DTI and MRI based Alzheimer's disease prediction using machine learning algorithms

Submitted by

Amira Mahjabeen

181-35-334

Department of Software Engineering

Daffodil International University

Supervised by

Mr. Shariful Islam

Lecturer

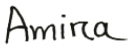
Department of Software Engineering

Daffodil International University

This Project report has been submitted in fulfillment of the requirements for the Degree of Bachelor of Science in Software Engineering.

DECLARATION

I hereby declare that this project was completed by me under the direction of Mr. Shariful Islam, Lecturer in the Department of Software Engineering at Daffodil International University. This also declares that neither this project nor any part of it has presented any degree awards anywhere else.



Amira Mahjabeen
ID: 181-35-334
Batch: 25
Department of Software Engineering
Faculty of Science & Information Technology
Daffodil International University

Certified by:

Mr. Shariful Islam
Lecturer
Department of Software Engineering
Faculty of Science & Information Technology
Daffodil International University

ACKNOWLEDGEMENT

First of all, we are grateful to the Almighty Allah for giving us the ability to complete the final thesis. We would like to express our gratitude to our supervisor Mr. Shariful Islam for the consistent help of my thesis and research work, through his understanding, inspiration, energy, and knowledge sharing. Her direction helped us to finding the solutions of research work and reach to our final theory. We would like to express my extreme sincere gratitude and appreciation to all of our teachers of Software Engineering department for their kind help, generous advice and support during the study. We are also express our gratitude to all of our friend's, senior, junior who, directly or indirectly, have lent their helping hand in this venture. Last but not the least, we would like to thank our family for giving birth to us at the first place and supporting me spiritually throughout my life.

ABSTRACT

Alzheimer's disease (AD) is a cognitive illness that commonly occurs in 65 years old or above people, which destroys the neurons and many parts of the brain. This disease has no cure, treatment can only slow the damage progression and finally, death occurs. So the early detection of this disease is essential. AD affects 40 million people worldwide. The number of AD patients is constantly increasing. So, it is necessary to identify the progression of AD. There are no single test has been developed to diagnose this disease. Different clinical methods-Mini-Mental State Examination (MMSE), Montreal Cognitive Assessment (MoCA), Alzheimer's disease Assessment Scale (ADAS), and neuroimaging techniques- positron emission tomography (PET), Magnetic Resonance Imaging (MRI), diffusion tensor imaging (DTI) is used to detect this disease. In this paper, Synthetic Minority Oversampling Technique (SMOTE) is used to oversample the dataset. And different machine learning algorithm such as support vector machine (SVM), K-Nearest Neighbor (K-NN), and Naïve Bayes (NB) to detect the stage of AD based on the different clinical data, cognitive data, brain data etc. 40 features were selected for training the model. Dataset obtained from Alzheimer's disease Neuroimaging Initiative (ADNI) database. The accuracy of the models are high. I obtain 85%, 93%, and 96% accuracy from k-nearest neighbor, support vector machine, and naïve bayes algorithm. Naïve Bayes provide highest accuracy.

LIST OF NOMENCLATURE

Terms	Nomenclature
AD	Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
ROI	Regions of Interest
LONI	Laboratory of Neuro Imaging
DTI	Diffusion Tensor Imaging
MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography
FDG	Fluorodeoxyglucose
APOE	Apolipoprotein E
MCI	Mild Cognitive Impairment
CN	Cognitive Normal
MMSE	Mini-Mental State Examination
MoCA	Montreal Cognitive Assessment
ADAS	Alzheimer's disease Assessment Scale
RAVLT	Rey Auditory Verbal Learning Test
WM	White Matter
SD	Standard Division
SVM	Support Vector Machine
K-NN	K-Nearest Neighbor
NB	Naïve Bayes
SMOTE	Synthetic Minority Oversampling Technique
DT	Decision Tree
CDRSB	Clinical Dementia Rating 'sum-of-boxes'
FA	Fractional Anisotropy
MD	Mean Diffusivity
AxD	Axial Diffusivity
RD	Radial Diffusivtiy

Table of figure

Table of Contents

CHAPTER I: INTRODUCTION

Alzheimer's disease (AD) is a type of dementia. In 2018, worldwide 50 million people were suffering from dementia which will increase to 152 million by 2050. In 2018, according to the World Alzheimer's Disease report dementia patient will appear every 3 seconds. AD accounts for 60%-80% among all dementia cases. Now the number of dementia patient are increasing. From 2010-2050, dementia sufferers are particularly common in low- and medium countries. AD is a brain shock that cannot be reversed. AD parents had a significant risk of depression. [1].

1.1 Background

AD is a neurodegenerative disorder that destroys the neurons in the brain and is responsible for memory loss and decreases thinking and learning skills. In the early stage of AD patients face mild cognitive difficulties, memory loss, and many difficulties in daily activities. Some patients face difficulties in language and exclusive functions such as thinking, planning, monitoring, control, working memory, time management, and organization. The patient finally dies after about 3 to 10 years. On average, an Alzheimer's patient survives 5.5 years [2]. Genetic mutation and Down syndrome is an uncommon genetic factor which is associate with Alzheimer's disease. Age, APOE, and family history are the risk factor for AD. About 3% people at age 65 to 74, 17% people at age 75 to 84, and 32% people at age 85 to above have AD. There are three forms of the APOE gene- e2, e3, e4. E4 gene is responsible for inheriting the disease. Genetic changes cause the abnormal beta-amyloid protein deposited around the brain cell and abnormal tau protein deposited inside the brain cell, which is the main reason for Alzheimer's disease [3]. Demographic information, living information, and

disease history are gathered by the questionnaire method. Mini-Mental State Examination (MMSE), Montreal Cognitive Assessment (MoCA), Alzheimer's disease Assessment Scale (ADAS), Rey Auditory Verbal Learning Test (RAVLT), positron emission tomography (PET), Magnetic Resonance Imaging (MRI), diffusion tensor imaging (DTI) are the clinical diagnosis to predict this disease. RAVLT, ADAS, MMSE and MoCA were used to approach cognitive performance and determine the progression of AD. PET, MRI, and DTI are neuroimaging techniques and successfully identify the progression of AD. PET analyzes glucose metabolism and amyloid deposition patterns to distinguish between AD and healthy people. MRI uses magnetic fields and radio waves to create a 3D representation of the internal brain structure. Structural MRI detects structural changes associated with AD. Functional MRI (rs-fMRI) measures the correlation between the blood oxygen level-dependent signal (BOLD) and neural activity to determine the functional connectivity across different regions of the brain. DTI measures the structural changes in white matter (WM). It can detect the water molecule diffusion in the brain and identify the abnormal spread [4-7].

1.2 Motivation

Alzheimer's disease symptoms occur gradually over a long duration. It effects in memory and cognitive skill and destroy many brain part. AD is the world's 4th leading risk factor for mortality and most costly disease. MCI is the early stage of AD. But after few years, MCI may transform into AD. Early treatment can help individuals delay the progression of MCI and improve the standard of living. But it can only slow the progression of the disease because there is no cure for this disease. This disease can take the life of a patent. About 50 million people suffer from AD worldwide in 2018. [8, 9].

1.3 Objective

Early detection can save patient from unexpected death. In this study, we aim to identify cognitive normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD) using different machine learning models. It is highly essential to early detect Alzheimer's disease to start treatment early. Our model will detect the three stages quickly and accurately.

1.4 Problem Statement

With the global population growing older, and the increasing number of dementia patients, Alzheimer's disease (AD) has emerged as a major social issue. The proportion of AD patients around the world would rise to 75.62 million by 2030, estimated by the International Federation of Alzheimer's Diseases. People with MCI have no issue but it can develop AD. Treatment can slow the progression. But it is not lifelong. It destroys different sections of the brain- the cerebral cortex, entorhinal cortex, hippocampus, parietal lobe, frontal lobe, occipital lobe, and temporal lobe in the brain.

1.5 Thesis Organization

The paper is divided into several chapters. The rest of this paper is arranged as follows: Section II, summary of previous works. Section III, description of the proposed method and algorithm used in this study. Section IV, evaluated and compared model result. Section V, highlight the future work and conclusion.

CHAPTER II: LITERATURE REVIEW

Supervised learning models- multivariate linear regression, logistic regression, and SVM used to differentiate the pathology of AD from aging-related cognitive impairment. In this study, the dataset was collected from ADNI. The dataset contains images, genetic assessment, medical records, and subject attributes data. Multivariate linear regression used for train the model and it provide continuous values. LR model used for increased efficiency. The kernel is selected and applied in the SVM classification algorithm depending on the number of attributes and training dataset [9]. An upgraded machine learning method Modified Random Forest (m-RF), has been used to identify AD progression. The accuracy of the model is 96.43%. m-RF is the upgraded version of the RF algorithm. This algorithm work for both regression and classification problem. Dataset is collected from the open-access database of OASIS-2. The dataset contains MRI images. Random state value 2 used for highest accuracy [10]. Another study, SVM and DT has been used to identify AD based on various parameters. But MMSE score, age, and education parameters are mainly used to identify AD. With an accuracy of 85%, SVM predicts the outcome. 83% accuracy for DT [11]. An algorithm was developed to identify MCI, and AD and compares them from healthy subjects based on a machine learning technique and dual-tasking gait assessment. Participants' single-task and dual-task measured and recorded by a computerized walkway and every participant has MoCA score. ProtoKinetics Movement Analysis Software (PKMAS) and GAITRite software are used to extract gait features from Zenomat and GAITRite systems and identify significant gait features for three classes – MCI vs. AD, healthy vs. MCI, and healthy vs. AD. SVM classifier used in those classes for training based on the selected gait features. The SVM classifier was also used to identify AD, MCI, and healthy for comparison based on the MoCA score. The accuracy was 78% for the

selected gait feature and 83% for the MoCA score [12]. Histogram, and random forest classifier are used for diagnosis of Alzheimer's disease. 8 different machine learning technique- naïve bayes classifier, logistic regression, support vector machine, artificial neural network, k-nearest neighbors, decision tree, random forest, and rotation forest are used to evaluate based on MRI images. 10 fold-cross validation technique is used to demonstrate the highest performance of the 8 machine learning technique. Performance are measured by accuracy, sensitivity, specificity, ROC area, and F-measure. To reduce the dimensionality of the data, a normalized histogram was obtained from each image [13]. To predict and classify various machine learning methods are used for identifying CN, early MCI, late MCI, and AD using the ADNI dataset. ADNI dataset provides MRI images, cognitive scale scores, clinical scale scores, and Clinical Dementia Rating (CDR) scale scores of CN persons, early and late MCI patients, and AD patients. FMRIB's Automated Segmentation Tool (FSL-FAST4) is used to segment 3D MRI images into the different tissues and regional cortical thickness of the left hemisphere and right hemisphere measured by FreeSurfer. Non-linear SVM classifier using the RBF kernel provided the best accuracy for 10 fold cross-validation [14]. Structural brain volume and cortical thickness have been measured from MRI scans by MRI analysis software (FreeSurfer) and used as features in AD identification. Swarm intelligence feature selection technique used for selecting the best features and building a multistage classifier based on different ML models KNN and SVM for detecting AD [15]. SVM, KNN, RF, gradient boosting, neural network models were used to detect AD based on cognitive and medical parameters. MRI scans are analyzed to produce numeric data and used in machine learning techniques. RF, and neural network [16].

CHAPTER III: METHODOLOGY

3.1 Dataset

The dataset which is used in this research was gathered from Alzheimer's neuroimaging Initiative (ADNI) (adni.loni.usc.edu). It establishes a standard process for identifying biomarkers. It provides any section of brain data. The dataset contains- Biomarker of Cerebral Spinal Fluid, Genetic biomarkers, Cognitive test results, and neuroimaging techniques- MRI, PET, and DIT. DTI data were obtained from the ADNI-2 and ADNI-3 projects (www.adni.org) and Laboratory of NeuroImaging, UCLA (Nir et al., 2013; link at adni.org). Datasets are provided to measure mild cognitive impairment (MCI), Alzheimer's disease (AD), and cognitive normal (CN). The dataset contains 177 person data and 266 features. The label feature is AD123. There 33 Cognitive normal person (1, CN, n=33), 36 Alzheimer disease patient (2, AD, n=36), 108 mild cognitive impairment patient (3, MCI, n=108). In figure 1, the count plot displays the number of samples of every class in the label.

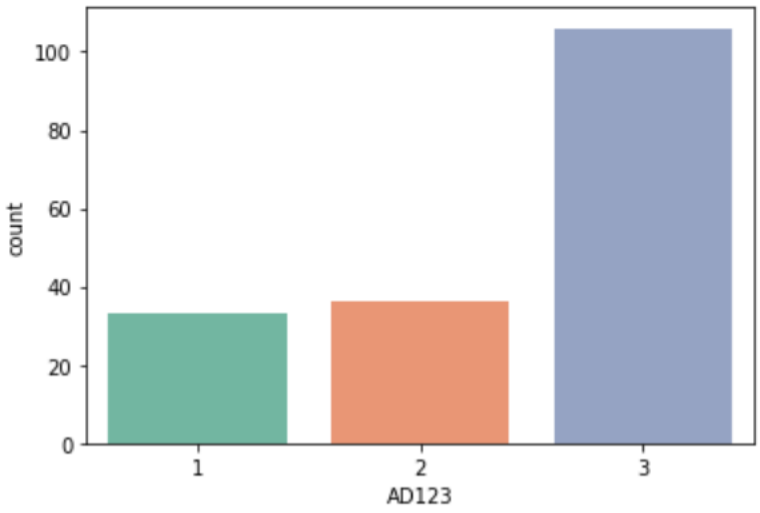


Fig1: Label Data (Cognitive normal, Alzheimer disease, mild cognitive impairment)

Table1: Individuals characteristics of the CN, AD, and MCI Groups

Characteristic	All (n = 177)	CN (n = 33)	AD (n = 36)	MCI (n = 108)
Gender, Male Female	 107 70	 19 14	 22 14	 66 42
Age, year mean±SD	73.34±7.01	73.54±5.60	74.35±8.27	72.93±6.96
Education mean±SD	16.08±2.76	16.75±2.90	15.30±2.60	16.14±2.73

The dataset contains data on people between the ages of 55 and 90. Table 1 shows the characteristics of individuals. MMSE score, MoCA score, CDRSB score, LDELTOT score, ADAS score, RAVLT score, FAQ, etc. are the Cognitive tests score data. MRI data, PET data, DTI data. Fractional anisotropy (FA), mean diffusivity (MD), radial diffusivity (RD), and axial diffusivity (AxD) are the four types of measurement of DTI. The level of directional of the diffusing process is analyzed by FA. The degree of diffusion rate is evaluated by MD. The degree of diffusion parallel to fiber tracts is AxD. The dataset also contains personal information. The water diffusion rate in the axis parallel to the axon fibers is RD. Those are the important measurement for detecting the three stages of the disease.

3.2 Method

In this study, different machine learning model has been used to identify CN, AD, and MCI. Various ML models are used to increase the accuracy of multiclass classification. The following steps are followed in the planned methodology's processing. (i) Data preprocessing, (ii) Feature selection, (iii) Built training set, (iv) Training algorithm. Figure 2 show the sequence of the method.

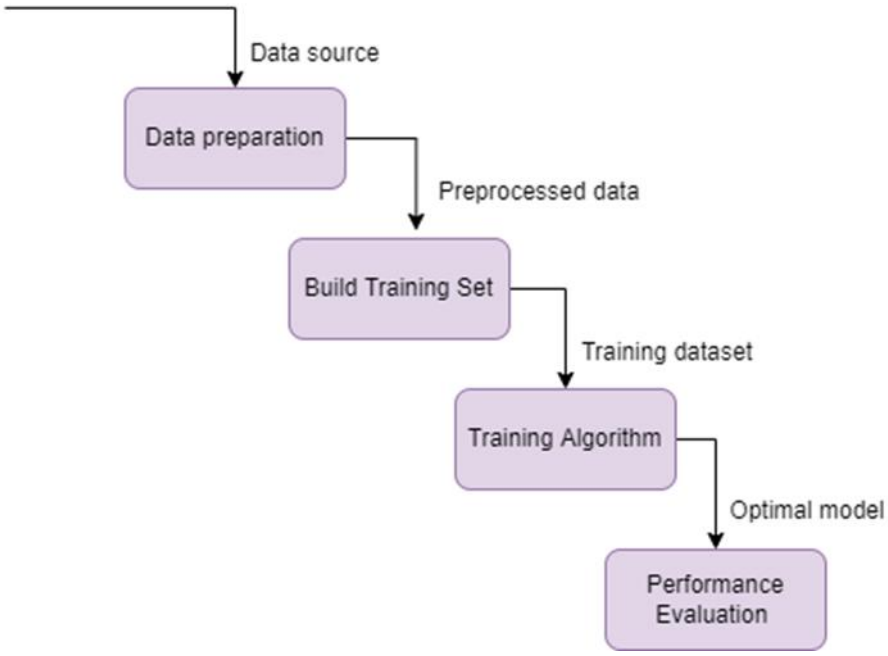


Fig2: Machine Learning Approach

3.2.1 Data preprocessing

The accuracy of the machine learning algorithm is affected by data preprocessing. A dataset may have missing data, duplicate data, and noisy data. Data preprocessing can transfer datasets into a useful format. There is various method to handle missing data. We can remove missing data. Rather than remove the missing value, we can be filling it with mean or median value or random value. There will be no problem if you remove the missing value in the big dataset. In this study, the dataset contains a small number of missing values. So mean or median can be used to fill the missing values and not affect the accuracy. The missing value is represented in 999999. I convert the missing value into NUN value. Figure 3 shows the null values from the dataset. The dataset has a low number of null values.

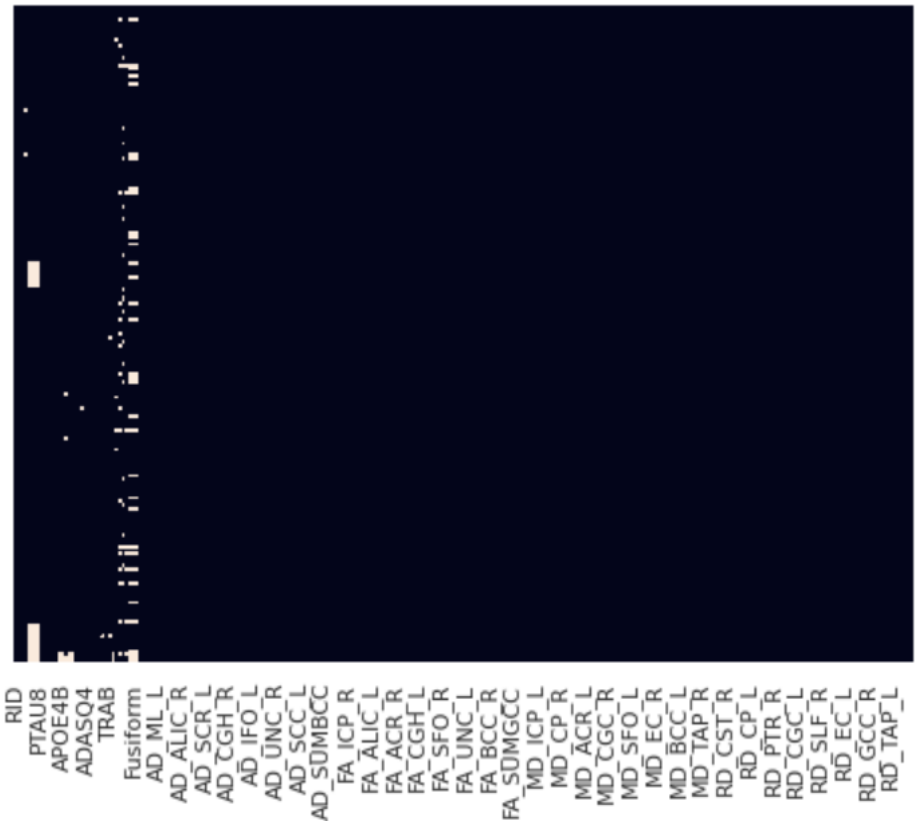


Fig3: Null values percentage

VISCODE, and ADMCI3 are categorical features. Label encoder is used for converting categorical features into numerical features. With the help of the label encoder technique, I transform all the categorical data into numeric form. And a dataset can contain different ranges of data. But the dataset needs to be on the same scale. Various feature scaling techniques are used to bound values in two numbers. Feature scaling speeds up the training process. MinMaxScaler is a feature scaling technique. The data used in this study are not scaled. It can reduce the accuracy of the model. After applying the MinMaxScaler technique, the features are scaled for a value range generally between 0<=x<=1.

ABETA12	TAU80	PTAU8	ADMCI3	APOE4	APOE4B	FDG	AV45	AV45AB12	.
0.000000	0.196498	0.176071	0.25	0.000000	0.000000	0.756851	0.034229	0.000000	.
0.000000	0.284319	0.279302	0.25	0.000000	0.000000	0.676969	0.145030	0.000000	.
0.000000	0.283271	0.225955	0.25	0.000000	0.000000	0.431478	0.082794	0.000000	.
0.000000	0.128469	0.083098	0.25	0.000000	0.000000	0.665559	0.155546	0.000000	.
0.000000	0.378564	0.300462	0.25	0.000000	0.000000	0.519412	0.176360	0.000000	.
...	...	...	...	...	...	...	...	...	.
1.000000	0.200470	0.190064	0.00	0.500000	1.000000	0.074266	0.724283	1.000000	.
1.000000	0.187859	0.207907	0.00	0.440858	0.881715	0.469546	0.729268	1.000000	.
1.000000	0.801530	0.777565	0.00	0.047732	0.095463	0.257300	0.634755	1.000000	.
1.000000	0.454229	0.483250	0.00	0.775629	0.775629	0.436292	0.478991	0.775629	.
0.356754	0.259925	0.243107	0.00	0.178377	0.356754	0.624122	0.401002	0.356754	.

Fig4: Scaled data

It is a multiclass classification problem and the dataset is imbalanced. Various techniques are used to handle an imbalanced dataset. We can use replacements to oversample the class distribution which is oversampling. Another is undersampling which removes rows from the given dataset at random. The dataset contains limited data. Rather than remove rows, the replacement technique will be better because the dataset is small. Figure 5 shows the label representation from the dataset. There are far fewer CN people and AD affected people than MCI affected people.

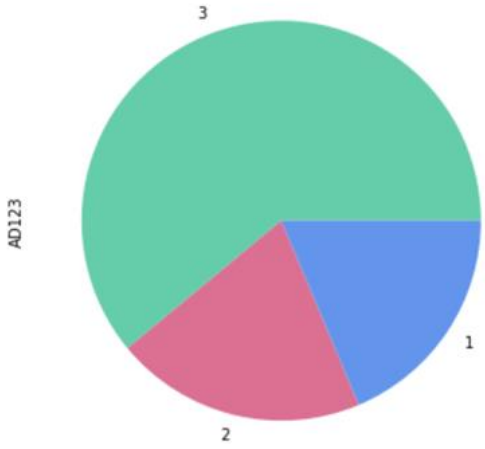


Fig5: Data representation

Synthetic Minority Oversampling Technique (SMOTE) sampling technique has been used to create data samples from the dataset and re-sample the features to match with the sample. I used SMOTE oversampling technique to increase the data for each label-CN, AD, and MCI**.** This technique is built based on the K-NN algorithm. Figure 6 show the data representation after applying SMOTE technique.

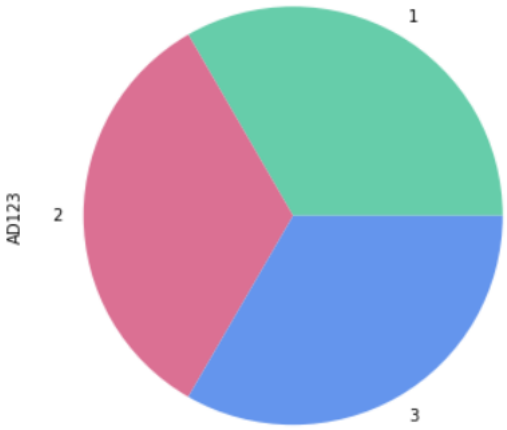


Fig6: Data representation after applying SMOTE technique

3.2.2 Feature selection

Feature selection is frequently used to reduce the dimension. It removes useless information. By removing the useless features, it helps to train the model faster. This technique also enhances accuracy by reducing the overfitting of the dataset. Important features can be selected by knowledge or different feature selection technique. ADNI dataset contains a large number of features. Therefore, it is impossible to select the necessary features based on the knowledge. So, different feature selection techniques can be utilized to identify the important features. In this study, the SelectKBest technique was used to figure out the important features. After using this technique, 30 significant features were selected for training. Table 2 shows the details of the selected features.

Table2: 30 selected feature

Feature	Definition
ADMCI3	In ADMCI3 there are 5 unique data. They are CN, EMCI, LMCI, SMC, and AD. They are the stages of Alzheimer's disease.
CDRSB	CDRSB is the Clinical Dementia Rating score. The score range is 0-5.
FAQ	Frequently Asked Question.
LDELTOT	LDELTOT means Delayed Recall Total. It is a cognitive test.
MMSE	MMSE means Mini-Mental State Exam. It is the most common and used cognitive test.
ADAS11	ADAS11 means Alzheimer's Disease Assessment Scale Cog-11. The range of the score is (0-70).
TRAB	
AV45AB12	It is preclinical AD. It is a stage of AD. It is identified by AV45.
ADAS13	ADAS13 means ADAS-Cog-13. The score range is 0-85.
ADASQ4	ADASQ4 is a cognitive test.
RAVLT_immediate	RAVLT is an effective cognitive test. RAVLT_immidiate is a stage of RAVLT stage. It is the total of the scores from the first five trials
ABETA12	ABETA12 means Amyloid beta12. It is a protein. Genetic mutilation increase this protein. Abnormal increase of this protein increase the risk of AD.
APOE4B	APOE4 gene makes apolipoprotein E protein. This gene is risk factor of AD. APOE4B is the type of APOE4 gene.
AV45	AV45 or Florbetapir is technique for detecting AD.
ABETA1700	Amyloid beta 1700.
APOE4	APOE4 is a gene. It has two copy. Those people who carries the two copy of this gene have strong risk of AD.
MD_FX_ST_L	It is DTI data and measured by MD. Fornix left is a WM tube. It connects the node. It helps for memory recall.
RD_FX_ST_L	WM fornix left value is measured by RD.

AD_FX_ST_L	It is also WM fornix left value but measured by AxD.
RD_CGH_L	Left Cingulum (hippocampus) measured by RD. Without it we cannot leave. It help us in learning. It also control emotion.
Entorhinal	It is a part of cerebral cortex. It is measured by MRI technique. Entorhinal cortex is effected by AD
Fusiform	It is found in the surface of lobe. It helps human to recognize face.
MidTemp	MidTemp is the middle temporal artery. It helps in memory function.
MD_SS_R	Right sagittal stratum is WM bundle. It is analysis by MD.
AD_SS_L	Left sagittal stratum is measured by AxD.
AD_UNC_L	Left uncinate fasciculus help us to understand language. It is measured by AxD.
PTAU8	It is tau protein. It is a risk factor of AD. It speed in the brain because of genetic mutation.
AD_CGH_L	Left Cingulum (hippocampus) is measure by AxD.
MD_UNC_L	Left uncinate fasciculus helps in language. It is measure by MD.
RD_UNC_L	Left uncinate fasciculus is measure by RD.

3.2.3 Training & testing set

The dataset were divided into two part. 80% data the training the model. And 20% data for test the model. The dataset used to build and optimize machine learning algorithms is referred to as "training" a machine learning model. To get a result, an algorithm is needed to train the model. The model provide output based on the input sample. The testing set has been used to model validation built from the execution of the training set, which can be displayed by a confusion matrix.

3.2.4 Training Model

Various machine learning model used for training the model based on the significant feature. Machine learning model works in the same way that the human brain performs. K-NN, SVM, and NB were used to train the model. Those algorithm work in different way. K-NN provide low accuracy compared with the other models.

Support Vector Machine

The SVM approach is used to find the hyper-plane of an N-dimensional structure that arranges datasets. Classification and regression problems can be solved by SVM. SVM makes a hyper-plane between two types of data. The hyper-plane divided the data point. For new data, it allocates the data into a class based on training. SVM kernel is a technique it converts data lower dimension to a higher dimension. For non-linear data, the polynomial kernel is used because a linear line cannot separate non-linear data. Then it is separated by a non-linear line. The radial base function is also a kind of non-linear function. It converts data lower to a higher dimension. For linear problems, SVM creates so many hyper-plane.

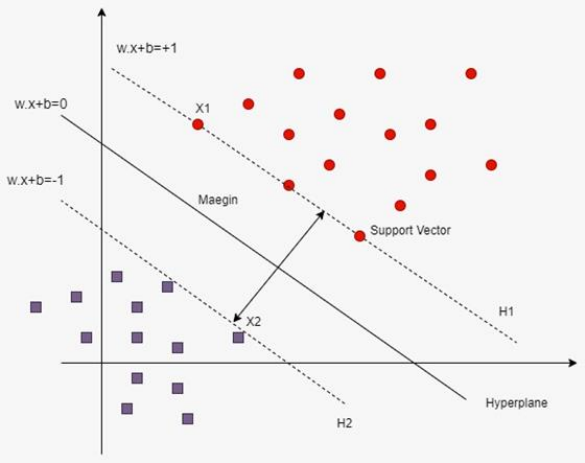


Fig7: Support Vector Machine

Naive Bayes

One of the most widely used classifiers is Naïve Bayes (NB). It's built on the Bayes theorem which is used to solve classification problems. This algorithm provide higher accuracy and faster training in big dataset. According to the Naïve Bayes classifier, the presence of each feature in a class is independent to the presence of any other feature. For training, naive Bayes requires significantly less data. NB calculate the probability of labels. And find probability for individuals feature for every class. Then calculate posterior probability. After that find the higher probability class. This is the equation of posterior probability –

$$P(y_i|X_1,X_2,...X_n)=\frac{P(X_1|y_i)\,P(X_2|y_i).....P(X_n|y_i)}{P(X_1)\,P(X_2).....P(X_n)}$$

K-nearest neighbors

K-NN can use in both classification and regression techniques commonly used in classification problems. It makes no guesses data. K is the number of neighbors. K is always an odd value. The Euclidean distance has been used to compute the distance between both the observed point and its neighbor. K-NN stores data and identifies new data based on its similarity to the existing data. It is an absence of any hypotheses about the underlying distribution. This algorithm makes training faster because no training data is needed in model generation. In the testing phase, all training data was used.

CHAPTER IV: RESULT & DISCUSSION

4.1 Performance Evaluation

In this study, I compare three different machine learning methods. The algorithms are the Support vector machine, K-nearest neighbors, Naïve Bayas algorithm, which was trained to solve the multiclass classifier problem. 30 significant features were selected by the SelectKBest method. The performance of classifiers was measured by different statistical metrics. Those are – (i) Accuracy,

(ii) Recall,

(iii) Precision, and

(iv) f1-score

Overall performance can be evaluated by Recall and Precision value. Recall referring to the accuracy of the model. On the other hand, precision indicates how well the models can correctly identify CN, AD, and MCI. The formula of the performance classifiers-

(i) Accuracy = (TP + TN) / (TP + TN + FP + FN)

(ii) Recall = TN / (TN + FP)

(iii) Precision = TP/ (TP + FN)

(iv) F1- measure = 2*TP / (2*TP + FP + FN)

Here, TP = TruePositive

TN = TrueNegative

FP = FalsePositive

FN = FalseNegative

4.2 Result & discussion

In this study, the model was designed for detecting cognitive normal, mild cognitive impairment, and Alzheimer's disease of individuals based on ADNI data. The dataset contains-cognitive data, clinical data, etc. I find out 30 significant features with the help of a feature selection technique. Then I find out the correlation between the selected 30 features and the label feature. Figure 6 show the correlation among the selected feature. From figure 6, AD_CGH_L, ADMCI3, ADASQ4, APOE4, AV45, AV45AB12, RAVLT_perc_forgetting, MD_CGH_L, RD_CGH_L are highly correlated with the label feature.

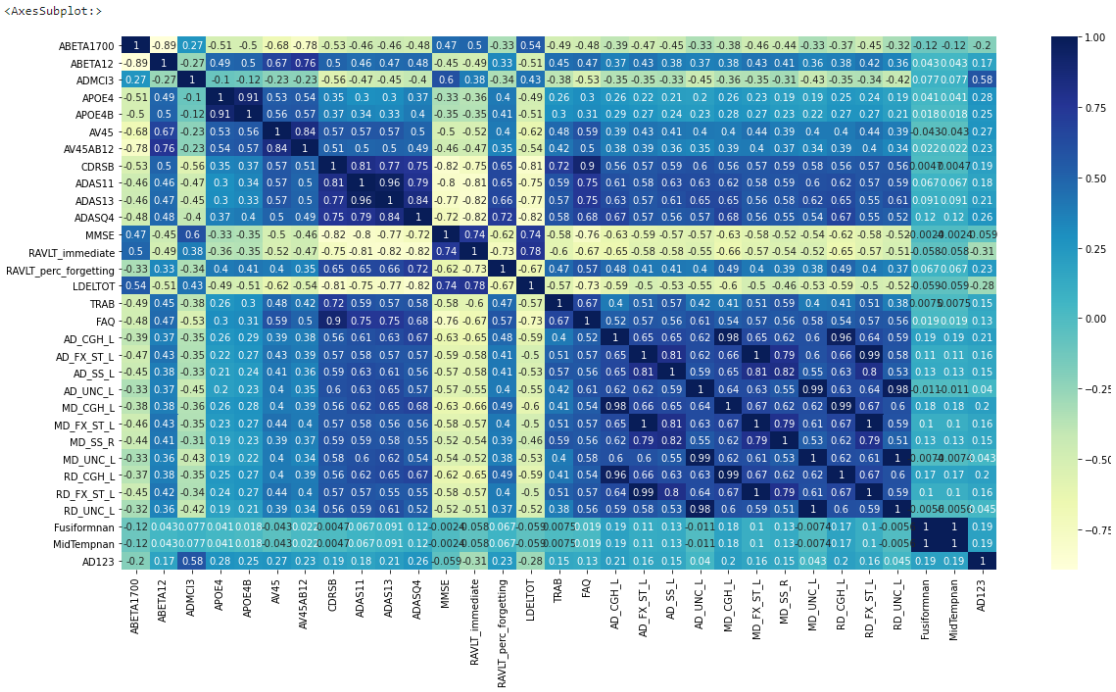


Fig8: Data correlation

For early detection of the disease, I implement machine learning models on the 30 significant features. I implement a support vector machine, naïve bayes, and k-nearest neighbor model to detect the disease. SVM, K-NN, and NB provide 93.84%, 84.61%, and 96.92% accuracy. This model successful can predict the result. I implement K-

nearest neighbors and the Naïve Bayes algorithm for comparison. For better accuracy, I used a min-max scalar to scale the dataset. And SMOTE technique is used to balance the dataset. K-NN provides 84% accuracy. To increase the accuracy of the K-NN model, I select the K value 15. NB gives 96.92% accuracy. I compared the three models in table 2 with the performance classifier score. NB provides better accuracy than KNN and SVM. 80% of data has been used for training the model. 20% of the data were used to evaluate the model. To enhance the accuracy, I used oversampling technique. Table 3 shows the performance of the model.

Table3: model performance

Algorithm	Stages	Precision	Recall	F1 score	Accuracy
K-NN	CN	0.72	0.86	0.78	0.84
	AD	1.00	1.00	1.00	
	MCI	0.82	0.67	0.74	
SVM	CN	0.90	0.90	0.90	0.93
	AD	1.00	1.00	1.00	
	MCI	0.90	0.90	0.90	
NB	CN	1.00	0.90	0.95	0.96
	AD	1.00	1.00	1.00	
	MCI	0.91	1.00	0.95	

Figures 7, 8, and, 9 show the confusion matrix of the support vector machine, naïve bayes, and k-nearest neighbor model. It compared the predicted value and actual value.

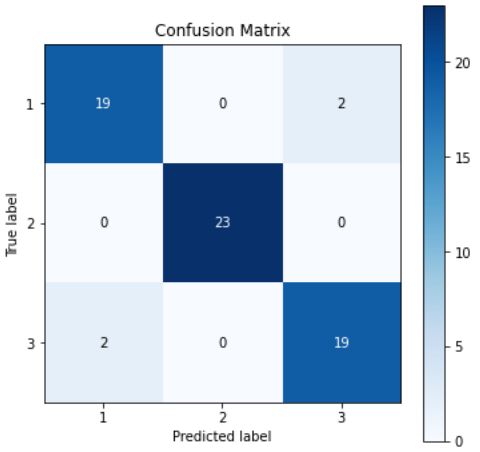


Fig9: Confusion matrix (SVM)

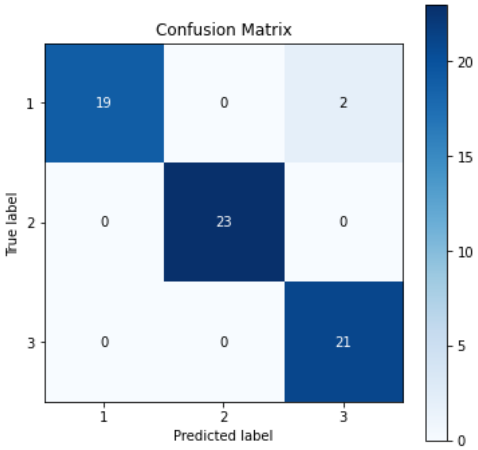


Fig10: Confusion matrix (NB)

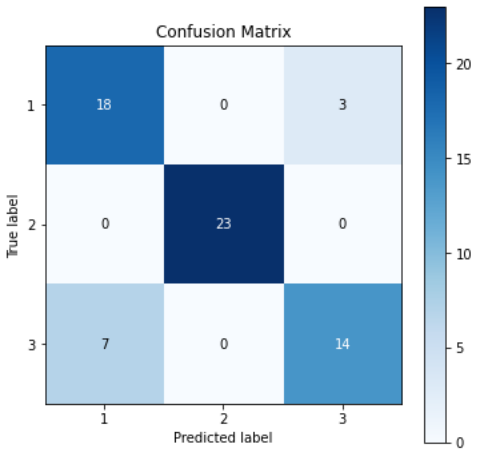


Fig11: Confusion matrix (K-NN)

CHAPTER V: CONCLUTION

In this contribution, I proposed a model which can automatically detect the stages of Alzheimer's disease. It will help the patient to faster detection. It is impossible to diagnose this disease with a single test. The dataset from ADNI has been utilized in this study. The dataset contains both neuroimaging diagnosis data and clinical psychological data. With the help of the SelectedKBest feature selection technique, I found the significant 30 features from neuroimaging diagnosis and clinical psychological data. I compared various machine learning models to find out the best model. Now it's easier to figure out the stages of the disease. NB provides the highest accuracy compared with another algorithm. The accuracy of the NB model is 96.92%. This thesis also has limitations. The dataset I used in this thesis was very poor. The accuracy can be more improved by a large number of the dataset. In the future, I will use brain images to detect the stages of Alzheimer’s disease.

REFERENCE

1. Liu, Lin; Zhao, Shenghui; Chen, Haibao; Wang, Aiguo (2019). A New Machine Learning Method for Identifying Alzheimer's Disease. Simulation Modelling Practice and Theory, (), 102023–. doi:10.1016/j.simpat.2019.102023
2. Fan, Z., Xu, F., Qi, X. et al. Classification of Alzheimer’s disease based on brain MRI and machine learning. Neural Comput & Applic 32, 1927–1936 (2020). https://doi.org/10.1007/s00521-019-04495-00
3. (2019). 2019 Alzheimer's disease facts and figures. Alzheimer's & Dementia, 15(3), 321–387. doi:10.1016/j.jalz.2019.01.010
4. A. Thushara, C. UshaDevi Amma, A. John and R. Saju, "Multimodal MRI Based Classification and Prediction of Alzheimer’s Disease Using Random Forest Ensemble," 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), 2020, pp. 249-256, doi: 10.1109/ACCTHPA49271.2020.9213211.
5. Kuang, Jie; Zhang, Pin; Cai, TianPan; Zou, ZiXuan; Li, Li; Wang, Nan; Wu, Lei (2020). Prediction of transition from mild cognitive impairment to <scp>Alzheimer's</scp> disease based on a logistic regressionâ□“artificial neural networkâ□“decision tree model. Geriatrics & Gerontology International, (), ggi.14097–. doi:10.1111/ggi.14097
6. De, Arijit; Chowdhury, Ananda S. (2020). DTI based Alzheimer disease classification with rank modulated fusion of CNNs and random forest. Expert Systems with Applications, (), 114338–. doi:10.1016/j.eswa.2020.114338
7. Talwar, P., Kushwaha, S., Chaturvedi, M. et al. Systematic Review of Different Neuroimaging Correlates in Mild Cognitive Impairment and Alzheimer’s

Disease. Clin Neuroradiol 31, 953–967 (2021). https://doi.org/10.1007/s00062-021-01057-7

8. Kishore, P., Usha Kumari, C., Kumar, M. N. V. S. S., & Pavani, T. (2020). Detection and analysis of Alzheimer's disease using various machine learning algorithms. Materials Today: Proceedings. doi:10.1016/j.matpr.2020.07.645b
9. Zhang, Fan; Li, Zhenzhen; Zhang, Boyan; Du, Haishun; Wang, Binjie; Zhang, Xinhong (2019). Multi-modal Deep Learning Model for Auxiliary Diagnosis of Alzheimer's Disease. Neurocomputing, (), S092523121930921X–. doi:10.1016/j.neucom.2019.04.093
10. Rohini, M., Surendran, D. Toward Alzheimer's disease classification through machine learning. Soft Comput 25, 2589–2597 (2021). https://doi.org/10.1007/s00500-020-05292-x
11. M. S. Ali, M. K. Islam, J. Haque, A. A. Das, D. S. Duranta and M. A. Islam, "Alzheimer's Disease Detection Using m-Random Forest Algorithm with Optimum Features Extraction," 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), 2021, pp. 1-6, doi: 10.1109/CAIDA51941.2021.9425212.
12. J. Neelaveni and M. S. G. Devasana, "Alzheimer Disease Prediction using Machine Learning Algorithms," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 101-104, doi: 10.1109/ICACCS48705.2020.9074248.
13. Ghoraani, Behnaz; Boettcher, Lillian N.; Hssayeni, Murtadha D.; Rosenfeld, Amie; Tolea, Magdalena I.; Galvin, James E. (2021). Detection of mild cognitive impairment and Alzheimerâ™s disease using dual-task gait

assessments and machine learning. Biomedical Signal Processing and Control, 64(), 102249–. doi:10.1016/j.bspc.2020.102249

14. Alickovic E., Subasi A., for the Alzheimer's Disease Neuroimaging Initiative (2020) Automatic Detection of Alzheimer Disease Based on Histogram and Random Forest. In: Badnjevic A., Škrbić R., Gurbeta Pokvić L. (eds) CMBEBIH 2019. CMBEBIH 2019. IFMBE Proceedings, vol 73. Springer, Cham. https://doi.org/10.1007/978-3-030-17971-7_14
15. Rallabandi, V.P. Subramanyam; Tulpule, Ketki; Gattu, Mahanandeeshwar (2020). Automatic classification of cognitively normal, mild cognitive impairment and Alzheimer's disease using structural MRI analysis. Informatics in Medicine Unlocked, (), 100305–. doi:10.1016/j.imu.2020.100305
16. Kruthika, K.R.; Rajeswari,; Maheshappa, H.D. (2018). Multistage classifier-based approach for Alzheimer's disease prediction and retrieval. Informatics in Medicine Unlocked, (), S2352914818301758–. doi:10.1016/j.imu.2018.12.003
17. P. Lodha, A. Talele and K. Degaonkar, "Diagnosis of Alzheimer's Disease Using Machine Learning," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697386.
18. Battista, P., Salvatore, C., Berlingeri, M., Cerasa, A., & Castiglioni, I. (2020). ARTIFICIAL INTELLIGENCE AND NEUROPSYCHOLOGICAL MEASURES: THE CASE OF ALZHEIMER'S DISEASE. Neuroscience & Biobehavioral Reviews. doi:10.1016/j.neubiorev.2020.04.0
19. Li, Wei; Lin, Xuefeng; Chen, Xi (2020). Detecting Alzheimer's disease Based on 4D fMRI: An exploration under deep learning framework. Neurocomputing, (), S0925231220301041–. doi:10.1016/j.neucom.2020.01.053

20. H. M. T. Ullah, Z. Onik, R. Islam and D. Nandi, "Alzheimer's Disease and Dementia Detection from 3D Brain MRI Data Using Deep Convolutional Neural Networks," 2018 3rd International Conference for Convergence in Technology (I2CT), 2018, pp. 1-3, doi: 10.1109/I2CT.2018.8529808.
21. M. Hon and N. M. Khan, "Towards Alzheimer's disease classification through transfer learning," 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017, pp. 1166-1169, doi: 10.1109/BIBM.2017.8217822.
22. Bhagya Shree, S.R., Sheshadri, H.S. Diagnosis of Alzheimer's disease using Naive Bayesian Classifier. Neural Comput & Applic 29, 123–132 (2018). https://doi.org/10.1007/s00521-016-2416-3