

BRAIN DISEASE DETECTION USING DEEP LEARNING ALGORITHMS

PROJECT REPORT

Submitted to Mahatma Gandhi University

*In partial fulfillment of the requirements for the award of
Degree in Master of Science in Computer Science*



By

AAFIYA PN

(Reg.No : 200011022418)

Department of Computer Applications

MES COLLEGE MARAMPALLY

(NAAC Reaccredited with A+ Grade, Affiliated to Mahatma Gandhi
University)

Marampally P.O, Aluva, Ernakulam, Kerala-683105

2020- 2022

DEPARTMENT OF COMPUTER APPLICATIONS

MES COLLEGE MARAMPALLY

(NAAC Reaccredited with A+ Grade, Affiliated to Mahatma Gandhi University)

Marampally P.O, Aluva, Ernakulam, Kerala-683105



CERTIFICATE

*This is to certify that the project report entitled **BRAIN DISEASE DETECTION USING DEEP LEARNING ALGORITHMS** is a bonafide record of work carried out by **Aafiya PN (Reg.No : 200011022418)** in partial fulfillment of the requirements for the award of Degree of Master of Science in Computer Science, Mahatma Gandhi University, Kottayam during the period 2020-2022.*

Project guide

Dr .Leena C Sekhar

Head of the Department

Dr. Leena C Sekhar

Submitted for the viva-voce held

on.....

External Examiner 1

External Examiner 2

DECLARATION

I hereby declare that the project entitled **BRAIN DISEASE DETECTION USING DEEP LEARNING ALGORITHMS** being submitted in partial fulfillment of the requirement for the award of Degree in Master of Science in Computer Science is the original work done by myself under the guidance of **Dr.Leena C Sekhar**, Associate Professor, Department of Computer Applications, MES College Marampally during the academic year 2020-2022. It has not formed the part of any other work submitted for award of any degree or diploma, either in this or any other University.

Place : Marampally

Signature of candidate

Date :

AAFIYA PN
(Reg.No : 200011022418)

ACKNOWLEDGEMENT

It has been said that gratitude is the memory of heart. Hence, I take this opportunity to express my gratitude to all those, whose contribution in this project can't be forgotten. First and foremost I thank The Almighty for his showers of blessings on this work and for endowing me the will to complete it on time. I am grateful to **Dr.Leena C Sekhar** , Head of the Department of Computer Application, MES College Marampally, for his able leadership and guidance in all the official matters regarding this work. I would like to thank, **Dr.Leena C Sekhar** , my guide and Associate Professor, Department of Computer Applications, MES College Marampally, for her inspiring and commendable support throughout the course and especially for this work. I also express my sincere thanks to our class tutor **Dr. Jaseena K.U**, Assistant Professor, Department of Computer Applications , for her expert guidance, constant encouragement and valuable suggestions. I would also like to express my gratitude to all other faculty members of the computer science department for their cooperation. I express my sincere thanks to my parents, teachers and friends those who had provide all the support both directly and indirectly during this work.

ABSTRACT

The nervous system comprised by the brain, spinal cord, and nerves is in effect the control center for the body. It reaches from our head to the (nerves in) tips of our fingers and toes. When it's working well, it allows us to function on all levels to walk, speak, breathe and swallow. The brain is amazingly complex. But when problems do occur within the brain, these can have a profound and even debilitating impact. In some cases, neurological conditions are often fatal. About one in six suffers from brain or neuropathy, and the annual cost of treatment exceeds a million. Historically, illnesses without effective treatment have been used to explain neurological and brain disorders. Brain damage is often fatal and very complex. Early detection of these diseases is important. Diagnosis is often made after the radiologist confirms a CT / MRI scan of the brain, which can take a long time. Computer-aided diagnosis (CAD) has revolutionized the way brain diseases are diagnosed. Diseases are identified using a variety of deep learning techniques and neural network ideas. The concept of image pattern recognition was also used. Convolutional Neural Network (CNN) with various combinations of deep learning algorithms are also effective.

Contents

Acknowledgement	ii
1 CHAPTER 1	
INTRODUCTION	viii
1.1 Introduction	viii
1.2 Problem definition	ix
1.3 Objectives	x
1.4 Report Organization	xi
2 CHAPTER 2	
LITERATURE REVIEW	xii
3 CHAPTER 3	
PROPOSED METHODOLOGY	xvii
3.1 Datasets	xvii
3.2 Methods Employed	xix
3.2.1 EXPLORATORY DATA ANALYSIS (EDA) . . .	xix
3.2.2 Scikit-learn	xxi
3.2.3 RANDOM FOREST	xxi
3.3 Proposed Framework	xxiv
4 CHAPTER 4	
EXPERIMENTAL RESULTS AND DISCUSSION	xxvii
4.1 Hardware and Software Code	xxvii
4.2 Data Preprocessing	xxvii
4.2.1 Check missing values by each column	xxvii

4.2.2	Dropped the 8 rows with missing values in the column, SES	xxviii
4.2.3	Imputation	xxviii
4.2.3.1	xxviii
4.2.3.2	xxviii
4.3	Splitting into Training and Testing	xxix
4.3.1	Dataset with imputation	xxix
4.3.2	splitting into three sets	xxix
4.3.3	Feature scaling	xxix
4.3.4	Dataset after dropping missing value rows	xxx
4.3.5	splitting into three sets	xxx
4.3.6	Feature scaling	xxx
4.4	Random Forest Classification	xxx
4.5	Result Analysis	xxxii
4.5.1	Result Tables	xxxii

5 CHAPTER 5

CONCLUSION AND FUTURE SCOPE xxxvi

6 REFERENCES xxxviii

List of Tables

4.1	Missing Values	xxxii
4.2	Performance Overview	xxxiii
4.3	Performance Evaluation	xxxv

List of Figures

2.1	LITERATURE REVIEW	xvi
3.1	COLUMN AND DESCRIPTION	xviii
3.2	EDA	xix
3.3	Gender and Dementia	xx
3.4	Mini Mental State Examination	xx
3.5	Random Forest Architecture	xxii
3.6	Precision and recall	xxv
3.7	Proposed Model	xxvi
4.1	EDUC vs SES	xxxiii
4.2	Confusion Matrix	xxxiv

CHAPTER 1

INTRODUCTION

1.1 Introduction

The goal of this research is to use several deep learning algorithms to build an automated early Alzheimer's disease diagnosis. Alzheimer's disease is a serious neurological brain ailment in which the patient's brain cells are destroyed and memory loss occurs. Early recognition and treatment of Alzheimer's disease signs can help the patient's condition to improve. Preceding research has looked into employing microwave methods in conjunction with 2D stochastic models that use deep neural networks or convolutional neural networks to identify strokes and hemorrhages. In this research Convolutional neural networks and multi-modal deep learning algorithms are used to identify strokes, hemorrhages, and Alzheimer's diseases and for early treatment. A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. Multimodal learning research focuses on developing models that combine multiple modes of data with varying structures such as sequential relationships between words in natural language and spatial pixel relationships in images. These models aspire to create joint representation of the input data that provide richer features for downstream tasks compared to models leveraging a single mode of data.

1.2 Problem definition

Alzheimer's disease (AD) is a neurodegenerative disease of unknown cause and etiology that affects mainly the elderly and is the most common cause of dementia. The earliest clinical symptom of AD is selective memory loss, and treatments are available to relieve some of the symptoms, but there is currently no cure. Magnetic resonance imaging (MRI) images of the brain are used to evaluate patients with suspected AD. MRI findings include both local and systemic contractions of brain tissue. Several studies suggest that MRI function can predict AD remission rates and guide treatment in the future. However, to reach this stage, clinicians and researchers need to employ machine learning techniques that can accurately predict the progression of patients from mild cognitive impairment to dementia. I suggest developing a robust model to help clinicians predict early-stage Alzheimer's disease.

1.3 Objectives

Alzheimer's disease is a serious neurological disease that can damage the brain cells, leading to memory loss. Early recognition and treatment of the symptoms of Alzheimer's disease can help improve the patient's condition. Magnetic resonance imaging (MRI) of the brain is used to screen patients with suspected AD. MRI results include both local and systemic contractions of brain tissue. Several studies show that MRI function can predict the likelihood of remission of Alzheimer's disease and help with further treatment. The main goal of this project is to develop an automated classification scheme for screening MRI scans of the brain using digital image processing and deep learning methods. This is a more accurate assessment method. The aim of this project is to extract the area from the MRI, determine the percentage of each, and plot it as a bar graph for future analysis. The proposed automated classification technique should achieve the best overall accuracy, precision, recall, and f1 score possible.

1.4 Report Organization

The material presented in this report is organized into six chapters. After this introductory chapter, chapter 2 explains the "Literature review" which describes similar project schemes. These project works and their outcomes are explained in the literature review. Chapter 3 gives an overall outline about the proposed methodology. It consist of several sub sections such as "Datasets", "Methods Employed", and "Proposed Framework". The sub section "Datasets" contain a brief discussion about the dataset which is used in this work. The other two sub section provide a broad view of project. Chapter 5 is discussing about "Experimental results and Discussion". Here, The accuracy of result is compared and significant code snippets are explained. Chapter 6 summarizes the project. It includes "Conclusion and Future scope". Finally, Chapter 6 "References" compiles the reference papers, articles, and website links.

CHAPTER 2

LITERATURE REVIEW

The brain is incredibly complex. It can process huge amounts of information quickly and efficiently. Some people with the nervous system problems can have a big impact on their health. Some of these problems can lead to serious neurological conditions, which can be fatal. In the era of modern medical science, Brain diseases is an incomplete research area. These researches are rapidly progressing. But, the base is the previous works and papers.

Zhang, Y., et al proposed a new CAD system for brain MRI imaging based on own brain and machine learning with two goals: accurate detection of both AD subjects and AD-related brain regions. First, they used the maximum intra-class variance (ICV) to select key slices from 3D volumetric data. Second, the researchers generated an eigenbrain set for each subject. Third, the most important eigenbrain was obtained by Welch's t-test (WTT). Finally, kernel support vector machines with different kernels trained by particle swarm optimization were used to make an accurate prediction of AD subjects. Discriminant regions that distinguish Alzheimer's disease (AD) from normal cognitive function (NC) were highlighted using coefficients of MIE above 0.98 quantiles. Experiments have shown that the proposed method can predict AD with competitive performance compared to existing methods, especially the accuracy of the polynomial kernel (92.36 ± 0.94) was better than the linear kernel 91.47 ± 1.02 and the radial basis function kernel (RBF). 86.71 ± 1.93

Jain, R., et al says there are several recent studies, that have used brain MRI scans and deep learning have shown promising results for diagnosis of Alzheimer's disease. However, most common issue with deep learning

architectures such as CNN is that they require large amount of data for training. In this paper, a mathematical model PFSECTL based on transfer learning is used in which a CNN architecture, VGG-16 trained on ImageNet dataset is used as a feature extractor for the classification task. Experimentation is performed on data collected from Alzheimer's Disease Neuroimaging Initiative database. The accuracy of the 3-way classification using the described method is 95. 73

Kim H,W.,et al have proposed a deep learning-based method for diagnosis of Alzheimer's disease (AD) that is less sensitive to different datasets for external validation, based upon F-18 fluorodeoxyglucose positron emission tomography/computed tomography (FDG-PET/CT). The accuracy, sensitivity, and specificity of our proposed network were 86.09, 80.00, and 92.96 using their dataset, and 91.02, 87.93, and 93.57 (respectively) using the ADNI dataset. They observed that their model classified AD and normal cognitive (NC) cases based on the posterior cingulate cortex (PCC), where pathological changes occur in AD. The performance of the GAP layer was considered statistically significant compared to the fully connected layer in both datasets for accuracy, sensitivity, and specificity ($p < 0.01$). In addition, performance comparison between the ADNI dataset and their dataset showed no statistically significant differences in accuracy, sensitivity, and specificity ($p > 0.05$). Their model demonstrated the effectiveness of AD classification using the GAP layer.

Venugopalan.J.,et al says, Nevertheless, research studies often have little effect on clinical practice mainly due to the following reasons: (1) Most studies depend mainly on a single modality, especially neuroimaging; (2) diagnosis and progression detection are usually studied separately as two independent problems; and (3) current studies concentrate mainly on optimizing the performance of complex machine learning models, while dis-

regarding their explainability. In this paper, they carefully developed an accurate and interpretable AD diagnosis and progression detection model. Their model provides physicians with accurate decisions along with a set of explanations for every decision. In the first layer, the model carries out a multi-class classification for the early diagnosis of AD patients. In the second layer, the model applies binary classification to detect possible MCI-to-AD progression from a baseline diagnosis. The proposed system helps to enhance the clinical understanding of AD diagnosis and progression processes by providing detailed insights into the effect of different modalities on the disease risk.

Zoetmulder, R., et al says most current Alzheimer's disease and mild cognitive disorders studies use single data modality to make predictions such as AD stages. Thus, they used deep learning to integrally analyze imaging (magnetic resonance imaging), genetic (single nucleotide polymorphisms), and clinical test data to classify patients into AD, MCI, and controls. They also developed a novel data interpretation method to identify top-performing features learned by the deep-models with clustering and perturbation analysis.

Moradi, Elaheh., et al First, developed a novel MRI biomarker of MCI-to-AD conversion using semi-supervised learning and then integrated it with age and cognitive measures about the subjects using a supervised learning algorithm resulting in what we call the aggregate biomarker. With the ADNI data, the MRI biomarker achieved a 10-fold cross-validated area under the receiver operating characteristic curve (AUC) of 0.7661 in discriminating progressive MCI patients (pMCI) from stable MCI patients (sMCI). Their aggregate biomarker based on MRI data together with baseline cognitive measurements and age achieved a 10-fold cross-validated AUC score of 0.9020 in discriminating pMCI from sMCI. The results pre-

sented in this study demonstrate the potential of the suggested approach for early AD diagnosis and an important role of MRI in the MCI-to-AD conversion prediction.

TITLE	AUTHORS	METHODS EMPLOYED
<i>Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning</i>	Yudong Zhang, Zhengchao Dong, Preetha Phillips, Shuihua Wang, Genlin Ji, Jiquan Yang and Ti-Fei Yuan	K-FoldCV,key-slice selection,RBF, eigen brains
<i>Convolutional neural network based Alzheimer's disease classification from magnetic resonance brain images</i>	Rachna Jain, Nikita Jain, Akshay Aggarwal , D. Jude Hemanth	PFSECTL, CNN,VGG-16
<i>Multi-slice representational learning of convolutional neural network for Alzheimer's disease classification using positron emission tomography</i>	Han Woong Kim, Ha Eun Lee, KyeongTaek Oh , Sangwon Lee, Mijin Yun and Sun K. Yoo	FDG-PET/CT, CNN
<i>Multimodal deep learning models for early detection ofAlzheimer's disease stage</i>	JananiVenugopalan, LiTong, Hamid Reza Hassanzadeh & May D. Wang	SVM, RANDOM FOREST, DOUBLE BRANCH CNN
<i>A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease</i>	Shaker El-Sappagh, Jose M.Alonso, S. M. Riazul Islam, Ahmad M. Sultan& Kyung Sup Kwak	ADNI-Study,RF BLACK BOX Study
<i>Automated Final Lesion Segmentation in Posterior Circulation Acute Ischemic Stroke Using Deep Learning</i>	Riaan Zoetmulder, Praneeta R. Konduri, Iris V. Obdeijn, Efstratios Gavves, Ivwana Išgum, Charles B.L.M. Majoie, Diederik W.J. Dippel, Yvo B.W.E.M. Roos, Mayank Goyal, Peter J. Mitchell, Bruce C. V. Campbell, Demetrius K. Lopes, Gernot Reimann, Tudor G. Jovin, Jeffrey L. Saver, Keith W. Muir, Phil White, Serge Bracad, Bailiang Chen, Scott Brown, Wouter J. Schonewille, Erik van der Hoeven, Volker Puetz and Henk A. Marquering.	ICC, BLAND ALTMAN ANALYSIS, FISHERS r-to-z TRANSFORMATION
Machine learning framework for early MRI-based Alzheimers conversion prediction in MCI-subjects	Elaheh Moradi, Antonietta Pepe, Christian Gaser, Heikki Huttunen, Jussi Tohka	MCI-to-AD conversion using semi-supervised learning

Figure 2.1: LITERATURE REVIEW

CHAPTER 3

PROPOSED METHODOLOGY

3.1 Datasets

The data used is MRI-related data from the Open Access Series of Imaging Studies (OASIS) project, which can be used to train different machine learning models to recognise patients with mild to moderate dementia. This data is accessible on both their website and Kaggle.

- using the longitudinal MRI data.
- The dataset consists of a longitudinal MRI data of 150 subjects aged 60 to 96.
- A minimum of one scan was performed on each subject.
- There is no left-handed person.
- The study included 72 subjects, of which 72 were labelled "Nondemented."
- At the time of their initial visits, 64 of the subjects were categorised as "Demented," and they stayed in that category for the duration of the study.
- At the time of their second visit, 14 subjects who had previously been classified as "Nondemented" were now considered to have dementia. The "Converted" category includes these.

COL	FULL-FORMS
EDUC	Years of education
SES	Socioeconomic Status
MMSE	Mini Mental State Examination
CDR	Clinical Dementia Rating
eTIV	Estimated Total Intracranial Volume
nWBV	Normalize Whole Brain Volume
ASF	Atlas Scaling Factor

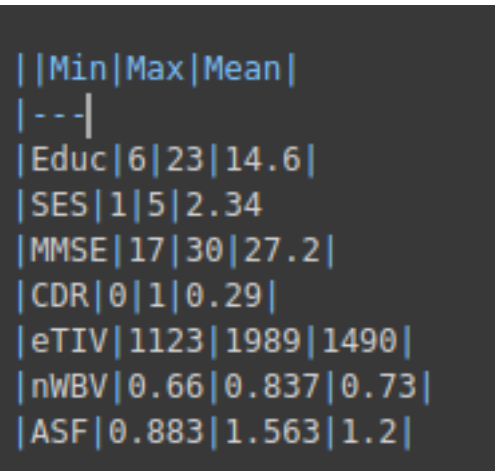
Figure 3.1: COLUMN AND DESCRIPTION

3.2 Methods Employed

3.2.1 EXPLORATORY DATA ANALYSIS (EDA)

This section has been primarily concerned with examining the connection between each MRI test feature and the patient's dementia. In order to explicitly state the relationship between the data and a graph before extracting or analysing the data, we carried out this exploratory data analysis process. We might use it to better comprehend the nature of the data and choose the best analysis technique for the model in the future.

For the purpose of implementing a graph, the minimum, maximum, and average values of each feature are as follows.

A terminal window with a dark background and light blue text. It displays a table with three columns: Min, Max, and Mean. The table lists statistics for various features: Educ, SES, MMSE, CDR, eTIV, nWBV, and ASF. The values are formatted with commas as thousands separators.

	Min	Max	Mean
Educ	6	23	14.6
SES	1	5	2.34
MMSE	17	30	27.2
CDR	0	1	0.29
eTIV	1123	1989	1490
nWBV	0.66	0.837	0.73
ASF	0.883	1.563	1.2

Figure 3.2: EDA

The graph indicates that men are more likely with dementia than women.

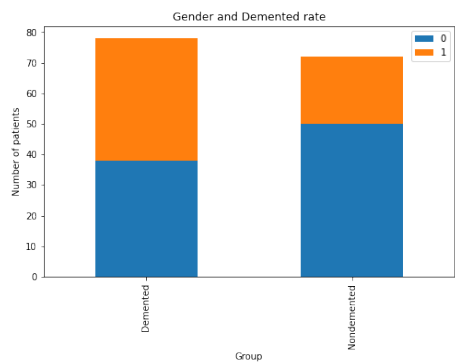


Figure 3.3: Gender and Dementia

The chart shows Nondemented group got much more higher MMSE scores than Demented group.

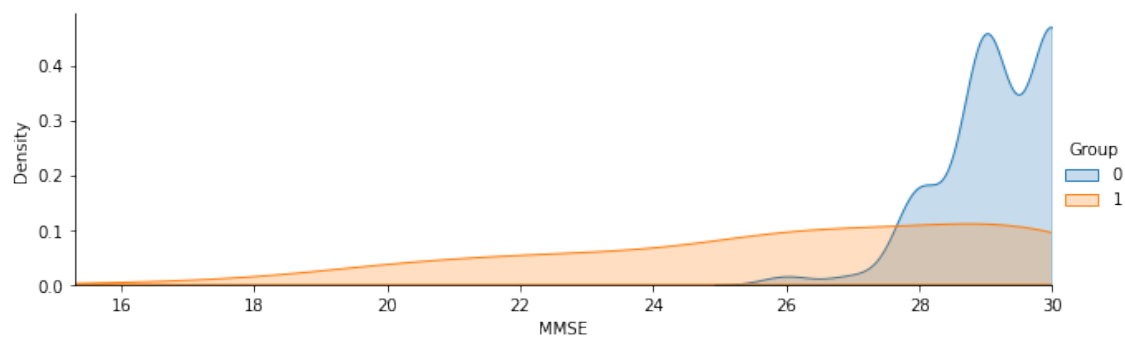


Figure 3.4: Mini Mental State Examination

- Men are more likely with demented, an Alzheimer's Disease, than Women.
- Demented patients were less educated in terms of years of education.
- Nondemented group has higher brain volume than Demented group.
- Higher concentration of 70-80 years old in Demented group than those in the nondemented patients.

3.2.2 Scikit-learn

Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. Scikit-learn is a NumFOCUS fiscally sponsored project.

3.2.3 RANDOM FOREST

Random Forest is one of the two primary traditional machine learning. One of the major advantages of the random forest algorithm is that it works well in small datasets. If the labels are manually painted labels, then it is suitable to use random forest for pixel classification problems. The second traditional algorithm is SVM(Support Vector Machines). In image processing problems, Random forest is much better than SVM. Here, ran-

dom forest is used for pixels classification.

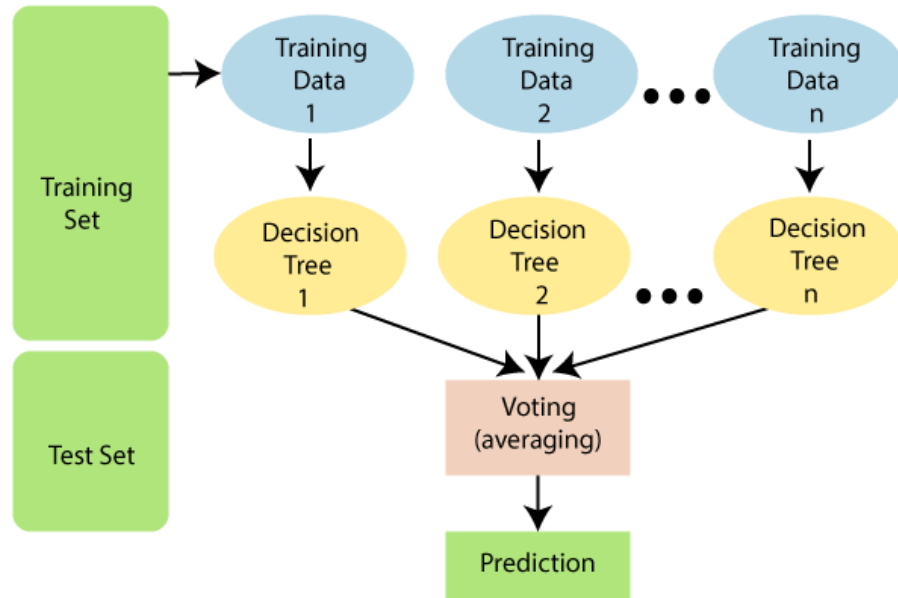


Figure 3.5: Random Forest Architecture

In "Random forest", The word forest indicates that it is a some sort of collection of trees. It is a collection of a bunch of decision trees. Random forest is a supervised machine learning method, that means, it requires labeled data. The model is trained based on that labeled data. The decision tree makes decision based on the pixel values of labeled pixel. For example, if the label value is greater than 15, that pixel is assumed as slough(just an assumption). It starts from the root node and ends at the leaf node. Between the root node and leaf node, there exist many internal nodes which depend up on the nature of the problem. In python, scikit-learn library provides a function to perform random forest classification and associated requirements. Gini impurity is used to identify the best split of decision tree. It is the probability of incorrectly classifying a randomly chosen el-

ement in the dataset. It is calculated as:

$$G(k) = \left[\sum_{i=1}^c P(i) * (1 - P(i)) \right] \quad (3.1)$$

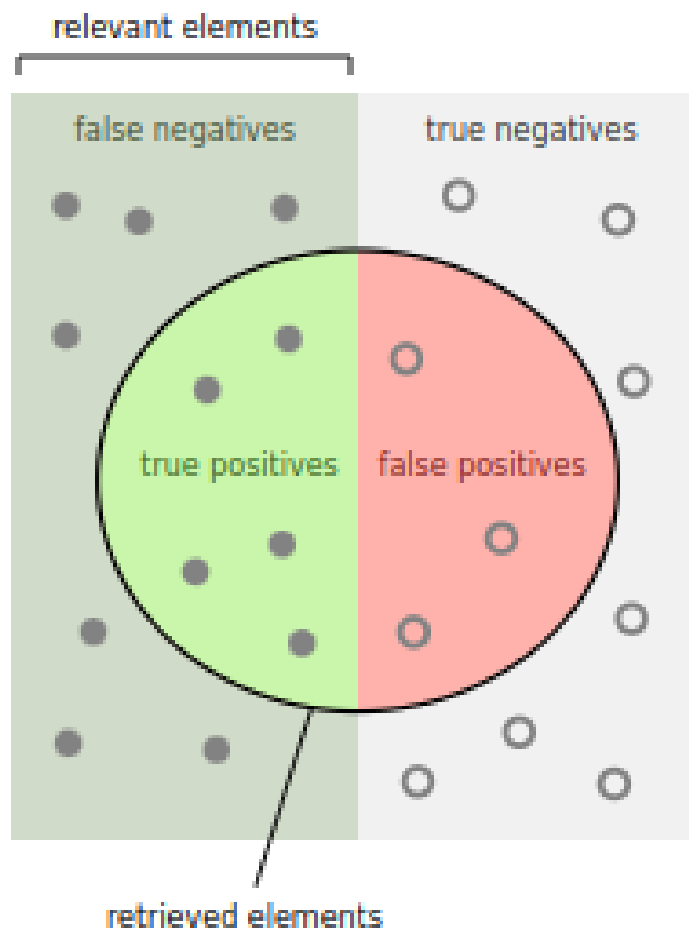
Where C is the number of classes and p(i) is the probability of randomly picking an element of class i. The primary disadvantage of decision tree is overfitting. It means, high training accuracy but low testing accuracy.

3.3 Proposed Framework

The proposed framework consists of different steps like Data preprocessing, Splitting into training and testing data sets, cross validation to figure out the best parameters for each model. In the end, we compare the accuracy, recall and AUC for each model.

The performance measure is calculated by area under the receiver operating characteristic curve (AUC). In medical diagnosis of non-life-threatening end-stage diseases, such as most neurodegenerative diseases, it is important to have a high true positive rate so that all patients with Alzheimer's disease are detected as soon as possible. However, we also want to keep the false positive rate as low as possible because we don't want to misdiagnose healthy adults with dementia and start medication. Therefore, AUC seemed to be an ideal choice for power measurement.

In the figure below, you can think relevant elements as actually demented subjects. Precision and Recall



How many retrieved items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are retrieved?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Figure 3.6: Precision and recall

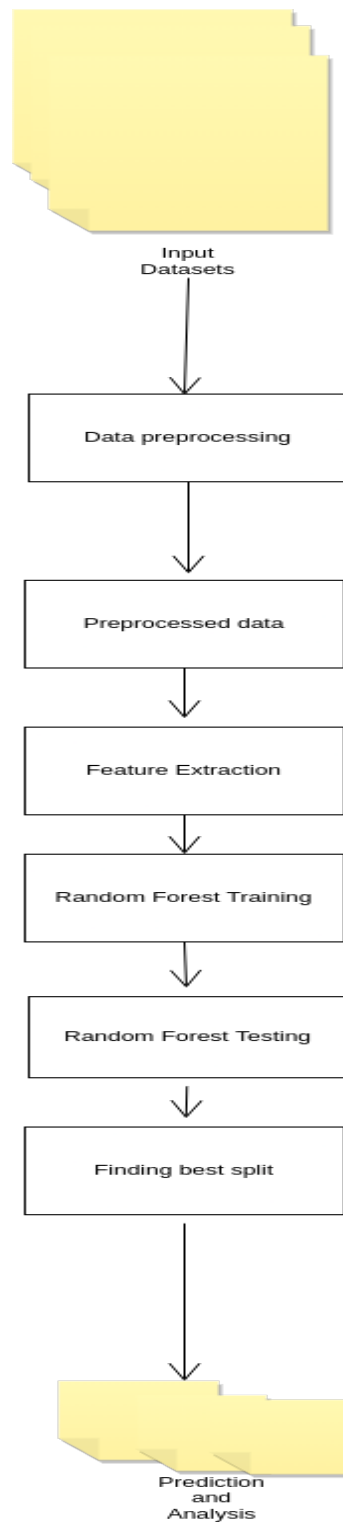


Figure 3.7: Proposed Model

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Hardware and Software Code

This project is implemented using python language. The Google Colab is used to run python code. Google Colab is a free Jupyter notebook environment that runs entirely in the cloud. Most importantly, it does not require a setup and the notebooks that you create can be simultaneously edited by your team members. Jupyter Notebook is an open-source web application that allows user to create and share documents that contain live code. The important code snippets are given below.

4.2 Data Preprocessing

Identified 8 rows with missing values in the SES column. We will deal with this problem in two approaches. One is to simply remove the row with the missing value. The other is to replace the missing value with the appropriate value. This is also known as "Imputation". Since there are only 150 data, I think the imputation will improve the performance of the model.

4.2.1 Check missing values by each column

```
pd.isnull(df).sum()
```

4.2.2 Dropped the 8 rows with missing values in the column, SES

```
df_dropna = df.dropna(axis=0, how='any')
pd.isnull(df_dropna).sum()
```

4.2.3 Imputation

Scikit-learn provides package for imputation, but we do it manually. Since the SES is a discrete variable, we use median for the imputation.

4.2.3.1

Draw scatter plot between EDUC and SES

```
x = df['EDUC']
y = df['SES']
ses_not_null_index = y[~y.isnull()].index
x = x[ses_not_null_index]
y = y[ses_not_null_index]
```

4.2.3.2

Draw trend line in red

```
z = np.polyfit(x, y, 1)
p = np.poly1d(z)
plt.plot(x, y, 'go', x, p(x), "r-")
plt.xlabel('Education Level(EDUC)')
plt.ylabel('Social Economic Status(SES)')
plt.show()
```

```
df.groupby(['EDUC'])['SES'].median()
df["SES"].fillna(df.groupby("EDUC")["SES"].transform("median"),
inplace=True)
pd.isnull(df['SES']).value_counts()
```

4.3 Splitting into Training and Testing

```
from sklearn.model_selection import train_test_split
from sklearn import preprocessing
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import cross_val_score
```

4.3.1 Dataset with imputation

```
Y = df['Group'].values
Features we use
X = df[['M/F', 'Age', 'EDUC', 'SES', 'MMSE', 'eTIV', 'nWBV',
'ASF']]
```

4.3.2 splitting into three sets

```
X_trainval, X_test, Y_trainval, Y_test = train_test_split( X, Y, ran-
dom_state=0)
```

4.3.3 Feature scaling

```
scaler = MinMaxScaler().fit(X_trainval)
X_trainval_scaled = scaler.transform(X_trainval)
```

```
X_test_scaled = scaler.transform(X_test)
```

4.3.4 Dataset after dropping missing value rows

```
Y = df_dropna['Group'].values X = df_dropna[['M/F', 'Age', 'EDUC',  
'SES', 'MMSE', 'eTIV', 'nWBV', 'ASF']]
```

4.3.5 splitting into three sets

```
X_trainval_dna, X_test_dna, Y_trainval_dna, Y_test_dna = train_test_split(  
X, Y, random_state=0)
```

4.3.6 Feature scaling

```
scaler = MinMaxScaler().fit(X_trainval_dna)  
X_trainval_scaled_dna = scaler.transform(X_trainval_dna)  
X_test_scaled_dna = scaler.transform(X_test_dna)
```

4.4 Random Forest Classification

- `n_estimators(M)`: the number of trees in the forest
- `max_features(d)`: the number of features to consider when looking for the best split
- `max_depth(m)`: the maximum depth of the tree.

```
best_score = 0
```

```
for M in range(2, 15, 2): combines M trees  
for d in range(1, 9): maximum number of features considered at
```


each split

for m in range(1, 9): maximum depth of the tree

train the model

n_jobs(4) is the number of parallel computing

```
forestModel = RandomForestClassifier(n_estimators=M, max_features=d,  
n_jobs=4, max_depth=m, random_state=0)
```

perform cross-validation

```
scores = cross_val_score(forestModel, X_trainval_scaled_dna, Y_trainval_dna,  
cv=kfolds, scoring='accuracy')
```

compute mean cross-validation accuracy

```
score = np.mean(scores)
```

if we got a better score, store the score and parameters

```
if score > best_score: best_score = score best_M = M best_d = d  
best_m = m
```

Rebuild a model on the combined training and validation set

```
SelectedRFModel = RandomForestClassifier(n_estimators=M, max_features=d,  
max_depth=m, random_state=0).fit(X_trainval_scaled_dna, Y_trainval_dna  
)
```

```
PredictedOutput = SelectedRFModel.predict(X_test_scaled) test_score  
= SelectedRFModel.score(X_test_scaled, Y_test) test_recall = recall_score(Y_test,  
PredictedOutput, pos_label=1)  
fpr, tpr, thresholds = roc_curve(Y_test, PredictedOutput, pos_label=1)  
test_auc = auc(fpr, tpr)  
print("Best accuracy on validation set is:", best_score)
```

```
print("Best parameters of M, d, m are: ", best_M, best_d, best_m)
print("Test accuracy with the best parameters is", test_score)
print("Test recall with the best parameters is:", test_recall)
print("Test AUC with the best parameters is:", test_auc)
```

```
m = 'Random Forest'
acc.append([m, test_score, test_recall, test_auc, fpr, tpr, thresholds])
```

4.5 Result Analysis

From Data Preprocessing we found that the dataset has 8 missing values

Table 4.1: Missing Values

Subject ID	0
Group	0
M/F	0
Age	0
EDUC	0
SES	8
MMSE	0
CDR	0
eTIV	0
nWBV	0
ASF	0

So after Dropping these values we imputed the dataset using median of EDUC and SES

4.5.1 Result Tables

The overall accuracy of the system is 92.105. A confusion matrix is a table that is often used to describe the performance of a clas-

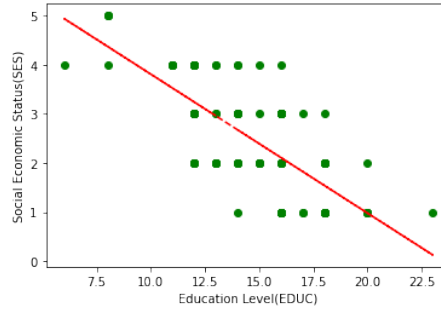


Figure 4.1: EDUC vs SES

sification model (or "classifier"). It measures the performance for machine learning classification problem where output can be two or more classes. Here, there are two classes: Demented,Nondemented.

Table 4.2: Performance Overview

	precision	recall	f1-score	support
0	1.00	0.86	0.92	21
1	0.85	1.00	0.92	17
accuracy			0.92	38
macro avg	0.93	0.93	0.92	38
weighted avg	0.93	0.92	0.92	38

Consider the Confusion matrix of the system. It contains 4 different combinations of predicted and actual values: True positive, False Positive, False Negative, and True Negative. It is very much useful for measuring Recall, Precision, Specificity, and Accuracy. True Positive means that, system predicted positive and its true. True Negative means that, system predicted negative and its true. False Positive means that, system predicted positive and its false. False Negative means that, system predicted negative and its false. The diagonal section of confusion matrix indicates True Positive.

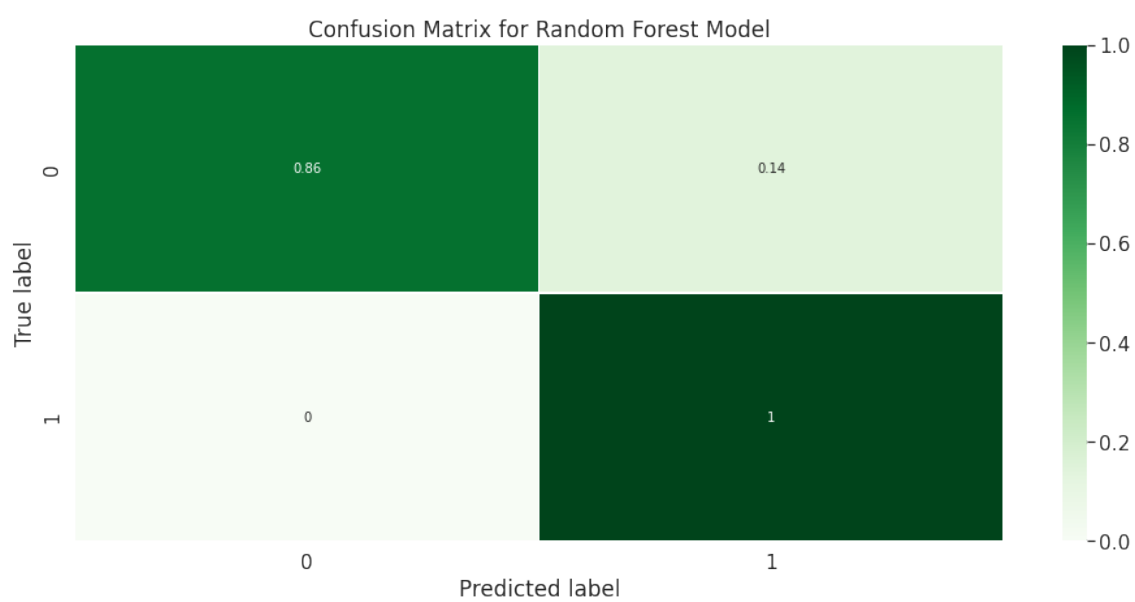


Figure 4.2: Confusion Matrix

Overall, the proposed method is able to generate efficient results in detection and classification of Alzheimers Disease. Consider the below table that dhowes the performance evaluation of various classifiers.

Table 4.3: Performance Evalutation

Index	Model	Accuracy	Recall	AUC
0	Logistic Regression (w/ imputation)	0.763158	0.70	0.766667
1	Logistic Regression (w/ dropna)	0.805556	0.75	0.750000
2	SVM	0.789474	0.75	0.791667
3	Decision Tree	0.815789	0.65	0.825000
4	Random Forest	0.921053	0.85	0.925000
5	AdaBoost	0.894737	0.85	0.897222

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

Alzheimer's disease is a serious neurological disease that can damage the brain cells, leading to memory loss. Early recognition and treatment of the symptoms of Alzheimer's disease can help improve the patient's condition. Magnetic resonance imaging (MRI) of the brain is used to screen patients with suspected AD. MRI results include both local and systemic contractions of brain tissue. Several studies show that MRI function can predict the likelihood of remission of Alzheimer's disease and help with further treatment. The proposed system provides cost effective technical support to the conventional detection method. The proposed version offers us a accuracy rate of 92.5 that is the maximum accuracy rate received in comparison to the other studies on this field. The distinctiveness of our technique is the reality that we'd be including metrics like MMSE and Education additionally in our version to train it to differentiate among ordinary healthy adults and people with Alzheimer's. MMSE is one of the gold requirements for figuring out dementia and as a result, we assume it's far an essential function to include.

The same fact additionally make our technique flexible enough to be carried out to different neurodegenerative illnesses which might be recognized the use of a aggregate of MRI functions and cognitive tests. There are obstacles in imposing a complicated version due to the quantity of the dataset. Even though the character of every characteristic is evident, the levels of every group's test value aren't categorized well. In different words, we ought to have diagnosed extra clearly the variations withinside the variables which may have

performed a function in the result. The expected value using the random forest model is better than the alternative models. It implies there may be a capacity for better prediction rate if we pay extra interest to develop the data cleaning and evaluation process. Moreover, the precise recall rating 1.0 of SVM 1.0. Indicates that the best and accuracy of the category may lower dramatically whilst we use specific dataset. The most important takeaway for us is that there are numerous key elements that are caused by Dementia and we ought to continue to test it and clear the process in different ways. For the further study, it's far vital for us to enhance our knowledge via greater sophisticated EDA method with a bigger pattern size. For instance, we'd strive now no longer most effective the age itself however additionally organization it into generation, or grade extent of brain tissue or examination scores. If the outcomes from this process are reflected withinside the data cleaning method and undoubtedly have an effect on the decision making of the model, the accuracy of the prediction model may be in addition improved.

REFERENCES

1. El-Sappagh, S., Alonso, J.M., Islam, S.M., Sultan, A.M. and Kwak, K.S., 2021. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Scientific reports*, 11(1), pp.1-26.
2. Jain, R., Jain, N., Aggarwal, A. and Hemanth, D.J., 2019. Convolutional neural network based Alzheimer's disease classification from magnetic resonance brain images. *Cognitive Systems Research*, 57, pp.147-159.
3. Kim, H.W., Lee, H.E., Oh, K., Lee, S., Yun, M. and Yoo, S.K., 2020. Multi-slice representational learning of convolutional neural network for Alzheimer's disease classification using positron emission tomography. *BioMedical Engineering OnLine*, 19(1), pp.1-15.
4. Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C. and Buckner, R.L., 2010. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *Journal of cognitive neuroscience*, 22(12), pp.2677-2684.
5. Elaheh Moradi, Antonietta Pepe, Christian Gaser, Heikki Huttunen, Jussi Tohka, Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects, In *NeuroImage*, Volume 104, 2015, Pages 398-412, ISSN 1053-8119, doi.org/10.1016/j.neuroimage.2014.10.002.
6. Venugopalan, J., Tong, L., Hassanzadeh, H.R. and Wang, M.D., 2021. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific reports*, 11(1), pp.1-13.

7. Zhang, Y., Dong, Z., Phillips, P., Wang, S., Ji, G., Yang, J. and Yuan, T.F., 2015. Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning. *Frontiers in computational neuroscience*, 9, p.66.
8. Zoetmulder, R., Konduri, P.R., Obdeijn, I.V., Gavves, E., Išgum, I., Majoie, C.B., Dippel, D.W., Roos, Y.B., Goyal, M., Mitchell, P.J. and Campbell, B.C., 2021. Automated final lesion segmentation in posterior circulation acute ischemic stroke using deep learning. *Diagnostics*, 11(9), p.1621.