

# **A Machine Learning Classification Framework for Early Prediction of Alzheimer's Disease**

**By**

Mohamed A.M. Mahyoub

**BSc (Hons), MPhil/PhD Candidate**

A thesis submitted in partial fulfilment of the  
requirements of Liverpool John Moores University  
for the degree of Doctor of Philosophy

January 2019

# ABSTRACT

People today, in addition to their concerns about getting old and having to go through watching themselves grow weak and wrinkly, are facing an increasing fear of dementia. There are around 47 million people affected by dementia worldwide and the cost associated with providing them health and social care support is estimated to reach 2 trillion by 2030 which is almost equivalent to the 18th largest economy in the world. The most common form of dementia with the highest costs in health and social care is Alzheimer's disease, which gradually kills neurons and causes patients to lose loving memories, the ability to recognise family members, childhood memories, and even the ability to follow simple instructions.

Alzheimer's disease is irreversible, unstoppable and has no known cure. Besides being a calamity to affected patients, it is a great financial burden on health providers. Health care providers also face a challenge in diagnosing the disease as current methods used to diagnose Alzheimer's disease rely on manual evaluations of a patient's medical history and mental examinations such as the Mini-Mental State Examination. These diagnostic methods often give a false diagnosis and were designed to identify Alzheimer's after stage two when the part of all symptoms are evident.

The problem is that clinicians are unable to stop or control the progress of Alzheimer's disease, because of a lack of knowledge on the patterns that triggered the development of the disease. In this thesis, we explored and investigated Alzheimer's disease from a computational perspective to uncover different risk factors and present a strategic framework called Early

Prediction of Alzheimer's Disease Framework (EPADf) that would give a future prediction of early-onset Alzheimer's disease.

Following extensive background research that resulted in the formalisation of the framework concept, prediction approaches, and the concept of ranking the risk factors based on clinical instinct, knowledge and experience using mathematical reasoning, we carried out experiments to get further insight and investigate the disease further using machine learning models. In this study, we used machine learning models and conducted two classification experiments for early prediction of Alzheimer's disease, and one ranking experiment to rank its risk factors by importance. Besides these experiments, we also presented two logical approaches to search for patterns in an Alzheimer's dataset, and a ranking algorithm to rank Alzheimer's disease risk factors based on clinical evaluation.

For the classification experiments we used five different Machine Learning models; Random Forest (RF), Random Oracle Model (ROM), a hybrid model combined of Levenberg-Marquardt neural network and Random Forest, combined using Fischer discriminate analysis (H2), Linear Neural Networks (LNN), and Multi-layer Perceptron Model (MLP). These models were deployed on a de-identified multivariable patient's data, provided by the ADNI (Alzheimer's disease Neuroimaging Initiative), to illustrate the effective use of data analysis to investigate Alzheimer's disease biological and behavioural risk factors. We found that the continues enhancement of patient's data and the use of combined machine learning models can provide an early cost-effective prediction of Alzheimer's disease, and help in extracting insightful information on the risk factors of the disease. Based on this work and findings we have developed the strategic framework (EPADf) which is discussed in more depth in this thesis.

# DECLARATION

I, Mohamed Mahyoub, hereby declare that this Thesis, submitted to the Liverpool John Moores University as the fulfilment of the requirements for the Doctor of Philosophy has not been submitted to any other universities and institutes. I confirm that the work described in this Ph.D. thesis is my own except for some sources that support our research, which is appropriately cited and indicated.

Mohamed Adel Mamoon Mahyoub

January 2019

# ACKNOWLEDGEMENTS

All praise is due to God (Allah S.W.T) for granting me the ability, strength, and determination to complete this PhD, through the hardships and challenges faced over the past few years. Here, I would like to express my gratitude and appreciation to the many people who helped me with the work or provided me the motivation I needed to continue. I would like to thank my family, wife and my friends for their love, patience, and support during all this time.

I am profoundly grateful to my supervisors (Dr Martin Ronald, Dr Thar Baker, and Dr Po Yang) and Liverpool John Moores University for their support, encouragement and for assisting me throughout the PhD. I would like to take this opportunity to thank them for their patience and to appreciate their support and their advice that helped me overcome the challenges.

I am also deeply obliged to Prof Dhyia Al-jumeily, Prof Abir Husain, Dr Hissam Tawfik, and Dr Davide Bruno for giving me valuable feedback and discussions on ideas that eventually became a part of this thesis.

During my research, the biggest challenge was the collection and interpretation of ADNI databases. Here, I would like to express appreciation to Professor Danielle Harvey who helped me understand the ADNI dataset, and to both Dr Ibrahim Idowu and Dr Mohamed Khalaf for helping me during the debugging of the code to get better results.

While doing my PhD, I needed to gain an industrial experience with clinical data, so I took a post at Med eTrax as Database Developer for a hospital database. During my

employment, I gained experience, and got the support I needed. Here, I would like to express my heartfelt gratitude to Mr Rob Connell for encouragement, and the time he dedicated to me. He gave me enough time to attend my meetings at the university and to write my thesis, and for this I am very grateful.

### **Acknowledgment for Using ADNI Database**

The data used in this research was provided by Alzheimer's disease Neuroimaging Initiative (ADNI). We thank Professor Danielle J Harvey from the University of California, Davis, who provided insight and expertise that greatly assisted us in understanding and using the datasets. Data used in preparation of this article were obtained from the Alzheimer's disease Neuroimaging Initiative (ADNI) database ([adni.loni.usc.edu](http://adni.loni.usc.edu)). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this thesis.

Data collection and sharing for this project was funded by the Alzheimer's disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada.

## THESIS DEDICATION

*I dedicate this thesis to my parents, grandparents, my better half Nadia Altairy, my son dear Mohamed Mahyoub Jr, my family, and to the young people of Yemen who missed out on their education due to the civil war and being forced to become child soldiers.*

*Since the start of my PhD studies until today, and Yemen has been embroiled in an ongoing civil war. During this time, I have reminded myself of the forgotten child-soldiers of Yemen. Their pain, struggle, and hardship were the burning fire inside me that made me appreciate life and the opportunities I have. I pray for a better future for them, for a better tomorrow for Yemen, and a better world for us all, where no child is left behind or harmed. Indeed, a fruitful education is our only hope for peace.*

# TABLE OF CONTENTS

ABSTRACT.....	I
DECLARATION .....	III
ACKNOWLEDGEMENTS.....	IV
THESIS DEDICATION .....	VI
TABLE OF CONTENTS.....	VII
LIST OF FIGURES .....	X
LIST OF TABLES.....	XII
THESIS ABBREVIATION .....	XIV
LIST OF PUBLICATIONS .....	XVI
Chapter 1 Introduction.....	1
1.1 Introduction .....	1
1.2 Motivation .....	2
1.2.1 Causes of Alzheimer’s Disease.....	2
1.2.2 Symptoms of Alzheimer’s disease.....	3
1.2.3 Diagnosis of Alzheimer’s disease.....	4
1.3 Problem Statement.....	5
1.4 Scope of the Research.....	5
1.5 Aim and Objectives .....	6
1.6 Research Contributions.....	8
1.7 Thesis Structure .....	11
Chapter 2 Background.....	14
2.1 Introduction .....	14
2.2 Dementia.....	14
2.3 Alzheimer’s disease.....	16
2.4 Biological Development and Causes of Alzheimer’s Disease .....	17
2.5 Symptoms and Diagnoses of Alzheimer’s Disease .....	20
2.5.1 Mini Mental Score Examination (MMSE) .....	22
2.6 Risk Factors of Alzheimer’s disease .....	24
2.6.1 Genetics and Family History .....	27



2.6.2	Medical History .....	30
2.6.3	Lifestyle and Diet.....	36
2.6.4	Characteristics.....	42
2.6.5	High Risk Factors .....	44
2.6.6	Summary.....	46
Chapter 3	Literature Review.....	48
3.1	Introduction .....	48
3.2	An Overview on Data Science.....	48
3.3	Concept of Machine Learning.....	50
3.3.1	Machine Learning Approaches .....	52
3.4	Learning Models.....	57
3.4.1	Artificial Neural Network.....	58
3.4.2	Non-Artificial Neural Network.....	65
3.4.3	Pre-processing.....	72
3.4.4	Data Exploratory.....	76
3.4.5	Feature Selection.....	80
3.5	Other Related Work.....	83
3.6	Summary.....	87
Chapter 4	The Machine Learning Classification Framework .....	89
4.1	Introduction .....	89
4.2	Proposed Framework.....	90
4.3	Logical Approaches to Predict Alzheimer’s disease .....	93
Chapter 5	Implementation of the Framework.....	96
5.1	Introduction .....	96
5.2	Data Access and Collection.....	97
5.3	Pre-processing .....	101
5.4	Data Analysis and Visualisation.....	103
5.4.1	Phase 1: Initial Experiment Dataset.....	103
5.4.2	Phase 2: Extending the Dataset and Ranking Risk Factors .....	109
5.4.3	Phase 3: Over-Sampling and Classification of Extended Data.....	113
5.5	Development of Crowdsourcing Risk Factor Ranking System.....	118
5.5.1	Crowd Contribution .....	119
5.5.2	Crowdsourcing System Screenshots.....	120
5.5.3	Converting Input to Weights.....	122

5.5.4	Crowdsourcing System Overview .....	124
5.6	Deployment of Machine Learning Towards Early Prediction.....	125
5.6.1	Phase 1: Initial Experiment Models.....	125
5.6.2	Phase 2: Model Used to Rank Risk Factors by Importance.....	126
5.6.3	Phase 3: Final Classification Experiment Models .....	127
5.7	Summary.....	128
Chapter 6	Results.....	129
6.1	Introduction .....	129
6.2	Performance Evaluation Metrics .....	129
6.3	Phase 1: Initial Experiment Results.....	131
6.3.1	Initial Experiment Training Results.....	131
6.3.2	Initial Experiment Test Results.....	135
6.4	Phase 2: Rank Risk Factors by Importance Results .....	137
6.4.1	Ranking with Random Forest Model.....	137
6.4.2	Ranking with Neural Network & PCA .....	140
6.4.3	Ranking with Support Vector Machines.....	142
6.4.4	Ranking with Multi-Layer Perceptron.....	144
6.4.5	Combined discussion of Risk Factor's ranking .....	146
6.5	Phase 3: Final Classification Results.....	147
6.5.1	Final Classification Training Results.....	150
6.5.2	Final Classification Test Results.....	152
6.6	Discussion.....	154
6.7	Summary.....	157
Chapter 7	Conclusion and Future Work.....	159
7.1	Introduction .....	159
7.2	Conclusion.....	159
7.3	Future Work.....	164
References	.....	166

# LIST OF FIGURES

Figure 1-1 Research and work carried out in this thesis .....	7
Figure 1-2 Early Prediction of Alzheimer's Disease Framework (EPADf).....	11
Figure 2-1 The most common types of dementia in the UK [15].....	15
Figure 2-2 Structure of the brain [19].....	18
Figure 2-3: Plaques and Tangles in the Brain [21] .....	18
Figure 2-4 Identified Risk Factors of Alzheimer 's disease .....	24
Figure 2-5 - Connection between APOE and Bloodstream Diseases .....	29
Figure 2-6 Hypothetical Examples of Alzheimer's Progression.....	31
Figure 2-7 Visualisation of Table 2-2.....	35
Figure 2-8 Different areas of study for AD .....	45
Figure 3-1 Supervised learning workflow [74].....	54
Figure 3-2 Unsupervised learning workflow [75] .....	55
Figure 3-3 Semi-Supervised learning workflow [75] .....	56
Figure 3-4 Reinforcement learning workflow [76][75].....	57
Figure 3-5 Artificial Neural Network .....	59
Figure 3-6 Support Vector Machine [92].....	66
Figure 3-7 Pre-processing Stage in Machine Learning Process [107].....	72
Figure 3-8 Synthetic Minority Over-Sampling Technique (SMOTE) [116].....	75
Figure 3-9 Principle Component Analysis [122].....	78
Figure 4-1 Diagram of our framework (EPADf) .....	90
Figure 4-3b Example of Coherence Development Patterns.....	94
Figure 4-3a Example of Sequential Development Patterns .....	94
Figure 5-1 Content of ADNI Data .....	98
Figure 5-2 Initial Experiment ADNI Data Volume .....	105
Figure 5-3 Explore of Data Using t-SNE on.....	106
Figure 5-4 Explore of Data Using PCA on MATLAB .....	107
Figure 5-5 Explore of Data Using ICA on MatLab .....	107
Figure 5-6 Explore of Data Using SPE on Matlab .....	108
Figure 5-7: Pearson Covariance Method (Phase 2) .....	111
Figure 5-8 Phase 2: T-distributed Stochastic Neighbourhood Embedding .....	112
Figure 5-9 Part 3: Before Over-sampling .....	114

Figure 5-10 Using SMOTE in WEKA 3.6 (Phase 3).....	115
Figure 5-11 Explore of Data Using t-SNE in Matlab (Phase 3) .....	116
Figure 5-12 Part 3: Exploration of Data Using PCA in MatLab .....	117
Figure 5-13 Part3: Explore of Data Using SPE in Matlab.....	117
Figure 5-14 Created with NodeXL ( <a href="http://nodexl.codeplex.com">http://nodexl.codeplex.com</a> ).....	118
Figure 5-15 Crowdsourcing System Homepage .....	120
Figure 5-16 Crowdsourcing System Heart Disease Page .....	121
Figure 5-17 Crowed-sourcing System Adding New Contribution .....	121
Figure 5-18 Crowed-sourcing Risk Factors Interconnections .....	122
Figure 5-19 CRFR System Overview .....	124
Figure 6-1 Training results for 5 different classifiers on Matlab.....	133
Figure 6-2 ROC training results for each classifier on Matlab.....	133
Figure 6-3 Test results for 5 different classifiers on MATLAB .....	135
Figure 6-4 ROC Test results for each classifier on Matlab .....	136
Figure 6-5 Overall results of the variable importance using the Random Forest model .....	139
Figure 6-6 Overall results of the variable importance using the pcaNnet model .....	141
Figure 6-7 Overall results of the variable importance using the svmLinear model.....	143
Figure 6-8 Overall results of the variable importance using the MLP model .....	145
Figure 6-9 : Training AUC for all models (Phase 3) .....	150
Figure 6-10: Training ROC curve for all models (Phase 3).....	150
Figure 6-11 : Test AUC for all models (Phase 3) .....	152
Figure 6-12 : Test ROC curve for all models (Phase 3) .....	152
Figure 6-13 Research and work carried out in this thesis .....	158

# LIST OF TABLES

Table 2-1 Alzheimer's Disease Risk Factors Summary .....	25
Table 2-2 AD Medical History Risk Factors & their Relationship .....	33
Table 2-3 Alzheimer's disease Medical History Risk Factors & the Relationship to Diet .....	39
Table 2-4 AD Medical History Risk Factors & the Relationship to Lifestyle Activities .....	40
Table 2-5 Summary of Table 2-3 and Table 2-4.....	41
Table 3-1 Handling Categorical Values.....	74
Table 3-2 Relationship between X1 and X2.....	81
Table 3-3 Summary and Critical Evaluation of Related Work.....	84
Table 3-4 Literature Review Summary.....	87
Table 4-1 Explaining How EPADf Works .....	92
Table 5-1 Dataset of attributes .....	100
Table 5-2 AD Risk Factors used in the initial experiment.....	104
Table 5-3 Final Dataset for Initial Experiment .....	105
Table 5-4 PCA Coefficient for Each Variable .....	108
Table 5-5 Extended Dataset of attributes (Phase 2).....	109
Table 5-6: Levels of the Class Attribute (Phase 2).....	110
Table 5-7 AD Risk Factors used in extended dataset .....	113
Table 5-8 Levels of the class attribute (Phase 3) .....	114
Table 5-9 Classes Representation on Graphs and Plot .....	116
Table 5-10 Initial Experiment Models .....	125
Table 5-11 Model Used to Rank Risk Factors by Importance.....	126
Table 5-12 Final Classification Experiment Models .....	128
Table 6-1 Confusion Matrix.....	130
Table 6-2 Metrics Calculation .....	130
Table 6-3 Training Overall Results (Phase 1).....	134
Table 6-4 Variable importance using Random Forest model .....	139
Table 6-5 Variable importance using pcaNNet .....	142
Table 6-6 Variable importance using svmLinear.....	144
Table 6-7 Variable importance using MLP .....	146
Table 6-8 Graphs Keys Representation .....	147
Table 6-9 : Mean Performance for Models (Test) (Phase 3) .....	148

Table 6-10 : Mean Performance for Models (Training) (Phase 3) .....	148
Table 6-11: Models Performance for all Classes (Training) (Phase 3) .....	151
Table 6-12 - Part 3: Models Performance for all Classes (Test).....	153
Table 6-13 Information on all Experiments .....	157
Table 7-1 Objectives and Work Carried Out .....	161

# THESIS ABBREVIATION

<b>ABBREVIATION</b>	<b>MEANING</b>
<b>AD</b>	Alzheimer's disease
<b>ADNI</b>	Alzheimer's disease Neuroimaging Initiative
<b>ANN</b>	Artificial Neural Networks
<b>ANN</b>	Artificial Neural Network
<b>ARV</b>	Average Rectified Value
<b>AUC</b>	Area Under the Curve
<b>BMI</b>	Body Mass Index
<b>BN</b>	Bayesian Network
<b>BPXNC</b>	Back-Propagation trained Feed-Forward Neural Network Classifier
<b>CDP</b>	Coherence Development Patterns
<b>EMCI</b>	Early Mild Cognitive Impairment
<b>FFNN</b>	Feed-Forward Neural Network
<b>FLNN</b>	Functional Link Neural Network
<b>FN</b>	False Negative
<b>FP</b>	False Positive
<b>GP</b>	General Practitioner
<b>H2</b>	Fischer Discriminate Analysis
<b>ICA</b>	Independent Component Analysis
<b>LEVNN</b>	Levenberg-Marquardt Learning Algorithm
<b>LMCI</b>	Late Mild Cognitive Impairment
<b>MCI</b>	Mild Cognitive Impairment
<b>ML</b>	Machine Learning

<b>MMSE</b>	Mini Mental Score Examination
<b>MRI</b>	Magnetic Resonance Imaging
<b>NC</b>	Normal Control
<b>NHS</b>	National Health Service
<b>PCA</b>	Principal Component Analysis
<b>PD</b>	Parkinson's Disease
<b>PET</b>	Positron Emission Tomography
<b>RFC</b>	Random Forest Classifier
<b>RL</b>	Reinforcement Learning
<b>SDP</b>	Sequential Development Patterns
<b>SMC</b>	Significant Memory Concern
<b>SPE</b>	Square Prediction Error
<b>T-SNE</b>	t-Distributed Stochastic Neighbour Embedding
<b>TREEC</b>	Trainable Decision Tree Classifier



# LIST OF PUBLICATIONS

## Research papers:

**Mohamed Mahyoub**, Martin Randles, Thar Baker, Po Yang, ADNI, Effective Use of Data Science Toward Early Prediction of Alzheimer's disease, Accepted for publication in The 16th IEEE International Conference on Smart City (SmartCity-2018) , 2018

**Mohamed Mahyoub**, Martin Randles, Thar Baker, Po Yang, ADNI, Comparison Analysis of Machine Learning Algorithms to Rank Alzheimer's disease Risk Factors by Importance, Accepted for publication in Data Science and Internet of Everything at Developments in E-systems Engineering (DeSE), IEEE Society, Cambridge, 2018

**Mohamed Mahyoub**, Martin Randles, Thar Baker, Po Yang, ADNI, Utilising Machine Learning Classification Techniques To Investigate Early Prediction Of Alzheimer's Disease, Journal Paper -2019 (Under submission as a Special Issue on Multidisciplinary Sciences and Engineering for The 16th IEEE International Conference on Smart City)

**Mohamed Mahyoub**, Martin Randles, Thar Baker, Po Yang, ADNI, Using Behavioural and Biological Markers for Early Prediction of Alzheimer's disease, Conference Paper -2019 (In progress)

# Chapter 1    **Introduction**

## **1.1 Introduction**

The brain is central to being human; our biological memory defines our identity and character, and it's the place where we store our entire life log. Without our memory we would disconnect with everything around us. The brain, as well as being a memory is also a regulator of our internal organs, and controls our decision making. As we get older this vital part of our body becomes under threat of decaying. People today, in addition to their concerns about getting old and having to go through watching themselves grow weak and wrinkly, they also have a growing fear of developing dementia. There are around 47 million people affected by dementia worldwide and the cost associated with providing health and social care support is equivalent to the 18th largest economy in the world [1]. Memory loss due to dementia, and the failure to recognise our surroundings is a very terrifying experience. The most common form of dementia is Alzheimer's disease. Alzheimer's disease gradually kills brain cells and as a results of that patients end up losing loving memories, the ability to recognise family members, childhood memories, and even the ability to follow simple instructions e.g. making their usual morning cup of coffee, remembering how to use the toilet, and maintaining normal self-hygiene [2][3][4].

Currently the disease has no known cure and scientists are still unsure of what is the actual cause of the disease. Knowing the patterns that cause Alzheimer's disease to develop and having the possibility to predict the disease at a very early stage could help us either prevent the disease or slow its progress. This thesis explores and investigates the disease from a computational perspective to uncover different patterns and present a framework called Early Prediction of Alzheimer's Disease Framework (EPADf) that would give a probable prediction of the onset of Alzheimer's disease.

## **1.2 Motivation**

Alzheimer's disease is a slow and fatal disease that progresses gradually. The pathological development for Alzheimer's disease is complicated since the knowledge about its causes is limited, and the current diagnostic methods are either expensive or unreliable. In this section we discuss the motivation for investigating Alzheimer's disease, from three different angles, as follows:

### **1.2.1 Causes of Alzheimer's Disease**

Alzheimer's disease starts as two abnormal protein fragments called "Plaques" and "Tangles". These two protein fragments develop in the brain and gradually kill brain cells [5]. The very first stage of Alzheimer's disease is the existence of abnormal clusters of protein fragments that build up between nerve cells called "Plaques". Plaques will then surround healthy brain cells and cause them to die, which then turns into other twisted strands of another protein called "Tangles" [5][6][7].

Since the presence of the protein fragments “Plaques” and “Tangles” indicates the development of Alzheimer’s disease, what actually causes these fragments to develop is still undiscovered. However, it is apparent that Alzheimer’s disease is caused by many different types of risk factors and has different patterns from one patient to another. These risk factors are either biological markers or behavioural markers, and fall into different categories i.e. genetics, lifestyle, medical history, demography and characteristics. The ostensible risk factors of Alzheimer’s disease are mainly in genetics information and in the medical history of the patients. Majority of these risk factors are strong indicators used in the diagnosis of the disease. In term of genetics, only 1% of Alzheimer’s disease patients directly inherit the gene that causes a genetic mutation (APP, PS1 or PS2) and triggers the development of onset Alzheimer’s disease. While another gene called APOE4 which is in 10-15% of people, increases the risk of developing Alzheimer’s disease. However, approximately three quarters of Alzheimer’s disease patients have no family history of the disease and still developed the disease because of many other risk factors [8][9]. Chapter 2 of this thesis contains an in-depth discussion of Alzheimer’s disease including its risk factors, their pathological contribution and their interrelationships.

### **1.2.2 Symptoms of Alzheimer’s disease**

When it comes to the onset of Alzheimer’s disease it usually shares a lot of its symptoms with other types of dementia, which makes it difficult to give an accurate diagnosis. Early symptoms might include changes in mood and personality, poor judgement, and becoming forgetful. These early symptoms force the patient to withdraw from work and social activities, and as the disease progresses further, other severe symptoms start to develop such as challenges in

planning and solving problems, confusion with time and place, and misplacing things and losing the ability to retrace steps. The most common early sign is difficulty remembering new memories because Alzheimer's disease affects the hippocampus area of the brain that is responsible for storing new memories. Chapter 2 of this thesis contains an in-depth discussion of Alzheimer's disease including its symptoms and description of what it is like to live as a person with Alzheimer's disease.

### **1.2.3 Diagnosis of Alzheimer's disease**

The diagnosis of Alzheimer's at a very early stage is very complex and presently not possible. Currently, the way Alzheimer's is diagnosed is through a very careful medical evaluation of clinical assessments. The medical evaluation to diagnose Alzheimer's disease includes an examination of medical history, mental status testing, a physical and neurological exam and other tests such as brain imaging. It is a very complex process to diagnosis Alzheimer's disease, but its complexity also depends on what stage the disease is being diagnosed at; it is more difficult to diagnose people with Alzheimer's at an early stage as most of the symptoms are unclear [10][11] [12].

There are multiple diagnosis standards and methods from using cognitive examination to analysis of brain imaging data. The most common cognitive test used to diagnose Alzheimer's disease is the Mini Mental Score Examination, also commonly known as MMSE. In a cognitive test, if all questions were answered correctly, then patient is classed as normal [13]. However, if any patient scores low, then the patient will be considered to have memory problems, but this might not necessarily mean it is Alzheimer's disease. There are a lot of challenges to

accurately diagnose Alzheimer's disease in a patient, accurate diagnosis and prediction can help slow its progress or tackle the disease all together. Chapter 2 of this thesis discusses the challenges to diagnose Alzheimer's disease patients, and current diagnosis standards.

### **1.3 Problem Statement**

Alzheimer's disease is affecting the lives of over 47 million of people and costing health care providers worldwide approximately \$1 trillion dollar every year. It is irreversible, unpreventable with no known cure or available methods to predict its onset development. In addition to this, little is known about its root cause, and because shares most of its symptoms with other types of dementia it is often misdiagnosed. Early onset prediction of Alzheimer's disease will help to track its development and provide more insightful knowledge about its root causes.

### **1.4 Scope of the Research**

The study focused on providing a framework for the early prediction of Alzheimer's disease by deploying machine learning models on real patient data provided by ADNI (Alzheimer's disease Neuroimaging Initiative). Based on the research problem of this study, the framework will use both behavioural and biological markers data exclusive of brain imaging, to provide a better understanding of Alzheimer's disease risk factors. The key areas of research that the study will focus on are the ability to predicted Alzheimer's disease before it kills brain cells, the possibility of improving the accuracy of onset diagnosis of the disease, and to identify different pathological development patterns for the disease using behavioural and biological markers dataset aided with clinical evaluation.

If we understand the correlation between biological markers and behavioural markers such as lifestyle and demography we will understand more about the disease and its onset development. Collecting large amount of data related to behavioural and biological data of Alzheimer's disease can be impossible for humans to find hidden patterns and correlate the features without the use of computer science. Therefore, in this study, we will examine the use of machine learning models to investigate patient's data and extract knowledge that can be used toward early prediction and accurate diagnosis of Alzheimer's disease. This will be conducted by analysing Alzheimer's disease risk factors in the available patient data, as well as by computing the clinical knowledge from a manual evaluation.

### **1.5 Aim and Objectives**

The aim of this thesis is to provide a framework called Early Prediction of Alzheimer's Disease Framework (EPADf) that utilises machine learning toward early and cost-effective prediction of Alzheimer's disease using a combination of behavioural and biological markers. EPADf is an evolving framework that is capable of analysing Alzheimer's disease, combine machine learning and manual evaluation to weight Alzheimer's disease risk factors, and predict the disease from an early stage. This framework will further the understanding of Alzheimer's disease risk factors and help clinicians with decision making. To achieve the research aim, the following objectives have been undertaken:

- To gain an in-depth understanding of Alzheimer's disease risk factors and the available Alzheimer's disease datasets provided by ADNI, and extract the relevant data related to the research aim.

- To prepare a dataset by applying pre-processing, balancing and filtering techniques.
- To conduct exploratory data analysis for further understand the data and select the relevant features that would assist the early prediction of Alzheimer’s disease.
- To develop a crowd-sourcing system to collect manual evaluation of risk factors’ interrelationships from an epidemical and pathological prospective.
- To deploy machine learning models on the baseline data to produce automatic weighting for the risk factors based on their correlation.
- To propose a new framework to detect Alzheimer’s disease before it causes severe brain damage.
- To dissemination of research findings and outcomes in international specialised venues and events.

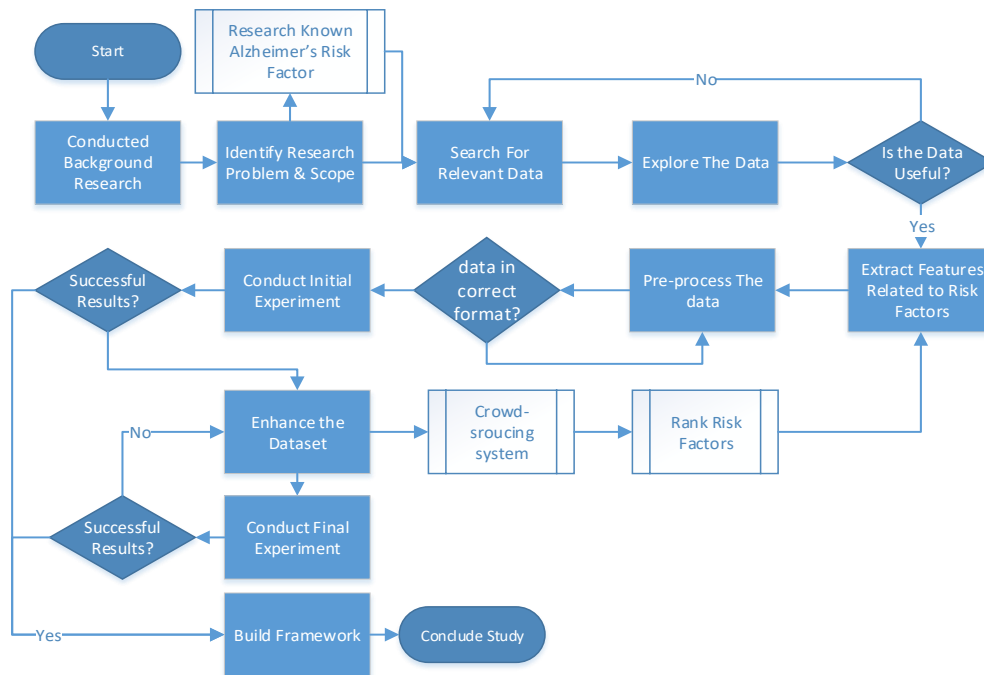


Figure 1-1 Research and work carried out in this thesis



The flowchart diagram illustrates a step by step of work conducted to complete this study and formulate the presented prediction framework (EPADf). The content of Chapter 4 and 5 will illustrate in-depth details on the proposed framework and the experiments carried out for this study.

## **1.6 Research Contributions**

This thesis proposes a new framework (EPADf) for early prediction of the onset of Alzheimer's disease to help clinicians with decision making and intervene before the disease fully develops and causes its victims to completely lose their living quality. Our approach provides a structured method to process, analyse and model Alzheimer's disease data using machine learning. On this basis there are other novel contributions which are discussed in turn in the following subsections.

The contributions of our work are detailed in the following list:

- Collection of extensive research related to understanding the fundamental requirements to investigate Alzheimer's disease using machine learning algorithms and the ranking of its risk factors. With this we also formulated logical approaches to illustrate how future research should be conducted.
- A baseline development for a strategic evolving framework (EPADf) for early prediction of Alzheimer's disease, which also lays a pathway towards early prediction and diagnosis of the disease. The framework is presented in Section (4.6), with all its sub-components which are also considered as further contributions.

- Solutions for skewed patient datasets to deal with limitations of imbalanced datasets to avoid biased results and overfittings. Methods such as under-sampling and over-sampling which were highlighted in Chapter 5, and Chapter 6.
- The development of a crowdsourcing risk factor ranking system (CRFR) as a subcomponent of our framework to collect clinical evolution of Alzheimer's risk factors and their contributors based on validated research and clinical understanding, and converting this knowledge to numerical sum using a mathematical algorithm presented in this study:
  - The idea is to produce a large fully connected network with risk factors and contributing factors as nodes of this network.
  - Based on the connections the CRFR system uses an algorithm to calculate the importance of each risk factor based on inwards and outwards influence for each risk factor. This was highlighted in section (5.5.1).
  - The CRFR system also has an easy to use platform where clinicians can input their experience and contribute to the knowledge confined in the network.
  - The produced ranking of this system provide a guideline for machine learning models to use during the learning process.
- Construction of a baseline dataset from the ADNI database and investigated the possibility of conducting early onset prediction of Alzheimer's disease from a machine learning prospective based on variables of risk factors that are not considered as late stage symptoms of Alzheimer's disease. This work illustrates the importance of underlying risk factors that is traditionally ignored.

- Defining and formalising the difference between the approach of using patient specific data (ADNI) and the approach of using clinical instinct and experience data (CRFR system) to investigate Alzheimer's disease.
- Evaluation of the performance of five different classification model algorithms, supported by two experiments, and compared against each other. The work also involved the following contributions:
  - Using five different architectures of machine learning classification models to investigate the possible prediction using only limited data in lifestyle, demography and common medical history.
  - Using different machine learning classification models to rank the risk factors variables in the dataset by importance. This presented an insightful knowledge about the dataset collected from ADNI, (see section 5.2).
  - Enhancement of the baseline dataset to include data from lifestyle, demography, medical history, genetics, and family dementia history. The work conducted here improved the classification process and set the path for future work following the footsteps for this thesis.

## Framework Overview

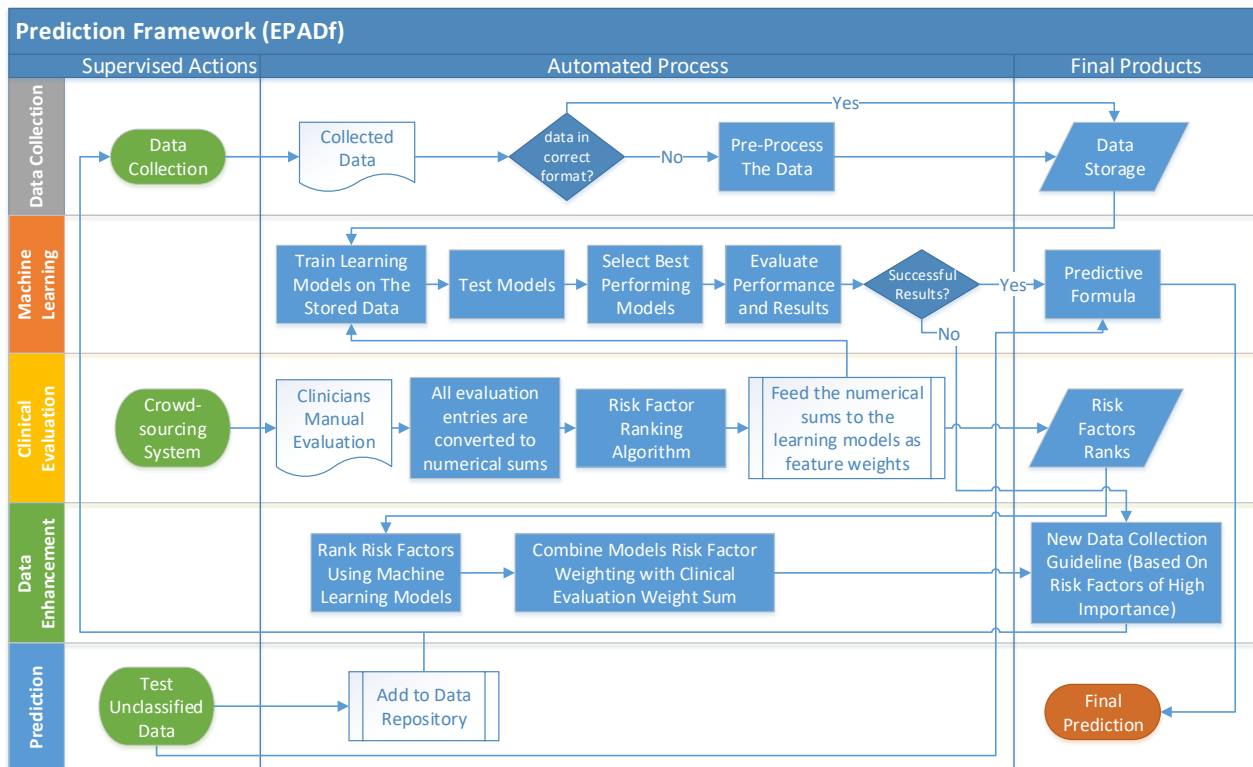


Figure 1-2 Early Prediction of Alzheimer's Disease Framework (EPADf)

The “Swimlane” diagram above shows a complete flowchart of the presented framework in this thesis, its subcomponent, and how it can be used for early prediction of Alzheimer’s disease.

### 1.7 Thesis Structure

This thesis is structured into 7 chapters giving an all-round view of the research problem, solution methodology, work carried out so far, and discussion of contribution and future work. Chapter 1 is an introduction section that delivers an overview of the research problem and scope of research. It highlights the negative impact of the Alzheimer’s disease burden on healthcare providers and the economy in general. The chapter argues the need to utilise

machine learning technology to predict the disease from an early stage to help clinicians with decision making. Predicting the disease from an early stage before it aggressively progresses and cause severe brain damage, would encourage clinicians and patients to work together to take actions that would either prevent it or slow down its progress, which in return will enable people to live longer and more independently. In this chapter the proposed predictive methodology and framework are presented with a descriptive content of aims and objectives, and a discussion of novel contributions claimed in this thesis.

A comprehensive content of research background is assembled in Chapter 2, which provides an in-depth understanding of dementia and with concentrated focus on Alzheimer's disease. The content discusses Alzheimer's disease risk factors, and their correlations and contribution to each other's development. It also discusses Alzheimer's disease symptoms and their impact on patients' lives, and the current methods used for diagnosis and treatment. Additionally, the discussion in this chapter is also extended to highlight challenges in this field, and the financial costs associated with Alzheimer's disease health and social care and its impact on the economy. Overall this chapter gives a wide understanding of dementia, specifically Alzheimer's disease and the need for decision support solutions to tackle the challenges.

Chapter 3 in this thesis discusses the science field of artificial intelligence, specifically its subset, field machine learning, and how it is used to aid and accelerate research in the interdisciplinary field of bioinformatics. The content of this chapter provides an in-depth understanding of machine learning and discusses and compares the capabilities, algorithms, and concepts of learning models used in this thesis. The modules used include different artificial neural network architectures, random forest, support vector machine, Naïve Bayes,

and decision trees. This chapter also includes discussion of related work to the utilization of machine learning to investigate Alzheimer's disease.

A comprehensive discussion of the proposed framework and its implementations are provided in Chapter 4 and 5, which includes the core methods and data pipeline modules used to provide an early prediction of Alzheimer's disease. The implementation chapter discusses the data pre-processing and the methods used for the processing, extraction, and selection of relevant features to the aim of this research, as well as the oversampling techniques used to balance the data. The implementation chapter also provides an exploration of the data sample and discusses the content and dimensions of the data. In general, these two chapters gives a detailed summary of the framework, the experimental set-up, and the machine learning techniques used to conduct the prediction of onset Alzheimer's and to systematically investigate the dataset.

To conclude the findings and proof of work, chapter 6 provides a comprehensive report of key experiments' results and further discusses the performance of each learning model employed in the work carried out in this thesis. This includes prediction results, model's accuracy calculation, the interpretation of the results, and a detailed summary of the contributions made in this thesis.

Chapter 7 of this thesis provides an overall discussion of results, a summary of the work carried out, and a conclusion of the content and contribution made in the thesis, with a final discussion of future work.

# Chapter 2     **Background**

## **2.1 Introduction**

Getting old and weak is something most people find hard to accept. It's a struggle that mature people over the age of 50 are facing but the concern doubles with the fear of losing their memory due to dementia. Elderly people who are affected by dementia are living the experience of watching themselves die slowly, fade away from their world, live in constant confusion, and no longer able to understand their surroundings. It is a horrible experience to be endured for dementia victims, their carers, and their families[14]. In this chapter we conducted an extensive background research on one of the most common types of dementia; Alzheimer's disease. We discussed Alzheimer's disease from different angles, covering its pathological development, the types of risk factors, symptoms, and the current diagnosis methods.

## **2.2 Dementia**

Dementia is becoming the challenge of our century, although the word itself does not refer to a specific disease, it is a term used to group several cognitive impairment diseases such as Alzheimer's disease, Parkinson disease, and many other types of dementia diseases. In a report published by Alzheimer's Society [15], discussed and listed the most common types of dementia in the UK. Figure 2-1 below shows the different types of dementia and percentage of people who are suffering from these diseases.

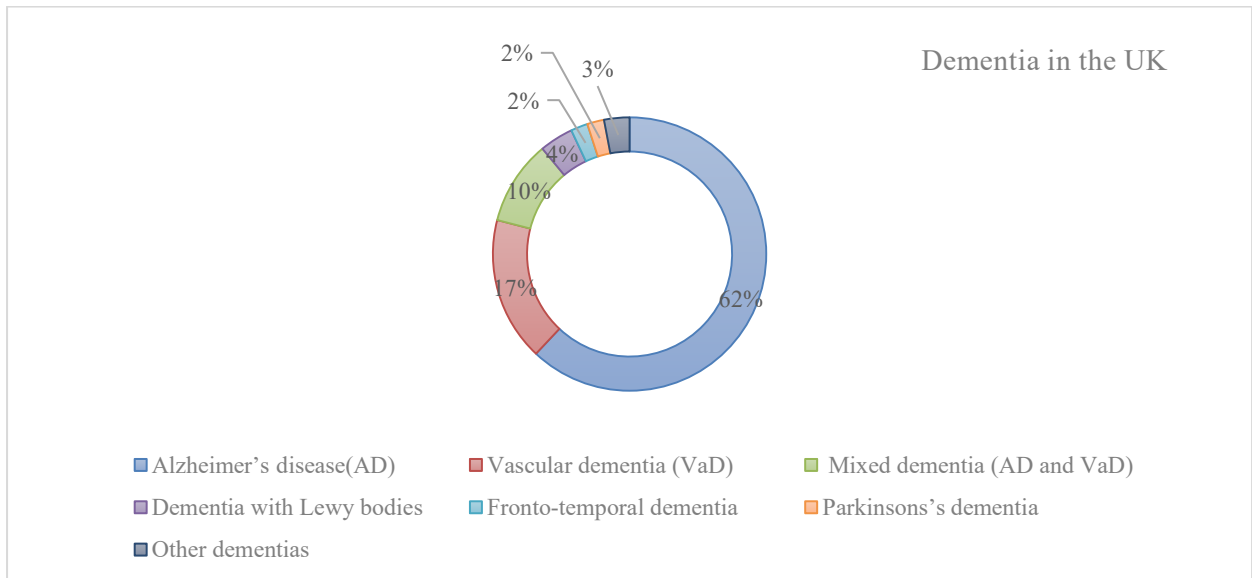


Figure 2-1 The most common types of dementia in the UK [15]

According to the World Alzheimer Report 2015 by Alzheimer's disease International; the estimated financial cost for dementia is 1 trillion dollars and will double to 2 trillion dollars by 2030. There are around 47 million people affected by dementia worldwide. The overall cost of the disease is higher than the market value of both Apple and Google combined. The cost associate with providing health and social care for dementia is equivalent to the 18<sup>th</sup> largest economy in the world. This comparison helps us to comprehend the massive impact of the disease on the economy [1].

This chapter of the thesis discusses the most common types of dementia i.e. Alzheimer's disease and the progress currently made in fighting the disease. It will highlight the effects of Alzheimer's disease on the economy and people, from both the social and biological perspectives, its symptoms, how it's currently diagnosed, and the work done to fight the disease.



## 2.3 Alzheimer's disease

Alzheimer's disease is the most common form of dementia with no known cure, it is a slow and fatal disease that affects the human brain. This disease develops in different stages, which can cause a full destruction in the functionality of the brain over time. Alzheimer's patients usually experience memory loss, difficulty in focusing and a struggle to learn. The disease can also cause changes in personality and affected patients suffers from many side effects, such as anger and depression. As a person's condition declines, they often withdraw from family and society. Gradually, bodily functions are lost, ultimately leading to death. Although the speed of progression can vary, the average life expectancy following diagnosis is approximately seven years. Fewer than 3% of individuals live more than 14 years after diagnosis.

Having Alzheimer's disease means losing loving memories, the ability to recognise family members, and childhood memories, or even the ability to follow simple instructions e.g. making their usual morning cup of coffee, remembering how to use toilet, and maintain self-hygiene [2][3][4]. Statistics show that around 1 in 10 people over the age of 65 have been affected by Alzheimer's. Unfortunately, there are no effective cures for this disease and no one is immune [16].

The reason Alzheimer's was chosen to be the disease to investigate, is that over 62% of dementia patients have this disease, and there are no current systems that can predict the disease from an early stage. Most of the work related to this proposed research is mainly focused on the diagnosis of Alzheimer's disease from a short term and biological perspective. Only few research works have been carried out to predict Alzheimer's disease before the clinical

diagnosis. The challenge is to look for the most accurate way to predict Alzheimer's disease at a very early stage before patients develop any of the symptoms. The success of such a challenge will help improve the research into "Early Treatments" and diagnosis [12].

## **2.4 Biological Development and Causes of Alzheimer's Disease**

To understand this fatal brain disease, one must understand how the human brain functions. The human brain consists of complex chemical and electrical processes to run the body functions. The brain is formed from billions of cells call neurons as shown in Figure 2-2. These electrically excitable cells communicate and transmit information with one another, through electrical and chemical signals. The brain is made up of an enormous number of neural networks, connected together to form a vastly large core component of the nervous system[17].

There are multiple types of neurons in the human body that are responsible for transmitting data. The brain cells or the neurons in the brain are responsible for storing and transmitting data. Dementia diseases generally affect the neurons in the brain and cause communication disorder in the neural networks.

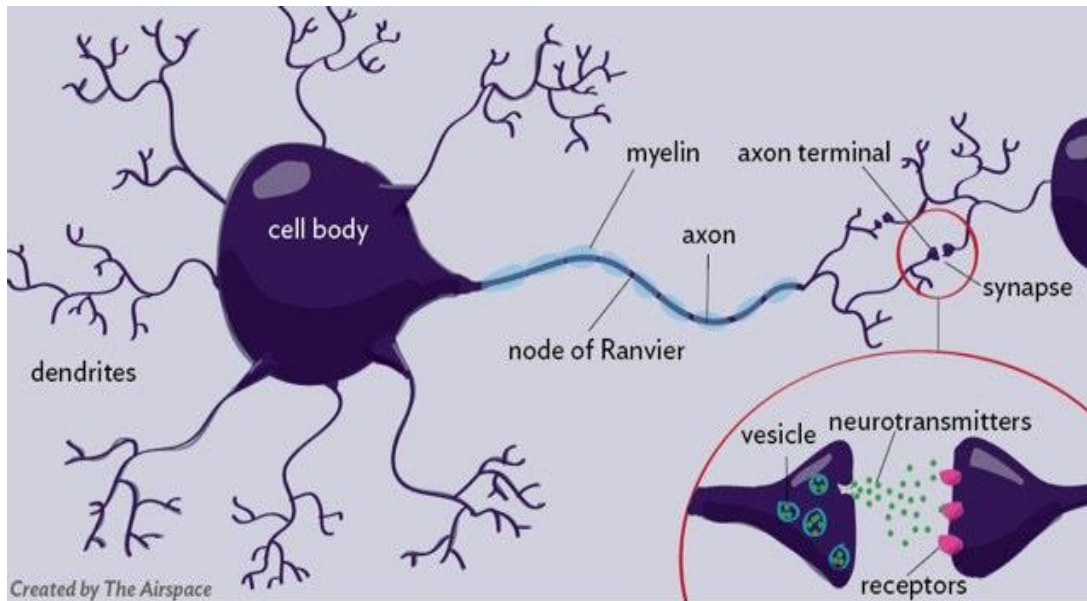
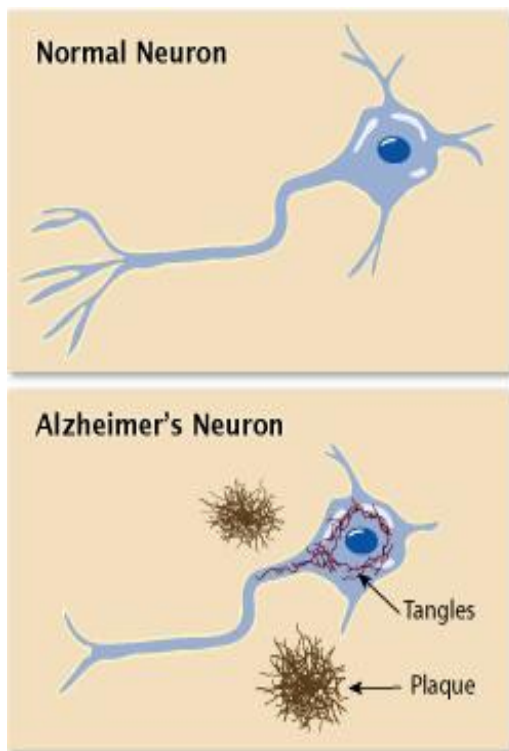


Figure 2-2 Structure of the brain [19]



Alzheimer's disease is a slow and fatal disease that develops gradually. It starts as two abnormal protein fragments called "Plaques" and "Tangles" (see Figure 2-3). These two protein fragments develop in the brain and gradually kill brain cells [5]. The very first stage of Alzheimer's disease is the existence of abnormal clusters of protein fragments that build up between nerve cells called "Plaques". Plaques then surround healthy brain cells and cause them to die, which then turns into other twisted strands of another protein called "Tangles" [5][6][7].

Figure 2-3: Plaques and Tangles in the Brain [21]

There is not enough evidence to backup the claim that Plaques and Tangles are the main causes for brain cells to die. However, according to the latest research Plaques and Tangles are the main suspects. The abnormal Plaque protein fragments are formed from small protein pieces called beta-amyloid that are clumped together. These protein pieces are usually found in the fatty membrane surrounding nerve cells.

Once both Plaques and Tangles have fully appeared in the brain they will start developing in the “hippocampus”; the part of the brain where human memory is first formed. Over time the "Plaques" and "Tangles" will start killing the cells in the “hippocampus” which will make it harder and harder for patients to develop new memory, specifically short-term memories of things that happened a few hours or days ago.

After the destruction of the hippocampus more Plaques and Tangles will spread to different parts of the brain, slowly and gradually kill brain cells wherever they go across the brain in different regions. This spread is what causes the different stages of Alzheimer’s disease. As the first region to be affected is the hippocampus, the patient starts to lose short term memories. The disease then spread, to the second region where language is developed, which causes patients to find it difficult to speak or make sentences.

After a few years the disease spreads to the front of the brain where logical thoughts are developed, and patients will struggle to plan tasks or solve problems. Gradually the disease will then spread to the region of the brain where emotions are regulated. When the disease affects the emotions region patients will experience constant changes of mood. When the emotions regulator region is fully affected, the "Plaques" and "Tangles" will then spread across

to the sense region, where understanding of surroundings is developed. Patients will struggle to analyse or make sense of things surrounding them and will experience hallucinations.

Near the end the disease will spread to the brain region where old memories are stored, and it will cause memory loss of the oldest memories of the person. After this stage and at the very end the disease will spread to the very last part of the brain where the heart and breathing balance is controlled. The patient will then gradually lose balance and coordination of their heart and stop breathing. Alzheimer's disease has no cure to this date and whole process of the disease is very slow and could take up to 8 to 10 years before the patient loses their life.

## **2.5 Symptoms and Diagnoses of Alzheimer's Disease**

The most frequent early symptom is having difficulty in remembering recent events [10]. As the disease advances, symptoms can include confusion, irritability, aggression, mood swings, changes in personality, anger, trouble with language, and long-term memory loss [3]. This disease develops in different stages, which can over time cause a full destruction in the functionality of the brain.

The diagnosis of Alzheimer's at a very early stage is very complex and presently not possible. Currently, the way Alzheimer's is diagnosed, is through a very careful medical evaluation. The medical evaluation to diagnose Alzheimer's disease includes an examination of medical history, mental status testing, physical and neurological tests, and other tests such as brain imaging. It is a very complex process to diagnosis Alzheimer's disease, but its complexity also depends on what stage the disease is being diagnosed at; it is more difficult to diagnose people with Alzheimer's at an early stage as most of the symptoms are unclear [10][16][12].

There are 10 main common signs that are used to identify Alzheimer's disease. According to the Alzheimer's association ([www.alz.org](http://www.alz.org)), the 10 suggested signs are as follows [14] [3] [18]:

- Memory loss that disrupts daily life
- Challenges in planning or solving problems
- Difficulty completing familiar tasks at home, at work or at leisure
- Confusion with time or place
- Trouble understanding visual images and spatial relationships
- New problems with words in speaking or writing.
- Misplacing things and losing the ability to retrace steps
- Decreased or poor judgment
- Withdrawal from work or social activities
- Changes in mood and personality

If a person has the majority of these symptoms they are believed to have Alzheimer's disease.

If a General Practitioner (GP) suspects that the patient may have Alzheimer's, then the patient will be asked to take the Mini Mental Score Examination test on a regular basis to confirm the diagnosis.

We have contacted the Department of Health to confirm if there are any Government-led programs to help diagnose dementia at a very early stage, such as a national examination for dementia for people over age 65, similar to the breast cancer program [19]. It was confirmed that there are no such government-led programs for early detection of Alzheimer's and dementia, and only in the middle of 2014 the NHS introduced "Dementia Enhanced Service

for GPs ” to detect and provide timely diagnosis of the disease [20]. Therefore, if a person has the majority of these symptoms, they are believed to have Alzheimer’s disease, or if a general practitioner (GP) suspects that the patient may have Alzheimer’s then the patient will be asked to take the Mini Mental Score Examination (MMSE) test on a regular basis to confirm the diagnosis. Alongside the Mini Mental Score Examination, the Alzheimer’s Association have released a paper “New Diagnostic Criteria and Guidelines for Alzheimer’s disease” to diagnose the disease at an earlier stage [21].

The usual way to diagnose Alzheimer’s disease, is firstly done by noticing some of its symptoms and when the disease advances the symptoms are easier to detect. If the symptoms of Alzheimer’s are clear and can easily be detected the patient is recommended to undertake several tests. Sometimes if the symptoms are unclear or abnormal doctors might refer patients to take a brain imaging test in order to confirm the existence of the Alzheimer’s disease and to differentiate it from other types of dementia or other brain problems [6] [12][13]. This work will explore the current diagnosis methods and the existing data collected from clinical trials during these diagnostic sessions. The novelty of our work will focus on using this data for early prediction of Alzheimer’s disease pre-diagnosis stage (see area 1 in Figure 2-8).

### **2.5.1 Mini Mental Score Examination (MMSE)**

Currently, doctors use a paper-based mental examination test to diagnose Alzheimer’s patients. The Mini Mental Score Examination (MMSE) is a series of questions and tests, each question is evaluated with points scored given based on the answers given. If all questions were answered correctly, the maximum score is 30 points. However, any score below 27 the patient

will be considered to have memory problems but not necessarily dementia. Below are the 6 different types of questions and tests included in the MMSE test [13]:

- **Orientation** – Such as identifying current location, time and date
- **Registration** – Repeat and learn individual words
- **Attention and Calculation** – Spell words backward or systematic mathematical equations
- **Recall** – Remember individual words learned previously during the Registration stage
- **Language** – Name different objects, repeat sentences, follow instructions and write sentences
- **Copying** – Draw simple shapes such as pentagons [22]

This test is commonly used for complaints of memory problems. However, as one of many other tests, MMSE test is also used by doctors to diagnosis Alzheimer's disease and other types of dementia [13].



## 2.6 Risk Factors of Alzheimer’s disease

Although, there are many opinions stating that Alzheimer’s disease is a heredity based disease, other research shows that many Alzheimer’s disease cases could be prevented by lifestyle changes such as exercise, eating healthily and not smoking. Scientists might not agree but the results of their research could potentially mean that they are right, and that Alzheimer’s disease can develop from lifestyle or could simply be a genetic disease.



Figure 2-4 Identified Risk Factors of Alzheimer’s disease

Figure 2-4 shows the currently recognised risk factors of Alzheimer’s disease, details of these risk factors and references to related studies are included in the following table (Table 2-1) :

Table 2-1 Alzheimer's Disease Risk Factors Summary

<b>Risk Factor</b>	<b>Description</b>	<b>Category</b>	<b>Citation</b>
<b>AGE</b>	In almost all cases symptoms of Alzheimer's start to show from the age of 50+.	Characteristic (Demography)	[23]
<b>APP, PS1, and PS2</b>	These three genes have been identified as causative genes of Alzheimer's Disease.	Genetics	[24]
<b>APOE</b>	Apolipoprotein E (APOE) gene increases a person's risk of developing Alzheimer's disease.	Genetics	[23][24]
<b>DIABETES</b>	A known cardiovascular risk factors is type 2 diabetes, it increases the risk of Alzheimer's disease in mid-life or later life.	Medical History	[23][25]
<b>OBESITY</b>	Obesity is one of the cardiovascular risk factors that increases the risk of Alzheimer's disease in mid-life.	Medical History	[23]
<b>STROKE</b>	Stroke is related to almost all of the cardiovascular disease that are considered to be high risk factors of Alzheimer's disease and dementia in general.	Medical History	[23][25][26]
<b>DEPRESSION</b>	People with history of depression in mid-life or later life have shown to have increased rates of dementia.	Medical History	[23] [27]
<b>HIV INFECTION</b>	People with HIV sometimes develop cognitive impairment.	Medical History	[23]
<b>DOWN'S SYNDROME</b>	Down's syndrome carries a gene that produces one of the key proteins (APP gene – Amyloid precursor protein) which is a causative gene of Alzheimer's Disease.	Medical History	[23]
<b>CHOLESTEROL</b>	Cholesterol is a fatty substance, which, causes the development of Alzheimer's Disease risk factors such as diabetes, high blood pressure, and other cardiovascular disease.	Medical History	[23]

<b>HEART DISEASE</b>	Heart disease shares ApoE as a genetic link with Alzheimer's disease, and it is also a vascular risk factor to onset of Alzheimer's.	Medical History	[25]
<b>HEAD TRAUMA</b>	A severe blow to the head increases the risk of later dementia such as Alzheimer's disease.	Medical History	[23]
<b>BLOOD PRESSURE</b>	High blood pressure is a risk factor of dementia, and beside this blood pressure can cause strokes and strokes are risk factors of dementia.	Medical History	[23]
<b>STRESS</b>	Stress affects the immune system, which is known to play an important role in the development of dementia.	Medical History	[28][29][30]
<b>POOR DIET</b>	An unhealthy diet can affect a person's risk of developing dementia and cardiovascular disease such as type 2 diabetes.	Lifestyle	[23][27]
<b>SUBSTANCES</b>	Drug abuse have been suggested as possible risk factor of dementia.	Lifestyle	[23][27]
<b>ALCOHOL</b>	Heavy and chronic drinking results in specific dementia-type symptoms.	Lifestyle	[23][27]
<b>LAZINESS (PHYSICAL INACTIVITY)</b>	Mid-life physical inactivity increases the risk of all-cause dementia.	Lifestyle	[23][27]
<b>SMOKING</b>	Smoking increases the risk of developing dementia, especially Alzheimer's disease.	Lifestyle	[23] [27]
<b>ALUMINIUM</b>	Research on Aluminium concentrations in water showed that he risk of Alzheimer's disease was 1.5 times higher in areas where the aluminium concentration exceeded 0.11 mg/l than in areas where concentrations were less than 0.01 mg/l.	Lifestyle	[31]
<b>LOW SOCIAL ACTIVITY</b>	Very few studies report the long-term effect of mid-life social isolation or loneliness on risk of dementia in to older age.	Lifestyle	[23] [27]

<b>LOW MENTAL ACTIVITY</b>	Mental activities in mid-life are associated with a lower risk of dementia in later life.	Lifestyle	[27][32][33]
<b>FEMALE</b>	Women are more likely to develop Alzheimer's disease than men.	Characteristic (Demography)	[23]
<b>ETHNICITY</b>	People from certain ethnicities are at higher risk of dementia than others.	Characteristic (Demography)	[23]
<b>NO EDUCATION</b>	Many research studies have associated lower education with a greater risk for dementia. Suggesting that the effect of education on risk for dementia may be best evaluated within the context of a lifespan developmental model.	Characteristic (Demography)	[23][34]
<b>EMPLOYMENT</b>	Low level of job control is associated with higher multivariate adjusted risk of dementia.	Characteristic (Demography)	[35]

## 2.6.1 Genetics and Family History

There are approximately 70 trillion cells in the human body, and the number varies from one to another. As people grow older or gain more weight, the number of cells increases and vice versa. Not all of these cells serve the same purpose; they fall into different categories such as blood cells, brain cells, tissues cells and other types of cells. Each type of cells has their own size, shape and responsibility to manage a particular function in the body. Each individual cell has a set of DNA sequence that shapes its amino acid and proteins, which determines how the cells would function. The human DNA is a sequence of a genetic code merged during fertilization, and it determines everything about a person, including how they going to look, and the genetic diseases that they might have in the future. Any changes to the genetic code of the DNA will mean that the functions of the cells will also change. Lifestyle and interaction

with other biological environments could cause slight changes in the function of some of our body cells; hence, why we get ill and catch different diseases. Investigating this field of the human body and its relation to Alzheimer's disease, will provide us with a solid knowledge about the disease.

There are multiple proteins (amino acids) in the human cells, which determine the shape and functionality of the cell. Each protein has its own sequence of genetic code (mRNA sequence) that forms its set of instructions on how the protein should work, any changes to this code will affect the tenacity of the protein and functionality of the cell. Proteins and amino acids are the visualisation of genetic codes or mRNA sequences, although, it is complicated to explain how they function exactly but perhaps if we compare it to computer code the analogy would be; genetic code and proteins work similarly to the front-end and the back-end of a software package.

The study of a particular cell will first start with the extraction of its DNA sequence (mRNA), then analyse the sequence to identify the different gene sets and amino acids. The genes set in the DNA sequence have already been named and previously defined by scientists. The genetic sequence helps scientists understand the behaviour and functionality of the cell. Abnormal activities or changes in the nature of the cell are also identified through its genetic code [36].

In the case of Alzheimer's disease, there are at least three genes that have been identified by scientists as causative genes. Mutations in these three genes, APP, PS1, and PS2 have been established to cause mainly early-onset Alzheimer's disease. If these three genes are found in the DNA variations sequence it means that the person who is carrying these genes will

definitely develop Alzheimer's disease in the future [24]. However, carrying these genes is a very rare occurrence. These three genes are not classified as risk factors of Alzheimer's disease, but more as hallmarks of the disease. There are several other risk factor genes that have been identified as contributors to the development of Alzheimer's disease, but the main contributor is the Apolipoprotein E gene (APOE), which has been identified as the highest risk factor gene of Alzheimer's disease [24]. Environmental risk factors can cause changes in the structure of a gene, resulting in an abnormal form that may be conveyed to succeeding generations.

The APOE gene is responsible for providing instructions for making APOE protein; this protein is a blood protein that carries Cholesterols and other fats through the bloodstream. High cholesterols are also believed to be one of the risk factors for the development of Alzheimer's disease. Maintaining cholesterol levels is essential

for the prevention of disorders such as high blood pressure, heart disease and strokes, which are also Alzheimer's disease risk factors. From Figure 2-5, we learn that APOE is an important risk factor as it has roots in other high and low risk factors and that increases the level of its risk frequency. However, if we study the effects of APOE in reverse, APOE is only responsible for carrying cholesterols and other fats; in this case, APOE is not the main risk factor in this

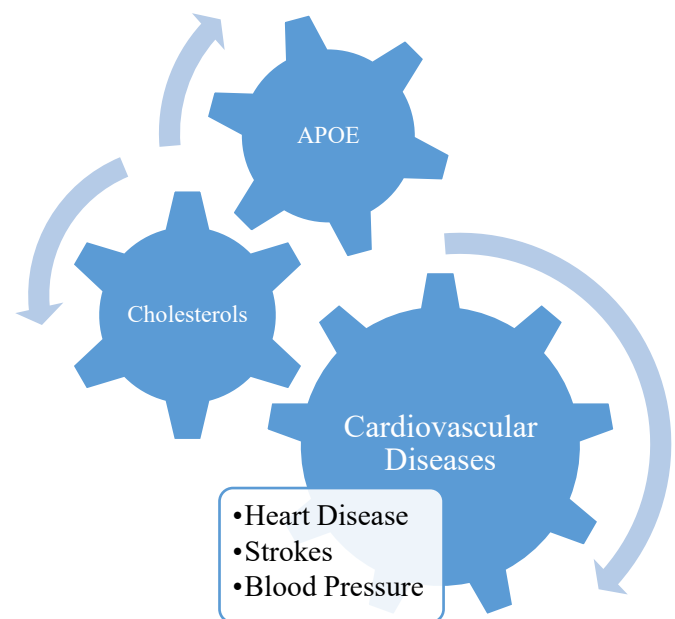


Figure 2-5 - Connection between APOE and Bloodstream Diseases

chain of risk factors. The main risk factor in this process is the source of fats and high cholesterol.

The APOE gene helps to facilitate the development of Alzheimer's disease but having this gene does not confirm the definite development of the disease. It is important to acknowledge that having this type of gene means there is a high risk of developing Alzheimer's disease if the lifestyle, body fat, and medical state are not managed properly.

## **2.6.2 Medical History**

One of the major aspects that doctors use in the evaluation process of diagnosing any disease is the medical record of the patient, as most diseases in a way can relate to each other or progress from one state to another state and become a disease in a new form or become more chronic. For example, high blood pressure leads to stroke, or an untreated premalignant condition will eventually turn into cancer. Therefore, the connection between Alzheimer's disease and other diseases or illnesses is an immense possibility.

When it comes to Alzheimer's disease, medical records that relate to heart and bloodstream diseases can significantly increase the risk of developing Alzheimer's disease or any other type of dementia. Scientists believe that medical history plays a vital part in the development of Alzheimer's disease. The majority of Alzheimer's disease patients share a medical record similar to other patients, for instance, many Alzheimer's disease patients have had blood circulation problems. Therefore, the risk factors of Alzheimer's disease in medical history can be entrenched at different degrees of disorders. In other words, a patient might develop

Alzheimer's disease if they have previously had high blood pressure followed by high cholesterol, or perhaps had both stress and diabetes. These are just hypothetical examples to demonstrate how medical history could contribute to the development of Alzheimer's disease.

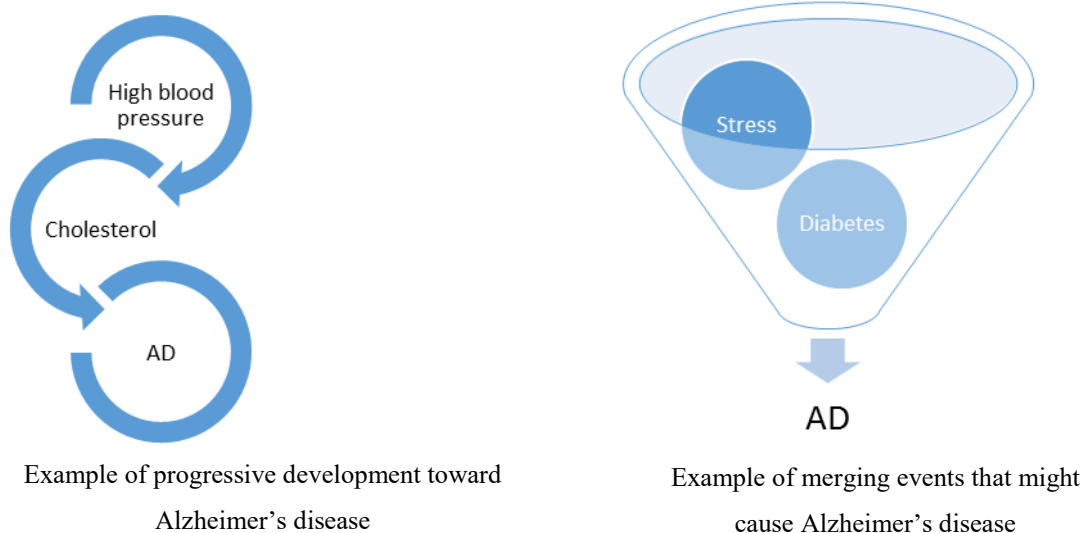


Figure 2-6 Hypothetical Examples of Alzheimer's Progression

Even though many Alzheimer's disease risk factors have been identified, the development of the disease remains a mystery. Alzheimer's disease develops over time through a complex series of brain changes that occur throughout the lifetime of the patient. These changes in the brain can be caused through diseases, genetic heredity, interaction with environment or bad and poor lifestyle. Identifying or categorising the level of risk to these causes is very difficult and complex as each cause can differ from person to person.



As there are several diseases, genes and other aspects that could be categorised as risk factors for Alzheimer's disease, it is a very complicated process to identify the causal risk factors of Alzheimer's disease. However, there are five types of risk factors; age, genetics, medical history, lifestyle and characteristics, as shown in Figure 4. It might be less difficult to point out which type of risk factor has the major impact, but it is a complex process to identify the relation between each type and the connection between all of the risk factors. What makes the disease far more complex is that it can develop from multiple causes and not just in one way. Therefore, it is important to use computer science in order to analyse Alzheimer's disease patient data to identify causal patterns in the risk factors and the relationship between them. These patterns are almost impossible to work out manually, therefore, the use of technical machine learning concepts will speed the learning on Alzheimer's disease and provide more information about its development.

Age and Genetics are the clearest risk factors. However, research suggests that a variation of medical history beyond genetics may play a role in the development of Alzheimer's disease. This has sparked a vast amount of research and become of great interest to many health providers and institutions. Below is a table of the diseases that are known as risk factor of Alzheimer's disease and their relationship to one another, which shows how they contributes to each other's development.

Table 2-2 AD Medical History Risk Factors & their Relationship

Risk Factor Diseases	Relationship (1)	Relationship (2)	Relationship (3)
1. Diabetes	Obesity [37]	Cholesterol [38]	Heart Disease [39]
2. Obesity	Cholesterol [40]	Depression [41]	Chronic Stress [42]
3. Stroke	Blood Pressure [43]	Head Trauma [44]	Chronic Stress [45]
4. Depression	Obesity [46]	Chronic Stress [47]	Stroke [48]
5. HIV Infection	Others	Others	Others
6. Down's Syndrome	Genetic	Genetic	Genetic
7. Cholesterol	Obesity [40]	Heart Disease [49]	Blood Pressure [50]
8. Heart Disease	Obesity [51]	Depression [52]	Blood Pressure [53]
9. Head Trauma	Chronic stress [54]	Stroke [44]	Blood Pressure [55]
10. Blood Pressure	Obesity [51]	Diabetes [56]	Chronic Stress [57]
11. Chronic Stress	Lifestyle	Lifestyle	Lifestyle

**Note:** Please note Table 2-2 is a results of basic manual analysis, and the relationship between the diseases is stated based on researched from multiple sources (as cited next to each relationship).

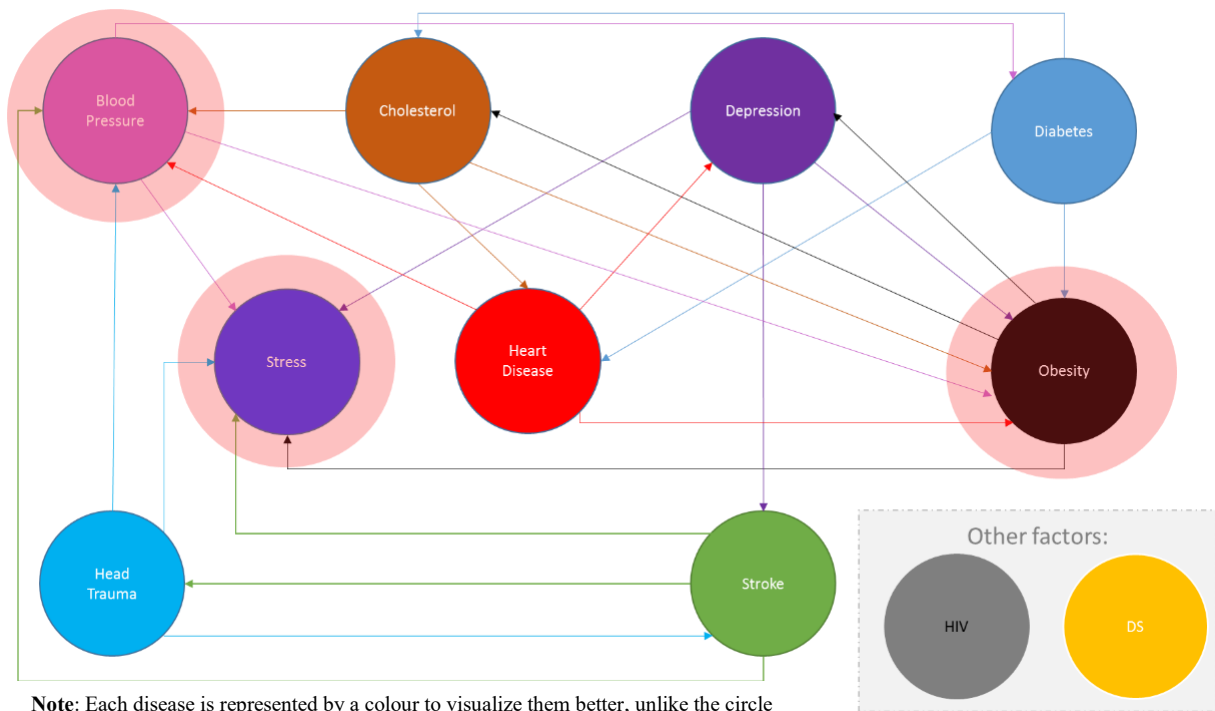
Biologists suggest that Alzheimer's disease is correlated to blood circulation and changes in the brain. Table 2-2 shows a preliminary demonstration of the correlation between the illnesses that are suggested to be Alzheimer's disease risk factors. For each illness, the top three possible illnesses that might be its initial cause, were selected and put into three new columns as shown in Table 2-2. Although, this is a preliminary analysis, the outcome has raised many questions, as well as a possible backing up some of the statements made by other researchers. For example, diabetes is classed as common risk factor for Alzheimer's disease, although, diabetes was only brought up once in Table 2-2 but this disease relates to the most significant diseases in the relationship columns in Table 2-2. Chronic stress, obesity, blood pressure and heart

disease are the most significant risk factors in Table 2-2, therefore, could these risk factors have high importance for increasing early risks of Alzheimer's disease?

In the correlation between the medical history risk factors in Table 2-2, the top four relatable diseases, chronic stress, obesity, blood pressure and heart disease are also very much caused by and related to life style risk factors. For example, in some cases stress, obesity, blood and heart disease, can be avoided by simply changing the diet and lifestyle, and staying physically active. After genetics, it is possible for lifestyle to be the biggest early risk factor of Alzheimer's disease.

Medical history plays a significant role in the development of Alzheimer's disease. Most of the diseases that play a role in developing Alzheimer's disease are not common among young people, which makes it difficult to identify Alzheimer's disease at an early stage. If the critical diseases that contribute to the development of Alzheimer's disease are caused by poor lifestyle and personal characteristics, then, having a good, healthy and quality lifestyle from a very young age could help to prevent the risk of getting Alzheimer's disease.

To visualise the connection between the risk factor diseases and Alzheimer's disease, Figure 8, is a network graph that demonstrates the link between the diseases. Each disease is represented in a circle (station); the connection between the diseases is demonstrated using the arrows (links), an outgoing link means that the disease that the link is pointing at is one of its causes and vice versa.



**Note:** Each disease is represented by a colour to visualize them better, unlike the circle size, the colours do not represent any significance.

Figure 2-7 Visualisation of Table 2-2.

In Figure 2-7, shows the main three stations in the network are obesity, blood pressure and stress. The reason why such a simple analysis triggers interest is because the outcome is immensely lifestyle-related and has a huge impact on the bloodstream. A protein called beta-amyloid is believed to be the cause of the development of plaques, which then triggers the development of tangles, and both together are hallmarks of, and responsible for the development of Alzheimer’s disease. However, not all brains with beta-amyloid will develop Alzheimer’s disease. The development of beta-amyloid is believed to be caused by fat and protein carried in the blood to the brain cells. The outcome of the analysis of Figure 8 shows that bloodstream and lifestyle-related diseases are high risk factors. The relation between obesity, blood pressure and stress together with other diseases developed over time could potentially explain the pathology of Alzheimer’s disease.

*“Alzheimer’s disease starts to manifest long before plaque formation becomes evident,” said Carla Shatz, PhD, professor of neurobiology and biology at Stanford University[58]. Dr Shatz also said, “I’ve always found it strange that these mice — and, in fact, all the mouse models for Alzheimer’s disease that we and other people study — seem not to have any problems with memory until they get old,” Shatz said. “These mice’s brains have high levels of beta-amyloid at a very early age.”*

The development of beta-amyloid happens in Alzheimer’s disease suspects at a very young age, and symptoms only become detectable at a much older age. In this case, medical history is one of the main sources to predict Alzheimer’s disease at a very early stage. The use of data analysis and machine learning will help in finding medical history patterns for the development of Alzheimer’s disease in current existing Alzheimer’s disease databases. The use of these patterns alongside other patterns found in lifestyle, characteristics and genetics could help with the diagnosis of Alzheimer’s disease 20-30 years before its symptoms are noticeable.

### **2.6.3 Lifestyle and Diet**

“A healthy body equals a healthy mind”. What we eat and what we do will certainly affect our mental health performance. Most of the common diseases of today are due to bad diet and nutrition. One of the questions in this thesis is how can lifestyle and diet contribute to the development of Alzheimer’s disease? Probably, the best way to investigate the answer is by looking at the patient’s lifestyle and their medical record to identify the lifestyle elements that trigger the development of Alzheimer’s disease risk factors. An example would be; a lifestyle of too much fat and unhealthy food with a lack of exercises will trigger the development of

both obesity and diabetes, which will then contribute to the development of many other blood and heart diseases. Heart problems, diabetes and other bloodstream diseases are high risk factors of Alzheimer's disease. Hence, lifestyle is the primary cause of heart problems, diabetes, stress and other bloodstream diseases; one can argue that besides genetics, the lifestyle and diet are primary contributors to the development of Alzheimer's disease.

This section explores the diverse lifestyles and diets that might contribute to the development of Alzheimer's disease. The process of developing Alzheimer's disease is still a mystery; however, the biological hallmarks of the disease are Plaques and Tangles, and the focus now is to investigate the development of these hallmarks. Plaques and Tangles are the hallmarks of Alzheimer's disease. Plaques, are abnormal clusters of protein fragments, built up between nerve cells. Tangles are dead brain cells, scientist know that tangles are formed after the development of Plaques around the healthy brain cells which cause them to die. For early prediction of the disease with the use of patterns found in diet and lifestyle, it will only make sense to investigate the elements that are responsible for the development of plaques, clusters of protein around the brain cells.

Brain cells or neurons are responsible for releasing the beta-amyloid protein, this process happens in almost all of us, and it has no preferred structure. In people with Alzheimer's disease, this protein folds in an abnormal structure around the cells to promote aggregation and triggers a cascade of pathologic events. What causes beta-amyloid to fold in such an abnormal way around the brain cells is a very interesting field of research to investigate; however, this is mainly down to biology and clinical trials. In this section, the focus will be on the type of food and activities that have an impact on our nervous system and specifically beta-amyloid protein.

Identifying risk factors in food and day-to-day activities is a very complex process and variant. Due to our inconsistent nature and diet, it will be difficult to identify risk patterns for Alzheimer's disease. Alternatively, a more simple and practical way to identify possible factors would be by investigating the diet and activities that are responsible for increasing the risk of developing medical risk factor diseases for Alzheimer's disease. For instance, if diabetes is suspected to be a high-risk factor for Alzheimer's disease, then it will only make sense to investigate the diets and activities that are responsible for diabetes, in order to identify the type of food and activities responsible for increasing the risk of Alzheimer's disease. Like the analysis given in Table 2-2 to investigate the association between risk factor diseases, in Table 2-3 and Table 2-4, we have listed all of the risk factor diseases alongside three known contributors of these diseases. In Table 2-3, risk factor diseases are listed alongside the top three foods habits that contribute to their pathology. In Table 2-4, risk factor diseases are listed together with the top three lifestyle habits that contribute to their development.

The aim of these two simple preliminary analyses is to learn about the signs that could indicate early development of Alzheimer's disease. The results of this preliminary analysis are unverified, and the data used is based on the information found on each disease from different sources [59][60][61]. Therefore, it will not be strictly accurate to claim the results of this analysis as a discovery. However, the outcome will help us find a pattern to our research that might be worthy of verifying in future work. Our future work will study early prediction of AD using risk factors and lifestyle.

Table 2-3 Alzheimer's disease Medical History Risk Factors & the Relationship to Diet

Disease	Diet 1	Diet 2	Diet 3
1. Diabetes	Alcohol	Candy (Sugary Food)	Cakes (Fat Food)
2. Obesity	Fast Meal (Oily Food)	Candy (Sugary Food)	Cakes (Fat Food)
3. Stroke	Red Meat (CAFO)	Fizzy Drinks	Processed Salt
4. Depression	Lack of sea food	Artificial sweeteners	Alcohol / Caffeine
5. HIV Infection	Transmission	Transmission	Transmission
6. Down's Syndrome	Genetic	Genetic	Genetic
7. Cholesterol	Fast Meal (Oily Food)	Cheese	Liver Dishes
8. Heart Disease	Fast Meal (Oily Food)	Alcohol	Candy (Sugary Food)
9. Head Trauma	Alcohol	MSG / Processed Salt	Caffeine
10. Blood Pressure	Processed Salt	Canned Food	Alcohol
11. Chronic Stress	Caffeine	Alcohol	Fast Meal (Oily Food)

**Note:** Please note the table above is the results of a basic premature analysis.

The information in Table 2-3 indicates a relationship between the food we eat and the development of different diseases that could possibly trigger the pathology of Alzheimer's disease in the long-term. Alcohol appeared 6 times in Table 2-3 (results in Table 2-5), which makes it the top leading diet habit to over half of the Alzheimer's disease risk factor diseases. Followed by fast meal / oil rich food, which was mentioned 4 times and it contributes to the high-risk factors of Alzheimer's disease; Obesity, Cholesterol, Heart Disease and Chronic Stress. Caffeine, candy and processed food and salt - these food categories have been mentioned 3 times each in the table.

The food addressed in this analysis is food most of us consume on a day-to-day basis. The results of this analysis could possibly be the reflection of our present-day food consumption. As a UK society and like most developed countries around the world, the food we bring home with us is most likely to fall into all these food categories in Table 2-3. For instance, alcohol consumption is one of the major problems we have in the UK, in 2014 there were over 1 million



hospital admissions related to alcohol consumption in the UK. The food mentioned in Table 2-3 includes a list of essentials that we can simply completely avoid as a society; however, it is an indication that too much of anything is unhealthy. A healthy well balanced diet should help us avoid most of the Alzheimer’s disease risk factor diseases and keep a clean medical record [59].

Like the analysis in Table 2-3, Table 2-4 demonstrates the impact of the lifestyle we choose on our health; the activities that we do as a part of our lifestyle. What we do can sometimes determine the habits or diet we get ourselves used to. There is no decision that we can make that does not come with some sort of balance or sacrifice, and almost every action we take has a consequence. What we choose to eat or not eat will eventually have an impact on our body health.

Table 2-4 AD Medical History Risk Factors & the Relationship to Lifestyle Activities

<b>Disease</b>	<b>Lifestyle 1</b>	<b>Lifestyle 2</b>	<b>Lifestyle 3</b>
1. Diabetes	Lack of exercise	Alcoholism	Smoking
2. Obesity	Food Addiction	Lack of exercise	Stressful Lifestyle
3. Stroke	Lack of exercise	Stressful Job	Family Problems
4. Depression	Family Problem	Loneliness	Smoking
5. HIV Infection	Unprotected Sex	Blood Contact	Transmission
6. Down's Syndrome	Genetic	Genetic	Genetic
7. Cholesterol	Lack of exercise	Alcoholism	Stressful Life
8. Heart Disease	Lack of exercise	Smoking	Food Addiction
9. Head Trauma	Alcoholism	Head Injuries	Stressful Lifestyle
10. Blood Pressure	Stressful Lifestyle	Family Problem	Smoking
11. Chronic Stress	Emotional Sensitivity	Expectations	Demands

**Note:** Please note the table above is the results of a basic premature analysis.

The second part of this preliminary investigation is identifying possible contributors to Alzheimer’s disease risk factors in our daily lifestyle. Table 2-4 shows a list of the high-risk factors found in the medical history of Alzheimer’s disease patients, alongside them the top three possible lifestyle contributors to their development. We learn that a stressful lifestyle with a sharp lack of exercise opens the opportunity for Alzheimer’s disease risk factors to develop. Lack of exercise and stressful activities are the high-risk factors in this analysis, followed by family problems, alcoholism and smoking.

Table 2-5 Summary of Table 2-3 and Table 2-4

<b>Lifestyle Factors</b>	<b>Occurrence</b>	<b>Diet Factors</b>	<b>Occurrence</b>
No exercise	5	Alcohol	6
Stressful Lifestyle	5	Oil Rich Food	4
Family Relation	3	Processed Food / Salt	3
Alcoholism	3	Caffeine	3
Smoking	3	Sugar Rich Food	3
Food Addiction	2	Fat Rich Food	2
Social Isolation	1	Red Meat (CAFO)	1
Emotional sense	1	Low Sea Diet	1
Head injuries	1	Fizzy Drinks	1
High Expectations	1	Artificial Additions	1
High Demands	1	Cheesy Food	1
		Canned Food	1
		Liver Dishes	1

In the analysis of diet and lifestyle, it’s clear that the top three main risk factors are the over consumption of alcohol, stressful activities, oily unhealthy food and the lack of physical activities. Avoiding these patterns in our diet and lifestyle will help us avoid development risk factor diseases, which will play a vital role in early prevention of Alzheimer’s disease risk. However, as the aim is to predict Alzheimer’s disease risk at a very early stage, it is important

to obtain such information about suspected individuals who might be at risk of developing Alzheimer's disease.

#### **2.6.4 Characteristics**

Life form is indeed a very complicated and sophisticated feature of planet earth. The extreme complexity of living organisms and their biology is a fascinating science with all of its hidden secrets. The perfection, in which life forms have been engineered with all of their intricacy, has attracted many scientists and researchers to investigate how they work for over thousands of years. This perfect engineering is called life. Each life form has a unique imprint, and a complexity of its own. When it comes to humankind, we are far more sophisticated and complicated in comparison to the rest of creation on planet earth. Not only biologically unique, but we are the only creation with such emotional and mental intelligence. The psychological and emotional complexity is what makes us more unpredictable and unique. Our psychological and emotional intelligence has divided us into nations with different languages, cultures and lifestyles.

Each human is unique and carries characteristics that define their life. These characteristics are personality, family, demography or environment, emotional intelligence and social life. Our characteristics are determined by our emotional and psychological intelligence. Who we are determines the things we do and the way we live our life. Our social life and environment also have a great influence on our emotional and psychological complexity. Living with a mental health problem such as Alzheimer's disease risk, causes and risk factors can vary between both biological changes and impacts of characteristics [62][63].

People suffering from Alzheimer's disease risk factors, would probably have a stressful type of personality with high expectations in every aspect of life. Both stress and high expectations create pressure on the mind and influence changes on the lifestyle. This part of the work will explore the impact of characteristics on people with potential to develop Alzheimer's disease risk. Characteristics influence changes on the daily lifestyle, which means a negative characteristic combined with unbalanced lifestyle can potentially cause all sorts of health problems. For instance, being a stressful person with a stressful job influences some people to smoking, unhealthy eating, overthinking or lack of sleep. These influences are the primary causes of several illnesses such high blood pressure, chronic stress, heart problems and other type of diseases. Characteristics can also trigger or help in the development of Alzheimer's disease risk medical risk factors at a very early stage [64] [3].

The characteristics that can be classified as high risks factors for Alzheimer's disease risk are; low education, stressful job and gender [65]. The number of females subjected to Alzheimer's disease risk is higher than males, and also the majority of Alzheimer's disease risk patients either have not been in higher education or spent the majority of their working life in a stressful job. An interesting research paper published by the American Psychological Association concluded that women are more likely to suffer from depression and stress than men, which probably explains why there are more women who are likely to develop Alzheimer's disease risk [66][67]. Stressful life and stressful jobs cause the majority of the medical risk factors of Alzheimer's disease risk. The demographic indication shows that people in western and developing countries are at greater risk of developing Alzheimer's disease risk than people in the Far East, Asia and Africa. This could possibly be due to the dynamic and demanding

lifestyle in countries with developing economies, where employability is high, high alcohol consumption and people live a more demanding lifestyle.

Alzheimer's disease risk is initiated by multiple variations of risk factors cumulating together over time, to cause the development of Alzheimer's disease risk biological hallmarks. Consequently, the most effective approach to prevent Alzheimer's disease risk would be as early as teenage life or childhood; long-term education, successful relaxed career, comfortable lifestyle and most importantly good balanced diet, all of these factors will help reduce the risk of getting Alzheimer's disease risk.

### **2.6.5 High Risk Factors**

The known high-risk factors of Alzheimer's disease are genetics and in the medical history of the patients. While genetics remains a primary high-risk factor as it is something people are born with, from the analysis in this chapter we understand that risk factors in medical history might not be considered primary as they are triggered contributing risk factors in diet, lifestyle and characteristics, and therefore the contributing risk factors should be considered as primary factors. However, in some cases, risk factors in diet, lifestyle and characteristics are also influenced by risk factors in the medical history. But this shows that the development of Alzheimer's disease is spread across a combination of medical history, lifestyle and diet. Sustaining a healthy lifestyle and diet might prevent the development of risk factors in medical history and overall the development of Alzheimer's disease.

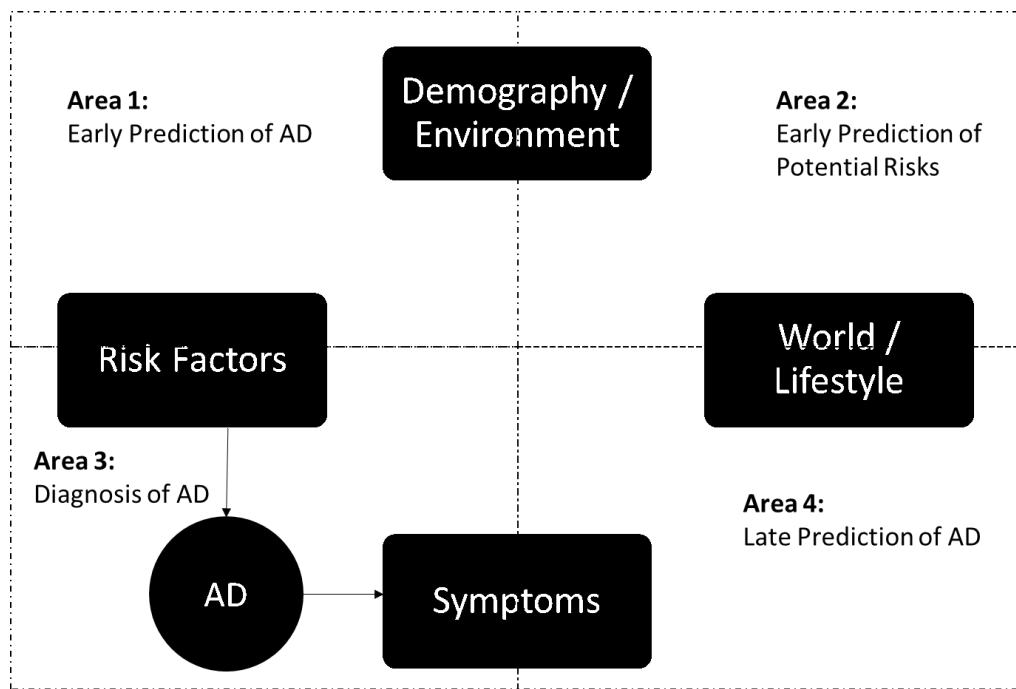


Figure 2-8 Different areas of study for AD

Figure 2-8 demonstrates how Alzheimer’s disease is studied from different areas. Area 1 explores the area in which Alzheimer’s disease can possibly be predicted at an early stage. At this point, the patient does not have any Alzheimer’s disease symptoms but has existing risk factor patterns in their medical history, which also interconnect with patterns in their characteristics and diet. In Area 2, it will be quite inaccurate to predict Alzheimer’s disease at this stage. This is because patients have not yet developed any of the high-risk factors nor have any symptoms. The data available in area 2 will consist of information on lifestyle, diet and characteristics, which will only be useful to predict early potential high risk of Alzheimer’s disease. Unlike area 2, both areas 3 and 4 are linked to late prediction and diagnosis of

Alzheimer's disease. At this point of the Alzheimer's disease pathology, the patient will already have most of the medical risk factors as well as visible symptoms of Alzheimer's disease.

### **2.6.6 Summary**

Alzheimer's disease costs its victims their lives, and it has a tremendous effect on health organisations and the economy all around the world. Statistics show that around 1 in 10 people over the age of 65 will be affected by Alzheimer's. Unfortunately, there are no effective cures for this disease and no one is immune [16]. Getting old and weak is something most people find hard to accept. It's a struggle that mature people over the age of 50 are facing but the concern doubles with the fear of losing their memory due to dementia. Elderly people who are affected by dementia are living the experience of watching themselves die slowly, fade away from their world, live in constant confusion, and no longer able to understand their surroundings. It is a horrible experience to endure for Alzheimer's disease victims, their carers, and their families. Having Alzheimer's disease means losing loving memories, the ability to recognise family members, and childhood memories, or even the ability to follow simple instructions e.g. making their usual morning cup of coffee, remembering how to use the toilet, and maintain self-hygiene [2][3][4].

Taking action to fight Alzheimer's disease will enable people to live longer and more independently. Unfortunately, what exactly causes Alzheimer's disease to develop is still undiscovered, however, research shows there are several known risk factors that contribute to its development. These risk factors fall into an array of categories; medical history, lifestyle, family dementia history, characteristics, and demography. Moreover, these risk factors are

classed as either behavioural markers or biological markers of Alzheimer's disease. Many research initiatives fighting Alzheimer's disease mostly focus on either new drug development or investigating the disease by studying its biological markers. With the expansion of computer science, several research approaches have emerged using the power of data science and machine learning to study Alzheimer's disease. Unfortunately, because of the challenges more specifically data limitations, researchers were inclined to carry out their study on biological markers of Alzheimer's disease and almost neglected its behavioural markers. Our research comprehensively studies Alzheimer's disease risk factors using both behavioural and biological markers to seek possible early prediction, or an onset diagnosis of the disease.



# Chapter 3      **Literature Review**

## **3.1 Introduction**

The world is almost fully dependent on the aid of computers. We use computers in almost everything such as agriculture, medical care, trade, travel, manufacturing, and communication. Computers are heavily used to aid us with decision making or to achieve tasks very quickly that the human mind wouldn't be able to achieve.

Data analysis and machine learning are interdisciplinary fields, where the former uses different scientific methods to collect, store, and extract data, and the latter provides systems with the ability to learn and improve from experiences using data without being explicitly programmed. These two fields of study are closely related and have tremendously impacted and accelerated the development of technology.

The different methods and techniques from the fields of data analysis and machine learning were applied throughout this research; therefore, this chapter presents a literature review of machine learning and data analysis, as well as an overview of the related work to the research problem in this thesis.

## **3.2 An Overview on Data Science**

Data is commonly used to extract knowledge and insights that would help us with decision making. However, to acquire useful information that would help us make decisions, we would

need to collect a data set relevant to our problems, then analyse the data in an effective way using scientific methods and algorithms to give us the information we need.

Manual extraction and analysis of data is extremely difficult, and, in some cases, it is impossible to do it without the help of machine learning tools and methods, especially, when we have large datasets. For example, companies such as Amazon and eBay use data science and machine learning to analyse a mixture of structured, semi-structured and unstructured data in search of valuable business information and insights [68] [69] [70].

The usefulness of data science is that it helps to uncover hidden patterns and unknown correlations in the data sets, which helps companies or corporations to understand market trends, customer preferences and other useful business information [69]. When working with large data sets, data scientists are responsible for the analysis, capture, duration, search, sharing, storage, transfer, visualisation, the privacy of this data, and extraction of useful meaningful knowledge [68] [71].

The research aim in this thesis is to present a framework to predict Alzheimer's disease at a very early stage, the experimentation in this thesis to demonstrate the framework uses risk factors data related to behavioural markers datasets acquired from the Alzheimer's disease Neuroimaging Initiative (ADNI). Scientific methods and computational models will be employed to ensure the data used is as accurate as possible, and clean from errors. Working on data for patients with Alzheimer's disease , such as ADNI [72], will require solid knowledge of data analysis and cleansing.

The ADNI database contains incomplete datasets [73], which means that before using such data and to make sense of the data it is important for this data to be cleaned. The data cleaning process ensures that the data is valid, clean, accurate, complete, consistent and uniform. Especially, when dealing with data for Alzheimer's disease patients that contain large sets of data with hundreds of variations, it is crucial for the data to be valid and of a high quality.

A good example of data cleansing and challenges in working with ADNI data is in Qu's work, titled: "A Predictive Model for Identifying Possible MCI to AD Conversions in the ADNI Database,"[73], he realized that there are some tests in the ADNI database in which a limited number of patients have participated and the corresponding values were marked as "-1000" for the rest of the subjects. Therefore, he had to clean the database and remove all the incomplete fields. It is crucial to have a clean dataset before applying data analytical tools such as principal component analysis (PCA), Pattern Recognition Tools (PR Tools) or any other machine learning tools.

### **3.3 Concept of Machine Learning**

This section discusses the study of Machine Learning (ML), one of the Artificial Intelligence subfields. Artificial Intelligence, often referred to as AI, is the field of study of intelligent behaviour and is a description given to smart software or machines that have the capability to think and learn independently. John McCarthy, who coined the term in 1955, defines it as "the science and engineering of making intelligent machines". Today AI is widely used by large corporations and businesses such as Tesla, Google and Apple, and Militaries around the world.

AI is developed on algorithms and artificial neural networks, which are inspired by biological neural networks in the central nervous systems of humans. The overall goal of AI is to develop systems that can learn and mimic the human response and behaviour in different circumstances. AI is highly complicated and very much a specialised field to study, which focuses on reasoning, knowledge, planning, learning (Machine learning), communication, perception and the ability to move and manipulate objects.

Machine Learning (ML), is a term used to describe the cover of providing computers with the ability to learn from experience from data and search for patterns without being programmed. It is widely used across almost all disciplines, for diverse purposes ranging from commercial use by businesses and healthcare to academia to conduct research studies. It is used by companies like Facebook to show personalised advertisements, or for image recognition to allow users to tag their friends. It is also used by gaming companies like the Nintendo Wii that uses real time image recognition and an algorithm called random forest to track users' movements, which, allow users to interact with the game by only moving their body and hands without a joystick. Machine learning is used by virtual reality technology companies to build virtual reality video games, and by mobile phone companies that provide a keyboard voice tool that most people are familiar with in modern smartphones, which uses machine learning algorithms for voice recognition to convert speech to text. Another example of the use of machine learning is in robotics, e.g. building walking dogs robots that use reinforcement machine learning algorithms that allow the robot to learn how to walk on its own.

Machine learning searches for patterns in data to enhance the performance of the system and change its actions accordingly, without human interference. This concept of learning from

experience without explicit programming will leave a huge impact on the future of technology and computer science in general.

Machine learning is a technology that the future will be built on. With 3.7 billion humans having access to the World Wide Web, the amount of data generated each day exceeds 2.5 quintillion bytes (equivalent to 2,328,305,664 Gigabytes). From 2010 to 2018 the growth of generated data has reached 50 times to an estimated 40,900 exabytes of data. These statistics are astonishing and show that with the growth of data there is an essential need for machine learning to process and analyse this large amount of data.

It is true that we humans are smarter than computers but when it comes to remembering, executing complex tasks and analysing data they're better than us and more accurate if designed correctly. The following sections in this chapter will discuss the different types of machine learning and the approaches, and their use for classification problems and prediction.

### **3.3.1 Machine Learning Approaches**

Machine Learning (ML) is a computational technique used for automated or semi-automated extraction of knowledge from large datasets. This is aimed to give computers the ability to learn from data and to classify or give productive values. It is inspired by human biological ability to learn and find answers to questions. Automated machine learning means that the computer can provide the insight into the data without human interference. Semi-automated learning is when many of the decisions made involve the human input.

ML has two different types of learning “signal” and “feedback”, and three different categories; supervised learning, unsupervised learning and reinforced learning. These three categories help systems to recognise patterns, learn from unknown data and interact with the environment. The complexity of this field has opened a window for innovation and research; as a result of these, different approaches were used such as the Decision Tree Learning for predictive modelling or the Bayesian Networks for graphical modelling.

The way that ML works is that it employs classification techniques (often referred to as classifiers) to combine a subset of the dataset into different classes depending on their variables (features). There are many classification and prediction algorithms techniques that can be employed for ML on a dataset. Even though they all perform differently and produce different results, they still have mutual procedures and characteristics.

To apply machine learning technique on a dataset, and enhance the overall outcome, the dataset needs to be split into three subsets, a subset for training, validation subset and testing subset. The reason for this is to provide the ML algorithm with a set of data to learn from by performing correlational tasks such as clustering, classifying and protecting the classes. The principal motivation behind having a validation subset of the dataset, for occurrence in Artificial Neural Network (ANN) is to locate the ideal number of hidden layers or to decide the definite ceasing point for the backpropagation technique. Finally, the testing subset is utilized to evaluate the execution of classifiers with obscure class names.

### 3.3.1.1 Supervised Learning

Supervised learning - or Model Training is when the Machine Learning model is trained on an existing labelled dataset that has already been labelled with the desired outcome. For instance, using medical data to classify patients with a certain disease who have already been diagnosed. This is used to help the model to learn the relationship between the attributes or variables of the dataset. After the model has been trained on the data, the test data would be used without any labels to predict the desired outcome. In simple terms model training is that the model is learning from past examples made up from inputs and outputs and applying what it has learned to future inputs in order to predict future outputs. The diagram below will demonstrate the normal process of a supervised machine learning (model training) technique.

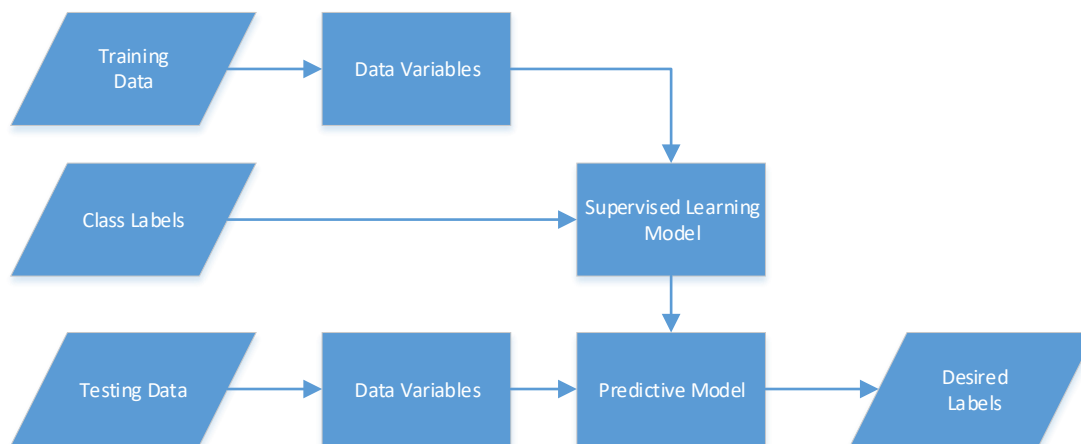


Figure 3-1 Supervised learning workflow [74]

### 3.3.1.2 Unsupervised Learning

Unsupervised Learning - Unlike supervised machine learning, unsupervised learning is also a machine learning technique, but it does not require the input of classes to learn from its algorithm during the training stage. In simple terms this technique is used to describe hidden structures from unlabelled data, also often referred to as clustering analysis, and used to draw conclusions from the entire dataset [27]. The diagram below demonstrates how unsupervised learning works.

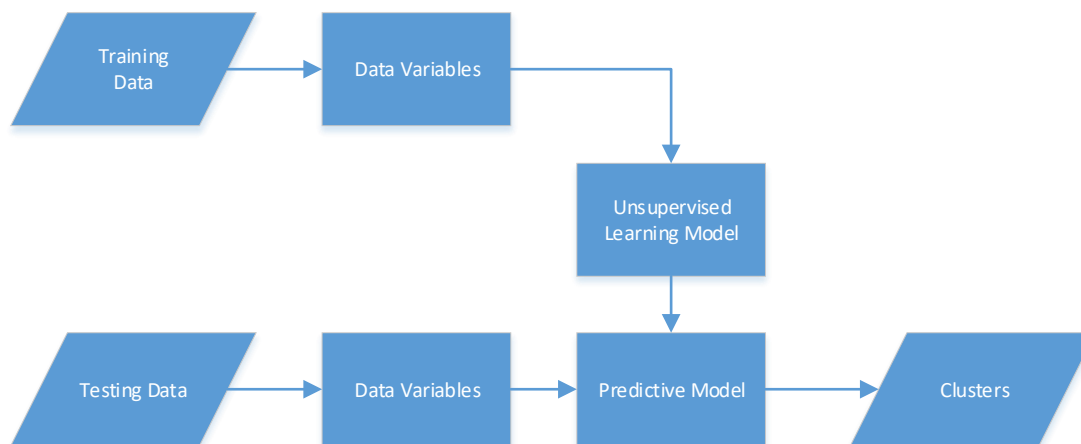


Figure 3-2 Unsupervised learning workflow [75]

### 3.3.1.3 Semi-Supervised Learning

Semi-Supervised learning is utilized to influence the notion of joining supervised learning (labelled dataset) and unsupervised learning (unlabelled dataset), which might have sway on changing the learning conduct. This type of learning is a significant advancement to machine learning as it enables the use of both labelled and unlabelled data which will make learning



from big data more accurate. The diagram below demonstrates how semi-supervised learning works.

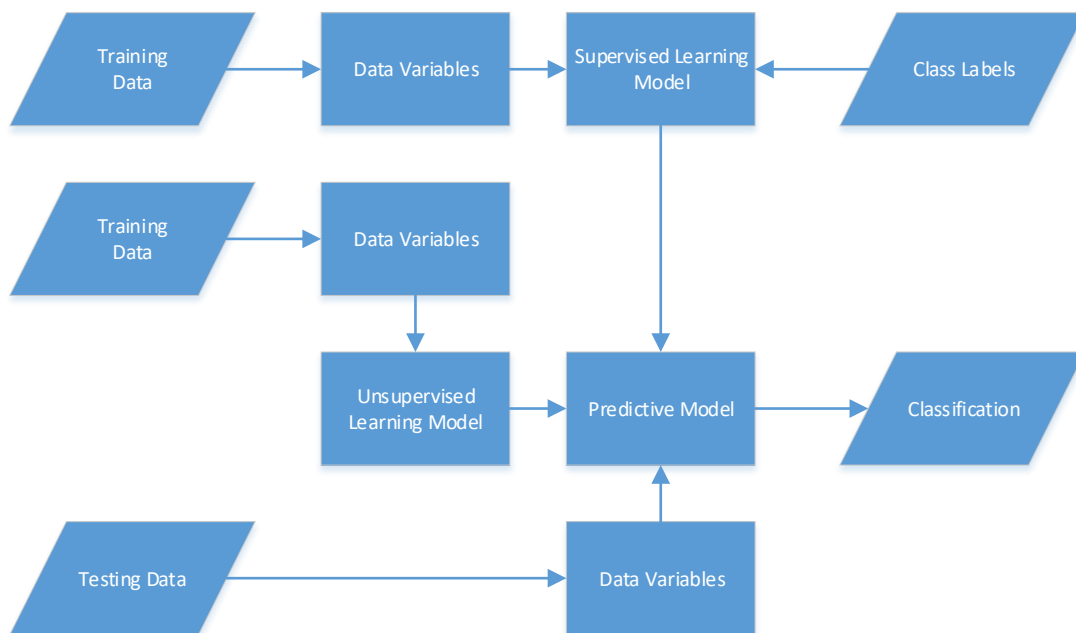


Figure 3-3 Semi-Supervised learning workflow [75]

### 3.3.1.4 Reinforcement Learning

Reinforcement Learning (RL) has a totally different concept to other types of machine learning techniques, inspired by the concept of consequence influence behaviour. Which means that it takes action because it knows other consequences will follow. The idea is that the algorithm of RL learns from the consequences of its actions. Unlike supervised learning where the model is explicitly taught. RL has three main learning rules from its consequences; if the consequences give a positive outcome (Reward) then its behaviour increases, if the consequences give a negative outcome (Punishment) then its behaviour decreases, and if the consequences are neutral then it extinguishes the behaviour.

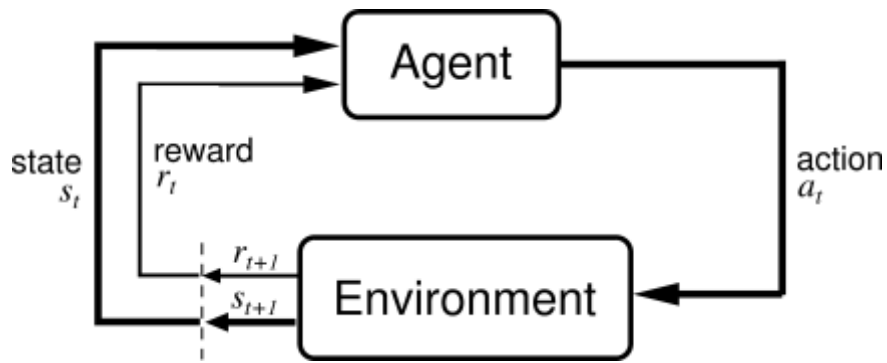


Figure 3-4 Reinforcement learning workflow [76][75]

### 3.4 Learning Models

The concept of machine learning is concluded in the use of algorithms based on computational models to train computers to extract knowledge from large sets of data. The data set represents historical data from reality, for example medical observation history for real patients. A machine learning model will have its own unique algorithm that would guide it to use this data set and discover different patterns to either classify the data or predict future data.

The algorithm that a model uses to manipulate or interpret the data is based on a set mathematical equation involving linear algebra, logarithmic, arithmetic, statistics, probability, and calculus. There are many machine learning models with multiple variations of algorithms and in some cases models are combined and classed as hybrid models.

Understanding the research problem and the type of data is very important before applying any machine learning model. To solve a problem or discover a pattern effectively you must apply the right type of models on the suitable datasets. For example, we cannot use classification models on a data set where we trying to predict a house price in a particular city. There are

models that are specifically designed to work best with classification problems such as teaching the machine to class emails as spam or not spam, these types of models are called classifiers such as Support Vector Machine (SVM). On the other hand there are also models that can be used to solve both classification and regression problems such as Artificial Neural Network.

This section will briefly describe the major learning models currently used and the different categories they fall under.

### **3.4.1 Artificial Neural Network**

The goal of ML is to give systems the ability to mine data, analyse data and learn from that data without being exclusively programmed. The Decision Tree and association rule learning are two ML methods that are used to identify the most important variable and discover data patterns. Another method that is typically used to find patterns and to represent non-linear and linear relationships in data is the artificial neural network learning algorithm.

Artificial Neural Networks (ANN) are sometimes described as simulated neural networks and are a group of connected neural networks created by software to work in a similar way to the biological neural networks of the human brain. The human brain is built from multiple nerve cells called neurons, connected together in a well-organized structure, this structure has inspired the science field of Neural networking and Artificial Intelligence [77]. Since scientists are now able to explain some of the functionalities of the brain in terms of learning, remembering, identifying objects and decision-making, ANN can be described as a mathematical concept that attempts to try to understand the functionality of the human brain. There are several neural networks software packages, which are used to simulate, research, visualise, develop and apply

the concept of artificial neural networks. One of the main software packages widely used by researchers to investigate ANN is a software called MatLab [78].

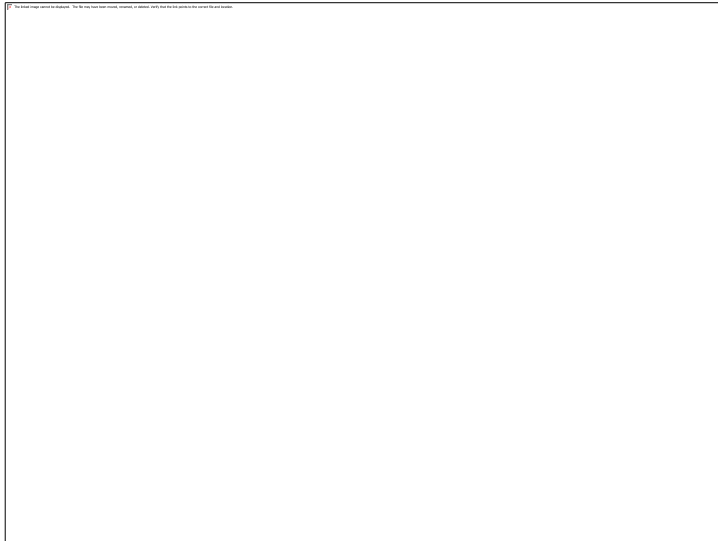


Figure 3-5 Artificial Neural Network

In machine learning the ANN concept is used to classify data and to model complex relationships between inputs and outputs, to identify models and structure in data. An example of using ANN in scientific research and patterns discovery is the work of Meysam

Torabi at the Iranian Sharif University

of Technology. The title of his work “Discrimination between Alzheimer’s disease and Control Group in MR-Images Based on Texture Analysis Using Artificial Neural Network” [79]. Torabi used ANN to classify MRI scans of 75 subjects; 50 normal subjects and 25 who have previously been diagnosed with AD (data obtained through Harvard University “The Whole Brain Atlas”). After extracting the features in the datasets, he used ANN to focus on brain texture in both normal and abnormal cases, comparing both types of subject in an attempt to diagnose AD. His work demonstrated that the neural network makes a significantly different output for AD patients compared with normal control group, which gives a strong diagnosis of AD with 95 percent proper response.

ANN 'learns' from observed data and requires good understanding of the underlying theory before using it, but is quite widespread and is not so straightforward. Before using this concept

to classify data, in order to have a robust ANN it is important to choose the correct model right learning algorithm.

An artificial neural network is formed from inputs, synapses (lines), hidden layers (neurons), and output layer. As illustrated in Figure 14, Synapses are responsible for taking values from inputs, calculating the input against a certain weight and passing it on to hidden layers and to the output layer. Increasing the number of hidden layers and synapses will form a deep learning neural network. ANN is a supervised learning model that can either be used to predict a single value for classification purposes or continues prediction for regression study.

### **3.4.1.1 Levenberg-Marquardt Feed-forward Neural Network**

Levenberg-Marquardt Feed-forward Neural Network is a feed-forward neural network that uses the Levenberg-Marquardt method to optimise its loss function. The Levenberg-Marquardt method is widely used for its training precision. Optimisation theory is a major field in the study of mathematics and since neural networks and machine learning algorithms heavily depend on mathematical equations there are parts of these algorithms where optimisation is needed in order to help the machine learning models such as neural network to learn and improve their predictions.

The learning of a feed-forward neural network model is measured through a loss function or sometimes known as cost function. The loss function measures the ability of the prediction algorithm to predict the expected outcome. An optimisation method is required to minimise the error of the loss function in order to find the minimum point where prediction is most accurate. There are several optimisation methods used but the most commonly used methods are the

gradient descent and the Gauss-Newton. The gradient descent takes small steps toward the minimum point by continuously updating parameters to reduce the sum of squared errors. The Gauss-Newton reduces the sum of squared errors by assuming the least squares function is locally quadratic and finding the minimum of the quadratic.

The Levenberg-Marquardt can be considered a combination of both methods; gradient descent and the Gauss-Newton. When the predicted output is far from the expected output, the Levenberg-Marquardt algorithm behaves like a gradient descent method, and when the predicted output is very close to the expected output, the algorithm behaves like a Gauss-Newton method. The Levenberg-Marquardt method is very effective and works well in practice [80]–[82].

### **3.4.1.2 Back-Propagation Feed-forward Neural Network**

The backpropagation algorithm was made famous in a 1986 paper by David Rumelhart, Geoffrey Hinton, and Ronald Williams. Their work described how a neural network with backpropagation works faster than the current learning approaches used back then. Today, backpropagation is widely used with neural network models. This section will briefly explain the use of backpropagation with a feed-forward neural network.

A feed-forward neural network is one of the Artificial Neural Network architectures where the connections are fed forward from input all the way to the output layer. Unlike the recurrent neural networks where the neurons form circles and feedback their output, feed-forward

prohibits any feedback between the layers. Often the feedback architecture would be confused with the back-propagation method; although, both are completely different things.

A feed-forward neural network with backpropagation means the training algorithm of the model consists of two steps; first, the input values are feedforward to the hidden layers then during the training optimization process the algorithm carries out an iterative adjustment to the weight of neurons. This is done to avoid overfitting and to reduce the differentiation between the expected results and the predicted results, over multiple training cycles until the model reaches the best level of prediction accuracy [83].

### **3.4.1.3 Multi-layer Perceptron Neural Network**

Muli-layer Perceptron Neural Network is a feedforward neural network architecture. A traditional single neural network would normally consist of one input layer, one hidden layer, and one output layer. This type of neural network architecture would only solve a linear problem as one neural network can only handle AND or OR boolean. When it comes to non-linear problems and the handling of XOR boolean it can only be achieved by increasing the number of nodes in the hidden layer which are often referred to as perceptrons.

The increase of perceptrons in a hidden layer results in a Multilayer Perceptron Neural Network (MLP). MLP was an evolution of the single perceptron neural network and can give better prediction results especially when using back-propagation algorithm. MLP would use  $x$  number of inputs and pass each input to each perceptron in the hidden layer with weight attached to it. The calculation of all weights between inputs and the perceptrons will be stored in a weight matrix, then the perceptron would calculate a weighted sum of all inputs denoted

as  $Z = \sum_{i=1}^n x_i w_i$ , where  $n$  is the number of inputs,  $x$  is the input,  $w$  is the given weight, and  $Z$  is the weighted sum.

$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$

Equation 3-1

Equation 3-1 represents the sigmoid function where  $x$  is a weighted sum, and  $e$  is the natural logarithm base or the Euler's number. Once the weighted sum is calculated, it will then be passed through to an activation function such as this sigmoid function in Equation 3-1 which will compute the value to be between 0 and 1. [84] [85]

Once the activation function generates an output the perceptron in the hidden layer would then calculate the classification error, which is the difference between the given output and the desired output. Based on that error, the model would then fire the back-propagation process which will feedback adjustment to all weights in the network.

The weights are adjusted back to front, meaning that the weights between the hidden layer and output layer will be adjusted first, then every other weight between input layer and the hidden layer will be adjusted. The weights between the layers are adjusted based on their responsibility for the error, therefore, the calculation of the output layer error is completed, followed by a calculation of error in in the hidden layer. The total of the output error is denoted as

$E = e_{o1} + e_{o2} \dots$  where  $E$  is the sum of all errors produced from the data points and  $e_{o1}$  is the individual error which is the difference between the target and the output results, and its calculation is denoted as  $e_{o1} = \frac{1}{2}(t_1 - o_1)^2$ , where  $t$  is the target and  $o$  is the output.



Once the output error is calculated, every perceptron in the hidden layer will have its own error sum, before the weights are adjusted. The calculation of the hidden layer perceptron error is denoted as  $e_{h1} = \frac{w_1}{w_1+w_2...} * e_o$ , where  $e_{h1}$  is the error for perceptron one in the hidden layer,  $w_1$  is the weight of the perceptron to the output layer,  $w_1 + w_2$  are the output layer weighted sum, and  $e_o$  is the error in the output layer. Then the weights between the layers will be adjusted based on their responsibility for the error by using Gradient Descent, sometimes called the Generalized Delta Rule [83], [86], [87].

#### **3.4.1.4 Deep Learning**

As the study of artificial intelligence (AI) has evolved and expanded to become a multidisciplinary field of research, machine learning as a subfield of AI has also expanded and has its own subfields of study such as supervised, semi-supervised, and unsupervised learning. One particular field of study that has become a major field in deep learning, is the study of Artificial Neural Networks. Artificial neural networks started as a simple study of a perceptron in a three layers network; A perceptron (hidden layer) connecting input (input layer) and output (output layer), to find patterns and to measure the connection between them. Ever since, researchers have contributed to this field of study and the concept of artificial neural networks has evolved over the years to have multiple architectures (models) and calculation methods.

Research on neural networks began by using neural networks to solve AND and OR linear problems. Then neural network architecture was expanded to include multiple perceptrons in the hidden layer, and this was to solve XOR nonlinear problems. The use of multiple perceptron has shown improvement in classification accuracy.

Today, the hot topic of research in neural networks is the field of Deep Learning which is part of a continuous evolution of artificial neural networks. Deep learning is a term often used to describe an artificial neural network that has more than one hidden layer or multiple networks connected together. Deep learning can be a supervised, a semi-supervised, or unsupervised learning. This type of learning has been adopted and used commercially in different industries; bioinformatics, aerospace, military, trade, and others such as exchange markets.

Most technology now depends on deep learning including voice recognition, image processing, natural language processing, computer vision, social network filtering, machine translation, drug design, medical image analysis, material inspection and game programs. The successful performance of deep learning networks has in some cases produced results comparable to or superior to human experts[88]–[91].

### **3.4.2 Non-Artificial Neural Network**

Although neural networks have become very popular and widely used and considered ideal to train machines to learn from experience, there are other learning algorithms that are not biological brain-inspired. These training models are mathematical algorithm-based and have shown in some cases that they can be very successful in solving classification problems. Non-artificial neural networks include Random Forest, Support Vector Machine, Naive Bayes, and Decision Tree. Some of these models will be discussed in the sections below.

#### **3.4.2.1 Support Vector Machine**

Support Vector Machine (SVM) model is a non-neural network learning algorithm, that it is commonly used in supervised learning to solve both regression and classification problems.

SVM uses associated learning algorithms to analyse data to find patterns in the data and separate the classes. The core concept of SVM, is to use hyperplane or a set of hyperplanes in a high-dimensional space to find the largest distance to the nearest training data point of any class.

SVM is commonly used for classification problems to distinguish between the features in the data. For example, if we have a large dataset of mugs and plates, containing a set of features which may include height, weight, width, and price, and the aim was to train a machine to differentiate between the two objects. The model would try to plot the data and find the line between the support vectors with widest margin (hyperplane) that splits the data best, as shown in the image below:

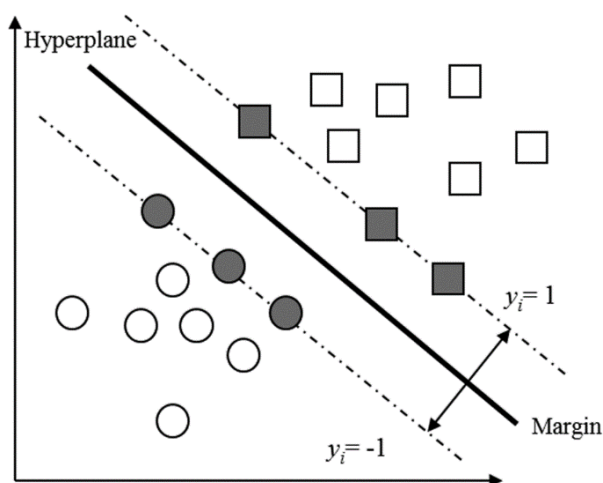


Figure 3-6 Support Vector Machine [92]

The SVM is trained to solve a constrained optimisation problem to maximise the margin between the two groups. This can be achieved by using the Lagrange Multipliers technique. Finding the widest margin between the two groups of a dataset means that any new data is more

likely to fall in either side of that line. This model supports higher dimensions data and handles multiple classes. When training an SVM model the model uses something called C Parameter variable to maximise its margins, whereby when the parameter is low it prioritises simplicity, but this can compromise the classification and allow the model to allow some mistakes in classification. Using higher value for C parameter minimises the misclassification but can sometimes cause overfitting problems. The best practices are to adjust the C parameter as necessary until it gives best results possible.

In a multi-class classification problem, the SVM uses two techniques to separate the classes. The first technique is One-Versus-Rest (OVR), where the model tries to separate a certain group of the data versus the rest of the data in a systematic way. The other technique is One-Versus-One (OVO), where the model tries to separate each class from every other class individually and create hyperplanes for each class. The OVR technique costs less as it performs fewer classification calculations compared to the OVO and takes longer to train. Although, the OVR model classification may be imbalanced, the OVO is less sensitive to imbalanced data.

The visualisation of the SVM results depends on the size of its dimensions; for 2D dimension a line is used to separate the groups, in data with 3D dimensions a plane is used to split the data, in 4D and higher dimensions visualisation becomes a very challenging problem, and that is where the use of the Kernel Trick comes in. The Kernel Trick helps split a 4D+ dimensional data using different methods depending on the classification problem. The main Kernel options are; Linear Kernel, Radial Basics Function, Polynomial, and sigmoid [92–98].

### 3.4.2.2 Naïve Bayes

The Naïve Bayes is one of the non-artificial neural networks and in fact, the model is a probability-based algorithm that uses the Bayes' Theorem to mathematically learn from data. The Bayes' Theorem or often called Bayes' rule, describes the probability of an event, based on knowledge of conditions that might be related to the event. For example, if Alzheimer's disease is related to weight, then, the Naive Bayes algorithm would use the feature weight as one of the key features to predict if people would have Alzheimer's or not. The model assumes that all of the features are independent, and all of the features independently contribute to the probability of an event. For example, if we have weight, diabetes, heart disease as features in a dataset to predict Alzheimer's disease the model would not acknowledge the relationship between weight and diabetes and would rather assume that each one is independently contributing to the probability that the person might get Alzheimer's disease.

$$P(C_K|X) = \frac{P(X|C_K)P(C_K)}{P(X)}$$

Equation 3-2

The Naïve Bayes algorithm is denoted in the equation above. The  $X$  represents the input vector  $X = (x_1, \dots, x_n)$  and  $n$  is the number of features.  $C$  represents the number of possibilities or the classes and  $K$  is the individual possible outcome or class. In the equation  $P(C_K)$  represents the class prior probability,  $P(X|C_K)$  represents the likelihood,  $P(X)$  is the evidence or the predictor prior probability, and  $P(C_K|X)$  then becomes the posterior probability of target class  $C_K$  given input  $X$ .

This model requires less training data and can learn very quickly even in multi-class prediction problems, especially on categorical datasets compare to a normalised numerical variable. However, like other classifiers this model also has some downs, and one of them is the assumption that all features are independent contributors because realistically most of the datasets used in machine learning have one or more features that are related to one another [99–103].

### **3.4.2.3 Decision Tree Learning**

Data classification and prediction are two concepts used when making decisions. Prediction of data is predicting an outcome of the data from previously classified existing data. For example, in the case of predicting if a patient’s health will improve or decline after taken a certain treatment, doctors will look at the existing current status of the patient, and treatment results of other similar patients to help them predict the final outcome, this would be data prediction. Whatever is the outcome of the treatments the doctors will take a decision to classify the resulting data, this would then be classification [103].

The Decision Tree Learning is one of many methods used in ML, it is based on the binary tree. Commonly used in statistics and data mining for different purposes, in ML it is used merely for predictive modelling and classification. The concept of the decision tree is a powerful method that uses a tree-like model similar to the binary tree, to model decisions and work out their possible consequences. Creating a set of connected trees would resemble a forest known as Random Forest or a Gradient Boosted Decision Trees (GBDT).

An example of how this method is used to classify data would be in the work of Sandhya Joshi, “Classification of Alzheimer’s disease (AD) and Parkinson's Disease (PD) by Using Machine Learning and Neural Network Methods” [104]. Joshi investigated a classification model for both AD and PD using machine learning methods and neural networks. He collected a dataset of AD and PD patients containing information about potential risk factors. The aim was to identify the strongest risk factors using the different classification methods and formulas. In the implementation of his work, he used the Decision Tree, Random Forest Tree and a few other ML methods for feature selection and reduction of variables in the data. The Decision Tree was used to extract knowledge from the data using IF-THEN rules. The use of this concept has made it very easy to understand more about the data “especially when the tree is large”[104].

One of the important steps in the process of predictive analytics and classification of data is the feature selection algorithms (sometimes called variable screening) for removing irrelevant, redundant, and noisy information from the data [102]. The decision tree helps identifying the most important variable in the dataset. When the decision tree is applied to the dataset, the top nodes on which the tree is split are the most important variables within the dataset and with this technique, feature selection become automatic. One of the advantages of Decision Tree Learning is the fact that it is easy to explain and the relationships between variables will not affect the overall tree performance and do not expect linear features or even features that interact linearly. The collection of big data often means that a large number of variables and irrelevant data will be stored in the datasets; in order to extract the relevant data, a feature selection technique such as the Decision Tree Learning is required.

### 3.4.2.4 Association Rule Learning

Unlike the decision tree learning, association rule learning is more complicated and more interesting. While decision tree learning is used as a part of the feature selection process to extract knowledge and most important variables in datasets, association rule learning is the concept of finding interesting relations between the variables. Association rule learning relies on using different measures of interestingness to identify strong rules in the database.

This technique is widely researched and used especially by large corporations and supermarket companies with a large inventory of items. This method is powerful as it helps them with prediction and forecasting of the sales, and with decision-making on what to put on sale. For example, one of the rules that can be found in the sales data is  $\{onion, potatoes\} \Rightarrow \{burger\}$ , which indicates that if onion and potatoes were sold together then the customer is more likely to buy a hamburger meat [105]. The advancement in using this method will help the progress of finding answers in any large and complex database. For example, the use of this method in medical research with biological data that are usually very large and complicated, it could make it easier for researchers to discover clues for treatments or signs of symptoms for early diagnosis.

The concept of finding interesting relations between variables is quite a popular and well researched area. An interesting paper was published by Le Queau, from the University of Calgary, titled: “Analysing Alzheimer’s disease gene expression dataset using clustering and association rule mining”[106]. Queau presents various data mining techniques for analysing Alzheimer’s disease Gene Expression Dataset using Clustering and association rule learning.



Biological data such as dataset for people with Alzheimer’s disease or Gene Expression Datasets, contain a lot of variables and typically are complex and hard to process manually. And sometimes because of environmental and experimental factors, the variability of the data can be wide and unpredictable. In this work Queau, demonstrated how the use of ML methods such as clustering association rule learning can be employed to discover or identify interesting patterns in the data.

### 3.4.3 Pre-processing

Real-world/row-data would normally be incompatible to be used by learning algorithms for several reasons; the data often is incomplete, has a lot of missing values and errors, inconsistent, and stored in multiple locations. Data pre-processing technique is used to solve all these data problems and transforming the raw data into an understandable format. Most of the time is spent on data-pre-processing and less time on the employment of the learning algorithms. The figure below demonstrates where data pre-processing takes place in machine learning:

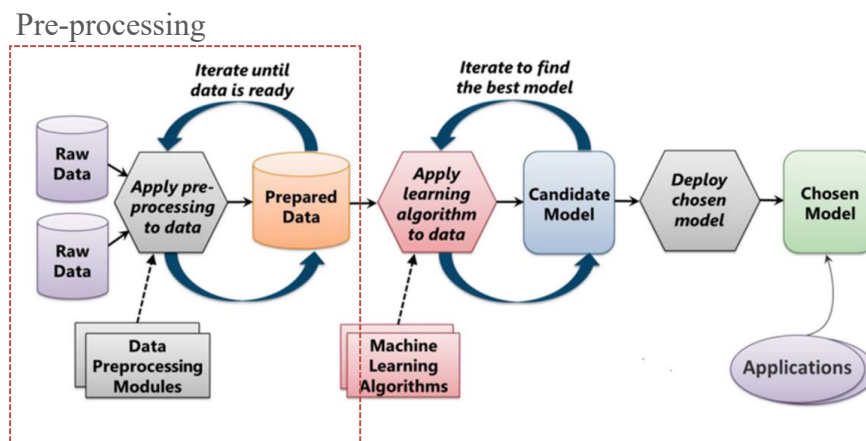


Figure 3-7 Pre-processing Stage in Machine Learning Process [107]

### **3.4.3.1 Data Collection**

The first step would be the gathering of the data often referred to as data collection. The collection of the data would often depend on an ETL (extract, transform, load) process. The data would be extracted from its multiple sources such as web pages, flat files or multiple databases, then transformed to an appropriate format and loaded to a unified location where machine learning would take place [108], [109].

### **3.4.3.2 Missing Values**

Missing values in a dataset would fail the performance of a learning algorithm and draw up an inaccurate inference about the data. Therefore, it is important to solve any missing values in the data. There are multiple techniques to deal with missing data but the two prominent ways are; either delete rows with missing values or use mean, median or mode to replace missing values. The first technique in some cases is acceptable to remove rows with missing values, but this way would reduce the data volume significantly, and also these values can contain crucial information. Depending on the problem and the data type, sometimes it is best to use the second technique and replace missing values with the total mean values as it can give better results [110],[111].

### **3.4.3.3 Categorical Values**

Machine learning algorithms are based on mathematical equations and would require numerical values. The data can often contain categorical values columns, for example, a dataset with 'country' column as a variable, this variable would cause some problem to the learning algorithm. Categorical values will need to be converted to numerical values in a way in which

the new numerical values would have equal importance. This is done by converting the categorical values to variables (columns) and filling the rows with 1/0 values.

Table 3-1 Handling Categorical Values

Before		After			
Country	Age	UK	IRAQ	YEMEN	Age
UK	19	1	0	0	19
Iraq	27	0	1	0	27
Yemen	20	0	0	1	20
UK	28	1	0	0	28
Iraq	30	0	1	0	30

Table 3-1 is an example of how to handle categorical values are handled in the pre-processing stage when conducting a machine learning investigation [112], [113].

### 3.4.3.4 Data Normalization & Rescaling

Other pre-processing methods that might be needed are the removal of unnecessary or repeated variables. Then finally before the exploration of the data, all the data numeric values must be rescaled to range between 0 and 1; this is called data normalization and can be achieved by subtracting the minimum value from all values in the column, then divided by all values by the maximum number. Below is the equation for data normalization where  $x = (x_1, \dots, x_n)$   $x = (x_1, \dots, x_n)$  and  $z_i$  the  $i^{\text{th}}$  normalized data [114], [115].

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Equation 3-3

### 3.4.3.5 Synthetic Minority Over-Sampling Technique (SMOTE)

The pre-processing of the data especially after the completion of data cleaning, normalization, and handling of missing values, can often result in an imbalanced dataset. Imbalanced data can result in compromising the learning process for some classifiers such as the Support Vector Machine, leading to biased prediction and affecting their accuracy. In some cases, where there is enough data, a quick solution would be using a technique called Random Under-sampling to remove data to ensure all classes have equal size. However, it is always beneficial to train models with as much data as possible, and removing data is not always advisable.

The alternative option would be to use a technique called Random Over-Sampling, which would randomly replicate minority data to balance the classes. This method prevents further loss of information from the data, however, the downside is that the data becomes prone to overfitting due to the duplication of the data. Therefore, the best alternative technique would be the deployment of another commonly used technique called Synthetic Minority Oversampling Technique (SMOTE).

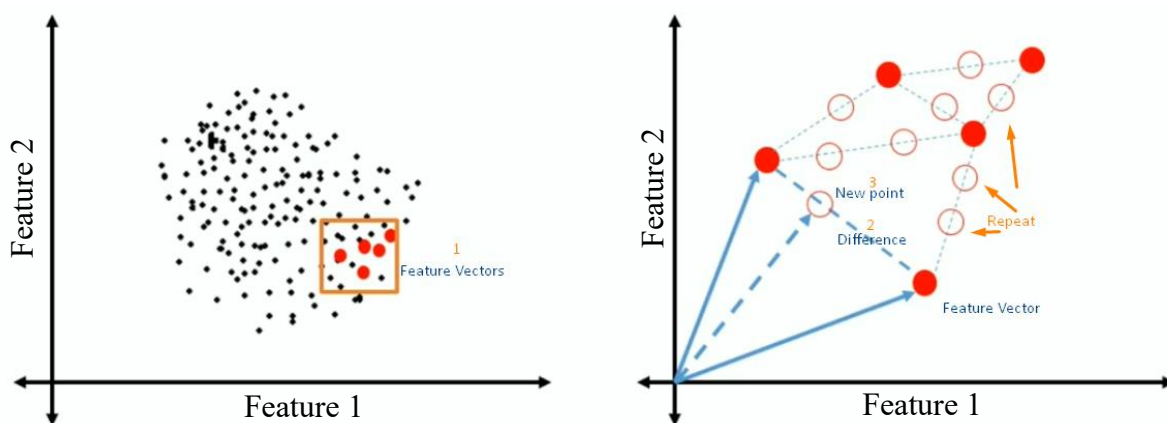


Figure 3-8 Synthetic Minority Over-Sampling Technique (SMOTE) [116]

SMOTE would first identify the feature vectors to resample, then take the difference between the feature vectors and their nearest neighbour. The difference would then be multiplied with a random number between 0 and 1, and the final step would be to find a new point on the line segment by adding the random number to the feature vector. This process would then be repeated for the identified feature vectors. The figure above demonstrates a theoretical process of SMOTE technique to resolve the imbalanced data problem [117–119].

### **3.4.4 Data Exploratory**

One of the most important tasks in machine learning is the data exploration analysis. This part of the process helps us to identify the right dataset needed for the study, by summarising the domain characteristics of the dataset, gain better understanding of the data, uncover relationships between the dataset variables, and extract the most important variables. This part of the process requires a good understanding of statistics and correlation algorithms to extract a good set of data that would help learning models to perform better. However, before data exploration, data pre-processing must take place in which the data is imported from its main sources, missing values handled correctly, converting it to data types that models would understand, and features are scaled correctly depending on the context.

Data exploration analysis is the initial step of the investigation of the problem; through the use of different method a good exploration of the data would give an indication on how the machine learning model would perform and what type of learning method is needed for the investigation. Data exploration analysis would often involve visualisation of the dataset using different methods such as box plots, histograms, scatter plot, and dimensionality reduction like

the principal component analysis technique. The use of this technique will help researchers to build different hypotheses, decide whether new data is needed for the study, know which learning model they can start with, and most importantly draw a roadmap for their data investigation [120], [121].

#### **3.4.4.1 Principal Component Analysis**

One of the major problems that can be faced in machine learning is having data with missing values, noise, or redundant information. These problems result in an over-fitting problem that misguides the classifiers and leads to unreliable results. One of the techniques that are often employed to reduce feature vector dimensions, and to avoid the over-fitting problem is called the Principal Component Analysis (PCA). PCA is used to eliminate ambiguity in the data by finding the correlations between features of multi-dimensional data. If the data has two dimensions then the correlation can easily be plotted on a 2-dimensional graph, and a 3-dimensional graph would be used to plot a dataset with 3 dimensions, however, when the data consist of a larger number of dimensions it becomes impossible to visualize the correlation in one single graph and extract the correlation between the features. Here is where PCA becomes a very useful technique to use to visualize the correlation. PCA calculates the correlation for two features at a time, then plots the correlations on a 2-dimensional graph. The idea is that each data point that is of the same class end up together in the shape of a cluster if they are highly correlated. Normally, each class of the dataset would be colour coded so then it would be easier and faster to distinguish between the data points in a single graph.

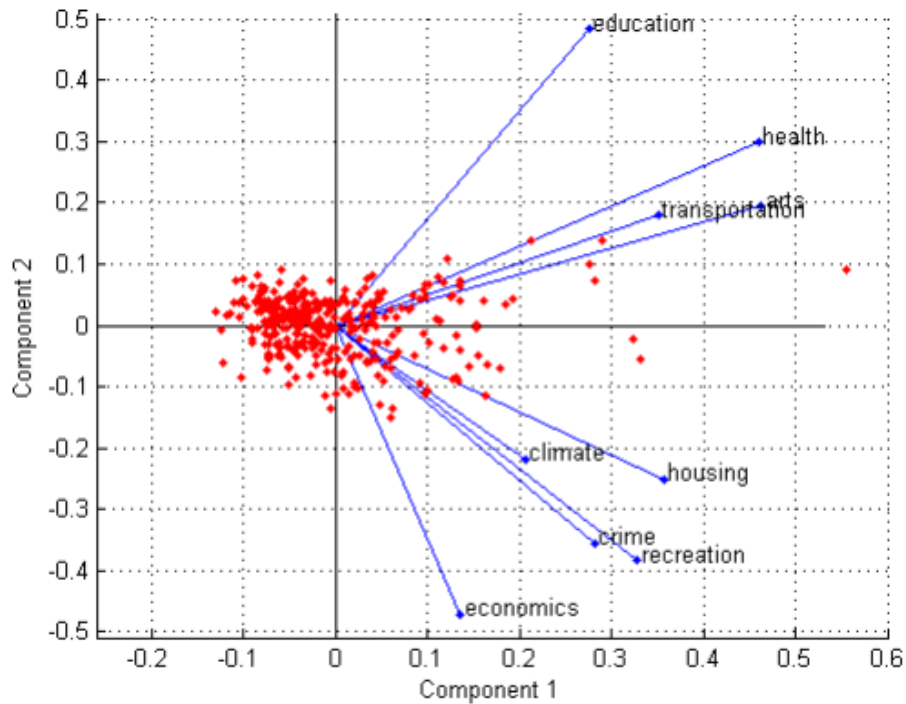


Figure 3-9 Principle Component Analysis [122]

Employing PCA before performing the training process on the data, will help understand what type of learning is needed and what type of model is required for the prediction i.e. if the learning would be linear or nonlinear, and if the prediction is possible or difficult. This is because the more clusters that are formed in the graph means there are features that will guide the learning model to compute a predictive formula [123],[124].

### 3.4.4.2 Independent Component Analysis

Independent Component Analysis (ICA) works in a similar concept to the Principal Component Analysis (PCA) technique, the main difference is that PCA is about finding a variable correlation in the data by maximizing variance and ICA tries to maximize the independence of data features. Unlike PCA, the ICA is designed to solve the blind source separation problem.

Using an example to explain this would be to imagine a dataset containing images of people's faces; PCA would focus on the direction of maximum variance which would be what features these images share that makes them correlated such as brightness and the average face. The ICA, focuses on independent features such as eyes, noses and mouth selectors [125], [126].

### **3.4.4.3 T-distributed Stochastic Neighbourhood Embedding**

A more modern and improved dimensional reduction technique compared to PCA, is the T-distributed Stochastic Neighbourhood Embedding (t-SNE). The t-SNE technique is commonly used for dimensional reduction and the visualisation of the data, it embeds high-dimensional data in a low-dimensional space while preserving the information in the high dimensional space. In simple terms, humans can only visualise 2D or 3D graphs, so if there was data with two dimensions such as height and weight of a person then we can easily plot this information on a 2D graph, and if we add an additional dimension such as gender then this would fit perfectly on a 3D graph, but if these dimensions were to increase to thousands of features then it would be impossible to draw and visualise this data in a graph without using dimension reduction techniques like t-SNE. The t-SNE reduces the dimensionality of data to two or three dimensions, and the algorithm does this with two steps; first, it constructs a probability distribution so that similar objects would have a high probability to be clustered together. Second, the algorithm defines a similar probability distribution over the points in the low-dimensional space, and it minimizes the probability difference between the two distributions with respect to the locations of the points in the space [127], [128].



### 3.4.5 Feature Selection

In ML the concept of feature selection is an essential data processing step to be taken before applying the learning algorithm. Feature selection is a technique that is frequently used in ML to select a subset of the features of a data set in order to build a robust model for learning. This technique helps to give researchers a clearer understanding about the data by telling them the important features of the data and their relationship with each other. There are three main goals of feature selection: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data [129].

Feature selection is a widely used technique by research with big data; researchers explore domains with hundreds to tens of thousands of variables or features. Therefore, many feature selection techniques are used to address these challenges in order to select relevant data and to remove irrelevant, redundant, and noisy information from the data [102]. There are many feature selection searching approaches and these search approaches are categorised in three different classes of methods based on how the selection algorithm and the model building are combined. The three classes of feature selection methods are; filter method, wrapper method and embedded method.

An example to explain the algorithm of the feature selection technique is if we assume that we have data about two types of dementia patient; Alzheimer's disease (AD) and Mild Cognitive Impairment (MCI) subjects. And we have a dataset with 100-dimensional feature vectors. The challenge is to identify the distinguishing features between these two conditions. We generally

know that for example memory test score is one of the distinguishing features between the two types of dementia, and that patients with Alzheimer’s disease generally will score less than MCI patients. However, in order to determine that this is a distinguishing feature and to identify other distinguishing features we will need to develop a feature selection algorithm on the dataset. Identifying the distinguishing features is what would be considered as the problem of the feature selection algorithm. In machine learning, redundant features act as noise, therefore, feature selection acts as noise removal and makes classification easier.

Another example to explain how feature selection works. Let us assume that we have three features X1 and X2 and X3, and  $X3 = 2 * X1 + 1$ , so then the value of X3 will change with the changes in the value of X1. Table 3-2 below demonstrates the changes:

Table 3-2 Relationship between X1 and X2.

Value of X <sub>1</sub>	Value of X <sub>2</sub>	Value of X <sub>2</sub>
1	2	3
2	5	5
10	4	21

The example above shows a direct relation between the two features, which in this case means that one of the features will need to be classed as redundant and removed. This is a very simple example to demonstrate redundancy in feature relationships during the process of feature selection.

A demonstration of how feature selection is used to help researchers; a feature selection technique was employed in the work of Dimitrios Ververidis from the VTT Technical Research

Centre of Finland. Titled: “Feature selection and time regression software: Application on predicting Alzheimer’s disease progress”[130]. Ververidis’ work is constructed on data obtained from Alzheimer’s disease Neuroimage Initiative (ADNI) database, which is publicly available[131]. The subset of ADNI used in his work consists of 2,712 neuropsychological and biomarker features measured over 819 subjects (patterns). 800 subjects were used in the experiments, as 19 out of 819 subjects had no label. In the 800 subjects: 185 subjects are AD patients, 389 are MCI patients, and 226 subjects are healthy. A wrapper method approach was selected for feature selection. He developed a software tool for features selection using Matlab. The tool employs a variant of the Sequential Forward Selection (SFS) algorithm for feature selection. The goal is to discriminate AD, MCI, and Healthy subjects, and to predict the progression of AD using biomarkers.

Most of the work related to this proposed research has mainly been focused on the diagnosis of Alzheimer’s disease from a short term and biological perspective. Only a few research studies have been carried out to predict Alzheimer’s disease before the clinical diagnosis. The challenge is to look for the most accurate way to diagnose Alzheimer’s disease at a very early stage before patients develop any of the symptoms. The success of such a challenge will help improve the research of “Early Treatments” [12].

The Alzheimer’s Association have released a paper “New Diagnostic Criteria and Guidelines for Alzheimer’s disease” to diagnose the disease at an early stage. [21] The current work in the investigation of the disease is mainly focusing on the biological markers and the changes in the brain. It is an easier task to diagnosis someone with Alzheimer’s after signs and symptoms

have reached the stage at which a diagnosis of clinically probable Alzheimer's disease can be made according to currently recommended criteria [132].

### **3.5 Other Related Work**

The works most closely related to this research are involving features that can be considered as late diagnosis of Alzheimer's disease taking for example the work of Gaël Chetelat and Jean-Claude Baron discussed in their *NeuroImage Journal* at the University of Cambridge, UK; titled: "Early diagnosis of Alzheimer's disease: Contribution of structural neuroimaging". Their focus was to use the structural brain imaging of people who have Alzheimer's and healthy people to find patterns or evidence that could potentially show people at risk [132]. To compare this research with the proposed research, they both share the similar concept of prediction and diagnosis. However, this proposed research will focus on searching for patterns and evidence that would help with prediction of Alzheimer's at an early stage through the use of several markers and not just structural brain imaging.

In 2005, Mei Sian Chong and Suresh Sahadevan discussed the diagnosis and prediction of Alzheimer's disease in a published journal called "Preclinical Alzheimer's disease: Diagnosis and prediction of progression". The journal summarises the current research on the accurate prediction - through the use of clinical assessment, psychometric testing, neuroimaging, and biomarkers - of which people with symptomatic predementia will develop clinical Alzheimer's disease.[10] The main focus of this research will be on predicting the disease through finding similarities and patterns between collected data for healthy people and infected patients. The aim is not just to look for biological markers but also to include behavioural markers as well.

Table 3-3 Summary and Critical Evaluation of Related Work

	Modality	Technique	Data Set Details	Pathologically proven data set	Accuracy	Limitation	Validation performed (No. of Folds )
(Klöppel, Stonnington et al. 2008)	MRI	Linear SVM	3-groups AD= 67 CTRL= 91	Yes	96%	Sample size is too small with no justification of missing values.	Leave one out Cross Validation
(Chaves, Ramírez et al. 2010)	SPECT	Apriori-AR mining	AD = 54 CTRL = 43	No	95.87%	Did not mention the how they limited the effect of missing values	Leave one out Cross Validation
(Chaves, Górriz et al. 2011)	SPECT	Apriori-AR mining	AD = 56 CTRL = 41	No	94.87%	The data may contain Missing values which will cause uncertainty	Leave one out Cross Validation
(Zhang, Wang et al. 2011)	MRI+ FDGPET + CSF	SVM	AD = 51 CTRL = 151	No	93.2%	Class Imbalance and missing values	10-fold Cross Validation
(Chaves, Ramirez et al. 2012)	FDG-PET + PiB-PET	Apriori-AR mining	AD = 19 CTRL = 84	No	94.74%	Unproven data with missing values	Leave one out Cross Validation
Robi Polikar et al. (2010)	EEG + MRI + PET	Ensemble based decision fusion	AD = 37 CTRL = 36	No	85.55%	Unproven data with missing values	5-fold Cross Validation
(Chaves, Ramírez et al. 2012)	SPECT PET	Apriori-AR mining	<b>SPECT:</b> AD = 55 CTRL = 42 <b>PET:</b> AD = 75 CTRL = 75	No	92.78%	Unproven data with missing values	Leave one out Cross Validation
(Westman et al., 2012)	CSF MRI	Apriori-AR mining+ SVM	AD = 96 CTRL = 273	No	91.8%	Class Imbalance and missing values	7-fold Cross Validation

(Chaves, Ramírez et al. 2012)	SPECT PET	Apriori-AR mining for feature selection PCA, SVM	<b>SPECT:</b> AD = 56 CTRL = 41 <b>PET:</b> AD = 75 CTRL = 75	No	91.75%	Unproven data with missing values	Leave one out Cross Validation
(Chaves, Ramírez et al. 2013)	SPECT PET	Apriori-AR mining	<b>SPECT:</b> AD = 56 CTRL = 41 <b>PET:</b> AD = 75 CTRL = 75	No	SPECT: 96.91% PET: 92%	Pathologically unproven data with no justification about missing values	Leave one out Cross Validation
A. Veeramuthu et al. (2014)	PET	AR mining	Not Given	No	91.33%	No dataset details, missing values or any preprocessing steps highlighted	No
Tong Tong et al. (2016)	MRI & Cognitive Tests	SVM & RF	NC= 229 SMCI = 129 PMCI = 171 uMCI = 98 AD = 191	Yes	84-92%	Class imbalance and the prediction accuracy is for conversion from MCI to AD	10-fold Cross Validation
Marwa Mostafa Abd El Hamid et al. (2017)	Genetic - SNPs	SVM	NC= 211 MCI = 365 AD = 175	Yes	76.70%	Not predictive of Alzheimer's Disease but it's toward identifying genetic biomarkers	10-fold Cross Validation
Minh Nguyen et al. (2018)	Not Stated	RNN	No Stated	Yes	0.86	Missing Data, Class imbalance and relies on brain imaging scans	No Stated
Solale Tabarestani et al. (2018/2019)	MRI, FDG-PET	MLP	NC= 341 LMCI = 529 EMCI = 255 AD = 333	Yes	Regression	Class imbalance and relies on brain imaging scans	10-fold Cross Validation
Emimal Jabason et al. (2018/2019)	MRI & Cognitive Tests	SVM	NC= 232 MCI = 991 AD = 647	Yes	98.78%	Class imbalance and relies on brain imaging scans	5-fold cross validation
Julian Fritsch et al. (2019)	Trans-literation Evaluation	n-gram Model	Cookie Theft picture from DementiaBank's Pitt Corpus	No	85.60%	It is an improvement to a method introduced by Wankerl et al.	Leave one out Cross Validation

Table 3-3 is a comprehensive review of related work and techniques used, from 2008 up to 2019. Part of the content in this table, between 2008 to 2014 was obtained from a published review paper titled “Early Diagnosis of Alzheimer’s disease using Machine Learning Techniques: A Review Paper” by Aunsia Khan and Muhammad Usman from Dept. of Computing, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Islamabad, Pakistan [133].

Another study towards early diagnosis of Alzheimer’s disease by detecting brain regions related to Alzheimer’s disease using 3D MRI scans is based on eigenbrain and machine learning by Yudong Zhang et al, 2015. Early diagnosis or detection of Alzheimer’s disease (AD) from the normal elder control (NC) is very important. However, the computer-aided diagnosis (CAD) was not ubiquitously used, and the classification performance did not reach the standard of practical use. They proposed a novel CAD system for MR brain images based on eigenbrains and machine learning with two goals: accurate detection of both AD subjects and AD-related brain regions. First, they used maximum inter-class variance (ICV) to select key slices from 3D volumetric data. Second, they generated an eigenbrain set for each subject. Third, the most important eigenbrain (MIE) was obtained by Welch's t-test (WTT). Finally, kernel support-vector-machines with different kernels that were trained by particle swarm optimization, were used to make an accurate prediction of AD subjects. Coefficients of MIE with values higher than 0.98 quantile were highlighted to obtain the discriminant regions that distinguish AD from NC [134].

In this thesis, analysing behavioural markers doesn’t necessarily mean the type of behavioural markers that would signify the existence of Alzheimer’s disease, but the research will

investigate patterns in markers shared by people with Alzheimer’s disease and normal people; these behaviour markers include life record beyond apparent symptoms of Alzheimer’s disease, such as frequent use of substances, type of diet, lifestyle or job career field.

### 3.6 Summary

In this chapter we conducted a literature review on machine learning and the possible techniques that we will use to conduct our experiments to investigate early prediction of Alzheimer’s disease. This literature review aims to give the reader an insight into why we are using computational modelling to conduct this research, the current existing models and methods used in this field, and more specifically more information about models and methods we have used during this study.

Overall, we discussed the concept of using machine learning, the methods used to pre-process and explore the datasets, different types of learning models to train the machine to predicted results, and we also discussed related work conducted by researchers in the same field. This chapter serves as a useful sources to get an up to date knowledge of research toward early prediction and diagnosis of Alzheimer’s disease.

Table 3-4 Literature Review Summary

Discussed Literature Review	Methods We Used From Literature Review	Reason for use
Machine Learning Approaches	We have used Supervised Learning approach for classification and Unsupervised Learning to visualise and explore the data.	The ADNI data in this study is multidimensional and has multiple classes, and the use of Supervised Learning approach allows learning models to be trained on a dataset that has already been labelled with the desired outcome. For the initial analysis of the data before the classification process Unsupervised Learning was used to describe



		hidden structures from the data, and to draw conclusions from the entire dataset [27][135].
<b>Learning Models</b>	We used a mixture of Artificial Neural Network Classifiers, and non-Artificial Neural Networks, as well as combining a hybrid model with a model from both categories to compare the performance and as an attempt to increase prediction accuracy.	In majority of the published work related to this study the performance of the Neural Network models such as the MLP, and Decision Trees based models such as the RF seem to perform best. This is further discussed in section 5.6.
<b>Data Collection</b>	We obtained access to set of data CSV files from ADNI (Alzheimer's Disease Neuroimaging Initiative).	For the purpose of this study and to present the framework and prediction concept we used this dataset to conduct our experimentation. Further data collection directly from patient will be discussed in the future work section. The details of this step is further discussed in section 5.2.
<b>Data Pre-processes</b>	<p>We selected the relevant features to our study based on a comprehensive study of Alzheimer's risk factors provided in Chapter 2. The data was then joined and undergone pre-processing procedures which includes:</p> <ul style="list-style-type: none"> <li>• Handling missing data using both approaches; data deletion, and oversampling techniques using SMOTE as discussed in section 3.4.3.2 and conducting an experiment with each approach respectively.</li> <li>• Categorical data has been converted to numerical values while preserving the importance of the data as demonstrated in section 3.4.3.2.</li> <li>• All of the data variables were normalised to numerical values ranging between 0 – 1.</li> </ul>	Taken into account the type of data we have it was necessary to take these pre-processing steps in order to prepare the data for the employment of classification models. This is further discussed in details in section 5.3.
<b>Data Exploratory</b>	As we have a multidimensional and multiclass dataset we used three of the most commonly used unsupervised learning models to explore the data and understand the type of study we are conducting on the data. The three technique used are PCA, ICA, and t-SNE.	Initial visualisation and analysis of the data was important to us in this kind of study in order for us to understand the behaviour of the data and distribution over the classes. We used some of the up to date and most commonly used techniques in this type of study, which is discussed in more details in section 5.4.
<b>Related Work</b>	Alzheimer's disease is being research from different fields, but one of the major fields of research is studying the disease using computer science specifically machine learning. We looked at published research on the use of machine learning toward early prediction and diagnosis of the disease, to ensure we have an improved study approach and an up to date methodology.	It was necessary and important to understand the current research in the field, to avoid the repetitiveness of research, and to contribute with new findings.

The next chapter will discuss the proposed framework to achieve the objectives in this thesis.

# Chapter 4    **The Machine Learning Classification**

## **Framework**

### **4.1 Introduction**

Although, there are many opinions stating that Alzheimer's disease is a heredity disease, there also many research studies that claim many Alzheimer's disease cases could be prevented by lifestyle changes such as exercise, eating healthily and not smoking. Scientists from different fields might not have a common agreement on the true causes of Alzheimer's disease but ultimately research will determine whether Alzheimer's disease is developed by a combination of risk factors including medical history, family dementia history and lifestyle, or is simply just genetically inherited disease.

Our study focuses on the early prediction of Alzheimer's disease by using a machine learning framework on a combination of behavioural and bio markers data. Here in this chapter we present our strategic framework called EPADf to predict onset Alzheimer's disease by frequently enhancing its dataset based on conclusions obtained from the machine learning and clinical evaluation of the risk factors.

This chapter also contains two logical approaches to conduct this study when using this framework. For the experimentations conducted in this study we considered Coherence Development Patterns (CDP), as the data set we used is limited and not a time series data (see section 4.3 for more information of these logical approaches).

## 4.2 Proposed Framework

The framework consists of five different component; Data Collection, Machine Learning, Clinical Evaluation, Data Enhancement, and Prediction. Each component of the framework has three internal process stages; Supervised Actions, Automated Process, and Final Products. In this section we will explain each component of the framework and elaborate on their internal process stages from a data flow prospective. The diagram below (Figure 4-1) illustrates the framework design using a swim lane diagram.

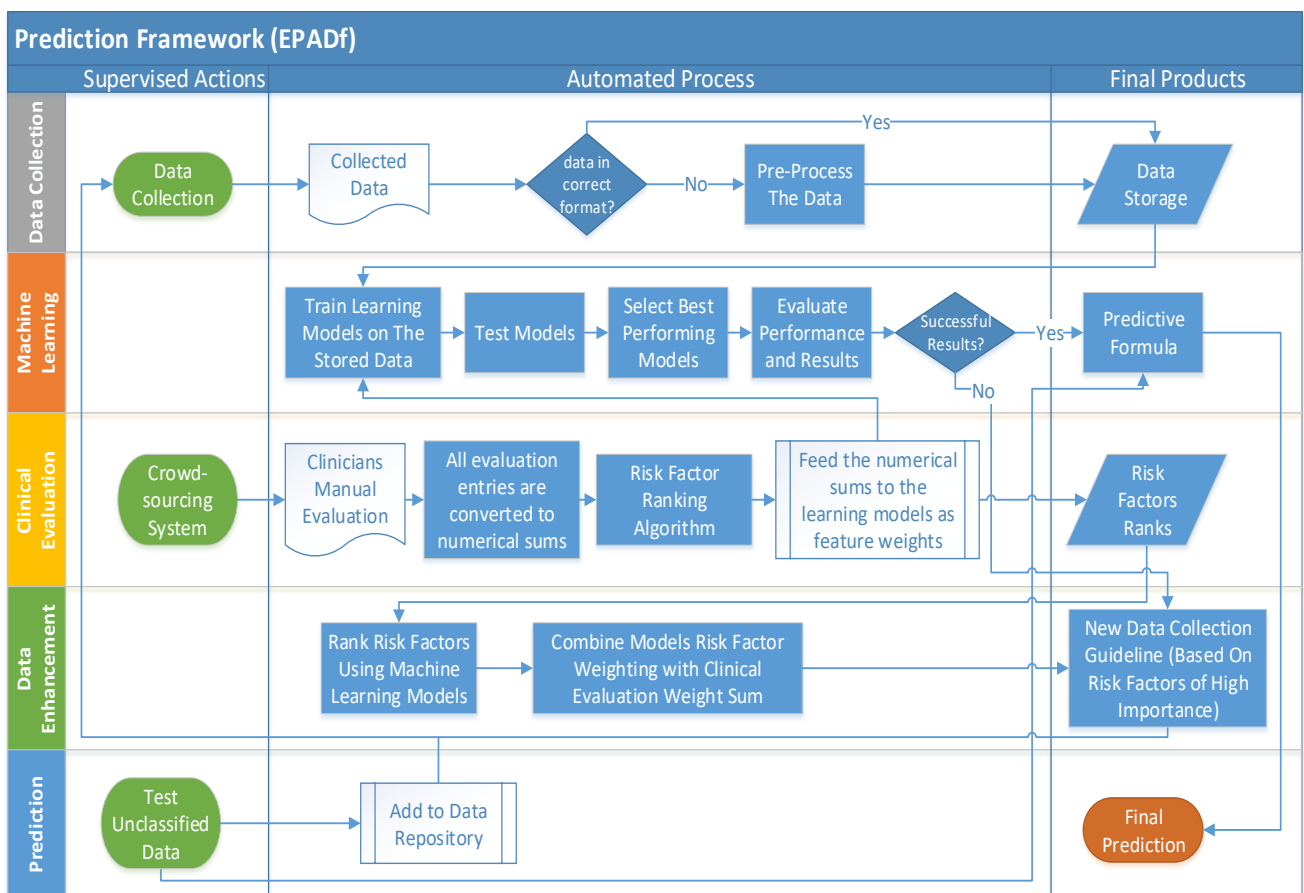


Figure 4-1 Diagram of our framework (EPADf)

## **Framework Components:**

1. **Data Collection:** This part focuses on the collection of the dataset. The data will contain patient specific data related Alzheimer's disease risk factor, focusing on behavioural and biological markers of the disease, obtained from patient's lifestyle, medical history, and demography. The collection of the data will depending on the provided Data Collection guideline produced by the Data Enhancement component of the framework. For our study we used ADNI dataset and selected initial features based on existing research on Alzheimer's risk factors (see section 2.6).
2. **Machine Learning:** Continuous learning technique employed to analyse the constructed dataset and provides a predictive formula, as well as feedback on the importance of the variables (Alzheimer's disease risk factors). This part of the framework and the employment of machine learning techniques can be done using MATLAB, R studio, Python and other integration services. For our study we used MATLAB to build the classifiers and employ machine learning models on the ADNI dataset.
3. **Clinical Evaluation:** A web-based sub-system is developed to calculate guided weighting for each risk factor. This system relies on validated discrete knowledge manually inputted by either system admin, or clinical professionals through crowdsourcing. This component influences the weighting used by machine learning techniques, as well as decision making when adding new variables to the dataset (see section 5.5).
4. **Data Enhancement:** The feedback from both machine learning techniques and clinical epidemic evaluation of the risk factors is used to determine which new data needs to be collected and what variables should be added to the baseline dataset.

5. **Prediction:** Whilst constantly learning from the datasets, the predictive formula will continuously be updated to provide as accurate a prediction as possible. The prediction formula will feed into a live system in which live patient data will be stored. This part of the framework will keep track of patient records and trigger warnings when establishing a possible prediction.

To understand the functionality of EPADf and the meaning of the Figure 4-1 in more depth, the table below (Table 4-1) provides a detailed description on each internal process stages of the framework:

Table 4-1 Explaining How EPADf Works

	Supervised Actions	Automated Process	Final Products
<b>1. Data Collection</b>	This section requires supervised interaction with the framework to determine what data is collected and combine with data previously used data. For example if previously we had data for patients with 10 features and we needed to collect more data as recommended in the ‘Clinical Evaluation’ component, then in this component we merge the old features with new features or to remove unnecessary features.	In this part of the component is responsible for pre-processing the data set by apply normalisation techniques, dealing with missing data and balancing the data.	This component ultimately produces a machine learning ready dataset.
<b>2. Machine Learning</b>	The deployment of machine learning models is automated.	The machine learning component uses the produced data by first component to train the learning models. In this component the framework will also consider the features ranking produced by the third component and auto rank the features in the dataset.	This component will continuously employ multiple learning models to classify the data, then consequently, choose the best model and present it’s learning as a predictive formula.
<b>3. Clinical Evaluation</b>	Clinicians use this system to evaluate the influence of risk factors and their interrelationships. For example an expert cardiologist puts their knowledge and clinical	The clinical evaluation input is converted to numerical values then rank risk factors importance by using an	This component will produce a ranking of risk factors based on clinician’s clinical experience, as well as the

	instinct on the relationship between fast food and blood pressure. The more we have of this kind of data the closer we get to mapping out the connection between the risk factors and rank their importance based on their contribution in the development of other risk factors.	algorithm as presented in section 5.5.3. This feature ranking will indicate which risk factor is more important and this will help with the enhancement of the dataset.	machine learning models learning experience.
<b>4. Data Enhancement</b>	For the production of the data enhancement guideline there are no supervised actions required in this component.	The framework ranks the risk factors using both machine learning as well as the clinical evaluation input.	The overall ranking will be presented as a data collection guideline to indicate what the most influential risk factors are. Based on this guideline new data related to the most important risk factors is to be collected and feed into the framework.
<b>5. Prediction</b>	To diagnose a patient and to predict the disease, in this component we feed unlabelled data and let the framework process the data and categorise it.	The inputted data will be fed through the predictive formula produced by the machine learning component. All new input of data in this component will also be added to the overall data storage.	The framework will use the unlabelled data and classify it by percentage to indicate if the patient is at risk of developing Alzheimer's Disease.

### 4.3 Logical Approaches to Predict Alzheimer's disease

The ADNI database store information about patients' medical history, partial lifestyle, genetics, characteristics and other information related to Alzheimer's disease. And it is important to extract knowledge about Alzheimer's disease by investigating the risk factors in existing databases. With effective use of computer science, specifically machine learning and data analysis techniques speeds up the investigation toward early prediction of the disease and provide information that are more accurate.

Here in this section of the thesis we present two logical approaches for the discovery of Alzheimer's disease development patterns. We called these two approaches, Sequential

Development Patterns (SDP) and Coherence Development Patterns (CDP). These two approaches are to be considered in the investigation of Alzheimer's disease development patterns. Due to the complexity of Alzheimer's disease development, it is assumed that the pathological development is triggered over time in two possible ways;

1. Risk factors develop in sequence, which then leads to the development of Alzheimer's disease.
2. A coherence of all risk factors developing at the same time to trigger the development of Alzheimer's disease.

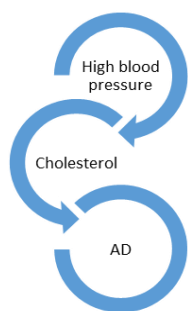


Figure 4-3a Example of Sequential Development Patterns

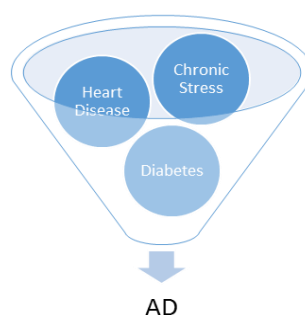


Figure 4-3b Example of Coherence Development Patterns

**Note:** Please note the examples used in Figure 4-3a and Figure 4-3b are hypothetical and are only used to demonstrate the concept of the two logical approaches.

**Sequential Development Patterns (SDP)** – The SDP approach focuses on finding patterns of Alzheimer's disease development based on the date at which the risk factor has occurred in the patient's record, as well as looking at the risk factors occurring before and after it. For example, to investigating the development of Alzheimer's disease in a dataset full of Alzheimer's disease patient medical records, we will start by looking at the diseases that have occurred before Alzheimer's disease, then investigating

each of the diseases in the same way (in reverse) and compare the results for each patient in order to find a sequential development pattern.

This approach could be used to investigate patterns in both lifestyle datasets and medical history. The success from using this approach may possibly designate a pattern that could perhaps help with early prediction and prevention of Alzheimer's disease. Such an approach will provide an in-depth insight into the sequential connection between risk factors of Alzheimer's disease. However, the down side of such an approach is that existing databases might not have the desired record needed to use to produce accurate results.

**Coherence Development Patterns (CDP)** – The CDP approach is unlike the SDP, this approach looks at the common combination of disease and risk factors that occur in most Alzheimer's disease patients' datasets. As shown in Figure 4-3, the CDP approach looks at the available information in the datasets and search for the combination of risk factors that are responsible for Alzheimer's disease development. The advantage of such an approach is that it will be possible to use it to investigate current existing Alzheimer's disease databases such as ADNI.

Due to the fact that it is not clear what causes Alzheimer's disease, these approaches are worth using for the investigation of Alzheimer's disease development patterns and it would be best to use both together. As for CDP, it provides a set of patterns combined with multiple possibilities, while SDP analyses the patterns in depth to find the highest possibilities that trigger Alzheimer's disease. In our study we used CDP approach as the dataset we obtained didn't contain enough information and structure to conduct time series study.



# Chapter 5    **Implementation of the Framework**

## **5.1 Introduction**

We obtained a dataset from Alzheimer’s disease Neuroimaging Initiative (ADNI) that contains data for over 1,880 subjects’ records related to some of the Alzheimer’s disease risk factors. Although, due to the limitation of the amount of risk factors’ data contained in the ADNI dataset it might not be possible to achieve an accurate prediction. We are looking to use this dataset to deploy a machine learning algorithm and rank the importance of the risk factors in the dataset.

Machine learning is a computational technique used for automated or semi-automated extraction of knowledge from large datasets. This is aimed to give computers the ability to learn from data and to classify or give productive values. It is inspired by the human biological ability to learn and find answers to questions. Machine learning has two different types of learning “signal” and “feedback”, and three different categories; Supervised Learning, Unsupervised Learning and Reinforced Learning. These three categories help systems to recognise patterns, learn from unknown data and interact with the environment. The complexity of this field has opened a window for innovation and research, as a result of this, different approaches were used such as the Decision Tree learning, Neural Network and the Bayesian Networks [136].

In this chapter we use supervised and unsupervised machine learning and experimented using 5 models. The study in these experiments includes classification to predict the disease based

on the dataset obtained and an analysis aiming to mathematically order which of the risk factors indeed have an impact.

This chapter is split into different sub-sections discussing the methods used to conduct the experiments by phases; all the way from data access and collection to the deployment of the machine learning model.

## **5.2 Data Access and Collection**

In this thesis we deployed the machine learning models on Alzheimer's disease dataset using MATLAB. The data was obtained from an organization called Alzheimer's disease Neuroimaging Initiative (ADNI), which was first launched in October 2004 with the aim to collect data that would help to find more sensitive and accurate methods to detect Alzheimer's disease [137]. Access was granted after the submission of a research proposal to ADNI, the data is unidentifiable data and available for researchers to use, no ethical approval was required.

The ADNI, was first launched in October 2004 with the aim to collect data that would help to find more sensitive and accurate methods to detect Alzheimer's disease [137]. In the first phase of ADNI (ADNI 1) the study gathered thousands of data for a limited number of Alzheimer's disease markers, this includes brain scans, genetic profiles, and biomarkers in blood and cerebrospinal fluid that are used to measure the development of Alzheimer's disease.

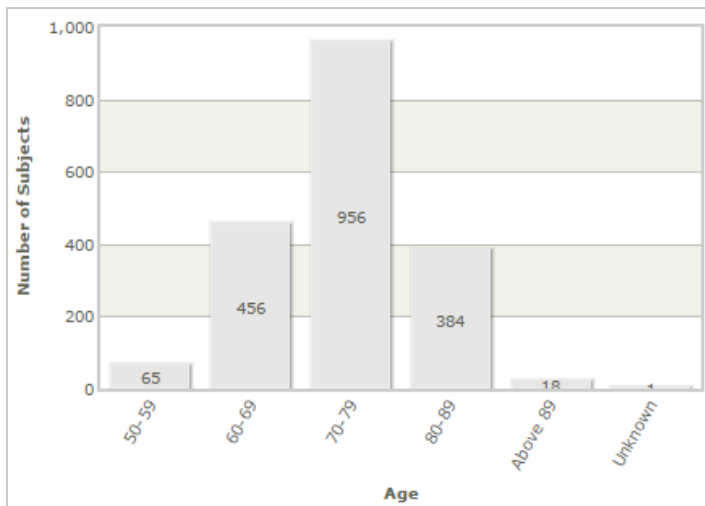


Figure 5-1 Content of ADNI Data

The collected brain-imaging data in the first study OF ADNI was gathered through the use of techniques such as positron emission tomography (PET), including FDG-PET (which measures glucose metabolism in the brain); PET using a radioactive compound (Florbetapir F 18) that measures brain

amyloid accumulation; and structural MRI [137]. The focus of this ADNI study was to detect Alzheimer’s disease in the brain to assist doctors with a more accurate diagnosis of the disease. However, as the study evolved, and Alzheimer’s disease diagnosis became easier the goal of ADNI has expanded toward the detection of the disease at a pre-dementia stage. With the successful standardized data collection methods used, the imaging and bio markers data was made available to scientists from around the world to conduct cohesive analysis and research into Alzheimer’s disease.

The second phase (ADNI2) began in 2011 as an expansion of the ADNI 1 goals, the aim was to help doctors and specialists to detect markers of Alzheimer’s disease and to accurately track progression of the disease. The aim was also to use this data for clinical trials and to measure the effectiveness of potential interventions. ADNI 2 phase was expanded to collect demographic, behavioural and biological data of Alzheimer’s disease patients. ADNI provides a large dataset of tests, measurements, and observations taken during the progression of normal

control to Alzheimer's disease patients. ADNI has data from approximately 1,880 patients; 843 females, 1,036 males and 1 unknown.

At the start of ADNI, data was collected from a small number of patients at different visits. In this study a patient's first visit will be referred to as baseline visit (BL). Some patients have returned for observations every 3 months since baseline, whilst some would turn up to some visits but not consistently.

When the second phase began the number of observations was extended in order to capture more information to help early detection of Alzheimer's disease. This resulted in some patients who had their data recorded in phase ADNI 1 having missing data for the new observations. Similarly, to ADNI 1 in phase ADNI 2 there were also patients who didn't have observations taken for them consistently on every visit. This shows that the datasets contain a lot of missing data which will cause inaccurate results if we were to use them in the experiment. For this exact reason the experiments done in this work will only be based on data from phase 2. Working with ADNI data for machine learning and data analysis was not a straightforward task, especially, with this amount of data that was collected at different phases on different visits, with missing rows, and inconsistent and imbalanced data.

### **Forming a Baseline Dataset**

ADNI provides comprehensive details regarding its data and the variables representation in the dataset. The ADNI data was stored in over 60 flat files, with approximately 70 columns each. We downloaded and extracted the relevant data to perform our first experiments then gradually

extracted more variables in attempts to improve the machine learning classifier’s performance. Overall, we obtained all the data files that contain lifestyle, demography, medical history and family dementia history. Here is an overall list of data attributes of the dataset some of which represent Alzheimer’s disease risk factors.

Table 5-1 Dataset of attributes

<b>Variable</b>	<b>Representation</b>	<b>Variable</b>	<b>Representation</b>
AGE	Subject age	SMOKING	Smoker or non-smoker
MUM_DEMENTIA	If mother had dementia	DRUG	Consumed drugs or substances
DAD_DEMENTIA	If father had dementia	ALCOHOL	If is alcohol drinker or not
GENDER	Male or Female	ALLERGY	If have any known allergy
RACE	White / Black / Mixed etc.	WORK_CAT	Work category
ETHNICITY	Hisp/Latino or Not, or Unknown	WEIGHT	Weight
EDUCATION	Education Category	HEIGHT	Height
MARRIED	Married/Divorced/Single...	APOE4	Have APOE4 gene or not
ENERGY	Experiencing lack of energy	DEPRESSION	Have history of feeling depressed
DIZZY	Experiencing Dizziness	CARDIOVASCULAR	Have any cardiovascular diseases
DROWSY	Experiencing feeling drowsy	BLOOD_PRESSURE	Blood pressure readings
VISION	Have vision problem	HEART_RATE	Heart rate readings
HEADACHE	Experiencing frequent headaches	MMSE	Mini Mental Score Examination

The data was in comma separated values files (CSV) and had several pre-processing challenges before deploying the machine learning models. The data is unidentifiable data and subject's information was replaced with a single numeric identifier called RID to help identifying the subject's data across the ADNI dataset files. We convert the extracted files to a relational SQL Server database, which helped us to easily extract correct and accurate relevant data to this experiment for each subject.

This was our first step in following a complete Extract, Transform, and Load (ETL) process to construct a baseline Alzheimer's disease dataset from the ADNI data files, before employing machine learning models. After extracting the ADNI dataset and converting it to a relational database, the relevant data to our experiment was further extracted and transformed using a combination of server-side scripting language and R. The transformation of the dataset involved data cleaning and balancing such as removing data with missing values and converting the text values to numeric representation [138]. The next section discusses all of the pre-processing steps in more detail.

### **5.3 Pre-processing**

The pre-processing stage was a difficult stage due to the fact the most machine learning classifiers, special neural networks fully depend on mathematical calculation. Therefore, the data must be converted from text to be all numerical. The ADNI dataset contained some numerical parts but a lot of it was text data. Beside this problem the data was imbalanced distribution over classes, had a lot of missing values, and categorical data. These problems were resolved as discussed in chapter 3 (see Table 3-4).

The work began to tackle these challenges step by step starting with dealing with categorical data values. An example of this would be the gender and marriage status of patients. For gender they had ‘Male’ or ‘Female’, and for marriage status the options were, ‘Single’, ‘Married’, ‘Widow’, and ‘Divorced’. This type of data will need to be numerical data in order to employing machine learning, however, if we were to convert for example ‘Single’ to 1, ‘Married’ to 2, ‘Widow’ to 3, and so on, this would also cause the machine learning model to be more confused as ‘Widow’ would automatically gain bigger weight than ‘Single’ and this is not necessarily true. As described in Section 3.4.3.3 the way this categorical data would be resolved is as follows:

<b>Original Format</b>		<b>New Format</b>				
Patient	Marriage Status	Patient	SINGLE	MARRIED	WIDOW	DIVORCED
1	Single	1	1	0	0	0
2	Single	2	1	0	0	0
3	Divorced	3	0	0	0	1
4	Widow	4	0	0	1	0
5	Married	5	0	1	0	0

This is done by converting the categorical values to variables (columns) and filling the rows with numerical values 1 for true and 0 for false, as show in the table above.

Now that the categorical data problem was resolved for every categorical column in the dataset, the second challenge was to normalise and rescale the rest of the data. Un-normalized data can cause the learning of the classification models to be conditioned by the values of the features, for example having a feature that ranges between 0 and 1, and another feature ranging between 100 – 500, a small variation in the first feature can be more influencing than a big variation in the second feature. Therefore, the data was converted to numerical values as shown in the

example above and high ranged data was normalised to range between 0 to 1, using this equation for data normalization  $z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$  where  $x = (x_1, \dots, x_n)$  and  $z_i$  the  $i^{\text{th}}$  normalized data.

The third challenge in the pre-processing stage is the handling of missing values, and imbalanced data. As discussed in section (3.4.3.2) and (3.4.3.5), this can be solved by either removing patients with missing data and balancing the data by further removal of patients, or alternatively replace missing data with average values and use Synthetic Minority Over-Sampling Technique (SMOTE) technique to balance the rest of the data [110],[111]. Since we had a reasonable amount of data, we removed patients with missing values and removed imbalanced classes in our initial experimentation then the latter for our second experimentation.

## **5.4 Data Analysis and Visualisation**

The machine learning experimentation work carried out in this thesis was conducted in three different parts; initial construction and investigation of a baseline dataset, enhancement of the dataset and ranking of variables, and deployment of machine learning models on the enhanced dataset. This section discusses the data before and after the pre-processing procedures, from an analytical approach with data visualisation techniques.

### **5.4.1 Phase 1: Initial Experiment Dataset**

This part of the research is focused on early prediction of Alzheimer's disease for the pre-dementia stage, the experiments are carried out on data that relates to the behavioural markers and the variables selected are matching the features discussed in the risk factors section (2.6)



of this thesis. Table 5-2 below shows the risk factors that we used for our initial experiment in phase one of the study:

Table 5-2 AD Risk Factors used in the initial experiment

<b>Medical History</b>	<b>Lifestyle</b>	<b>Demography</b>
Diabetes	Alcohol	Age
Cholesterol	Smoking	Education Field
Heart Disease	BMI	Race

After the performance of data pre-processing and the removal of missing data the total data volume was reduced to 650 subjects from 1880, and the distance between the data volume for each class has changed and resulted in the data being imbalanced. Figure 5-2 shows an exploration of the data after the deletion of subjects with missing data. This resulted in a very small number of subjects who are classed as Significant Memory Concern (SMC) compare to the rest of the classes, which makes the dataset imbalanced.

Running the experiment on imbalanced data will not give a correct accuracy as it will mislead the artificial agents to give insufficient results. To make the data more balanced, all subjects with SMC class were removed from the dataset. The section shows an overview of the final dataset which is used in the following experiments.

**Data with imbalanced class:**

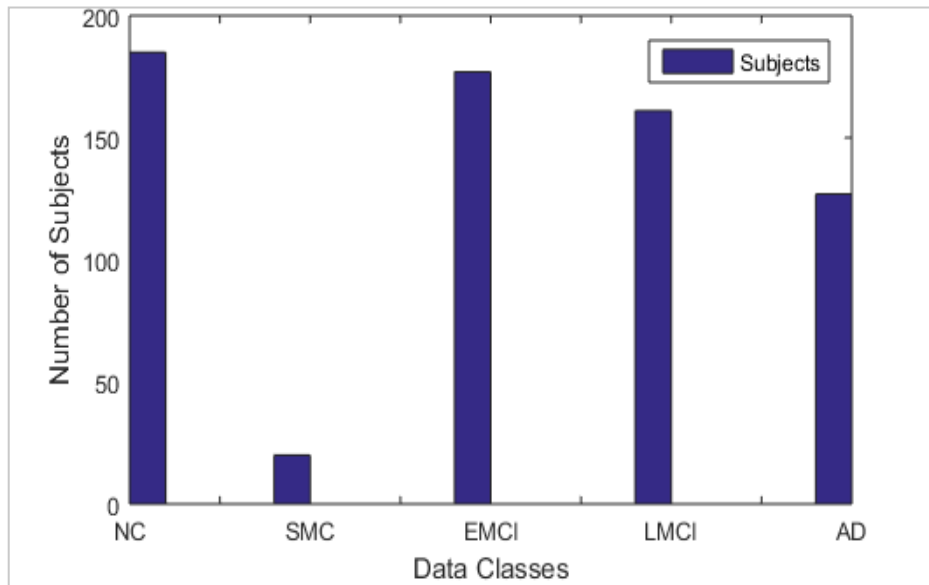


Figure 5-2 Initial Experiment ADNI Data Volume

**Data summary after removal of imbalanced class:**

Table 5-3 Final Dataset for Initial Experiment

<b>Class</b>	<b>Data Volume</b>
NC	185
EMCI	177
LMCI	161
AD	127
<b>Total</b>	<b>650</b>

During this phase the number of subjects was reduced by almost 50%, now with a total remaining number of 650 subjects to be studied. Table 5-3 gives an overview of the dataset for each of the four remaining classes (Labels).

To explore the dataset suitably, different data analysis toolboxes on MatLab, MiniTab 16 were deployed, this includes t-Distributed Stochastic Neighbour Embedding (t-SNE), Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Square Prediction Error (SPE).

The t-SNE is a Machine Learning algorithm commonly used for dimensionality reduction in data visualisation. t-SNE is applied on the dataset, giving the results shown in Figure 5-3, it shows mixed clusters detached apart. Ideally the perfect result that we had hoped for is that each cluster will contain a majority of one class. However, as shown in Figure 5-3 the clusters have an almost equal mixture from all classes, which, means that the algorithm struggled to differentiate between the categories of the subjects. Although, this algorithm shows a large distance between the clusters which means on a dimensional level it managed to differentiate between the variables (risk factors).

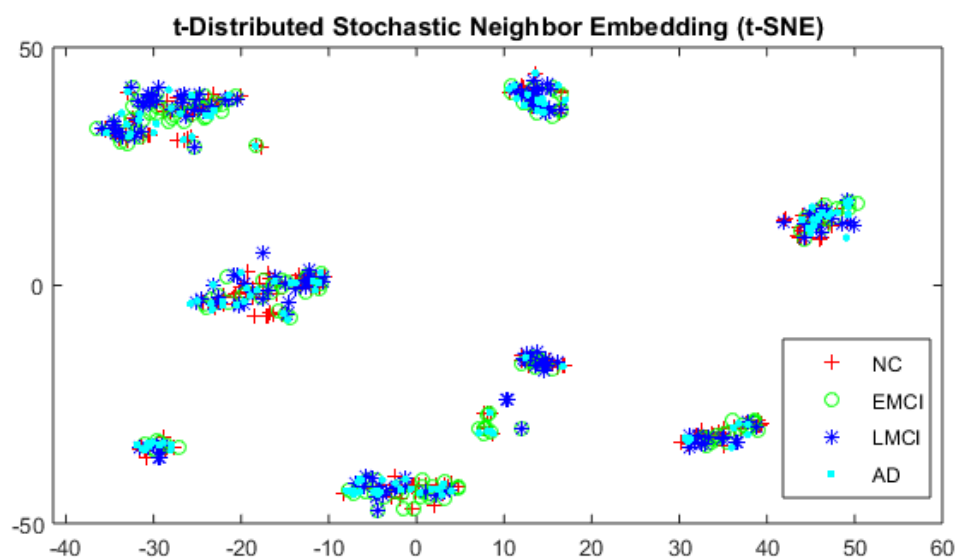


Figure 5-3 Explore of Data Using t-SNE on

Other useful methods used to explore and visualize this dataset are the Principal Component Analysis (PCA) and Independent Component Analysis (ICA), both Figure 5-4 and Figure 5-5 illustrate the use of these techniques on MATLAB. PCA is used to emphasize the variation, dimensions reduction and bring out the strongest patterns in the dataset. ICA is a method for

separating a multivariate signal into additive subcomponents; for more information see section (3.4.4.1 and 3.4.4.2).

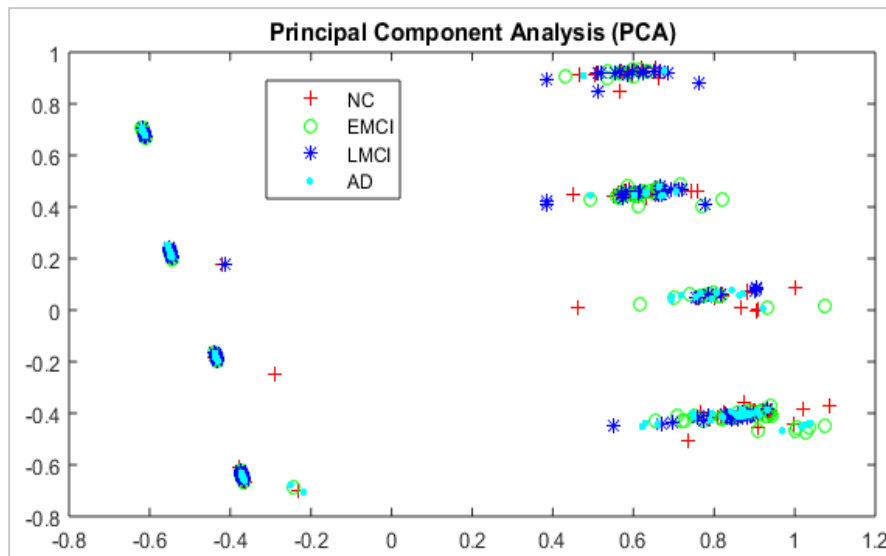


Figure 5-4 Explore of Data Using PCA on MATLAB

Ideally, what we want to see in both Figure 5-4 and Figure 5-5 is four clusters and each cluster has only one shape representing a specific data class. But in these figures we see that both models struggled to separate the classes this is because the data lacks of features that are unique to specific classes, which, ultimately indicates at this level the classification process might not present a positive prediction.

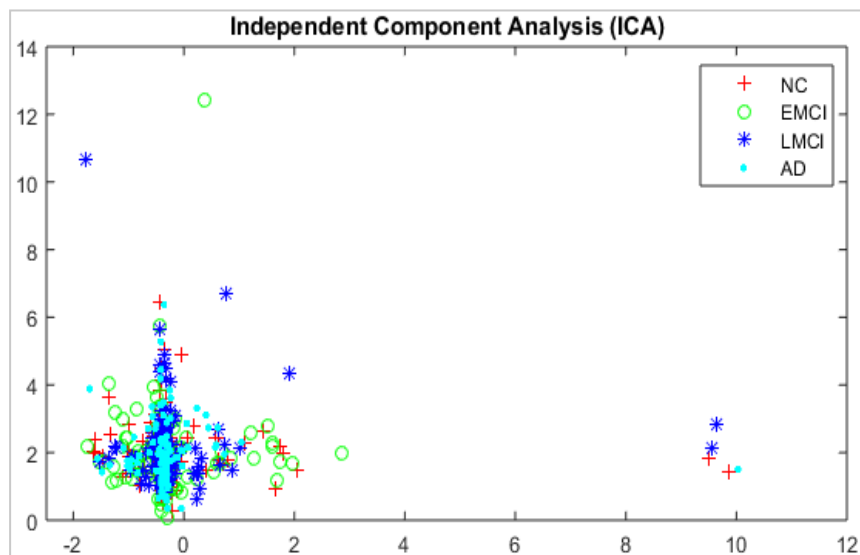


Figure 5-5 Explore of Data Using ICA on MatLab

Furthermore, the data was explored using the Square Prediction Error (SPE) plot to measure the quality of a predictor. The graphs and coefficients result in Table 5-4 show an apparent indication that the type of study will be a nonlinear regression.

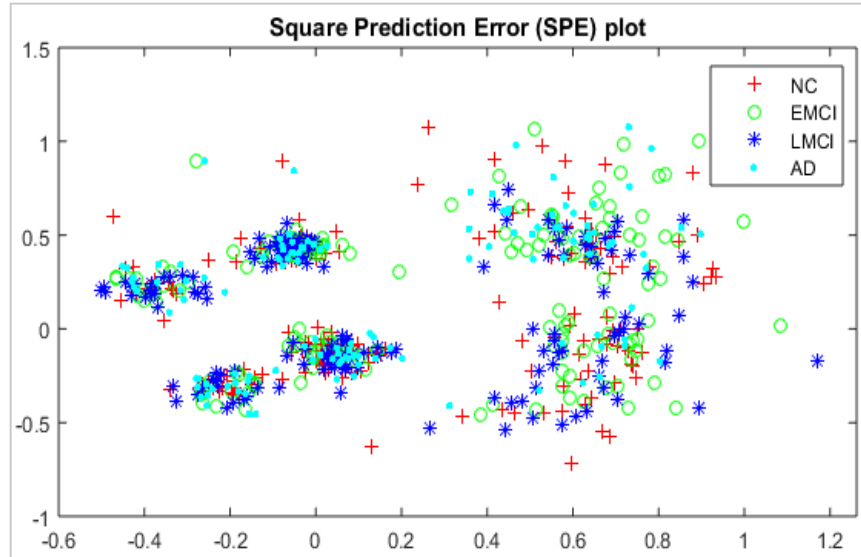


Figure 5-6 Explore of Data Using SPE on Matlab

Table 5-4 PCA Coefficient for Each Variable

Variables / Components	PC1	PC2	PC3	PC4
Diabetes	0.001793	0.027300	0.136991	0.094909
Cholesterol	-0.071206	0.105367	-0.247695	-0.485408
Smoking	0.446737	0.293874	0.043807	-0.054670
Smoking Years	0.367998	0.264838	0.038639	-0.003285
Smoking Per Day	0.439407	0.278448	0.050468	-0.052177
Quit Smoking Period	0.391725	0.239360	0.061360	-0.015511
Heart Disease	0.036513	-0.039123	0.116216	0.451622
Alcohol	0.305641	-0.402629	-0.307793	0.006958
Alcohol Duration	0.299719	-0.378212	-0.278685	0.025604
Alcohol Duration Since End	0.289615	-0.396444	-0.285074	0.011434
Gender	-0.122825	0.249310	-0.410394	-0.188994
Race	0.038631	-0.033482	0.023795	-0.186449
Education	-0.019303	-0.134047	0.073096	-0.044850
AGE	0.065078	0.053068	0.087619	0.378552
BMI	-0.060827	0.131281	-0.312178	0.455934
Weight	0.090059	-0.214031	0.358343	0.073775
Height	0.108802	-0.292391	0.478053	-0.293493

The coefficient data shows an apparent indication that the experiment result of the predictors will not provide a clear outcome. From the data exploration it is very unclear which variable is the highest predictive factor, which shows that the classifiers might not give a very clear classification or definite predictive value. In this case future experiments involving more variables and underlying data will provide a better outcome.

### 5.4.2 Phase 2: Extending the Dataset and Ranking Risk Factors

After the negative classification in the initial experiment we extended the dataset to include more variables and employed machine learning models to rank the variables by importance, in an attempt to visualise the data and investigate the dataset further. Table 5-5 contains all of the variables we used during this part of the investigation:

Table 5-5 Extended Dataset of attributes (Phase 2)

Variable	Representation	Variable	Representation
AGE	Subject age	SMOKING	Smoker or non-smoker
MUM_DEMENTIA	If mother had dementia	DRUG	Consumed drugs or substances
DAD_DEMENTIA	If father had dementia	ALCOHOL	If is alcohol drinker or not
GENDER	Male or Female	ALLERGY	If have any known allergy
RACE	White / Black / Mixed etc.	WORK_CAT	Work category
ETHNICITY	Hispanic / Latino or Not, or Unknown	WEIGHT	Weight
MARRIED	Married/Divorced/Single...	APOE4	Have APOE4 gene or not
ENERGY	Experiencing lack of energy	DEPRESSION	Have history of feeling depressed
DIZZY	Experiencing Dizziness	CARDIOVASCULAR	Have any cardiovascular diseases
DROWSY	Experiencing feeling drowsy	BLOOD_PRESSURE	Blood pressure readings
VISION	Have vision problem	HEART_RATE	Heart rate readings

HEADACHE	Experiencing frequent headaches		
----------	---------------------------------	--	--

In the same steps as part one, before we conduct the experiment, we ran a few essential steps to summarize and understand the data better. We explored the dimensions of the new dataset after the extraction process and compared it with the initial extract of approximately 1,880 patients that are of interest to us. The final dimensions were 1,635 instances and 26 attributes (some subject were removed because they had no data related to the extended attribute). Then we ensured that our data was all numeric and understandable by the models as well as exploring the classes' distribution. The final dataset had 5 levels of class attributes "AD", "NC", "EMCI", "LMCI", and "SMC". The aim is to find the importance of each variable (risk factor) for each of the 5 class levels. This experiment was not to predict or to classify the patients, so we didn't need to remove classes that had fewer patients. Table 5-6 demonstrates the levels of class attributes and the data distribution of instances.

Table 5-6: Levels of the Class Attribute (Phase 2)

<b>Class</b>	<b>Representation</b>	<b>Frequency</b>	<b>Percentage</b>
NC	Normal Control	267	20.36%
AD	Alzheimer's disease	310	23.64%
SMC	Early Mild Cognitive Impairment	232	17.69%
EMCI	Late Mild Cognitive Impairment	432	32.95%
LMCI	Significant Memory Concern	70	5.33%

The first analysis of the dataset was to plot the correlations summary using the 'Pearson' covariance method [139][140]. The 'Pearson' correlation method is the most common way to quantify attribute relationships, which measures the relationship from -1 to +1. Having +1 means a strong positive relationship, 0 represents no relationship, and -1 represents strong negative relationship. If the relationship is positive it means that as one variable increases the

related variable increases also and vice versa. In a negative relationship it works differently as it means if a variable increases the related variable will decrease.

Figure 5-7 illustrates the correlation plot for the Pearson correlation method. The output shows an interesting correlation between the variables. Showing positive correlations between demographic attributes such as gender and weight, gender and height, and work and education. The negative correlations are mainly between symptoms and medical history attributes.

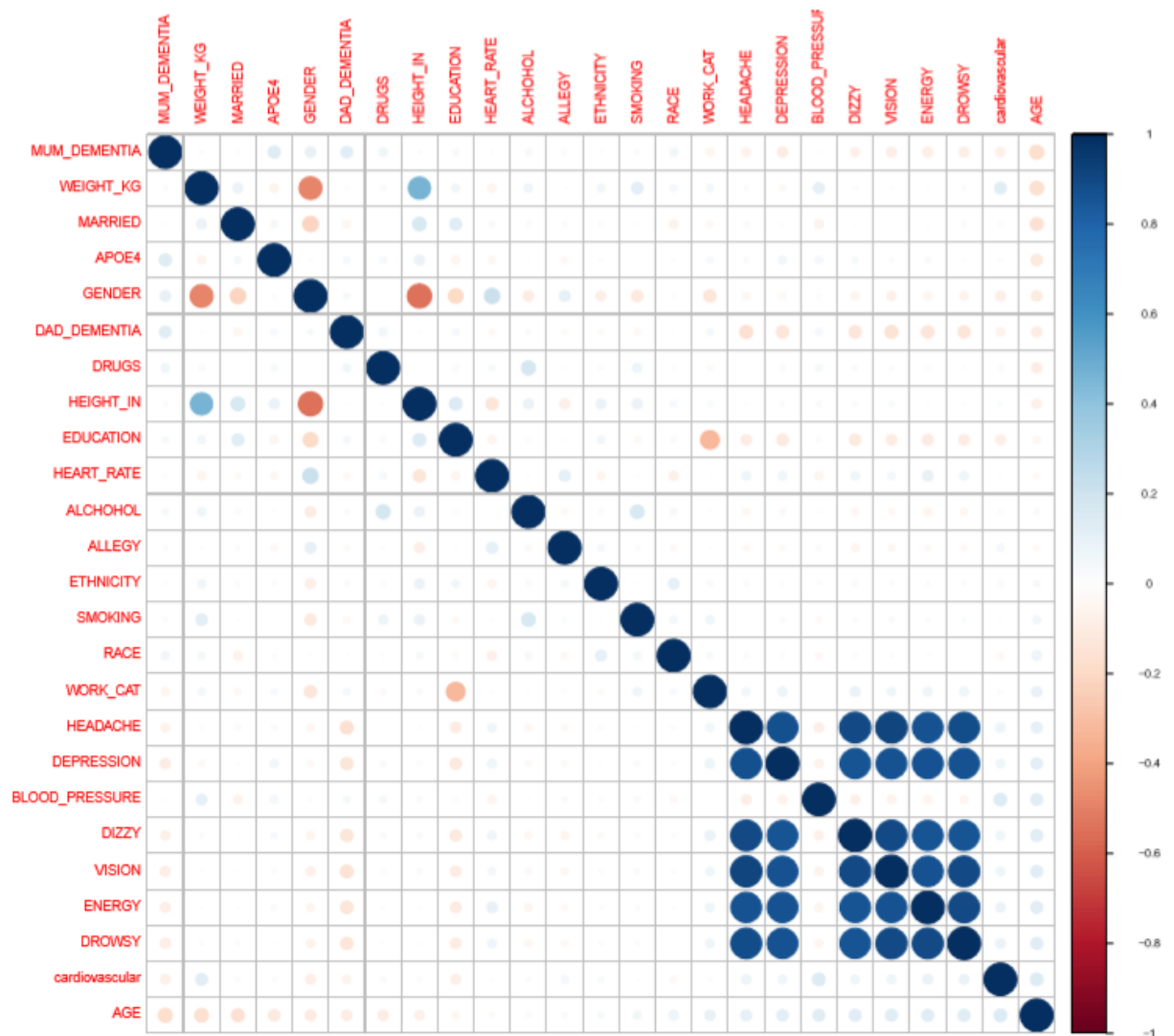


Figure 5-7: Pearson Covariance Method (Phase 2)



To further examine the dataset, we employed a technique called T-distributed Stochastic Neighbourhood Embedding (tSNE). Here is the outcome of the tSNE technique:

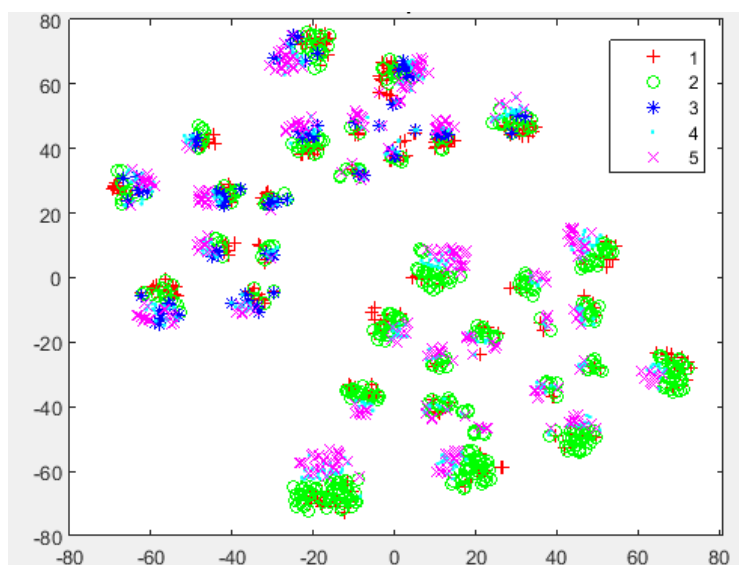


Figure 5-8 Phase 2: T-distributed Stochastic Neighbourhood Embedding

Figure 5-8 shows the datasets with 5 levels of class attributes labels. This plot illustrates the class dispersion problem with different types of colour, where points from the 5 classes of dataset are clustered. Ideally, the 5 classes are decomposed using a clustering technique; each cluster can determine a new class label for the testing set. This shows a real example using t-SNE of the class distribution problem: clusters with the same class points are spread across the variable values. In this case, the machine learning models specifically with RFC are trained on the original dataset with the class labels. The main point behind using this t-SNE is to represent dimensionality reduction that is suitable to visualise our datasets with high dimensional. t-SNE scales depend on the total number of objects  $N$ , it is appropriate to a limited number of datasets with a few thousand instances. We applied this technique with our datasets up to 1,635 instances. [127]

### 5.4.3 Phase 3: Over-Sampling and Classification of Extended Data

As ADNI increase their dataset subjects, further data was added to the baseline dataset and this time more variables were added related to demographic data, behavioural markers, genetic, and some medical history data. The variables selected are matching the features discussed in the risk factors section (2.6) of this thesis.

The selected variables were related to:

Table 5-7 AD Risk Factors used in extended dataset

Medical History	Lifestyle	Genetic	Demography	Family History
Diabetes	Alcohol	APOE4	Age	Dad Dementia
Cholesterol	Smoking		Education Field	Mum Dementia
Allergy	Drug Use		Ethnicity	
Heart Disease	BMI		Race	
Depression	Energy Level		Marriage Status	
Drowsiness, Dizziness, Headaches	Cognitive Test		Work Category	

Different to the initial experiment, this time because the data was extended to cover more variables, the removal of the data would cause more than one class to be imbalanced and some classes had all their subjects removed completely. For this reason, during the implementation of data pre-processing we replaced missing values with mean values of the data, maintaining the number of subjects at 1,737 patients.

During the pre-processing stage, we resolved the problem of having categorical data, which resulted in variable expansion of dataset matrix from 28 features to 70 variables (see section 5.3), Table 5-8 demonstrates the levels of class attributes and the data distribution of instances.

Table 5-8 Levels of the class attribute (Phase 3)

Class	Representation	Frequency	Percentage
NC	Normal Control	417	24 %
AD	Alzheimer’s disease	342	19.68%
EMCI	Early Mild Cognitive Impairment	310	17.84%
LMCI	Late Mild Cognitive Impairment	562	32.35%
SMC	Significant Memory Concern	106	6.10%

The dataset was still imbalanced even if we didn’t remove missing values, because the ADIN2 contained an imbalanced number of subjects for each class. The solution was to use an over sampling technique call Synthetic Minority Over-Sampling Technique (SMOTE), see section (3.4.3.5) for more details on this technique.

**Before over-sampling:**

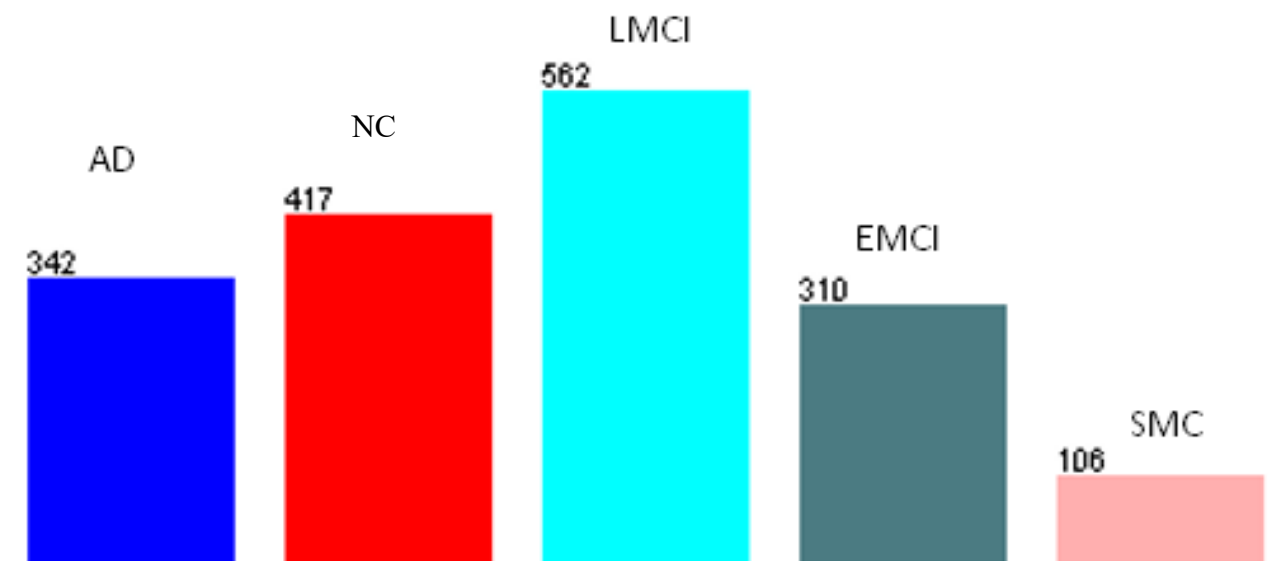


Figure 5-9 Part 3: Before Over-sampling

### After Synthetic Minority Over-Sampling Technique (SMOTE):

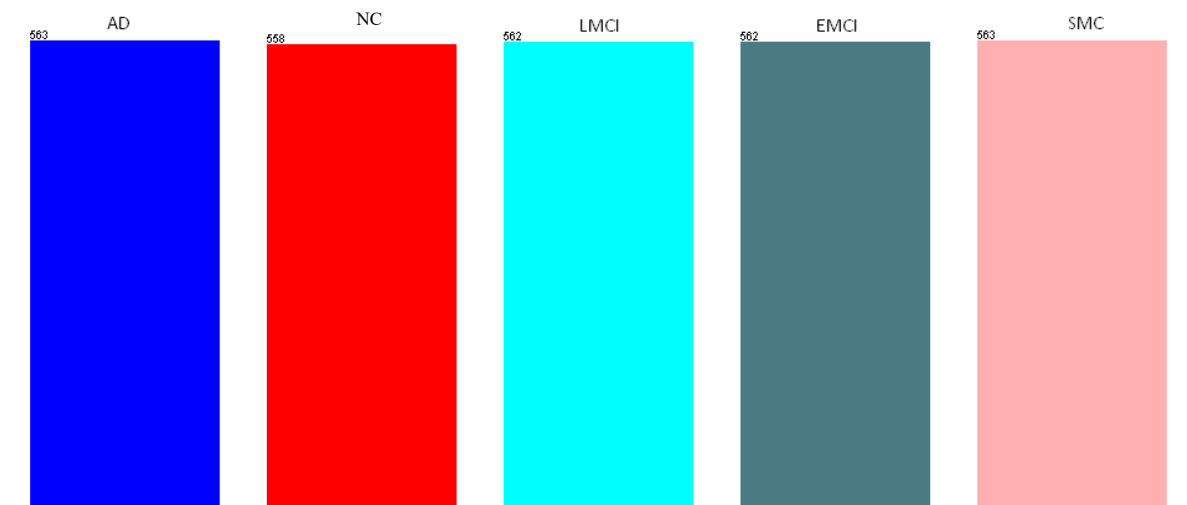


Figure 5-10 Using SMOTE in WEKA 3.6 (Phase 3)

After over-sampling the dataset to balance the classes, the data volume increases to 2,808 subjects, giving approximately 20% of the data to every class label.

For dimensionality reduction, and to visualize the data, t-SNE was applied on the dataset, giving the results shown in Figure 5-11, it shows one large mixed cluster. Ideally the perfect results that we had hoped for is that each cluster will contain a majority of one class. However, as shown in Figure 5-11; the clusters have almost equal mixtures from all classes. Which means that the algorithm struggled to differentiate between the categories of the subjects. The results presented here are nonlinear, if we zoom in further, we find multiple clusters from the same class which means on a dimensional level it managed to differentiate between the variables (risk factors) and could possibly indicate an improvement in the classification. The table below explains the labels on the graphs.

Table 5-9 Classes Representation on Graphs and Plot

Class	AD	NC	LMCI	EMCI	SMC
Number on Graphs	1	2	3	4	5

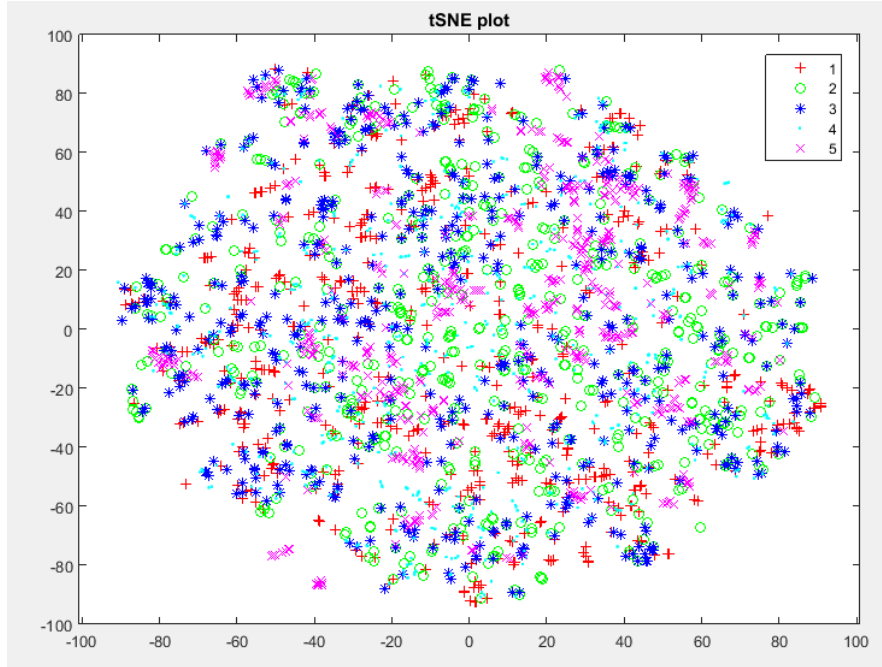


Figure 5-11 Explore of Data Using t-SNE in Matlab (Phase 3)

Other useful methods used to explore and visualize this dataset is the Principal Component Analysis (PCA) and Independent Component Analysis (ICA), both Figure 5-12 and Figure 5-13 illustrate the use of these techniques in Matlab. PCA is used to emphasize the variation, dimensions reduction and bring out the strongest patterns in the dataset. ICA is a method for separating a multivariate signal into additive subcomponents, for more information see section (3.4.4.1 and 3.4.4.2).

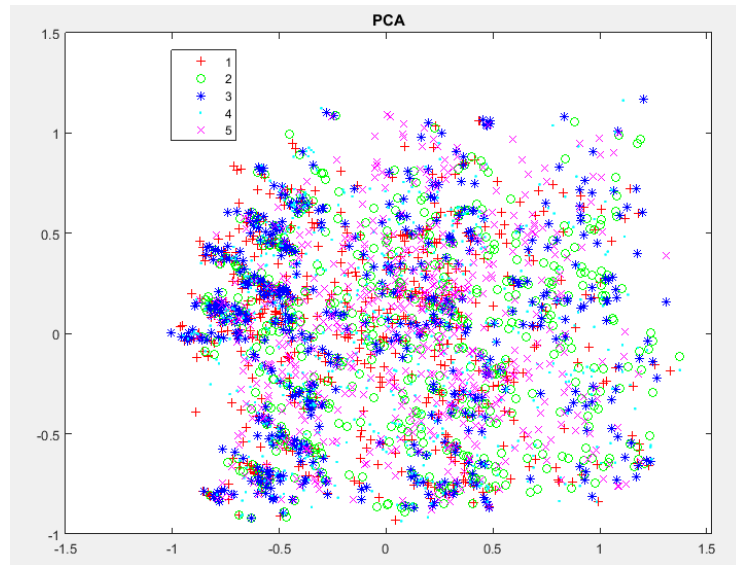


Figure 5-12 Part 3: Exploration of Data Using PCA in MatLab

Furthermore, the data was explored using the Square Prediction Error (SPE) plot to measure the quality of a predictor. The graphs and coefficients results show an apparent indication that the type of study will be a nonlinear regression with a slight indication of classification improvement.

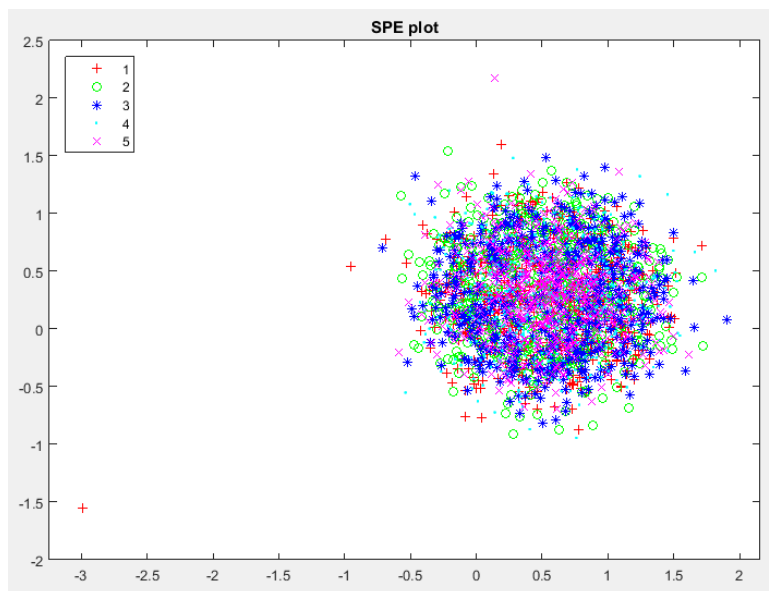


Figure 5-13 Part3: Explore of Data Using SPE in Matlab

## 5.5 Development of Crowdsourcing Risk Factor Ranking System

The current diagnosis of Alzheimer’s disease is conducted through manual evaluation based on clinician’s experience and clinical instinct. Some of this knowledge is not only based on judgment of apparent symptoms of the disease, but it’s also based on their knowledge of a patient’s medical history as well as their knowledge about Alzheimer’s disease and current ongoing research. Since there is no existing certainty in knowing the main causes of the disease yet, most current evolution is based on assumptions. This section of the thesis presents a present a sub-component of the presented framework EPADf, this component is a web-based system called “Crowdsourcing Risk Factor Ranking” (CRFR), used to collect clinical evaluation for Alzheimer’s disease and its risk factors. Using machine learning only to evaluate the disease might not be reliable enough, as traditionally initial weightings in machine learning are distributed randomly. The use of the crowdsourcing system will unify the clinical instinct and evaluations of clinicians, and also produce weighted values for each of the risk factors.

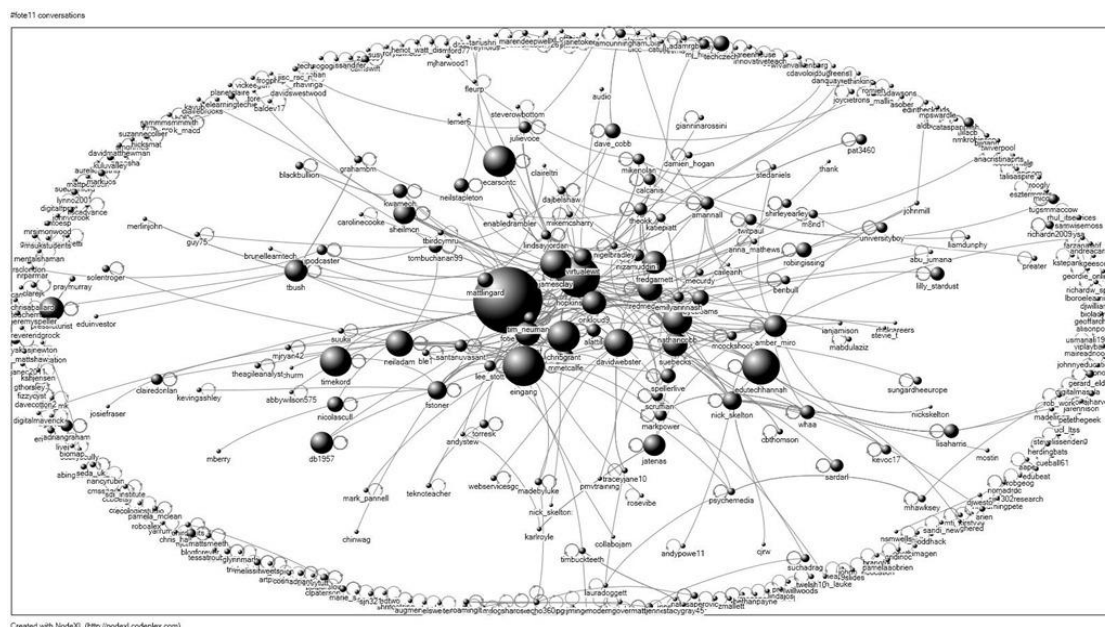


Figure 5-14 Created with NodeXL (<http://nodexl.codeplex.com>)

The image above has no relation to this research, it is visualisation of a social network user's connections and influences. However, taking the same concept and applying it to the interconnection and influence of Alzheimer's disease risk factors on each other, this, would highlight the most influential risk factors and rank them as important contributors. This can be done by developing a crowdsourcing system to map the interconnection between the risk factors. The system would act as a discrete biological knowledge bank to aid with decision making and machine learning tools with early prediction of disease. The extracted knowledge would improve the prediction and diagnosis of Alzheimer's disease using the relationship between risk factors across all of the risk factors categories (Genetics, Lifestyle, Medical History and Characteristics). The aim for this system in relation to the proposed framework is to provide a supervised dataset to boost the accuracy of the machine learning algorithm when working with bioinformatics datasets.

The overall objective is to have a system with all Alzheimer's disease risk factors (clinically inputted) and their relationships added to it (in the concept of a large network). Using a mathematical model to rank and classify each risk factor and giving them a weight. This system will provide a reliable analysis for each risk factor.

### **5.5.1 Crowd Contribution**

Based on current research and input from clinicians the system would have a list of all the factors currently considered as risk factors of Alzheimer's disease, and contributing risk factors to the development of main risk factors. For example, if 'diabetes' is considered an Alzheimer's disease risk factor and 'fast food' is the risk factor for 'diabetes', then in this case 'diabetes' is a main risk factor and 'fast food' is a contributing factor.



Experts (clinicians) would enter their contribution by selecting one of the listed risk factors or one of the contributing factors, then enter their clinical instinct evaluation (input) based on a valid research. If the risk factor or contributing factor is not listed, then they will be able to add it to the list.

To add a risk or a contributing factor they will need to enter its name (unique), select one of the categories this factor falls into e.g. lifestyle, medical history or genetics, and then select the type i.e. if it is a risk factor or a contributor to the risk factors. To add clinical knowledge on a factor the user must select the factor name from the list, select the other factor that this factor contributes to, add the risk level 1-10, add the development stage (1-Mild, 2-Moderate, 3-Severe), then finally add the reference to back up their contribution. The following section 5.5.2 provides a visual user guide for this process.

## 5.5.2 Crowdsourcing System Screenshots

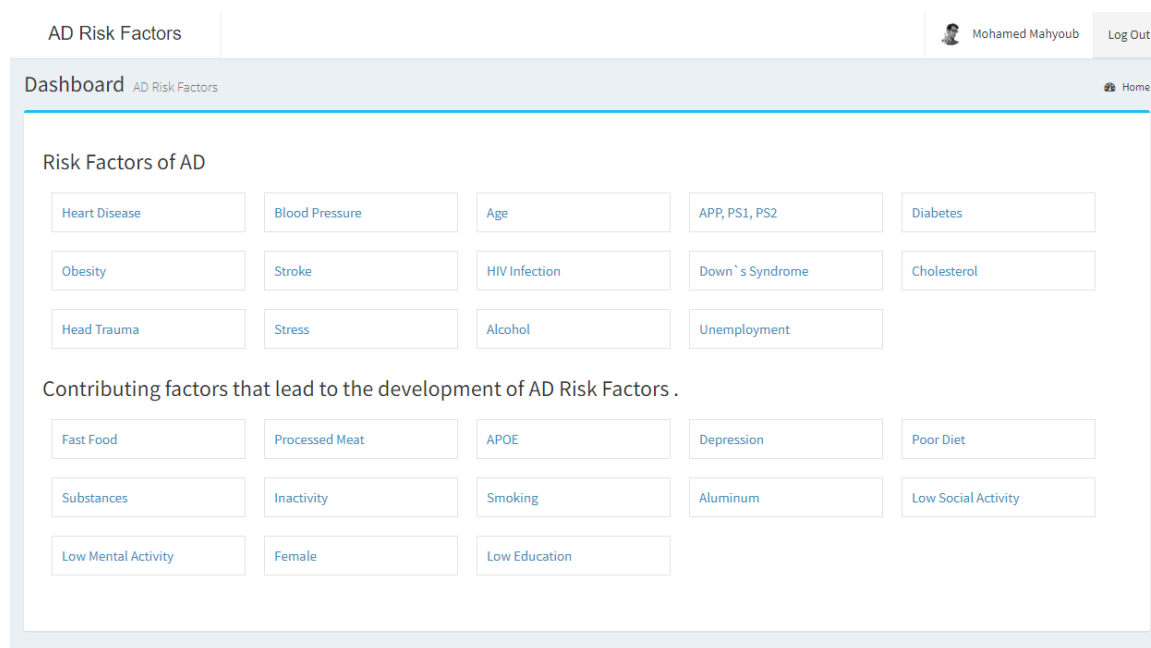


Figure 5-15 Crowdsourcing System Homepage

The users would login to the system, then select one of the risk/contributing factors to contribute to as shown in Figure 5-15. The screenshot below (Figure 5-16) shows the page loaded after the user clicks on one of the factors. For example, “Heart Disease”:

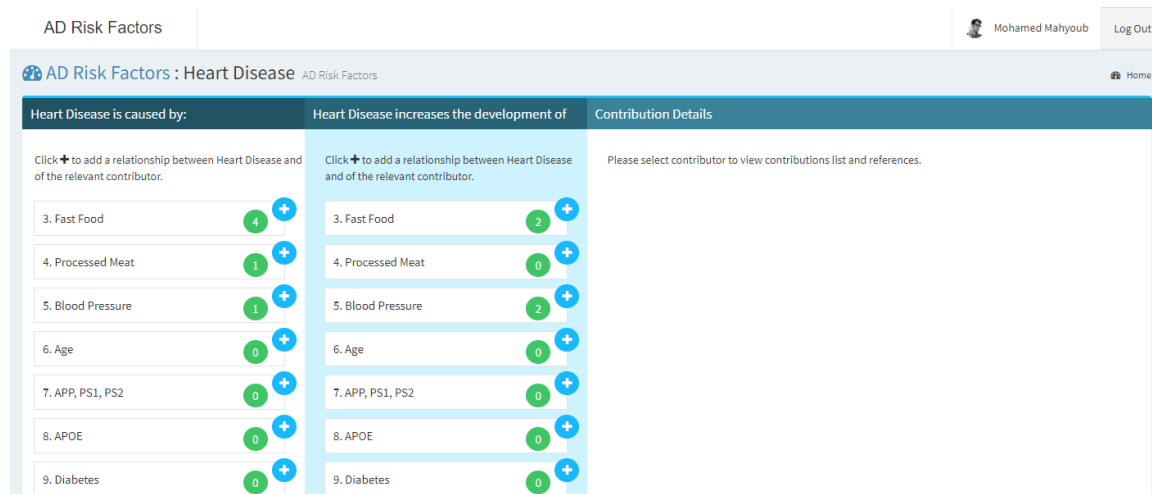



Figure 5-16 Crowdsourcing System Heart Disease Page

The first section on the right lists the contributing risk factors for Heart Disease, the middle section lists the risk factors that heart disease contributes to, and the third section details the contribution between heart disease and any factor selected from the list. To add new clinical input the user would click on  sign and fill in the little form displayed.

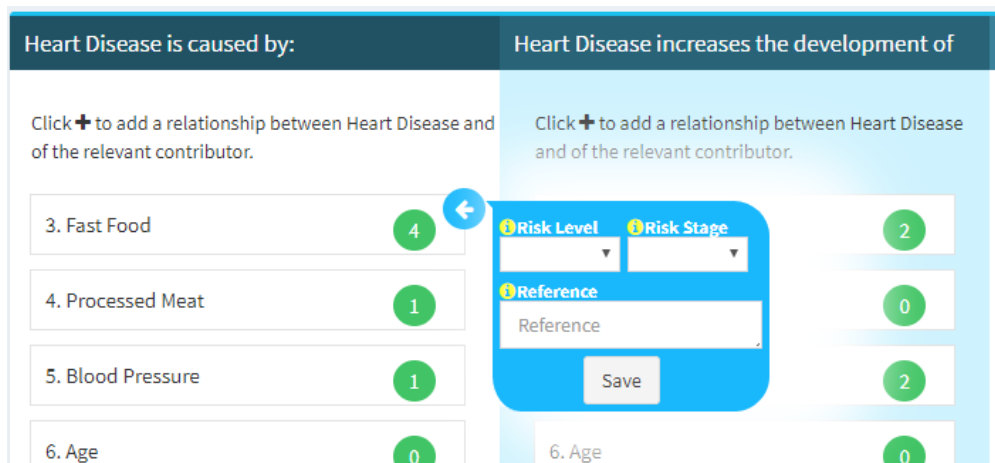


Figure 5-17 Crowded-sourcing System Adding New Contribution

The green circle **4** (as shown in Figure 5-17) represents the number of clinical contributions added. When clicked it will display the contribution details as shown in Figure 5-18 below:

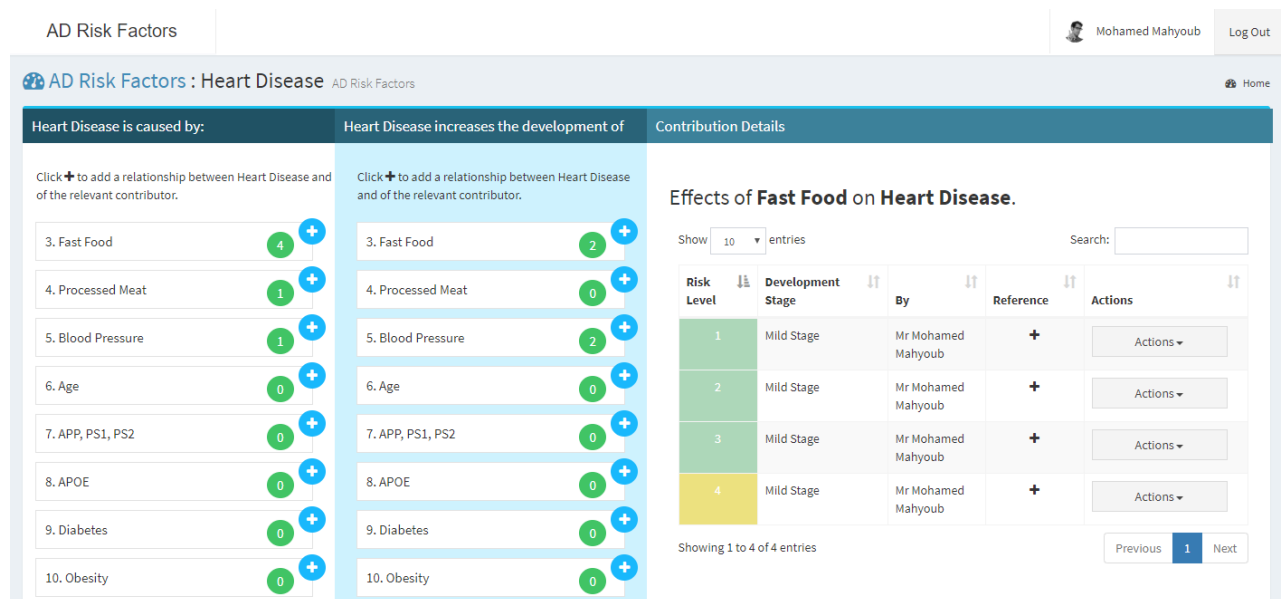


Figure 5-18 Crowded-sourcing Risk Factors Interconnections

### 5.5.3 Converting Input to Weights

Depending on the number of pieces of clinical knowledge contributed over time the more contributions the more comprehensive and useful the extracted weighting would be. Eventually, this concept could possibly present new high-risk factors for the early development of Alzheimer’s disease. The contributing factors could possibly have higher relevance to Alzheimer’s disease more than the currently considered high-risk factors.

Each factor will get a ranking through a mathematical formula that would calculate the overall weight by average number of clinical inputs; considering its contributions levels, outwards and

inwards contributions, and the weighting of factors that contribute to it. This is denoted in the mathematical **Error! Reference source not found.** below:

$$f(x) = \frac{\left( \sum_{i=1}^c \frac{\left( \sum_{j=1}^n \frac{i + R_j + f(j)}{n} \right) + \left( \sum_{j=1}^m \frac{i + R_j + f(j)}{m} \right)}{n + m} \right) - \left( \sum_{i=1}^c \frac{\left( \sum_{j=1}^v \frac{i + R_j + f(j)}{v} \right) + \left( \sum_{j=1}^z \frac{i + R_j + f(j)}{z} \right)}{v + z} \right)}{Y} \quad (1)$$

The function  $f(x)$  returns a numerical ranking value for a risk factor, where  $x$  is the risk factor, we need to calculate its ranking in the network. The first part of the equation is to calculate a value representing the connections that  $x$  contributes to other factors (Outwards), and then calculate a value representing the connections contributed to  $x$  (Inwards). Subtract the inward from the outward then divided it by  $Y$ , which, represents the number of risk factors in the network to get the overall ranking.

$$Outward = \sum_{i=1}^c \frac{\left( \sum_{j=1}^n \frac{i + R_j + f(j)}{n} \right) + \left( \sum_{j=1}^m \frac{i + R_j + f(j)}{m} \right)}{n + m} \quad (2)$$

The calculation for Outward starts by looping through the risk stages  $c$ , then identifying the connections at that level, loop through the connections and get their connections to the main risk factors  $n$ , then for every  $n_j$  get its risk level  $R_j$ , plus risk stage  $c_i$ , plus the rank of the risk factor  $f(j)$ , divided by  $n$ , then repeat the same process to get connections to other contributing factors  $m$ , add the values together and divide it by  $n + m$  to get the inward value. Though it is worth noting that use of  $f(j)$  inside  $f(x)$  could cause a potential loop that will need to be limited.

$$Inward = \sum_{i=1}^c \frac{\left( \sum_{j=1}^v \frac{i + R_j + f(j)}{v} \right) + \left( \sum_{j=1}^z \frac{i + R_j + f(j)}{z} \right)}{v + z} \quad (3)$$

The calculation for Inward connection is conducted in the same process at Outward, except that it measures with risk factors from the opposite direction. The calculation starts by looping

through the risk stages denoted as  $c$ , then identifying the connections at that level, loop through the connections and get their connections to the main risk factors  $v$ , then for every  $n_j$  get its risk level  $R_j$ , plus risk stage  $c_i$ , plus the rank of the risk factor  $f(j)$ , divided by  $n$ , then repeat the same process to get connections to other contributing factors  $z$ , add the values together and divide it by  $v + z$  to get the inward value.

### 5.5.4 Crowdsourcing System Overview

The CRFR component is designed to work as a complete independent system and doesn't necessarily need to depend on the EPADf framework to serve its purpose. However, it has functionality that enables it to communicate and integrated with the framework through a Representational State Transfer (REST) application programming interface (API). This allows the system to send and receive data from external sources. Figure 5-19 is a flowchart that provides an overview of how this component is built.

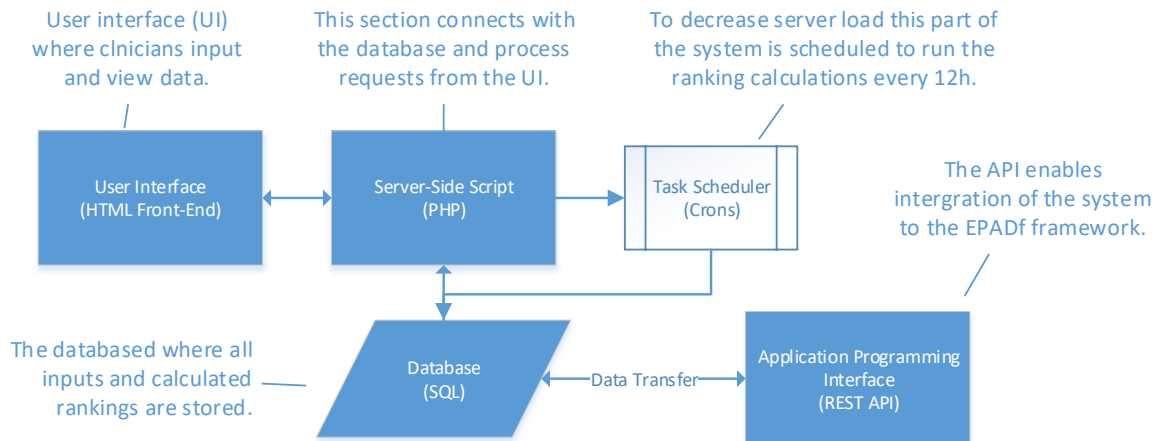


Figure 5-19 CRFR System Overview

## 5.6 Deployment of Machine Learning Towards Early Prediction

This section is a follow up for section (5.3) and section (5.4), after the pre-processing and analysis of the data. Machine learning models were deployed to investigate the possibility for early prediction results. Here we list the different models used during the investigations. The results will be presented in the results in Chapter 6.

### 5.6.1 Phase 1: Initial Experiment Models

During both the training stage and test stage for the initial experiment, the five different classifiers were applied consecutively for approximately 30 simulation runs for a better accuracy. We used five different Machine Learning Classifiers; Random Oracle Model, Random Forest Classifier, Fischer Discriminate Analysis, Multi-Layer Perceptron and Linear Neural Networks. Table 5-10 provides a description for each classifier. For this initial experiment we used 70% of the data to train the classifiers, 10% for validation, and 20% was used to test the performance of the classifiers.

Table 5-10 Initial Experiment Models

<b>Models</b>	<b>Description</b>	<b>Type</b>
<b>Random Forest (rf)</b>	Implements Breiman's random forest algorithm (based on Breiman and Cutler's original Fortran code) for classification and regression.	Classification, Regression
<b>Random Oracle Model</b>	Pseudorandom number generator	Classification
<b>H2 - Levenberg- Marquardt learning neural network and Random Forest, combined using Fischer discriminate analysis</b>	A hybrid model combined of Levenberg Marquardt learning algorithm and Random Forest, combined using Levenberg neural network. It uses	Classification, Regression

	gradient descent with momentum and adaptive learning rate backpropagation.	
<b>Linear Neural Networks</b>	Used for batch training with weight and bias learning rules.	Classification, Regression
<b>Multi-Layer Perceptron (MLP)</b>	A fully connected feedforward artificial neural network.	Classification, Regression

### 5.6.2 Phase 2: Model Used to Rank Risk Factors by Importance

Four candidate models are used to explore the dataset to rank Alzheimer’s disease risk factors by importance and validate the accuracy with 10-fold cross validation and total of 30 simulations. The models compared are Random Forest (RF), Neural Networks with a Principal Component Analysis (pcaNNet), Support Vector Machines with Linear Kernel (svmLinear), and Multi-Layer Perceptron (MLP). Table 5-11 provides a description for each classifiers.

Table 5-11 Model Used to Rank Risk Factors by Importance

<b>Models</b>	<b>Description</b>	<b>Type</b>
<b>Random Forest (rf)</b>	Implements Breiman's random forest algorithm (based on Breiman and Cutler's original Fortran code) for classification and regression.	Classification, Regression
<b>Neural Networks with a Principal Component Analysis (pcaNNet)</b>	Run Principal Component Analysis on a dataset, then use it in a neural network model.	Classification, Regression

<b>Support Vector Machines with Linear Kernel (svmLinear)</b>	Based on the concept of decision planes that define decision boundaries and performs linear regression in the high-dimension feature space.	Classification, Regression
<b>Multi-Layer Perceptron (MLP)</b>	A fully connected feedforward artificial neural network.	Classification, Regression

Variables importance gives an indication of the possible predictive factors in the dataset by calculating their statistical significance. The calculation happens with respect to its impact on the generated model. On the other hand, since the variable importance is based on the contribution that predictors make to the model, this also helps us determine which variables are not necessarily required to be in the dataset. The four machine learning models used to explore the ADNI dataset to rank Alzheimer’s disease risk factors by importance are the Random Forest (RF), Neural Networks with a Principal Component Analysis (pcaNNet), Support Vector Machines with Linear Kernel (svmLinear), and Multi-Layer Perceptron (MLP). We validated the accuracy with 10-fold cross validation and total of 30 simulations. The models would use the dataset to train with and build a predictive model then validate the impact of each variable against the built model and scale the best performing variables.

### **5.6.3 Phase 3: Final Classification Experiment Models**

For the final experiment we used the same classifiers as the initial experiment for both the training stage and test stage, the five different classifiers were applied consecutively for 30 simulation runs. We used five different Machine Learning Classifiers; Random Oracle Model, Random Forest Classifier, Fischer Discriminate Analysis, Multi-Layer Perceptron and Linear Neural Networks. Table 5-12 provides a description for each classifiers. For this final



experiment we used 70% of the data was used for training the classifiers, 10% for validation, and 20% was used to test the performance of the classifiers.

Table 5-12 Final Classification Experiment Models

<b>Models</b>	<b>Description</b>	<b>Type</b>
<b>Random Forest (rf)</b>	Implements Breiman's random forest algorithm (based on Breiman and Cutler's original Fortran code) for classification and regression.	Classification, Regression
<b>Random Oracle Model</b>	Pseudorandom number generator	Classification
<b>H2 - Levenberg-Marquardt learning neural network and Random Forest, combined using Fischer discriminate analysis</b>	A hybrid model combined of Levenberg Marquardt learning algorithm and Random Forest, combined using Levenberg neural network. It uses gradient descent with momentum and adaptive learning rate backpropagation.	Classification, Regression
<b>Linear Neural Networks</b>	Used for batch training with weight and bias learning rules.	Classification, Regression
<b>Multi-Layer Perceptron (MLP)</b>	A fully connected feedforward artificial neural network.	Classification, Regression

## 5.7 Summary

In this chapter we provided and in-depth discusses on the methods that we intend to use to achieve the objectives in this thesis. We presented the proposed framework for early prediction with a comprehensive discussion of its sub-components, which, includes the learning models and deployment phases, the web-based crowdsourcing system, and logical approaches toward early predictions. As well as a descriptive analysis and visualisation of the data we plan to use as a baseline of the propose framework. The next chapter will present the results of the conducted experiments and discussion of these results.

# Chapter 6     **Results**

## **6.1 Introduction**

In the previous chapter, we presented a machine learning models deployment strategy as well as a visual exploration of the data. This chapter is a follow-up for Chapter 5 section (5.3), section (5.4), and section (5.6). After the pre-processing and analysis of the data, machine learning models were deployed to investigate the possibility for early prediction results. Here are the results for all the three stages of the research.

## **6.2 Performance Evaluation Metrics**

We compared the performance of the classifiers by measuring and evaluating their performances by calculating their decision threshold parameter. Our classifier evaluation consists of two stages; training evaluation (in sample) and testing evaluation (out-of-sample). We used a number of methods to measure the accuracy of the classifiers, which include sensitivity, specificity, F1 score, precision, Youden's J statistic, and the overall classification accuracy, then we presented the overall true and false values for each model using Area under the Curve (AUC) and Receiver Operating Characteristic (ROC) plots. These methods used to measure the classifiers results, are dependent on the overall count of these possible outcomes; True Positive (TP) and True Negative (TN) results that represents patients that are correctly classified for positive and negative respectively, False Positives (FP) represents number of patient that have the disease and have been misclassified, and final possible outcome is False Negatives (FN), which represent positive number of patients that don't have the disease and

have been incorrectly classified. Table 6-1 is a contingency table known as confusion matrix showing the four possible outcomes, and

Table 6-2 describes the methods used that utilised confusion matrix outcomes to evaluate the performance of the classifiers.

Table 6-1 Confusion Matrix

		<b>Actual Results</b>	
		Positive (1)	Negative (0)
<b>Predicted Results</b>	Positive (1)	TP	FP
	Negative (0)	FN	TN

Table 6-2 Metrics Calculation

<b>Metric Name</b>	<b>Calculation</b>
Sensitivity	$TP/(TP+FN)$
Specificity	$TN/(TN+FP)$
Precision	$TP/(TP+FP)$
F1 score	$2 * (Precision*Recall)/(Precision+Recall)$
Youden's J statistic (J Score)	Sensitivity + Specificity – 1
Accuracy	$(TP+TN)/(TP+FN+TN+FP)$
Area Under Curve (AUC)	$0 \leq \text{Area under the ROC Curve} \leq 1$
ROC	sensitivity vs (1 – specificity)

### 6.3 Phase 1: Initial Experiment Results

The initial experiment had training results of 0.92 sensitivity, 0.935 specificity and 0.771 precision. However, during the test stage the final output was 0.741 sensitivity, 0.515 specificity and 0.286 precision. The results of this experiment did not give a clear classification or definite predictive value. Involving more variables and underlying data could provide a better outcome. This section was an initial experiment to examine the performance of the classifiers. The aim of all three stages of the research is the classification and ranking of the importance of Alzheimer’s disease risk factors using Machine Learning predictive models and classifications techniques.

#### 6.3.1 Initial Experiment Training Results

During both the training stage and test stage, the five different classifiers were applied consecutively for 30 simulation runs (1,000 iterations per simulation) for a better accuracy. The contrast between the outcome of the training experiment and the test experiment is obvious. As expected, the classifiers have performed better during the training stage because the class labels

were provided to the training model. The hybrid classifier Levenberg- Marquardt learning neural network and Random Forest, combined using Fischer discriminate analysis (referenced as H2) performed the best while other classifiers did not have dramatic differences during the test stage. Figures and Tables in section (6.3) show the outcome of both training and test experiments. From a constructive perspective the initial investigation provides a needed foundation to draw a roadmap for further work and it has become apparent that more variables related to Alzheimer's disease risk factors are required to improve the accuracy of the classifiers.

Figure 6-1 below is a simplified chart displaying the AUC training results for all of the classifiers used showing each class in the dataset in a different colour. The y-axis of the graph displays the accuracy of the classifier between 0 and 1. Where 1 means that the performance of classifier was successfully, and 0 means the classifier failed to perform of the data. The x-axis is displaying the classes in the data for each of the classifying model.

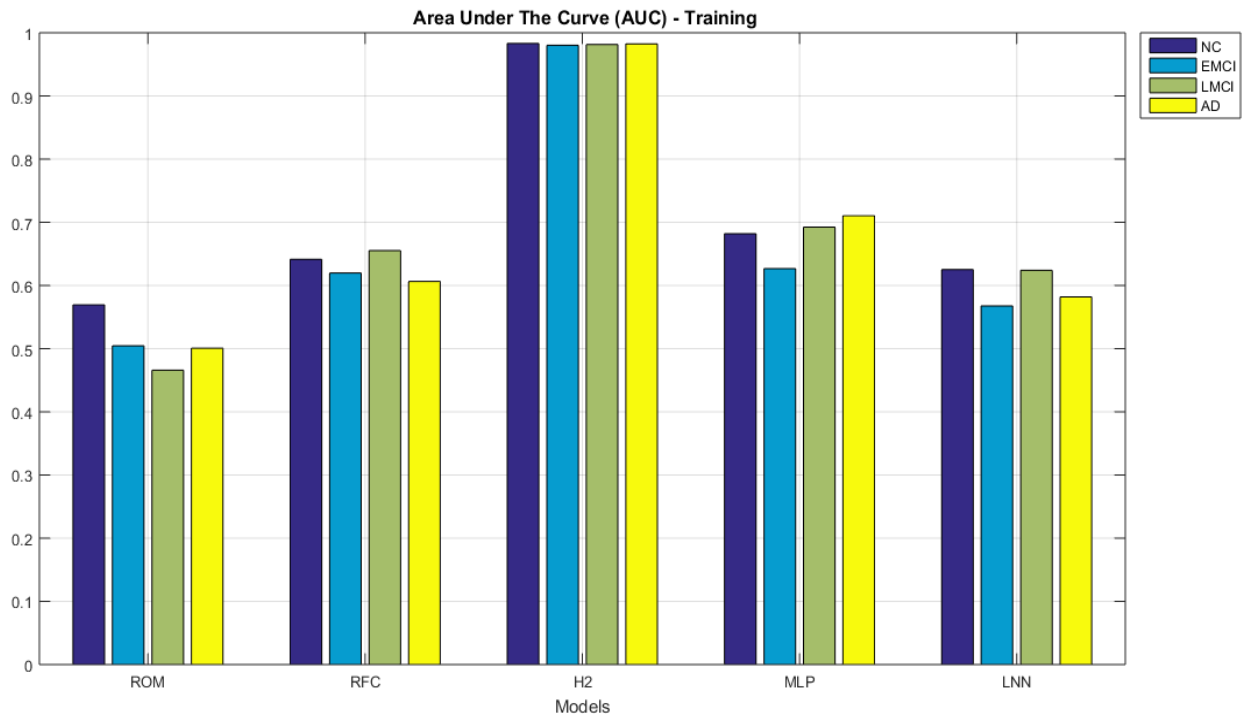


Figure 6-1 Training results for 5 different classifiers on Matlab

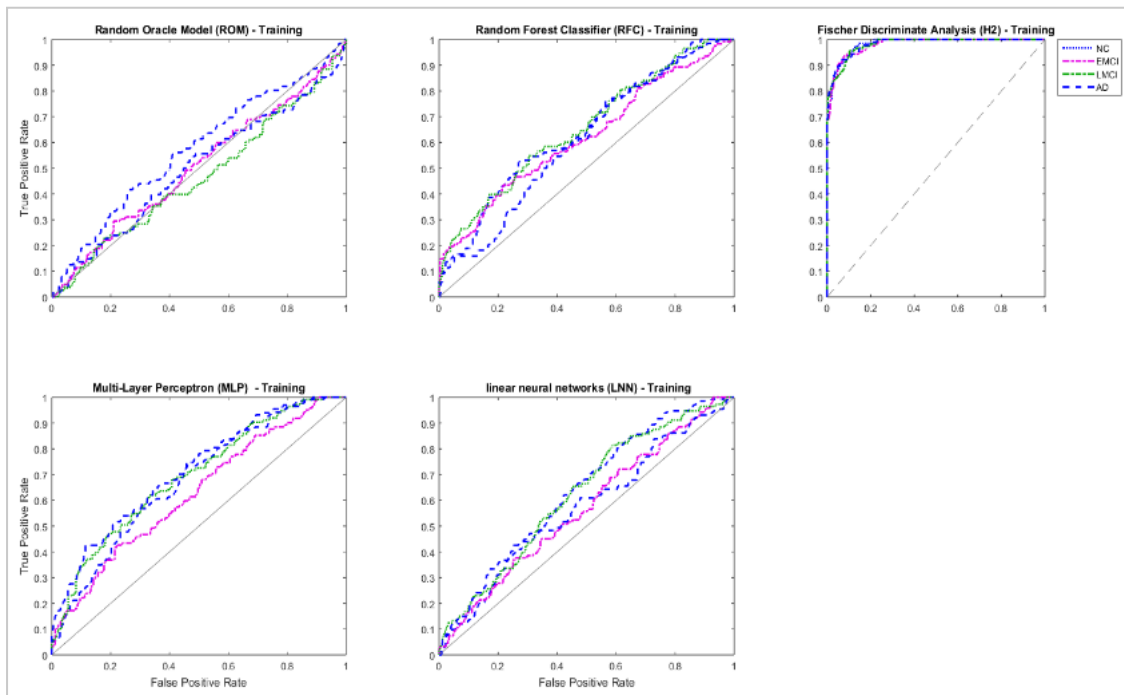


Figure 6-2 ROC training results for each classifier on Matlab

Figure 6-2 is a group of ROC plots displaying a ROC curve for each model, with True Positive (TP) rate against the False Positive (FP) rate, where TP is on y-axis and FP is on the x-axis. The closer the curve to 1 on the y-axis the better the performance of the classifier.

Table 6-3 Training Overall Results (Phase 1)

Model	Class	Sensitivity	Specificity	Precision	F1 score	J	Accuracy	AUC
ROM	NC	0.553	0.591	0.356	0.433	0.144	0.58	0.569
	EMCI	0.492	0.547	0.284	0.36	0.0383	0.532	0.505
	LMCI	0.389	0.623	0.254	0.308	0.0122	0.565	0.466
	AD	0.545	0.526	0.216	0.31	0.0713	0.53	0.501
RFC	NC	0.53	0.728	0.443	0.483	0.258	0.67	0.642
	EMCI	0.557	0.613	0.345	0.426	0.17	0.598	0.62
	LMCI	0.549	0.696	0.373	0.444	0.245	0.659	0.655
	AD	0.58	0.58	0.249	0.348	0.16	0.58	0.606
H2	NC	0.909	0.938	0.857	0.882	0.847	0.93	0.983
	EMCI	0.934	0.925	0.82	0.874	0.859	0.927	0.98
	LMCI	0.956	0.898	0.755	0.844	0.853	0.912	0.981
	AD	0.92	0.935	0.771	0.839	0.855	0.932	0.982
MLP	NC	0.568	0.699	0.436	0.493	0.267	0.661	0.682
	EMCI	0.574	0.581	0.335	0.423	0.155	0.579	0.627
	LMCI	0.619	0.669	0.383	0.473	0.288	0.656	0.692
	AD	0.644	0.657	0.308	0.416	0.3	0.654	0.71
LNN	NC	0.636	0.565	0.375	0.472	0.202	0.586	0.625
	EMCI	0.508	0.587	0.312	0.386	0.0955	0.566	0.568
	LMCI	0.655	0.545	0.323	0.433	0.2	0.573	0.624
	AD	0.471	0.695	0.268	0.342	0.166	0.652	0.582

Table 6-3 above is displaying the metrics calculation for evaluation methods used to evaluate the performance of the classifiers on each class of the dataset.

### 6.3.2 Initial Experiment Test Results

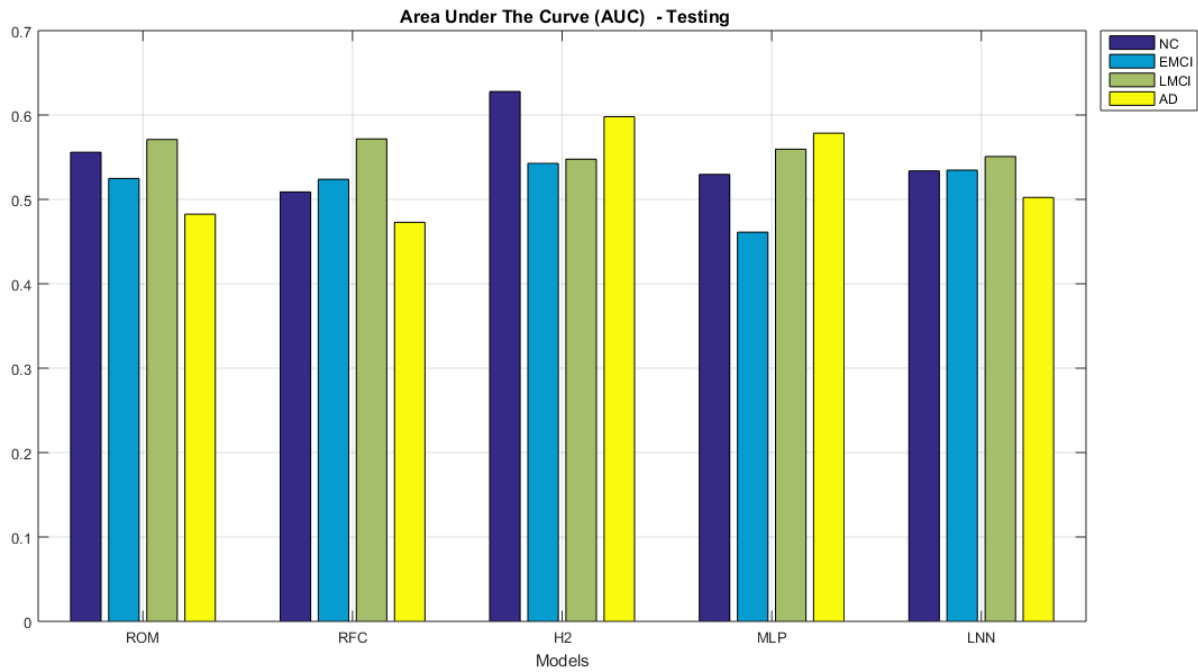


Figure 6-3 Test results for 5 different classifiers on MATLAB

Similar to Figure 6-1, Figure 6-3 above is a simplified chart displaying the AUC test results for all of the classifiers used. The y-axis of the graph displays the accuracy of the classifier between 0 and 1. We notice that when testing the models with unlabelled data after they have complete their training, the models performance significantly reduces, especially the H2 classifier.



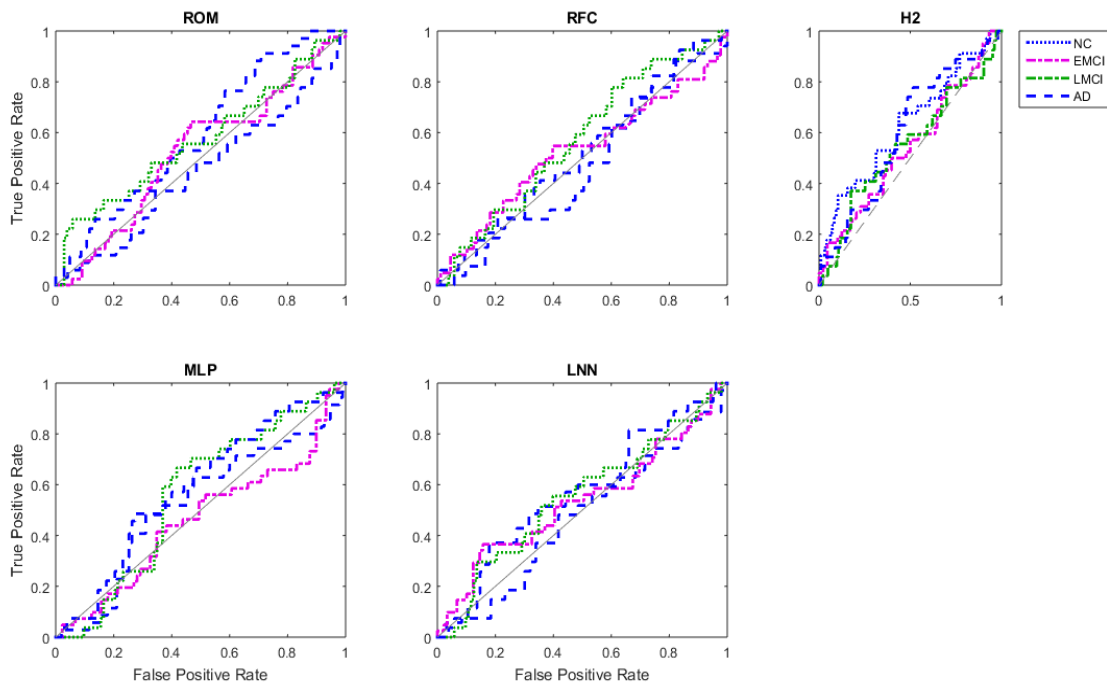


Figure 6-4 ROC Test results for each classifier on Matlab

Figure 6-4 is a group of ROC plots displaying a ROC curve for each model, with True Positive (TP) rate against the False Positive (FP) rate, where TP is on y-axis and FP is on the x-axis. As shown in this figure almost all of the classifiers had almost a straight line instead of a curve which, means they struggle to perform.

Before the experiment there were no prior expectations, the experiment results of the classifiers might not give a very clear classification or definite predictive value, but we expect to see which classifier performed best on the data and which risk factor is more likely to be a predictive factor to the rest of the dataset variables. Unfortunately, we could not tell which variables were potential high-risk factors, (predictive feature) from this experiment. However, we have noticed that the H2 hybrid classifier has performed better compared to other classifiers during both training and testing phases. The hybrid Fischer Discriminate Analysis classifier

(H2) gave training results of 0.92 sensitivity, 0.935 specificity and 0.771 precision. During the test stage the final output of this classifier was 0.741 sensitivity, 0.515 specificity and 0.286 precision. The results of this experiment did not give a clear classification or definite predictive value. Which means involving more variables and underlying data could provide a better outcome. This was conducted in part three of the thesis.

## **6.4 Phase 2: Rank Risk Factors by Importance Results**

The four machine learning models used to explore the ADNI dataset to rank Alzheimer's disease risk factors by importance are the Random Forest (RF), Neural Networks with a Principal Component Analysis (pcaNNet), Support Vector Machines with Linear Kernel (svmLinear), and Multi-Layer Perceptron (MLP). We validated the accuracy with 10-fold cross validation and a total of 30 simulations. The models would use the dataset to train with and build a predictive model then validate the impact of each variable against the built model and scale the best performing variables. This section illustrates the output of the four models used to calculate the importance of some Alzheimer's disease behavioural and biological risk factors in the ADNI dataset.

### **6.4.1 Ranking with Random Forest Model**

The first model employed was the Random Forest (RF), which, is an ensemble learning method often used for both classification and regression problems. The outcome of this model shows that the overall top three most important variables are APOE4 gene, Age, and lack of energy. Followed by significant importance for father dementia history and vision problems. Other risk factors also have some sort of importance and the lowest ones were smoking, ethnicity,

cardiovascular problems, and alcohol consumption. We believe the reason for some of the least importance factors in this result i.e. smoking and alcohol could either be true or because the duration of consumption was not involved. For each individual class the variable importance was different.

The most interesting part of this outcome is the results for subjects with Alzheimer's disease, normal control, and subjects with significant memory concern as they have some sort of logical correlations with current research and support our study to investigate and predict Alzheimer's disease at a very early stage. Alzheimer's disease subjects had APOE4, drowsiness, education level, work category, and weight as highest importance factors, and ethnicity, height, and mother's dementia history as the least important factors. Subjects with significant memory concern had low energy, vision problem, feeling dizzy and drowsy, and headache as highest importance factors, and weight, drugs, cardiovascular as the least important factors. Normal control subjects had APOE4, age, low energy, drowsiness and height as highest importance factors, and weight, smoking, depression as the least important factors.

In the significant memory concern group of subjects, the high importance factors are all head and brain related problems which to some logical extent shows some positive results. However, having significant memory concern does not necessarily mean the onset of Alzheimer's disease, and therefore, the important variables for these subjects do not necessarily need to correlate significantly with Alzheimer's disease subjects. Additionally, this class had a very low distribution and there is a probability that the results would change if there were a much bigger number of subjects. Table 6-4 shows the top five most important risk factors for each class of the dataset using the Random Forest model, alongside this Figure 6-5 is showing the

level of importance for each variable in each class of the dataset. The discussion in this section is based on the results produces by this model.

Table 6-4 Variable importance using Random Forest model

AD	NC	EMCI	SMC	LMCI	Overall
APOE4	APOE4	AGE	ENERGY	ENERGY	APOE4
DROWSY	AGE	DAD	VISION	VISION	AGE
EDUCATION	ENERGY	WEIGHT	DIZZY	HEADACHE	ENERGY
WORK	DROWSY	ENERGY	DROWSY	DIZZY	DAD_DEMENTIA
WEIGHT	HEIGHT	HEADACHE	HEADACHE	DEPRESSION	VISION

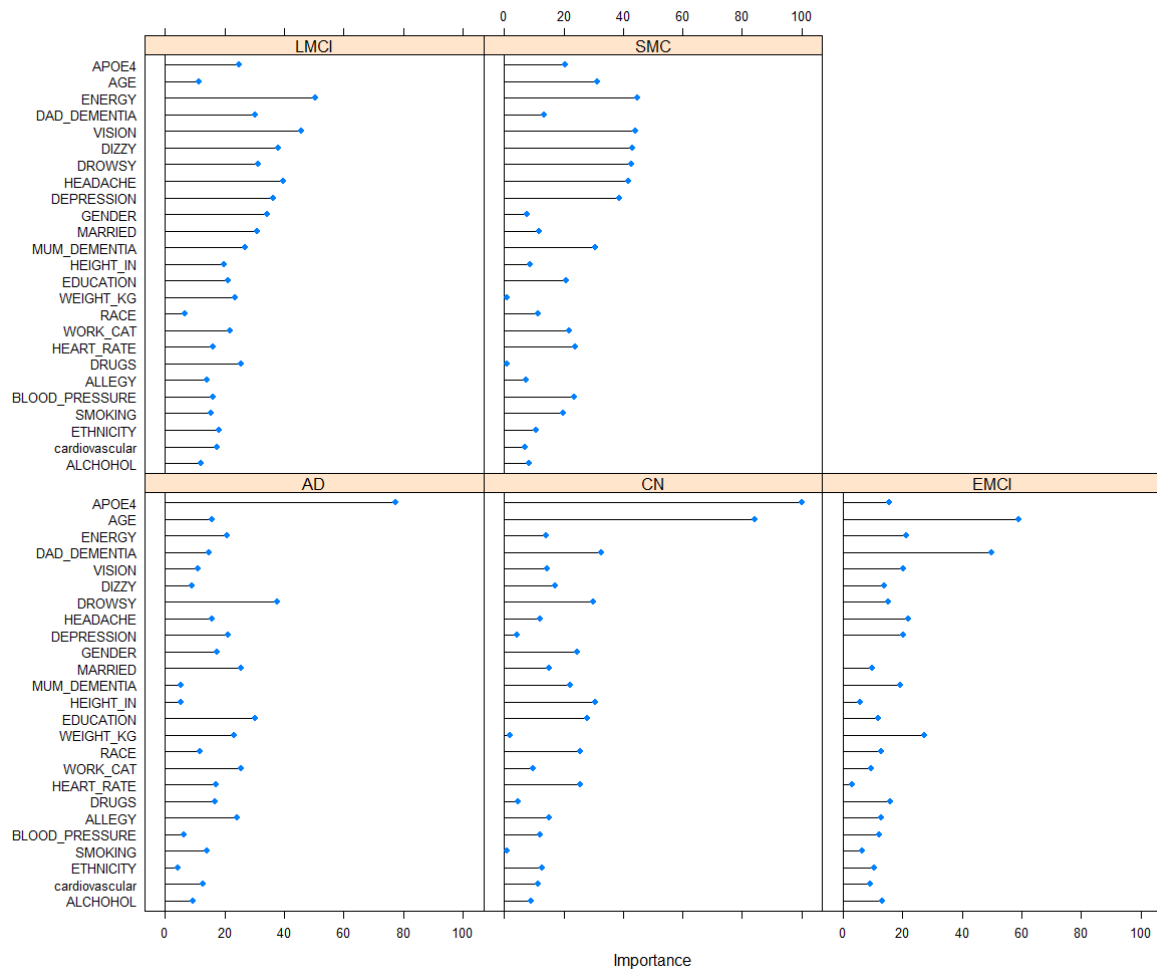


Figure 6-5 Overall results of the variable importance using the Random Forest model

## 6.4.2 Ranking with Neural Network & PCA

The second model was the Neural Network model with feature extraction method using principal component analysis (pcaNnet), which, is a neural network model that runs principal component analysis on the data to compute the cumulative percentage of variance for each principal component. It is a common method often used for both classification and regression problems [128].

The outcome of this model shows that the overall top three most important variables are dizziness, depression, and drowsiness as can be seen in Table 6-5. Followed by respectful importance for vision problems and low energy. Other risk factors also have some sort of importance and the lowest ones were alcohol consumption, race, ethnicity, and drug use. Again, we believe the reason for some of the least importance factors in this result i.e. smoking, drugs and alcohol could either be insignificant or because the duration of consumption was not involved. Moreover, in the results for this model each individual class had different variable importance. The interesting part in the performance of this model is the results for subjects with Alzheimer's disease, subjects with significant memory concern, and early mild cognitive impairment subjects as they have some sort of correlations.

Alzheimer's disease subjects had APOE4, age, weight, father's dementia history, education and depression as highest importance factors, and ethnicity, smoking, drugs and alcohol as the least important factors. Subjects with significant memory concern had age, APOE4, weight, depression and father's dementia history as highest importance factors, and ethnicity, smoking, and drugs as the least important factors. Early mild cognitive impairment subjects had

dizziness, depression, drowsiness, vision and energy as highest importance factors, and ethnicity, alcohol, and drugs as the least important factors. In the early mild cognitive impairment group of subjects the high importance factors are all head and brain related problems which to some logical extent shows some positive results as they are also symptoms of dementia [3][3][5].

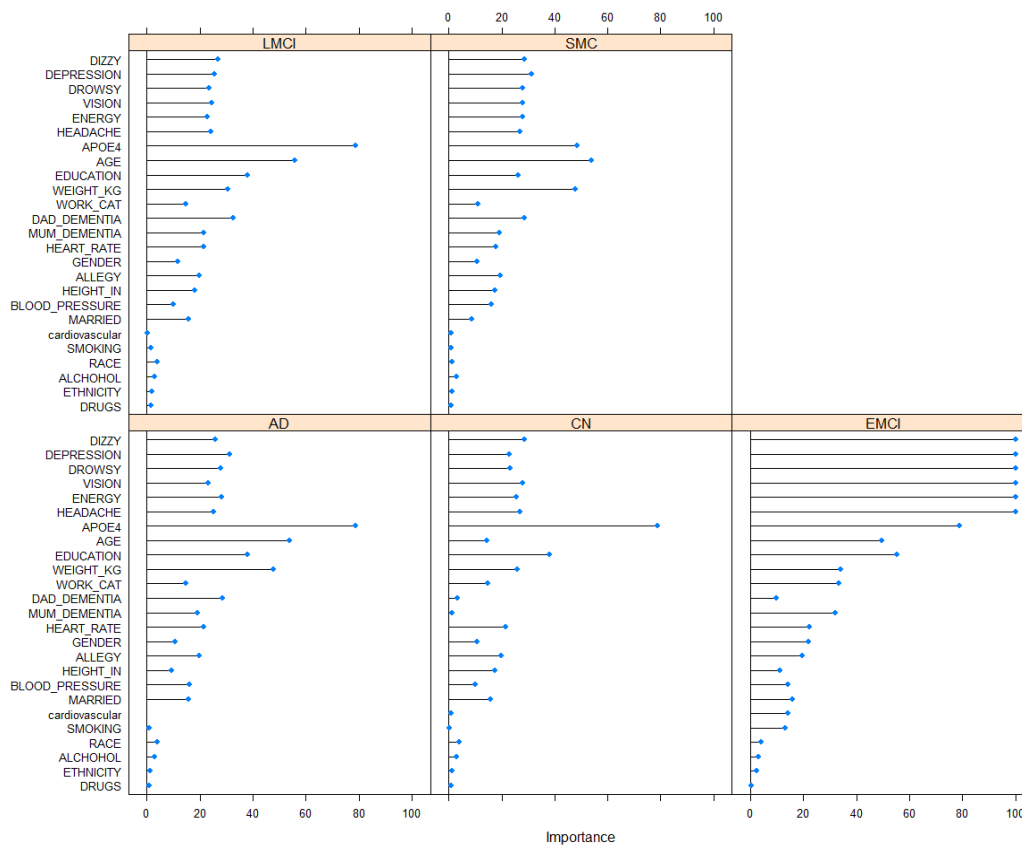


Figure 6-6 Overall results of the variable importance using the pcaNnet model

The discussion in this section is based on the results produced by Neural Network with PCA model. Figure 6-6 is showing the level of importance for each variable in each class of the dataset and the top 5 results in this figure have been summarised in Table 6-5 below.

Table 6-5 Variable importance using pcaNNet

AD	NC	EMCI	SMC	LMCI	Overall
APOE4	APOE4	DIZZY	AGE	APOE4	DIZZY
AGE	EDUCATION	DEPRESSION	APOE4	AGE	DEPRESSION
WEIGHT	VISION	DROWSY	WEIGHT	EDUCATION	DROWSY
DAD_ DEMENTIA	DIZZY	VISION	DEPRESSION	DAD_ DEMENTIA	VISION
EDUCATION & DEPRESSION	WEIGHT	ENERGY & HEADACHE	DAD_ DEMENTIA	WEIGHT	ENERGY

### 6.4.3 Ranking with Support Vector Machines

The third model used was the Support Vector Machines with Linear Kernel (svmLinear), which is a supervised machine learning model that constructs hyperplanes high dimensional data and can be used for both classification and regression analysis. It is a common algorithm often used to solve machine learning problems and has been widely applied in the biological and other sciences [141].

The outcome of this model shows that the overall top three most important variables are drowsiness, low energy, and vision as seen in Table 6-6. Followed by some high importance for dizziness and headache. Other risk factors also have some sort of importance and the lowest the ones were drug use, smoking, ethnicity, and alcohol consumption. Again, we believe the reason for some of the least importance factors in this result i.e. smoking, drugs and alcohol could either be insignificant or because the duration of consumption was not involved. Like the results of other models each class in this model had comparable variable importance except subjects with early mild cognitive impairment as their variable importance significantly differs from the rest of the classes.

Alzheimer’s disease subjects had APOE4, age, weight, education and depression as highest importance factors, and ethnicity, smoking, drugs and alcohol as the least important factors.

Which is almost same as the results produced by pcaNent model.

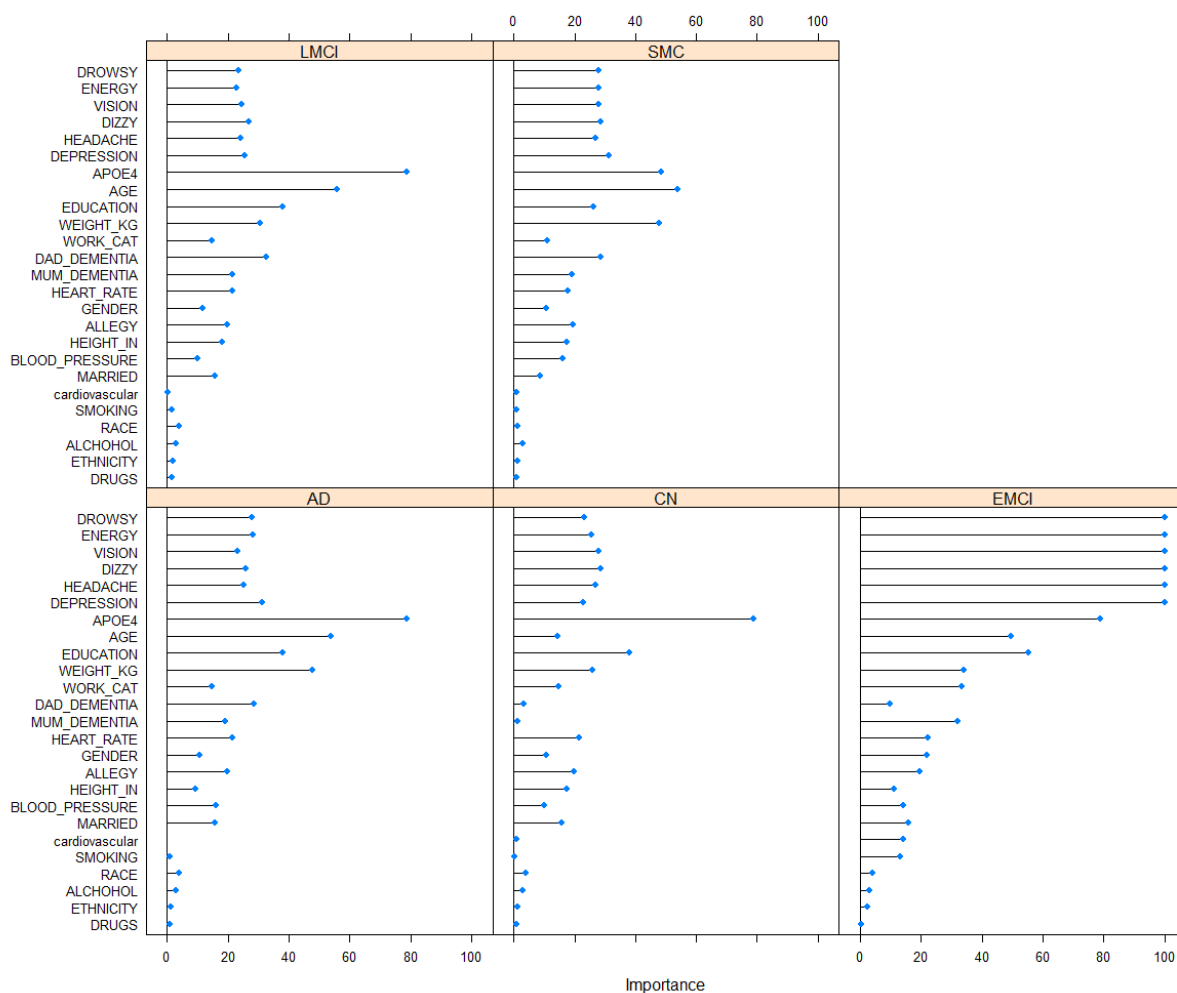


Figure 6-7 Overall results of the variable importance using the svmLinear model

The discussion in this section is based on the results produces by svmLinear model. Figure 6-7 is showing the level of importance for each variable in each class of the dataset and the top 5 results in this figure have been summarised in Table 6-6 below.



Table 6-6 Variable importance using svmLinear

AD	NC	EMCI	SMC	LMCI	Overall
APOE4	APOE4	DROWSY	AGE	APOE4	DROWSY
AGE	EDUCATION	ENERGY	APOE4	AGE	ENERGY
WEIGHT	VISION	VISION	WEIGHT	EDUCATION	VISION
EDUCATION	DIZZY	DIZZY	DEPRESSION	DAD_DEMENTIA	DIZZY
DEPRESSION	WEIGHT	HEADACHE & DEPRESSION	DAD_DEMENTIA	WEIGHT	HEADACHE

#### 6.4.4 Ranking with Multi-Layer Perceptron

The last model examined was the Multi-Layer Perceptron (MLP), which is a supervised learning feedforward artificial neural network model that uses a minimum of 3 layers of nodes. It uses a technique called backpropagation to perform its training and it is often used for both classification and regression analysis [86].

The results produced by this model show that the top three most important variables are headache, low vision, and depression as shown in Table 6-7. Followed by some importance for energy and dizziness. Other risk factors also have some sort of importance and the lowest ones were drug use, smoking, ethnicity, and alcohol consumption. The least important variables are almost the same as the least important variables recognised by other models. Although, they had some sort of importance in the results produced by other models here we see that they are almost of zero importance. Perhaps in a future experiment we will try adding the duration of the consumption and see if they would have higher importance. Like the results of other models each class in this model had similar variable importance except subjects with early mild

cognitive impairment as their variable importance significantly differs from the rest of the classes. However, what is apparent here is that almost across all classes, factors that are related to symptoms of Alzheimer’s disease are recognised to have some importance.

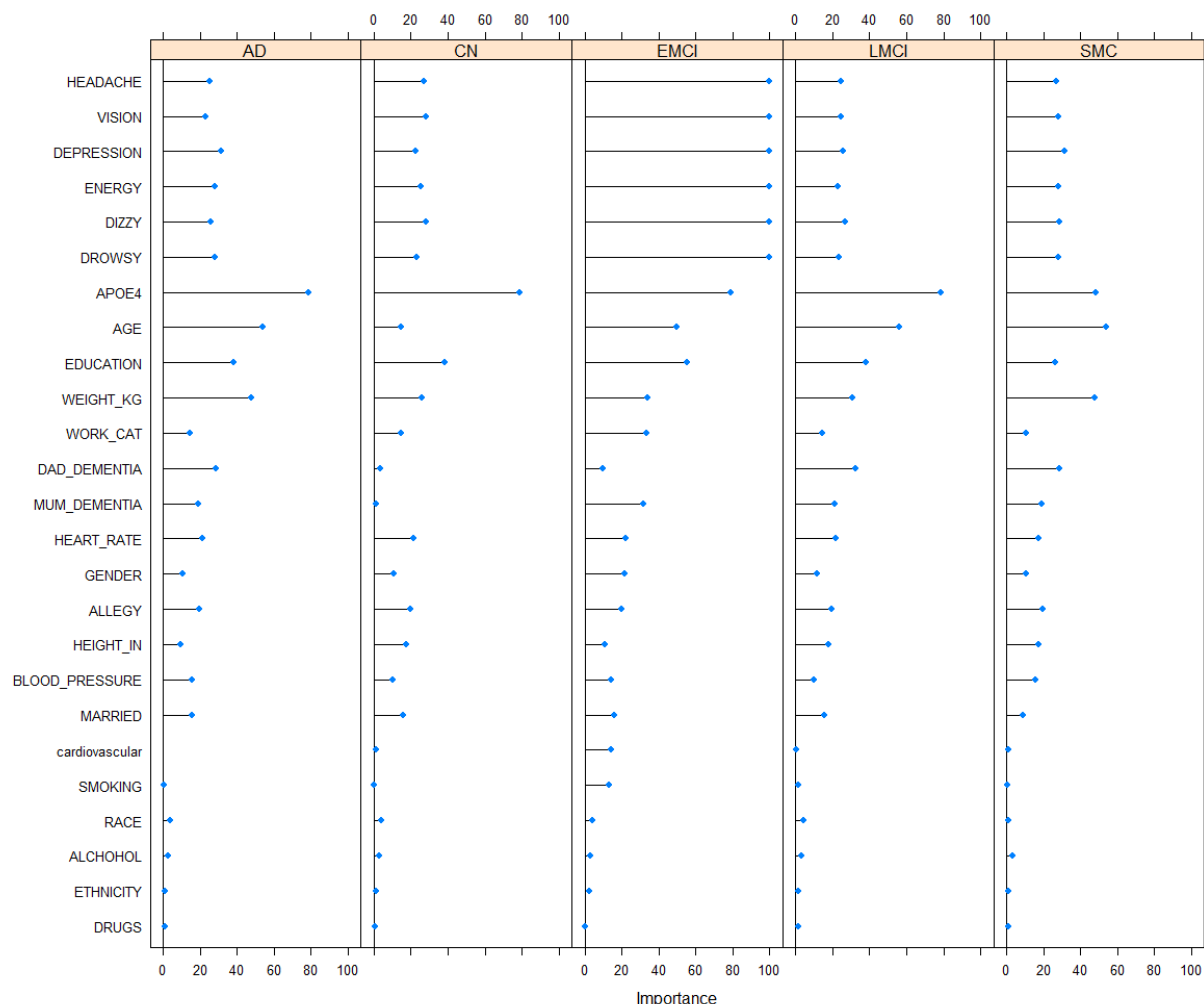


Figure 6-8 Overall results of the variable importance using the MLP model

The discussion in this section is based on the results produced by MLP model. Figure 6-8 is showing the level of importance for each variable in each class of the dataset and the top 5 results in this figure have been summarised in Table 6-6 below.

AD	NC	EMCI	SMC	LMCI	Overall
APOE4	APOE4	DROWSY	AGE	APOE4	HEADACHE
AGE	EDUCATION	ENERGY	APOE4	AGE	VISION
WEIGHT	VISION	VISION	WEIGHT	EDUCATION	DEPRESSION
EDUCATION	DIZZY	DIZZY	DEPRESSION	DAD_DEMENTIA	ENERGY
DEPRESSION	WEIGHT	HEADACHE & DEPRESSION	DAD_DEMENTIA	WEIGHT	DIZZY

Table 6-7 Variable importance using MLP

### 6.4.5 Combined discussion of Risk Factor's ranking

The experiment employed four different machine learning models on the dataset extract to rank variables based on importance. These models include the Random Forest (RF), Neural Networks with a Principal Component Analysis (pcaNNet), Support Vector Machines with Linear Kernel (svmLinear), and Multi-Layer Perceptron (MLP). The overall output of the models combined suggests that all variables had some sort of importance but variables that had highest importance are energy, vision, dizziness, depression and headaches (top 5 from the combined overall results). This is very interesting as current research shows that social inactivity and sport inactivity are related to Alzheimer's disease and could possibly be contributing risk factors [142][59]. Lack of energy could indicate that the subject had a low sporty lifestyle, and correspondingly depression could also indicate a lazy and anti-social lifestyle.

The stated results above were mainly influenced by the results for early mild cognitive impairment subjects as the importance level was very high for that class. However, there are other high importance variables beside energy and vision, APOE4, age, and father's dementia

history have scored high importance also. This also indicates that genetics also play a vital role in the development of Alzheimer’s disease and as such, current research recognises the important role of APOE4 in relation to Alzheimer’s disease [143].

### 6.5 Phase 3: Final Classification Results

During both the training stage and test stage, the five different classifiers were applied consecutively for 30 simulation runs (1,000 iterations per simulation). The classifiers have performed better during the training stage because the class labels were provided to the training model. Figures and Tables in section (5.3) show the outcome of both training and tests during this experiment.

Results from the data analysis and visualisation process in section (5.4.3) show that some slight apparent structure is present within the data. Though both PCA and t-SNE plots show that the two techniques struggled to separate the classes to clear clusters, this is an indication that the data will be challenging for the classifiers to create a clear separation between the classes. This section presents the results showing how the classifiers performed compared to each other.

Table 6-8 Graphs Keys Representation

Class	AD	NC	LMCI	EMCI	SMC
Keys on Graph	1	2	3	4	5

Table 6-8 above shows the representation of the classes’ keys on the graphs presented in this section.

Here we analyse the results from the experiment as listed in Table 6-9 and Table 6-10, showing overall performance results for classifiers during the training and testing stages, separately. To

demonstrate the performance results further, further visualisations are presented using ROC plots as shown in Figure 6-10 and Figure 6-12, and AUC comparison plots as shown in Figure 6-9 and Figure 6-11.

Table 6-9 : Mean Performance for Models (Test) (Phase 3)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
<b>ROM</b>	0.4646	0.5704	0.2138	0.2918	0.03508	0.5494	0.492
<b>RFC</b>	0.671	0.6224	0.3224	0.4324	0.2932	0.6318	0.6808
<b>H2</b>	0.8502	0.806	0.5458	0.6592	0.6562	0.8148	0.8812
<b>MLP</b>	0.8282	0.7742	0.5026	0.6186	0.6026	0.7846	0.8642
<b>LNN</b>	0.706	0.7238	0.4194	0.5208	0.42996	0.7206	0.7628

Table 6-10 : Mean Performance for Models (Training) (Phase 3)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
<b>ROM</b>	0.5064	0.5168	0.2078	0.2948	0.0232	0.5146	0.509
<b>RFC</b>	0.6536	0.66	0.3312	0.4386	0.313	0.6588	0.6952
<b>H2</b>	0.9484	0.9528	0.8358	0.8882	0.9008	0.952	0.9876
<b>MLP</b>	0.8748	0.8776	0.6534	0.7458	0.7526	0.877	0.9398
<b>LNN</b>	0.7522	0.7734	0.4696	0.575	0.5258	0.769	0.821

It is clear from the performance results that the dataset contains significant non-linear relationships, which made the learning process very difficult for the test models. The best performing classifier was the hybrid model combination of Levenberg- Marquardt learning neural network and Random Forest, combined using Fischer discriminate analysis (H2), as

shown in Table 6-12, followed by the Multi-layer Perceptron Model (MLP), both demonstrated their capabilities in fitting the training data, then generalising to unseen examples during testing.

The mean AUC for the hybrid model for all the classes yielded an area of 0.987 during training, in comparison to 0.881 during the test sample. Though the classifier struggled with three classes (2-3) during the test and performed best for all the classes during the training. The Multi-layer Perceptron model (MLP) also displayed similar results but struggled with the same three classes as the hybrid model during both training and testing. The MLP had a mean AUC for all the classes yielding at 0.939 during training, in comparison to 0.864 during the test sample. The third best classifier that also had improved results compared to the initial experiment is the Linear Neural Networks (LNN) model. Though it had a good AUC for class 1 during the both training and testing, overall the mean AUC for all classes yielded at 0.821 during training and 0.762 for testing.

This experiment found that the classifiers H2, MLP, and LNN have improved in classification after the enhancement of the dataset, while the Random Oracle Model (ROM) and Random Forest Classifier (RFC) classifiers made small and insignificant improvement giving a mean AUC yielding at 0.509 and 0.695 for training and 0.492 and 0.680 for testing, respectively. The RCF is founded on decision tree primitives, and possibility of overfitting problems have caused it to struggle with the classification.

## 6.5.1 Final Classification Training Results

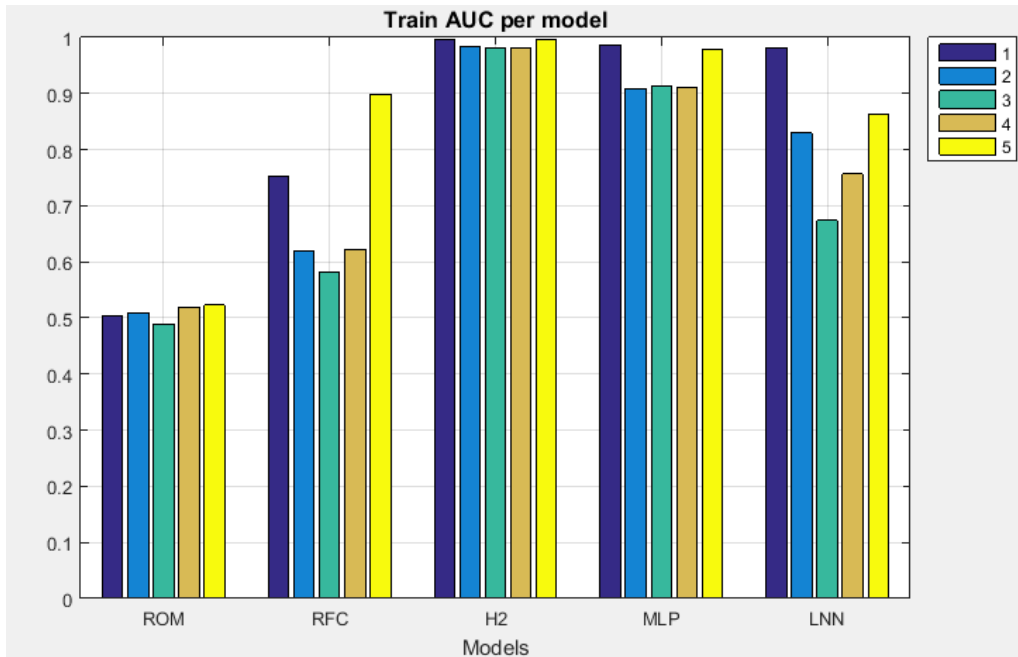


Figure 6-9 : Training AUC for all models (Phase 3)

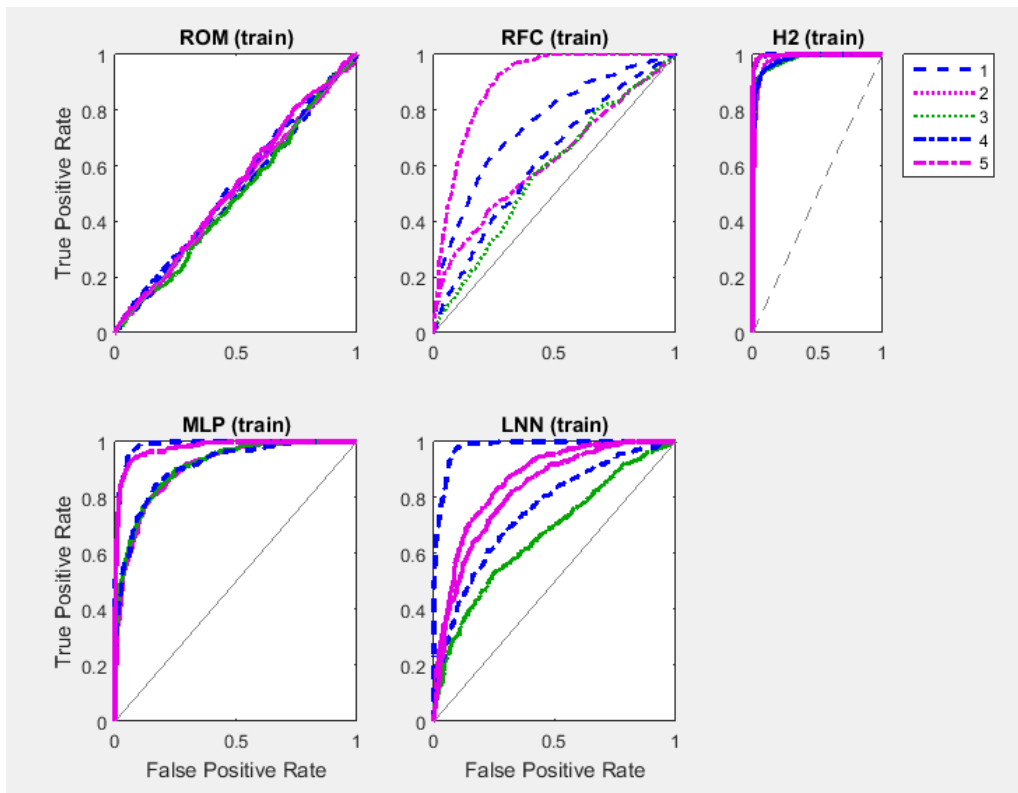


Figure 6-10: Training ROC curve for all models (Phase 3)

Table 6-11: Models Performance for all Classes (Training) (Phase 3)

Model	Class	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
<b>ROM</b>	Class 1	0.512	0.49	0.193	0.281	0.0018	0.494	0.505
	Class 2	0.489	0.545	0.217	0.301	0.0342	0.534	0.509
	Class 3	0.472	0.517	0.195	0.276	-0.0108	0.508	0.489
	Class 4	0.514	0.539	0.218	0.306	0.0534	0.534	0.519
	Class 5	0.545	0.493	0.216	0.31	0.0374	0.503	0.523
<b>RFC</b>	Class 1	0.683	0.706	0.357	0.469	0.389	0.702	0.753
	Class 2	0.531	0.645	0.278	0.365	0.176	0.622	0.62
	Class 3	0.567	0.59	0.255	0.351	0.156	0.585	0.582
	Class 4	0.606	0.586	0.267	0.371	0.191	0.59	0.623
	Class 5	0.881	0.773	0.499	0.637	0.653	0.795	0.898
<b>H2</b>	Class 1	0.974	0.974	0.898	0.934	0.947	0.974	0.996
	Class 2	0.933	0.934	0.785	0.853	0.867	0.934	0.983
	Class 3	0.921	0.952	0.825	0.87	0.872	0.946	0.98
	Class 4	0.929	0.936	0.783	0.85	0.865	0.934	0.982
	Class 5	0.985	0.968	0.888	0.934	0.953	0.972	0.997
<b>MLP</b>	Class 1	0.96	0.943	0.8	0.873	0.903	0.946	0.987
	Class 2	0.799	0.855	0.588	0.677	0.654	0.844	0.908
	Class 3	0.835	0.824	0.54	0.656	0.66	0.826	0.914
	Class 4	0.847	0.833	0.558	0.673	0.68	0.836	0.911
	Class 5	0.933	0.933	0.781	0.85	0.866	0.933	0.979
<b>LNN</b>	Class 1	0.955	0.923	0.748	0.839	0.878	0.929	0.982
	Class 2	0.769	0.732	0.425	0.548	0.501	0.739	0.83
	Class 3	0.535	0.75	0.346	0.42	0.285	0.707	0.673
	Class 4	0.679	0.712	0.37	0.479	0.391	0.705	0.756
	Class 5	0.823	0.75	0.459	0.589	0.574	0.765	0.864



## 6.5.2 Final Classification Test Results

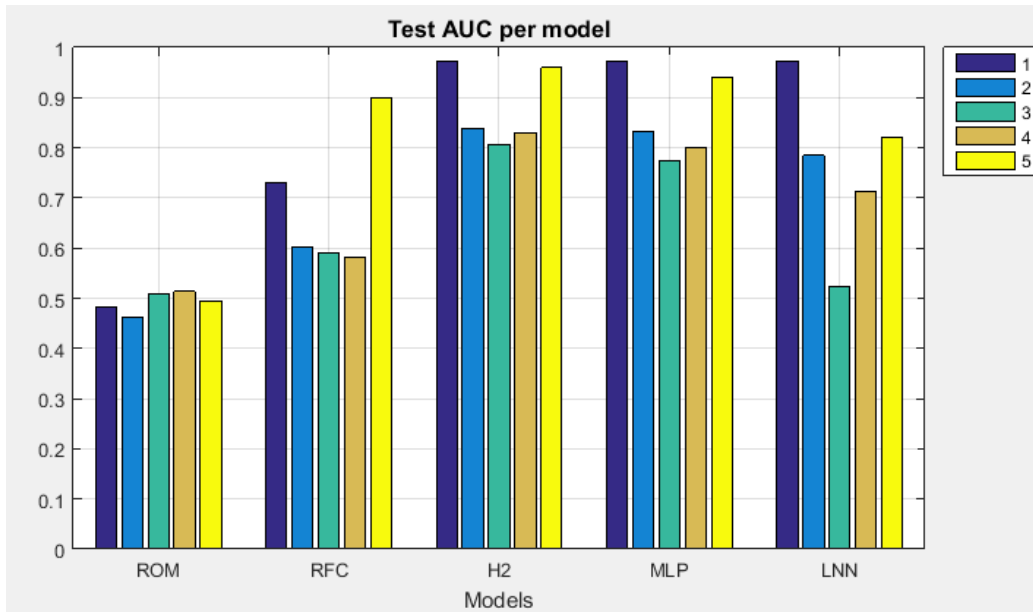


Figure 6-11 : Test AUC for all models (Phase 3)

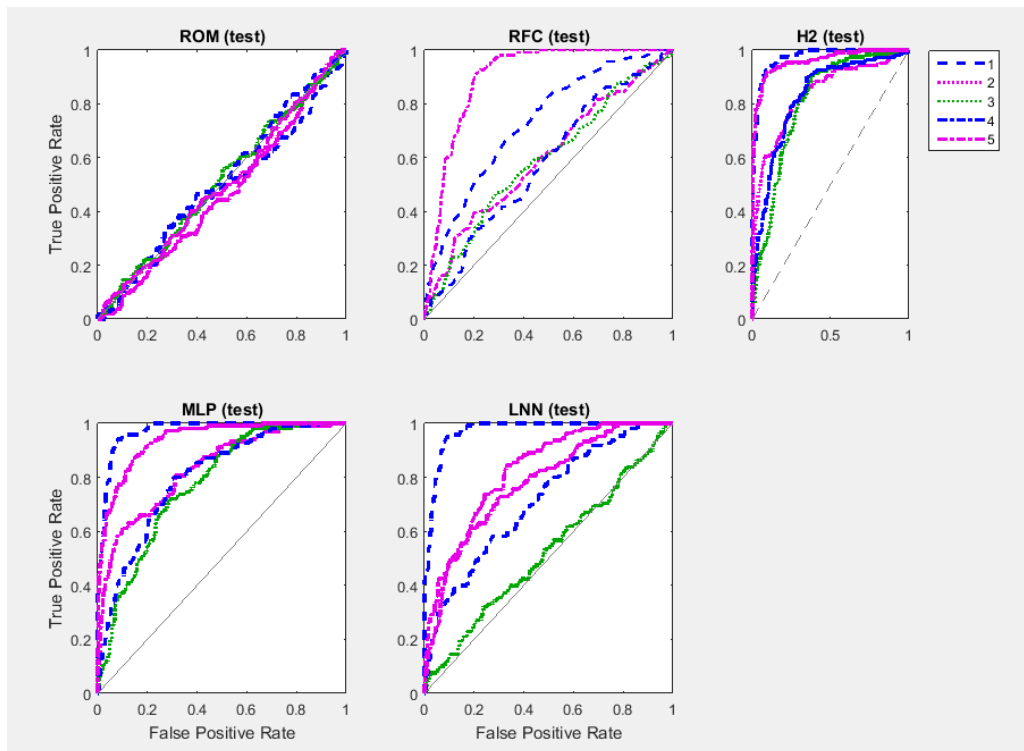


Figure 6-12 : Test ROC curve for all models (Phase 3)

Table 6-12 - Part 3: Models Performance for all Classes (Test)

<b>Model</b>	<b>Class</b>	<b>Sensitivity</b>	<b>Specificity</b>	<b>Precision</b>	<b>F1</b>	<b>J</b>	<b>Accuracy</b>	<b>AUC</b>
<b>ROM</b>	Class 1	0.417	0.639	0.239	0.304	0.0561	0.592	0.481
	Class 2	0.433	0.536	0.175	0.249	-0.0312	0.517	0.462
	Class 3	0.564	0.491	0.226	0.323	0.0551	0.506	0.509
	Class 4	0.464	0.605	0.223	0.301	0.069	0.578	0.513
	Class 5	0.445	0.581	0.206	0.282	0.0264	0.554	0.495
<b>RFC</b>	Class 1	0.667	0.673	0.357	0.465	0.34	0.672	0.732
	Class 2	0.596	0.538	0.227	0.329	0.134	0.549	0.601
	Class 3	0.556	0.604	0.27	0.363	0.159	0.594	0.591
	Class 4	0.627	0.499	0.234	0.341	0.126	0.524	0.582
	Class 5	0.909	0.798	0.524	0.664	0.707	0.82	0.898
<b>H2</b>	Class 1	0.942	0.907	0.734	0.825	0.849	0.914	0.973
	Class 2	0.788	0.753	0.421	0.548	0.541	0.759	0.838
	Class 3	0.812	0.703	0.419	0.552	0.515	0.725	0.806
	Class 4	0.791	0.758	0.444	0.569	0.549	0.765	0.83
	Class 5	0.918	0.909	0.711	0.802	0.827	0.911	0.959
<b>MLP</b>	Class 1	0.942	0.918	0.758	0.84	0.86	0.923	0.973
	Class 2	0.806	0.69	0.369	0.506	0.496	0.711	0.832
	Class 3	0.72	0.713	0.401	0.515	0.434	0.715	0.775
	Class 4	0.8	0.694	0.389	0.524	0.494	0.715	0.8
	Class 5	0.873	0.856	0.596	0.708	0.729	0.859	0.941
<b>LNN</b>	Class 1	0.95	0.912	0.745	0.835	0.862	0.92	0.972
	Class 2	0.728	0.703	0.355	0.478	0.431	0.708	0.785
	Class 3	0.534	0.517	0.227	0.319	0.0508	0.52	0.524
	Class 4	0.582	0.729	0.344	0.432	0.311	0.701	0.713
	Class 5	0.736	0.758	0.426	0.54	0.495	0.754	0.82

## 6.6 Discussion

Following the extensive background research that resulted in the formalisation of the framework concept, prediction approaches, and the concept of ranking the risk factors based on clinical instinct, knowledge and experience using a mathematical algorithm, we conducted three experiments to get further insight and investigate the disease further using machine learning models.

According to the Alzheimer's Association there are no current working methods to diagnose Alzheimer's disease at a very early stage and the "current diagnosis of Alzheimer's relies largely on documenting mental decline" [144]. The method used to diagnose Alzheimer's disease is by using the Mini Mental Score Examination test and in some cases a brain scan. Unfortunately, methods such as brain scans detect Alzheimer's disease at a very late stage when all of the symptoms appear [3]. However, the more knowledge gained on Alzheimer's disease, the closer scientists get to solve its mysterious cause.

We conducted three experiments in this study on a dataset which was extracted from a larger dataset provided by Alzheimer's disease Neuroimaging Initiative (ADNI). All subjects were unidentifiable, and the extracted data contained information that is related to Alzheimer's disease possible risk factors. The risk factors are categorised as behavioural risk factors such as lifestyle, demographic and characteristics, and biological risk factors such as medical history, genetics, and symptoms of sickness (see Chapter 2).

Our initial investigation was purely used as an approach to begin a long journey into early prediction of Alzheimer's disease. The outcome of the work carried out, supports the need to

involve more underlying data related to both behavioural and biological markers of Alzheimer's disease as the classifying models struggled to differentiate the dataset subject classes. The overall aim of the study is to improve the accuracy of early diagnosis of Alzheimer's disease, and start the foundation of a predictive dynamic framework, that aims to help with early prediction of Alzheimer's disease, collection of valuable relevant data, support Healthcare Professionals with diagnosis decision-making and provide an insight into Alzheimer's disease.

Identifying the causes of Alzheimer's disease is a very challenging task as it is caused by multiple risk factors. The use of Machine Learning to assist in the diagnosis and prediction process of Alzheimer's disease will help us learn more about the disease and its behaviour. Using Machine Learning algorithms, computers can analyse and extract patterns from multivariable datasets far more quickly compared to humans. Early prediction of Alzheimer's disease is a very challenging path of study since there is a limited amount of knowledge revealed on its underlying causes. However, that does not mean it is not possible. Alzheimer's disease is like any other disease; it is caused by abnormal genetic mutation of the cells. At some point in the life of an Alzheimer's disease subject, they must have been exposed over time to risk factors that are responsible for the development of Alzheimer's disease. The apparent element in the study of Alzheimer's disease, is that besides being caused by genetic disorder from birth it is also caused by over time genetic mutation as a side effect of multiple high-risk factors such as lifestyle, medical vascular diseases and genetics type.

Overall our study aimed to explore the ADNI dataset and underlying connection between the risk factors. Though the outcome is based on the ADNI dataset and it cannot claim that the

summary of it is medically of significance. For our second phase of investigation we employed four different machine learning models on the dataset extract to rank variables based on importance. These models include the Random Forest (RF), Neural Networks with a Principal Component Analysis (pcaNNet), Support Vector Machines with Linear Kernel (svmLinear), and Multi-Layer Perceptron (MLP). The overall output of the models combined suggests that all variables had some sort of importance but variables that had highest importance are energy, vision, dizziness, depression and headaches. This is very interesting as current research shows that social inactivity and sport inactivity are related to Alzheimer's disease and could possibly be contributing risk factors [17] [18]. Lack of energy could indicate that the subject had a low sporty lifestyle, and correspondingly depression could also indicate a lazy and anti-social lifestyle. Risk factors of Alzheimer's disease are further discussed in one of our previous works and we proposed a framework to predict onset Alzheimer's disease [19]. The stated results above were mainly influenced by the results for early mild cognitive impairment subjects as the importance level was very high for that class. However, there are other high importance variables beside energy and vision, APOE4, age, and father's dementia history have scored high importance also. This also indicates that genetics also plays a vital role in the development of Alzheimer's disease and as such current research recognises the important role of APOE4 in relation to Alzheimer's disease [20].

In our final investigation we employed the same classifying models on an enhanced dataset in efforts to improve their performance. Because of the limitation of the data we needed to use an over-sampling technique and involved more variables related to different risk factors including medical history, lifestyle, genetic type, demography and family history. The prediction was

successful for two of the models, however, this is based on this result and the data variable, and we can't state that the journey towards early onset prediction is complete. The data variable contained a cognitive test score and this for some Alzheimer's disease patients is not an onset indicator. Though without this the classifiers still improved in the classification since the initial set of investigations on the baseline dataset, which indicates that more information about the risk factors and their connection is the key toward early onset prediction.

## 6.7 Summary

In this chapter we provided and in-depth discusses of the results generated from the machine learning models deployment of the ADNI dataset. The overall work conducted is summaries in Figure 6-13, which is a flowchart highlighting the work from start to end. Chapter 2 discusses the background research, Chapter 3 is the literature review in which discusses techniques and methods we used when building the presented framework, in Chapter 4 we discussed the presented framework and Chapter 5 comprehensively discusses the implementation of the framework. In this chapter we presented the results from the experiments, however the details of the experiments are spread across Chapter 5 and Chapter 6, in Table 6-13 we provide references to the location where the details on the experiments are discussed.

Table 6-13 Information on all Experiments

<b>Experiments</b>	<b>Data Analysis</b>	<b>Experimental Setup</b>	<b>Experimental Results</b>
Phase 1	Details in section 5.4.1	Details in section 5.6.1	Details in section 6.3
Phase 2	Details in section 5.4.2	Details in section 5.6.2	Details in section 6.4
Phase 3	Details in section 5.4.2	Details in section 5.6.3	Details in section 6.5

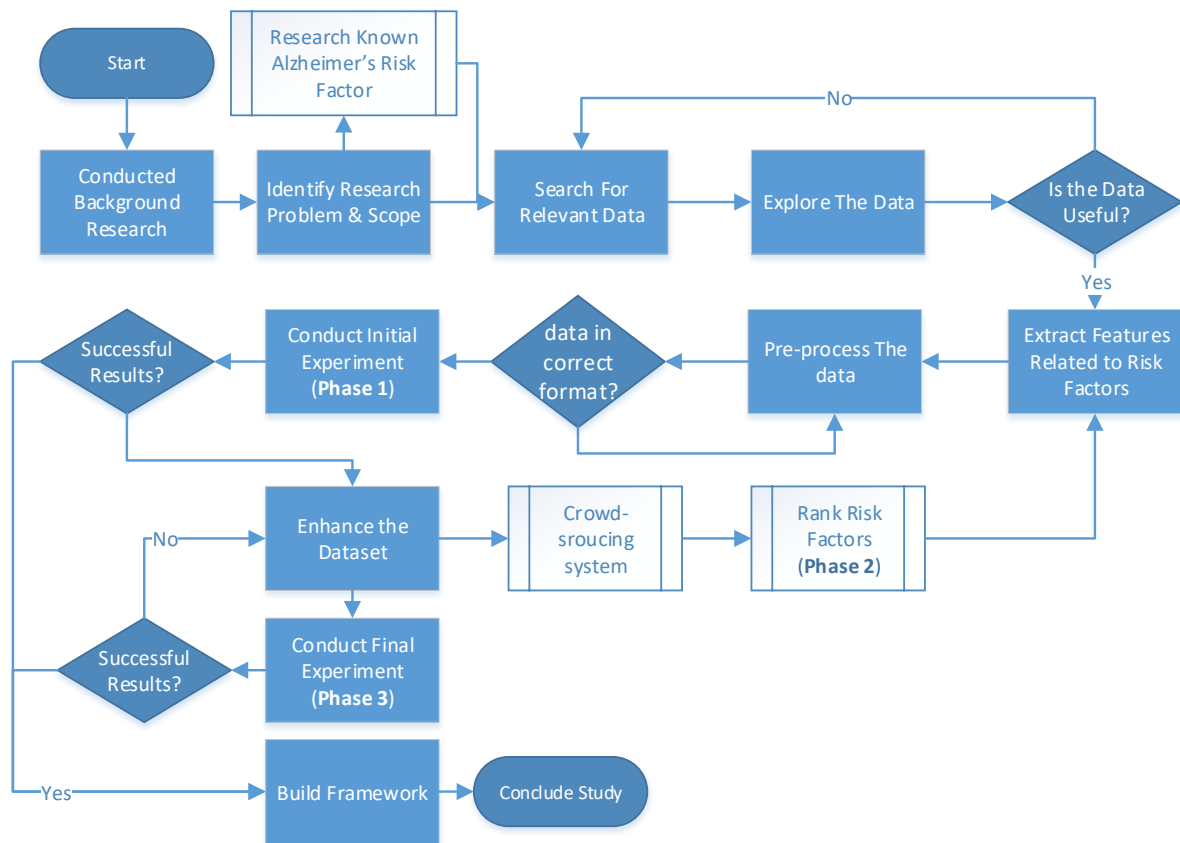


Figure 6-13 Research and work carried out in this thesis

The Phase 1 results in this chapter are for our initial experiment on a limited baseline dataset extract from the ADNI data, these results helped us understand the data further as well as measure the performance of the learning models to classify the dataset. Phase 2 results in this chapter are for our second experiment on an extended dataset we used to rank the risk factors and to extract further more knowledge with a comprehensive discussion of the ranking results. Phase 3 of the results in this chapter are for our final experiment on the extended dataset to measure the performance of the machine learning models in providing an early prediction of Alzheimer's disease. In this chapter we also provided a discussion of for all of the 3 phases of our experimentations. The next chapter will present our conclusion and recommendations for future work.

# Chapter 7      **Conclusion and Future Work**

## **7.1 Introduction**

Here, we give our conclusion on the research problem and the work we have carried out to this point. The lessons we have learnt and the work we carried out toward the achievement of the objectives in this thesis will positively be a roadmap for future work toward practical and structured onset prediction and a cost-effective diagnosis of Alzheimer's disease.

## **7.2 Conclusion**

In conclusion, it is difficult to manually diagnose Alzheimer's disease, or any other types of dementia at an early stage before most of its symptoms are noticeable. General Practitioners (GP) would usually rely on diagnosing Alzheimer's disease through manual evaluation by analysing its symptoms using several standards and procedures. For example, in the United Kingdom the Department of Health recommends GPs to first use their clinical instinct to conduct a manual evaluation, but this often results in a misdiagnoses of the disease, and even if the diagnosis was correct this normally happens after the disease have advanced rapidly into the brain and there's nothing available to reverse or stop its development.

The motivation behind this study is the fact that there have been no major breakthroughs and scientists are still unsure of what is the actual cause of Alzheimer's disease and don't have any cure for it, even though there is a valuable amount of knowledge and information that have been gained on the disease, yet these knowledge is not enough to pin point the actual causes of



the disease, and like any disease, it is important that we know its risk factors and avoid them. Researchers from different fields such as biology, physiology, neurology, computer science and others have been exploring Alzheimer's disease for decades, and their work indicates that the pathology of Alzheimer's disease progresses through different channels, and the main categories of its risk factors are age, genetics, medical history, lifestyle, and characteristics / demography (see Chapter 2).

Health Service providers and major research institutions around the world have published multiple lists of risk factors that are potentially responsible for the development of Alzheimer's disease. However, these risk factors do not mean that they are the only reasons behind the development of Alzheimer's disease as research shows that risk factors varies from one person to another and there are other hidden factors that are currently unknown.

These problems and challenges imposes the need for an early detection of Alzheimer's disease, and the presence of methods that would enable us to learn more about the hidden patterns that leads to its pathology. Hence, at the start of this study we had one aim and seven objectives that we needed to accomplish to achieve our aim to predict early onset of Alzheimer's disease. In this study we present a framework which we believe with enough data and continues use will predict Alzheimer's as early as possible. We carried out work in this study that supports our claim; started with an extensive research on Alzheimer's disease, different risk factors and their relevance to the pathology of Alzheimer's disease which can be found in Chapter 2, then we carried out a literature review study on different learning methods and techniques to use as well as related work in this field which we included in Chapter 3. We presented logical approaches for early prediction and our proposed framework to be used as a future pathway

toward early prediction of Alzheimer’s disease (see Chapter 4), and during the implementation of this framework we investigated and analysed the ADNI database from a machine learning perspective for potential early onset prediction of Alzheimer’s disease, as well as presented a sub-component of the framework which is a crowdsourcing system with an algorithm to rank disease risk factors and contributors based on clinical knowledge and experience (see Chapter 5). To build the machine learning component of the framework we investigated the performance of five machine learning classification models with different architectures on two different datasets, as well as conducting an experiment to rank the risk factors based on importance and evaluated the results for each machine learning models (see Chapter 5 and 6). Table 7-1 summaries the methods used to achieve the objectives of the thesis, here, in the table below we have the objectives listed in the first column and the corresponding work carried out and methods used to achieve it in the second column.

Table 7-1 Objectives and Work Carried Out

<b>Objective</b>	<b>Work Carried Out</b>
To gain an in-depth understanding of Alzheimer’s disease risk factors and the available Alzheimer’s disease datasets provided by ADNI, and extract the relevant data related to the research aim.	An extensive background research was carried out on the disease as well as opening dialogues with professionals and experts on dementia and cognitive health. The work conducted for this objective is included in Chapter 2 and its sub-section, where we provided background research and analysis of Alzheimer’s disease risk factors. .
To prepare a dataset by applying pre-processing,	Real data was required to investigate Alzheimer Disease using Machine Learning. No ethical approval was needed, but access was granted to use anonymised Alzheimer’s

<p>balancing and filtering techniques.</p>	<p>patients' data from ADNI. Data was extracted from multiple comma separated files (CSV) and unified to one set of data to be used as a baseline dataset for the experimentations. The data collected required a lot of pre-processing before the employment of machine learning models. This involved cleaning, normalization, balancing, and dealing with categorical data. The work carried out to achieve this objective is included in Chapter 5 in section 5.2 and 5.3.</p>
<p>To conduct exploratory data analysis for further understand the data and select the relevant features that would assist the early prediction of Alzheimer's disease.</p>	<p>Before conducting the experimentations we conducted an analysis and visualisation of the data in order to form an understanding of what type of learning and models were to be used, this can be found in Chapter 5 in section 5.4.</p>
<p>To develop a crowd-sourcing system to collect manual evaluation of risk factors' interrelationships from an epidemical and pathological prospective.</p>	<p>A crowdsourcing risk factor ranking system (CRFR) as a subcomponent of our framework was developed to collect clinical evolution of Alzheimer's risk factors and their contributors. The system uses a ranking algorithm to convert the collected knowledge to numerical sum, the work on this objective can be found in Chapter 5 section 5.5.</p>
<p>To deploy machine learning models on the baseline data to produce automatic weighting for the risk factors based on their correlation.</p>	<p>To get further insight and investigate the disease further using machine learning models, we conducted three experiments in this study, two of the experiments were for classification and prediction of Alzheimer's disease why one of the experiments was to rank the risk factors in the</p>

	dataset by importance. The work performed here is split into three different phases, the details and discussion of this work can be found in Chapter 5, Chapter 6, and in Table 6-13 we provide references to the location where the details on each experiment are discussed.
To propose a new framework to detect Alzheimer's disease before it causes severe brain damage.	Based on our overall work during this study we formulated and presented a strategic framework called Early Prediction of Alzheimer's Disease Framework (EPADf) that would give a future prediction of early onset Alzheimer's disease. Chapter 4 includes a discussion on the framework and the logical approaches we used during this study.
To dissemination of research findings and outcomes in international specialised venues and events.	We produced research papers during this study and published them in international specialised venues. The list of publications can be found in page XVI.

Our framework is a working concept we have implemented it on a limit dataset but the work we carried out to this point to build its sub components show a working method to predict the disease using behavioural and bio markers data that can be obtained through assessment or cost effective methods, as well as a working method to rank and analysis these risk factors and extract the hidden knowledge in the dataset. Owe conclude that we full implementation of such framework in an active environment with continuous feed of data will predict the development of Alzheimer's disease, uncover hidden patterns and help people avoid the early contributing risk factors. This framework can also be applied to predict other disease and uncover the hidden pattern in their pathological development. This is discuss further in the future work section.

### 7.3 Future Work

This is the end of a long journey for me, but without a doubt the start of an even longer journey toward the final aim of fully understanding Alzheimer's disease, well, at least from a computational perspective. This thesis concludes our work to date and motivates us for future works. This section will contain future recommendations, limitations, and challenges that could be used to further develop the work of this thesis.

With the results of our experiential study, we consider further work directions, including improvements to the framework. Our framework aims to predict early onset of Alzheimer's disease, and what we have done thus far opens doors for potential research directions to investigate Alzheimer's disease further. The framework consists of multiple sub-components and perhaps the first point of improvements would be to collaborate with clinicians, doctors, and researchers to make most of the crowdsourcing system and input their knowledge into its platform, this would open doors to further understand the disease and even to understand the details of each risk factor and their contributors. This component of the framework can also be used to collect clinical evaluation about other disease and not only Alzheimer's disease, it has the potential of expanding to become like an epidemic knowledge graph that provides an insight into diseases and give machines direct access to clinical evaluation.

The success of such crowdsourcing system will open doors for further research and mining of hidden pattern behind the development of diseases, and it will also allow further improvements and development of this framework. A potential research direction would be the building of a data collection strategy in active real world environment such as hospital, home-care and local

clinics. For example using data that is related to patient's diet, lifestyle, and complete medical history. From my experience while working at Med eTrax, we developed an in-hospital electronic observation system and an at-home electronic observation systems that keeps track of a patient's wellbeing. The information is collected through a mobile application that uses sensors to get data as well as assessment questionnaires answered by the patients or carers. The most promising project currently under development by the company is the non-invasive point of care that uses sensors to get real-time full blood spectrum scan. Effective utilisation of such data and solutions to integrate them to the proposed framework here would provide an evolving system. These solutions can be used for data collection by designing a set of assessments to target biomarkers in Alzheimer's patients from their medical history, lifestyle, DNA, demography and diet.

## References

- [1] “World Alzheimer Reports | Alzheimer’s Disease International.” [Online]. Available: <http://www.alz.co.uk/research/world-report>. [Accessed: 21-Mar-2016].
- [2] B. Reisberg, J. Borenstein, S. P. Salob, S. H. Ferris, and et al, “Behavioral symptoms in Alzheimer’s disease: Phenomenology and treatment.,” *J. Clin. Psychiatry*, 1987.
- [3] M. Fernández, A. L. Gobartt, and M. Balañá, “Behavioural symptoms in patients with Alzheimer’s disease and their association with cognitive impairment.,” *BMC Neurol.*, vol. 10, no. 1, p. 87, Jan. 2010.
- [4] Alzheimer’s Association, “Dementia Types | Signs, Symptoms, & Diagnosis.” [Online]. Available: <http://www.alz.org/dementia/types-of-dementia.asp>.
- [5] Z. Nagy *et al.*, “Relative Roles of Plaques and Tangles in the Dementia of Alzheimer’s Disease: Correlations Using Three Sets of Neuropathological Criteria,” *Dement. Geriatr. Cogn. Disord.*, vol. 6, no. 1, pp. 21–31, 1995.
- [6] National Institute on Aging, “Progress Report on Alzheimer ’ S Disease,” 2000.
- [7] “Brain Plaques and Tangles.” [Online]. Available: [https://www.alz.org/braintour/plaques\\_tangles.asp](https://www.alz.org/braintour/plaques_tangles.asp). [Accessed: 11-May-2015].
- [8] T. D. Bird, *Alzheimer Disease Overview*. 1993.
- [9] T. D. Bird, “Genetic aspects of Alzheimer disease.,” *Genet. Med.*, vol. 10, no. 4, pp. 231–9, Apr. 2008.
- [10] M. S. Chong and S. Sahadevan, “Preclinical Alzheimer’s disease: Diagnosis and prediction of progression,” *Lancet Neurol.*, vol. 4, no. 9, pp. 576–579, 2005.
- [11] “2014 Alzheimer’s Disease - Facts and Figures.” [Online]. Available: [http://www.alz.org/downloads/Facts\\_Figures\\_2014.pdf](http://www.alz.org/downloads/Facts_Figures_2014.pdf). [Accessed: 21-Mar-2016].
- [12] L. F. M. Scinto and K. R. Daffner, Eds., *Early Diagnosis of Alzheimer’s Disease*. Totowa, NJ: Humana Press, 2000.
- [13] Alzheimer’s Society, “The Mini Mental State Examination (MMSE),” *Alzheimer’s Society*, 2015. [Online]. Available: [http://www.alzheimers.org.uk/site/scripts/documents\\_info.php?documentID=121](http://www.alzheimers.org.uk/site/scripts/documents_info.php?documentID=121). [Accessed: 07-May-2015].
- [14] N. Choices, “What causes dementia? - Dementia guide - NHS Choices,” *NHS UK*, 2015. [Online]. Available: <http://www.nhs.uk/conditions/dementia-guide/pages/causes-of-dementia.aspx>. [Accessed: 11-May-2015].
- [15] T. Hubbard-Green, “Demography,” *Alzheimer’s Society*, 2015. [Online]. Available: [http://www.alzheimers.org.uk/site/scripts/documents\\_info.php?documentID=412](http://www.alzheimers.org.uk/site/scripts/documents_info.php?documentID=412).

- [Accessed: 11-May-2015].
- [16] “2011 Alzheimer’s disease facts and figures.,” *Alzheimers. Dement.*, vol. 7, no. 2, pp. 208–44, Mar. 2011.
- [17] M. A. Arbib, *The handbook of brain theory and neural networks*. THE MIT PRESS, 1995.
- [18] D. M. Holtzman, J. C. Morris, and A. M. Goate, “Alzheimer’s disease: the challenge of the second century.,” *Sci. Transl. Med.*, vol. 3, no. 77, p. 77sr1, Apr. 2011.
- [19] D. of Health, “The NHS introduced Dementia Enhanced Service,” 2015. [Online]. Available: <https://twitter.com/DHgovuk/status/560731492629090305>.
- [20] N. England, “FACILITATING TIMELY DIAGNOSIS AND SUPPORT FOR PEOPLE WITH DEMENTIA,” 2014.
- [21] “Clinical Criteria for Alzheimer’s Diagnosis | Research Center | Alzheimer’s Association.” [Online]. Available: [http://www.alz.org/research/diagnostic\\_criteria/](http://www.alz.org/research/diagnostic_criteria/). [Accessed: 17-May-2015].
- [22] L. Kurlowicz and M. Wallace, “The Mini Mental State Examination,” *Alzheimer’s Soc.*, pp. 1–2, 1999.
- [23] F. 450lp, “Risk factors for dementia,” 2016.
- [24] Y.-P. Tang and E. S. Gershon, “Genetic studies in Alzheimer’s disease.,” *Dialogues Clin. Neurosci.*, vol. 5, no. 1, pp. 17–26, Mar. 2003.
- [25] “Alzheimer’s disease and diabetes.”
- [26] J. A. Luchsinger, M.-X. Tang, Y. Stern, S. Shea, and R. Mayeux, “Diabetes Mellitus and Risk of Alzheimer’s Disease and Dementia with Stroke in a Multiethnic Cohort,” 2001.
- [27] Phe, “The effect of midlife risk factors on dementia in older age,” 2017.
- [28] K. Bisht, K. Sharma, and M.-È. Tremblay, “Chronic stress as a risk factor for Alzheimer’s disease: Roles of microglia-mediated synaptic remodeling, inflammation, and oxidative stress,” *Neurobiol. Stress*, vol. 9, pp. 9–21, Nov. 2018.
- [29] “Is there a link between stress and dementia risk? | Alzheimer’s Society.” [Online]. Available: <https://www.alzheimers.org.uk/blog/there-link-between-stress-and-dementia-risk>. [Accessed: 28-Apr-2019].
- [30] M. S. Greenberg, K. Tanev, M.-F. Marin, and R. K. Pitman, “Stress, PTSD, and dementia.,” *Alzheimers. Dement.*, vol. 10, no. 3 Suppl, pp. S155-65, Jun. 2014.
- [31] C. . Martyn, C. Osmond, J. . Edwardson, D. J. . Barker, E. . Harris, and R. . Lacey, “GEOGRAPHICAL RELATION BETWEEN ALZHEIMER’S DISEASE AND ALUMINIUM IN DRINKING WATER,” *Lancet*, vol. 333, no. 8629, pp. 59–62, Jan. 1989.
- [32] Y. Stern, “Cognitive reserve in ageing and Alzheimer’s disease.,” *Lancet. Neurol.*, vol. 11, no. 11, pp. 1006–12, Nov. 2012.
- [33] M. Gatz, “Educating the brain to avoid dementia: can mental exercise prevent



- Alzheimer disease?,” *PLoS Med.*, vol. 2, no. 1, p. e7, Jan. 2005.
- [34] E. S. Sharp and M. Gatz, “Relationship between education and dementia: an updated systematic review.,” *Alzheimer Dis. Assoc. Disord.*, vol. 25, no. 4, pp. 289–304, 2011.
- [35] H.-X. Wang, M. Wahlberg, A. Karp, B. Winblad, and L. Fratiglioni, “Psychosocial stress at work is associated with increased dementia risk in late life,” *Alzheimer’s Dement.*, vol. 8, no. 2, pp. 114–120, Mar. 2012.
- [36] A. J. Griffiths, J. H. Miller, D. T. Suzuki, R. C. Lewontin, and W. M. Gelbart, “An Introduction to Genetic Analysis. 7th edition.” W. H. Freeman, 2000.
- [37] A. J. Hartz, D. C. Rupley, R. D. Kalkhoff, and A. A. Rimm, “Relationship of obesity to diabetes: Influence of obesity level and body fat distribution,” *Prev. Med. (Baltim).*, vol. 12, no. 2, pp. 351–357, Mar. 1983.
- [38] “Study sheds light on link between cholesterol and diabetes — Oxford Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU).” [Online]. Available: <https://www.ctsuo.ox.ac.uk/news/study-sheds-light-on-link-between-cholesterol-and-diabetes>. [Accessed: 28-Apr-2019].
- [39] “Diabetes and heart disease | Cardiovascular disease | Diabetes UK.” [Online]. Available: [https://www.diabetes.org.uk/guide-to-diabetes/complications/cardiovascular\\_disease](https://www.diabetes.org.uk/guide-to-diabetes/complications/cardiovascular_disease). [Accessed: 28-Apr-2019].
- [40] J. Kaiser, “Cancer. Cholesterol forges link between obesity and breast cancer.,” *Science*, vol. 342, no. 6162, p. 1028, Nov. 2013.
- [41] D. Nemiary, R. Shim, G. Mattox, and K. Holden, “The Relationship Between Obesity and Depression Among Adolescents.,” *Psychiatr. Ann.*, vol. 42, no. 8, pp. 305–308, Aug. 2012.
- [42] M. Razzoli and A. Bartolomucci, “The Dichotomous Effect of Chronic Stress on Obesity.,” *Trends Endocrinol. Metab.*, vol. 27, no. 7, pp. 504–515, 2016.
- [43] “How High Blood Pressure Can Lead to Stroke | American Heart Association.” [Online]. Available: <https://www.heart.org/en/health-topics/high-blood-pressure/health-threats-from-high-blood-pressure/how-high-blood-pressure-can-lead-to-stroke>. [Accessed: 28-Apr-2019].
- [44] “Brain Injuries Increase Risk of Stroke.” [Online]. Available: <https://www.webmd.com/stroke/news/20110728/brain-injuries-increase-risk-of-stroke#1>. [Accessed: 28-Apr-2019].
- [45] “How does stress lead to heart attacks and stroke - British Heart Foundation.” [Online]. Available: <https://www.bhf.org.uk/information-support/heart-matters-magazine/news/behind-the-headlines/stress-and-heart-disease>. [Accessed: 28-Apr-2019].
- [46] “Genetic links between depression and obesity explored - NHS.” [Online]. Available: <https://www.nhs.uk/news/obesity/genetic-links-between-depression-and-obesity/>. [Accessed: 28-Apr-2019].

- [47] “The Stress-Depression Connection | Can Stress Cause Depression?” [Online]. Available: <https://www.webmd.com/depression/features/stress-depression#1>. [Accessed: 28-Apr-2019].
- [48] E. M. Whyte, B. H. Mulsant, J. Vanderbilt, H. H. Dodge, and M. Ganguli, “Depression After Stroke: A Prospective Epidemiological Study,” *J. Am. Geriatr. Soc.*, vol. 52, no. 5, pp. 774–778, May 2004.
- [49] “High Cholesterol - Causes & Treatments - British Heart Foundation.” [Online]. Available: <https://www.bhf.org.uk/informationsupport/risk-factors/high-cholesterol>. [Accessed: 28-Apr-2019].
- [50] “5 Diseases Linked To High Cholesterol.” [Online]. Available: <https://www.webmd.com/cholesterol-management/guide/diseases-linked-high-cholesterol>. [Accessed: 28-Apr-2019].
- [51] L. Akil and H. A. Ahmad, “Relationships between obesity and cardiovascular diseases in four southern states and Colorado,” *J. Health Care Poor Underserved*, vol. 22, no. 4 Suppl, pp. 61–72, 2011.
- [52] N. Frasure-Smith and F. Lespérance, “Recent Evidence Linking Coronary Heart Disease and Depression,” *Can. J. Psychiatry*, vol. 51, no. 12, pp. 730–737, Oct. 2006.
- [53] “High Blood Pressure and Hypertensive Heart Disease.” [Online]. Available: <https://www.webmd.com/hypertension-high-blood-pressure/guide/hypertensive-heart-disease#1>. [Accessed: 28-Apr-2019].
- [54] S.-K. C. Kwon *et al.*, “Stress and traumatic brain injury: a behavioral, proteomics, and histological study,” *Front. Neurol.*, vol. 2, p. 12, 2011.
- [55] “Traumatic Brain Injury – Causes, Symptoms and Treatments.” [Online]. Available: <https://www.aans.org/Patients/Neurosurgical-Conditions-and-Treatments/Traumatic-Brain-Injury>. [Accessed: 28-Apr-2019].
- [56] B. M. Y. Cheung and C. Li, “Diabetes and hypertension: is there a common metabolic pathway?,” *Curr. Atheroscler. Rep.*, vol. 14, no. 2, pp. 160–6, Apr. 2012.
- [57] S. Kulkarni, I. O’Farrell, M. Erasi, and M. S. Kochar, “Stress and hypertension,” *WMJ*, vol. 97, no. 11, pp. 34–8, Dec. 1998.
- [58] “Scientists reveal how beta-amyloid may cause Alzheimer’s | News Center | Stanford Medicine.” [Online]. Available: <https://med.stanford.edu/news/all-news/2013/09/scientists-reveal-how-beta-amyloid-may-cause-alzheimers.html>. [Accessed: 17-Aug-2015].
- [59] N. Scarmeas *et al.*, “Physical Activity, Diet, and Risk of Alzheimer Disease,” *JAMA*, vol. 302, no. 6, p. 627, Aug. 2009.
- [60] J. A. Luchsinger and R. Mayeux, “Dietary factors and Alzheimer’s disease,” *Lancet Neurol.*, vol. 3, no. 10, pp. 579–587, 2004.
- [61] C. Annweiler *et al.*, “Higher Vitamin D Dietary Intake Is Associated With Lower Risk of Alzheimer’s Disease: A 7-Year Follow-up,” *Journals Gerontol. Ser. A Biol. Sci.*

- Med. Sci.*, vol. 67, no. 11, pp. 1205–1211, Nov. 2012.
- [62] A. Chatterjee *et al.*, “Personality Changes in Alzheimer’s Disease,” *Arch. Neurol.*, vol. 49, no. 5, pp. 486–491, May 1992.
- [63] S. L. Ball, A. J. Holland, J. Hon, F. A. Huppert, P. Treppner, and P. C. Watson, “Personality and behaviour changes mark the early stages of Alzheimer’s disease in adults with Down’s syndrome: findings from a prospective population-based study,” *Int. J. Geriatr. Psychiatry*, vol. 21, no. 7, pp. 661–673, Jul. 2006.
- [64] B. S. McEwen and R. M. Sapolsky, “Stress and cognitive function,” *Curr. Opin. Neurobiol.*, vol. 5, no. 2, pp. 205–216, 1995.
- [65] M. Zhang *et al.*, “The prevalence of dementia and Alzheimer’s disease in Shanghai, China: Impact of age, gender, and education,” *Ann. Neurol.*, vol. 27, no. 4, pp. 428–437, Apr. 1990.
- [66] American Psychological Association, “Women and Depression Briefing Sheet.”
- [67] jwh, “SUMMIT ON WOMEN AND DEPRESSION,” *Summit Dates*.
- [68] C. Dobbins, M. Merabti, P. Fergus, and D. Llewellyn-Jones, “The big data obstacle of lifelogging,” in *Proceedings - 2014 IEEE 28th International Conference on Advanced Information Networking and Applications Workshops, IEEE WAINA 2014*, 2014, pp. 922–926.
- [69] “What is big data analytics? - Definition from WhatIs.com.” [Online]. Available: <http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics>. [Accessed: 11-May-2015].
- [70] “IBM big data platform - Bringing big data to the Enterprise.” IBM Corporation, 19-Feb-2015.
- [71] C. L. Giles, “Scholarly big data,” in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*, 2013, pp. 1–2.
- [72] “ADNI | Alzheimer’s Disease Neuroimaging Initiative.” [Online]. Available: <http://adni.loni.usc.edu/>. [Accessed: 11-Nov-2015].
- [73] X. Qu, B. Yuan, and W. Liu, “A Predictive Model for Identifying Possible MCI to AD Conversions in the ADNI Database,” in *2009 Second International Symposium on Knowledge Acquisition and Modeling*, 2009, vol. 3, pp. 102–105.
- [74] “Real-time human interaction with supervised learning algorithms for music composition and performance.” [Online]. Available: [https://scholar.google.com/citations?view\\_op=view\\_citation&hl=en&user=fEMWXvkAAAAJ&citation\\_for\\_view=fEMWXvkAAAAJ:Y0pCki6q\\_DkC](https://scholar.google.com/citations?view_op=view_citation&hl=en&user=fEMWXvkAAAAJ&citation_for_view=fEMWXvkAAAAJ:Y0pCki6q_DkC). [Accessed: 12-Feb-2016].
- [75] “The Elements of Statistical Learning.” [Online]. Available: [http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII\\_print4.pdf](http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf). [Accessed: 12-Feb-2016].

- [76] “9 Temporal-Difference Learning.” [Online]. Available: <https://web.stanford.edu/group/pdplab/pdphandbook/handbookch10.html>. [Accessed: 15-Feb-2016].
- [77] C. Gershenson, “Artificial Neural Networks for Beginners,” *Networks*, vol. cs.NE/0308, p. 8, 2003.
- [78] “Neural Network Toolbox - MATLAB - MathWorks United Kingdom.” [Online]. Available: <http://uk.mathworks.com/products/neural-network/>. [Accessed: 11-May-2015].
- [79] M. Torabi, R. D. Ardekani, and E. Fatemizadeh, “Discrimination between alzheimer’s disease and control group in MR-images based on texture analysis using artificial neural network.” pp. 79–83, 2006.
- [80] Jingdong Zhao, Zongwu Xie, Li Jiang, Hegao Cai, Hong Liu, and G. Hirzinger, “Levenberg-Marquardt Based Neural Network Control for a Five-fingered Prosthetic Hand,” in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pp. 4482–4487.
- [81] H. Liu, “On the Levenberg-Marquardt training method for feed-forward neural networks,” in *2010 Sixth International Conference on Natural Computation*, 2010, pp. 456–460.
- [82] H. P. Gavin, “The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems,” 2019.
- [83] A. Ghaffari, H. Abdollahi, M. R. Khoshayand, I. S. Bozchalooi, A. Dadgar, and M. Rafiee-Tehrani, “Performance comparison of neural network training algorithms in modeling of bimodal drug delivery,” *Int. J. Pharm.*, vol. 327, no. 1–2, pp. 126–138, Dec. 2006.
- [84] “Sigmoid Function - an overview | ScienceDirect Topics.” [Online]. Available: <https://www.sciencedirect.com/topics/computer-science/sigmoid-function>. [Accessed: 29-Apr-2019].
- [85] D. Pedomonti, “Comparison of non-linear activation functions for deep neural networks on MNIST classification task.”
- [86] R. Kruse, C. Borgelt, F. Klawonn, C. Moewes, M. Steinbrecher, and P. Held, “Multi-Layer Perceptrons,” Springer, London, 2013, pp. 47–81.
- [87] M. . Gardner and S. . Dorling, “Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences,” *Atmos. Environ.*, vol. 32, no. 14–15, pp. 2627–2636, Aug. 1998.
- [88] H. Schulz and S. Behnke, “Deep Learning,” *KI - Künstliche Intelligenz*, vol. 26, no. 4, pp. 357–363, Nov. 2012.
- [89] J. Schmidhuber, “Deep Learning in Neural Networks: An Overview,” 2014.
- [90] “CS231n Convolutional Neural Networks for Visual Recognition.” [Online]. Available: <http://cs231n.github.io/neural-networks-1/#classifier>. [Accessed: 29-Apr-

- 2019].
- [91] “Single-layer Neural Networks (Perceptrons).” [Online]. Available: <https://www.computing.dcu.ie/~humphrys/Notes/Neural/single.neural.html>. [Accessed: 29-Apr-2019].
- [92] S. Du, C. Liu, and L. Xi, “A Selective Multiclass Support Vector Machine Ensemble Classifier for Engineering Surface Classification Using High Definition Metrology,” *J. Manuf. Sci. Eng.*, vol. 137, no. 1, p. 011003, Feb. 2015.
- [93] L. R. Trambaiolli, A. C. Lorena, F. J. Fraga, P. a M. Kanda, R. Anghinah, and R. Nitri, “Improving Alzheimer’s disease diagnosis with machine learning techniques.,” *Clin. EEG Neurosci.*, vol. 42, no. 3, pp. 160–165, 2011.
- [94] I. A. Illan *et al.*, “Machine learning for very early Alzheimer’s Disease diagnosis; a 18F-FDG and PiB PET comparison,” in *IEEE Nuclear Science Symposium & Medical Imaging Conference*, 2010, pp. 2334–2337.
- [95] D. K. Srivastava and L. Bhambhu, “DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE,” 2005.
- [96] S. Shahbudin, M. Zamri, M. Kassim, S. A. C. Abdullah, and S. I. Suliman, “Weed classification using one class support vector machine,” in *2017 International Conference on Electrical, Electronics and System Engineering (ICEESE)*, 2017, pp. 7–10.
- [97] J. A. K. Suykens, J. A. K. Suykens, and J. Vandewalle, “Least Squares Support Vector Machine Classifiers LS-SVM Applications View project Machine learning for approximating the solution of dynamical systems View project Least Squares Support Vector Machine Classifiers,” 1999.
- [98] H. Zhang, “The Optimality of Naive Bayes.”
- [99] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger, “Tackling the Poor Assumptions of Naive Bayes Text Classifiers.”
- [100] Stanford.edu, “Naive Bayes text classification.” [Online]. Available: <https://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html>. [Accessed: 07-Feb-2019].
- [101] E. Frank and R. R. Bouckaert, “Naive Bayes for Text Classification with Unbalanced Classes.”
- [102] K. Javed, H. A. Babri, and M. Saeed, “Feature Selection Based on Class-Dependent Densities for High-Dimensional Binary Data,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 3, pp. 465–477, Mar. 2012.
- [103] “Data Mining Classification & Prediction.” [Online]. Available: [http://www.tutorialspoint.com/data\\_mining/dm\\_classification\\_prediction.htm](http://www.tutorialspoint.com/data_mining/dm_classification_prediction.htm). [Accessed: 11-May-2015].
- [104] S. Joshi, D. Shenoy, G. G. V. Simha, P. L. Rrashmi, K. R. Venugopal, and L. M. Patnaik, “Classification of Alzheimer’s Disease and Parkinson’s Disease by Using

- Machine Learning and Neural Network Methods,” *Mach. Learn. Comput. (ICMLC), 2010 Second Int. Conf.*, 2010.
- [105] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” *ACM SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, 1993.
- [106] B. Le Queau, O. Shafiq, and R. Alhadjj, “Analyzing Alzheimer’s disease gene expression dataset using clustering and association rule mining,” in *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*, 2014, pp. 283–290.
- [107] “Intro to Azure Machine Learning.” [Online]. Available: <https://www.slideshare.net/dshevani/intro-to-azure-machine-learning>. [Accessed: 30-Apr-2019].
- [108] A. Tashakkori and C. Teddlie, *Handbook of mixed methods in social & behavioral research*. SAGE Publications, 2003.
- [109] A. Berro, I. Megdiche, and O. Teste, “A Content-Driven ETL Processes for Open Data,” Springer, Cham, 2015, pp. 29–40.
- [110] M. Humphries, “Missing Data & How to Deal: An overview of missing data.”
- [111] G. Papageorgiou, S. W. Grant, J. J. M. Takkenberg, and M. M. Mokhles, “Statistical primer: how to deal with missing data in scientific research?†,” *Interact. Cardiovasc. Thorac. Surg.*, vol. 27, no. 2, pp. 153–158, Aug. 2018.
- [112] “Simple Methods to deal with Categorical Variables in Predictive Modeling.” [Online]. Available: <https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling/>. [Accessed: 30-Apr-2019].
- [113] “Categorical Data.” [Online]. Available: <http://www.stat.yale.edu/Courses/1997-98/101/catdat.htm>. [Accessed: 30-Apr-2019].
- [114] S. G. K. Patro and K. K. sahu, “Normalization: A Preprocessing Stage,” *IARJSET*, pp. 20–22, Mar. 2015.
- [115] B. K. Singh, N. I. T. Raipur, K. Verma, and A. S. Thoke, “Investigations on Impact of Feature Normalization Techniques on Classifier’s Performance in Breast Tumor Classification,” 2015.
- [116] “A-Z Machine Learning using Azure Machine Learning (AzureML) | Udemy.” [Online]. Available: <https://www.udemy.com/machine-learning-using-azureml/?couponCode=COUPON090>. [Accessed: 30-Apr-2019].
- [117] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
- [118] R. C. Bhagat and S. S. Patil, “Enhanced SMOTE algorithm for classification of imbalanced big-data using Random Forest,” in *2015 IEEE International Advance Computing Conference (IACC)*, 2015, pp. 403–408.
- [119] J. Wang, M. Xu, H. Wang, and J. Zhang, “Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding,” in *2006 8th international*

- Conference on Signal Processing*, 2006.
- [120] N. Andrienko and G. Andrienko, *Exploratory analysis of spatial and temporal data : a systematic approach*. Springer, 2006.
- [121] F. L. Gewers *et al.*, “Principal Component Analysis: A Natural Approach to Data Exploration.”
- [122] IBRAHIM OLATUNJI IDOWU, “Classification Techniques Using EHG Signals for Detecting Preterm Births,” *A thesis Submitt. Partial fulfilment Requir. Liverpool John Moores Univ. degree Dr. Philos.*, vol. 91, no. May, pp. 399–404, 2017.
- [123] B. A. Hoverstad, A. Tidemann, and H. Langseth, “Effects of data cleansing on load prediction algorithms,” in *2013 IEEE Computational Intelligence Applications in Smart Grid (CIASG)*, 2013, pp. 93–100.
- [124] S. Sehgal, H. Singh, M. Agarwal, V. Bhasker, and Shantanu, “Data analysis using principal component analysis,” in *2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom)*, 2014, pp. 45–48.
- [125] A. Hyvärinen, J. Karhunen, and E. Oja, “Independent Component Analysis,” 2001.
- [126] A. Hyvärinen, “A Short Introduction to Independent Component Analysis with Some Recent Advances.”
- [127] L. Van Der Maaten and G. Hinton, “Visualizing Data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.
- [128] Y. Zhang, H. Li, A. Hou, and J. Havel, “Artificial neural networks based on principal component analysis input selection for quantification in overlapped capillary electrophoresis peaks,” *Chemom. Intell. Lab. Syst.*, vol. 82, no. 1–2, pp. 165–175, May 2006.
- [129] I. Guyon, A. Elisseeff, and A. M. De, “An Introduction to Variable and Feature Selection,” *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.
- [130] D. Ververidis, M. Van Gils, J. Koikkalainen, and J. Lotjonen, “Feature selection and time regression software: Application on predicting Alzheimer’s disease progress.” pp. 1179–1183, 2010.
- [131] “LONI Image Data Archive (IDA).” [Online]. Available: <https://ida.loni.usc.edu/login.jsp?project=ADNI&page=HOME>. [Accessed: 15-Jul-2015].
- [132] G. Chetelat and J. C. Baron, “Early diagnosis of Alzheimer’s disease: Contribution of structural neuroimaging,” *Neuroimage*, vol. 18, no. 2, pp. 525–541, 2003.
- [133] A. Fred *et al.*, *IC3K 2015 : 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management : proceedings : 12-14 November 2015, Lisbon, Portugal*. .
- [134] Y. Zhang *et al.*, “Detection of subjects and brain regions related to Alzheimer’s disease using 3D MRI scans based on eigenbrain and machine learning,” *Front. Comput. Neurosci.*, vol. 9, p. 66, Jun. 2015.

- [135] R. Sathya Professor, J. Nivas College, and A. Abraham Professor, “Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification,” 2013.
- [136] “Machine Learning.” [Online]. Available: <http://www.springer.com/computer/ai/journal/10994>. [Accessed: 11-May-2015].
- [137] “ADNI Overview - Alzheimer’s Disease Neuroimaging Initiative .” [Online]. Available: <http://www.adni-info.org/Scientists/ADNIOverview.html>. [Accessed: 22-Jan-2016].
- [138] E. Rahm and H. Do, “Data cleaning: Problems and current approaches,” *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, 2000.
- [139] “Pearson’s correlation coefficient,” 2015.
- [140] L. Lyons, *A practical guide to data analysis for physical science students*. Cambridge: Cambridge University Press, 1991.
- [141] “SVM and kernel machines: linear and non-linear classification.”
- [142] B. D. James, R. S. Wilson, L. L. Barnes, and D. A. Bennett, “Late-life social activity and cognitive decline in old age.,” *J. Int. Neuropsychol. Soc.*, vol. 17, no. 6, pp. 998–1005, Nov. 2011.
- [143] J. Kim, J. M. Basak, and D. M. Holtzman, “The role of apolipoprotein E in Alzheimer’s disease.,” *Neuron*, vol. 63, no. 3, pp. 287–303, Aug. 2009.
- [144] “Alzheimer’s & Dementia Testing Advances | Research Center,” *Alzheimer’s Association*, 2015. [Online]. Available: [http://www.alz.org/research/science/earlier\\_alzheimers\\_diagnosis.asp#Biomarkers](http://www.alz.org/research/science/earlier_alzheimers_diagnosis.asp#Biomarkers). [Accessed: 17-May-2015].
- [1] “World Alzheimer Reports | Alzheimer’s Disease International.” [Online]. Available: <http://www.alz.co.uk/research/world-report>. [Accessed: 21-Mar-2016].
- [2] B. Reisberg, J. Borenstein, S. P. Salob, S. H. Ferris, and et al, “Behavioral symptoms in Alzheimer’s disease: Phenomenology and treatment.,” *J. Clin. Psychiatry*, 1987.
- [3] M. Fernández, A. L. Gobartt, and M. Balañá, “Behavioural symptoms in patients with Alzheimer’s disease and their association with cognitive impairment.,” *BMC Neurol.*, vol. 10, no. 1, p. 87, Jan. 2010.
- [4] Alzheimer’s Association, “Dementia Types | Signs, Symptoms, & Diagnosis.” [Online]. Available: <http://www.alz.org/dementia/types-of-dementia.asp>.
- [5] Z. Nagy *et al.*, “Relative Roles of Plaques and Tangles in the Dementia of Alzheimer’s Disease: Correlations Using Three Sets of Neuropathological Criteria,” *Dement. Geriatr. Cogn. Disord.*, vol. 6, no. 1, pp. 21–31, 1995.
- [6] National Institute on Aging, “Progress Report on Alzheimer ’ S Disease,” 2000.
- [7] “Brain Plaques and Tangles.” [Online]. Available: [https://www.alz.org/braintour/plaques\\_tangles.asp](https://www.alz.org/braintour/plaques_tangles.asp). [Accessed: 11-May-2015].
- [8] T. D. Bird, *Alzheimer Disease Overview*. 1993.
- [9] T. D. Bird, “Genetic aspects of Alzheimer disease.,” *Genet. Med.*, vol. 10, no. 4, pp.



- 231–9, Apr. 2008.
- [10] M. S. Chong and S. Sahadevan, “Preclinical Alzheimer’s disease: Diagnosis and prediction of progression,” *Lancet Neurol.*, vol. 4, no. 9, pp. 576–579, 2005.
- [11] “2014 Alzheimer’s Disease - Facts and Figures.” [Online]. Available: [http://www.alz.org/downloads/Facts\\_Figures\\_2014.pdf](http://www.alz.org/downloads/Facts_Figures_2014.pdf). [Accessed: 21-Mar-2016].
- [12] L. F. M. Scinto and K. R. Daffner, Eds., *Early Diagnosis of Alzheimer’s Disease*. Totowa, NJ: Humana Press, 2000.
- [13] Alzheimer’s Society, “The Mini Mental State Examination (MMSE),” *Alzheimer’s Society*, 2015. [Online]. Available: [http://www.alzheimers.org.uk/site/scripts/documents\\_info.php?documentID=121](http://www.alzheimers.org.uk/site/scripts/documents_info.php?documentID=121). [Accessed: 07-May-2015].
- [14] N. Choices, “What causes dementia? - Dementia guide - NHS Choices,” *NHS UK*, 2015. [Online]. Available: <http://www.nhs.uk/conditions/dementia-guide/pages/causes-of-dementia.aspx>. [Accessed: 11-May-2015].
- [15] T. Hubbard-Green, “Demography,” *Alzheimer’s Society*, 2015. [Online]. Available: [http://www.alzheimers.org.uk/site/scripts/documents\\_info.php?documentID=412](http://www.alzheimers.org.uk/site/scripts/documents_info.php?documentID=412). [Accessed: 11-May-2015].
- [16] “2011 Alzheimer’s disease facts and figures.,” *Alzheimers. Dement.*, vol. 7, no. 2, pp. 208–44, Mar. 2011.
- [17] M. A. Arbib, *The handbook of brain theory and neural networks*. THE MIT PRESS, 1995.
- [18] D. M. Holtzman, J. C. Morris, and A. M. Goate, “Alzheimer’s disease: the challenge of the second century.,” *Sci. Transl. Med.*, vol. 3, no. 77, p. 77sr1, Apr. 2011.
- [19] D. of Health, “The NHS introduced Dementia Enhanced Service,” 2015. [Online]. Available: <https://twitter.com/DHgovuk/status/560731492629090305>.
- [20] N. England, “FACILITATING TIMELY DIAGNOSIS AND SUPPORT FOR PEOPLE WITH DEMENTIA,” 2014.
- [21] “Clinical Criteria for Alzheimer’s Diagnosis | Research Center | Alzheimer’s Association.” [Online]. Available: [http://www.alz.org/research/diagnostic\\_criteria/](http://www.alz.org/research/diagnostic_criteria/). [Accessed: 17-May-2015].
- [22] L. Kurlowicz and M. Wallace, “The Mini Mental State Examination,” *Alzheimer’s Soc.*, pp. 1–2, 1999.
- [23] F. 450lp, “Risk factors for dementia,” 2016.
- [24] Y.-P. Tang and E. S. Gershon, “Genetic studies in Alzheimer’s disease.,” *Dialogues Clin. Neurosci.*, vol. 5, no. 1, pp. 17–26, Mar. 2003.
- [25] “Alzheimer’s disease and diabetes.”
- [26] J. A. Luchsinger, M.-X. Tang, Y. Stern, S. Shea, and R. Mayeux, “Diabetes Mellitus and Risk of Alzheimer’s Disease and Dementia with Stroke in a Multiethnic Cohort,” 2001.

- [27] Phe, “The effect of midlife risk factors on dementia in older age,” 2017.
- [28] K. Bisht, K. Sharma, and M.-È. Tremblay, “Chronic stress as a risk factor for Alzheimer’s disease: Roles of microglia-mediated synaptic remodeling, inflammation, and oxidative stress,” *Neurobiol. Stress*, vol. 9, pp. 9–21, Nov. 2018.
- [29] “Is there a link between stress and dementia risk? | Alzheimer’s Society.” [Online]. Available: <https://www.alzheimers.org.uk/blog/there-link-between-stress-and-dementia-risk>. [Accessed: 28-Apr-2019].
- [30] M. S. Greenberg, K. Tanev, M.-F. Marin, and R. K. Pitman, “Stress, PTSD, and dementia.,” *Alzheimers. Dement.*, vol. 10, no. 3 Suppl, pp. S155–65, Jun. 2014.
- [31] C. . Martyn, C. Osmond, J. . Edwardson, D. J. . Barker, E. . Harris, and R. . Lacey, “GEOGRAPHICAL RELATION BETWEEN ALZHEIMER’S DISEASE AND ALUMINIUM IN DRINKING WATER,” *Lancet*, vol. 333, no. 8629, pp. 59–62, Jan. 1989.
- [32] Y. Stern, “Cognitive reserve in ageing and Alzheimer’s disease.,” *Lancet. Neurol.*, vol. 11, no. 11, pp. 1006–12, Nov. 2012.
- [33] M. Gatz, “Educating the brain to avoid dementia: can mental exercise prevent Alzheimer disease?,” *PLoS Med.*, vol. 2, no. 1, p. e7, Jan. 2005.
- [34] E. S. Sharp and M. Gatz, “Relationship between education and dementia: an updated systematic review.,” *Alzheimer Dis. Assoc. Disord.*, vol. 25, no. 4, pp. 289–304, 2011.
- [35] H.-X. Wang, M. Wahlberg, A. Karp, B. Winblad, and L. Fratiglioni, “Psychosocial stress at work is associated with increased dementia risk in late life,” *Alzheimer’s Dement.*, vol. 8, no. 2, pp. 114–120, Mar. 2012.
- [36] A. J. Griffiths, J. H. Miller, D. T. Suzuki, R. C. Lewontin, and W. M. Gelbart, “An Introduction to Genetic Analysis. 7th edition.” W. H. Freeman, 2000.
- [37] A. J. Hartz, D. C. Rupley, R. D. Kalkhoff, and A. A. Rimm, “Relationship of obesity to diabetes: Influence of obesity level and body fat distribution,” *Prev. Med. (Baltim).*, vol. 12, no. 2, pp. 351–357, Mar. 1983.
- [38] “Study sheds light on link between cholesterol and diabetes — Oxford Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU).” [Online]. Available: <https://www.ctsuo.ox.ac.uk/news/study-sheds-light-on-link-between-cholesterol-and-diabetes>. [Accessed: 28-Apr-2019].
- [39] “Diabetes and heart disease | Cardiovascular disease | Diabetes UK.” [Online]. Available: [https://www.diabetes.org.uk/guide-to-diabetes/complications/cardiovascular\\_disease](https://www.diabetes.org.uk/guide-to-diabetes/complications/cardiovascular_disease). [Accessed: 28-Apr-2019].
- [40] J. Kaiser, “Cancer. Cholesterol forges link between obesity and breast cancer.,” *Science*, vol. 342, no. 6162, p. 1028, Nov. 2013.
- [41] D. Nemiary, R. Shim, G. Mattox, and K. Holden, “The Relationship Between Obesity and Depression Among Adolescents.,” *Psychiatr. Ann.*, vol. 42, no. 8, pp. 305–308, Aug. 2012.

- [42] M. Razzoli and A. Bartolomucci, “The Dichotomous Effect of Chronic Stress on Obesity.,” *Trends Endocrinol. Metab.*, vol. 27, no. 7, pp. 504–515, 2016.
- [43] “How High Blood Pressure Can Lead to Stroke | American Heart Association.” [Online]. Available: <https://www.heart.org/en/health-topics/high-blood-pressure/health-threats-from-high-blood-pressure/how-high-blood-pressure-can-lead-to-stroke>. [Accessed: 28-Apr-2019].
- [44] “Brain Injuries Increase Risk of Stroke.” [Online]. Available: <https://www.webmd.com/stroke/news/20110728/brain-injuries-increase-risk-of-stroke#1>. [Accessed: 28-Apr-2019].
- [45] “How does stress lead to heart attacks and stroke - British Heart Foundation.” [Online]. Available: <https://www.bhf.org.uk/information-support/heart-matters-magazine/news/behind-the-headlines/stress-and-heart-disease>. [Accessed: 28-Apr-2019].
- [46] “Genetic links between depression and obesity explored - NHS.” [Online]. Available: <https://www.nhs.uk/news/obesity/genetic-links-between-depression-and-obesity/>. [Accessed: 28-Apr-2019].
- [47] “The Stress-Depression Connection | Can Stress Cause Depression?” [Online]. Available: <https://www.webmd.com/depression/features/stress-depression#1>. [Accessed: 28-Apr-2019].
- [48] E. M. Whyte, B. H. Mulsant, J. Vanderbilt, H. H. Dodge, and M. Ganguli, “Depression After Stroke: A Prospective Epidemiological Study,” *J. Am. Geriatr. Soc.*, vol. 52, no. 5, pp. 774–778, May 2004.
- [49] “High Cholesterol - Causes & Treatments - British Heart Foundation.” [Online]. Available: <https://www.bhf.org.uk/information-support/risk-factors/high-cholesterol>. [Accessed: 28-Apr-2019].
- [50] “5 Diseases Linked To High Cholesterol.” [Online]. Available: <https://www.webmd.com/cholesterol-management/guide/diseases-linked-high-cholesterol>. [Accessed: 28-Apr-2019].
- [51] L. Akil and H. A. Ahmad, “Relationships between obesity and cardiovascular diseases in four southern states and Colorado.,” *J. Health Care Poor Underserved*, vol. 22, no. 4 Suppl, pp. 61–72, 2011.
- [52] N. Frasure-Smith and F. Lespérance, “Recent Evidence Linking Coronary Heart Disease and Depression,” *Can. J. Psychiatry*, vol. 51, no. 12, pp. 730–737, Oct. 2006.
- [53] “High Blood Pressure and Hypertensive Heart Disease.” [Online]. Available: <https://www.webmd.com/hypertension-high-blood-pressure/guide/hypertensive-heart-disease#1>. [Accessed: 28-Apr-2019].
- [54] S.-K. C. Kwon *et al.*, “Stress and traumatic brain injury: a behavioral, proteomics, and histological study.,” *Front. Neurol.*, vol. 2, p. 12, 2011.
- [55] “Traumatic Brain Injury – Causes, Symptoms and Treatments.” [Online]. Available:

- <https://www.aans.org/Patients/Neurosurgical-Conditions-and-Treatments/Traumatic-Brain-Injury>. [Accessed: 28-Apr-2019].
- [56] B. M. Y. Cheung and C. Li, "Diabetes and hypertension: is there a common metabolic pathway?," *Curr. Atheroscler. Rep.*, vol. 14, no. 2, pp. 160–6, Apr. 2012.
- [57] S. Kulkarni, I. O'Farrell, M. Erasi, and M. S. Kochar, "Stress and hypertension.," *WMJ*, vol. 97, no. 11, pp. 34–8, Dec. 1998.
- [58] "Scientists reveal how beta-amyloid may cause Alzheimer's | News Center | Stanford Medicine." [Online]. Available: <https://med.stanford.edu/news/all-news/2013/09/scientists-reveal-how-beta-amyloid-may-cause-alzheimers.html>. [Accessed: 17-Aug-2015].
- [59] N. Scarmeas *et al.*, "Physical Activity, Diet, and Risk of Alzheimer Disease," *JAMA*, vol. 302, no. 6, p. 627, Aug. 2009.
- [60] J. A. Luchsinger and R. Mayeux, "Dietary factors and Alzheimer's disease," *Lancet Neurol.*, vol. 3, no. 10, pp. 579–587, 2004.
- [61] C. Annweiler *et al.*, "Higher Vitamin D Dietary Intake Is Associated With Lower Risk of Alzheimer's Disease: A 7-Year Follow-up," *Journals Gerontol. Ser. A Biol. Sci. Med. Sci.*, vol. 67, no. 11, pp. 1205–1211, Nov. 2012.
- [62] A. Chatterjee *et al.*, "Personality Changes in Alzheimer's Disease," *Arch. Neurol.*, vol. 49, no. 5, pp. 486–491, May 1992.
- [63] S. L. Ball, A. J. Holland, J. Hon, F. A. Huppert, P. Treppner, and P. C. Watson, "Personality and behaviour changes mark the early stages of Alzheimer's disease in adults with Down's syndrome: findings from a prospective population-based study," *Int. J. Geriatr. Psychiatry*, vol. 21, no. 7, pp. 661–673, Jul. 2006.
- [64] B. S. McEwen and R. M. Sapolsky, "Stress and cognitive function," *Curr. Opin. Neurobiol.*, vol. 5, no. 2, pp. 205–216, 1995.
- [65] M. Zhang *et al.*, "The prevalence of dementia and Alzheimer's disease in Shanghai, China: Impact of age, gender, and education," *Ann. Neurol.*, vol. 27, no. 4, pp. 428–437, Apr. 1990.
- [66] American Psychological Association, "Women and Depression Briefing Sheet."
- [67] jwh, "SUMMIT ON WOMEN AND DEPRESSION," *Summit Dates*.
- [68] C. Dobbins, M. Merabti, P. Fergus, and D. Llewellyn-Jones, "The big data obstacle of lifelogging," in *Proceedings - 2014 IEEE 28th International Conference on Advanced Information Networking and Applications Workshops, IEEE WAINA 2014*, 2014, pp. 922–926.
- [69] "What is big data analytics? - Definition from WhatIs.com." [Online]. Available: <http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics>. [Accessed: 11-May-2015].
- [70] "IBM big data platform - Bringing big data to the Enterprise." IBM Corporation, 19-Feb-2015.

- [71] C. L. Giles, "Scholarly big data," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*, 2013, pp. 1–2.
- [72] "ADNI | Alzheimer's Disease Neuroimaging Initiative." [Online]. Available: <http://adni.loni.usc.edu/>. [Accessed: 11-Nov-2015].
- [73] X. Qu, B. Yuan, and W. Liu, "A Predictive Model for Identifying Possible MCI to AD Conversions in the ADNI Database," in *2009 Second International Symposium on Knowledge Acquisition and Modeling*, 2009, vol. 3, pp. 102–105.
- [74] "Real-time human interaction with supervised learning algorithms for music composition and performance." [Online]. Available: [https://scholar.google.com/citations?view\\_op=view\\_citation&hl=en&user=fEMWXvkAAAAJ&citation\\_for\\_view=fEMWXvkAAAAJ:Y0pCki6q\\_DkC](https://scholar.google.com/citations?view_op=view_citation&hl=en&user=fEMWXvkAAAAJ&citation_for_view=fEMWXvkAAAAJ:Y0pCki6q_DkC). [Accessed: 12-Feb-2016].
- [75] "The Elements of Statistical Learning." [Online]. Available: [http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII\\_print4.pdf](http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf). [Accessed: 12-Feb-2016].
- [76] "9 Temporal-Difference Learning." [Online]. Available: <https://web.stanford.edu/group/pdplab/pdphandbook/handbookch10.html>. [Accessed: 15-Feb-2016].
- [77] C. Gershenson, "Artificial Neural Networks for Beginners," *Networks*, vol. cs.NE/0308, p. 8, 2003.
- [78] "Neural Network Toolbox - MATLAB - MathWorks United Kingdom." [Online]. Available: <http://uk.mathworks.com/products/neural-network/>. [Accessed: 11-May-2015].
- [79] M. Torabi, R. D. Ardekani, and E. Fatemizadeh, "Discrimination between alzheimer's disease and control group in MR-images based on texture analysis using artificial neural network." pp. 79–83, 2006.
- [80] Jingdong Zhao, Zongwu Xie, Li Jiang, Hegao Cai, Hong Liu, and G. Hirzinger, "Levenberg-Marquardt Based Neural Network Control for a Five-fingered Prosthetic Hand," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pp. 4482–4487.
- [81] H. Liu, "On the Levenberg-Marquardt training method for feed-forward neural networks," in *2010 Sixth International Conference on Natural Computation*, 2010, pp. 456–460.
- [82] H. P. Gavin, "The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems," 2019.
- [83] A. Ghaffari, H. Abdollahi, M. R. Khoshayand, I. S. Bozchalooi, A. Dadgar, and M. Rafiee-Tehrani, "Performance comparison of neural network training algorithms in modeling of bimodal drug delivery," *Int. J. Pharm.*, vol. 327, no. 1–2, pp. 126–138,

- Dec. 2006.
- [84] “Sigmoid Function - an overview | ScienceDirect Topics.” [Online]. Available: <https://www.sciencedirect.com/topics/computer-science/sigmoid-function>. [Accessed: 29-Apr-2019].
- [85] D. Pedamonti, “Comparison of non-linear activation functions for deep neural networks on MNIST classification task.”
- [86] R. Kruse, C. Borgelt, F. Klawonn, C. Moewes, M. Steinbrecher, and P. Held, “Multi-Layer Perceptrons,” Springer, London, 2013, pp. 47–81.
- [87] M. . Gardner and S. . Dorling, “Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences,” *Atmos. Environ.*, vol. 32, no. 14–15, pp. 2627–2636, Aug. 1998.
- [88] H. Schulz and S. Behnke, “Deep Learning,” *KI - Künstliche Intelligenz*, vol. 26, no. 4, pp. 357–363, Nov. 2012.
- [89] J. Schmidhuber, “Deep Learning in Neural Networks: An Overview,” 2014.
- [90] “CS231n Convolutional Neural Networks for Visual Recognition.” [Online]. Available: <http://cs231n.github.io/neural-networks-1/#classifier>. [Accessed: 29-Apr-2019].
- [91] “Single-layer Neural Networks (Perceptrons).” [Online]. Available: <https://www.computing.dcu.ie/~humphrys/Notes/Neural/single.neural.html>. [Accessed: 29-Apr-2019].
- [92] S. Du, C. Liu, and L. Xi, “A Selective Multiclass Support Vector Machine Ensemble Classifier for Engineering Surface Classification Using High Definition Metrology,” *J. Manuf. Sci. Eng.*, vol. 137, no. 1, p. 011003, Feb. 2015.
- [93] L. R. Trambaiolli, A. C. Lorena, F. J. Fraga, P. a M. Kanda, R. Anghinah, and R. Nitri, “Improving Alzheimer’s disease diagnosis with machine learning techniques.,” *Clin. EEG Neurosci.*, vol. 42, no. 3, pp. 160–165, 2011.
- [94] I. A. Illan *et al.*, “Machine learning for very early Alzheimer’s Disease diagnosis; a 18F-FDG and PiB PET comparison,” in *IEEE Nuclear Science Symposium & Medical Imaging Conference*, 2010, pp. 2334–2337.
- [95] D. K. Srivastava and L. Bhambhu, “DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE,” 2005.
- [96] S. Shahbudin, M. Zamri, M. Kassim, S. A. C. Abdullah, and S. I. Suliman, “Weed classification using one class support vector machine,” in *2017 International Conference on Electrical, Electronics and System Engineering (ICEESE)*, 2017, pp. 7–10.
- [97] J. A. K. Suykens, J. A. K. Suykens, and J. Vandewalle, “Least Squares Support Vector Machine Classifiers LS-SVM Applications View project Machine learning for approximating the solution of dynamical systems View project Least Squares Support Vector Machine Classifiers,” 1999.

- [98] H. Zhang, “The Optimality of Naive Bayes.”
- [99] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger, “Tackling the Poor Assumptions of Naive Bayes Text Classifiers.”
- [100] Stanford.edu, “Naive Bayes text classification.” [Online]. Available: <https://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html>. [Accessed: 07-Feb-2019].
- [101] E. Frank and R. R. Bouckaert, “Naive Bayes for Text Classification with Unbalanced Classes.”
- [102] K. Javed, H. A. Babri, and M. Saeed, “Feature Selection Based on Class-Dependent Densities for High-Dimensional Binary Data,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 3, pp. 465–477, Mar. 2012.
- [103] “Data Mining Classification & Prediction.” [Online]. Available: [http://www.tutorialspoint.com/data\\_mining/dm\\_classification\\_prediction.htm](http://www.tutorialspoint.com/data_mining/dm_classification_prediction.htm). [Accessed: 11-May-2015].
- [104] S. Joshi, D. Shenoy, G. G. V. Simha, P. L. Rrashmi, K. R. Venugopal, and L. M. Patnaik, “Classification of Alzheimer’s Disease and Parkinson’s Disease by Using Machine Learning and Neural Network Methods,” *Mach. Learn. Comput. (ICMLC), 2010 Second Int. Conf.*, 2010.
- [105] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” *ACM SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, 1993.
- [106] B. Le Queau, O. Shafiq, and R. Alhadjj, “Analyzing Alzheimer’s disease gene expression dataset using clustering and association rule mining,” in *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*, 2014, pp. 283–290.
- [107] “Intro to Azure Machine Learning.” [Online]. Available: <https://www.slideshare.net/dshevani/intro-toazuremachinelearning>. [Accessed: 30-Apr-2019].
- [108] A. Tashakkori and C. Teddlie, *Handbook of mixed methods in social & behavioral research*. SAGE Publications, 2003.
- [109] A. Berro, I. Megdiche, and O. Teste, “A Content-Driven ETL Processes for Open Data,” Springer, Cham, 2015, pp. 29–40.
- [110] M. Humphries, “Missing Data & How to Deal: An overview of missing data.”
- [111] G. Papageorgiou, S. W. Grant, J. J. M. Takkenberg, and M. M. Mokhles, “Statistical primer: how to deal with missing data in scientific research?†,” *Interact. Cardiovasc. Thorac. Surg.*, vol. 27, no. 2, pp. 153–158, Aug. 2018.
- [112] “Simple Methods to deal with Categorical Variables in Predictive Modeling.” [Online]. Available: <https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling/>. [Accessed: 30-Apr-2019].
- [113] “Categorical Data.” [Online]. Available: <http://www.stat.yale.edu/Courses/1997->

- 98/101/catdat.htm. [Accessed: 30-Apr-2019].
- [114] S. G. K. Patro and K. K. sahu, “Normalization: A Preprocessing Stage,” *IARJSET*, pp. 20–22, Mar. 2015.
- [115] B. K. Singh, N. I. T. Raipur, K. Verma, and A. S. Thoke, “Investigations on Impact of Feature Normalization Techniques on Classifier’s Performance in Breast Tumor Classification,” 2015.
- [116] “A-Z Machine Learning using Azure Machine Learning (AzureML) | Udemy.” [Online]. Available: <https://www.udemy.com/machine-learning-using-azureml/?couponCode=COUPON090>. [Accessed: 30-Apr-2019].
- [117] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
- [118] R. C. Bhagat and S. S. Patil, “Enhanced SMOTE algorithm for classification of imbalanced big-data using Random Forest,” in *2015 IEEE International Advance Computing Conference (IACC)*, 2015, pp. 403–408.
- [119] J. Wang, M. Xu, H. Wang, and J. Zhang, “Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding,” in *2006 8th international Conference on Signal Processing*, 2006.
- [120] N. Andrienko and G. Andrienko, *Exploratory analysis of spatial and temporal data : a systematic approach*. Springer, 2006.
- [121] F. L. Gewers *et al.*, “Principal Component Analysis: A Natural Approach to Data Exploration.”
- [122] IBRAHIM OLATUNJI IDOWU, “Classification Techniques Using EHG Signals for Detecting Preterm Births,” *A thesis Submitt. Partial fulfilment Requir. Liverpool John Moores Univ. degree Dr. Philos.*, vol. 91, no. May, pp. 399–404, 2017.
- [123] B. A. Hoverstad, A. Tidemann, and H. Langseth, “Effects of data cleansing on load prediction algorithms,” in *2013 IEEE Computational Intelligence Applications in Smart Grid (CIASG)*, 2013, pp. 93–100.
- [124] S. Sehgal, H. Singh, M. Agarwal, V. Bhasker, and Shantanu, “Data analysis using principal component analysis,” in *2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom)*, 2014, pp. 45–48.
- [125] A. Hyvärinen, J. Karhunen, and E. Oja, “Independent Component Analysis,” 2001.
- [126] A. Hyvärinen, “A Short Introduction to Independent Component Analysis with Some Recent Advances.”
- [127] L. Van Der Maaten and G. Hinton, “Visualizing Data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.
- [128] Y. Zhang, H. Li, A. Hou, and J. Havel, “Artificial neural networks based on principal component analysis input selection for quantification in overlapped capillary electrophoresis peaks,” *Chemom. Intell. Lab. Syst.*, vol. 82, no. 1–2, pp. 165–175, May 2006.



- [129] I. Guyon, A. Elisseeff, and A. M. De, “An Introduction to Variable and Feature Selection,” *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.
- [130] D. Ververidis, M. Van Gils, J. Koikkalainen, and J. Lotjonen, “Feature selection and time regression software: Application on predicting Alzheimer’s disease progress.” pp. 1179–1183, 2010.
- [131] “LONI Image Data Archive (IDA).” [Online]. Available: <https://ida.loni.usc.edu/login.jsp?project=ADNI&page=HOME>. [Accessed: 15-Jul-2015].
- [132] G. Chetelat and J. C. Baron, “Early diagnosis of Alzheimer’s disease: Contribution of structural neuroimaging,” *Neuroimage*, vol. 18, no. 2, pp. 525–541, 2003.
- [133] A. Fred *et al.*, *IC3K 2015 : 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management : proceedings : 12-14 November 2015, Lisbon, Portugal. .*
- [134] Y. Zhang *et al.*, “Detection of subjects and brain regions related to Alzheimer’s disease using 3D MRI scans based on eigenbrain and machine learning,” *Front. Comput. Neurosci.*, vol. 9, p. 66, Jun. 2015.
- [135] R. Sathya Professor, J. Nivas College, and A. Abraham Professor, “Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification,” 2013.
- [136] “Machine Learning.” [Online]. Available: <http://www.springer.com/computer/ai/journal/10994>. [Accessed: 11-May-2015].
- [137] “ADNI Overview - Alzheimer’s Disease Neuroimaging Initiative .” [Online]. Available: <http://www.adni-info.org/Scientists/ADNIOverview.html>. [Accessed: 22-Jan-2016].
- [138] E. Rahm and H. Do, “Data cleaning: Problems and current approaches,” *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, 2000.
- [139] “Pearson’s correlation coefficient,” 2015.
- [140] L. Lyons, *A practical guide to data analysis for physical science students*. Cambridge: Cambridge University Press, 1991.
- [141] “SVM and kernel machines: linear and non-linear classification.”
- [142] B. D. James, R. S. Wilson, L. L. Barnes, and D. A. Bennett, “Late-life social activity and cognitive decline in old age.,” *J. Int. Neuropsychol. Soc.*, vol. 17, no. 6, pp. 998–1005, Nov. 2011.
- [143] J. Kim, J. M. Basak, and D. M. Holtzman, “The role of apolipoprotein E in Alzheimer’s disease.,” *Neuron*, vol. 63, no. 3, pp. 287–303, Aug. 2009.
- [144] “Alzheimer’s & Dementia Testing Advances | Research Center,” *Alzheimer’s Association*, 2015. [Online]. Available: [http://www.alz.org/research/science/earlier\\_alzheimers\\_diagnosis.asp#Biomarkers](http://www.alz.org/research/science/earlier_alzheimers_diagnosis.asp#Biomarkers). [Accessed: 17-May-2015].