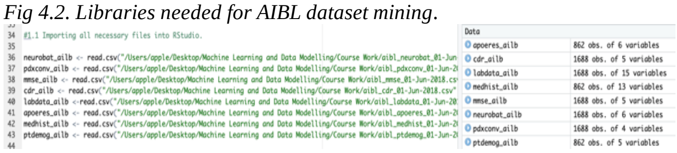
Data mining is a powerful method when examining massive data sets to find significant patterns, correlations, and links between various variables. It entails applying statistical and machine learning methods, including clustering, classification, regression, association rule mining, and anomaly detection, to extract meaningful information from data. It applies to several industries, including commerce, medicine, finance, marketing, and social media. Data mining in business may be used to discover consumer preferences, industry trends, and future development possibilities. It may be used in healthcare to examine medical information and find illness risk factors. Data mining in finance may assist in identifying fraud and evaluating investment risks. It may be used in marketing to study customer behavior and specific target demographics.

The AIBL dataset mining begins with loading all the necessary libraries, an option in R to enable verbose output when using the 'tiny text' package. 'tinytex' is a LaTeX distribution for R that allows users to compile LaTeX documents within R as displayed in fig 4.2.



```
#1. Loading of all relevant libraries needed.

"```{r, include=FALSE}"
options(tinytex.verbose = TRUE)

library(Boruta)
library(missForest)
library(Publish)
library(caTools)
library(ggplot2)
library(corrplot)
library(smotefamily)
library(ROCR)
library(editrules)
library(skimr)
library(randomForest)
library(dplyr)
library(MLmetrics)
library(data.table)
library(caret)
```

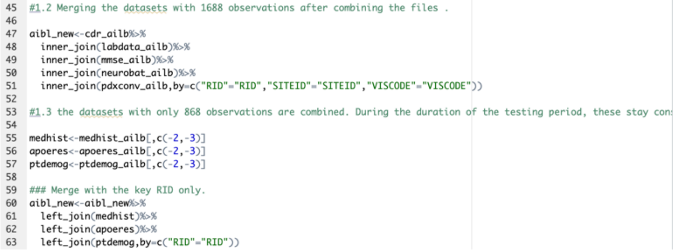
Fig 4.2. Libraries needed for AIBL dataset mining.



```
#1.1 Importing all necessary files into RStudio.

neurobat_ailb <- read.csv("/Users/apple/Desktop/Machine Learning and Data Modelling/Course Work/aibl_neurobat_01-Jun
pdxconv_ailb <- read.csv("/Users/apple/Desktop/Machine Learning and Data Modelling/Course Work/aibl_pdxconv_01-Jun-20
mmse_ailb <- read.csv("/Users/apple/Desktop/Machine Learning and Data Modelling/Course Work/aibl_mmse_01-Jun-2018.cs
cdr_ailb <- read.csv("/Users/apple/Desktop/Machine Learning and Data Modelling/Course Work/aibl_cdr_01-Jun-2018.csv"
labdata_ailb <-read.csv("/Users/apple/Desktop/Machine Learning and Data Modelling/Course Work/aibl_labdata_01-Jun-20
apoeres_ailb <- read.csv("/Users/apple/Desktop/Machine Learning and Data Modelling/Course Work/aibl_apoeres_01-Jun-20
medhist_ailb <- read.csv("/Users/apple/Desktop/Machine Learning and Data Modelling/Course Work/aibl_medhist_01-Jun-20
ptdemog_ailb <- read.csv("/Users/apple/Desktop/Machine Learning and Data Modelling/Course Work/aibl_ptdemog_01-Jun-20
```

Data	
apoeres_ailb	862 obs. of 6 variables
cdr_ailb	1688 obs. of 5 variables
labdata_ailb	1688 obs. of 15 variables
medhist_ailb	862 obs. of 13 variables
mmse_ailb	1688 obs. of 5 variables
neurobat_ailb	1688 obs. of 6 variables
pdxconv_ailb	1688 obs. of 4 variables
ptdemog_ailb	862 obs. of 5 variables

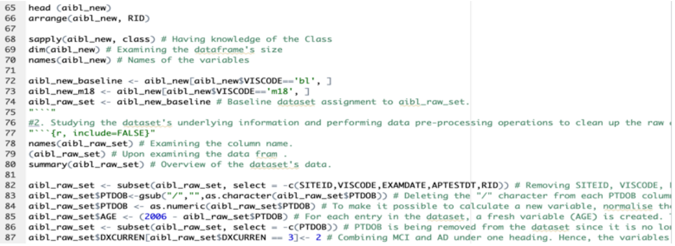
Fig 4.2.1. Importing of all raw datasets needed for AIBL dataset mining.



```
#1.2 Merging the datasets with 1688 observations after combining the files .

aibl_new<-cdr_ailb%>%
  inner_join(labdata_ailb)%>%
  inner_join(mmse_ailb)%>%
  inner_join(neurobat_ailb)%>%
  inner_join(pdxconv_ailb,by=c("RID"="RID","SITEID"="SITEID","VISCODE"="VISCODE"))

#1.3 the datasets with only 868 observations are combined. During the duration of the testing period, these stay con

medhist<-medhist_ailb[,c(-2,-3)]
apoeres<-apoeres_ailb[,c(-2,-3)]
ptdemog<-ptdemog_ailb[,c(-2,-3)]

### Merge with the key RID only.
aibl_new<-aibl_new%>%
  left_join(medhist)%>%
  left_join(apoeres)%>%
  left_join(ptdemog,by=c("RID"="RID"))
```

Fig 4.2.2. Merging of all Imported raw datasets needed for AIBL dataset mining.

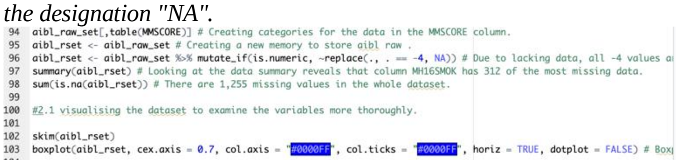
Any superfluous characteristics like visit code (VISCODE), RID (unique participant identification), VISCODE2 (visit code), APTESTDT, and day of evaluation (EXAMDATE) were subsequently removed from the data after the raw dataset was shown in RStudio with the variable names printed. The PTDOB column was cleaned up of symbols like "/," and the variables were forced to take on the proper data type. The AGE feature was developed (feature engineering) for analytical purposes, and the intended output variables, HC and Non-HC, were reduced from three-factor levels to two. Also, the variables in the MMSCORE columns were grouped and forced, and any missing values, denoted by "-4", were located and designated with "NA".



```
head (aibl_new)
arrange(aibl_new, RID)

sapply(aibl_new, class) # Having knowledge of the Class
dim(aibl_new) # Examining the dataframe's size
names(aibl_new) # Names of the variables

aibl_new_baseline <- aibl_new[aibl_new$VISCODE=='bl', ]
aibl_new_m18 <- aibl_new[aibl_new$VISCODE=='m18', ]
aibl_raw_set <- aibl_new_baseline # Baseline dataset assignment to aibl_raw_set.
"```"
#2. Studying the dataset's underlying information and performing data pre-processing operations to clean up the raw
"```{r, include=FALSE}"
names(aibl_raw_set) # Examining the column name.
(aibl_raw_set) # Upon examining the data fram .
summary(aibl_raw_set) # Overview of the dataset's data.

aibl_raw_set <- subset(aibl_raw_set, select = -c(SITEID,VISCODE,EXAMDATE,APTESTDT,RID)) # Removing SITEID, VISCODE,
aibl_raw_set$PTDOB<-gsub("/","",as.character(aibl_raw_set$PTDOB)) # Deleting the "/" character from each PTDOB colum
aibl_raw_set$PTDOB <- as.numeric(aibl_raw_set$PTDOB) # To make it possible to calculate a new variable, normalise th
aibl_raw_set$AGE <- (2006 - aibl_raw_set$PTDOB) # For each entry in the dataset, a fresh variable (AGE) is created.
aibl_raw_set <- subset(aibl_raw_set, select = -c(PTDOB)) # PTDOB is being removed from the dataset since it is no lo
aibl_raw_set$DXCURREN[aibl_raw_set$DXCURREN == 3]<- 2 # Combining MCI and AD under one heading. Hence, the variables
```

Fig 4.3.1. Combining MCI and AD under one heading. Hence, the variables are now 1 for Healthy Control (HC) and 2 for non-HC (Non-Healthy Control).

As shown below in fig 4.3.2. The process of identifying outliers included creating a boxplot of all the data. Inconsistencies in the dataset, such as rule breaches and the existence of unique characters, were also checked. Tukey's box-and-whisker technique for outlier identification was used to find the outliers, and they were all given the designation "NA".



```
aibl_raw_set[,table(MMSCORE)] # Creating categories for the data in the MMSCORE column.
aibl_rset <- aibl_raw_set # Creating a new memory to store aibl raw .
aibl_rset <- aibl_raw_set %>% mutate_if(is.numeric, ~replace(., . == -4, NA)) # Due to lacking data, all -4 values a
summary(aibl_rset) # Looking at the data summary reveals that column MH16SMOK has 312 of the most missing data.
sum(is.na(aibl_rset)) # There are 1,255 missing values in the whole dataset.

#2.1 visualising the dataset to examine the variables more thoroughly.

skim(aibl_rset)
boxplot(aibl_rset, cex.axis = 0.7, col.axis = "#0000FF", col.ticks = "#0000FF", horiz = TRUE, dotplot = FALSE) # Box
```

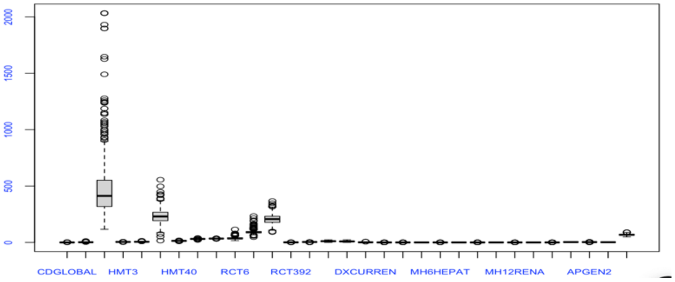
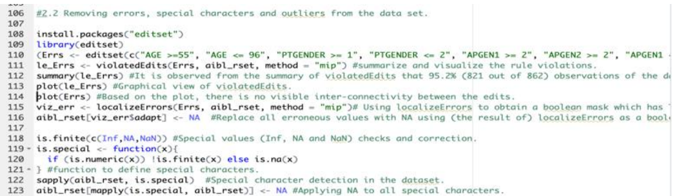


Fig 4.3.2. Boxplot of the characteristics for a visual representation of outliers.



```
#2.2 Removing errors, special characters and outliers from the data set.

install.packages("editset")
library(editset)
(Errs <- editset(c("AGE >=55", "AGE <= 96", "PTGENDER >= 1", "PTGENDER <= 2", "APGEN1 >= 2", "APGEN2 >= 2", "APGEN1
le_Errs <- violatedEdits(Errs, aibl_rset, method = "mip") #summarize and visualize the rule violations.
summary(le_Errs) #It is observed from the summary of violatedEdits that 95.2% (821 out of 862) observations of the d
plot(le_Errs) #Graphical view of violatedEdits.
plot(Errs) #Based on the plot, there is no visible inter-connectivity between the edits.
viz_err <- localizeErrors(Errs, aibl_rset, method = "mip")# Using localizeErrors to obtain a boolean mask which has
aibl_rset[viz_err$adapt] <- NA  #Replace all erroneous values with NA using (the result of) localizeErrors as a bool

is.finite(c(Inf,NA,NaN)) #Special values (Inf, NA and NaN) checks and correction.
is.special <- function(x){
  if (is.numeric(x)) !is.finite(x) else is.na(x)
} #function to define special characters.
sapply(aibl_rset, is.special)  #Special character detection in the dataset.
aibl_rset[mapply(is.special, aibl_rset)] <- NA #Applying NA to all special characters.
```

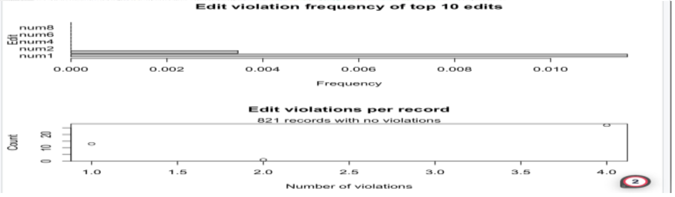
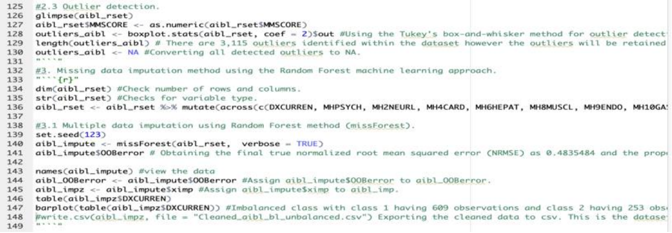


Fig 4.3.3. Removing errors, special characters and outliers from the dataset.

Random Forest method(missForest) was used to impute non-parametric missing data. This is because, unlike other approaches like hot-deck or k-NN, Random Forest is appropriate for datasets with various data types, such as the AIBL data. It also estimates the out-of-bag (OOB) imputation error and may run simultaneously, reducing calculation time, as shown below in fig 4.4 with an unbalanced bar plot. After four (4) iterations with a set seed value of 123, the model's actual normalized root mean squared error (NRMSE) was 0.4835484, and its percentage of erroneously categorized values (PFC) was 0.2660085.



```
#2.3 Outlier detection.
glimpse(aibl_rset)
aibl_rset$MMSCORE <- as.numeric(aibl_rset$MMSCORE)
outliers_aibl <- boxplot.stats(aibl_rset, coef = 2)$out #Using the Tukey's box-and-whisker method for outlier detect
length(outliers_aibl) # There are 3,115 outliers identified within the dataset however the outliers will be retained
outliers_aibl <- NA #Converting all detected outliers to NA.
"```"
#3. Missing data imputation method using the Random Forest machine learning approach.
"```{r}"
dim(aibl_rset) #Check number of rows and columns.
str(aibl_rset) #Checks for variable type.
aibl_rset <- aibl_rset %>% mutate(across(c(DXCURREN, MHPSYCH, MH2NEURL, MH4CARD, MH6HEPAT, MH8MUSCL, MH9ENDO, MH10GA

#3.1 Multiple data imputation using Random Forest method (missForest).
set.seed(123)
aibl_impute <- missForest(aibl_rset,  verbose = TRUE)
aibl_impute$OOBerror # Obtaining the final true normalized root mean squared error (NRMSE) as 0.4835484 and the prop

names(aibl_impute) #view the data
aibl_OOBerror <- aibl_impute$OOBerror #Assign aibl_impute$OOBerror to aibl_OOBerror.
aibl_impz <- aibl_impute$ximp #Assign aibl_impute$ximp to aibl_imp.
table(aibl_impz$DXCURREN)
barplot(table(aibl_impz$DXCURREN)) #Imbalanced class with class 1 having 609 observations and class 2 having 253 obs
#write.csv(aibl_impz, file = "Cleaned_aibl_bl_unbalanced.csv") Exporting the cleaned data to csv. This is the datase
"```"
```

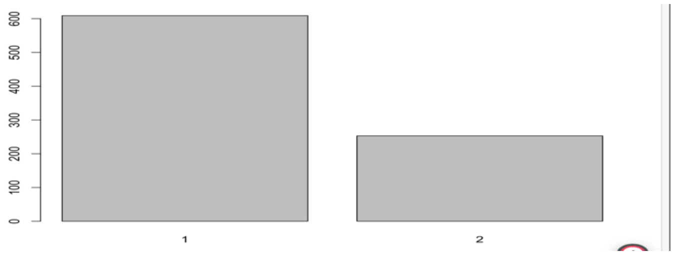
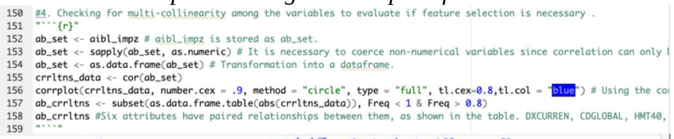


Fig 4.4. Unbalanced Class, with class one having 609 observations and class two having 253.

The necessity for feature selection was determined by doing a multi-collinearity check among the variables using Pearson's correlation coefficient, as shown below in fig 4.5. The six factors with strong multiple-collinearity were discovered as LDELTOTAL, CDGLOBAL, HMT40, DXCURREN, HMT3, and LIMMTOTAL after the interactions were plotted using R's "corrplot" function.



```
#4. Checking for multi-collinearity among the variables to evaluate if feature selection is necessary .
"```{r}"
ab_set <- aibl_impz # aibl_impz is stored as ab_set.
ab_set <- sapply(ab_set, as.numeric) # It is necessary to coerce non-numerical variables since correlation can only
ab_set <- as.data.frame(ab_set) # Transformation into a dataframe.
crrltns_data <- cor(ab_set)
corrplot(crrltns_data, number.cex = .9, method = "circle", type = "full", tl.cex=0.8,tl.col = "blue") # Using the co
ab_crrltns <- subset(as.data.frame.table(abs(crrltns_data)), Freq < 1 & Freq > 0.8)
ab_crrltns #Six attributes have paired relationships between them, as shown in the table. DXCURREN, CDGLOBAL, HMT40,
"```"
```

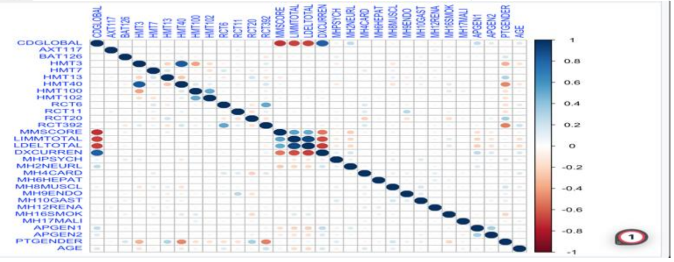
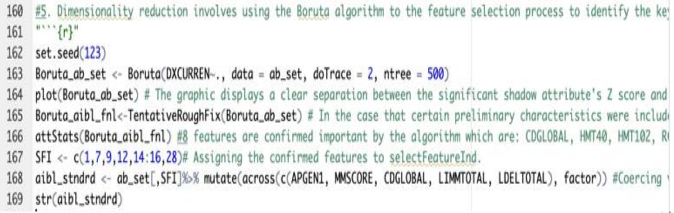


Fig 4.5. The AIBL Dataset's 31 features' correlation plot.

The Boruta method was used, as shown below in fig 4.6. to identify the key traits. The dimensionality reduction algorithm identified eight characteristics as being crucial; they are as follows: RCT20, CDGLOBAL, HMT102, LIMMTOTAL, HMT40, MMSCORE, APGEN1, and LDELTOTAL. In addition, the multi-collinearity check reveals that HMT3 has been eliminated have lesser significance than HMT40, which both exhibited a high degree of correlation. Moreover, it is noted that the last eight characteristics include both the linked features LIMMTOTAL and LDELTOTAL. Throughout the model training and testing, the backward elimination approach will be used to decide whether to eliminate either of the variables.



```
#5. Dimensionality reduction involves using the Boruta algorithm to the feature selection process to identify the ke
"```{r}"
set.seed(123)
Boruta_ab_set <- Boruta(DXCURREN~., data = ab_set, doTrace = 2, ntree = 500)
plot(Boruta_ab_set) # The graphic displays a clear separation between the significant shadow attribute's Z score and
Boruta_aibl_fnl<-TentativeRoughFix(Boruta_ab_set) # In the case that certain preliminary characteristics were includ
attStats(Boruta_aibl_fnl) #8 features are confirmed important by the algorithm which are: CDGLOBAL, HMT40, HMT102, R
SFI <- c(1,7,9,12,14:16,28)# Assigning the confirmed features to selectFeatureInd.
aibl_stndrd <- ab_set[,SFI]%>% mutate(across(c(APGEN1, MMSCORE, CDGLOBAL, LIMMTOTAL, LDELTOTAL), factor)) #Coercing
str(aibl_stndrd)
```

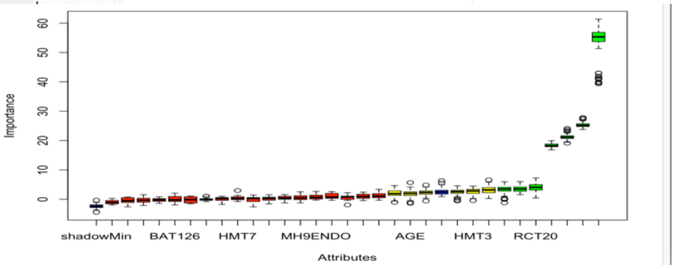
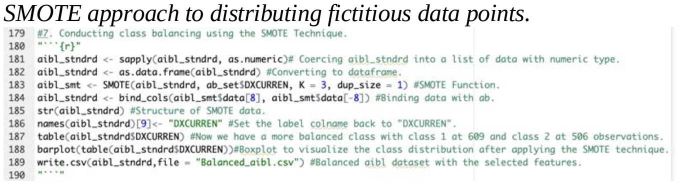


Fig 4.6. Boruta plot with the most significant shadow attribute's Z score delineated from the lesser-significant characteristics.

It was observed that the unbalanced dataset had been reduced to 862 observations of 8 variables, with class one healthy control having 609 observations and non-healthy control having 253 observations; it was seen that the target variable had an unbalanced class. The more *balanced class(1115 observations with nine(9) columns) was changed due to the class observations, with class one having 609 observations and class two having 506 observations using the setting k = 3 and Synthetic Minority Oversampling Method (SMOTE), as shown in figure 4.7. The k-Nearest Neighbor method is used by the SMOTE approach to distributing fictitious data points.*



```
#7. Conducting class balancing using the SMOTE Technique.
"```{r}"
aibl_stndrd <- sapply(aibl_stndrd, as.numeric)# Coercing aibl_stndrd into a list of data with numeric type.
aibl_stndrd <- as.data.frame(aibl_stndrd) #Converting to dataframe.
aibl_smt <- SMOTE(aibl_stndrd, ab_set$DXCURREN, K = 3, dup_size = 1) #SMOTE Function.
aibl_stndrd <- bind_cols(aibl_smt$data[8], aibl_smt$data[-8]) #Binding data with ab.
str(aibl_stndrd) #Structure of SMOTE data.
names(aibl_stndrd)[9]<- "DXCURREN" #Set the label colname back to "DXCURREN".
table(aibl_stndrd$DXCURREN) #Now we have a more balanced class with class 1 at 609 and class 2 at 506 observations.
barplot(table(aibl_stndrd$DXCURREN))#Boxplot to visualize the class distribution after applying the SMOTE technique.
write.csv(aibl_stndrd,file = "Balanced_aibl.csv") #Balanced aibl dataset with the selected features.
"```"
```

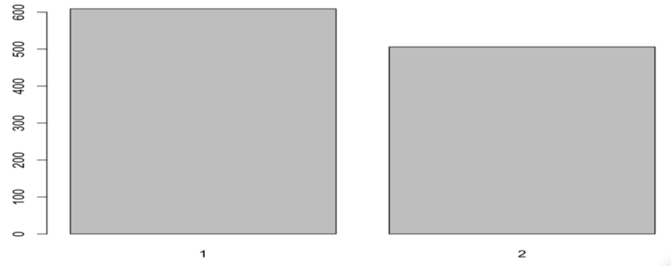
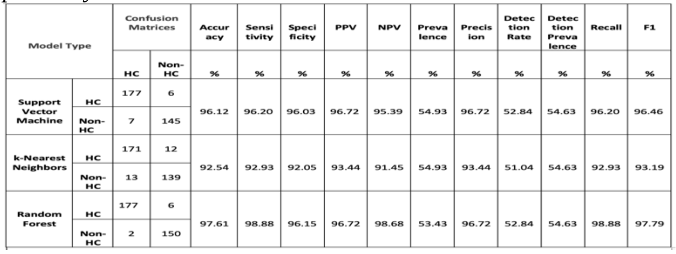


Fig 4.7. A plot of the target balanced class distribution per category.

Three machine learning techniques K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM), was used for model training. The cleaned dataset was divided between training and testing halves at a ratio of 70:30 each. The radial basis function (RBF-based) kernel SVM technique was chosen because of its excellent classification accuracy and capacity to categorize non-linearly separable issues. It effectively addresses "over-fitting" by generating a hyper-plane that separates a dataset into homogenous partitions on both sides. Similarly, the k-nearest neighbors (KNN) technique is a straightforward machine-learning approach that may be used to address classification and regression issues. While it is simple to apply, training takes longer when the data is enormous. Last but not least, the Random Forest (RF) method effectively correctly forecasts outcomes, mainly when each tree in the ensemble is uncorrelated from the others. For regression work, the average of each decision tree's outputs is calculated, while a classification task produces the mode of the categorical variable.

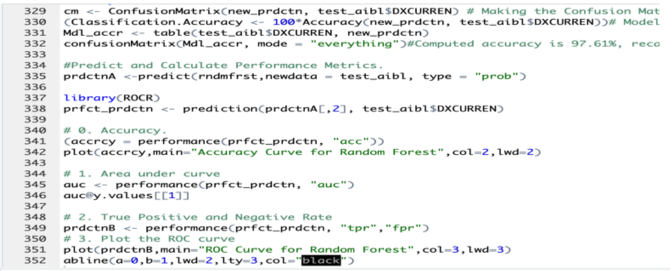
5. ***Results And Discussions***

The three trained algorithms' findings in the RStudio program are emphasized as illustrated in Fig. 5.1. The F-1 Score, sensitivity, and prediction accuracy suggested by the Random Forest model are all high. Recall/ Sensitivity will serve as a superior assessment measure due to the nature of the issue because our goal is to create a model that accurately identifies each patient and generates the fewest false negatives. The recall score for the model was 98.88% after 335 records were evaluated using the eight features we had chosen as the outcome of our dimensionality reduction method. Of those 335 records, 327 were identified correctly, while eight were incorrectly classified. Also, the model is a good classifier because of the Area Under Curve (AUC) score of 0.9936008 and the ROC curve's proximity to 1.0.



Model Type		Confusion Matrices		Accuracy	Sensitivity	Specificity	PPV	NPV	Prevalence	Precision	Detection Rate	Detection Prevalence	Recall	F1
		HC	Non-HC	%	%	%	%	%	%	%	%	%	%	%
Support Vector Machine	HC	177	6	96.12	96.20	96.03	96.72	95.39	54.93	96.72	52.84	54.63	96.20	96.46
	Non-HC	7	145											
k-Nearest Neighbors	HC	171	12	92.54	92.93	92.05	93.44	91.45	54.93	93.44	51.04	54.63	92.93	93.19
	Non-HC	13	139											
Random Forest	HC	177	6	97.61	98.88	96.15	96.72	98.68	53.43	96.72	52.84	54.63	98.88	97.79
	Non-HC	2	150											

Fig 5.1. The three model predictions analysis results.



```
cm <- ConfusionMatrix(new_prdctn, test_aibl$DXCURREN) # Making the Confusion Mat
(Classification.Accuracy <- 100*Accuracy(new_prdctn, test_aibl$DXCURREN))# Model
Mdl_accr <- table(test_aibl$DXCURREN, new_prdctn)
confusionMatrix(Mdl_accr, mode = "everything")#Computed accuracy is 97.61%, reca

#Predict and Calculate Performance Metrics.
prdctnA <-predict(rndmfrst,newdata = test_aibl, type = "prob")

library(ROCR)
prfct_prdctn <- prediction(prdctnA[,2], test_aibl$DXCURREN)

# 0. Accuracy.
(accrcy = performance(prfct_prdctn, "acc"))
plot(accrcy,main="Accuracy Curve for Random Forest",col=2,lwd=2)

# 1. Area under curve
auc <- performance(prfct_prdctn, "auc")
auc@y.values[[1]]

# 2. True Positive and Negative Rate
prdctnB <- performance(prfct_prdctn, "tpr","fpr")
# 3. Plot the ROC curve
plot(prdctnB,main="ROC Curve for Random Forest",col=3,lwd=3)
abline(a=0,b=1,lwd=2,lty=3,col="black")
```

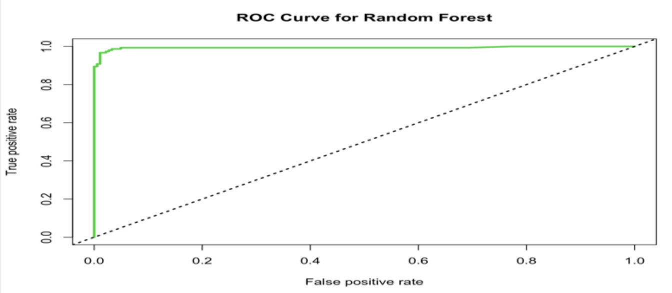


Fig 5.2. ROC curve displaying Random Forest model performance.

6. ***Conclusion***

The study's goal was accomplished, and the ML strategy used to analyze the AIBL dataset was effective. By fitting just eight features from the 31 characteristics offered in the baseline dataset, the chosen learned model could predict new data with an accuracy of 97.61% and a recall of 98.88%. Analytically, it was impossible to choose between dropping LIMMTOTAL or LDELTOTAL since the backward elimination method revealed that either or both variables are significant characteristics. Thus, it is suggested that future data collection should limit itself to gathering data on only seven biomarkers, with either LDELTOTAL or LIMMTOTAL included as selected by clinical research domain experts. Ensuring that only the relevant biomarkers important to the research are captured will lower the cost of data collecting. Future research utilizing the AIBL dataset should consider the machine learning(ML) strategy described in this study and include more data gathered over many months. Last but not least, this study experienced no computing constraints and may be repeated using the same approaches in the future.

References

[1] P. Balaji *et al*, "Hybridized Deep Learning Approach for Detecting Alzheimer's Disease," *Biomedicines*, vol. 11, *(1)*, pp. 149, 2023.

[2] X. Ding *et al*, "A hybrid computational approach for efficient Alzheimer's disease classification based on heterogeneous data," *Scientific Reports*, vol. 8, *(1)*, pp. 9774, 2018.

[3] S. V. Savita and M. Sabharwal, "Alzheimer's disease detection through machine learning," *Annals of the Romanian Society for Cell Biology*, pp. 2782-2792, 2021.

[4] M. T. Heneka *et al*, "Immediate and long-term consequences of COVID-19 infections for the development of neurological disease," *Alzheimer's Research & Therapy*, vol. 12, pp. 1-3, 2020.

[5] J. Cai *et al*, "An embedded feature selection and multi-class classification method for detection of the progression from mild cognitive impairment to alzheimer's disease," *Journal of Medical Imaging and Health Informatics*, vol. 10, *(2)*, pp. 370-379, 2020.

[6] Alzheimer's Association, "2016 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 12, *(4)*, pp. 459-509, 2016.

[7] A. F. Jorm and D. Jolley, "The incidence of dementia: a meta-analysis," *Neurology,* vol. 51, *(3),* pp. 728-733, 1998.

[8] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet],* vol. 9, pp. 381-386, 2020.

[9] H. A. Al-Odan and A. A. Al-Daraiseh, "Open source data mining tools," in *2015 International Conference on Electrical and Information Technologies (ICEIT),* 2015, .

[10] W. A. Mohotti and S. C. Premaratne, "Analysing Sri Lankan lifestyles with data mining: two case studies of education and health," *Kelaniya Journal of Management,* vol. 6, *(1),* 2017.

[11] A. Rubinski et al, "Polygenic effect on tau pathology progression in Alzheimer's disease," Ann. Neurol., 2022.

[12] M. Parsa, M. R. Alam and A. Mihailidis, "Towards AI-powered language assessment tools," 2021.

[13] J. JIAO and Y. LI, "Review of typical machine learning platforms for big data," Journal of Computer Applications, vol. 37, (11), pp. 3039, 2017.

[14] S. W. Moon et al, "Brain structure and allelic associations in Alzheimer's disease," CNS Neuroscience & Therapeutics, 2022.

[15] A. Simonetti et al, "Neuropsychiatric symptoms in elderly with dementia during COVID-19 pandemic: definition, treatment, and future directions," Frontiers in Psychiatry, vol. 11, pp. 579842, 2020.

[16] A. Rubinski et al, "Polygenic effect on tau pathology progression in Alzheimer's disease," Ann. Neurol., 2022.

[17] A. Varzandian et al, "Classification-biased apparent brain age for the prediction of Alzheimer's disease," Frontiers in Neuroscience, vol. 15, pp. 673120, 2021.

[18] J. Wen et al, "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation," Med. Image Anal., vol. 63, pp. 101694, 2020.