Pre-detection of Dementia Using Machine Learning Mechanism

M.S.U Weerasinghe

2021

Pre-Detection of Dementia Using Machine Learning Mechanism

A dissertation submitted for the Degree of Master of Computer Science

M.S.U Weerasinghe
University of Colombo School of Computing
2021

UCSC

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis. This thesis has also not been submitted for any degree in any university previously.

Student Name: M S U Weerasinghe

Registration Number: 2018MCS097

Index Number: 18440972

Shalika.

____

Signature of the Student & Date: 2021 - 11 - 30

This is to certify that this thesis is based on the work of Mr. M S U Weerasinghe under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by,

Supervisor Name: Prof. G. Kapila. A. Dias

____

Signature of the Supervisor & Date: 2021 - 11 - 30

ABSTRACT

Dementia is a neurological disorder that affects millions of people worldwide. Dementia is also not a normal part of aging, and there is no cure or effective treatment for it. The number of persons suffering from dementia is rapidly increasing. According to the 2015 World Alzheimer Report, there are 46.8 million people worldwide who have been diagnosed with dementia, with that figure anticipated to rise to 74.7 million by 2030 and 131.5 million by 2050. The number of people diagnosed with dementia in Sri Lanka is continuously increasing. According to government projections, there are already more than 0.2 million dementia patients, with that number anticipated to climb to 0.5 million by 2050. As a result, dementia has emerged as a serious medical condition that needs to be addressed for the sake of society's well-being.

Dementia is a difficult condition to manage, and it must be dealt with quickly. According to the World Alzheimer's Association, Dementia is one of the most financially costly diseases in the world since there is no proper treatment to cure the disease and the cause of the disease is not identified correctly.

Recently there is a strong interest in machine learning mechanism which provides a better classification accuracy than the conventional classification methods. Based upon the recent studies we design and perform some experiments to investigate the possibility of early diagnosis of dementia from machine learning mechanism using the clinical data.

In this project, we compare machine learning algorithms to clinical data from dementia patients in order to develop a better approach for detecting dementia at an early stage. We primarily use three main advanced machine learning algorithms: SVM, decision tree, and random forest. We train the preprocessed clinical data with the above advance algorithms and takes the highest accurate algorithm after comparing the results along with the confusion matrix of each other. After comparing the confusion matrix results, we choose random forest algorithm as the most accurate machine learning algorithm and it used to trained machine learning model.

In this project we developed a mobile application for the public people and this mobile application is integrated with the machine learning model. So it is very helpful for the general public to identify their day-to-day mental capabilities with this mobile application. This will evaluate the person's dementia level and aid in the early detection of dementia. As a result, this research will be a significant step forward in the prevention of dementia in Sri Lanka.

We expect that this dissertation will help researchers to get better understand about how machine learning can be used in early dementia detection.

ACKNOWLEDGEMENTS

First and foremost, I have to thank my research supervisor, Prof. G. Kapila. A. Dias. This study would not have been accomplished without his guidance and dedicated engagement at every step throughout the year.

I'd also like to express my gratitude to all of the authors of those publications and articles that assisted me in a variety of ways. I'd like to extend my gratitude to my friends, who have been helpful to me in a variety of ways. In addition, I would like to express my gratitude to Mr. Nishan Gunawardena, Mr. Chandima Arangala, Mr. Darshana Thennakoon, and Miss Dilanka Wagachchi for their assistance in making this dissertation a success.

Most importantly, none of this could have happened without my family. My loving mother, father, grandmother, and siblings, without your guidance and help this study is not possible. This thesis stands as evidence of your love and inspiration.

Finally, I'd like to thank the Open Access Series of Imaging Studies (OASIS) and its collaborators for their tireless efforts in sharing their data. Without them, this thesis and the scientific work described here would not have been possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

FTD	Fronto temporal Dementia
CT	Computed tomography
MRI	Magnetic resonance imaging
OASIS	Outcome and Assessment Information Set
PET	Positron Emission Tomography
CSF	Cerebro Spinal Fluid
VD	Vascular Dementia
PD	Parkinson's disease without Dementia
FTD	Front temporal Dementia
AD	Alzheimer's disease
MMSE	Mini-Mental State Exam
SVM	Support vector machine
ASM	Attribute Selection Measure
CDRs	Clinical dementia ratings
FAQ	Frequently Asked Questions
BOMC	Blessed orientation memory concentration
EEG	Electroencephalography
MCI	Mild cognitive impairment
CERAD	Consortium to Establish a Registry for Alzheimer's disease
FCM	Fuzzy c-means clustering algorithm
PNN	probabilistic neural network
ANOVA	Analysis of Variance
SES	Socioeconomic status
ASF	Atlas Scaling Factor feature attribute
eTIV	Estimated Total Intracranial Volume feature attribute
NL	Normal
CART	Classification and regression tree
EDU	Education
ROC	Receiver operating characteristic
EHR	Electronic health record

CHAPTER 1 – INTRODUCTION

1.1 Background and Motivation

Dementia is a neurological disease that affects the brain. It's a catch-all term for memory loss and other cognitive impairments severe enough to interfere with daily life. Furthermore, dementia is not a normal aspect of aging, and there is no known cure or effective treatment for it. Dementia is a term used to describe a set of symptoms that can emerge when a few brain cell groupings stop functioning properly. Dementia is thought to be caused by the normal progression of proteins surrounding neuronal synapses. Amyloid is one of the proteins included, and deposits of it form plaques surrounding neurotransmitters. The other protein is tau, which forms tangles inside neurotransmitters when it is stored (Zhang, et al., 2015).It usually takes place in certain areas of the brain and has an impact on one's ability to think, recall, and communicate. As a result of this harm, brain cells' ability to communicate with one another is disrupted. When brain cells are unable to communicate properly, this could affect one's thinking, behavior, and feelings. It has an impact on a number of cerebrum functions. Memory issues are frequently the first indicators of dementia disease. For example, it could be forgetting the titles of sites and products and overlooking late talks or experiences (Marcus, et al., 2007).

The brain is divided into several unique areas, each of which performs a different function. When cells in a certain area are destroyed, that area is unable to operate normally. Different varieties of dementia are linked to different types of brain cell destruction in specific brain regions. Unfortunately, no one test has yet been developed to identify this condition. Cerebrum examinations alone cannot be used to determine whether or not a person is suffering from it. Dementia, which most commonly affects older persons, is a significant financial and social burden on families and society due to their intellectual incapacity. There is currently no viable medication for dementia that can halt or stop its growth. The importance of focusing on the disease's early stages, quick intervention, and delay cannot be overstated.

The number of people living with dementia is growing very fast. According to the 2015 World Alzheimer Report, the global number of persons diagnosed with dementia is 46.8 million, with an expected increase of 74.7 million by 2030 and 131.5 million by 2050 (Wu, et al., 2017). In Sri Lanka, the number of people diagnosed with dementia is rapidly growing. According to government estimates, there are more than 0.2 million dementia patients, with that number expected to rise to 0.5 million by 2050 (statista.com, n.d.). These figures will increase in the coming years, as the boomer generation reaches the age of 65 and beyond when dementia is most prevalent. As a result, dementia has emerged as a critical disease to be concerned about for the welfare of society.

Early diagnosis of many disorders allows for effective treatment and has a positive impact on the healing process. Early diagnosis of dementia is critical, as therapy is more effective if it is provided as soon as feasible. People with dementia can take charge of their disease, plan for the future, and live well with dementia if they are diagnosed early and have access to the correct services and support. It can assist persons with dementia in gaining access to important information, resources, and support, as well as maximizing their abilities and perhaps benefiting from available drug and non-drug treatments.

So far, the clinician has concluded that a person has dementia based on reports of social propensity, analysis of previous clinical records, and use of neurological examination and cognitive tests. To rule out other causes of dementia, further tests such as hematological, CT, and MRI should be conducted. To detect dysfunctions in human cognitive domains, neuropsychological tests are essential.

In recent years, machine learning algorithms like as support vector machines, independent component analysis, and penalized regression have been used to diagnose dementia. Some of these techniques have been demonstrated to be quite effective in diagnosing dementia from neuroimages, even outperforming professional radiologists in some cases. Machine learning algorithms, for example, have been shown in recent studies to be more accurate than experienced physicians in predicting dementia.

The implementation of a computerized assistance system that can complete the diagnostic process independently of the expert in this procedure saves time and reduces the most common human errors.

1.2 Statement of the problem

The following is the present study's problem statement.

To what extent can a classifier predict the progression of subjects from clinical data to dementia and the progression's corresponding moment and which machine learning algorithm is most suitable for the classifier?

1.3 Research Aims and Objectives

1.3.1 Aim

In the early stages of dementia, obtaining a solid diagnosis remains a challenge. We wanted to create and test a new machine learning-based strategy for assisting in the preliminary diagnosis of dementia using an informant-based questionnaire.

1.3.2 Objectives

The following project tasks have been identified in order to achieve the project objectives.

- Collect the clinical data and pick the data points, extract the relevant information from data, identify the key values from the collected data set.

- Detect the effective feature attributes which impact the diagnosis of dementia from the already diagnosed clinical patients.

- Preprocess the collected data by using data reduction, data discretization, data cleaning, and data transformation.

- Design and create a dataset on dementia disorder, using the collected effective feature attribute values and their weights.

- Create models using the supervised learning classifiers with the collected data. Then pick the most suitable supervised classifier to create the model.

- Test the trained dataset with known data for verification of the implemented model and get the highest prediction rate.

- Design and implement a mobile-based application for healthy persons. This mobile-based application consists of a series of questions designed to test a range of everyday mental skills.

- Integrate the mobile application with the implemented supervised model and evaluate the user inputs. This is used to evaluate the dementia level as a percentage value of a healthy person.

Other than the above mention major objectives we can mention the below points as the alternative objectives of this study.

Not only are accurate dementia classifications helpful to individuals, but they are also crucial for medical doctors and other medical personnel. The manual diagnosis of dementia, which may require multiple pieces of information such as a neuropsychological test score, laboratory research results, informed informant accounts, and so on, is time-consuming in clinical diagnosis. The practitioner's professional-level determines the efficiency and accuracy of the diagnostic. It will be much more difficult to classify and diagnose dementia in several remote areas where professional experts are limited.

In the future, there will be enough resources to pre-detect dementia using machine learning. Also, new researchers will be able to find the advanced techniques, methods, and architecture for the pre-detection of dementia at an early stage. It will be more accurate and more efficient.

Those researchers, on the other hand, will be relying on earlier study, and previous findings will be able to be compared to the most recent findings. It will also aid in the development of the new strategy. When examining the project's long-term goals, this research project has developed a new method for dementia pre-detection using machine learning techniques. It will be useful in future research on machine learning techniques for dementia pre-detection.

1.4 Scope

The objective of this research is to create a dementia pre-detection model. In this initiative, we look at the clinical data of dementia patients who have been diagnosed. Data mining techniques will be used to evaluate and preprocess the data collected.

Machine Learning is an advanced computing technique that can improve medical data processing and make diagnostic decisions automatically. In this project supervised learning classifiers will be used to create a machine learning model and will then pick the most accurate supervised classifier to create the final model. Machine Learning is an advanced computing technique that can improve medical data processing and make diagnostic decisions automatically. In this project supervised learning classifiers will be used to create a machine learning model and will then pick the most accurate supervised classifier to create the final model. With the Covid-19 situation in the world we online available dementia patient's clinical data to build the machine learning model and we mainly use the OASIS dataset for our project.

A mobile-based application will be developed for healthy persons to evaluate their dementia level as a percentage value. This mobile-based application consists of a series of questions designed to test a range of everyday mental skills. Mobile application integrates with the created supervised model a consumable API. This dementia disorder predicting approach will help to evaluate the dementia level of the user and classify the healthy person from dementia diagnosed person.

1.5 Report Outline

- Chapter 2 deals with the background of the thesis and the literature review of the study. It presents relevant information regarding dementia such as risk factors, pathophysiology of dementia, signs and symptoms, diagnosis and treatments. The relevant topics of machine learning and methodologies are addressed in the latter portion of this chapter. Finally, the importance of related research by many scholars that is pertinent to this thesis will be highlighted.

- A full overview of approach will be presented in Chapter 3. The design will be described first. After that, the experiment will be described in detail.

- All of the implementation specifics are found in Chapter 4. It will be introduced the dataset that was utilized in the project. Not only that, but various machine learning technologies, tools, and implementations will be thoroughly examined. Finally, the most relevant code snippets utilized throughout the project will be displayed.

- In Chapter 5, results of each experiment will be illustrated using diagrams and confusion matrices. Afterwards, the results will be discussed.

- The conclusion and future work are included in the final chapter. The conclusions will be presented, as well as some potential paths of study, according to the discussion. Finally, the work that has to be done in the future will be discussed.

CHAPTER 2 - LITERATURE REVIEW

2.1 Introduction

An overview of the study was presented in the previous chapter. It contains the study's background, motivation, scope, and objectives, as well as research questions and a brief description of the methodology. This chapter covers the machine learning approaches such as SVMs, Logistic Regression, and Random Forest that were used during the project and comprehensive overview of dementia. Finally, previous research that is relevant to this study will be discussed.

2.2 Dementia

2.2.1 Introduction

Dementia is a generic word for a loss of capacity to recall, think, or make decisions that interferes with daily tasks. Dementia affects roughly 2% of people over the age of 65, and it can affect up to 35% of those over the age of 85 (Prince, et al., 2015). The number of persons suffering from dementia is rapidly increasing as life expectancy rises. Dementia was projected to affect 26.6 million people in 2006 (Lozano, et al., 2010). By 2050, this population is predicted to exceed 100 million.

Early-onset dementia refers to dementia that develops before the age of 65, whereas late-onset dementia refers to dementia that develops after the age of 65. Dementia is a frequent ailment among adults aged 65 and up. Dementia is a disease that affects 1 out of every 14 persons over the age of 65 and 1 out of every 6 people over the age of 80.In addition, one out of every 20 cases of dementia infection affects people aged 40 to 65. This is referred to as early or youthful beginning dementia ailment.

According to experts, the situation is expected to worsen in the next years. As a result, dementia is predicted to become a growing socioeconomic concern in our societies, in addition to causing a significant psychological burden on patients and their families.

While some of its symptoms may resemble those of advanced aging, it is vital to remember that dementia is not a natural component of the aging process. Dementia symptoms progressively intensify as the disease progresses. There is currently no treatment for dementia that is curative. Rather, the goal is to reduce the disease's course, relieve symptoms, treat behavioral issues, and improve the overall quality of life. An early dementia diagnosis is critical for slowing the progression of dementia.

Dementia's specific cause has yet to be discovered. Neuritic plaques and neurofibrillary tangles, on the other hand, are linked to dementia. The current trend is to discover dementia biomarkers. Recent research has centered on identifying biomarkers that can be used to identify and visualize dementia disease processes. Dementia is diagnosed in clinical practice based on the disease's usual symptoms. As mentioned earlier the only approach to get the absolute confirmation of dementia is a post-mortem brain tissue examination.

2.2.2 Risk Factors

Since the actual origin of dementia is unknown, there are several risk factors to consider when it comes to dementia disease.

- Age

  Age is a significant element that has a close link to dementia. After the age of 65, the likelihood of developing dementia doubles every five years, reaching over 50% by the age of 85 (Alzheimer's Association, 2015).

- Family History

  Another risk factor to consider is your family history. According to recent studies, everyone who has a parent or sibling with dementia is at a higher chance of developing the condition (Alzheimer's Association, 2015).

- Genetics

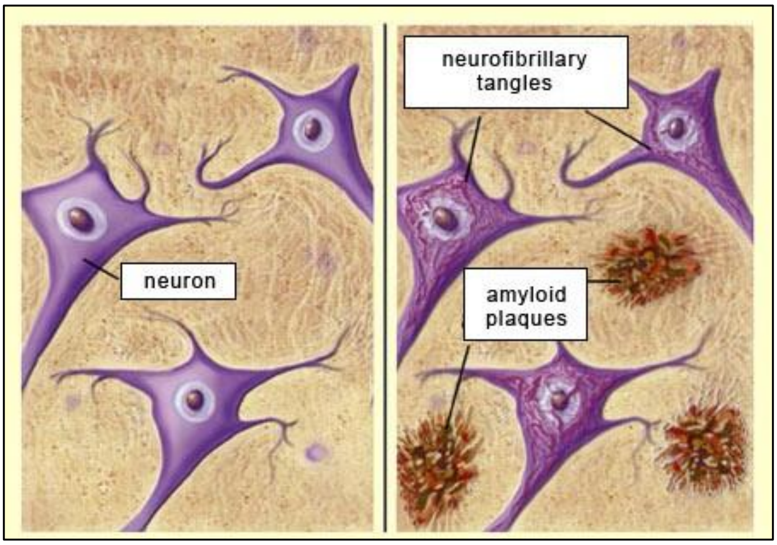
  Dementia is caused by two different types of genes. The two genetic groups are risk genes and deterministic genes. The risk genes make it more likely that you'll get the disease, but there's no guarantee that you'll get it. However, the deterministic genes cause the disease directly, ensuring that anyone who inherits them will get dementia (Alzheimer's Association, 2015).

2.2.3 Pathophysiology

There are over 86 billion neurons in the brain. To build communication networks, each nerve cell connects to a large number of others. Specific functions are assigned to different groups of nerve cells. Some people are involved in the processes of thinking, learning, and remembering. Others assist us in seeing, hearing, and smelling. Brain cells work like miniature factories to carry out their specialized function. Dementia causes parts of a cell's factories to malfunction. They're not clear where the problem began. Backups and malfunctions in one system, like in a real industry, generate issues in other areas as well (Alzheimer's Association, 2015).

The accumulation of incorrectly folded amyloid beta (β-amyloid) protein in the brains of dementia patients has been recognized as a protein misfolding illness (Alzheimer's Association, 2015). The sticky amyloid fragments cluster together and form plaques. These plaques prevent cells from processing messages (i.e. communicating) (Arvesen, 2015). Positron Emission Tomography (PET) and Cerebro Spinal Fluid (CSF) can both be used to measure the protein.

Marilyn S, et al stated Current evidence suggests that markers of amyloid pathology (i.e., CSF and PET) precede evidence of neuronal injury. This does not rule out the possibility that Aβ is the cause of the sickness. It does appear, however, that these various types of biomarkers convey different types of information regarding the progression of the disease in the brain (Albert, et al., 2011).Figure 2.1 shows how protein plaques in the brain of a dementia patient disrupt signals.



Healthy brain Dementia brain

Figure 2.1 - Plaques and tangles in the cerebral cortex in a Healthy brain (left) and a Dementia brain (right) *(alzheimers asspciation, 2021)*

Neuropathology

The loss of neurons and synapses in the cerebral cortex and certain subcortical regions is referred to as dementia. As seen in Figure 2.3, the disease causes a loss of neurons and synapses, resulting in plainly evident changes in brain tissues. The brain's structure becomes irregular as dementia progresses. According to recent studies, it is one of the disease's most sensitive aspects (Laakso, et al., 1996) (Apostolova, et al., 2012). The temporal lobe, parietal lobe, and sections of the frontal cortex and cingulate gyrus are all affected. The human brain's lobes are represented in Figure 2.2.

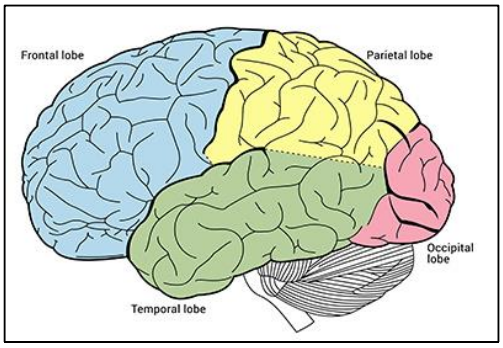
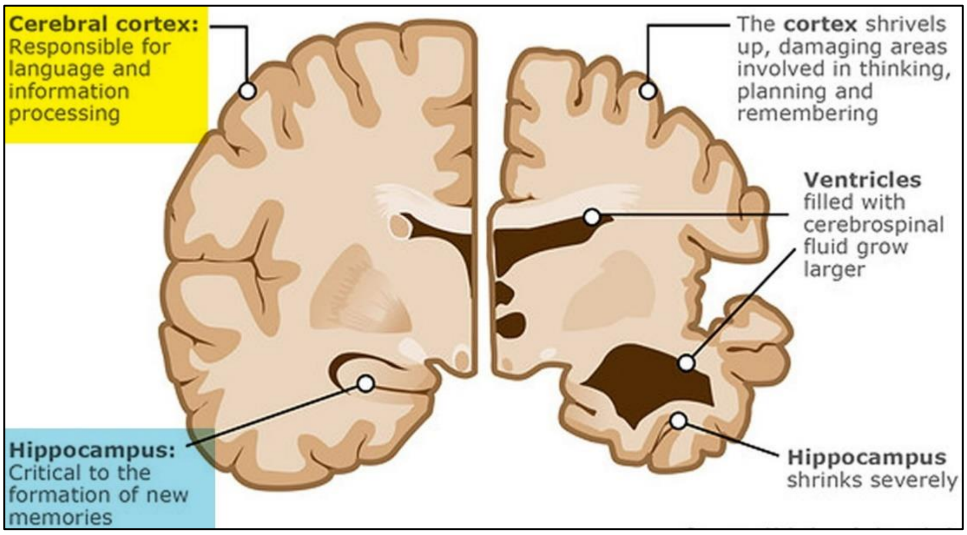


Figure 2.2 - Lobes of the human brain *(The Rag Tree., 2020)*

The study of brain degeneration can be done using a variety of neuroimaging techniques. There is a distinct brain area decline in the brain imaging of people with dementia when compared to the brain images of healthy adults. Early in the course of dementia, the hippocampi of patients with the condition atrophies. When a person's hippocampus shrinks, the ability to generate new memories is lost.

In addition, the brain shrinks, and the ventricles expand. Those parts of the brain that are engaged in thinking, planning, and remembering are harmful (Alzheimer's Association, 2015).

Figure 2.3 shows a graphical representation of a brain that shows the differences between a dementia brain and a healthy brain. Figure 2.4 also depicts an MRI picture of a healthy and a dementia brain.


Cerebral cortex: Responsible for language and information processing
Hippocampus: Critical to the formation of new memories
The **cortex** shrivels up, damaging areas involved in thinking, planning and remembering
Ventricles filled with cerebrospinal fluid grow larger
Hippocampus shrinks severely

Healthy brain Dementia brain

Figure 2.3- Graphical view of a healthy brain (left) and a dementia brain (right) *(Alzheimer's Association, 2015)*

Using a 1.5-T MR imager, Laakso et al. compared hippocampal volumes in 59 patients with mild to moderate Alzheimer's disease, nine patients with Vascular Dementia (VD), 12 patients with idiopathic Parkinson's disease without Dementia (PD), 8 patients with Parkinson's Disease with Dementia (PDD), and 34 elderly control subjects. They discovered that all of the patient groups had significantly smaller hippocampal volumes (on both sides) than the control group, as well as significantly lower absolute volumes (Laakso, et al., 1996). They found that all patient groups had smaller hippocampal volumes (on both sides) than the control group, with absolute volumes in a group of Parkinson's disease and dementia patients being significantly lower.

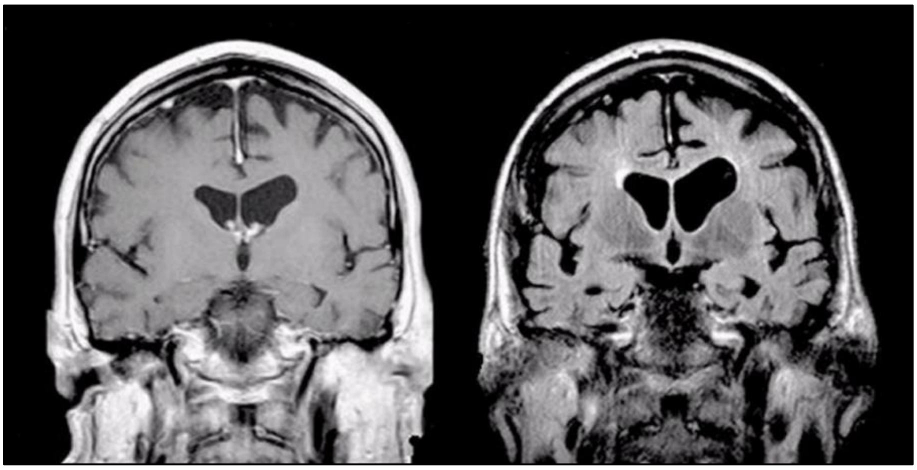


Figure 2.4 - MRI view of Healthy brain (left) and Dementia brain (right)

The researchers arrived at the conclusion that hippocampal atrophy, ventricle enlargement, and cortex shrinkage are sensitive dementia characteristics, but their specificity limits their application in clinical practice.

2.2.4 Signs and symptoms

Dementia symptoms may develop over time. A person with dementia usually lives four to eight years after being diagnosed, although they can live for up to 20 years depending on other circumstances (Alzheimer's Association, 2020). Many years before clinical symptoms appear, changes associated with dementia begin to emerge. Dementia can be classified into four main types based on the symptoms.

Types of dementia

1. Alzheimer’s Disease

Alzheimer's disease is the most common cause of dementia among older persons. It usually takes several years to develop. Early stages may be mistaken for moderate amnesia, which is a natural aspect of growing older. The inability to develop new memories for recent events is frequently one of the first indicators. Finding the right words, figuring out problems, making judgments, judging distance, and locating familiar sites are all difficult jobs. Alzheimer's disease is characterized by the formation of abnormal protein clumps in the brain, which damage nerve cells.

2. Vascular Dementia

Another common type of dementia with a less well-known name is vascular dementia. It can happen unexpectedly, such as after a stroke that affects major blood vessels. A succession of tiny strokes or damage to blood vessels in the brain can also cause it to progress slowly or over time. The signs and symptoms of vascular dementia might be difficult to distinguish.

Symptoms of vascular dementia include,

- Loss of memory.
- Communication and disorientation concerns build up.
- Alterations in the person's walking style.

Depending on which section of the brain is injured, more specific symptoms will appear. Problems with planning, concentrating, or short bursts of severe bewilderment are some of the possibilities.

Vascular dementia is caused by a decrease in blood supply to the brain due to narrowing or blockages in blood arteries, resulting in brain cell destruction. Symptoms of vascular dementia might appear quickly or develop over time.

3. Lewy body Dementia.

Movements may be altered by Lewy body disease, causing people to shuffle while they walk and increase their risk of falling. Some of the symptoms are similar to those seen in people with Parkinson's disease. Lewy body disease patients may have times of extreme bewilderment. They may experience hallucinations, in which they see or hear things that aren't there. Swallowing and sleep patterns might also be disturbed - people may find it easy to fall asleep during the day but have trouble sleeping at night.

There are aberrant protein clumps that develop up in the brain with Lewy bodies. Protein deposits in the brain disrupt nerve cell connections, causing alterations in movement, thinking, behavior, and attentiveness.

4. Front temporal Dementia

This type of dementia can alter one's behavior, personality, and ability to communicate. FTD has been connected to motor neuron disease in some circumstances. Early symptoms of FTD differ depending on which part of the brain is damaged.

- Personality transformations dietary alterations.
- Lack of social awareness.
- Lack of personal awareness.
- In the early stages, difficulty communicating or comprehending others might be a major symptom.

FTD is most frequent in persons between the ages of 45 and 65, but it can also affect people in their later years.

2.2.5 Diagnosis

Dementia is diagnosed by examining at the patient's medical history, behavioral observations, and family history. For a clinical diagnosis of potential or likely AD, these criteria necessitate that the existence of cognitive impairment and a suspected dementia syndrome be validated by neuropsychological testing (Stonnington, et al., 2008). However, diagnosing dementia necessitates a complete medical examination, which includes the tests listed below.

- Medical history.
- Mental status testing.
- Physical and neurological testing.
- Tests such as blood tests and brain imaging.

Additionally, neuropsychological tests such as the Mini-Mental State Examination (MMSE) are frequently used to determine the cognitive abnormalities needed for disease diagnosis. Doctors use brain scans like Magnetic Resonance Imaging, Computed Tomography (CT), and Positron Emission Tomography to rule out other possible causes of the disease- (Farina, et al., 2012).Recently, experts have been concentrating their efforts on the early detection of dementia. For a variety of reasons, an early and precise diagnosis is advantageous.
Even if the underlying dementia process cannot be stopped or reversed, starting treatment early in the disease process may help to preserve everyday functioning for some time.

Furthermore, obtaining an early diagnosis will benefit patients and their families by providing multiple opportunities, such as those listed below (National Institute on Aging, 2021).

- Planning for the future
- Caring of financial and legal matters
- Potential safety issues
- Making living arrangements
- Supporting networks

2.2.6 Treatment

For the time being, there is no cure for dementia disease. Researchers look forward to inventing better dementia treatments in the near future. The various treatments can help to decrease the disease's progression and manage its symptoms. When it comes to severe dementia, however, slowing the progression is challenging. As a result, the treatments will be more effective if they are implemented as soon as feasible before irreversible brain damage occurs (National Institute on Aging, 2021). Symptoms such as cognitive and psychological disorders, as well as behavioral issues, are the focus of current treatments. In some circumstances, the subject's environment may also change to provide him or her with a better background.

2.3 Machine Learning

2.3.1 Introduction

Machine learning is a major subfield of computer science that allows computers to learn without having to be explicitly programmed. Machine learning is important in fields where programming algorithms lack the potential to solve issues, such as artificial intelligence, computer vision, data mining, and data science. Machine learning can be divide into three broad categories which are,

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Supervised learning is a type of machine learning that uses labeled data. Each input has a corresponding output value. Create a model that can predict the responses of a new dataset using this supervised learning method. The test data collection is usually used to validate the model. In the supervised learning method, a larger training dataset is required to obtain a powerful predictive model. There are two categories of supervised learning which are classification and regression. For categorical response values, classification is employed, while for continuous response values, regression is used.

Support Vector Machines, Random Forest, Logistic regression, Naive Bayes classifier, Decision trees, and linear regression are examples of supervised learning algorithms. Supervised learning methods are widely used in a variety of applications, including financial credit scoring, biological applications, and voice and visual pattern recognition.

Unsupervised learning, on the other hand, has no known output. There is no error because the input is unlabeled. Cluster analysis is the most often used unsupervised learning technique. The similarity is used to model the clusters. Hierarchical clustering, k-Means clustering, Gaussian mixture models, and Hidden Markov models are examples of unsupervised learning approaches. Unsupervised learning approaches such as sequence analysis and genetic grouping are employed in bioinformatics.

Reinforcement Learning helps computers and software agents to automatically determine normal behavior in a given scenario to maximize their performance (Moni, 2021). In practice, Reinforcement Learning is used in a wide range of computer science applications, including directing robotic arms to identify the most efficient motor combination and playing logic games.

2.3.2 Machine Learning Algorithms

An algorithm is a mechanism used to construct a machine learning model from data in machine learning. Machine learning algorithms are used to recognize patterns. Algorithms are either trained on data or fitted to a collection of data. (Brownlee, 2019).

There are several characteristics of machine learning algorithms.

- Machine learning algorithms can be described using math and pseudo code.
- The efficiency of machine learning algorithms can be analyzed and described.

There are a variety of machine learning algorithms available. The machine learning model in this research project was developed using the algorithms mentioned below.

- Support vector machine algorithm
- Random forest algorithm
- Decision tree algorithm

Support Vector Machine Algorithm (SVM)

Support Vector Machines are supervised learning models with related learning algorithms that analyze data and classify it based on the findings. Both classification and regression problems can be solved using the support vector machine approach. It is, however, mostly used to solve classification problems.

SVM uses a hyper plane to divide data into categories. In other words, when given labelled training data, SVM generates an ideal hyper plane for categorizing testing data (opencv.org, 2021). Figure 2.5 shows that there are multiple lines that divide a set of two-dimensional points into two classes.

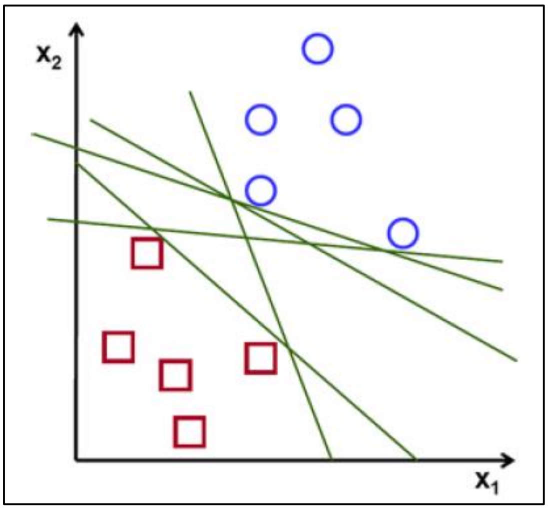


Figure 2.5 - Multiple line separation of SVM *(opencv.org, 2021)*

The SVM method, on the other hand, identifies the hyper plane with the shortest minimum distance between the training examples. As a result, the best separating hyper plane maximizes the training data margin. (Figure – 2.6)

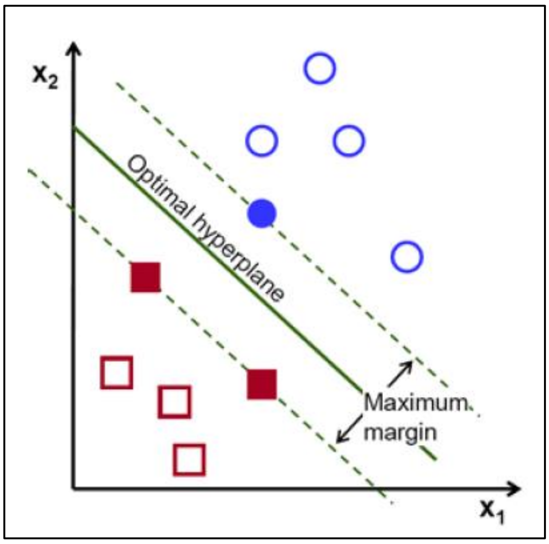


Figure 2.6 - Optimal hyper plane of the SVM ***(opencv.org, 2021)***

The purpose of the SVM algorithm is to maximize the margin. It will choose the hyper plane with the least value from all feasible hyper planes that match the conditions. Because it is the one with the largest profit margin.

Decision Tree Algorithm

A decision tree is a supervised learning technique that can be used to solve classification and regression problems, however it is most often used to address classification problems. In this tree-structured classifier, internal nodes represent dataset properties, branches represent decision rules, and each leaf node delivers the conclusion.
A Decision tree's two nodes are the Decision Node and the Leaf Node. Decision nodes are used to make any decision and have numerous branches, whereas Leaf nodes are the output of such decisions and do not have any further branches.

The features of the given dataset are used to make judgments or run tests. It's a diagram that shows how to find all possible answers to an issue or make a choice based on specified factors. It's called a decision tree because it begins with the root node and grows into a tree-like structure with additional branches, much like a tree. A decision tree asks a question and then divides the tree into sub trees based on the answer (Yes/No) (javatpoint.com, 2021). The general structure of a decision tree is shown in Figure 2.7.

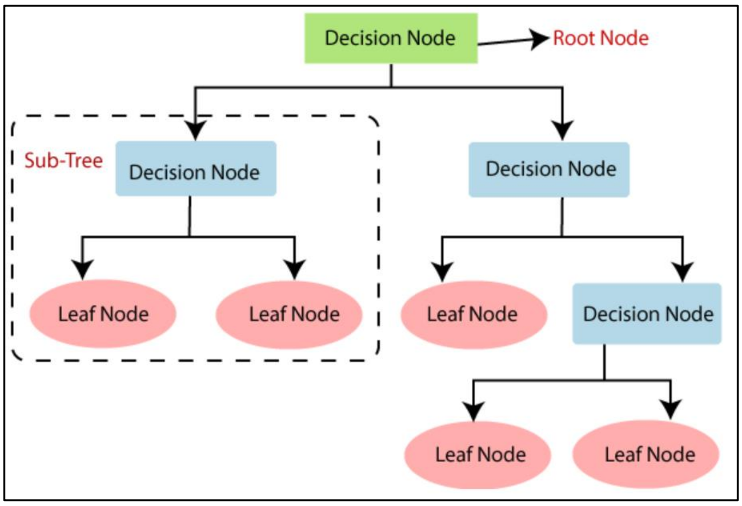


Figure 2.7 - General structure of a decision tree ***(javatpoint.com, 2021)***

In a decision tree, the mechanism for deciding the class of a given dataset begins at the root node. Based on the comparison, this algorithm compares the values of the root property to the values of the record (actual dataset) attribute, then follows the branch and leaps to the next node. The method compares the value of the property with the values of the other sub-nodes before moving on to the next node. It continues in this manner until it reaches the leaf node of the tree. (javatpoint.com, 2021).

The most difficult aspect of creating a Decision tree is deciding which attribute is ideal for the root node and sub-nodes. To overcome such circumstances, a technique known as Attribute Selection Measure, or ASM, might be applied. Using this measurement, we can simply find the best characteristic for the tree's nodes. There are two common ASM approaches to calculate the Gain Ratio.

- Information Gain
- Gini Index

The assessment of changes in entropy after segmenting a dataset based on an attribute is known as information gain. It establishes how much information a feature provides about a class. It divided the node and constructed a decision tree based on the value of the information gathered. In a decision tree approach, the node/attribute with the highest information gain is divided first, with the goal of maximizing the value of information gain. (javatpoint.com, 2021).

Information Gain= Entropy(S) - [(Weighted Average) * Entropy (each feature)

Entropy is a metric for determining the degree of impurity in a particular property. It denotes the randomness of data. Entropy can be calculated as below (javatpoint.com, 2021).

S = Total number of samples
P (yes) = Probability of yes
P (no) = Probability of no

Entropy(s) = -P (yes) * log2 P (yes) - P (no) * log2 P (no)

The Gini index is a measure of impurity or purity used while creating a decision tree. In comparison to a high Gini index, an attribute with a low Gini index should be preferred. The classification and regression tree (CART) algorithm uses the Gini index to construct binary divides, and it exclusively creates binary splits. The following formula can be used to compute the Gini index.

P_i = Probability of an object being classified to a particular class.

Gini Index $= 1 - \sum_{i=1}^{n} (P_i)^2$

Random Forest Algorithm

The supervised learning method is used by Random Forest, a well-known machine learning algorithm. It can be used for both classification and regression problems in machine learning. It's based on supervised methods, which is a technique for combining multiple classifiers to tackle a complex problem and improve the model's performance (javatpoint.com, 2021).

Random Forest is a classifier that averages the results of several decision trees applied to distinct subsets of a dataset to improve the dataset's projected accuracy. The random forest collects forecasts from each tree and predicts the final output based on the majority votes of predictions, rather than relying on a single decision tree (javatpoint.com, 2021). The random forest approach is explained in detail in Figure 2.8.

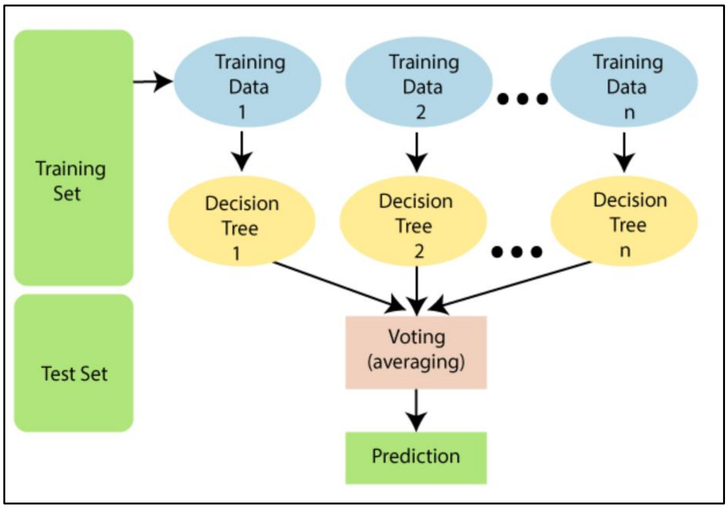


Figure 2.8 - Steps of the random forest algorithm *(javatpoint.com, 2021)*

There are two stages to generating a random forest. The first step is to build the random forest by combining N decision trees, and the second step is to make predictions for each tree created in the first phase.

2.3.3 Machine Learning Model

A computer software that has been trained to recognize specific patterns is known as a machine learning model. Machine learning involves training a model on a set of data and then providing it with an algorithm to reason about and learn from that data. A model is a representation of what a machine learning algorithm has learned. After running a machine learning algorithm on training data, the machine learning model represents the rules, numbers, and other algorithm-specific data structures needed to make predictions (https://docs.microsoft.com/, 2021).

Machine learning model = Model Data + Prediction Algorithm

Machine learning models are created to extract insights from data that may subsequently be used to make better decisions. Based on training data, algorithmic models tell you which outcome is most likely to hold for the target variable. They create a representation of the relationships and pick out patterns between all of the dataset's distinct attributes that may be applied to comparable data in the future, allowing them to make decisions based on those patterns and correlations (https://www.datarobot.com/, n.d.).

2.4 Related Work

2.4.1 Introduction

Many investigations on an automatic diagnosis of dementia disease utilizing various methodologies have been conducted in recent years. Several pattern classifiers have been tried for subject discrimination utilizing various clinical data.

In those recent investigations, various feature extraction and classification algorithms were applied. This section will provide a quick overview of related research in this field.

2.4.2 Related Work Approaches

Data mining has been utilized by many researchers to diagnose a variety of ailments. Jyothi and Sony (2011) have used classifiers to predict cardiac diseases, such as naive Bayes, k-nearest neighbor, and decision tree (Soni, 2011).

Williams (2013) used support vector machines (SVM), decision trees, and neural networks to record clinical dementia ratings (CDRs), and naive Bayes replaced missing values with the average value to get the best accuracy and correlation (Williams, J.A.; Weakley, A.; Cook, D.J.; Schmitter-Edgecombe, M., 2013).

Voxel-based morphometry applied to MRI images from an oasis medical dataset was presented by Chyzhyk and Savio (2010) (Chyzhyk & Savio, 2010).

Bhagya and Sheshadri (2014) compared numerous classifiers in the detection of Alzheimer's disease, including naive Bayes, J48 decision tree method, random forest, and JRip. The results revealed that naive Bayes, Jrip, and random forest outperformed other algorithms. The problem with this article was that it used a data set with 250 participants that had not been preprocessed (Shree & Sheshadri, 2014).

In a study of several statistical and machine learning methods, Chen and Herskovits (2010) found that a Bayesian network classifier and an SVM performed best in assessing participants with little or no dementia (Chen, R.; Herskovits, E.H, 2010).

Using the Alzheimer's disease neuroimaging initiative dataset, Tohka and his colleagues examined several feature selection algorithms using machine learning for dementia in anatomical brain MRI. SVM and logistic regression are used as classifiers, and it is shown that after using feature selection procedures, reducing age improves the average accuracy of all classifiers (Tohka, J; Moradi, E; Huttunen, H, 2016).

Shankle and his team used C4.5 rules, C4.5, naive bayes to analyze data collected from two simple cognitive and functional skills tests, FAQ and blessed orientation memory concentration (BOMC) a cognitive assessment, in order to improve dementia screening (Shankle, et al., 1996).

Joshi, Shenoy, Venugopal, and Patnaik (2009) employed machine learning and neural network techniques to improve the accuracy of current dementia screening tools including the MMSE and the Functional Activities Questionnaire. The findings revealed that accuracy can be increased by combining both tests with machine learning (Joshi, S.; Shenoy, P.D.; Venugopal, K.R.; Patnaik, L.M., 2009).

Trambaiolli and his colleagues previously used electroencephalography (EEG) data to differentiate between those with normal cognition and those with Alzheimer's or MCI, using the SVM algorithm to learn the EEG pattern of Alzheimer's patients. As a conclusion, EEG Epochs had a high level of accuracy (79.9%), while SVM had an accuracy of approximately 87 percent (Trambaiolli, L.R.; Lorena, A.C.; Fraga, F.J.; Kanda, P.A.; Anghinah, R.; Nitrini, R, 2011).

Cho and Chen proposed a hierarchical double-layer system for early dementia diagnosis. This model predicts an early diagnosis of dementia using a Bayesian network in the top layer following diagnostic prediction with the FCM and PNN algorithm in the base layer when a cognitive test such as the MMSE or CERAD is done. FCM and PNN accuracy in this model were 74 percent and 69 percent, respectively (Cho & Chen, 2012).

Vemuriet and his team have created a technique for diagnosing Alzheimer's disease using SVM on structural MRI data (Vemuri, et al., 2008). Kloppel and his team (2008) employed SVM classification to diagnose AD using real-world MRI images (Kloppel, et al., 2008).Hoosh and his team proposed a strategy for dividing MMSE-KC data into normal and abnormal categories, and CERAD-K is used to categorize mild cognitive impairment and dementia (Hooshyar, et al., 2008).

2.4.3 Comparison between related projects and proposed study

Following table 2.1 represents the comparison between existing projects and proposed project.

Table 2.1 - Comparison between related projects and proposed project

Project	Related project feature details	Proposed project
Trambaiolli and Team (2011)	In related projects, they have used not only dementia patient's clinical data. They use other data such as MRI and EEG data.	This proposed project uses 450 clinical data records of dementia and non-dementia persons in the oasis data dictionary.

Williams, Weakley and Team (2013)	In related projects, they have used not only machine learning. They have used neural networks, deep learning, and other approaches.	In this proposed project we only use machine learning to train the supervised model.

2.4.4 Research gap between related projects and proposed study

Following table 2.2 represents the research gap between existing projects and proposed study.

Table 2.2 - Research gap between related projects and proposed project

Study	Findings	Research Gap
Trambaiolli, Lorena and team study (2011) (Trambaiolli, L.R.; Lorena, A.C.; Fraga, F.J.; Kanda, P.A.; Anghinah, R.; Nitrini, R, 2011)	This study carried out through the MRI and EEG clinical data of dementia patients.	This proposed project uses 450 clinical data records of dementia and non-dementia persons in the oasis data dictionary.
Tohka and team study (2016) (Tohka, J; Moradi, E; Huttunen, H, 2016)	This study carried out through the neuroimaging Initiative dataset.	
Joshi, Shenoy and team (2009) (Joshi, S.; Shenoy, P.D.; Venugopal, K.R.; Patnaik, L.M., 2009)	This study carried out through the neural network and machine learning approach.	In this proposed project we only use machine learning to train the supervised model. It will help this study get most accurate conclusions without any dependencies.
Williams, Weakley and team (2013) (Williams, J.A.; Weakley, A.; Cook, D.J.; Schmitter-Edgecombe, M., 2013)	This study carried out through neural networks, deep learning, and other approaches.	
Chen and team (2010) (Chen, R.; Herskovits, E.H, 2010)	This study carried out SVM and Bayesian-network classifier to implement the machine learning model.	In this proposed project we use SVM, Random forest and decision tree classifier and get the most accurate outcome to build the model.

Noria and team (2019) (Noria, et al., 2019)	This study carried out using the EHR data of the demented patients.	To train the machine learning model in this work, we used data from both demented and non-demented people. It is quite beneficial to supervised models in terms of obtaining highly accurate outputs.

2.5 Summary

The background study is described in full in this chapter. It covers dementia, the technology that underlying machine learning, and several machine learning algorithms. Furthermore, this chapter discussed the various ways used by recent studies to develop an automatic dementia disease diagnosis. The research methodology and several pre-processing methods used during the project will be described in the next chapter.

CHAPTER 3 – METHODOLOGY

3.1 Introduction

This chapter will serve as a foundation for the different methods that will be used throughout the project. In addition, this chapter will go through the research technique, study design, and data pre-processing procedures in depth.

3.2 Research Method

There were two phases to the study in this thesis. To begin, we gathered, preprocessed, and verified the data for this study. Second, for pre-detection, a machine learning-based technique will be applied. Different experiments were used to examine those two periods. As a result, this research takes an experimental approach. Three experiments with three distinct machine learning algorithms were carried out to answer the research study in order to produce the most accurate machine learning model.

3.3 Clinical Data Collection

For this study, we need to collect clinical data. As a result, we should collect clinical data from dementia patients at the beginning of the study by visiting dementia diagnostic patient care centers. We should provide clinical patients a questionnaire and gather the necessary supporting data.

Unfortunately, due to the pandemic situation, collecting data from clinical patients was not feasible. As a result, we gather information from publicly available data sets that have already been used in research projects. So, after reviewing the publicly available data sets, I decided to complete this study using the OASIS Data Dictionary.

I use clinical data from young, middle-aged, non-demented, and demented older persons from the OASIS Data Dictionary, which is freely available. Some of the collected data is invalid, partial, or noisy. As a result, we should use data mining techniques to preprocess the obtained data at the beginning of the project.

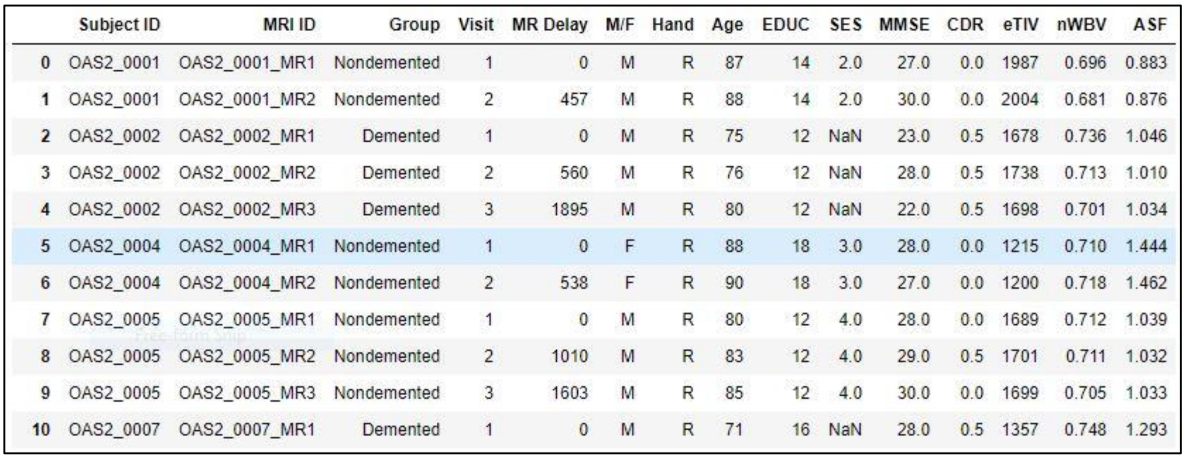
3.3.1 Clinical Data Preprocess

Finding incomplete, erroneous, inaccurate, or unnecessary bits of data and modifying, updating, or eliminating them from a record set, table, or database is known as data preparation.

The acquired data includes data from patients with hearing and vision impairments, as well as data loss due to errors or omissions in the data collecting technique, as well as data loss owing to data loss due to errors or errors in the data collection procedure. Because only a few machine learning algorithms overlook missing values during data training, and most algorithms are impacted by such gaps, we may remove missing values and mistakes using data preparation.

Data normalization combines values from diverse ranges into a single range, preventing an attribute with a larger range of values from having a higher weight than one with a smaller range. During the data preprocessing phase, any missing or incorrectly entered data is eliminated. Due to differences in the data range of each feature, normalization is also performed to set the range of data to 0-1. (which may impair machine learning algorithms).

Figure 3.1 depicts the data set after the original clinical data was preprocessed.



	Subject ID	MRI ID	Group	Visit	MR Delay	M/F	Hand	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
0	OAS2_0001	OAS2_0001_MR1	Nondemented	1	0	M	R	87	14	2.0	27.0	0.0	1987	0.696	0.883
1	OAS2_0001	OAS2_0001_MR2	Nondemented	2	457	M	R	88	14	2.0	30.0	0.0	2004	0.681	0.876
2	OAS2_0002	OAS2_0002_MR1	Demented	1	0	M	R	75	12	NaN	23.0	0.5	1678	0.736	1.046
3	OAS2_0002	OAS2_0002_MR2	Demented	2	560	M	R	76	12	NaN	28.0	0.5	1738	0.713	1.010
4	OAS2_0002	OAS2_0002_MR3	Demented	3	1895	M	R	80	12	NaN	22.0	0.5	1698	0.701	1.034
5	OAS2_0004	OAS2_0004_MR1	Nondemented	1	0	F	R	88	18	3.0	28.0	0.0	1215	0.710	1.444
6	OAS2_0004	OAS2_0004_MR2	Nondemented	2	538	F	R	90	18	3.0	27.0	0.0	1200	0.718	1.462
7	OAS2_0005	OAS2_0005_MR1	Nondemented	1	0	M	R	80	12	4.0	28.0	0.0	1689	0.712	1.039
8	OAS2_0005	OAS2_0005_MR2	Nondemented	2	1010	M	R	83	12	4.0	29.0	0.5	1701	0.711	1.032
9	OAS2_0005	OAS2_0005_MR3	Nondemented	3	1603	M	R	85	12	4.0	30.0	0.0	1699	0.705	1.033
10	OAS2_0007	OAS2_0007_MR1	Demented	1	0	M	R	71	16	NaN	28.0	0.5	1357	0.748	1.293

Figure 3.1 - Preprocessed clinical data set

We must choose the attributes of the obtained data for the machine learning model according to their weight during data preparation. For each featured property, we utilize the scientific approach ANOVA test to pick features.

3.4 Feature Selection

3.4.1 ANOVA Test

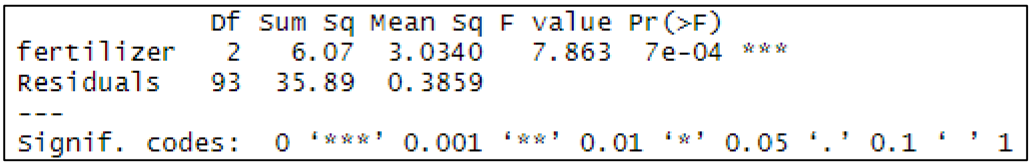
ANOVA is a statistical test for determining how one or more categorical independent variables influence changes in a quantitative dependent variable. The ANOVA method determines if the means of the groups differ at each level of the independent variable. The null hypothesis (H0) of an ANOVA is that there is no difference in means, whereas the alternate hypothesis (Ha) is that there is a difference in means.

To choose the weighted feature attributes from a dataset, the ANOVA test is used. ANOVA examines whether any of the group means differ from the overall mean of the data by comparing the variance of each group to the overall variance of the data. If one or more groups differ from the null hypothesis' expected range of variance, the test is statistically significant.

One-way ANOVA and Two-way ANOVA are the two most common types of ANOVA testing. To determine the weighted features in the clinical data set, we used a one-way ANOVA technique in this study.

3.4.2 Results of a ANOVA Test

In this sub section will describe the results of an ANOVA test. Below Figure 2.5 shows the sample result output of the one-way ANOVA test. The rest values in the output result are described further below.



```
             Df Sum Sq Mean Sq F value Pr(>F)    
fertilizer    2   6.07  3.0340   7.863  7e-04 ***
Residuals    93  35.89  0.3859                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3.2 - Sample ANOVA test output result

- The **Df** column shows the independent variable's degrees of freedom as well as the residuals' degrees of freedom.

- The sum of squares is displayed in the Sum **Sq** column (The total variation between the group means and the overall mean).

- The mean of the sum of squares is computed by dividing the sum of squares by the degrees of freedom for each parameter in the **Mean Sq** column.

- The F test statistic is shown in the F-value column. This is obtained by dividing the mean square of each independent variable by the mean square of the residuals. The higher the F value, the more likely the independent variable's variation is real rather than coincidental.

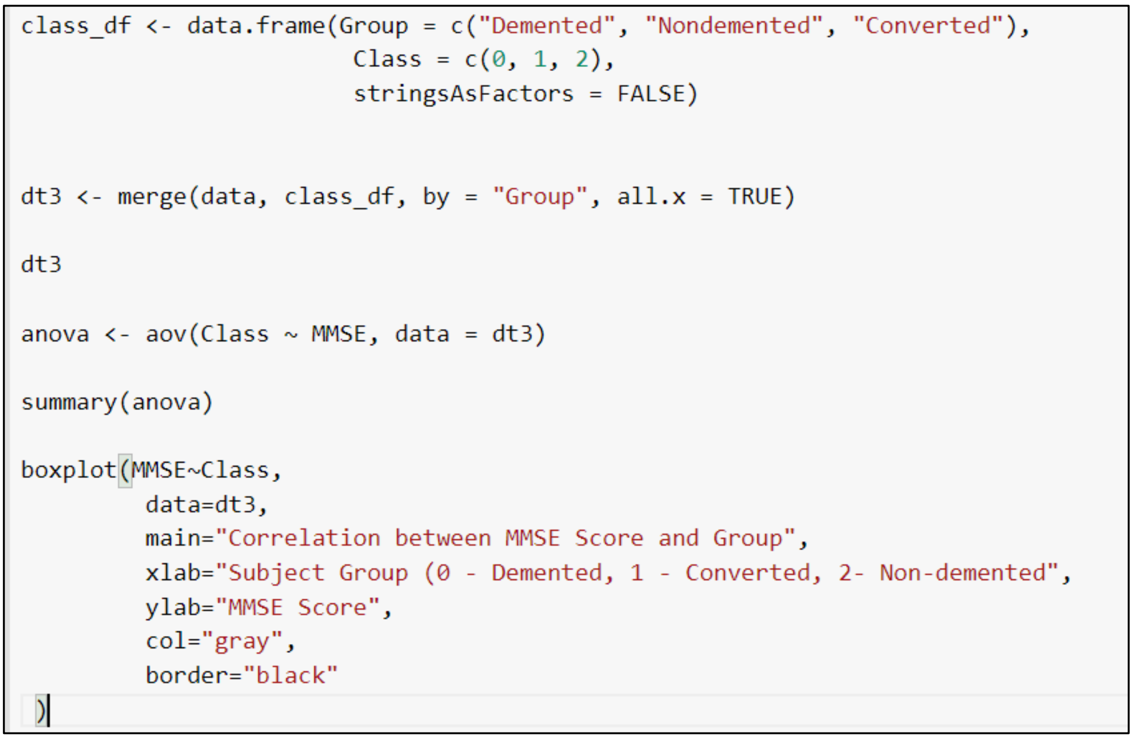
- The Pr (>F) column contains the p-value. Statistic’s If the null hypothesis of no difference in group means had been true, the F-value obtained from the test would have been more likely.

3.4.3 ANOVA Test On Feature Selection of Clinical Data

The most difficult part of machine learning is deciding which features to use to train the model. Only the qualities that are substantially reliant on the response variable are required. ANOVA assists us in completing our task of identifying the greatest attributes.

The aov() function in R can be used to do an ANOVA. This will compute the ANOVA test statistic and determine if there is significant variation among the groups established by the independent variable's levels. After applying the ANOVA test for the clinical data set it gives the below feature attributes as the most weighted feature attributes for the study. The statistic output of the ANOVA test will be more described in the evaluation and result section.

The ANOVA test function is shown in Figure 3.3, and it is used to compute the feature weight of the MMSE (Mini mental state exam) attribute feature. According to the ANOVA test, the following feature attributes are the machine learning model's higher weighted feature attributes. They are described in the Table – 3.1.



```
class_df <- data.frame(Group = c("Demented", "Nondemented", "Converted"),
                        Class = c(0, 1, 2),
                        stringsAsFactors = FALSE)


dt3 <- merge(data, class_df, by = "Group", all.x = TRUE)

dt3

anova <- aov(Class ~ MMSE, data = dt3)

summary(anova)

boxplot(MMSE~Class,
         data=dt3,
         main="Correlation between MMSE Score and Group",
         xlab="Subject Group (0 - Demented, 1 - Converted, 2- Non-demented",
         ylab="MMSE Score",
         col="gray",
         border="black"
 )
```

Figure 3.3 - ANOVA Test calculation on MMSE feature attribute

Table 3.1 - ANOVA Test Weighted Feature Attribute List

Feature Attribute	**Description**
M / F	Gender
SES	Socioeconomic status as assessed by the Hollingshead Index of Social Position and classified into categories from 1 (highest status) to 5 (lowest status)
Age	Age in years
Education	Education Level
MMSE	Mini Mental State Examination (range is from 0 = worst to 30 = best)

3.5 Supervised Machine Learning Model

In this paper, we describe a model that uses a machine learning technique to learn data and label it as normal or dementia. The proposed model is a two-level hierarchical model, similar to the dementia diagnosis approach used at the dementia support center. After data preprocessing, normalization, and feature selection, machine learning techniques are used to categorize normal persons and dementia-diagnosed people.

In supervised learning, instances are given known labels for the appropriate outputs, whereas in unsupervised learning, instances are not given labels. Many machine learning applications require supervised activities. As a result, in this research, we will primarily focus on the labeling strategies that have been used.

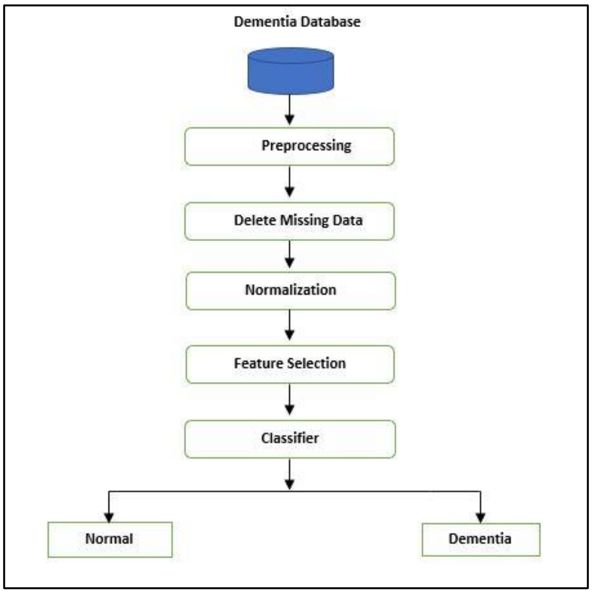


Figure 3.4 - Machine learning model implementation approach

3.5.1 Classification

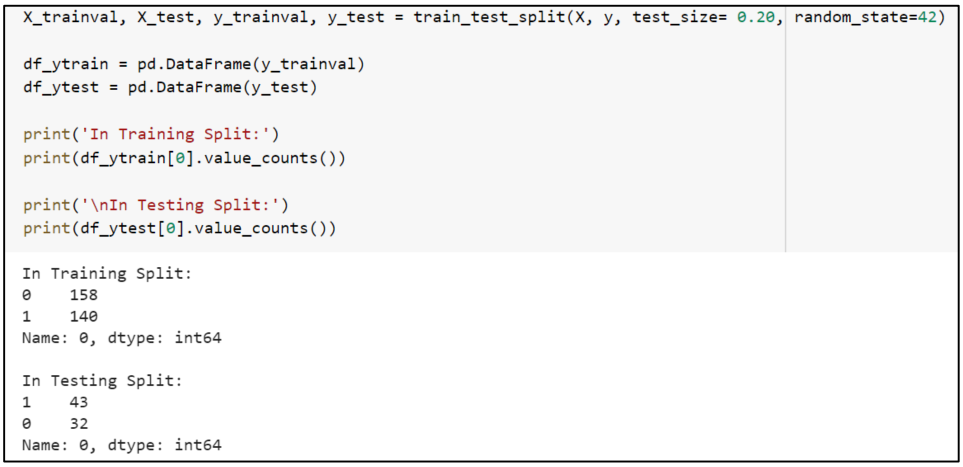
In this study, we mainly use three classification techniques to classify the preprocessed clinical data set. We use these three classification techniques as three experiments in the study. SVM, decision tree, and Random forest, machine learning algorithms are used to classify the preprocessed clinical data set. We use these mentioned machine learning algorithms as three main experiments in the study. After analyzing the results we are going to select the most suitable classification technique to implement the machine learning model.

3.6 Experiment

We must choose the most appropriate machine learning algorithm to implement the machine learning model in this study. So, to obtain the most accurate result, we employ the same clinical data and runtime environment. These three experiments are based on three main machine learning algorithms. The three experiments' scientific approach is described below.

3.6.1 Experiment 1 – Evaluation with SVM algorithm

In experiment 1, the preprocessed clinical data is trained using the SVM method. We calculate the confusion matrix values after training the clinical data set to find the best training algorithm. This preprocessed clinical data set contains 373 patient clinical data records, and for each of the three experiments mentioned in this paper, we used an 80/20 training and testing data ratio.

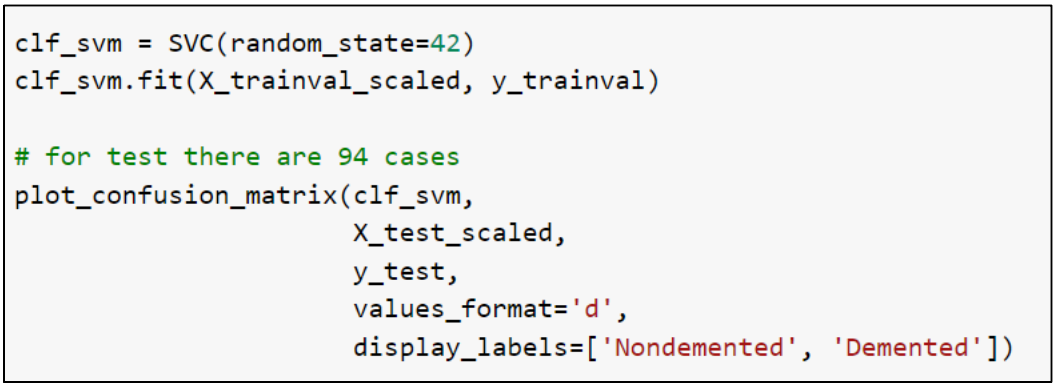


```
X_trainval, X_test, y_trainval, y_test = train_test_split(X, y, test_size= 0.20, random_state=42)

df_ytrain = pd.DataFrame(y_trainval)
df_ytest = pd.DataFrame(y_test)

print('In Training Split:')
print(df_ytrain[0].value_counts())

print('\nIn Testing Split:')
print(df_ytest[0].value_counts())
```

```
In Training Split:
0    158
1    140
Name: 0, dtype: int64

In Testing Split:
1    43
0    32
Name: 0, dtype: int64
```

Figure 3.5 - Train and test ratio of clinical data

First of all the data set needed to be separated into two sets as the training set and the testing set. Two data sets were loaded into the memory. Then clinical data were further processed before classifying using SVM. The figure 3.6 shows the method source code that use to train the clinical data set using SVM algorithm.



```
clf_svm = SVC(random_state=42)
clf_svm.fit(X_trainval_scaled, y_trainval)

# for test there are 94 cases
plot_confusion_matrix(clf_svm,
                      X_test_scaled,
                      y_test,
                      values_format='d',
                      display_labels=['Nondemented', 'Demented'])
```

Figure 3.6 - Method used to train the clinical data set using SVM algorithm

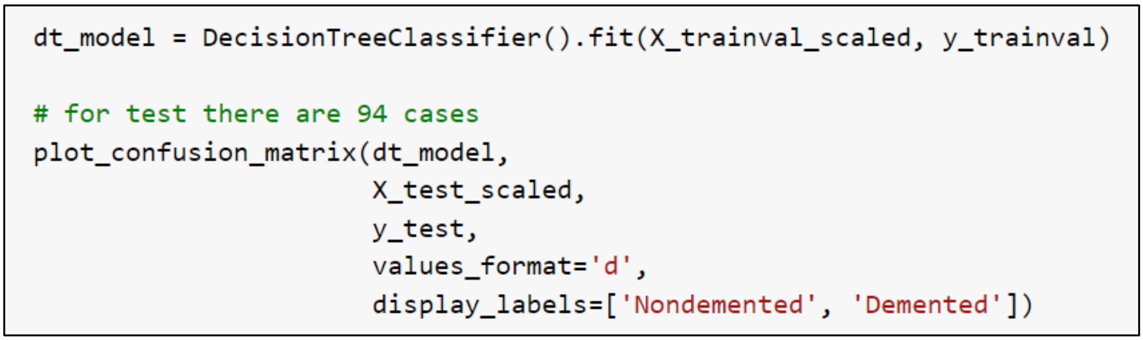
After training the data set with SVM algorithm we calculate the confusion matrix to compare with the confusion matrix output of other machine learning algorithms.

3.6.2 Experiment 2 – Evaluation with Decision tree algorithm

In experiment 2, the preprocessed clinical data is trained using the decision tree algorithm technique. We calculate the confusion matrix values after training the clinical data set to find the best training algorithm as in the experiment 1. As mentioned in the experiment 1 for each experiment we use used an 80/20 training and testing data ratio.

As in the experiment 1, First of all the data set needed to be separated into two sets as the training set and the testing set. Two data sets were loaded into the memory.

Then clinical data were further processed before classifying using decision tree algorithm. The figure 3.7 shows the method source code that use to train the clinical data set using decision tree algorithm.



```
dt_model = DecisionTreeClassifier().fit(X_trainval_scaled, y_trainval)

# for test there are 94 cases
plot_confusion_matrix(dt_model,
                      X_test_scaled,
                      y_test,
                      values_format='d',
                      display_labels=['Nondemented', 'Demented'])
```

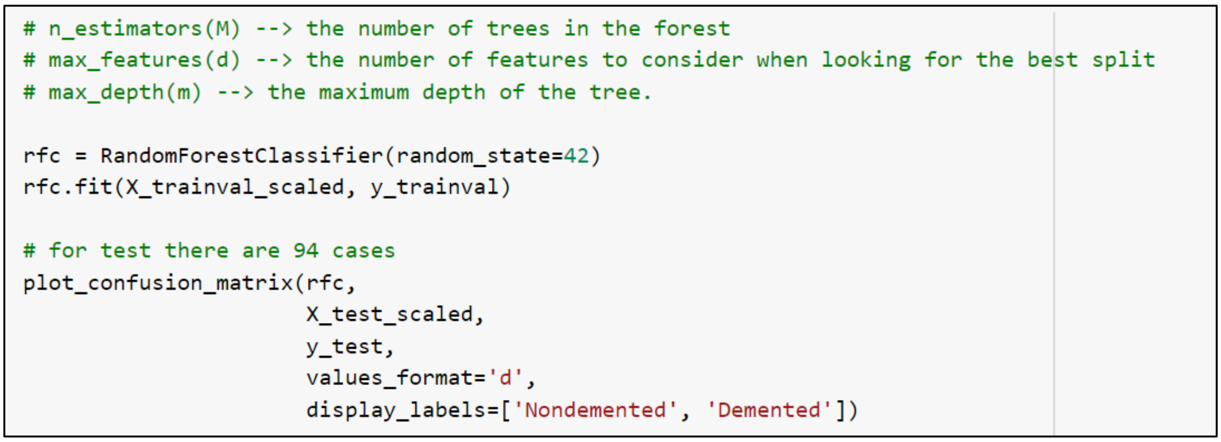
Figure 3.7 - Method used to train the clinical data set using decision tree algorithm

As in the experiment 1, after training the data set with decision tree algorithm we calculate the confusion matrix to compare with the confusion matrix output of other machine learning algorithms.

3.6.3 Experiment 3 – Evaluation with Random forest algorithm

In experiment 3, the preprocessed clinical data is trained using the random forest algorithm technique. We calculate the confusion matrix values after training the clinical data set to find the best training algorithm as in the above two experiments. As mentioned in the above experiments, for experiment 3 we use used an 80/20 training and testing data ratio.

As in the experiment 1, First of all the data set needed to be separated into two sets as the training set and the testing set. Two data sets were loaded into the memory. Then clinical data were further processed before classifying using random forest algorithm. The figure 3.8 shows the method source code that use to train the clinical data set using decision tree algorithm.



```
# 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.

rfc = RandomForestClassifier(random_state=42)
rfc.fit(X_trainval_scaled, y_trainval)

# for test there are 94 cases
plot_confusion_matrix(rfc,
                      X_test_scaled,
                      y_test,
                      values_format='d',
                      display_labels=['Nondemented', 'Demented'])
```

Figure 3.8 - Method used to train the clinical data set using random forest algorithm

As in the experiment 1and 2, after training the data set with decision tree algorithm we calculate the confusion matrix to compare with the confusion matrix output of other machine learning algorithms.

3.7 Tools and technologies

3.7.1 R Language

R is a programming language and environment for statistical computing and graphics. R is a powerful statistical program that includes linear and nonlinear modeling, traditional statistical tests, time-series analysis, classification, clustering, and graphical tools. One of R's advantages is how easy it is to construct well-designed publication-quality graphs using mathematical symbols and calculations when needed. Small design choices in graphics have been carefully picked as defaults, but the user retains complete discretion (r-project.org, 2021).

We use the R programming language to perform ANOVA testing on feature attributes in this study. It aids in the identification of the clinical data set's weighted feature properties and then we can apply them to machine learning model.

3.7.2 Google colab Tool

Google Colab is a Google product. It's essentially a cloud-based notebook environment that's free to use. It offers features that allow you to edit documents in the same way that Google Docs allows you to. Many popular and high-level machine learning libraries are supported by Colab and can be quickly loaded into your notebook (research.google.com, 2019).

To implement the machine learning model in this study, we used Google Colab. So, with Google Colab's strong support for machine learning libraries, it's a lot easier to develop the model.

3.8 Summary

The implementation of two experiments is described in-depth in this chapter. Aside from methodological specifics, this chapter discusses the clinical data set as well as the many tools and technologies that we used during the study. The results and evaluation of each experiment will be discussed in the next chapter.

CHAPTER 4 - EVALUATION AND RESULTS

4.1 Introduction

The results of the research discussed in the preceding chapter are presented in this chapter. As previously stated, all experiments were conducted with a dataset split of 80% training and 20% testing. Confusion matrices will be used to show the results. A confusion matrix is a table that is used to describe an algorithm's or classification model's performance. Confusion matrices were used by many researchers to evaluate their classification model because they provide a better understanding of the actual and predicted classifications.

In machine learning, the confusion matrix is one of the best error evaluation measures. The confusion matrix for a two-class classifier is shown in the table below (Table 4.1).

Furthermore, different measurements such as sensitivity, specificity, and accuracy are employed in this chapter to stress the correctness of the outcomes in each experiment.

Table 4.1 - Confusion Matrix

		Real Class	
		Positive (Class 1)	Negative (Class 0)
Predicted Class	Positive (Class 1)	True Positives - TP	False Positives - FP
	Negative (Class 0)	False Negatives - FN	True Negatives - TN

- True Positives (TP): These are the no of correct predictions that outputs as class 1.
- True Negatives (TN): These are the no of correct predictions that outputs as class 0.
- False Positives (FP): These are the no of incorrect predictions that outputs as class 1.
- False Negatives (FN): These are the no of incorrect predictions that outputs as class 0.

The true positive rate is defined as the sensitivity or recall of a test. It represents the proportion of correctly identified positives, and it may be calculated as follows.

$$\text{Sensitivity (Recall)} \quad = \frac{TP}{TP + FN}$$

The true negative rate, which measures the proportion of correctly identified negatives, is referred to as specificity. The following formula can be used to compute the specificity.

$$\text{Specificity} \quad = \frac{TN}{TN + FN}$$

Another important measure that determines the overall accuracy of a classifier is accuracy. The following formula can be used to calculate accuracy.

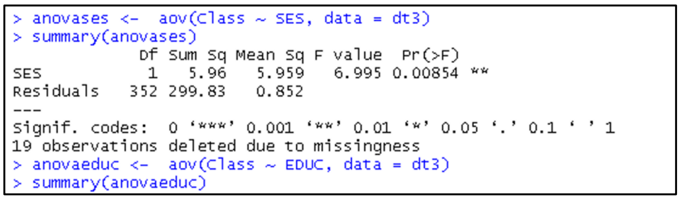
$$\text{Accuracy} \quad = \frac{TP + TN}{TP + TN + FP + FN}$$

In this chapter, the above three measures will be used to evaluate the performance of each experiment. The results of each experiment will be described in the sections that follow.

4.2 Feature Extraction Using ANOVA Test Results

After the preprocessing of clinical data set we have to select the high weighted feature attributes for the study. As a scientific approach we use ANOVA test for the selection of weighted feature attributes. Below mentioned the ANOVA test result of the each feature attribute.

1. Social Economic Status Feature Attribute (SES)



```
> anovases <-  aov(Class ~ SES, data = dt3)
> summary(anovases)
             Df Sum Sq Mean Sq F value  Pr(>F)   
SES           1   5.96   5.959   6.995 0.00854 **
Residuals   352 299.83   0.852                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
19 observations deleted due to missingness
> anovaeduc <-  aov(Class ~ EDUC, data = dt3)
> summary(anovaeduc)
```

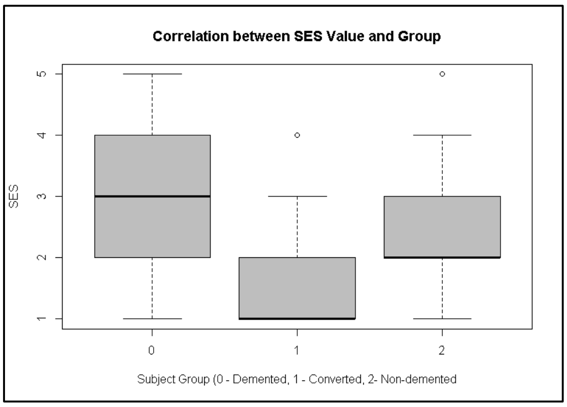
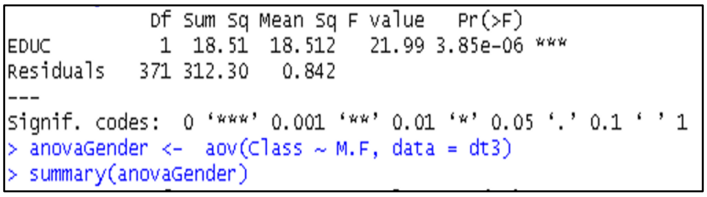


Figure 4.1 - ANOVA Test result on SES Feature attribute

Figure 4.1 shows the ANOVA test result on SES feature attribute shows that null hypothesis on the SES feature attribute can be rejected (F = 6.995, p = 0.00854 < 0.01).So there is clear impact of SES feature attribute on dementia diagnosed patients.

2. Education Feature Attribute (EDU)



```
             Df Sum Sq Mean Sq F value   Pr(>F)    
EDUC          1  18.51  18.512   21.99 3.85e-06 ***
Residuals   371 312.30   0.842                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anovaGender <-  aov(Class ~ M.F, data = dt3)
> summary(anovaGender)
```

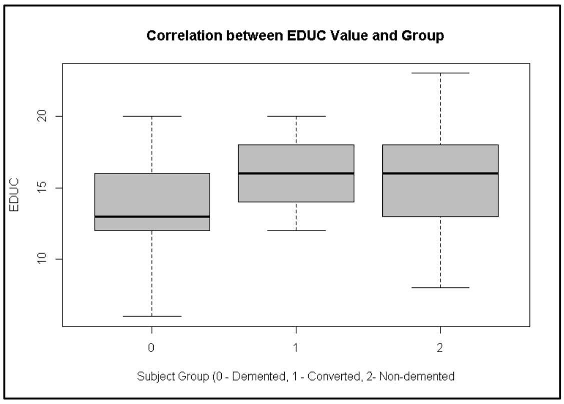
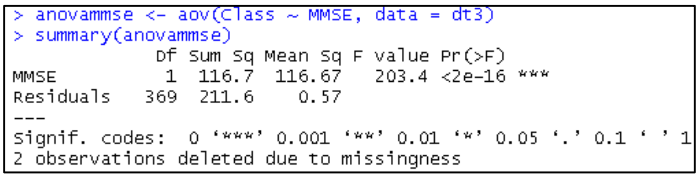


Figure 4.2 - ANOVA Test result on Education Feature attribute

Figure 4.2 shows the ANOVA test result on education feature attribute shows that null hypothesis on the education feature attribute can be rejected (*F* = 21.99, *p* = 3.85e-06 < 0.01).So there is clear impact of education feature attribute on dementia diagnosed patients.

3. <u>Mini Mental State Exam feature attribute (MMSE)</u>



```
> anovammse <- aov(Class ~ MMSE, data = dt3)
> summary(anovammse)
             Df Sum Sq Mean Sq F value Pr(>F)    
MMSE          1  116.7  116.67   203.4 <2e-16 ***
Residuals   369  211.6    0.57                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
2 observations deleted due to missingness
```

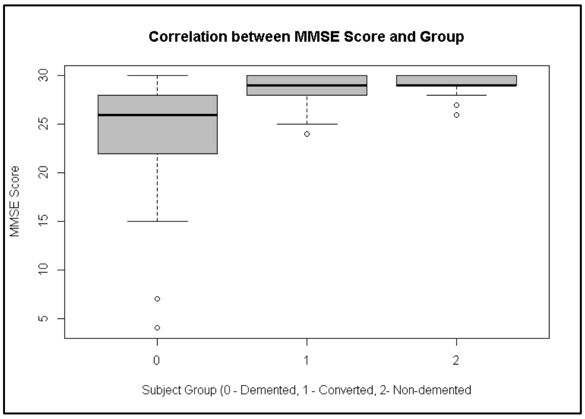
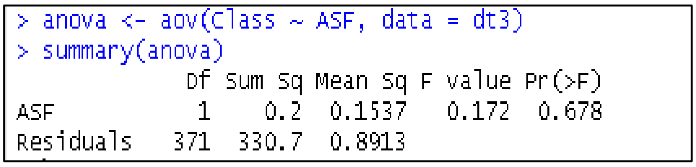


Figure 4.3 - ANOVA Test result on MMSE Feature attribute

Figure 4.3 shows the ANOVA test result on education feature attribute shows that null hypothesis on the MMSE feature attribute can be rejected (*F* = 203.4, *p* = 2e-16 < 0.01).So there is clear impact of MMSE feature attribute on dementia diagnosed patients

4. <u>Atlas Scaling Factor feature attribute (ASF)</u>



```
> anova <- aov(Class ~ ASF, data = dt3)
> summary(anova)
             Df Sum Sq Mean Sq F value Pr(>F)
ASF           1    0.2  0.1537   0.172  0.678
Residuals   371  330.7  0.8913               
```

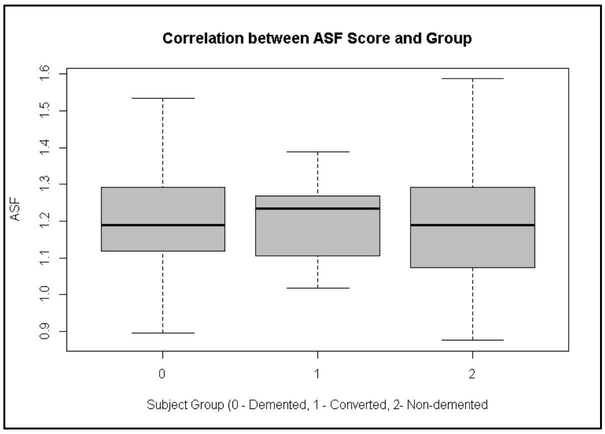
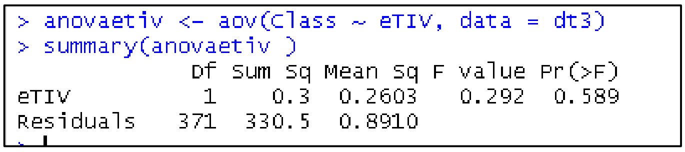


Figure 4.4 - ANOVA Test result on ASF Feature attribute

ANOVA test result on ASF feature attribute shows that null hypothesis on the ASF feature attribute cannot be rejected (*F* = 0.172, *p* = 0.678 > 0.01). As a result, this feature is unable to contribute in the creation of a machine learning model.

5. Estimated Total Intracranial Volume feature attribute (eTIV)



```
> anovaetiv <- aov(Class ~ eTIV, data = dt3)
> summary(anovaetiv )
             Df Sum Sq Mean Sq F value Pr(>F)
eTIV          1    0.3  0.2603   0.292  0.589
Residuals   371  330.5  0.8910
```

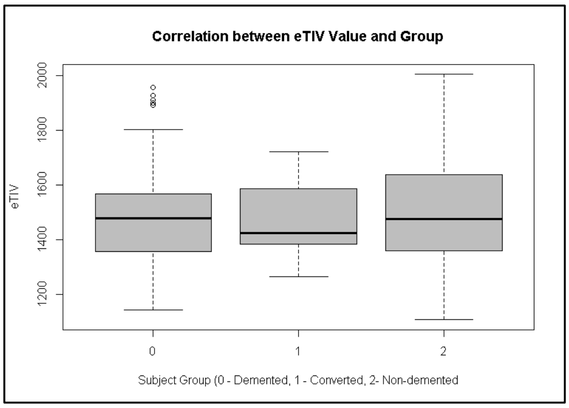


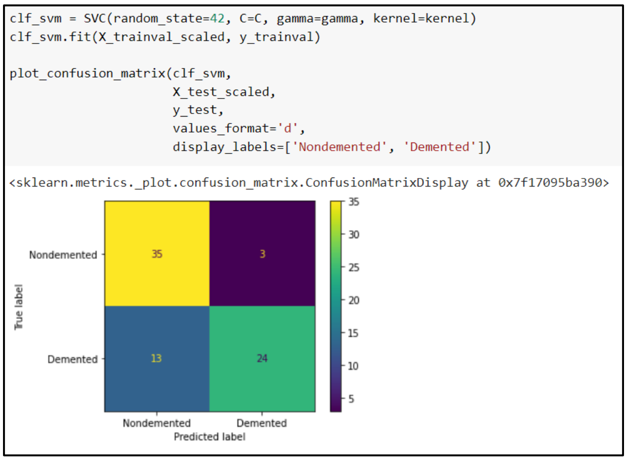
Figure 4.5 - ANOVA Test result on eTIV Feature attribute

Figure 4.5 shows the ANOVA test result on eTIV feature attribute shows that null hypothesis on the eTIV feature attribute cannot be rejected (*F* = 0.292, *p* = 0.589 > 0.01).As a result, this feature is unable to contribute in the creation of a machine learning model

We use feature characteristics indicated in the methodology section for the machine learning model implementation after assessing the ANOVA test output results on feature attributes (considering the F value and p-value of the output result). As a result, the ANOVA test assists us in confirming the selected hypothesis for each feature attribute.

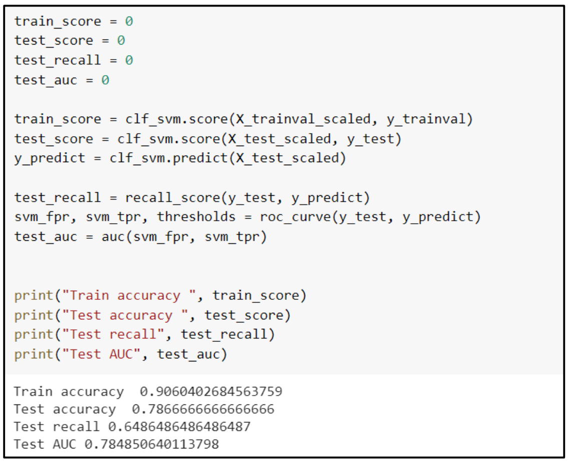
4.3 Experiment 1 – Accuracy testing using SVM Algorithm

For the experiment 1, we use the preprocessed clinical data set contains 373 patient clinical data records, and we used an 80/20 training and testing data ratio. Because SVM is a binary classifier that does not do well in multiclass classifications, only two classes were employed in this experiment (Dementia and Healthy - NL) (Mayoraz & E, 2006).The output of the Experiment one shows in the figure 4.6 and figure 4.7



```
clf_svm = SVC(random_state=42, C=C, gamma=gamma, kernel=kernel)
clf_svm.fit(X_trainval_scaled, y_trainval)

plot_confusion_matrix(clf_svm,
                      X_test_scaled,
                      y_test,
                      values_format='d',
                      display_labels=['Nondemented', 'Demented'])
```

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7f17095ba390>
```



```
train_score = 0
test_score = 0
test_recall = 0
test_auc = 0

train_score = clf_svm.score(X_trainval_scaled, y_trainval)
test_score = clf_svm.score(X_test_scaled, y_test)
y_predict = clf_svm.predict(X_test_scaled)

test_recall = recall_score(y_test, y_predict)
svm_fpr, svm_tpr, thresholds = roc_curve(y_test, y_predict)
test_auc = auc(svm_fpr, svm_tpr)


print("Train accuracy ", train_score)
print("Test accuracy ", test_score)
print("Test recall", test_recall)
print("Test AUC", test_auc)
```

```
Train accuracy  0.9060402684563759
Test accuracy  0.7866666666666666
Test recall 0.6486486486486487
Test AUC 0.784850640113798
```

Figure 4.6 - Experiment 1 – SVM result output

Figure 4.7 - Experiment 1 – SVM confusion matrix

The Table 4.2 shows the confusion matrix of the experiment 1.

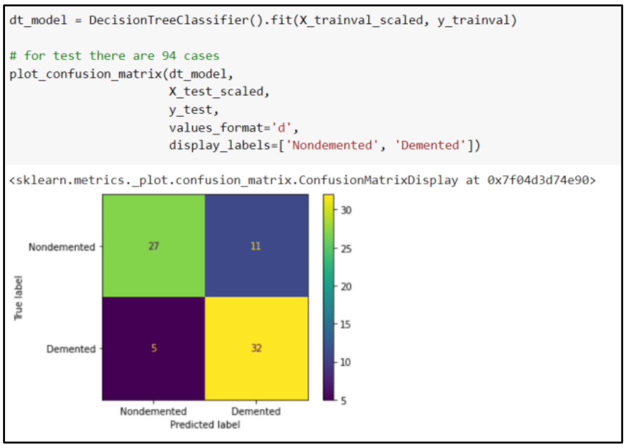
Table 4.2 - Confusion matrix of experiment 1

Used Algorithm	Train Accuracy	Test Accuracy	Test Recall(Sensitivity)
SVM Algorithm	90.6%	78.6%	64.8%

According to the confusion matrix illustrated in table 4.2 train accuracy is 90.6%, test accuracy is 78.6% and the sensitivity is 64.8%.

4.4 Experiment 2 – Accuracy testing using Decision Tree Algorithm

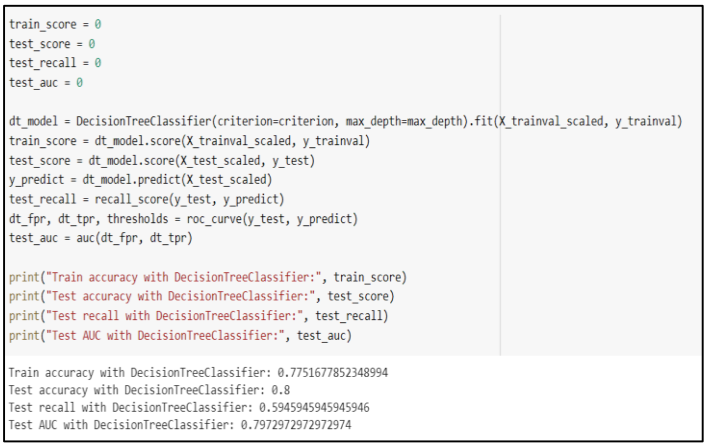
For the experiment 2, we also use the preprocessed clinical data set contains 373 patient clinical data records, and we used an 80/20 training and testing data ratio. As in the experiment 1, we only two classes were employed in this experiment (Dementia and Healthy - NL). The output of the Experiment one shows in the figure 4.8 and figure 4.9



```
dt_model = DecisionTreeClassifier().fit(X_trainval_scaled, y_trainval)

# for test there are 94 cases
plot_confusion_matrix(dt_model,
                      X_test_scaled,
                      y_test,
                      values_format='d',
                      display_labels=['Nondemented', 'Demented'])
```

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7f04d3d74e90>
```

Figure 4.8 - Experiment 2 – Decision tree result output



```
train_score = 0
test_score = 0
test_recall = 0
test_auc = 0

dt_model = DecisionTreeClassifier(criterion=criterion, max_depth=max_depth).fit(X_trainval_scaled, y_trainval)
train_score = dt_model.score(X_trainval_scaled, y_trainval)
test_score = dt_model.score(X_test_scaled, y_test)
y_predict = dt_model.predict(X_test_scaled)
test_recall = recall_score(y_test, y_predict)
dt_fpr, dt_tpr, thresholds = roc_curve(y_test, y_predict)
test_auc = auc(dt_fpr, dt_tpr)

print("Train accuracy with DecisionTreeClassifier:", train_score)
print("Test accuracy with DecisionTreeClassifier:", test_score)
print("Test recall with DecisionTreeClassifier:", test_recall)
print("Test AUC with DecisionTreeClassifier:", test_auc)
```

```
Train accuracy with DecisionTreeClassifier: 0.7751677852348994
Test accuracy with DecisionTreeClassifier: 0.8
Test recall with DecisionTreeClassifier: 0.5945945945945946
Test AUC with DecisionTreeClassifier: 0.7972972972972974
```

Figure 4.9 - Experiment 2 – Decision tree confusion matrix

The Table 4.3 shows the confusion matrix of the experiment 2.

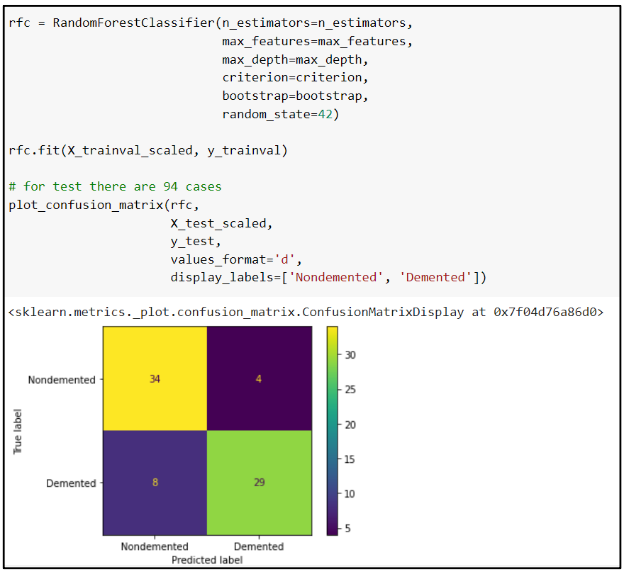
Table 4.3 - Confusion matrix of experiment 2

Used Algorithm	Train Accuracy	Test Accuracy	Test Recall(Sensitivity)
Decision Tree Algorithm	77.5%	80.0%	59.4%

According to the confusion matrix illustrated in table 4.3 train accuracy is 77.5%, test accuracy is 80.0% and the sensitivity is 59.4%.

4.5 Experiment 3 – Accuracy testing using Random Forest Algorithm

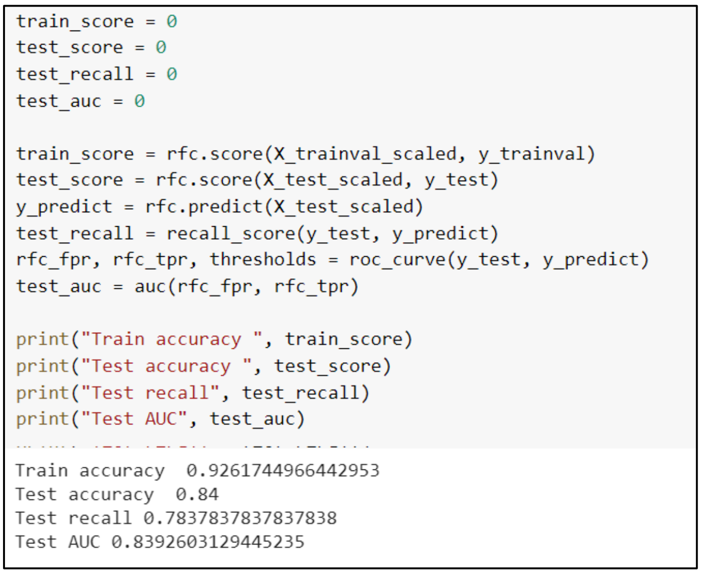
For the experiment 3, we also use the preprocessed clinical data set contains 373 patient clinical data records, and we used an 80/20 training and testing data ratio. As in the experiments 1 and 2, we only two classes were employed in this experiment (Dementia and Healthy - NL). The output of the Experiment one shows in the figure 4.10 and figure 4.11



```
rfc = RandomForestClassifier(n_estimators=n_estimators,
                             max_features=max_features,
                             max_depth=max_depth,
                             criterion=criterion,
                             bootstrap=bootstrap,
                             random_state=42)

rfc.fit(X_trainval_scaled, y_trainval)

# for test there are 94 cases
plot_confusion_matrix(rfc,
                      X_test_scaled,
                      y_test,
                      values_format='d',
                      display_labels=['Nondemented', 'Demented'])
```

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7f04d76a86d0>
```

Figure 4.10 - Experiment 3 – Random forest result output



```
train_score = 0
test_score = 0
test_recall = 0
test_auc = 0

train_score = rfc.score(X_trainval_scaled, y_trainval)
test_score = rfc.score(X_test_scaled, y_test)
y_predict = rfc.predict(X_test_scaled)
test_recall = recall_score(y_test, y_predict)
rfc_fpr, rfc_tpr, thresholds = roc_curve(y_test, y_predict)
test_auc = auc(rfc_fpr, rfc_tpr)

print("Train accuracy ", train_score)
print("Test accuracy ", test_score)
print("Test recall", test_recall)
print("Test AUC", test_auc)
```

```
Train accuracy  0.9261744966442953
Test accuracy  0.84
Test recall 0.7837837837837838
Test AUC 0.8392603129445235
```

Figure 4.11 - Experiment 3 – Random forest confusion matrix

The Table 4.4 shows the confusion matrix of the experiment 3.

Table 4.4 - Confusion matrix of experiment 3

Used Algorithm	Train Accuracy	Test Accuracy	Test Recall(Sensitivity)
Random Forest Algorithm	92.5%	84.0%	78.3%

According to the confusion matrix illustrated in table 4.4 train accuracy is 92.5%, test accuracy is 84.0% and the sensitivity is 78.3%.

4.6 Discussion

This section will discuss about the results of each experiment and how those results can be used to answer the research questions.

Many studies on early dementia diagnosis have been published in the last decade. Those research, on the other hand, were carried out with distinct datasets and methodology. As a result, comparing such results is difficult because different methodologies cannot be compared with different datasets.

When looking at all of the previous studies on dementia early detection, only one machine learning algorithm was utilized with clinical data from dementia patients. As a result, in this work, we used SVM, decision trees, and random forests to test the early diagnosis approaches with clinical data.

- Research problem statement: To what extent can a classifier predict the progression of subjects from clinical data to dementia and the progression’s corresponding moment and which machine learning algorithm is most suitable for the classifier?

According to the results obtained from the every experiment we can conclude the output of each experiments as in the below table 4.5.

Table 4.5 – confusion matrix summary of all three experiments

Used Algorithm	Train Accuracy	Test Accuracy	Test Recall(Sensitivity)
SVM Algorithm	90.6%	78.6%	64.8%
Decision Tree Algorithm	77.5%	80.0%	59.4%
Random Forest Algorithm	**92.5%**	**84.0%**	**78.3%**

As a result, we can answer the study's research question based on the results of each experiment in table 4.5. Machine learning algorithms can always be used to predict dementia utilizing clinical data from dementia patients. We primarily employ the five clinical data parameters of dementia patients specified in the 3.4 section in this investigation.

We use three main machine learning algorithms in this work to find the best method for pre-detection of dementia using clinical data from patients. As in the concluded confusion matrix, the SVM and Decision tree algorithms are not suitable for use as classification algorithms in a machine learning model to pre-detect dementia. According to results Random forest algorithm gives a training accuracy of 92.5%, test accuracy of 84.0% and sensitivity of 78.3%. As a result, when compared to the SVM and decision tree algorithms, implementing a machine learning model with the random forest algorithm should have the best accuracy and sensitivity.

4.7 <u>Summary</u>

Each experiment's results and evaluation are presented in this chapter. Confusion matrices were used to describe the results and evaluations. Finally, each of the results was discussed to answer the research question.

CHAPTER 5 - CONCLUSION AND FUTURE WORK

5.1 Conclusion

This thesis shows how machine learning algorithms and clinical data from dementia patients can be utilized to predict the dementia at an early stage. Experimental and exploratory approaches have been taken to compare the results from different techniques and methods.

Dementia has recently emerged as an essential disease to be concerned about for the development of society. In addition, there is no cure for dementia condition. As a result, it is critical to diagnose dementia early in order to use existing treatments to decrease the disease's course. Still there is no clear method to get the absolute confirmation of the disease. Furthermore, in the recent few decades, many investigations on dementia illness pre-detection utilizing various methodologies have been conducted. Early detection of dementia disease, on the other hand, is critical for the introduction of disease-modifying medications. As a result, the focus of this research was on using machine learning approaches to pre detect dementia disease.

For this study, we primarily use the clinical data of dementia patients. We design three key experiments after preprocessing the clinical data to find the best machine learning algorithm for the machine learning model.

In the first experiment, we utilize the SVM algorithm to train the machine learning model and examine the confusion matrix outcomes after preprocessing clinical data. According to the confusion matrix from experiment 1, it gives the train accuracy is 90.6 %, the test accuracy is 78.6 %, and the sensitivity is 64.8 %.

In the experiment 2, we utilize the decision tree algorithm to train the machine learning model and examine the confusion matrix outcomes after preprocessing clinical data. . According to the confusion matrix from experiment 2, it gives the train accuracy is 77.5 %, the test accuracy is 80.0 %, and the sensitivity is 59.4 %.

In the experiment 3, we utilize the random forest algorithm to train the machine learning model and examine the confusion matrix outcomes after preprocessing clinical data. . According to the confusion matrix from experiment 3, it gives the train accuracy is 92.5 %, the test accuracy is 84.0 %, and the sensitivity is 78.3 %.

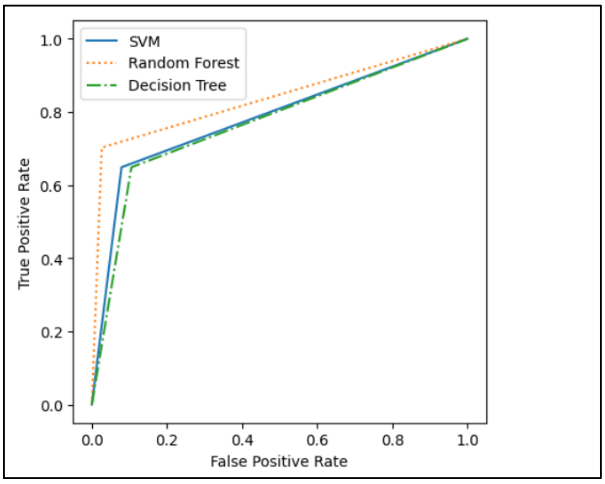


Figure 5.1 - ROC of SVM, decision tree and random forest compared with accuracy

Figure 5.1 shows the ROC value of SVM, decision tree and random forest algorithm with compared to the accuracy.

So, when compared to the SVM and decision tree algorithms, implementing a machine learning model with the random forest algorithm should have the best accuracy and sensitivity. We can conclude that with the clinical data of dementia patients we can pre-detect the diagnosis of dementia disease and as a machine learning algorithm random forest algorithm gives the most accurate results.

Considering train accuracy, test accuracy, and sensitivity of the above experiments this study got very promising results. Therefore proposed machine learning-based method will be useful in identifying early diagnosis of dementia with the clinical data of dementia patients.

5.2 <u>Contribution</u>

Our research objectives, as indicated in section 1.3.2, have been met, and the contribution of this work will be discussed.

We collected clinical data and selected data points for this study, extracted essential information from the data, and identified key values from the collected data set. We identify the effective feature features that influence dementia diagnosis in clinical patients who have already been diagnosed. We used data reduction, data discretization, data cleaning, and data transformation to preprocess the gathered data in this study. Using the obtained effective feature attribute values and their weights, we build and create a dataset on dementia disorder.

In this study, we conducted three trials and compared the confusion matrix findings to the machine learning technique used. Among them, we use a random forest algorithm to create the machine learning model.

This research aids the general population in detecting dementia at an early stage. We construct a mobile-based application for the general public that is integrated with the machine learning model we developed. This mobile app consists of a series of questions that are aimed to assess a variety of everyday mental capabilities. This will evaluate the person's dementia level and aid in the early detection of dementia. As a result, this research will be a significant step forward in the prevention of dementia in Sri Lanka.

5.3 <u>Future works</u>

Following potential avenues of investigation can suggest as future experiments.

- With this COVID 19 situation we do this study with the limited set of data clinical dataset. So I would like to suggest to do this study with large no of clinical data set in the same way.

- In this study, we only concern about the clinical data attributes of the dementia patient. It would be beneficial to improve this study by include non-clinical data variables such as occupation, marital status, and the number of children in the family.

- As a future enhancement I would like to suggest using different experiments and then further can be improved by fine-tuning the machine learning model

REFERENCES

1. Albert, M. .. et al., 2011. *The diagnosis of mild cognitive impairment due to Alzheimer's disease Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease.* s.l.:s.n.

2. Alzheimer's Association, 2015. *what-is-alzheimers.* [Online]
   Available at: https://www.alz.org/alzheimers-dementia/what-is-alzheimers

3. Alzheimer's Association, 2020. https://www.alz.org/alzheimers-dementia/stages.

4. alzheimers asspciation, 2021. *Brain Plaques and Tangles - Alzheimer's Association.*. [Online] Available at: https://www.alz.org/alzheimers-dementia/what-is-alzheimers/brain_tour_part_2

5. Apostolova, L. G. et al., 2012. Hippocampal atrophy and ventricular enlargement in normal aging, mild cognitive impairment and Alzheimer Disease.. Volume 26.

6. Arvesen, E., 2015. *Automatic Classification of Alzheimer's Disease from Structural MRI.* s.l.:s.n.

7. Brownlee, J., 2019. *Difference Between Algorithm and Model in Machine Learning.* [Online]
   Available at: https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/

8. Chen, R.; Herskovits, E.H, 2010. *Machine-learning techniques for building a diagnostic model for very mild dementia.*, s.l.: s.n.

9. Cho, P. C. & Chen, W. H., 2012. A double layer dementia diagnosis system using machine learning techniques..

10. Chyzhyk, D. & Savio, A., 2010. Feature extraction from structural MRI images based on VBM: data from OASIS database..

11. Farina, N. et al., 2012. Vitamin E for Alzheimer's dementia and mild cognitive impairment. *National Library Of Medicine.*

12. Hooshyar, D., Park, K. W. & Lim, H. S., 2008. Early Diagnosis of Dementia from Clinical Data by Machine Learning Techniques..

13. https://docs.microsoft.com/, 2021. *What is a machine learning model?.* [Online]
    Available at: https://docs.microsoft.com/en-us/windows/ai/windows-ml/what-is-a-machine-learning-model

14. https://www.datarobot.com/, n.d. *Machine Learning Model | What are Machine Learning Models.* [Online]
    Available at: https://www.datarobot.com/wiki/model/

15. javatpoint.com, 2021. *Decision Tree Classification Algorithm.* [Online]
    Available at: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

16. javatpoint.com, 2021. *Random Forest Algorithm.* [Online]
Available at: https://www.javatpoint.com/machine-learning-random-forest-algorithm

17. Joshi, S.; Shenoy, P.D.; Venugopal, K.R.; Patnaik, L.M., 2009. *Evaluation of different stages of dementia employing neuropsychological and machine learning techniques.,* s.l.: s.n.

18. Kloppel, S. et al., 2008. Automatic classification of MR scans in Alzheimer's disease..

19. Laakso, M. P. et al., 1996. Hippocampal volumes in Alzheimer's disease, Parkinson's disease with and without dementia, and in vascular dementia.

20. Lozano, R. et al., 2010. *Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010.* s.l.:s.n.

21. Marcus, D. S. et al., 2007. cross-sectional MRI data in young, middle aged, nondemented, and demented older adults.

22. Mayoraz, E. & E, A., 2006. *Support Vector Machines for Multiclass Classification.* [Online].

23. Moni, R., 2021. *Reinforcement Learning algorithms — an intuitive overview.* [Online]
Available at: https://smartlabai.medium.com/reinforcement-learning-algorithms-an-intuitive-overview-904e2dff5bbc

24. National Institute on Aging, 2021. *About Alzheimer's Disease: Diagnosis.* [Online]
Available at: https://www.nia.nih.gov/alzheimers/topics/diagnosis

25. Noria, V. S. et al., 2019. Machine learning models to predict onset of dementia: A label learning.

26. opencv.org, 2021. *Introduction to Support Vector Machines — OpenCV 2.4.13.1 documentation..* [Online]
Available at:
http://docs.opencv.org/2.4/doc/tutorials/ml/introductiontosvmintroductiontosvm.html.

27. Prince, M. et al., 2015. *World Alzheimer Report 2015 The Global Impact of Dementia An AnAlysIs of prevAlence, IncIDence,cosT AnD TrenDs.* s.l.:s.n.

28. research.google.com, 2019. [Online]
Available at: https://research.google.com/colaboratory/faq.html

29. r-project.org, 2021. [Online]
Available at: https://www.r-project.org/about.html

30. Shankle, W. R., Datta, P., Dillencourt, M. & Pazzani, M., 1996. Improving dementia screening tests with machine learning methods.

31. Shree, B. R. & Sheshadri, H. S., 2014. An initial investigation in the diagnosis of Alzheimer's disease using various classification techniques..

32. Soni, J., 2011. Predictive Data Mining for medical diagnosis An Overview of Heart Disease Prediction. *International Journal of Computer Applications.*

33. statista.com, n.d. *https://www.statista.com/statistics/738359/sri-lanka-projected-number-of-people-with-dementia/.* [Online].

34. Stonnington, C. M. et al., 2008. Automatic classification of MR scans in Alzheimer’s disease. Volume 131.

35. The Rag Tree., 2020. *The Human Brain–The Quaggas of Creativity 2.* [Online] Available at: https://cathay12.wordpress.com/2012/08/26/the-human-brain-the-quaggas-of-creativity-2/

36. Tohka, J; Moradi, E; Huttunen, H, 2016. Comparison of feature selection techniques in machine learning for anatomical brain mri in dementia.

37. Trambaiolli, L.R.; Lorena, A.C.; Fraga, F.J.; Kanda, P.A.; Anghinah, R.; Nitrini, R, 2011. *Improving Alzheimer’s disease diagnosis with machine learning techniques,* s.l.: s.n.

38. Vemuri, P. et al., 2008. Alzheimer's disease diagnosis inindividual subjects using structural MR images..

39. Williams, J.A.; Weakley, A.; Cook, D.J.; Schmitter-Edgecombe, M., 2013. *Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia.,* s.l.: s.n.

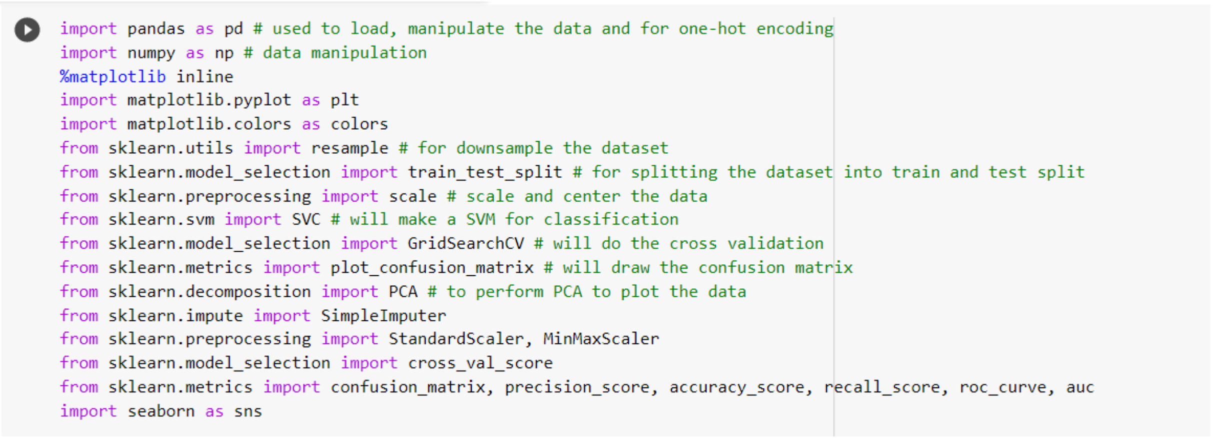
40. Williams, et al., 2013. Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia..

41. Wu, Y.-T.et al., 2017. The changing prevalence and incidence of dementia over time - current evidence.

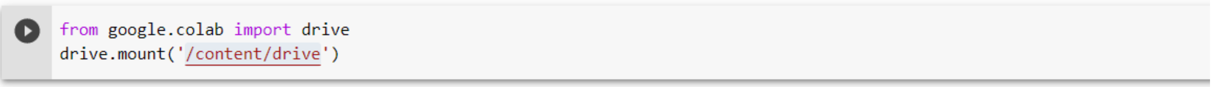
42. Zhang, Y. et al., 2015. Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning.

Appendix A

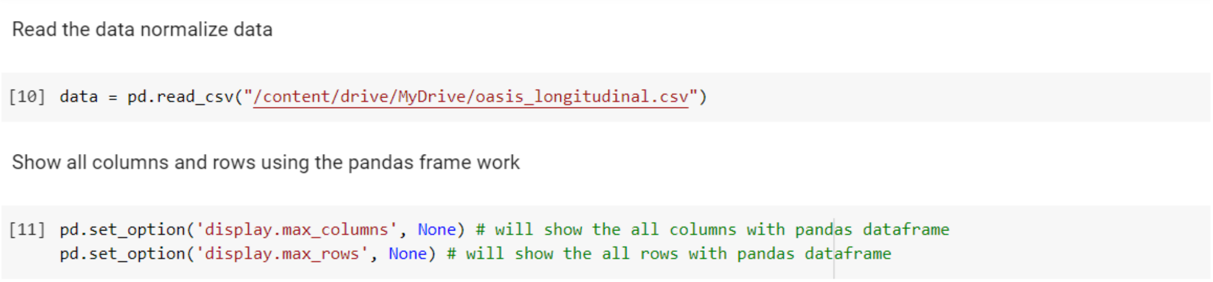
Machine Learning Model Implementation



```
import pandas as pd # used to load, manipulate the data and for one-hot encoding
import numpy as np # data manipulation
%matplotlib inline
import matplotlib.pyplot as plt
import matplotlib.colors as colors
from sklearn.utils import resample # for downsample the dataset
from sklearn.model_selection import train_test_split # for splitting the dataset into train and test split
from sklearn.preprocessing import scale # scale and center the data
from sklearn.svm import SVC # will make a SVM for classification
from sklearn.model_selection import GridSearchCV # will do the cross validation
from sklearn.metrics import plot_confusion_matrix # will draw the confusion matrix
from sklearn.decomposition import PCA # to perform PCA to plot the data
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from sklearn.model_selection import cross_val_score
from sklearn.metrics import confusion_matrix, precision_score, accuracy_score, recall_score, roc_curve, auc
import seaborn as sns
```



```
from google.colab import drive
drive.mount('/content/drive')
```

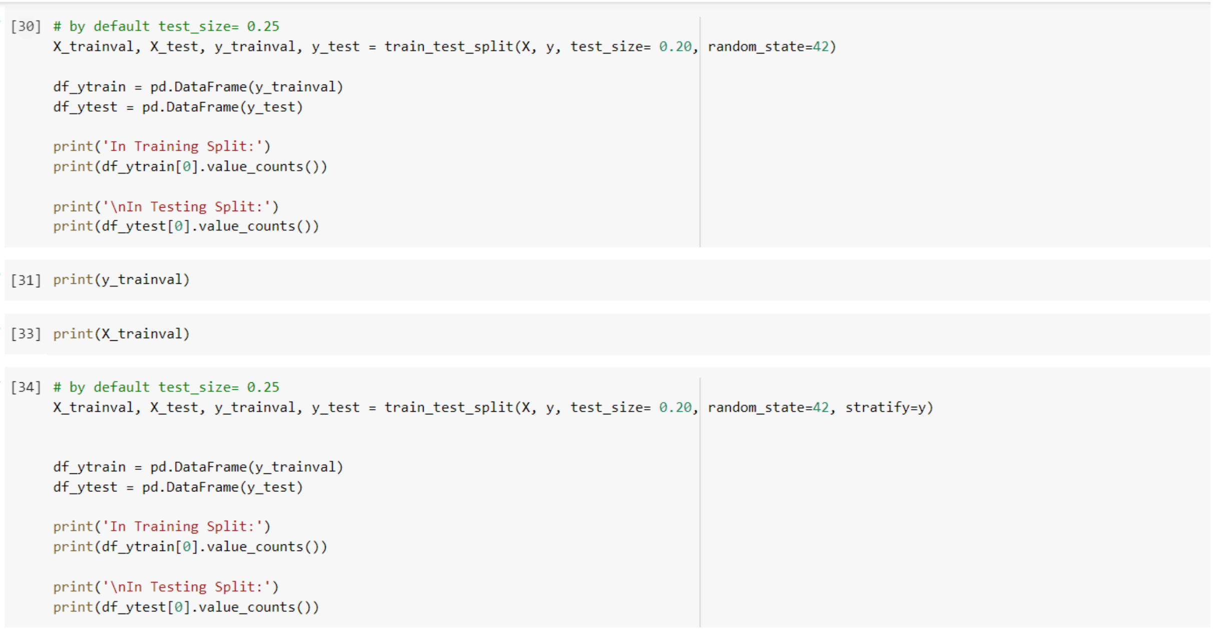


Read the data normalize data

```
[10] data = pd.read_csv("/content/drive/MyDrive/oasis_longitudinal.csv")
```

Show all columns and rows using the pandas frame work

```
[11] pd.set_option('display.max_columns', None) # will show the all columns with pandas dataframe
     pd.set_option('display.max_rows', None) # will show the all rows with pandas dataframe
```

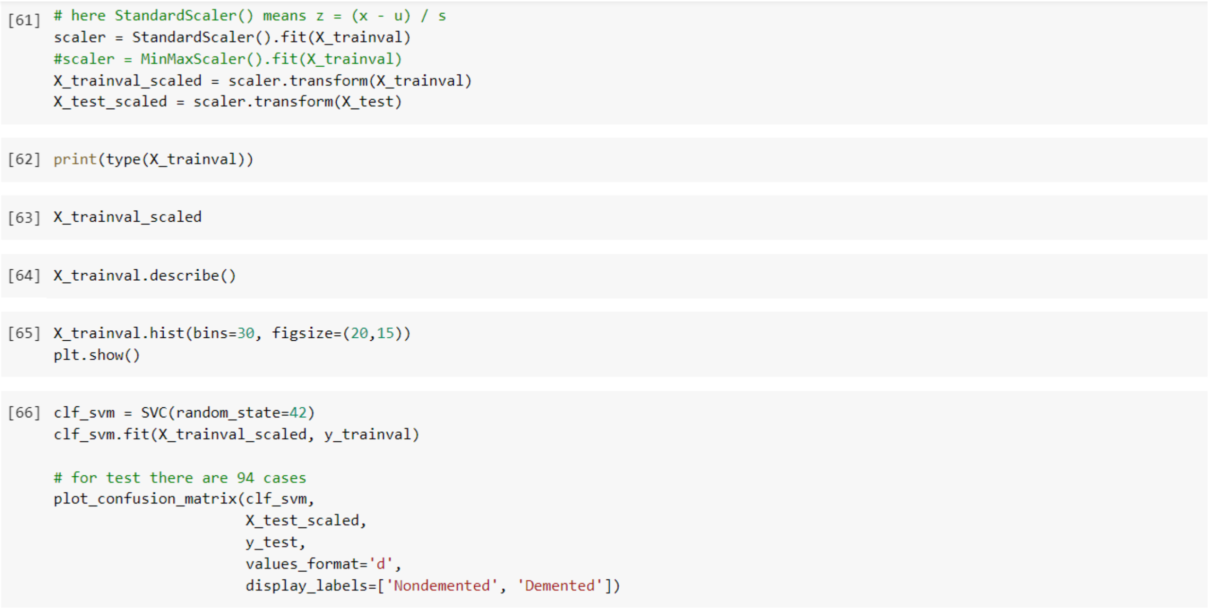


```
[30] # by default test_size= 0.25
     X_trainval, X_test, y_trainval, y_test = train_test_split(X, y, test_size= 0.20, random_state=42)

     df_ytrain = pd.DataFrame(y_trainval)
     df_ytest = pd.DataFrame(y_test)

     print('In Training Split:')
     print(df_ytrain[0].value_counts())

     print('\nIn Testing Split:')
     print(df_ytest[0].value_counts())
```

```
[31] print(y_trainval)
```

```
[33] print(X_trainval)
```

```
[34] # by default test_size= 0.25
     X_trainval, X_test, y_trainval, y_test = train_test_split(X, y, test_size= 0.20, random_state=42, stratify=y)


     df_ytrain = pd.DataFrame(y_trainval)
     df_ytest = pd.DataFrame(y_test)

     print('In Training Split:')
     print(df_ytrain[0].value_counts())

     print('\nIn Testing Split:')
     print(df_ytest[0].value_counts())
```



```
[61] # here StandardScaler() means z = (x - u) / s
     scaler = StandardScaler().fit(X_trainval)
     #scaler = MinMaxScaler().fit(X_trainval)
     X_trainval_scaled = scaler.transform(X_trainval)
     X_test_scaled = scaler.transform(X_test)
```

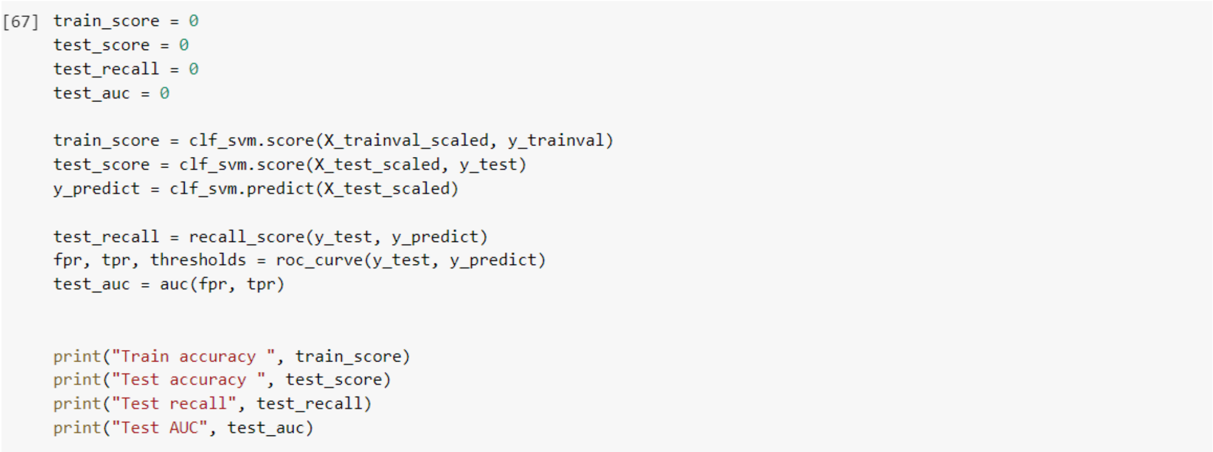
```
[62] print(type(X_trainval))
```

```
[63] X_trainval_scaled
```

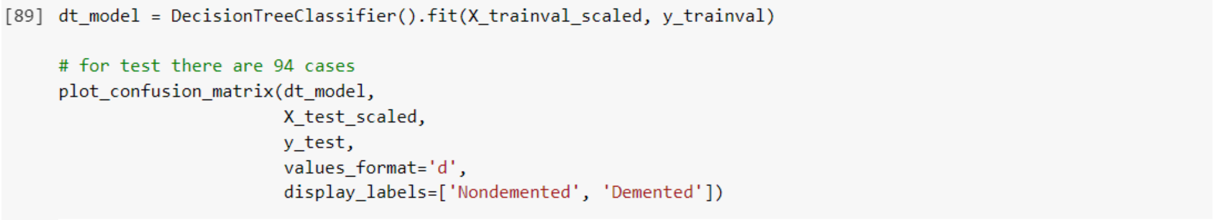
```
[64] X_trainval.describe()
```

```
[65] X_trainval.hist(bins=30, figsize=(20,15))
     plt.show()
```

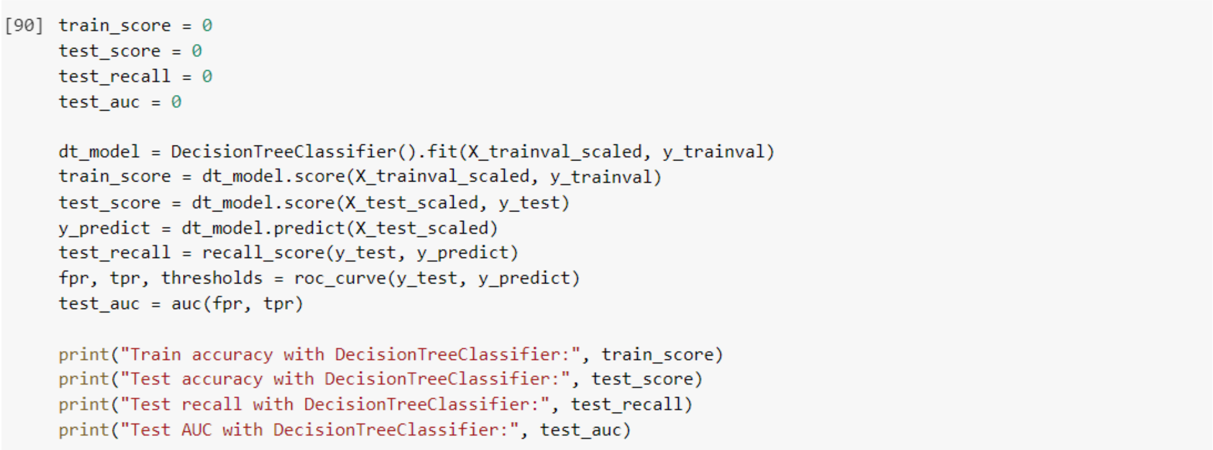
```
[66] clf_svm = SVC(random_state=42)
     clf_svm.fit(X_trainval_scaled, y_trainval)

     # for test there are 94 cases
     plot_confusion_matrix(clf_svm,
                           X_test_scaled,
                           y_test,
                           values_format='d',
                           display_labels=['Nondemented', 'Demented'])
```



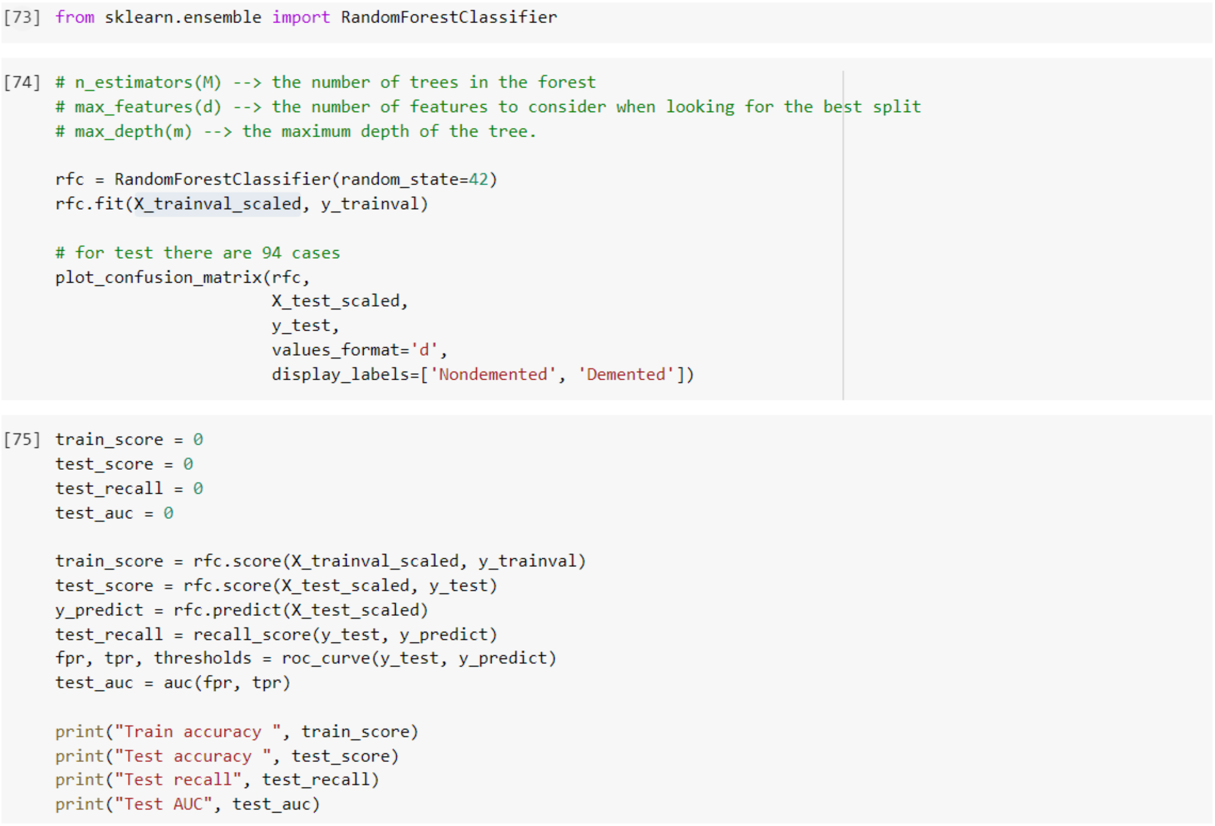
```
[67] train_score = 0
     test_score = 0
     test_recall = 0
     test_auc = 0

     train_score = clf_svm.score(X_trainval_scaled, y_trainval)
     test_score = clf_svm.score(X_test_scaled, y_test)
     y_predict = clf_svm.predict(X_test_scaled)

     test_recall = recall_score(y_test, y_predict)
     fpr, tpr, thresholds = roc_curve(y_test, y_predict)
     test_auc = auc(fpr, tpr)


     print("Train accuracy ", train_score)
     print("Test accuracy ", test_score)
     print("Test recall", test_recall)
     print("Test AUC", test_auc)
```



```
[89] dt_model = DecisionTreeClassifier().fit(X_trainval_scaled, y_trainval)

     # for test there are 94 cases
     plot_confusion_matrix(dt_model,
                           X_test_scaled,
                           y_test,
                           values_format='d',
                           display_labels=['Nondemented', 'Demented'])
```



```
[90] train_score = 0
     test_score = 0
     test_recall = 0
     test_auc = 0

     dt_model = DecisionTreeClassifier().fit(X_trainval_scaled, y_trainval)
     train_score = dt_model.score(X_trainval_scaled, y_trainval)
     test_score = dt_model.score(X_test_scaled, y_test)
     y_predict = dt_model.predict(X_test_scaled)
     test_recall = recall_score(y_test, y_predict)
     fpr, tpr, thresholds = roc_curve(y_test, y_predict)
     test_auc = auc(fpr, tpr)

     print("Train accuracy with DecisionTreeClassifier:", train_score)
     print("Test accuracy with DecisionTreeClassifier:", test_score)
     print("Test recall with DecisionTreeClassifier:", test_recall)
     print("Test AUC with DecisionTreeClassifier:", test_auc)
```



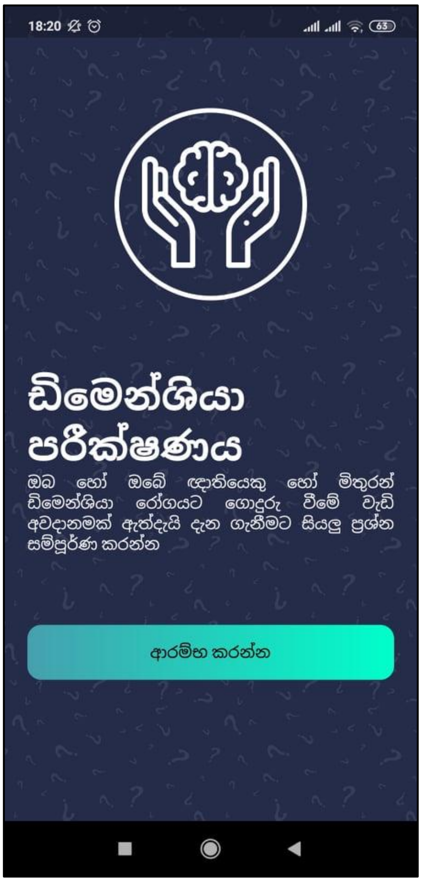
```
[73] from sklearn.ensemble import RandomForestClassifier
```

```
[74] # 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.

     rfc = RandomForestClassifier(random_state=42)
     rfc.fit(X_trainval_scaled, y_trainval)

     # for test there are 94 cases
     plot_confusion_matrix(rfc,
                           X_test_scaled,
                           y_test,
                           values_format='d',
                           display_labels=['Nondemented', 'Demented'])
```

```
[75] train_score = 0
     test_score = 0
     test_recall = 0
     test_auc = 0

     train_score = rfc.score(X_trainval_scaled, y_trainval)
     test_score = rfc.score(X_test_scaled, y_test)
     y_predict = rfc.predict(X_test_scaled)
     test_recall = recall_score(y_test, y_predict)
     fpr, tpr, thresholds = roc_curve(y_test, y_predict)
     test_auc = auc(fpr, tpr)

     print("Train accuracy ", train_score)
     print("Test accuracy ", test_score)
     print("Test recall", test_recall)
     print("Test AUC", test_auc)
```

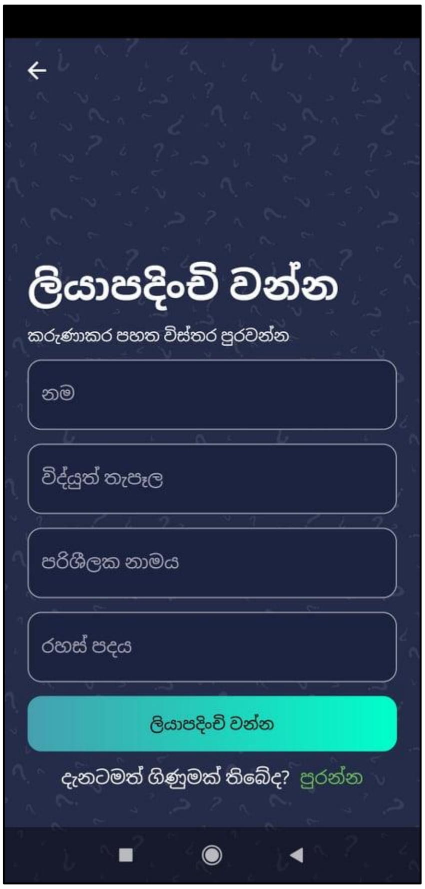
Appendix B

Mobile Application for General Public



ඩිමෙන්ශියා පරීක්ෂණය

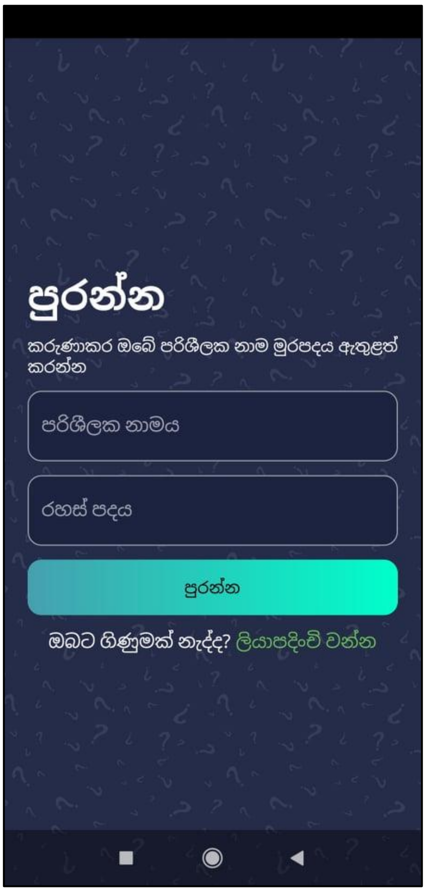
ඔබ හෝ ඔබේ ඥාතියෙකු හෝ මිතුරන් ඩිමෙන්ශියා රෝගයට ගොදුරු වීමේ වැඩි අවදානමක් ඇත්දැයි දැන ගැනීමට සියලු ප්‍රශ්න සම්පූර්ණ කරන්න



ලියාපදිංචි වන්න

කරුණාකර පහත විස්තර පුරවන්න

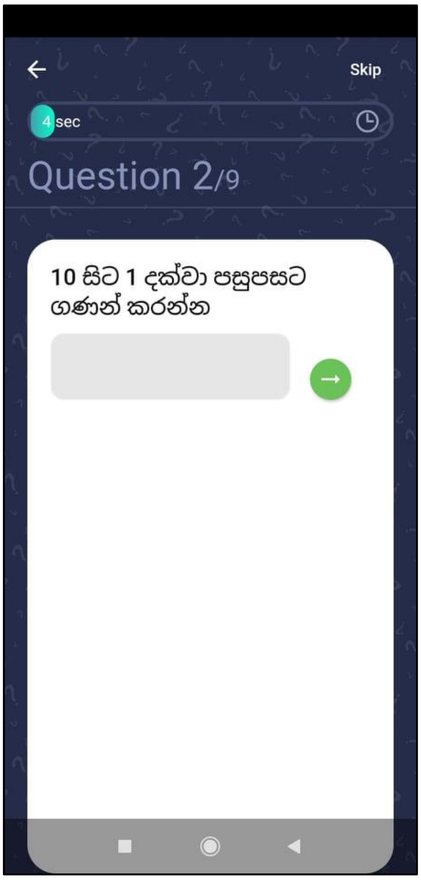
දැනටමත් ගිණුමක් තිබේද? පුරන්න



පුරන්න

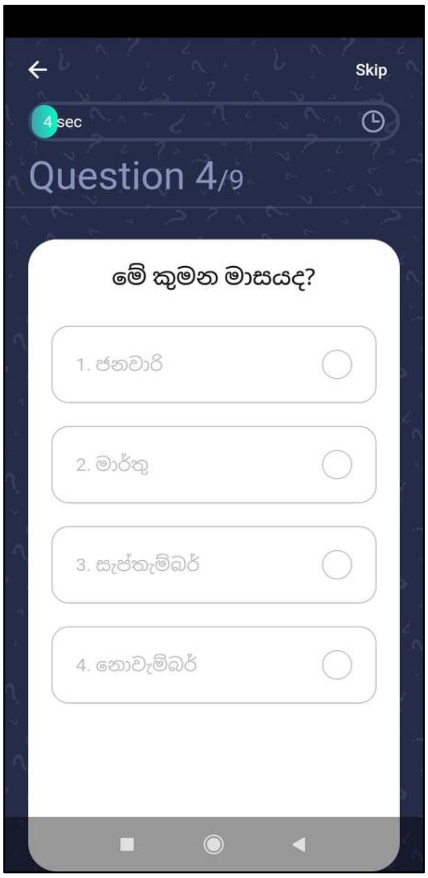
කරුණාකර ඔබේ පරිශීලක නාම මුරපදය ඇතුළත් කරන්න

ඔබට ගිණුමක් නැද්ද? ලියාපදිංචි වන්න



Question 2/9

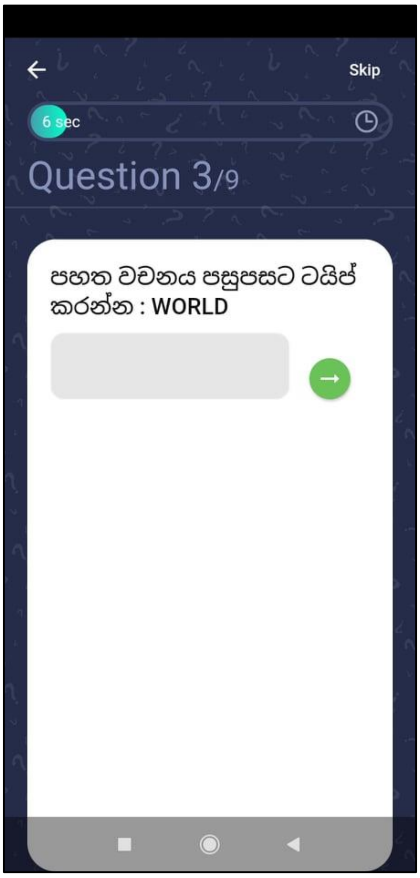
10 සිට 1 දක්වා පසුපසට ගණන් කරන්න



Question 4/9

මේ කුමන මාසයද?

1. ජනවාරි ○
2. මාර්තු ○
3. සැප්තැම්බර් ○
4. නොවැම්බර් ○



Question 3/9

පහත වචනය පසුපසට ටයිප් කරන්න : WORLD