**CLASSIFICATION OF ALZHEIMER’S DISEASE USING MACHINE LEARNING APPROACH**

ANTHONY OMOWUMI

*MSc Data Science (B00892305), Ulster University* Magee Campus, Northern Ireland.

[anthony-o@ulster.ac.uk](mailto:anthony-o@ulster.ac.uk)

***Abstract - The most common cause of Dementia in the United Kingdom (UK) is Alzheimer’s Disease (AD). Alzheimer’s Disease is a progressive condition with symptoms developing over the years and it affects multiple brain functions. The aim of the study is to employ machine learning techniques in the classification of individuals as Healthy Control (HC) or Non-healthy Control (Non-HC) which include Mild Cognitive Impairment and AD combined as non-HC, as well as the identification of relevant and important predictors/features. The dataset used for the study was sourced from the Australian Imaging, Biomarkers, and Lifestyle Flagship Study of Ageing (AIBL). The R programming language was used for data preprocessing, mining, and model fitting. The data set was divided into 70% (780) and 30% (335). 3 supervised learning- Random Forest, Support Vector Machine (Untuned and Tuned), and K-Nearest Neighbor, and 1 unsupervised learning- K-Means for Clustering was done. The best-predicting model was Random Forest with an Accuracy of 97%, Recall of 96%, AUC of 98.99%, and ROC Curve closer to 1, fitting only 12 features out of the 31 features.***

***Keywords: Alzheimer’s Disease, Machine Learning, Data preprocessing, Data mining, and Model Fitting***

1. **INTRODUCTION**

The most common cause of Dementia in the United Kingdom (UK) is Alzheimer’s Disease (AD). The term *"dementia"* refers to a collection of symptoms brought on by a persistent deterioration in brain function that may impair one's memory, ability to think clearly, and other mental faculties while Alzheimer’s Disease is a progressive condition with symptoms developing over the years and it affects multiple brain functions.

The prevalence of Alzheimer's disease is highest among those over 65 years while the Age factor increases the chance of developing Alzheimer's disease, Alzheimer’s is not a normal part of aging [2]. The National Health Service, UK cites that Alzheimer’s disease affects an estimated 1 in 14 people over 65 and 1 in every 6 people over 80. However, around 1 in every 20 people with Alzheimer's disease is under the age of 65. This is called early- or young-onset Alzheimer's disease [1]. Also, it has been identified that women are most likely to be affected by dementia than men in their lifetime basically because of their longer life expectancy [4].

A group of Global researchers (Alzheimer’s Disease Neuroimaging Initiative) since the year 2004, has been in a longitudinal study on the investigation and development of treatments that can inhibit or stop the progression of AD and their overall goal is to validate biomarkers for the use in AD clinical treatment trials yet, prevention method and cures have not been discovered [3]. Apart from age being the most cause of AD, other risk factors include genetic and Vascular factors. The apolipoprotein E gene has three allelic variants (alleles ε2, ε3, and ε4), with allele ε4 being associated with an increased risk of Alzheimer's disease. APOE ε4 acts as a genetic risk modifier rather than being required or sufficient to develop AD. The well-known effect of age on AD is modified by APOE, as the age of onset is lower in APOE ε4 positives. Moreover, it has been hypothesised that APOE and vascular risk factors interact [5]. Making an early diagnosis of Alzheimer's disease and identifying early morphological abnormalities in the brain are crucial (AD). High-resolution magnetic resonance imaging (MRI) can be performed to aid in the disease's diagnosis and prognosis of the disease [7].

Recent research has shown a connection between COVID-19 instances, which include neurological symptoms, and biochemical indicators of Alzheimer's disease. The level of markers in the blood was suggestive of neurodegeneration is examined in a study by US researchers on 310 individuals admitted with COVID-19 utilising their blood samples. It was discovered that Neurofilament Light chain (Nfl) and a type of tau protein linked to Alzheimer's disease were strongly associated with the presence of neurological symptoms during COVID-19 infection. It is inferred that the coronavirus could accelerate the progression of dementia in some people which needs further study for its long terms effect [8]. Many novel approaches are proposed by researchers for early detection, towards early interventions, however, Machine Learning techniques have been found to be useful for the diagnosis of Alzheimer’s disease in the past decade [10].

ML algorithms attempt to create a model using statistical and predictive features that forecast desired results for the existing data. These models can find patterns and relationships in the provided data, then attempt to make predictions without or with limited human involvement [11]. Australian Imaging, Biomarkers and Lifestyle Flagship Study of Ageing made the data set used for this study available to the public (AIBL). Healthy Control, Mild Cognitive Impairment, and Alzheimer's disease are among the categories in a 4.5+ years Longitudinal study that involved 1000+ individuals and had a minimum age requirement of 60 [7]. The goal of this study is to employ machine learning techniques and algorithms in the classification of individuals as Healthy Controls (HC) or Non-healthy Controls (Non-HC), as well as the identification of relevant and important predictors/features. Mild Cognitive Impairment (MCI) and Alzheimer's disease (AD) are combined into one as Non-HC with respect to this study.

1. **RELATED WORK**

A recent study aimed at making use of Machine Learning algorithms to process data obtained by Alzheimer’s Disease Neuroimaging Initiative (ADNI) for the detection of Alzheimer’s disease in its primitive stage. MRI pictures were used and analysed to obtain numerical data to identify subjects with Alzheimer's disease and evaluate images of brain regions associated with the disease, which was then processed using machine learning algorithms Five Machine Learning algorithms were implemented which include Support Vector Machine (SVM), Gradient boosting Neural Network (NN), K-Nearest Neighbor (KNN), and Random Forest (RF). Neural Network and Random Forest have performance accuracy of 98.36 and 97.86, performed better than SVM 97.56, Gradient Boosting 97.26, and KNN 95.00. It was concluded that Alzheimer's disease can be identified at an early stage using these algorithms, and by receiving the proper therapy at these stages, the risk of Alzheimer's patients developing new complications is reduced [11].

1. **METHODS AND MATERIALS**

This study uses the Australian Imaging, Biomarkers, and Lifestyle Flagship Study of Ageing (dataset) that was launched on 14th November 2006, a four and half year prospective longitudinal study of cognition. The study cohort includes volunteer participants of about 1000+ with 60 years of age as the minimum across two centers in Australia. Each participant volunteer was examined to discover the biomarkers, cognitive characteristics, health, and lifestyle factors that would determine the subsequent development of symptomatic Alzheimer’s Disease (AD). The volunteers include patients with either Healthy Control, Mild Cognitive Impairment (MCI), or Alzheimer’s disease (AD) [9]

The R programming language was used to explore the dataset. R is a popular statistical and data mining tool that is used by researchers and data scientists alike. Necessary libraries were imported for the analysis of the dataset [12].

1. ***Dataset Description***

The AIBL non-imaging dataset consists of 862 (Participants) records and 32 Features. The features include categories like demographic, Medical History, ApoE genotypes, Neuropsychology assessments, blood analysis, and clinical diagnostic results of the participants. Below is a brief description of the AIBL data features.

*Demographic* includes the Age in years and gender (Female/Male) of the participants in the study.

*Medical History* includes the psychiatric (MH\_PSYCH), neurologic (MH\_NEURL), cardiovascular (MH\_CARD), hepatic (MH\_HEPAT), musculoskeletal (MH\_MUSCL), Endocrine-metabolic (MH\_ENDO), gastrointestinal (MH\_GAST), renal-genitourinary (MH\_RENA), smoking (MH\_SMOK), Malignancy (MH\_MALI) of the participants. Medical history is recorded as a binary feature Yes/No.

*ApoE Genotypes* include 2 alleles genotype in which each allele holds one of the three genotypes: ε2, ε3, and ε4.

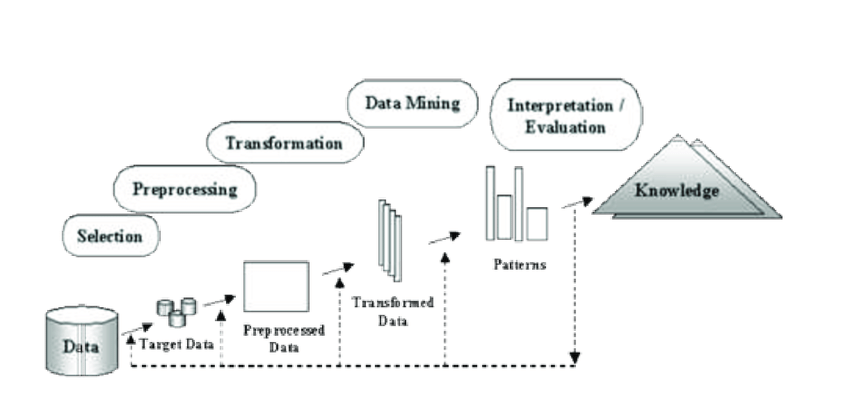
*Neuropsychological Assessments* include clinical dementia rating (CDR) with 5 categories, mini-metal state exam (MMSE) with 4 categories, logical memory immediate recall (LMIR), and logical memory delayed recall (LMDR).

*Blood Analysis:* includes thyroid stim. Hormone (AXT117), Vitamin B12 (BAT126), red blood cell (HMT3), white blood cell (HMT7), platelets (HMT13), haemoglobin (HMT40), mean corpuscular haemoglobin (HMT100), mean corpuscular haemoglobin concentration (HMT102), urea nitrogen (RCT6), serum glucose (RCT11), cholesterol (high performance) (RCT120), creatinine (rate blanked) (RCT329).

*Clinical Diagnosis:* includes the outcome of diagnostic results which is in 3 categories, Healthy Control (HC), Mild Cognitive Impairment (MCI), and Alzheimer’s Disease (AD). MCI and AD are combined into one (non-HC) in this study analysis to make it 2 categories.

1. ***Methodology***

In this study, an efficient data mining approach was used to analyse the AIBL dataset “Fig 1”. Appropriate data selection was done based on the study objectives and processed likewise.



*Fig 1: Basic Data Mining Approach*

1. ***Data Preprocessing and Transformation***

Because raw data is frequently not in a format that can be analysed, therefore data preprocessing and transformation is a crucial stage in data mining. In this study, a variety of techniques were used to improve the data's quality by looking for errors, outliers, and missing values. These methods are crucial since the accuracy of the results is related to the quality of the data. The following are steps taken to make sure that the data used was accurate, complete, and consistent, across all qualities.

***Data Selection & Cleaning:***Important libraries for the data analysis were imported into R studio. The data was selected andimported after which some descriptive statistical analysis was performed on the dataset, this shows the mean, median, min, max, and Interquartile range of the dataset. The min and max values show the data attributes are of different ranges of scale which confers a must to scale the data columns before Model fitting. The data attribute names were visualized, and the RID column was dropped because it is just a unique Identifier for each participant and has little or no contribution to the analysis. The AD value represented by 3 was changed into 2 to make MCI and AD one value equal to 2, known as the Non-Healthy Control (Non-HC). The label/class/outcome in the analysis is HC and Non-HC that is, 1 and 2. Data Attribute, MMSCORE was coerced as a factor to categorize its variables to their levels. All missing values, -4, 7 were replaced with NA respectively.

***Outlier & Error Detection****:* This was done using a boxplot (Tukey and Whisker plot) after special characters and errors and data inconsistency had been checked and replaced with NA in the dataset. The box plot shows BAT126 has a higher presence of outliers while others fall into the mild and relative showing no presence of outliers. “Fig 2”. Nothing was done to the column as the data attributes would be subjected to Feature Selection and Normalized respectively before Model fitting.

Chart

Description automatically generated

*Fig 2: Showing the Presence of Outliers using Boxplot in R*

***Missing Value Imputation****:* A non-parametric missing values imputation using Random Forest. This is chosen because it can be used for a mixed type of data (both numerical and categorical types of data), and it is robust to noise, outliers, and non-linearity of data which makes it more efficient than KNN or other methods. It yields an OUT-OF-BAG (OOB) Error along its computation. This error is used to measure the quality of the algorithm. Setting the seed to 12345 to allow for reproducibility of the same random samples when run over again, the OOB error in terms of NRMSE (Normalized Root Mean Square Error for continuous variables) and PFC (Proportion of Falsely Classified for categorical variables) for this imputation after 3 iterations, NMRSE was 0.4859566 and PFC was 0.2674520. With this algorithm, the NAs were filled.

***Correlation Co-efficient***: Histogram and Boxplot were used to visualise the relationship between Age and Diagnosis “Appendix A”. It was inferred that participants with higher Ages have a higher chance of being diagnosed as non-HC. Correlation Matrix was used to check for multicollinearity between the variables in determining a need for variable selection “Fig 3”

Chart, scatter chart

Description automatically generated

*Fig 3: Correlation plot of the features and collinearity*

The correlation plot shows 6 variables with a high positive correlation between CDGLOBAL, HMT40, HMT3, LIMMTOTAL, and LDELTOTAL with the Response variable (DIAGNOSIS). Further analysis was conducted to find the important features (Feature selection)

***Feature Selection:***  this was done using Boruta Algorithm. This was conducted to choose the important features. With the seed set to 123, the algorithm was run (500 iterations) on the data features, the features selected as important were 9 attributes, and 3 were tentatively unconfirmed. The final Boruta was run to confirm the tentative attribute “Fig 4”. The concluded important features were PTGENDER, MH2NEURL, APGEN1, HMT3, HMT40, HMT102, RCT20, RCT392, CDGLOBAL, MMSCORE, LIMMTOTAL, and LDELTOTAL. Prior to the feature selection, the HMT3 HMT40, LIMMTOTAL, and LDELTOTAL both have strong correlations with each other. However, the algorithm selected them as an important variable, hence the inclusion in the model fitting.

Chart

Description automatically generated

*Fig 4: Boruta Plot showing the important features in green, unimportant ones in red and shadow attributes in blue.*

***Data Distribution***: The outcome variable (Diagnosis) was checked for proportionality and found to be imbalanced with class 1(HC) having 609 and class 2 (non-HC) having 253. Using Synthetic Minority Over-Sampling Technique (SMOTE) and k set to 3, the data was balanced to 609 (HC) and 506 (non-HC) respectively “Appendix A”. SMOTE uses the nearest neighbor technique to allocate synthetic data points to balance the minority class. Also, the selected numeric columns were normalized on a similar scale. This improves the performance and stability of the training model *A picture containing square

Description automatically generated*

*Fig 5: Class (Diagnosis) balancing outcome using SMOTE.*

***Model Training & Testing:*** The data was divided into 70% (780) Train and 30% (335) Test. The model fitting was done using 3 supervised machine learning algorithms: Random Forest (RF), Support Vector Machine (SVM – Untuned and Tuned SVM), and K-Nearest Neighbor (KNN). RF is known for its high accuracy performance and does not overfit even many features. It is an ensemble learning method that fits well for both classification (data labels that are discrete) and regression (data labels that are continuous) hence the choice for the model because of its good efficiency [14]. However, Radial Basis Function Kernel in SVM, also known as the Gaussian RBF kernel, was chosen among other kernels because it is localized and has a finite response along the complete x-axis, its robustness to overfitting and handling of noisy data effectively. It is known for projecting non-linearly separable data into a higher dimensional space so it can be separable using a hyperplane and it can be classified using a basic idea of linear SVM. Also, KNN was chosen due to its quick calculation time although the prediction stage might be slow for large datasets, high accuracy although depending on the quality of data, and no assumptions about data which makes it crucial in non-linear data cases. Lastly, an unsupervised learning algorithm namely K-Means. Observations can be grouped into k groups using K-means clustering depending on how similar they are. This algorithm requires a set of unlabeled data because it’s an unsupervised type of learning. K-means clustering is a simple and efficient method. It is capable of handling very large data sets although sensitive to outliers.

1. **EXPERIMENTAL RESULTS AND DISCUSSION**

The test algorithms’ results are found below in “Fig 6”. The Random Forest algorithm has a higher and better prediction with an Accuracy of 97%, Sensitivity of 96%, and F1 Score of 97%. Apart from the accuracy of RF being better than other Models built, the Sensitivity/Recall defines the ability to accurately classify the disease as positive, and few false negative results misclassifications have a higher value in RF than other models which is the aim of the experiment. A total number of 335 observations was used for testing the algorithm with 12 features selected out of the 31 features. 325 was accurately classified while 10 was misclassified and shows Recall having a score of 96%. The ROC curve of the model was almost closer to 1 “Fig 8”. The AUC score is 98.99%. This indicates the RF model is an excellent classifier. Also, the K means algorithm was able to distinctively create the clusters after the optimal value for k was chosen using the Within Sum of Squares (WSS/Elbow Plot). The cluster means was uniquely classified with little or no overlapping of the clusters “Fig 7”.

*A picture containing table

Description automatically generated*

*Fig 6: Model Performance Evaluation chart*

Chart

Description automatically generated

*Fig 7: Cluster Plot after k optimal value has been obtained by wss/elbow method using k=2*

*Chart, line chart

Description automatically generated*

*Fig 8: The ROC Curve for Random Forest Model*

1. **CONCLUSION**

The ML techniques used in analysing the AIBL dataset were successful in their aim. 3 Supervised Algorithms with RF, SVM (Untuned and Tuned) and KNN, and 1 unsupervised algorithm, K-Means were used while the Random Forest was selected as the better-predicting model with an accuracy of 97%, Recall/Sensitivity of 96% by fitting only 12 features out of 31 features. It is therefore advisable to focus on getting 12 feature biomarkers for this analysis as considered by domain experts in clinical research. This would reduce the cost of data collection and computational expenses.

**ACKNOWLEDGEMENT**

The Australian Imaging, Biomarker & Lifestyle (AIBL) team provided the longitudinal data used in this research and acknowledges the team for making the dataset available. The machine learning techniques and methodologies used in this research were adapted as well from lessons Dr. Muskaan Singh, the module coordinator, taught in the classroom.

**REFERENCE**

|  |
| --- |
| [1] Alzheimer's Disease - National Health Service, 5 July 2021. [Online]. Available: https://www.nhs.uk/conditions/alzheimers-disease/. [Accessed 16 March 2023]. |
| [2] Alzheimer's Research UK, May 2022. [Online]. Available: https://www.alzheimersresearchuk.org/wpcontent/plugins/mof\_bl\_0.2.9/downloads/WAD-0522-0524\_WEB.pdf. [Accessed 16 March 2023]. | | |

[3] "ADNI Study Design," 2017. [Online]. Available: https://adni.loni.usc.edu/study-design/#background-container. [Accessed 17 March 2023].

[4] Prevalence by age in the UK - Dementia Statistics Hub, 5 August 2018. [Online]. Available: https://www.dementiastatistics.org/statistics/prevalence-by-age-in-the-uk/. [Accessed 17 March 2023].

[5] Scheltens. P., Van der Flier WM, "Epidemilogy and Risk Factors of Dementia," *Journal of Neurology, Neurosurgery & Psychiatry ,* Vols. V2 - V7, p. 76, 2005.

[6] Chen, J.H. et al "Risk factors for dementia," *Journal of the Formosan Medical Association,* vol. 108, no. 10, pp. pp. 754-764, 2009.

[7] Long. X., et al and Alzheimer’s Disease Neuroimaging Initiative *Prediction and classification of Alzheimer disease based on quantification of MRI deformation.,* vol. 12, no. 3, p. p.e0173372, 2017.

[8] COVID-19 Linked to Worsening Brain Health and Markers of Alzheimer's Disease by Alzheimer’s Research UK, 29 July 2021. [Online]. Available: https://www.alzheimersresearchuk.org/covid-linked-brain-health-alzheimers-disease/. [Accessed 20 March 2023].

[9] "Australian Imaging, Biomarkers & Lifestyle Flagship Study of Ageing (AIBL)," 2 March 2011. [Online]. Available: https://aibl.csiro.au/. [Accessed 20 March 2023].

[10] Tanveer, M., et al "Machine Learning Techniques for the Diagnosis of Alzheimer's Disease," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), ,* vol. 16, no. 1, pp. pp.1-35, 2020.

[11] P. Lodha, A. Talele and K. Degaonkar, "Diagnosis of Alzheimer's Disease Using Machine Learning," in *Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, India, 2018.

[12] Python vs. R: What’s the Difference? (2021). Available at: https://www.ibm.com/cloud/blog/python-vs-r (Accessed: 20 March 2023).

[13] Handling Missing Values with Random Forest (2022) Available at https://www.analyticsvidhya.com/blog/2022/05/handling-missing-values-with-random-forest (Accessed: 11 April 2023)

[14]Random Forest Algorithm in Machine Learning- An Overview (2022) Available at: https://www.mygreatlearning.com/blog/random-forest-algorithm (Accessed: 12 April 2023)

**APPENDIX A**

Chart, box and whisker chart

Description automatically generated

*Boxplot to visualize relationship between age and diagnosis.*

*Chart, histogram

Description automatically generated*

*Age Distribution using Histrogram*

*Chart, radar chart

Description automatically generated*

*K-means Cluster plot when k was set to 3*

Chart, histogram

Description automatically generated

*Using the Elbow Plot/WSS to get the optimal value for k*

Chart, line chart

Description automatically generated

*Optimal value for k chart (k=13)*