Software Engineering Development of Machine Learning Approaches for Alzheimer's Disease Classification

by

Abbas Saad Alatrany

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

July 2025

Declaration

I, Abbas Saad Alatrany, declare that this thesis, submitted to Liverpool John Moores University to fulfil the Doctor of Philosophy requirements, has not been submitted to other universities and institutes. I confirm that the work described in this PhD thesis is my own except for some sources that support my research, which are appropriately cited and indicated.

Abbas Alatrany

Abstract

Alzheimer's disease (AD) is a progressive and degenerative neurological disorder that profoundly impacts daily life. As the most common form of dementia, accounting for up to 80% of all cases, AD is marked by a gradual decline in memory, thinking, and behavior. What often begins with mild symptoms progresses to severe cognitive and physical impairments that compromise independence and quality of life. Despite affecting more than 55 million people worldwide, the precise causes of AD remain unclear and no cure currently exists, though treatments can help manage symptoms and slow progression.

This thesis investigates the classification and prediction of AD by applying machine learning (ML) and data analytics techniques to genetic and multi-source datasets. A key challenge in AD research lies in the immense size of genetic data, which makes analysis computationally intensive. To overcome this, transfer learning is introduced—an approach not previously applied in this domain. Convolutional Neural Networks (CNNs) were first trained on genome-wide association study (GWAS) data from the Alzheimer's Disease Neuroimaging Initiative, and deep transfer learning was subsequently used to refine the model with a separate AD GWAS dataset. The final feature set extracted from this process was classified using a Support Vector Machine, achieving an accuracy of 89% and demonstrating the effectiveness of the proposed strategy.

Beyond predictive accuracy, high-dimensional data raises challenges for interpretability. To address this, the thesis develops a hybrid feature selection method combining association testing, principal component analysis, and the Boruta algorithm to identify key predictors of AD. The selected features were then applied to wide and deep neural network models, which maintained high accuracy despite the dimensionality reduction—highlighting the robustness of the approach.

Expanding beyond genetic data, a further methodology was applied using the multi-source dataset from the National Alzheimer's Coordinating Center was analysed, encompassing 45,923 participants, 1,023 variables, and 169,408 records across baseline and follow-up visits. Using the Boruta algorithm, a relevant subset of features was extracted, and among the tested classifiers, Random Forest achieved strong and balanced performance.

Finally, recognising that the "black-box" nature of ML models can limit clinical adoption, this work emphasises interpretability. Extended experiments uncovered meaningful patterns and risk factors for AD, with the Clinical Dementia Rating tool emerging as a particularly significant predictor. These findings not only strengthen the predictive framework but also provide clinically relevant insights into AD progression and risk profiling.

Acknowledgements

In the Name of Allah, the Most Beneficent, the Most Merciful

All praise and thanks are due to Allah for granting me the strength, patience, and guidance to embark on and complete this PhD journey. Without His countless blessings, none of this would have been possible.

I extend my deepest gratitude to my supervisors—Prof. Dhiya Al-Jumeily, Prof. Abir Hussain, Dr. Waisq Khan, Dr. Yasser Abdullah, and Dr. Fariba Sharifian—for their invaluable guidance, encouragement, and unwavering support throughout this research. Your insights, expertise, and constructive feedback were pivotal in shaping this work. I am truly fortunate to have had your mentorship, and I owe much of this accomplishment to your belief in me.

I would also like to express my heartfelt appreciation to all my teachers and mentors throughout my life, from my earliest education to my postgraduate studies. Your dedication and passion for teaching have inspired me and shaped the path I have taken.

Special thanks to the staff at the Iraqi Embassy in London, particularly the Cultural Attaché, for their assistance throughout this journey. I also extend my gratitude to the staff at the Postgraduate Research Administration at the Faculty of Engineering and Technology, Liverpool John Moores University, for their continuous support.

To my friends and colleagues, thank you for your camaraderie, motivation, and thoughtful discussions. Your support has been a source of comfort and encouragement during challenging times. A special mention goes to Dr. Muhammed Al-Asadi and his family, Dr. Miran Al-Rammahi, Dr. Ziad Ali, and Dr. Omar Aldhaibani for their invaluable support.

To my family—words cannot fully express my gratitude. To my parents, Prof. Saad Alatrany and Mrs. Azhar Alatrany, who instilled in me the values of education and hard work, your sacrifices and prayers have been the foundation of all my achievements. To my siblings, Ali, Dhuha, and Mujtaba, thank you for your unwavering love and encouragement.

To my wife, Narjis Al-Musawi, your patience, understanding, and endless encouragement have been my greatest source of strength. To my wonderful son, Laith, you are my light and inspiration.

Dedication

To **Iraq**, my homeland, the cradle of civilisation, where the echoes of history resonate through the valleys of Mesopotamia and the mighty rivers of the Tigris and Euphrates.

This work is dedicated to the resilient spirit of its people, who endure with courage, grace, and unwavering hope despite immense challenges.

May this thesis be a small contribution to the pursuit of knowledge and progress, inspired by Iraq's rich heritage and the dreams of its future generations.

For **Iraq**, with love and eternal gratitude.

Datasets Acknowledgements

Alzheimer's Disease Neuroimaging Initiative Dataset

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 Pharmac euticals, 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. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

The National Alzheimer's Coordinating Center Dataset

The NACC database is funded by NIA/NIH Grant U24 AG072122. NACC data are contributed by the NIA-funded ADRCs: P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez, MD), P30 AG062421 (PI Bradley Hyman, MD, PhD), P30 AG066509 (PI Thomas Grabowski, MD), P30 AG066514 (PI Mary Sano, PhD), P30 AG066530 (PI Helena Chui, MD), P30 AG066507 (PI Marilyn Albert, PhD), P30 AG066444 (PI John Morris, MD), P30 AG066518 (PI Jeffrey Kaye, MD), P30 AG066512 (PI Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI

Andrew Saykin, PsyD), P30 AG072975 (PI David Bennett, MD), P30 AG072978 (PI Neil Kowall, MD), P30 AG072977 (PI Robert Vassar, PhD), P30 AG066519 (PI Frank LaFerla, PhD), P30 AG062677 (PI Ronald Petersen, MD, PhD), P30 AG079280 (PI Eric Reiman, MD), P30 AG062422 (PI Gil Rabinovici, MD), P30 AG066511 (PI Allan Levey, MD, PhD), P30 AG072946 (PI Linda Van Eldik, PhD), P30 AG062715 (PI Sanjay Asthana, MD, FRCP), P30 AG072973 (PI Russell Swerdlow, MD), P30 AG066506 (PI Todd Golde, MD, PhD), P30 AG066508 (PI Stephen Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, PhD), P30 AG072931 (PI Henry Paulson, MD, PhD), P30 AG066546 (PI Sudha Seshadri, MD), P20 AG068024 (PI Erik Roberson, MD, PhD), P20 AG068053 (PI Justin Miller, PhD), P20 AG068077 (PI Gary Rosenberg, MD), P20 AG068082 (PI Angela Jefferson, PhD), P30 AG072958 (PI Heather Whitson, MD), P30 AG072959 (PI James Leverenz, MD). The NACC database is funded by NIA/NIH Grant U24 AG072122. NACC data are contributed by the NIA-funded ADRCs: P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez, MD), P30 AG062421 (PI Bradley Hyman, MD, PhD), P30 AG066509 (PI Thomas Grabowski, MD), P30 AG066514 (PI Mary Sano, PhD), P30 AG066530 (PI Helena Chui, MD), P30 AG066507 (PI Marilyn Albert, PhD), P30 AG066444 (PI John Morris, MD), P30 AG066518 (PI Jeffrey Kaye, MD), P30 AG066512 (PI Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI Andrew Saykin, PsyD), P30 AG072975 (PI David Bennett, MD), P30 AG072978 (PI Neil Kowall, MD), P30 AG072977 (PI Robert Vassar, PhD), P30 AG066519 (PI Frank LaFerla, PhD), P30 AG062677 (PI Ronald Petersen, MD, PhD), P30 AG079280 (PI Eric Reiman, MD), P30 AG062422 (PI Gil Rabinovici, MD), P30 AG066511 (PI Allan Levey, MD, PhD), P30 AG072946 (PI Linda Van Eldik, PhD), P30 AG062715 (PI Sanjay Asthana, MD, FRCP), P30 AG072973 (PI Russell Swerdlow, MD), P30 AG066506 (PI Todd Golde, MD, PhD), P30 AG066508 (PI Stephen Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, PhD), P30 AG072931 (PI Henry Paulson, MD, PhD), P30 AG066546 (PI Sudha Seshadri, MD), P20 AG068024 (PI Erik Roberson, MD, PhD), P20 AG068053 (PI Justin Miller, PhD), P20 AG068077 (PI Gary Rosenberg, MD), P20 AG068082 (PI Angela Jefferson, PhD), P30 AG072958 (PI Heather Whitson, MD), P30 AG072959 (PI James Leverenz, MD).

Contents

List of Figures

List of Tables

Publications

Journal Articles

Alatrany, A.S., Khan, W., Hussain, A. et al. An explainable machine learning approach for Alzheimer's disease classification. Sci Rep 14, 2637 (2024). https://doi.org/10.1038/s41598-024-51985-w.

Ogden, Ruth, et al. "Distortions to the passage of time for annual events: Exploring why Christmas and Ramadan feel like they come around more quickly each year." Plos one 19.7 (2024): e0304660.

S. Alatrany, W. Khan, A. J. Hussain, J. Mustafina, and D. Al-Jumeily, "Transfer Learning for Classification of Alzheimer's Disease Based on Genome Wide Data," (in eng), IEEE/ACM Trans Comput Biol Bioinform, vol. Pp, Jan 3 2023, doi: 10.1109/tcbb.2022.3233869.

Ansari, S., **Alatrany, A. S.**, Alnajjar, K. A., Khater, T., Mahmoud, S., Al-Jumeily, D., Hussain, A. J. (2023). A survey of artificial intelligence approaches in blind source separation. Neurocomputing, 561, 126895.

Alatrany AS, Khan W, Hussain A, Al-Jumeily D, for the Alzheimer's Disease Neuroimaging Initiative (2023) Wide and deep learning based approaches for classification of Alzheimer's disease using genome-wide association studies. PLoS ONE 18(5): e0283712. https://doi.org/10.1371/journal.pone.0283712

S. Alatrany, A. J. Hussain, J. Mustafina and D. Al-Jumeily, "Machine Learning Approaches and Applications in Genome Wide Association Study for Alzheimer's Disease: A Systematic Review," in IEEE Access, vol. 10, pp. 62831-62847, 2022, doi: 10.1109/ACCESS.2022.3182543.

Conference Articles

Alatrany, A. Hussain, J. Mustafina, and D. Al-Jumeily, "A Novel Hybrid Machine Learning Approach Using Deep Learning for the Prediction of Alzheimer Disease Using Genome Data," in International Conference on Intelligent Computing, 2021: Springer, pp. 253-266.

A. Alatrany, A. Hussain, M. Jamila and D. Al-Jumeiy, "Stacked Machine Learning Model for Predicting Alzheimer's Disease Based on Genetic Data," 2021 14th International Conference on Developments in eSystems Engineering (DeSE), 2021, pp. 594-598, doi: 10.1109/DeSE54285.2021.9719449.

Alatrany, A.S., Hussain, A., Alatrany, S.S.J., Al-Jumaily, D. (2022). Application of Deep Learning Autoencoders as Features Extractor of Diabetic Foot Ulcer Images. In: Huang, DS., Jo, KH., Jing, J., Premaratne, P., Bevilacqua, V., Hussain, A. (eds) Intelligent Computing Methodologies. ICIC 2022. Lecture Notes in Computer Science(), vol 13395. Springer, Cham. https://doi.org/10.1007/978-3- 031-13832-$4_1$1

A. S. Alatrany, A. Hussain, S. S. J. Alatrany, J. Mustafina, and D. Al-Jumeily, "Comparison of Machine Learning Algorithms for classification of Late Onset Alzheimer's disease," in 2023 15th International Conference on Developments in eSystems Engineering (DeSE), 9-12 Jan. 2023 2023, pp. 60-64, doi: 10.1109/DeSE58274.2023.10099655.

Abbreviations

AD	**A**lzheimer's **D**isease
EOAD	**E**arly **O**nset **AD**
LOAD	**L**ate **O**nset **AD**
AI	**A**rtificial **I**ntelligence
ML	**M**achine **L**earning
NC	**N**ormal **C**ontrol
SVM	**S**upport **V**ector **M**achine
ANN	**A**rtificial **N**eural **N**etwork
DL	**D**eep **L**earning
TL	**T**ransfer **L**earning
GWAS	**G**enome **W**ide **A**ssociation **S**tudy
SNP	**S**ingle **N**ucleotide **P**olymorphism
MCI	**M**ild **C**ognitive **I**mpairment
CNS	**C**entral **N**ervous **S**ystem
DNA	**D**eoxyribonucleic **A**cid
RF	**D**Random **F**orest
CV	**C**ross **V**alidation
BSWiMS	**B**ootstrap **S**tage- **Wi**se **M**odel**S**election
LASSO	**L**east **A**bsolute **S**hrinkage and **S**election **O**perator
RPART	**R**ecursive **P**artitioning **A**nd **R**egression **t**rees
AUC	**A**rea **U**nder the **C**urve
AUC-ROC	**A**rea **U**nder a **R**eceiver **O**perating **C**haracteristic **C**urve
DLG	**D**eep-**L**earning **G**enomics
MRI	**M**agnetic **R**esonance **I**maging
RNN	**R**ecurrent **N**eural **N**etwork

KNN	**K**-**N**earest **N**eighbour Algorithm
PCA	**P**rincipal **C**omponent **A**nalysis
PC	**P**rincipal **C**omponent
CAR	**C**lass **A**ssociation **R**ules
SIRUS	**S**table and **I**nterpret **RUL**e **S**et
TP	**T**rue **P**ositive
FP	**F**alse **P**ositive
FN	**F**alse **N**egative
PRS	**P**olygenic **R**isk **S**core
ADNI	**A**lzheimer's **D**isease **N**euroimaging **I**nitiative
QC	**Q**uality **C**ontrol
HWE	**H**ardy-**W**einberg **E**quilibrium
MAF	**M**inor **A**llele **F**requency

Chapter 1

Introduction

1.1 Introduction

Alzheimer's disease (AD) is the most prevalent kind of dementia, accounting for 80% cases of dementia [1]. It impairs memory, thinking, conduct, and overall capacity to do everyday tasks such as eating and bathing etc. The illness can generally be classified into two subcategories: early-onset AD (EOAD) and late-onset AD (LOAD) [2]. The EOAD is almost entirely a genetic disease with heritability ranging from 92% to 100% [3] where the affected first-degree relatives account for 35% to 60% of EOAD patients. Usually, the EOAD patients experience their first symptoms between the age of 30 and 65, with the majority of EOAD patients diagnosed between the ages of 45 and 60 years [4]. In contrast to EOAD, the LOAD affects elderly people (usually over 65 years of age) acounting 90-95% of the AD cases [5]. LOAD appears to be a more complicated illness induced by genetic as well as the environmental and lifestyle factors.

AD poses significant challenges to individuals and their families, but these challenges can be decreased with the use of medical systems. The development of medical information systems has been of paramount importance to medical societies all over the world. Such developments have been aimed at improving the utilisation of technology in medical applications. To this end, new and emerging technologies such as expert systems and various Artificial Intelligence (AI) methods and techniques have been employed and developed to enhance the decision support tools used in the medical field. The field of scientific research employs Machine Learning (ML) models as a strong powerful technique according to [6]. Through this technology computers learn from data to develop predictive models for medical diagnosis and prognosis. Medical field processes can become automated through these models which deliver both fast and precise diagnosis and treatment solutions. ML models enable the analysis of big data to detect patterns which leads to important discoveries about disease causes and mechanisms. These models assist medical professionals to make improved decisions through data analysis which leads to better patient care [7].

During the last ten years ML analytics have gained widespread use for AD through the application of Support Vector Machine (SVM), Artificial Neural Network (ANN) and Deep Learning (DL), a specialised subset of ANN that focuses on deeper architectures, being the most widely used classification techniques. The ability of DL to handle big datasets makes them ideal for feature extraction which leads to better learning model performance [8] [9], [10], [11], [12]. The accuracy of Alzheimer's classification improves

through ensemble methods which combine predictions from multiple models [13]. The method provides advantages through reduced overfitting risk which results in better outcomes. Researchers have investigated Transfer Learning (TL) [14] as a method to apply knowledge obtained from one task to another. The application of TL to Alzheimer's could provide advantages because it enables the utilisation of existing data from other diseases for Alzheimer's research. The ML techniques show promise to enhance our understanding of AD diagnosis which will lead to better care for affected patients.

1.2 Problem Statement

AD represents a degenerative neurological disorder which causes permanent damage to individuals and their families as well as healthcare organisations and economic systems across the world. Dementia affects 55 million people worldwide and disproportionately affects elderly individuals. The annual number of dementia diagnoses reaches 10 million according to available statistics. The disease remains incurable and no person can avoid developing AD [15].

The disease produces a slow deterioration of memory functions which eventually results in patients losing their ability to identify family members and remember things and follow basic instructions. The disease progression brings both physical destruction to patients and emotional suffering to their families because they lose their important memories and essential abilities.

The diagnosis of AD at an early stage combined with precise identification remains essential for initiating proper interventions which lead to better treatment results. ML techniques demonstrate promising capabilities for AD classification through the analysis of neuroimaging data, genomic information and clinical record information. The development of effective ML models for AD classification faces substantial obstacles in the current research landscape. The development of ML models for AD classification faces three major obstacles which include managing large datasets and dealing with imbalanced classes and maintaining model interpretability.

The research addresses these obstacles through the development of new ML algorithms and methodologies which were aimed for AD classification. The study will focus on: Developing robust feature selection techniques to effectively leverage genetic and multisource data. Developing new methods to achieve precise cognitive state classification

with both high accuracy and balanced performance. The research applies interpretable ML models to achieve transparent decision-making while delivering meaningful insights about the decision processes. The research aims to improve current AD classification methods using ML techniques which will lead to earlier diagnosis with better accuracy. By integrating ML tools into clinical practice, this study seeks to enhance patient care and contributing to the global fight against AD.

Research Questions:
Can ML algorithms accurately classify individuals with AD from Normal Controls (NC) using genetic markers?
Can AI algorithms be developed to identify novel biomarkers in AD data that have not been previously identified by traditional statistical methods?
How can TL and domain adaptation be employed to improve the generalisation and robustness of AI models for AD classification across different populations and datasets?
What are the most effective neural network architectures for classification of AD using GWAS data?
Can AI be used to differentiate between different classes of individual's cognitive state?
Can AI predict individual's cognitive state in the future?
How can interpretable AI techniques be applied to AD data analysis to enhance the understanding and transparency of the decision-making process?

1.3 Aim and Objectives

This research project focus on developing ML-based approaches for classifying and predicting AD using genetic markers and multi-source data. The proposed approach is designed to continuously evolve and improve its analysis of AD risk factors, as well as accurately predict the onset of the disease at an early stage. This research aims to enhance the knowledge of AD risk factors and assist clinicians in making informed decisions. Additionally, the proposed approach is intended to be cost-effective, enabling early detection and management of AD. The objectives are as follows:

1. Conduct a comprehensive review and identify gaps in the literature.
2. Identify and access appropriate open-source datasets from AD institutions.
3. Determine best practices for Quality Control (QC) in order to eliminate bias and

inaccurate data.

4. Solve the issue of dimensionality reduction by using appropriate feature selection approach(s) and address the data size challenges associated with the GWAS.

5. Develop effective and reliable ML approaches for classification and prediction of AD.

6. Extract patterns in data to serve for the ML model interpretability and highlight the most important factors related to AD.

1.4 Contributions to Knowledge

This research project presents new approaches for analysing AD datasets to enable the classification of cases and controls. The approaches presented provide a reliable data pipeline for pre-processing multi sources and genetic data, selecting relevant features, and utilising ML algorithms for data modeling. As a result, the research discusses various contributions:

Major Contributions

- **Robust Feature Selection for SNPs in AD Classification** Developed a hybrid dimensionality reduction and feature selection approach to identify the most promising Single Nucleotide Polymorphisms (SNPs) for AD classification. Enhanced model performance and robustness through the selection of distinguishing genetic features. (Related to Objectives 3 and 4)

- **New Methodology for Multiclass Classification of AD** Proposed a novel methodology to classify individuals into NC, Mild Cognitive Impairment (MCI), or AD. Achieved high and balanced accuracy using a small number of features, demonstrating efficiency and potential for real-world diagnostic use. The method outperforms existing approaches and is generalisable to other chronic diseases. (Related to Objective 5)

- **Application of Deep Transfer Learning in GWAS** Pioneered the use of deep TL to address data size challenges in GWAS for AD. Explored and compared various TL techniques for improved model generalisation and accuracy. (Related to Objective 5)

Minor Contributions

- **Systematic Review on ML in GWAS for AD** Conducted a comprehensive systematic review on the application of ML algorithms in the analysis and interpretation of GWAS data for AD. Covered topics including supervised and unsupervised learning, deep learning techniques, evaluation metrics, and comparison of existing approaches. (Related to Objectives 1 and 2)
- **Neighbour SNP Selection Approach** Proposed an approach to evaluate the impact of neighbouring SNPs on classification accuracy in GWAS-based models. (Related to Objective 4)
- **Extraction of Human-Readable Rules from ML Models** Enhanced model interpretability by extracting human-readable rules from AD data. Facilitated understanding of significant contributing factors and underlying data patterns. (Related to Objective 6)

1.5 Structure of the Thesis

The reminder of the of this thesis organised as follows:
Chapter 2 - Literature Review: This chapter discusses what AD is and the common types, as well as the treatments currently used to mitigate the severity of the disease. It provides information on the risk factors associated with AD and outlines diagnosis strategies. Chapter 2 delves further into the literature review related to ML and current algorithms used to analyse AD datasets.
Chapter 3 – Machine Leaning Overview: This chapter provides an in-depth exploration of the various ML models, learning algorithms, and classification techniques that can be utilised to solve a variety of problems in the area of AD. It will discuss the different types of models, algorithms, and techniques available. Additionally, it will discuss the various types of feature extraction techniques that can help with the problem of curse of dimensionality. Finally, it will provide an overview of the approaches and techniques that can be used for rule mining and extracting human readable rules from the data.
Chapter 4 – Deep TL in GWAS: In this chapter, for the first time in literature, TL was used to build a classification models for AD and NC from GWAS data. After QC

steps and feature selection, a pre-trained model of GWAS data on human and animals is transferred to be used on another human GWAS dataset. This is conducted to solve the data size associated with GWAS (extremely high numbers of features and only small sample size).

Chapter 5 – Wide and deep learning approaches in GWAS: in this chapter, different neural network architectures have been utilised, wide (single hidden layer with high number of neurons) and deep (multiple hidden layers with small number of neurons in each layer) learning are used in order to increase the classification accuracy of AD and test which architecture can better classify AD through different experiments.

Chapter 6 - ML for AD classification: In this chapter, access to the National Alzheimer's Coordinating Center dataset has been granted to build reliable and high accuracy ML models for AD classification and prediction state of cognitive state of a person four years ahead. The methodology employed in the chapter consist of several stages starting from data cleaning to feature selection and classifiers constructing. In addition, rule mining techniques are employed to extract human-readable rules to understand factors that influence the risk of AD.

Chapter 7 - Conclusion and future work: The conclusion section of the research presents the overall findings of the study and discusses its outcomes highlighted in this chapter. Future work is also discussed that can be done to improve the research domain.

1.6 Chapter Summary

The chapter highlights the growing prevalence of AD and the challenges faced in its early detection and diagnosis. It emphasises the need for advanced AI techniques to analyse large and complex datasets associated with AD, such as multi sources and genetic data. The research objectives are then presented, outlining the specific goals and inquiries of the study. These objectives include developing novel algorithms for feature selection, data preprocessing, and utilising ML algorithms to model AD data accurately.

Lastly, the chapter concludes by providing a brief outline of the subsequent chapters in the thesis. This roadmap highlights the structure and content of the thesis, indicating how the research will unfold, and what readers can expect from the subsequent chapters.

Chapter 2

Literature Review

Parts of the research defined in this chapter has been published in IEEE Access.

S. Alatrany, A. J. Hussain, J. Mustafina and D. Al-Jumeily, "Machine Learning Approaches andApplications in Genome Wide Association Study for Alzheimer's Disease: A Systematic Review," in IEEE Access, vol. 10, pp. 62831-62847, 2022, doi: 10.1109/ACCESS.2022.3182543. [16]

2.1 Introduction

Elderly people affected by AD are experiencing a progressive decline in cognitive abilities, slowly losing their memories, independence, and ability to understand their surroundings. This challenging experience is endured by the individuals affected, as well as their caregivers and families. Therefore, it is crucial to conduct extensive research into this condition and its associated risk factors, symptoms, and diagnostic methods, to provide the best possible care and support to those affected.

In this chapter, an in-depth background research is conducted into AD, exploring the pathological development of the illness, the types of risk factors, the symptoms of the condition, and the latest diagnosis methods. A comprehensive overview of the use of ML in the classification and prediction of AD was also provided.

2.2 Medical Diagnosis

A medical diagnosis is the process of determining which condition or disease is responsible for a person's symptoms. This typically requires gathering information from the patient's medical history and conducting a physical examination. During the diagnosis process, one or more diagnostic procedures, such as medical tests, may also be performed to aid in identifying the underlying cause of the symptoms.

Diagnosing a medical condition can be challenging due to the nonspecific nature of many symptoms. For instance, a reddened skin symptom (erythema) can indicate many different conditions, making it impossible for a healthcare practitioner to determine the underlying problem by looking at it alone. Therefore, differential diagnosis, which involves comparing and contrasting different possible explanations, is necessary.

This process involves identifying all possible diseases or conditions that could cause the signs or symptoms and then eliminating or at least reducing the likelihood of each entry through further medical tests and other processes. The aim is to reach a point where only one condition or disease remains probable after all other possibilities have been ruled out or deemed less likely. [17].

No single test can be used to diagnose AD disease. Instead, physicians assess a combination of medical history, symptoms, physical exams and tests [21]. By reviewing the medical history and physical exam, other conditions causing similar symptoms may be

eliminated.

Currently, diagnosing AD in its earliest stages is a complex process. The diagnosis of Alzheimer's requires a comprehensive medical assessment which includes an evaluation of medical records, mental health testing, physical and neurological exams, and imaging tests such as brain scans. The difficulty of diagnosing this condition depends on which stage it is found; recognising it in its early stages is far more challenging.

The initial step in diagnosing AD is recognising its signs and symptoms, which become more noticeable as the condition progresses. If the symptoms are obvious, patients will then be encouraged to do some tests. In cases where the indications are not straightforward or peculiar, doctors may suggest that a brain imaging test such as magnetic resonance imaging is done to affirm the existence of AD and distinguish it from other kinds of dementia or neurological issues [18].

2.3 Overview of Neuron Structure

A thorough examination of this fatal brain condition requires first studying the normal functioning of the human brain and the structure of neurons. Human brain operations depend on complex electrical and chemical systems which control all bodily activities. Neurons function as electrically excitable cells which conduct information by transmitting electrical and chemical signals. The brain contains vast neural networks which make up a substantial portion of the nervous system [19].

The human nervous system contains two fundamental elements.

- The Central Nervous System (CNS) functions as the body's control centre and contains both the brain and spinal cord. The system functions to receive sensory information then process it while generating motor responses.
- The Peripheral Nervous System (PNS) includes every neuron together with neuron components that exist outside the CNS. The system performs two main functions which include sending sensory data to the CNS and controlling motor outputs to produce body responses [20].

A neuron functions as a distinct unit which accepts information from other neurons then processes the input and delivers the output to additional neurons. The human nervous

system consists of three neuronal types which serve different functions including sensory neurons motor neurons and interneurons. The nervous system obtains information through sensory neurons which operate as its primary information gathering agents. These cells obtain environmental information together with internal data which they deliver to the brain for processing. The motor neurons act as transmission agents because they deliver instructions from the brain and spinal cord to muscles and organs and glands for physical action. Interneurons serve as connectors within the nervous system. The neural network functions through interneurons which receive signals from certain neurons before transferring them to other neurons to maintain system communication. These three neuron types create a complex web of communication which allows the human body to detect and understand and generate reactions to environmental stimuli.

To gain a deeper understanding of this severe brain disorder, it is essential to first understand how the human brain operates and the structure of neurons. The human brain functions through intricate chemical and electrical processes that regulate bodily functions. Neurons, which are electrically excitable cells, communicate and transmit information through electrical and chemical impulses. These neurons form vast networks within the brain, which together constitute a significant part of the nervous system [19].

2.3.1 Anatomy of a Neuron

The main functions of neurons which consist of information reception and processing occur at the dendrites together with the cell body as illustrated in Figure 2.1. The signals that reach neurons have two functions: excitatory signals can activate the neuron to generate electrical impulses and inhibitory signals work to stop the neuron from producing electrical signals.

Most neurons possess many dendritic trees which receive various input signals and each neuron has multiple dendritic sets that process thousands of signals. All received signals determine how likely a neuron will produce an electrical impulse. The neuron fires an impulse when the total input value surpasses a particular threshold point.

A neuron depends on its axon to distribute electrical signals to other neurons. The electrical signals begin at the soma which serves as the main body of the neuron before moving through the axon. The axon extends beyond the soma into multiple branches which create connections that reach other neurons. It comprises A critical feature of the

neuron includes two main parts which are the axon hillock for electrical signal generation and the axon terminals that create synaptic connections with other neurons. The size of the human hair represents the typical scale of the axon thickness although the length and dimensions change based on the specific neuron type.

The formation of connections between neurons occurs through dendritic and cell body attachment which creates synaptic junctions. Synapses function as essential communication points which transfer information from the presynaptic neuron to the postsynaptic neuron [21].

Most synapses use chemical messengers known as neurotransmitters to transfer information between neurons. The termination of electrical signals in an axon leads to neurotransmitter molecule release from the presynaptic cell. The neurotransmitter molecules pass through the synaptic cleft space to bind with receptors on the postsynaptic cell membrane thus sending inhibitory or excitatory signals [22].

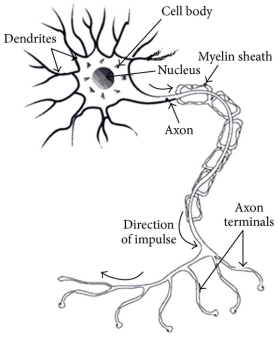


FIGURE 2.1: Structure of biological neuron. Taken from [23]

The human body contains various types of neurons responsible for relaying information. In the brain, neurons play a key role in storing and communicating information. Diseases like dementia damage these neurons, disrupting communication within neural networks. AD is a progressive and fatal illness that develops gradually. It begins with the formation of two abnormal protein fragments called plaques and tangles (see Figure 2.2). These proteins build up in the brain and damage its cells. In the early stages of AD, clusters of protein fragments (plaques) form between nerve cells, surrounding and damaging healthy brain cells. Over time, these plaques lead to the formation of twisted strands of another protein, known as tangles [24].

Although there is no conclusive evidence that plaques and tangles are the main cause of neuron death, they are considered the leading suspects in current research [25]. Plaques

are made up of beta-amyloid, small protein fragments that clump together, often in the fatty coating surrounding nerve cells [26].

Plaques and tangles first appear in the hippocampus, the brain region responsible for forming memories [27]. As they accumulate, they damage and kill cells in this area, making it harder for individuals with AD to form new short-term memories, such as recalling events from a few hours or days ago.

As the disease progresses, plaques and tangles spread to other areas of the brain, causing further neuron death. This progression explains the stages of AD. Since the hippocampus is affected early, people with AD typically struggle with short-term memory loss. Later, language skills are impaired, making it difficult to speak or form sentences [28].

Over time, the disease affects the part of the brain responsible for logical thinking, making it hard for individuals to plan activities or solve problems. Eventually, it spreads to areas that control emotions, leading to mood swings and anxiety [29].

When plaques and tangles reach the emotional centres of the brain, patients experience ongoing mood changes. As the damage continues, these proteins affect the sensory areas, which are responsible for understanding the environment. At this stage, individuals may have trouble recognizing their surroundings and can experience delusions.

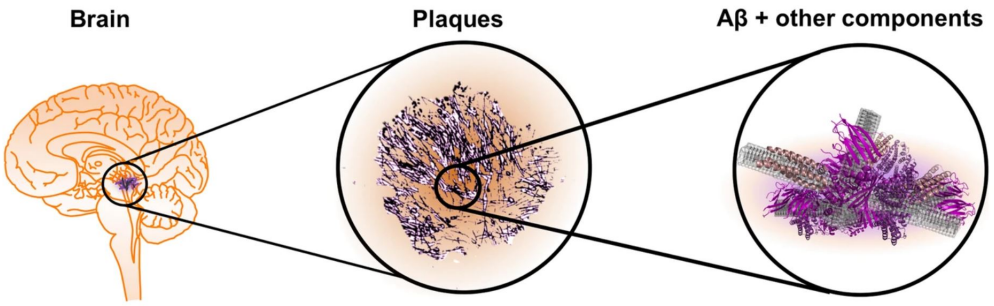


FIGURE 2.2: Pathological hallmarks of AD brains. Taken from [30]

2.4 Risk factors for Alzheimer's Disease

Age is found to be one of the well-known risk factors for AD. It is mostly seen in those over 65 years old, and it is rare for younger individuals to develop the disease. The aging process can cause multiple organs and cell systems to be affected as well as decrease in brain volume and weight. Thus it is difficult to distinguish from normal aging process when it comes to early AD detection [31].

Research has revealed that genetics are a significant contributor to the onset of AD. Of all AD cases, 70% have been linked to genetic influences. Furthermore, most cases of EOAD appear to be inherited [32].

Environmental factors like air pollution, diet, and infections can cause oxidative stress and inflammation, thus increasing the likelihood of AD [33]. Additionally, older people with AD typically suffer from medical issues such as obesity and diabetes; all of which are linked to a higher risk of AD [34].

2.5 Genome Wide Associations Study

GWAS is a powerful method for uncovering the genetic basis of complex diseases which have traditionally been difficult to study. By using genome-wide data, GWAS is able to detect and identify risk factors for diseases such as AD which would have otherwise been difficult to detect. By examining the genetic architecture of a disease, researchers can better understand its underlying mechanisms, which could in turn lead to the development of new treatments and therapies. This kind of analysis is particularly important for understanding the genetic basis of diseases, as it can help us to identify which genetic variants are associated with the phenotype and how strongly they are associated. By understanding the relationship between genetic variants and disease, it can better predict an individual's likelihood of developing the disease [35].

Recent advances in Deoxyribonucleic Acid (DNA) sequencing technology have made it possible to sequence the entire human genome within a single day. This has been made possible by the development of next-generation sequencing technology, which is more cost-effective and rapid than previous methods such as Sanger sequencing [36].

GWAS have had a significant impact on the field of human genetics, but there are still some challenges associated with computational and statistical methods that can make conducting such an analysis difficult. One of the main challenges is the scalability of the dataset [37]; GWAS datasets can contain hundreds of thousands to millions of SNPs with hundreds of individuals. This means that the algorithms used for GWAS need to be extremely scalable in order to avoid using huge amounts of computational resources and reducing the time it takes to conduct GWAS.

In addition, GWAS is successful in detecting single SNPs associated with a phenotype under study. Nevertheless, GWAS could not be able to detects multi-SNPs interactions

in chronical diseases such as AD. This is due to the complexity of the disease, multiple SNPs could together influence the initialise the disease instead of a single SNP. Each of these SNPs might have only a weak relation to the disease making it difficult to be detected alone. Therefore, multi-SNPs analysis needs to be utilised [35].

2.6 Machine learning in Alzheimer Disease

The application of ML can help early detection and diagnosis of AD. ML algorithms may be able to identify individuals at risk of developing AD before the onset of significant symptoms by analysing patterns in brain scans or cognitive test scores [38] [39] [40] [41] [42] This could allow for earlier intervention, potentially slowing disease progression.

2.6.1 Machine learning in GWAS for Alzheimer's Disease

There are several related work addressing the use of ML in GWAS data for AD. Araujo et al. [43] suggested to use biologically motivated SNP selection as an input to Random Forest (RF) for predicting patient risk of developing AD. The findings reveal that non-disease-related SNPs perform similarly to or better than disease-related SNPs. As the identification of novel relevant markers is the most important effort in GWAS. These findings suggest that SNPs from unrelated sets might be new candidates for AD [43]. Similarly, authors in [44] proposed an approach to find SNPs linked with AD in a GWAS data collection of 550 controls and 861 cases. The authors employed single-locus analysis to filter the data depending on a p-value threshold, resulting in a subset of SNPs that were used by RF to perform a multi-locus analysis. The single-locus analysis yielded 199 SNPs. Using 10-fold Cross Validation (CV) in RF modelling, these SNPs, together with other SNPs that associated to AD, produced a predictive subgroup for AD prediction with an average error of 9.8% [44].

In addition, RF was also included in a methodology proposed by [45]. The method begins by performing a p-value assessment to identify a cut-off point that separates the SNPs into two separate groups: those that are informative and those that are irrelevant. The informative SNPs group is further broken down into two sub-groups: highly informative SNPs and weak informative SNPs. These two sub-groups are the only SNPs considered when sampling the SNP subspace for constructing the trees of the forest.

Whenever a node is split at a tree, the feature subspaces always contain highly informative SNPs. This ensures that the trees of the forest are built with the most informative SNPs, providing the forest with a solid foundation for making accurate predictions. The approach also helps to reduce the dimensionality of the data, meaning that the model can be trained with less data and still achieve good results [45].

In a similar manner, SVM classifiers of various kernels were applied to GWAS data based on chosen 21 variations most related with AD using two techniques, Correlation-based and Chi-squared. The authors indicated that the results for SVM trained model utilising an RBF kernel reached the highest accuracy of 76.70 percent [46].

Chang et al. [47] introduced GenEpi, a computational package that uses L1-regularized regression to find epistasis related with phenotypes. GenEpi uses a two-stage modelling methodology to identify both within-gene and cross-gene epistasis. The ML model was trained and evaluated on genetic data of 364 individuals. A total of 24 SNPs from 12 genes were used in the final model. This model has achieved a 2-fold CV accuracy of 0.83 and a leave-one-out CV accuracy of 0.83 [47].

FRESA.CAD's (Feature Selection Algorithms for Computer Aided Diagnosis) benchmarking tool was employed to forecast a person's risk of acquiring AD by contrasting and evaluating several ML models, including Bootstrap Stage-Wise Model Selection (BSWiMS), Least Absolute Shrinkage and Selection Operator (LASSO), and Recursive Partitioning and Regression Trees (RPART). The Area Under a Receiver Operating Characteristic Curve (AUC-ROC) varied between 0.60 and 0.70. The BSWiMS, LASSO, and RPART performed similarly. However, the authors indicator that an ensemble model of all approaches performed the best, with a AUC-ROC score of 0.719 [48].

Romero-Rosales et al. present a comparison of three ML models: genetic algorithm, stepwise, as well as LASSO approach for developing models for AD classification trained on data of 813 cases and 1,017 controls. Their initial results conclude that LASSO outperformed the other two methods. Their hypothesis is to use the markers of the incorrectly classified samples to train the model, so that it can better identify and classify similar samples in the future. By doing so, the authors found that the accuracy of LASSO improved by around 5%, reaching 0.84 Area Under the Curve (AUC) [49].

Aflakparast et al. [50] present a novel strategy, cuckoo search epistasis, for detecting epistatic interactions in case–control studies. This technique combines a Bayesian scoring function with a heuristic search algorithm. The algorithm was able to find SNPs that were reported to be associated with AD according to the literature [50].

Stokes et al. assess the efficacy of label propagation, a multivariate graph-based approach for effectively ranking SNPs in GWAS. The top-ranked SNPs were evaluated in terms of classification performance, and prior evidence of being linked with AD. label propagation performed much better in categorisation than other control approaches. There were 14 SNPs in one dataset among the 25 top-ranked SNPs found by label propagation that had evidence in the literature of being linked with AD [51].

Li et al. present a novel Deep-Learning Genomics (DLG) model and apply it to the multitasking categorisation of AD progression. For the DLG model, the ResNet framework was employed using 1461 patients' genotyping data including 366 NC, 473 MCI and 622 Alzheimer's cases. The results of the DLG model were compared to those obtained using a basic Convolutional Neural Network (CNN) structure. When applied to the course of AD, the authors claimed that DLG model can obtain improved accuracy and sensitivity [52].

Moore et al. present Crush, a stochastic search technique to explore relations between genes in genome-wide data as an application of multifactor dimensionality reduction. Applying the approach to an AD GWAS dataset, results showed that Crush multifactor dimensionality reduction was capable of identifying a collection of interacting genes with biological linkages to AD [53].

By leveraging tree-based ML methods and a set of 145 SNPs associated with AD that were previously documented in DisGeNET which is an invaluable platform that offers comprehensive data on human genes and their associated variants that are related to diseases. The authors indicated that the ML models were able to accurately classify cases of LOAD and healthy control subjects with high performance. The model achieved an accuracy of 0.80 and an AUC-ROC of 0.91 when using gradient boosting algorithm [54].

In reference [55], the authors accessed GWAS data of 431 participants (divided into 304 AD cases and 127 NC) from ADNI with the intention to find non-linear SNPs epistasis interactions. After data pre-processing 447,538 SNPs retrieved for subsequent analysis. To reduce the number of SNPs, they applied three association test methods, to test the association of each SNP with AD. A p-value of 0.01 used as a threshold for signification SNPs. 3,502 significant SNPs resulted from intersection of the three association tests, this number then was further reduced to 1050 SNPs using TuRF algorithm. An ensemble learning approach includes Classification and Regression Trees, extreme gradient boosting and RF is utilised, by integrating the top 20 ranking SNPs from each method to identify key SNPs that are involved in non-linear epistasis interactions in GWAS of

AD. Multifactor dimensionality reduction used to find from 2-way up to 5-way SNPs interaction from the identified SNPs. The accuracy of 5-way models varied between 0.8674 and 0.8758. Their proposed framework for identifying disease-causing genes can identify high-risk genes and epistasis interactions.

TABLE 2.1: Summary table of the methods used in various studies applying machine learning to GWAS data for Alzheimer's Disease.

Study	ML Model	Dataset Size	Feature Selection	Novelty	Advantages	Disadvantages
[43]	Random Forest	205/169	Used set of known genes from AlzGene database	Biologically motivated SNP selection	Novel markers from unrelated SNPs	Selected SNPs questioned
[44]	Random Forest	550/861	Logistic regression	P-value filtering, 199 SNPs, 10-fold CV	9.8% prediction error	Loss of potential SNPs during filtering
[45]	Random Forest	188/176	SNPs of top 10 AD candidate genes listed on the AlzGene	Informative SNP selection via p-values	Reduces dimensionality, high accuracy	Relies on p-value cutoff
[46]	SVM	241/132	Wilcoxon test	21 SNPs via correlation/chi-squared	Accuracy of 76.7%	May miss complex SNP signals
[47]	L1-regularized regression	230/241	2 test	Two-stage epistasis detection	Accuracy of 0.83 (CV and LOOCV)	Computational complexity
[48]	BSWiMS, LASSO, RPART	1017/813	P-value using summary statics	FRESA.CAD benchmarking	AUC-ROC of 0.719	Models alone AUC-ROC ranged 0.60–0.70
[49]	LASSO, Genetic Algorithm	1017/813	2 test	Retrain with misclassified samples	LASSO AUC improved to 0.84	Possible retraining bias
[50]	Cuckoo search Bayesian scoring	550/861	2 test	Epistasis detection in GWAS	Validated SNPs via literature	Computational complexity
[51]	Label propagation	938/1291	Chi-square test	SNP graph-based ranking	High performance	High computational cost
[52]	ResNet-based DLG	622/366	Chi-square test	Multi-task AD progression model	Improved accuracy/ sensitivity over CNN	Needs large data, model tuning
[53]	Crush (Stochastic + MDR)	1461		Gene interaction detection	Identified biologically linked genes	Stochastic variance
[54]	Gradient boosted decision trees	75,000/738	SNPs from DisGeNET	Tree-based model with known SNPs	Accuracy 0.80, AUC-ROC 0.91	Relies on known SNP databases
[55]	Ensemble model	127/304	TuRF	5-way SNP interactions	Accuracy up to 0.8758	Complex pipeline, resource-intensive

The summary table from the literature review shows how different studies have addressed the difficulties of creating reliable ML models from GWAS data for AD. The majority of these studies have one main weakness which is their dependence on small datasets. The models may produce biased results because of inadequate training data which restricts their ability to generalise and perform well. GWAS research has extensively documented this issue because obtaining large datasets needs both substantial financial support and expert participation.

The majority of studies included validation methods to reduce bias while improving model accuracy in their research. The most popular validation method used was cross-validation but the authors of [43] chose out-of-bag error and [53] did not implement any validation strategy. The majority of studies maintained balanced case-control ratios to address class imbalance issues. The study based on UK Biobank data [54] demonstrated a major class imbalance issue because it contained many more control samples than Alzheimer's disease cases. Different studies applied different methods to address class imbalance and dataset bias which resulted in inconsistent and sometimes inadequate methodological approaches.

The analysis of validation methods used in reviewed studies shows internal validation through cross-validation was common yet external validation with independent datasets was rarely found. The majority of studies used internal validation techniques which included 10-fold, 2-fold and Leave-One-Out Cross-Validation (LOOCV) to assess model performance. Study [44] used 10-fold CV to validate a Random Forest model which reduced overfitting and produced more dependable performance metrics. The research by Chang et al. [47] used both 2-fold CV and LOOCV to achieve a 0.83 accuracy score which demonstrated strong internal model stability. Despite these advantages, there were notable limitations. The majority of research failed to present essential statistical measures which included accuracy score variance and standard deviation across different folds. The absence of these critical metrics prevents complete evaluation of model stability together with performance variability which are essential for assessing ML model robustness. The main weakness of these studies was their failure to conduct external validation through testing models on separate datasets. The studies failed to validate their results through external datasets which makes it difficult to determine how well the models would perform with different populations or real-world clinical data.

The studies from the literature also shows that robustness and data reliability receive inadequate attention despite being essential elements for machine learning applications

particularly those working with complex high-dimensional data like GWAS. The majority of studies do not explain their methods for handling missing data which remains a frequent and critical issue in genomic datasets. The studies [43], [44], [45], and [55] do not provide information about missing value treatment which makes readers uncertain about the management of potential data gaps. The studies [47], [49] and [50] presented more transparent approaches by implementing reference-based imputation and median substitution methods for data handling. The studies lack standardisation because researchers did not establish clear reporting practices. The process of hyperparameter tuning which is essential for building robust models remains poorly documented. The majority of studies either selected parameters manually or omitted description of their process while [49] and [51] used cross-validation or empirical testing for model tuning. The failure to document this process restricts both the interpretability and replicability of the results from machine learning models that work with small sample sizes and extensive feature sets.

2.6.2 Challenges of Using ML in GWAS

Machine learning models encounter multiple issues with applying GWAS data because of its inherent properties. The main challenge in GWAS datasets arises from their high dimensionality because they contain between hundreds of thousands and millions of SNPs while the sample number stays relatively small. Models become vulnerable to overfitting because of the large feature space relative to sample size which creates the "curse of dimensionality" problem when they select noise instead of meaningful biological signals. Model performance becomes misleading when dimensionality reduction and feature selection techniques are not applied because of which models fail to generalise.

The small sample sizes found in multiple studies from the review act to worsen this situation. The scarcity of available samples reduces statistical power for models and produces false positive results. The genetic effects in Alzheimer's disease and other complex diseases become challenging to detect because they remain subtle and multiple interacting factors influence them. The small number of participants in studies creates challenges to divide data according to important variables such as age, sex or ancestry which may result in confounding variables that affect model prediction accuracy.

The quality and preprocessing of data together with GWAS data dimensions affect the reliability and reproducibility of machine learning results. The reviewed studies used different levels of preprocessing method complexity in their pipelines especially when it came to SNP filtering, imputation, normalization and feature selection. Multiple studies implemented biological feature selection approaches to decrease data dimensions while keeping important genetic data points. Araujo et al. [43] and [45], for example, used SNPs derived from the AlzGene database, leveraging prior biological knowledge to guide model input selection. The studies [44], [46], and [50] used p-value thresholds and ^{2} statistical tests to select informative SNPs which reduced noise and focused on potential causal variants.

Preprocessing strategies showed inconsistent approaches despite the implemented efforts. The full documentation of essential preprocessing procedures such as genotype imputation and quality control filters like minor allele frequency thresholds and Hardy-Weinberg equilibrium checks remains incomplete in various studies making it difficult to reproduce their results. The large variations in dataset size from under 200 cases [45] to more than 75,000 controls in [54] indicate substantial differences in data availability and quality. The absence of standardization in pipelines creates biases and reduces the ability to compare model performance metrics due to heterogeneity.

The application of ML to GWAS data faces a major limitation because it lacks the ability to interpret results. Complex models including Random Forests [43], [44], [45] and Support Vector Machines [46] and DL approaches like ResNet [52] achieved competitive accuracy yet they operate as "black boxes" which prevents understanding the biological mechanisms behind their predictions. The inability to understand how specific SNPs and their interactions contribute to results creates an obstacle for clinical and translational research to discover biomarkers and develop hypotheses. The implementation of interpretability techniques through LASSO [48] [49]and epistasis-focused algorithms [47] [50] remains limited and lacks thorough biological validation.

2.6.3 Machine Learning in Multi-sources Data for Alzheimer's disease

In this section various ML strategies have been implemented for the diagnosis of AD using muti-modal data, including SVMs, ANNs, RFs and DL models. In [56], SVMs were utilised to differentiate between Alzheimer's and MCI based on Magnetic Resonance

Imaging (MRI) scans, yielding a classification accuracy of 90.5% for classification of AD and NC [56]. Additionally, ANNs and RFs were applied to classify Alzheimer's from functional MRI images with an accuracy of 87.5% using ANNs and 90% using RFs.

DL models have been applied to AD, such as CNNs and Recurrent Neural Networks (RNNs). Different CNN architectures and methodologies were used to identify AD, MCI and NC from MRI scans with high accuracies [57] [58] [59]. Other works utilized an RNN model to predict the progression of AD [60] [61]. While in [62], the authors proposed a framework, combining CNN and RNN for longitudinal analysis of structural MR images, yielded an AD classification accuracy of 91.33% when compared to NC.

In addition to MRI scans, ML has been used in the diagnosis of AD through the analysis of other data sources such as electroencephalography and cerebrospinal fluid cerebrospinal fluid biomarkers. For example, one study combined SVMs with electroencephalography data to distinguish between AD and MCI with 96% accuracy [63]. Another research project combined SVMs with cerebrospinal fluid biomarkers to differentiate between AD and non-demented controls, achieving 93.2% accuracy [64].

The combination of GWAS with neuroimaging and clinical records data has become a fundamental research method for AD because it helps scientists understand how genetic factors interact with disease symptoms. This method overcomes the restrictions of standard GWAS because it considers the diverse characteristics of AD pathology.

GWAS research has proven effective when combined with multimodal imaging data to improve understanding of AD progression. In the study [65] implemented a new multimodal neuroimaging phenotype which combined cortical amyloid burden with hippocampal volume to run a GWAS that revealed the LCORL gene variant protects against AD progression from mild cognitive impairment. The research demonstrates how uniting genetic information with imaging biomarkers reveals protective genetic elements and reveals disease process mechanisms.

Scientists have investigated the combination of GWAS with transcriptomic and imaging data to understand the molecular basis of AD. For example in [66] researchers, developed the GEIDI federated model which detects genetic and transcriptomic effects on brain structural MRI measurements while generating genotype-dependent personalised inferences. The integration of multiple data types shows promise for both understanding AD heterogeneity and developing personalized medical approaches.

The integration of GWAS with multimodal data continues to face challenges despite recent progress. The combination of large datasets with different formats and the requirement for extensive well-identified participant groups creates major analytical obstacles. The analysis of diverse datasets requires strong analytical frameworks and tools because their processing complexity is high. The ongoing methodological developments and research initiatives improve our ability to merge GWAS with multimodal data which leads to promising breakthroughs in AD research and therapeutic development.

2.7 Chapter Summary

Over time, AD causes a gradual decline in cognitive functioning and memory, leading to a loss of the ability to recognise family members, retain memories, and even follow simple instructions.

Although the cause of AD is still unknown, there are several risk factors that have been identified which may contribute to its development. These risk factors include medical history, lifestyle choices, family dementia history and personal characteristics. By taking action to address these risk factors, it may be able to prolong the onset of AD and allow people to live longer, more independent lives.

The use of data science and ML to study AD has become increasingly popular, with several research approaches exploring this field. The available research has taken a comprehensive approach to studying AD risk factors, looking at both behavioural and biological markers. By doing this, the hope to gain an understanding of the early signs of the onset of the disease, or even to be able to predict when an individual may be at risk of developing AD in the future. This chapter has studied a wide variety of approaches, including genetics, cognitive abilities, and other biological markers, to determine whether there are any patterns that could be used to help predict or diagnose the disease in its earlier stages. The research is ongoing, and it is hoped to be able to provide more insight into the factors that contribute to the onset of AD. Next chapter will address ML models employed in the current work.

Chapter 3

Overview of Machine Learning

3.1 Introduction

ML analytics have become an increasingly important tool in the medical field, offering the potential to enhance the prediction of outcomes for diseases such as AD. This chapter will provide an overview of the different learning algorithms that form the foundation of ML, including supervised and unsupervised learning approaches. Additionally, the background of ML models, feature selection methods and rule extraction procedures will be discussed.

The use of ML analytics to classify and predict AD has been made possible by the use of multimodal data. This data includes imaging, genetics, laboratory assessments, and other types of patient information. By combining these types of data, ML models can be used to identify patterns and correlations to improve the accuracy of diagnosis and prognosis. ML can also be used to identify biomarkers that can be used to detect early stages of the disease, as well as to identify potential treatments.

3.2 Machine Learning

ML is a subfield of AI, simulates human learning by allowing computers to recognise and gain knowledge from observations made from the actual world. ML was explored as a separate discipline in the 1990s [67]. Apart from computer science, ML analytics have been applied in a variety of fields, including business [68][69][70], advertising [71][72], and medicine [73][74][75]. The ML classification process is illustrated in Figure 3.1. First and most important step for any ML model development is the existence of an appropriate dataset (i.e. accurate, complete, reliable, relevant). After data collection or requesting, the data is split into two sets: a training set used to train the ML model for the required task (for example classification), on the other hand, the test is used to evaluate the trained model and test how well it generalises on unseen samples. Some models use a another sub set of the dataset called validation set which is used to validate the model performance during training. Then several pre-processing steps are usually conducted to ensure the quality of the data is up to the required standards. In addition, the performance of a classification algorithm is evaluated by counting the number of test instances that the model correctly or incorrectly predicted.

Learning is the process of gaining information, because of their ability to reason, humans naturally learn from their experiences. Conventional computers, on the other hand, do not learn by thinking rather by following algorithms. There are several ML analytics presented in the literature nowadays. They may be divided into groups based on how they approach the learning process, supervised, unsupervised, semi-supervised, and reinforcement learning are the four primary learning algorithms [76]. In sections (3.2.1 and 3.2.2) supervised and unsupervised learning will be discussed, respectively.

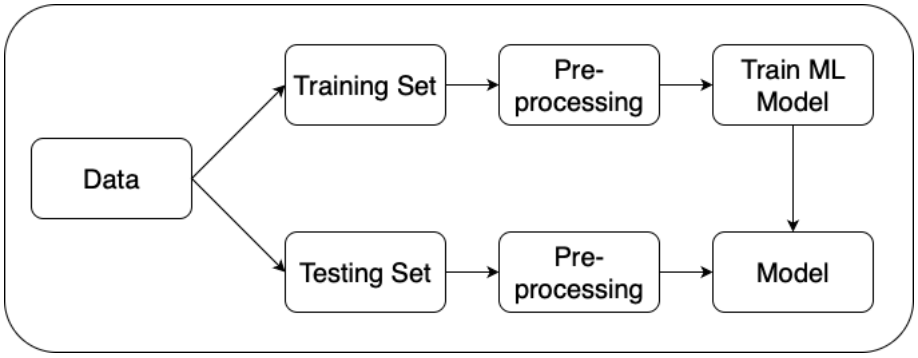


FIGURE 3.1: A general process of machine learning

3.2.1 Supervised Machine Learning

A supervised learning algorithm involves the use of labelled data [77]. For example, using genetic data to classify individuals as case or control. The learning approach allow the ML classifiers to learn the relationship between the features of the dataset and the output. After using both the features and outputs for training, the model is then tested on unseen individuals features to predict the class label. Figure 3.2 illustrates a workflow of a typical supervised learning process.

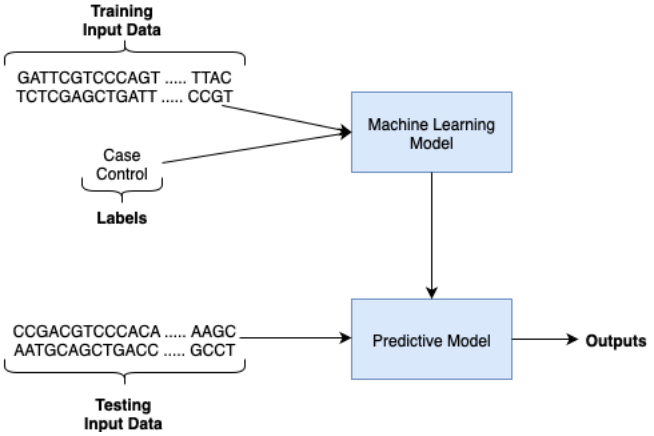


FIGURE 3.2: Supervised machine learning process

3.2.2 Unsupervised Machine Learning

With unsupervised learning, outputs are not available for training data samples, and ML algorithms are used to extract useful information from inputs [78]. Cluster analysis is one of the main uses of unsupervised learning to find patterns within the dataset [79]. For instance, in the area of genetics, unsupervised learning can be used to cluster genes that have a common characteristics [80]. Figure 3.3 shows the workflow of an unsupervised learning approach.

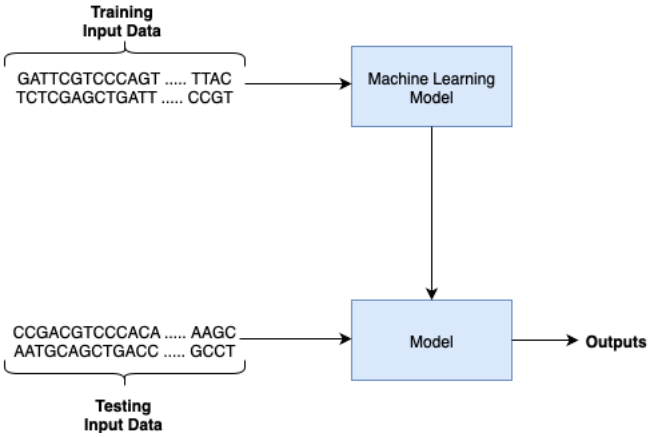


FIGURE 3.3: Unsupervised machine learning process

3.3 Machine Learning Models

ML is a method of training computers to extract knowledge from large datasets by using algorithms based on computational models. Each ML model has its own algorithm for using the data set and discovering different patterns. They can either classify the data

or predict future data using their algorithm. The choice of selecting a ML model will mostly depend on the studied problem and type of data. For instance, a classification model should not be used to solve a regression problem. In this section, ML models utilised in the current thesis will be briefly described.

3.3.1 Random Forest

A RF is a ML algorithm that constructs an ensemble of decision trees to create a powerful and robust classifier. RF uses the bagging technique, where multiple trees are trained on bootstrapped datasets—random samples with replacement from the original dataset. For classification tasks, the algorithm aggregates the predictions of individual trees by majority voting, while for regression tasks, it averages their outputs [81].

To ensure diversity among trees and reduce overfitting, RF also randomly selects a subset of features to train each tree. Typically, this subset size is the square root of the total number of features for classification or the logarithm of the total features for regression. This process helps decorrelate the trees, improving the overall model's generalisation. When a new data point is introduced, it is passed through all trees in the forest, and the model combines their predictions through aggregation to produce the final output. The bagging and bootstrapping mechanisms ensure that the model avoids overfitting and achieves robustness by leveraging multiple uncorrelated trees.

While RF is computationally efficient, particularly for high-dimensional datasets, reducing the number of features can further enhance its training speed and interpretability. Figure 3.4 illustrates the RF process, including dataset sampling, tree training, and result aggregation. For more detailed mathematical insights, including tree-splitting criteria like Gini impurity or information gain, refer to [82].

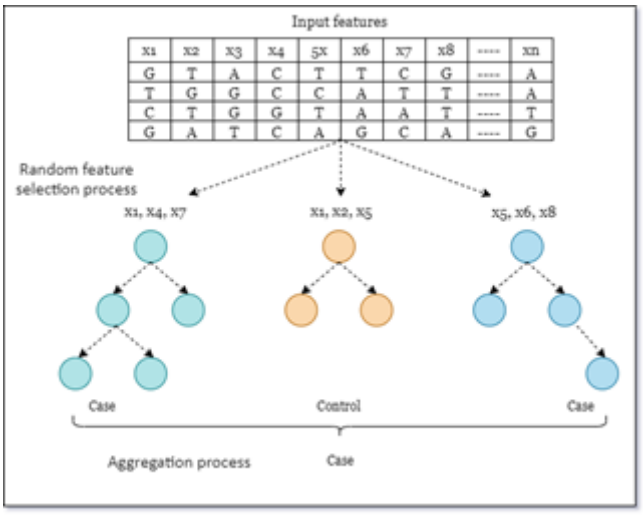


FIGURE 3.4: Random Forest in genetics application

3.3.2 Support Vector Machines

Support Vector Machines (SVMs) are supervised learning algorithms commonly used for classification and regression tasks [83]. They are particularly effective in high- dimensional spaces, such as genetic data, where the number of features often exceeds the number of observations. The core objective of SVMs is to identify a hyperplane that separates data points into different classes while maximising the margin the distance between the hyperplane and the nearest data points (support vectors) from each class. The dimensionality of the hyperplane is determined by the number of features in the dataset. For example, with two features, the hyperplane is a line, while with three features, it becomes a plane. As the number of features grows, the hyperplane exists in higher-dimensional spaces, making it challenging to visualize. However, support vectors—data points closest to the hyperplane play a critical role in defining its position and orientation.

For non-linear classification problems, SVMs use the "kernel trick" to map data into higher- dimensional spaces indirectly. This approach enables SVMs to identify non-linear decision boundaries without explicitly performing high-dimensional transformations, thereby reducing computational complexity [84]. However, the choice of kernel function (e.g., linear, polynomial, or radial basis function) and parameter optimisation is critical to the model's success.

Figure 3.5 illustrates the classification process of an SVM model separating cases and controls. For a deeper understanding of the mathematical foundations refer to [85].

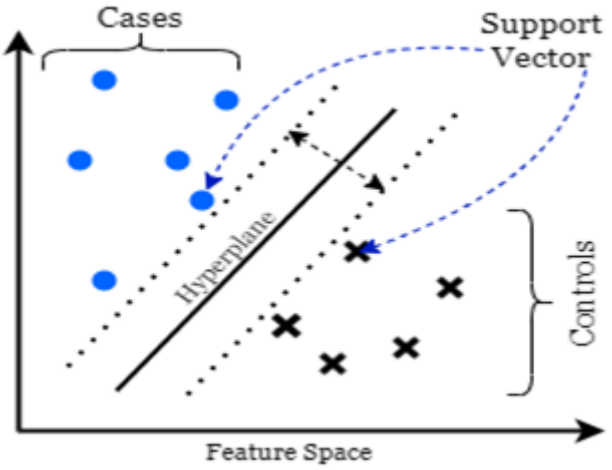


FIGURE 3.5: An example of SVM model

3.3.3 K-Nearest Neighbour

K-Nearest Neighbour (KNN) is a supervised machine learning algorithm widely used for classification and regression tasks across various domains [86]. It operates on the principle of assigning a given data point to a class based on the majority class among its kk nearest neighbors in the feature space. These nearest neighbors are identified using a distance metric, such as Euclidean or Manhattan distance, which measures the proximity between points in the dataset.

KNN is a non-parametric algorithm, meaning it makes no assumptions about the underlying distribution of the data. This flexibility allows it to adapt well to diverse datasets. Additionally, KNN is simple to implement and effective for tasks like pattern recognition and predictive modeling. However, KNN is computationally intensive for large datasets, as it requires calculating the distance between the query point and all training points. Despite these challenges, KNN remains a versatile algorithm for various applications, particularly when interpretability and simplicity are priorities.

3.3.4 Naive Bayes

Naive Bayes is a widely used supervised learning algorithm based on Bayes' rule, which calculates the probability of a class given a set of features [87]. The algorithm assumes that features are conditionally independent given the class label, an assumption that simplifies computations but is often violated in real-world scenarios [88]. Despite this, Naive Bayes performs well in many practical applications

One of the key advantages of Naive Bayes is its computational efficiency, making it suitable for large datasets and problems with high-dimensional feature spaces. It is particularly popular in domains like text classification, spam filtering, and sentiment analysis due to its simplicity and effectiveness.
However, Naive Bayes can struggle with correlated features, as the independence assumption may lead to suboptimal predictions in such cases. Additionally, it may perform poorly on imbalanced datasets.

3.3.5 Artificial Neural Networks

In the same manner as the brain is composed of a network of neurons, Neural Networks (NN) are made up of connected units or nodes and are also known as artificial neurons. NNs are essentially dense networks of interconnected layers, which are further divided into perceptrons [89], which draw their basic functionalities from neurons. A perceptron consists of three parts: input, a processing unit and an output as shown in Figure 3.6. A perceptron receives signals from preceding ones and pass their output to following perceptrons after processing their inputs. At each neuron, the weighted sum of the input is passed to an activation function to generate an output [90].

The three layers L1, L2 and L3 shown in Figure 3.7 provide a number of neurons that

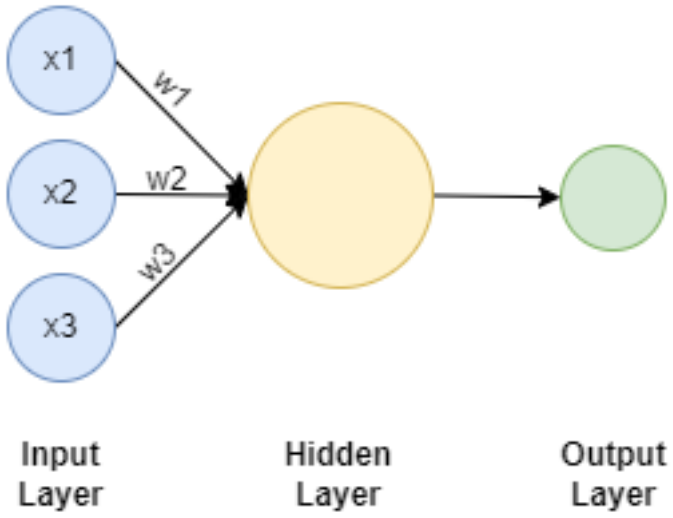


FIGURE 3.6: An illustration of a perceptron

are interconnected with each other. By using layer L2 as a reference layer, it receives the input from L1 and pass its output to L3. The inputs of the layer is represented by a vector X=[x1,x2,x3]. While the outputs are represented by vector y =[y1,y2,y3,y4] each

of which symbolised the outputs of each perceptron of L2. Alongside side the input a weight is also given in a layer. Within a network, a weight determines how the input data will be transformed. The weight matrix is represented by W, each element of the matrix is represented by W[r,q], were r and q represent row and column, respectively. Input elements and corresponding perceptrons in a given layer are connected by weights in W. the index r in the weight matrix represents the element of the input to L2, while index q identifies the perceptron in L2 where the input is entering.

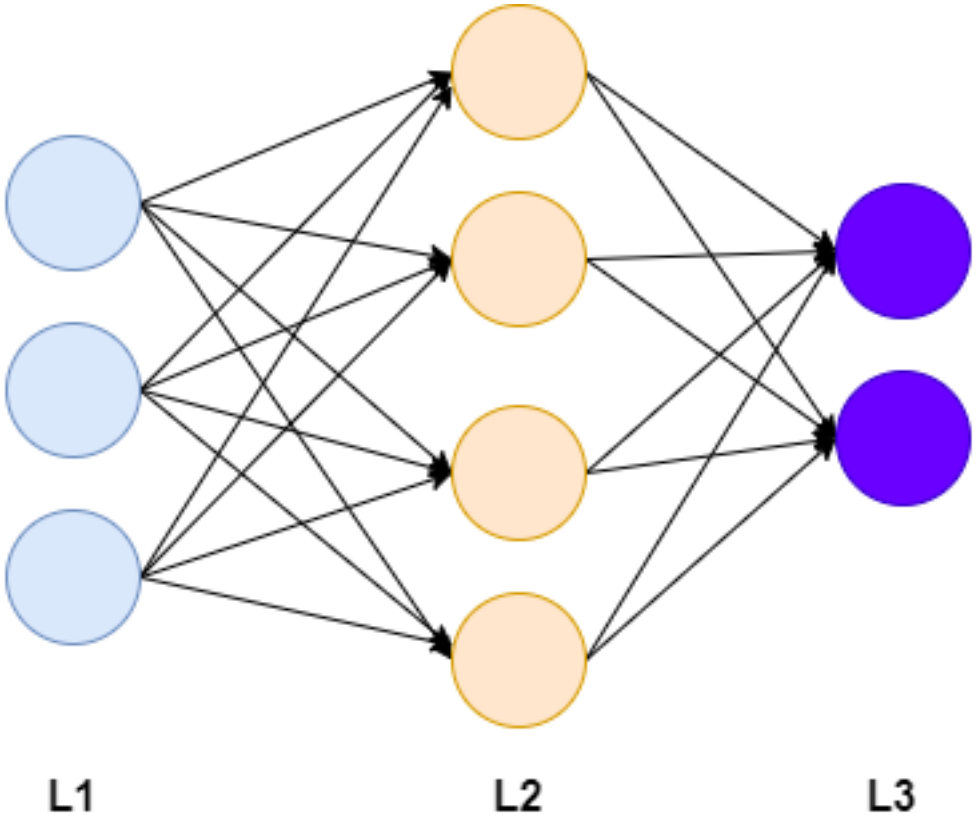


FIGURE 3.7: A neural network with one hidden layer

The procedure known as forward propagation and can be summaries in the following equation:

$$Z = W^T.X + b \tag{3.1}$$

Final step in forward propagation is to utilise an activation function for non-linear transformation.

During forward propagation, each layer of the network gives out a vector output, which is used as an input vector by the next layer, and so on until last layer. The final output of the network is produced at the final layer. During the first iteration of the

forward propagation, weights and biases are randomly initialised. These are known as the parameters of the network. The parameters are tuned (i.e. changed) according to the dataset via another process called back propagation.
In order to tune the parameters, the error of the output of network needs to be calculated in respect to the expected output. To mathematically represent the error a loss function is constructed. It is a function that maps parameters on to a scalar value that indicates how well they achieve the intended outputs. The loss function output a large value when the predictions of the network are poor. In contrast, a small error is produced when the predictions are as desired. Loss function is mathematically expressed as:

$$Loss(y, \hat{y}) = \sum_{i=1}^{n} (y - \hat{y})^2 \tag{3.2}$$

Where y represents desired output, n is the total number of samples in the dataset and y is the predicted output and given by:

$$f(b + \sum_{i=1}^{n} x_i w_i) \tag{3.3}$$

Where $F(z)$ is the activation function at a perceptron. The final step is to determine the optimal parameters that result in the minimum value of the loss function. This achieved by gradient descent algorithm. One gradient descent iteration changes the parameters W and b.

3.3.5.1 Multi-Layer Perceptron (MLP)

MLP considered as a simple type of NNs comparing to other types. They are feedforward networks in which the connections of layers are in one direction. The input layer passes the input signal to the next layer and the process continued until it reaches the output layer, in which an output is produced. Figure 3.8 shows basic structure of an MLP network consisting of two hidden layers and a single output node. There is no limit or constraints on the number of inputs, outputs, layers, or nodes per layers. The output of such neural networks depends totally on the current input therefore nodes are memoryless.

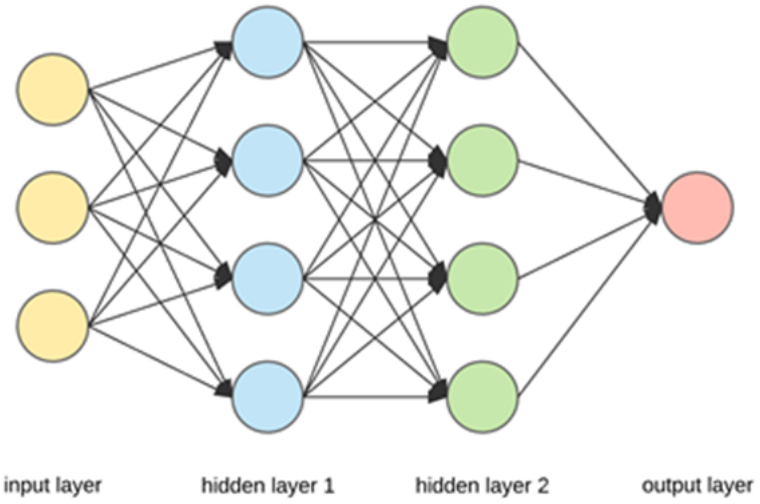


FIGURE 3.8: A MLP architecture with two hidden layers

3.3.5.2 Convolution Neural Networks (CNN)

CNNs are a widely recognised neural network architecture known for their efficiency across various domains. They reduce the number of connections between the input and hidden layers by applying filters over the input matrix, resulting in neurons in the hidden layer being connected to localised regions of the input. For improved results, multiple hidden layers are typically added. Each layer can utilise a unique filter, enabling the extraction of diverse patterns from the input data [91]. The initial layers focus on capturing low-level features such as edges and gradient directions. As the network progresses through the hidden layers, it adapts to identifying high-level features, resulting in a neural network capable of a deeper understanding of the input data. Figure 3.9 illustrates the structure of a CNN, which consists of several key layers. The three primary layers are summarized as follows [92]:

Convolutional Layer:

This is the first layer following the input layer. It extracts features from the input by performing a mathematical convolution operation between the input data and a filter of size N×NN×N. The filter slides over the input matrix, producing a feature map that highlights important characteristics such as edges. This feature map is then passed to subsequent layers for further feature extraction.

Pooling Layer:

The pooling layer reduces the dimensions of the feature map generated by the convolutional layer, lowering computational costs and preventing overfitting. Common pooling

methods include max pooling, which retains the maximum value in a region, and average pooling, which calculates the average.

Fully Connected Layer:

As in ANNs, neurons in the fully connected layer are connected to all neurons in the preceding and succeeding layers. This layer maps the learned features to the final output, aiding in tasks like classification and prediction.

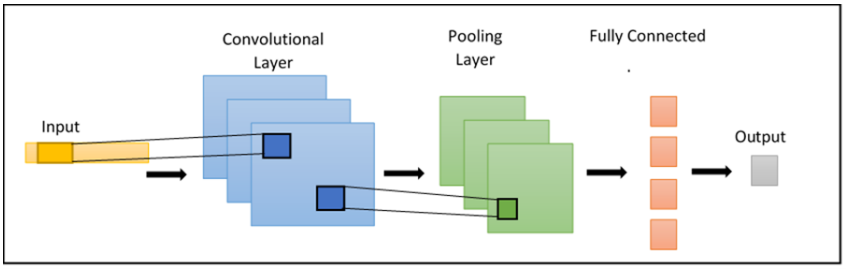


FIGURE 3.9: A typical CNN architecture

3.3.5.3 ML Models Selection

The selection of SVM, RF, KNN, NB and CNN models for AD classification stems from their common usage in related studies and the lack of a universally accepted best model for this classification task. The models differ in their underlying mechanisms, levels of complexity, and learning strategies which makes them suitable for testing across diverse datasets. The nature of Alzheimer's-related data can vary significantly from clinical features to genetic and imaging data, and it is important to explore how different models perform in capturing relevant patterns. This study adopts a comparative approach to evaluate both traditional ML and DL techniques rather than relying on a single method. DL, such as CNNs, has gained popularity in recent work involving complex biomedical data like genomic sequences, and its inclusion allows for a direct comparison against classical algorithms. Since model performance can be highly dependent on the data characteristics, and that there are no established guidelines for model selection in this domain, experimenting with a variety of algorithms ensures a more robust and unbiased assessment. This comprehensive approach increases the likelihood of identifying the most effective model for accurate and reliable AD classification. futhermore, Table 3.1 provide a comparison of the ML classifiers.

Table 3.1: Comparison of ML classifiers.

Model	Strengths	Weaknesses
SVM	Good for high-dimensional data	Sensitive to parameter tuning
	Effective with clear margin separation	Less efficient with large datasets
RF	Handles non-linearity well	Can be computationally intensive
	Robust to overfitting	Less effective on sparse data
	Interpretable	
KNN	Simple and intuitive	Sensitive to noisy data and irrelevant features
	Effective with well-separated classes	Poor performance on large datasets
Naive Bayes	Fast and efficient	Can underperform if features are correlated
	Works well with small datasets	Assumes feature independence (often unrealistic)
	Handles categorical features	
CNN	Excellent at learning spatial/hierarchical features	Requires large datasets
	Suitable for complex data like images/genomics	High computational resources

3.3.5.4 Risk of Bias in ML Development

ML models achieve high performance across numerous tasks. These models encounter multiple challenges including bias together with overfitting and data quality problems which become especially significant in healthcare settings. The presence of bias in ML models stems from unbalanced datasets which include disproportionate control to case ratios and population stratification in GWAS which generates incorrect associations. The solution to these problems requires applying stratified sampling to the data and multiple evaluation metrics for assessing models from different viewpoints. The performance of ML models depends heavily on the amount of data used for training and testing because missing values and genotyping errors significantly affect model results thus requiring proper data quality measures to select only high-quality SNPs and samples. Machine learning development often faces the problem of overfitting which occurs when models perform exceptionally well on training data yet fail to deliver good results on test data. The recommended solution to reduce this problem includes cross-validation and an independent test set and an external data set for better results. The reduction of methodological bias requires complete documentation of data pre-processing methods and model analysis procedures.

3.3.5.5 Computational Challenges of Training Deep Learning Models on Genomic Data.

The increasing interest in using DL models for genomic data analysis faces important computational challenges that need attention. The large dimensionality of genomic datasets requires extensive preprocessing steps which makes model training more complicated. The effective generalisation of DL models particularly CNNs requires extensive amounts of labeled data which is often difficult to obtain in biomedical research. The training process for these models requires significant computational resources because it needs GPUs or TPUs and extended training periods. The process becomes more complicated because of hyperparameter tuning and architecture selection and managing overfitting. The deployment of DL for genomic analysis faces significant practical challenges because of its computational requirements which become problematic in environments with limited resources. The potential of CNNs to detect complex patterns in genetic data requires evaluation against their computational expenses and practical implementation in research and clinical environments.

3.4 Transfer Learning (TL)

The input feature space and distribution of training and testing data in traditional ML are the same. ML classifiers performance can be reduced when the training and test data have different distributions. It can be difficult and costly to obtain training data whose features and distribution characteristics match those of the test data. This necessitates results in creation of a new design methodology called TL. In order to improve learning in the target domain, TL involves gaining knowledge from a dataset (source domain), and then transferring that knowledge to a new dataset (target domain) [93]. The weights of a pretrained model trained at "problem A" transferred to solve "problem B" [94].
for the TL process,given a source domain (D_s) and source task (T_s) to improve the performance on a target domain (D_t) with target task (T_t) where $D_s \neq D_t$ or $T_s \neq T_t$. A domain consists of a feature space X and a marginal probability distribution $P(X)$. Thus, the condition $D_s \neq D_t$ can be further extended as $X_s \neq X_t$ or/and $P(X_s) \neq P(X_t)$. The TL is heterogeneous when the source dataset and target dataset come from different domains, with different marginal distributions, predictive distributions, and

feature spaces. Homogeneous TL is defined, on the other hand, when the source and target datasets are less different from one another [95].

3.5 Feature Selection Algorithms

An important step in developing a ML approach is feature selection which is to reduce the number of original dataset features. Models can perform better if the number of input variables is reduced to only important ones, resulting in both a reduction in computation costs and, in most cases, improving the performance [96]. In the next sub-sections, feature selection algorithms used in this work will be explained.

3.5.1 Principal Component Analysis

Principal Component Analysis (PCA) is a powerful statistical method widely used in research for reducing the number of variables in a dataset and selecting important features [97]. The main idea of PCA is simple: it reduces the number of variables while retaining as much information as possible.

Before applying PCA, it is important to standardise the data, especially for continuous variables. This step ensures that variables with larger values don't dominate over those with smaller values, helping to avoid biased results.

PCA works by creating new variables called principal components (PCs). These PCs are linear combinations of the original variables, and they follow a specific order. he first PC captures the largest amount of variance in the data. This is essentially the average squared distance of the data points from the origin when projected onto a line. The second PC captures the next highest variance while being completely uncorrelated with the first PC.

This process continues, creating as many PCs as there are variables in the dataset. The PCs are ordered by how much variance they explain, with the first PC being the most important. Once all the PCs are generated, the next step is to decide how many to keep. Less important components (those explaining little variance) can be discarded to simplify the analysis. While PCA is commonly used for feature extraction, it can also serve effectively as a feature selection technique, as demonstrated in studies such as [98] and [99]. Feature selection using PCA involves identifying the original features (e.g.,

SNPs) that contribute most significantly to the principal components that capture the highest variance in the data. This ensures that the resulting feature vector reflects the most informative and varied aspects of the dataset [100].

In this context, the component loadings—which represent the correlation coefficients between the original GWAS features and the principal components—are key. By applying component rotations, PCA maximizes the sum of variances of these squared loadings. The absolute sum of these rotations is then used to compute an importance score for each feature, effectively allowing features to be ranked based on their contribution to the variance captured by the selected components.

The number of principal components used to calculate feature importance is determined by the cumulative variance they explain, ensuring that only the most informative components are retained. A more detailed explanation of this PCA-based feature ranking method can be found in [97].

3.5.2 Boruta Algorithm

The Boruta algorithm [101] is a feature selection technique designed to identify relevant features in a dataset. It leverages the principles of the RF classifier by introducing randomness to assess feature importance. Specifically, the algorithm compares the importance scores of the original features with those of randomised (shuffled or permuted) versions of the features.

Through an iterative process, Boruta selects or rejects features based on their importance scores, continuing this process until a stable set of relevant features is identified. This method is particularly advantageous in high-dimensional datasets, where effective feature selection is critical for optimising model performance.

This approach allows Boruta to capture both linear and non-linear interactions, and to maintain sensitivity to weak but relevant signals, which is particularly valuable in complex diseases like AD where multiple SNPs may contribute small but meaningful effects. Table 3.2 shows the strengths and limitations of PCA and Boruta algorithms.

TABLE 3.2: Strengths and Limitations of PCA and Boruta Algorithms

Method	Strengths	Limitations
PCA	Reduces dimensionality efficiently Captures variance structure	Does not consider disease labels (unsupervised) Limited biological interpretability of components
Boruta	Identifies all relevant features Incorporates feature importance from Random Forests Supports interpretability and biological insight	Computationally intensive for large GWAS Depends on performance of base classifier May select redundant features

3.6 Rule Extraction Techniques

State-of-the-art ML algorithms (such as tree-based classifiers or NNs) are known for their powerful predictive performance. However, this high level of performance comes from complex prediction mechanisms. Because of the long processing computations to reach an output, these models considered as black-box. This block-box issue limits the application of these intelligent models, especially within areas where the interpretability of a decision is of high importance such as healthcare. Alternatively, an approach called rule extraction can find patterns in data and help in explain how these models reach a final decision. Following sub-sections will highlight briefly the rule extraction algorithms utilised in the current thesis.

3.6.1 Class Rule Mining

As a special case of conventional rule mining [102], Class Association Rules (CARs) can be used, where target classes are only used as a consequence. CARs are commonly used to identify common patterns in large datasets that can be readily interpreted by humans. In most cases, confidence (c) and support (s) metrics are used to determine the strength of a rule (X) and therefore the strength of its association where support is mathematically defined in Eq 3.4 [103]:

$$s(X \Rightarrow Y) = \frac{frq(X \cup Y)}{N} \tag{3.4}$$

Where N indicates how many observations/records there are in the dataset. In a rule, confidence (c) represents the probability that factor Y occurs with factor X present and defined mathematically as in Eq 3.5.

$$c(X \Rightarrow Y) = \frac{frq(X \cup Y)}{frq(X)} \tag{3.5}$$

Rules are typically evaluated by varying thresholds for the 's' and 'c' criteria [104]. These metrics might, however, misinterpret the significance or importance of an association. The reason for this is that only the popularity of X is considered, not that of Y. An additional measure called lift accounts for the popularity of each constituent item (i.e., X and Y), which indicates how the X affects the Y, and is calculated as follows (Eq 3.6):

$$lift(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X) * s(Y)} \tag{3.6}$$

X and Y are independent when $lift(XY) = 1$, whereas $lift(XY) > 1$ indicates that they are positively dependent. Further detailed information on CAR can be found in [105].

3.6.2 Stable and Interpretable Rule Set (SIRUS)

Stable and Interpretable RUle Set (SIRUS) [106] is a rule extraction algorithm designed to produce interpretable models by leveraging a modified version of the RF algorithm. SIRUS generates a large set of potential rules and selects those that exceed a specified redundancy threshold, which is controlled by the tuning hyperparameter p0.
The optimal value of p0 is determined using cross-validation, which identifies the number of relevant rules to extract. This process evaluates the frequency with which a rule appears across the trees in the RF model, ensuring that only consistently relevant rules are retained. Rules meeting this criterion are included in the final rule set. For a detailed explanation and mathematical formulation, refer to the original work [106].

3.7 Machine Learning Evaluation

ML models get assessed through evaluation metrics which measure their success in data prediction or classification tasks. These evaluation metrics enable researchers to analyse model accuracy, precision, recall and additional performance characteristics. The selection of evaluation metrics depends on the analysis type and characteristics of the examined data set. The evaluation metrics used in this work include:

Accuracy: Accuracy measures the correct predictions over the total instances to provide a general idea about the model's prediction ability. The single use of accuracy proves insufficient for imbalanced datasets which contain unequal class distributions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.7}$$

Precision: Precision determines the correct positive instances True Positives (TP) relative to the total number of positive predictions which includes TP and False Positives (FP). It centres on the precision of positive predictions and proves valuable when the consequences of FP are significant.

$$Precision = \frac{TP}{TP + FP} \tag{3.8}$$

Recall (Sensitivity): Recall or sensitivity and true positive rate define the ratio between correctly identified positive instances (TP) and all actual positive instances (TP + False Negatives (FN)). It evaluates the model's proficiency in recognising all positive instances and holds significance when the repercussions of FN are substantial.

$$Recall = \frac{TP}{TP + FN} \tag{3.9}$$

F1 Score: The F1 score calculates the average of precision and recall to provide a balanced evaluation metric. A single value combines both precision and recall to provide a complete evaluation metric. The F1 score proves its value when class distributions are unbalanced between positive and negative instances.

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.10}$$

AUC-ROC: The AUC-ROC serves as a prominent metric that ML practitioners use for performance evaluation. The AUC-ROC curve shows how true positive rate (sensitivity) relates to false positive rate (1-specificity) when using different classification thresholds. The overall classification performance becomes better with an increased AUC-ROC value.

Researches have extensively studied classification algorithms through various techniques according to a state-of-the-art study [107]. The research community has not established a definitive performance metric that surpasses all others. A specific performance metric provides strong assessment of a classifier for certain perspectives but demonstrates weak performance for other aspects. The use of multiple evaluation techniques provides essential insight into a classifier's complete performance evaluation.

In healthcare, ML models play a crucial role in making decisions related to human health. However, the accuracy of models is not the only concern. The necessity of explainability becomes critical when making model decisions in AD classification because these decisions affect medical understanding and future intervention strategies. DL models together with complex ensemble methods demonstrate excellent predictive accuracy but fail to explain their results which creates issues regarding trust and transparency and limits clinical use. Post-hoc explainability methods SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) have become increasingly important to address these issues. SHAP as a cooperative game theory-based framework allows for unified output explanation through individual prediction feature contribution analysis. SHAP proves essential in AD models for determining the effect of specific SNPs and phenotypic elements on disease classification predictions. This method delivers both global and local interpretability that proves essential for patient-specific predictions because it shows general model feature dependencies and specific prediction reasoning. The LIME approach creates understandable explanations for complex models by developing a basic explainable linear regression model to represent the behavior around each particular prediction. The localised explanation method lets researchers and clinicians understand particular model decisions while the overall model structure remains undecipherable.

Rule extraction techniques provide additional interpretability through human-readable representations of complex model knowledge by converting learned patterns into decision paths and rules. Decision tree surrogates and rule sets from ensemble trees and logic-based decompositions prove best for genomics-based AD models because they enable

the interpretation of "if-then" conditions between SNP patterns and disease likelihood. Rule-based representations combine transparency benefits with the capability to link computational model outputs to domain expertise so researchers can check model behaviors against established biological mechanisms. SHAP, LIME and rule extraction techniques together make black-box models ML learning deployment for AD diagnosis.

3.8 Chapter Summary

In this chapter, different types of ML architectures such as supervised and unsupervised have been discussed. Additionally, ML-based feature selection techniques have been reviewed and their significance in constructing strong models has been highlighted. Moreover, this chapter has focused on the extraction of human-readable rules to facilitate understanding of the decisions taken by models. Next chapter will dive into the use of TL to classify cognitive state of individuals based on genome wide data.

Chapter 4

Transfer Learning for Classification of Alzheimer's Disease Based on Genome Wide Data

The research defined in this chapter has been published in IEEE/ACM Transactions on Computational Biology and Bioinformatics.

S. Alatrany, W. Khan, A. J. Hussain, J. Mustafina, and D. Al-Jumeily, "Transfer Learning for Classification of Alzheimer's Disease Based on Genome Wide Data," (in eng), IEEE/ACM Trans Comput Biol Bioinform, vol. Pp, Jan 3 2023, doi: 10.1109/tcbb.2022.3233869. [108]

4.1 Introduction

AD has been identified as a complex condition that is caused by a combination of hereditary and environmental factors [109]. Since there is no definitive cure for AD, studying the genes that are involved in its progression can help with early identification, close monitoring of at-risk patients, and the implementation of early treatment and prevention strategies.

GWAS is a commonly used strategy for determining the association between common DNA sequence variants and a phenotype. These large-scale studies collect genetic diversity in the form of SNPs across the human genome, and each variation is statistically assessed to identify links to a well-defined trait that is being investigated [110]. The most prevalent strategy used in GWAS is the case-control design, where the cases refer to a cohort that has been affected by the disease under study, while the controls refer to healthy (i.e., normal) subjects.

Studies in the literature have reported that genetic factors play a significant role in AD. In 2013, one of the largest AD GWAS studies identified 19 risk loci that were related to AD [111]. More recent studies have identified additional risk loci (now totaling 40) [112] [113] [114] which clearly demonstrate the significant contribution of GWAS towards understanding the genetic components associated with AD.

Although classical ML methods have achieved significant success in various practical applications, they face certain challenges in specific real-world scenarios. Supervised ML typically relies on a large volume of labeled training data that shares the same distribution as the test data. However, obtaining enough labeled data is often prohibitively costly, time consuming, or even infeasible [94]. To address this issue, TL has become a widely used approach. TL leverages the knowledge gained from solving one problem using a large dataset and applies it to other problems with relatively smaller datasets. This involves initially training a base model on a larger dataset for a specific task, followed by fine-tuning it on a smaller dataset in the target domain [94].

Although there have been many studies [48] [49] [115] [116] [117] that have used ML in the area of GWAS, there are some limitations to these studies. The study in [115] reveals that the research has a small sample size which limits the generalisation of the results. The SVM model achieved promising sensitivity at 70% but its moderate specificity at

61% shows there is potential to enhance predictive performance. In [116], researchers used multimodal data to create a predictive model for AD conversion from mild cognitive impairment through a multistep training process. The initial training step of this method transforms each modality into optimised feature vectors independently for MCI conversion prediction. The independent optimisation approach restricts the model from utilising all possible cross-modal interactions. The SWAT-CNN framework proposed in [117] demonstrates effective AD classification but requires significant computational resources for genome fragmentation and repeated CNN models training which restricts its practical use in limited computational environments. In this chapter, multiple types of TL are used for a reliable classification of AD using GWAS data. In contrast to other existing literature, the proposed study in the current chapter comprises the following novelties:

a) To the best of my knowledge, this is the first study to use deep TL to address the data size challenges associated with GWAS.

b) A comprehensive analysis of multiple types of the TL models has been proposed.

c) A robust feature selection approach is utilised to identify the most promising SNPs contributing to the AD classification.

4.2 Review of TL in Bioinformatics

TL is typically divided into three main subcategories: inductive, transductive, and unsupervised TL. These categories are defined based on the differences in context between the source and target domains and the tasks [93]. TL has been in found extensive application in various fields of bioinformatics [118] [119] [120]. For instance, Zhao et al. [121] introduced a TL-based polygenic risk score (PRS) method called TL-PRS. In this approach, an ML model trained on a large GWAS dataset from one ancestry group is fine-tuned to align with the target dataset. This method was applied to individuals of South Asian and African ancestry in the UK Biobank, focusing on seven quantitative traits and two dichotomous traits. Compared to the standard PRS method, TL-PRS demonstrated an average relative improvement in predicted R-squared of 25% for South Asian samples and 29% for African samples.

In the context of using TL in GWAS, the authors of [122] developed a TL-Multi which

is an approach developed to create PRS for populations outside of Europe. This method involves using a TL framework to gain insights from the European population, thereby improving the learning accuracy of the target data [122]. Similarly, Muneeb et. al. [123] proposed the prediction of genotype-phenotype using DL models through TL while utilising simulated data.

Unlike the previously mentioned studies, particularly those employing TL in GWAS with simulated data or focusing on gene-level classification of AD, the approach proposed in this chapter applies TL to real GWAS data. Initially, a CNN model is trained on one GWAS dataset and subsequently used to extract features from another GWAS dataset related to AD. These extracted features are then input into a SVM model to classify individuals as healthy or unhealthy at the SNP level.

The implementation of transfer learning in GWAS for AD shows great promise as a method to increase model performance when dealing with restricted labeled data. The pre-training of models on extensive related datasets enables them to learn general features which can then be adapted to specific AD-related genomic data for better generalisation and reduced need for extensive training from scratch. The acquisition of large well-annotated datasets in AD research faces significant obstacles. CNNs demonstrate exceptional capability in modelling genomic data because they detect local patterns, and hierarchical structures present in sequence-based data. The success CNNs achieve in image recognition enables them to detect motifs and dependencies and spatial relationships in genomic sequences which could be linked to disease phenotypes. The CNN architecture performs automatic feature extraction which gives it an advantage over traditional GWAS methods that use manually defined SNP features. The combination of CNNs with TF enables them to use learned representations from large biological datasets to improve their robustness against noise and variability in smaller AD-specific datasets. The combination of CNN architectures with TF creates a strong framework which enables researchers to discover intricate nonlinear connections between genetic variants to better understand AD genetics.

4.3 Materials and Methods

The proposed approach exploits TL where multiple datasets are used to train a deep ML model and transfer the learned knowledge efficiently to target domain. Detailed

experiments were conducted to analyse the effectiveness of varying types of TL and to investigate the impact of knowledge transfer from one dataset to another in GWAS analysis. The proposed approach is composed of several components that include quality control, association test, feature selection and classification. A detailed description of each task is provided in the following sub-sections.

4.3.1 Datasets

The following three datasets comprise the GWAS data sets used in this study:

Dataset A: ADNI GWAS dataset

The Alzheimer's Disease Neuroimaging Initiative (ADNI) [124] dataset is a large-scale, longitudinal dataset that has been collected as part of a collaborative research effort to better understand and track the progression of AD. The dataset includes comprehensive clinical, imaging, genetic, and biomarker data from individuals with AD, mild cognitive impairment, and healthy controls.

The ADNI dataset has been collected from multiple sites across North America and has undergone rigorous QC measures to ensure data reliability and consistency. It encompasses various data modalities, including structural and functional MRI positron emission tomography scans, cerebrospinal fluid biomarker measurements, genetic data, and clinical and cognitive assessments.

The ADNI dataset provides a valuable resource for researchers studying AD and related neurodegenerative disorders. It allows for the exploration of disease progression, identification of biomarkers and development of diagnostic and prognostic models. The longitudinal nature of the dataset enables the investigation of changes over time and the examination of factors contributing to disease progression or conversion from MCI to AD.

Within the ADNI dataset, there is a subset specifically designed for GWAS Studies related to AD. The ADNI GWAS subset includes genetic data from participants. These genetic variants are analysed to identify associations with AD risk and disease progression.

Researchers can utilise the ADNI GWAS subset to perform large-scale genetic analyses, such as identifying genetic risk factors, exploring genetic interactions, and investigating

the relationship between genetic variants and various phenotypic measures. To fulfil the objectives of the proposed study, GWAS data from ADNI1 were accessed, where individuals with CN or AD were chosen. A total of 388 subjects were identified, producing 174 cases and 214 controls.

The dataset is originally presented in plink file format, with three files: 'bim', 'bed', and 'fam' files. In the 'fam' file, subject characteristics are recorded, while SNP (feature) characteristics are stored in the 'bim' file, including location, name, and allele representation. Finally, 'bed' files contain machine codes that are unreadable to humans and comprise 8-bit codes representing the genotype codes, as well as map the information between fam and bim files. In this study, SNPs were used as features to classify the individuals into CN or AD cases. Table 4.1 shows the statistics of the dataset including age and Mini-mental State Examination (MMSE) is 30-point questionnaire used to measure cognitive impairment. Table 4.1 also shows that most cases carry at least one of copy of APOE4 gene.

Dataset B: AD GWAS Dataset

TABLE 4.1: Characteristics statistics of Alzheimer's disease and normal subjects of ADNI dataset.

	Age	**Sex (M/F)**	**Education (years)**	**MMSE**	**APOE4**	**ADAS11**
Cases	75.35	92/82	15	23	1	18.11
Controls	75.66	115/99	16	29	0	5.83

The second dataset used in this study is a GWAS case-control dataset obtained from [125]. The inclusion criteria for participants are: a) self-reported European ethnicity, b) compliance with National Alzheimer's Coordinating Centre standards, and c) late-onset AD confirmed by board-certified neuropathologists in cases and no neuropathology in controls. Plaque and tangle assessments, which are unique structures that affect cells in the brain and could contribute to the pathophysiology of the disease, were conducted on all cases and controls. Samples with a history of stroke, Lewy bodies, or any other neurological disorder were excluded. The final dataset includes 191 males and 173 females partitioned into 176 cases and 188 controls, each with genotyping information of 502,627 SNPs. The DNA of participants was genotyped via the Affymetrix GeneChip Human Mapping 500K Array Set. Detailed information regarding the dataset can be found in the primary study [125].

Dataset C: AdaptMap goat GWAS dataset

Unlike the two datasets mentioned above, which contain human records, the third dataset we use in this study is AdaptMap [126]. It comprises 4,653 animals representing 169 populations from 35 countries spread across six continents. The animals were genotyped using an Illumina GoatSNP50 BeadChip with 53,347 SNPs [127].

The application of TF to combine animal genetic data with human GWAS provides a new method for studying complex diseases such as AD. The direct application of models trained on animal GWAS data to human AD datasets remains scarce but the concept receives backing from related studies. Biomedical research has depended on animal models to advance knowledge about human diseases including neurodegenerative conditions [128] [**?**] [129]. Reference [130] explain how high-throughput animal models enable researchers to functionally verify GWAS signals while demonstrating how animal studies can help advance human disease research. The Dlgap2 gene emerged as a protective candidate in a genetically diverse mouse model of AD which later received verification through human GWAS studies [131]. The use of animal data remains valid because animal models share genetic and physiological features with humans which enable researchers to discover disease mechanisms and therapeutic targets.

4.3.1.1 SNPs as features

SNPs are the most common type of genetic variation found in individuals' DNA sequences [132]. They are single base pair differences occurring at specific positions in the genome, where one nucleotide (adenine, cytosine, guanine, or thymine) is substituted by another. SNPs can be present throughout the human genome and can vary in their frequency within populations.

SNPs play a crucial role in human genetics and have been extensively studied to understand their association with various traits, diseases, and drug responses [133]. They can influence phenotypic differences among individuals, including susceptibility to certain diseases, response to treatments, and variations in physical and physiological characteristics.

The Human Genome Project and subsequent advancements in DNA sequencing technologies have greatly contributed to the identification and cataloging of millions of SNPs across the human genome. These SNPs serve as markers that allow researchers to investigate genetic variation within populations and explore their relationship with specific

traits or diseases.

SNPs can have different effects on gene function and protein synthesis. They can occur in coding regions, affecting the amino acid sequence of a protein and potentially altering its function. SNPs can also be found in non-coding regions, including regulatory regions that control gene expression. Changes in regulatory SNPs can impact the binding of transcription factors and alter the levels of gene expression [133]. Due to their abundance and wide distribution in the genome, SNPs have become a key focus in genetic association studies. GWAS often rely on genotyping large numbers of SNPs to identify associations between specific genetic variants and diseases or traits of interest. By studying the patterns of SNPs across populations and their correlation with phenotypic traits, researchers can gain insights into the genetic basis of complex diseases, genetic diversity, and population history.

4.3.1.2 Data Representation

This section outlines the representation of genetic data and the binary transformations used to convert the information into a format suitable for statistical manipulation. In this study, bi-allelic SNPs are used. These SNPs can be represented in two ways: as individual alleles or as genotypes. The allele representation refers to each of the two possible variants at a single SNP position. The genotype representation, on the other hand, describes the combination of the two alleles present in an individual at that position. Examples of this type of representation can be seen in Table 4.2 and Table 4.3. To make the data more suitable for statistical manipulation, binary transformations are employed. This involves converting the genetic data into a binary format, which can be used to generate a numerical value. This numerical value can then be used to perform statistical tests and analyse the data.

Table 4.2: Allele Representation

SNP	
Allele 1	Allele 2

Table 4.3: Genotype Representation

SNP		
AA	AB	BB

SNPs are classified not only by their frequency and their alleles. Alleles are identified as either dominant or recessive, or commonly referred to as major or minor, respectively, as demonstrated in Table 4.4. It's important to note that the genotypic expression within the focus population determines which allele is dominant or minor. The dominant allele or major allele is the one that is presented in most of the population, and it generally masks the contribution of the recessive allele. In other words, the dominant allele is more likely to be expressed than the recessive allele.

TABLE 4.4: Dominant and recessive allele representation

SNP			
Allele 1		Allele 2	
Dominant	Recessive	Dominant	Recessive
A	B	A	B

When combined, these alleles create one of three genotype states as displayed in Table 4.3. These states can also be described in terms of their respective characteristics, as represented in Table 4.5. If both alleles are dominant (AA), the genotype is referred to as dominant homozygous. Conversely, if both alleles are recessive (BB), the genotype is referred to as homozygous recessive. Finally, if allele 1 is dominant and allele 2 is recessive (AB), then the genotype is referred to as heterozygous.

TABLE 4.5: Homozygous and Heterozygous representation

SNP		
Dominant Homozygous	Heterozygous	Recessive Homozygous
AA	AB	BB

4.3.1.3 Data Format

GWAS data is typically stored in three PLINK files: the .fam file, the .bim file, and the .bed file. The .fam file is a PLINK binary file that contains crucial information about the study subjects, such as their characteristics and identity codes. The data of .fam file is provided in two tables: Table 4.6 and Table 4.7. The .bim file, which contains information on each SNP, is shown in Tables 4.8 and 4.9. The .bed file (Table 4.10) contains the genotype data encoded in a binary format. All three files are required to perform statistical analysis in PLINK. It is important to note that the .fam file is used

to link the subject data to the genotype data in the other two files.

Table 4.6: Format of a PLINK .FAM file

IID	FID	PID	MID	Sex	Phenotype
1	Fam1	0	0	2	1
2	Fam2	0	0	2	2
3	Fam3	0	0	1	1

Table 4.7: Variables Description of PLINK .FAM file.

Variable	Description
IID	Individual ID
FID	Family ID
PID	Individual's father's IID (Paternal ID)
MID	Individual's mother's IID (Maternal ID)
SEX	1 for male, 2 for female
Phenotype	1 for control, 2 for case

Table 4.8: Format of a PLINK .BIM file

Chromosome	SNP ID	POS	BP	Allele 1	Allele 2
8	rs4734674	110.849	103963502	C	T
10	rs17436819	94.9892	77437096	G	A
10	rs386976	46.6912	19972134	C	T

Table 4.9: Variables Description of PLINK .BIM file

Variable	Description
Chromosome	Chromosome Code
SNP ID	SNP ID
POS	Position of SNP
BP	Base-pair coordinate
Allele 1	Usually minor allele
Allele 2	Usually major allele

Table 4.10: Genotype data description in PLINK .BED file

Genotype Code	Description
00	Homozygous for first allele
01	Missing
10	Heterozygous
11	Homozygous for second allele

4.3.1.4 Features Encoding

The representation of SNP data as categorical or numerical features is an important consideration when choosing a classification algorithm. categorical features can be represented as discrete categories, such as AA, AG or GG, and numerical features as integer values, such as 0, 1 and 2. Algorithms such as the Decision Tree and RF are able to work with categorical features, while others such as the SVM and ANN are only able to work with numerical features. Therefore, it is necessary to encode the SNPs as numerical features in order to use these more advanced algorithms.

Label encoding is a method of encoding categorical features that can be used when dealing with genetic data. It involves assigning each genotype a numerical value that corresponds to the number of minor alleles in the genotype. For example, homozygous major alleles are encoded as 0, heterozygous alleles are encoded as 1, and homozygous minor alleles are encoded as 2. This encoding technique helps to preserve all the information while also minimising the number of generated features. This is especially important when dealing with large datasets with many different genotypes. By using label encoding, the data can be processed and analysed more efficiently [134].

Another encoding algorithm is one-hot encoding scheme is a convenient and popular way to represent genotype information in binary form. It is especially useful in detecting gene-gene interactions, as shown in [135]. In this scheme, three features are created for each SNP, and each feature encodes whether its corresponding genotype is present or not. This means that only one of the three features is set to 1, while the other two are set to 0. By taking this approach, the genotype information is effectively represented in an easy-to-understand way, which makes it a popular choice for encoding genotype information. Table 4.11 shows an example of both encoding methods.

TABLE 4.11: Encoding methods for SNP data with two alleles A (major allele) and B (minor allele). The label encoding represents each genotype through minor allele count. While one-hot encoding represents SNP with three feature, one for each genotype.

SNP	Label Encoding	Genotype (One-hot Encoding)		
		AA	AB	BB
AA	0	1	0	0
AB	1	0	1	0
BB	2	0	0	1

4.3.2 Quality Control

In the proposed study, individuals and SNPs were subjected to QC and filtering procedures in accordance with conventional QC protocols and guidelines as shown in [136] using PLINK software.
In GWAS, QC measures are employed to identify and eliminate low-quality DNA samples and markers. These measures help ensure the reliability and accuracy of the genetic data used in the study. Some common QC steps in GWAS include:

Sample QC: This involves assessing the quality of DNA samples used in the study. QC measures may include checking for sample-related biases. Samples that do not meet predefined quality thresholds may be excluded from further analysis which are defined as below:
1) Genotyping Call Rate: The genotyping call rate measures the proportion of successfully genotyped markers for each DNA sample. Samples with low call rates may indicate poor DNA quality or technical issues during genotyping and can lead to unreliable results. Setting a minimum threshold for genotyping call rate helps exclude samples with insufficient data [137].
2) Relatedness and Duplicate Samples: Duplicate samples or samples from closely related individuals can introduce bias and inflate false-positive associations. QC includes identifying and removing duplicate or closely related samples using methods like identity-by-descent estimation or genetic relatedness calculation [138].
3) Genetic Ancestry and Population Stratification: Genetic differences between populations can lead to spurious associations in GWAS. Samples showing significant differences in genetic ancestry from the study population may indicate population stratification and require appropriate adjustments [139].

Marker QC: Marker QC focuses on evaluating the quality of genetic markers used in the study. Markers with low quality are removed using the following:
1) Call Rate and Missing Data: The marker call rate assesses the proportion of successfully genotyped samples for each genetic marker. Markers with low call rates may indicate technical issues or genotyping errors. Similarly, markers with excessive missing data may compromise statistical power and introduce bias. QC filters are applied to exclude markers with low call rates or high rates of missing data [138].

2) Hardy-Weinberg Equilibrium (HWE): The HWE test assesses whether the observed genotype frequencies of markers in a population conform to the expected frequencies under certain assumptions. Markers deviating significantly from HWE expectations may suggest genotyping errors, population stratification, or other biases. QC filters can identify and exclude markers that deviate from HWE [140].

3) Minor Allele Frequency (MAF): MAF refers to the frequency of the less common allele at a genetic marker in a population. Very low MAF can limit the statistical power to detect associations. QC measures may exclude markers with low MAF, depending on the study design and sample size [35].

By implementing these QC measures, the integrity, and reliability of GWAS results can be improved by removing low-quality DNA samples and markers, reducing false positives. This ensures more accurate and robust genetic associations in GWAS studies.

The QC steps are essential to the process of obtaining reliable and accurate results from a genetic dataset are summarised in Table 4.12. By using these criteria, it is possible to identify and remove any SNPs or individuals that could potentially produce biased results.

TABLE 4.12: GWAS Quality Control Steps Description

QC Step	**Description**
SNPs missingness	Missing SNPs in a large percentage of the Individuals are excluded.
Individuals' missingness	Individuals with a high rate of genotype missingness are excluded.
Sex discrepancy	Check sex of individuals depending on their X chromosome homozygosity
Autosomes Chromosomes	Only selecting SNPs of 1 to 22 Chromosomes
MAF	SNPs above a MAF threshold are included.
HWE	SNPs that deviate from HWE are excluded.
Relatedness	Generates a list of persons with relatedness degree greater than a specified threshold.
Population stratification	Individuals from different populations present in the study.

For Dataset A, there were initially 620,901 SNPs before genotyping trimming. Based on the HWE test, 72,490 markers were excluded (with p = 0.1); 61,065 markers failed the HWE test in cases, whereas 72,490 markers failed the HWE test in controls. The missingness test failed 31,368 SNPs (GENO ¿ 0.1). A total of 154,598 SNPs failed the frequency test (MAF ¡ 0.1). In total, 411,077 SNPs remained after frequency and

genotyping trimming. One individual was removed for low genotyping (MIND ¿ 0.1). After all QC stages, a total of 398 individuals and 411,077 SNPs were left for subsequent analysis.

To filter out low-quality genetic markers in Dataset B, several QC were performed. First, SNPs with a missing genotype rate greater than 5% were eliminated. Markers also filtered for HWE with a p-value less than 0.001 and a MAF less than 0.05. Individuals were subjected to QC processes, including checks for missing genotyping data (less than 5%), relatedness, and sex-homozygosity. After these QC steps, 356,499 SNPs remained for subsequent analysis.

For Dataset C, SNPs and samples are filtered out for missing genotype data (0.1) and the MAF was less than 0.05. a total of 51117 SNPs and 2765 samples pass filters and QC.

4.3.3 Association Analysis

Association analysis is an incredibly versatile concept, encompassing a wide array of methodologies that are used to identify and analyse relationships between variables. This can include anything from simple statistical filtering to more complex relationship modelling techniques, utilising both univariate and multivariate data. Association analysis can be applied to a variety of fields to uncover valuable insights from data.

Univariate analysis is one of the most important methods used in this field of research, as it allows researchers to measure the association between a single variable (X) and a response (Y). An example of this is the analysis of the association between a single genetic variant and the expressed phenotype. This type of analysis is not a definitive answer to a question, but rather a tool used to provide insight into potential relationships between two variables. Univariate analysis is often used to filter data statistically and to identify patterns that can be further analysed. For instance, if a researcher is looking for a certain genetic variant associated with a specific phenotype, univariate analysis can be used to narrow down the data set to those variants that are most likely to be associated with the phenotype. Once the data has been filtered, further analysis can be done to confirm the association.

In standard practice, a GWAS as is commonly conducted will involve a univariate analysis of the data. This analysis can be performed using a variety of methods, such as Fisher's Exact Test, Chi-Squared Test, or Logistic Regression [141]. The choice of which

test to be utilised depends on the type of data being used, and its size [37]. For family-based association testing, the Transmission Disequilibrium Test [142] is commonly used, while logistic regression is typically used for population-based association testing with unrelated samples. When it comes to sample size, Fisher's exact test is more appropriate for small samples [37] while Pearson's Chi-Squared test can be used with larger samples. Because the resulting information from association analysis varies, it is critical to select the appropriate method for the given context. The association of all SNPs (in Dataset A and Dataset B) within the study with disease status of binary variables (0/1) for case and control patients was assessed using logistic regression. An association test between SNPs and the AD was carried out to decrease the computationally enormous number of genetic variants. The SNPs are sorted in ascending order by p-value, and only the first 5000 SNPs are retrieved for further analysis.

Quantile-quantile plots are often employed in standard case-control approaches to assess the success of QC protocols. While this is not a precise measure, it can certainly act as a flag for any issues in the data such as population stratification. Figures 4.1 and 4.2 illustrate a tail-end deviation from the assumed values when the null hypothesis is assumed. Generally, the deviation begins at ¿3, which indicates that the QC process was satisfactory for both datasets.

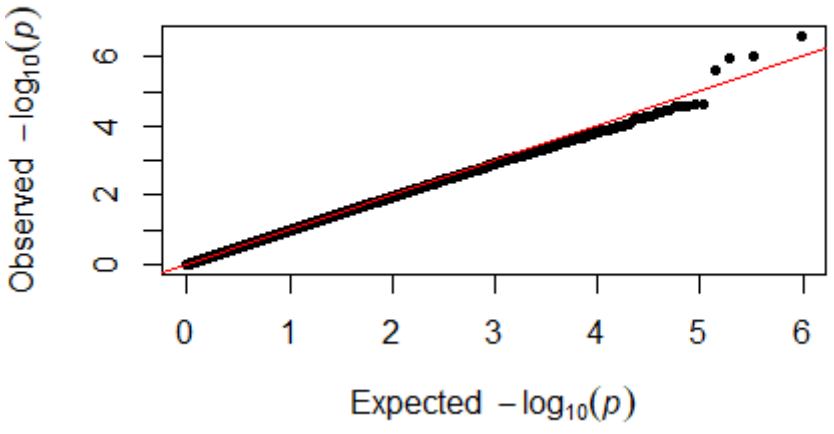


FIGURE 4.1: Quantile-quantile plot shows the deviation from the null hypothesis line for Dataset A.

Figures 4.3 and 4.4 depict Manhattan plots for Dataset A and B respectively, which

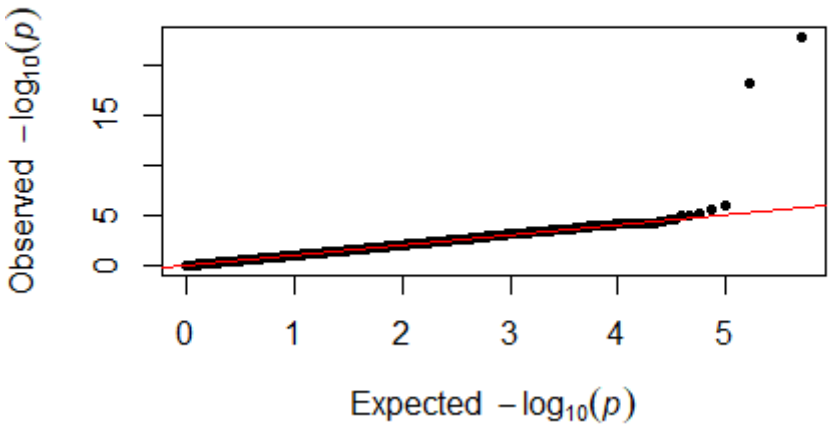


FIGURE 4.2: Quantile-quantile plot shows the deviation from the null hypothesis line for Dataset B.

reflect the -log10 p-values generated through the standard case-control approach. Those values that surpass the blue threshold are marked as 'suggestive significance'. In GWAS literature, a significance threshold 5108 is commonly used as a reference value of a convincing association, yet none of the SNPs in Dataset A exceeded that threshold. On the other hand, in Dataset B, two SNPs (rs429358 and rs4420638) surpassed the GWAS threshold, indicating higher significance. Thus, the Manhattan plots provide a useful visual representation of the significance of the SNPs, particularly with respect to the different thresholds.

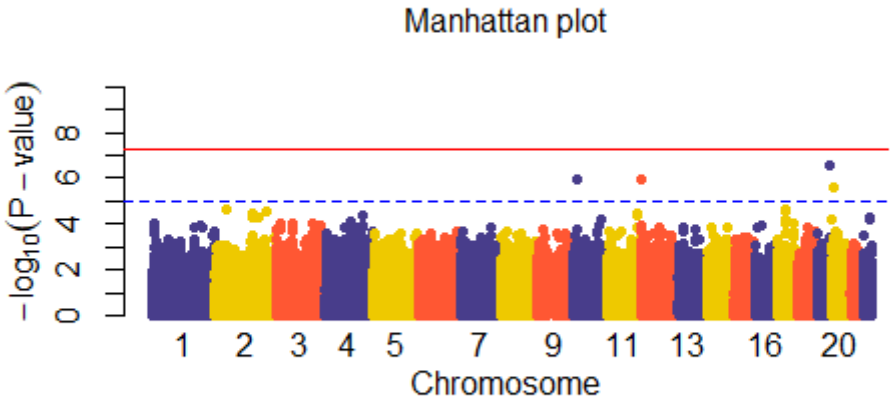


FIGURE 4.3: Manhattan plot of standard case-control shows association of between genotypes and AD for Dataset A.

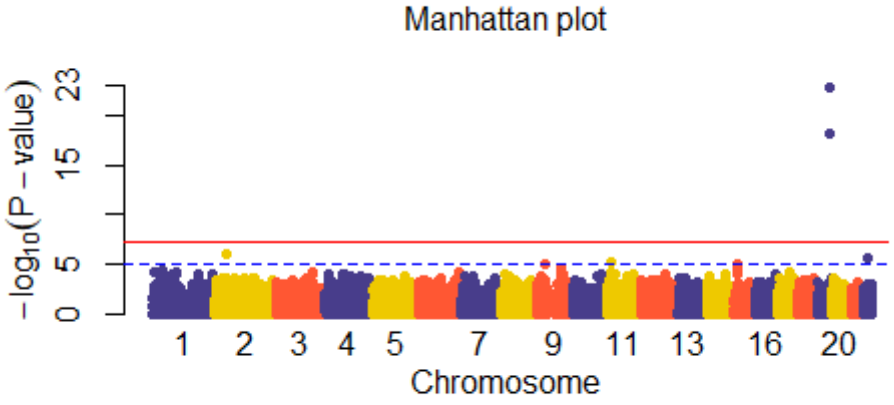


FIGURE 4.4: Manhattan plot of standard case-control shows association of between genotypes and AD for Dataset B.

4.3.4 Feature selection

GWAS involve high-dimensional data, making direct interpretation challenging, as most SNPs are irrelevant or lack meaningful information. Therefore, identifying the most important SNPs is essential. This offers three key benefits: First, it simplifies the interpretability of the ML model, Second, it reduces the model's variance, thereby minimizing the risk of overfitting, and Finally, it decreases the computational cost of training the

ML model by working with fewer features. At this stage, the results of the association analysis are used to select features that are significantly linked to the target phenotype. In this study, the Gini measure, which is one of the RF methods for measuring feature relevance, is used as a feature selector. A substantial number of SNPs are identified as irrelevant with extremely low significance values. Therefore, any SNPs with a Gini value of 0.0009 or higher are included in the feature set for classification. The significance criterion of 0.0009 was chosen by a trial-and-error approach because it can identify the SNPs that reflect favourable results in the classification task. RF selected a total of 120 SNPs as important features for the classification task of both of Dataset A and Dataset B in which each dataset has 60 significant SNPs. However, for Dataset C, only 57 SNPs are significant according to the analysis by Bertolini et al. [143].

Table 4.13 shows the top 10 SNPs chosen by RF during the feature selection stage. The rs429358 SNP exists within the APOE 4 allele that represents the main genetic risk factor for developing late-onset AD. People who possess one or two copies of the 4 allele face a significantly elevated danger of developing AD. The 4 allele leads to higher amyloid-beta accumulation and reduced clearance that both play a role in Alzheimer's disease development.

The rs4420638 SNP exists close to the APOC1 gene and commonly occurs together with APOE 4 inheritance. Research indicates that the G allele of rs4420638 increases the chances of developing Alzheimer's disease and cognitive deterioration. The APOE gene's proximity to this variant suggests that its AD associations stem from genetic linkage disequilibrium with the 4 allele.

The rs7718940 SNP exists within the APOE gene sequence. The SNP exists in a position that could indicate linkage disequilibrium with rs429358 which leads to an indirect association with AD risk.

The rs862245 SNP exists as an intronic variant within the APOC1 gene. The gene expression or splicing processes may be influenced by this variant although direct AD associations remain less established. The close location of this variant to APOE indicates possible relevance through linkage disequilibrium.

The rs153864 SNP exists in the ATP6AP1L gene which functions as a vacuolar ATPase to maintain lysosomal acidification [144]. The SNP shows potential links to AD through its association with lysosomal dysfunction [145] although this relationship needs further investigation.

The KIF2A gene contains two SNPs known as rs37032 and rs16890651 which encode

kinesin family proteins that regulate microtubule dynamics and neuronal development [146]. The association between KIF2A variants and AD needs additional research to determine their involvement because these variants have been linked to other neurological conditions. These SNPs indicating that the model is effective in identifying the most promising features relevant to the disease.

TABLE 4.13: Characteristics of the top 10 SNPs being selected as important features

SNP	Location	Function	Gene
rs2937774	5:74124992		
rs26642	5:62488562	Intron Variant	IPO11
rs153864	5:62425115	Intron Variant	IPO11
rs7718940	5:86207592		
rs862245	5:82289918	Intron Variant	ATP6AP1L
rs429358	19:44908684	Coding Sequence Variant	APOE
rs4420638	19:44919689	Downstream Transcript Variant	APOC1
rs12374530	5:63761206		
rs37032	5:62388203	Genic Downstream Transcript Variant	KIF2A
rs16890651	5:62333712	Intron Variant	KIF2A

4.3.5 Transfer Learning

In order to improve learning in the target domain, TL involves gaining knowledge from a dataset (source) and transferring that knowledge, in the form of a pre-trained model, to a new dataset (target). There are different settings of TL depending on the difference in task and domain of source and target datasets. For inductive TL, the target task is different from the source task, regardless of whether the source and target domains are similar or different. In contrast, transductive TL is used when the tasks in the source and target are the same, but the domains differ. For example, in this study, the source and target datasets for human GWAS vary in terms of genotyping platforms. Since genotyping platforms influence marker selection strategies and the number of markers, these variations affect the data, necessitating a transfer learning approach to account for these differences [147].

4.3.6 Experiment Design

The proposed model framework is illustrated in Figure 4.5. GWAS data undergoes preprocessing and filtering to retain only high-quality samples and markers, utilizing

appropriate QC methods across all datasets. Logistic regression-based association testing identifies SNPs strongly linked to the disease. Additionally, the RF algorithm is employed to select key features and reduce dimensionality, ensuring the feature count aligns with the number of available observations.

Through trial-and-error testing, convolutional layers between 2 and 4 were selected. Excessive layers risk overfitting the model, while too few may restrict its functionality [148]. Following best practices from related literature [117] [149] [150], convolutional layers and other DL model hyperparameters were determined [151] [152] [153]. For SVM and RF classifiers, a grid search was performed to optimise user-defined hyperparameters. The structures of the DL models are detailed in Table 4.14.

In this study, TL is applied in three ways:

- Classification: Using the pre-trained model to classify new observations directly.
- Fine-tuning: Adjusting the classifier, or part of it, by retraining on a new dataset.
- Feature extraction: Feeding the output of the final layer of the pre-trained model into a ML model.

After completing data processing and filtering, several experiments were conducted to evaluate TL's effectiveness in GWAS:

Experiment 1 (EXP1): Transductive TL was applied to train a model using source and target datasets from different domains but with the same task. Dataset A was used as the source dataset and partitioned into 80% for training and 20% for testing. A base CNN was pre-trained on Dataset A after testing multiple CNN architectures, with the best-performing architecture selected (details in Table 4.14.A). The trained base CNN was then used for prediction, fine-tuning, and feature extraction on Dataset B, the target dataset.

Experiment 2 (EXP2): Inductive TL was applied using source and target datasets from different domains with different tasks. The model was trained on GWAS data from Dataset C (animal data) to classify goats into 11 subcontinental breeds, as detailed in Table 4.14.B. The model trained on Dataset C was then fine-tuned and applied to classify individuals in Dataset B.

Experiment 3 (EXP3): Inductive TL was extended by using the pretrained model from Experiment 2. The model, trained on Dataset C, was fine-tuned and applied to classify individuals in Dataset A, treating Dataset A as the target dataset for this experiment.

Experiment 4 (EXP4): Inductive TL was further applied using the pretrained model from Experiment 2. This time, the model was fine-tuned and applied to an aggregated dataset that combined individuals from Dataset A and Dataset B, treating the combined dataset as the target for this experiment.

All experiments employed ML analytics built with the Scikit-learn Python library [154]. The PyPlink library [155] was used for genotype data processing in Python. DL models were developed using Keras with TensorFlow as the backend [156].

TABLE 4.14: Architectures of the Proposed CNNs; (A) for exp1 and (B) for Exp2, 3 and 4

CNN Model A		**CNN Model B**	
Layer Type	Description	Layer Type	Description
Conv1D	F = 16, K = (5,), ReLu	Conv1D	F = 16, K = (5,), ReLu
Conv1D	F = 16, K = (3,), ReLu	Conv1D	F = 32, K = (3,), ReLu
Pool1D	Max Pooling (2,)	Pool1D	Max Pooling (2,)
Dropout	10%	Dropout	10%
Reshape	Flatten	Conv1D	F = 32, K = (3,), ReLu
Dense	F = 64, Sigmoid	Pool1D	Max Pooling (2,)
Dropout	10%	Dropout	10%
Dense	F=2, softmax	Conv1D	F = 32, K = (3,), ReLu
		Pool1D	Max Pooling (2,)
		Dropout	10%
		Reshape	Flatten
		Dense	F = 64, Sigmoid
		Dropout	10%
		Dense	F=2, softmax

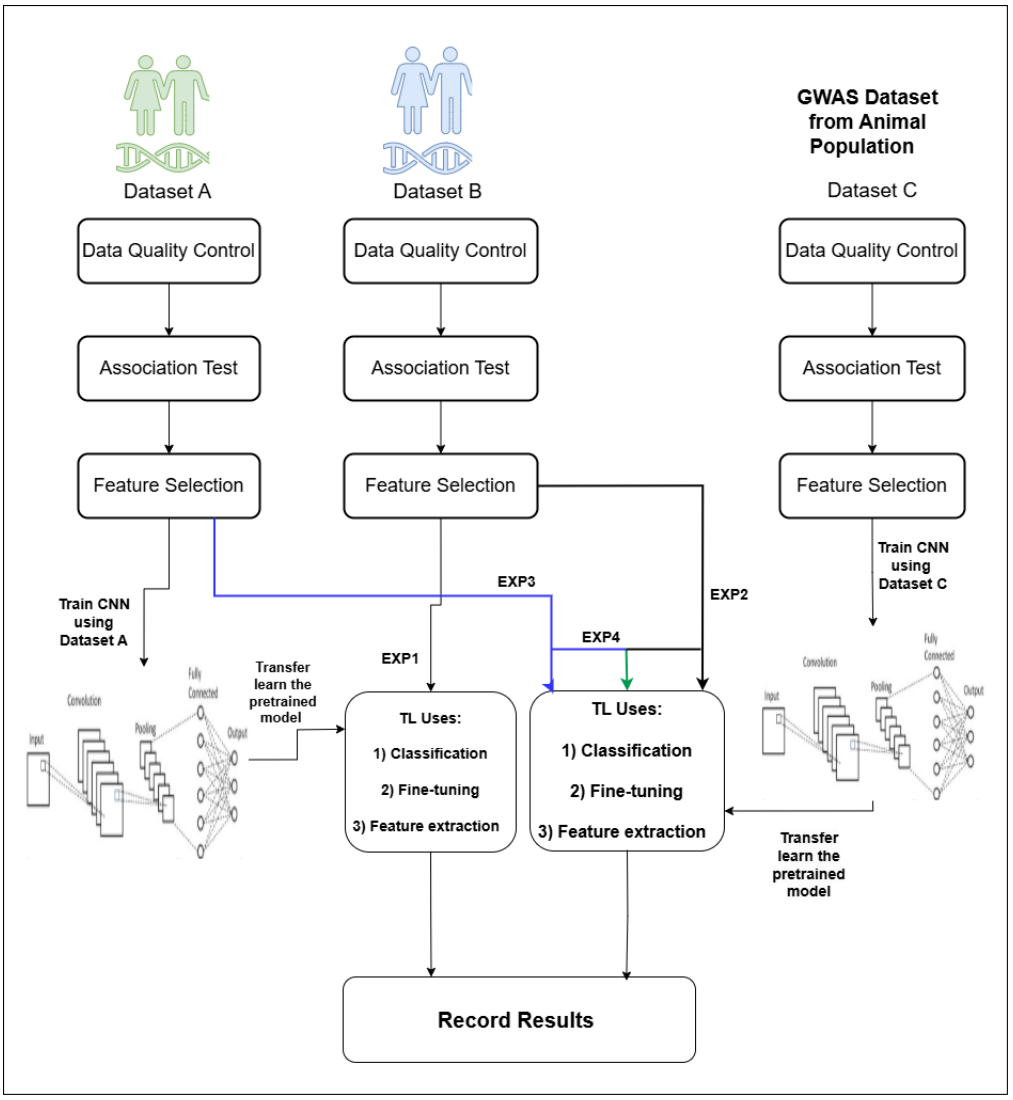


FIGURE 4.5: The proposed Transfer Learning Framework. On left side, quality control and feature selection are conducted on human data (Dataset A), then a CNN model is trained on Dataset A as a base model to be transfer to Dataset B for EXP 1. On right side, a CNN is trained on animal data (Dataset C) as a base model to be transferred to both Dataset A and Dataset B for EXP 2,3 and 4.

4.4 Results and Discussions

4.4.1 Evaluation Criteria

In this study, a deep TL model was trained to distinguish between healthy and LOAD subjects using GWAS data. To evaluate the performance of the proposed approach, standard metrics were employed, including accuracy, precision, recall, F1 score [157], and AUC [158], using a held-out testing set comprising 20% of the patients in all experiments. The results presented in the following sections are reported based on the testing set.

4.4.2 Transductive Transfer Learning Based AD classification (EXP1)

A CNN model is trained and tested on Dataset A in EXP1. The pre-trained model was saved for TL purposes, so that it could be reused in the target domain, Dataset B. The pre-trained model was first used after training only the fully connected layers to classify the samples in Dataset B. Then unfroze pre-trained model's layers and trained the transfer model on 80% of the observations in Dataset B, and tested on the remaining observations in Dataset B. Finally, the fine-tuned model was used as a feature extractor and served as an input to ML classifiers (i.e., SVM and RF in this case).

The results obtained during this experiment are presented in Table 4.15, which shows that the highest accuracy (89.04%) and F1 score (88.57%) were achieved by customising the pre-trained model as a feature extractor and feeding it into an SVM with rbf kernel. However, utilising the pre-trained model for the prediction task did not generalise well to the target dataset and showed a significant decrease in accuracy up to 39%. Despite the drop in accuracy, the model achieved a high recall score compared to other models in EXP1. This suggests that the accuracy metric alone is not enough to evaluate the true performance of a model, especially in cases of biasedness towards a specific class.

It should be noted that the choice of kernel type also influenced the model's performance. There was an improvement of around 2% in both accuracy and F1-score when using rbf kernel compared to a linear kernel. Similarly, more balanced performance was achieved using FE+SVM with rbf kernel in terms of precision and recall, which was not the case for other models.

TABLE 4.15: Results of EXP1 Transductive Transfer Learning (Transfer from Dataset A to Dataset B)

Model Use	Accuracy	Precision	Recall	F1-score	AUC
classification	0.39	0.43	0.75	0.55	0.42
Fine-tuning	0.76	0.80	0.69	0.74	0.76
FE+RF	0.89	0.96	0.80	0.87	0.89
FE+SVM with linear Kernel	0.87	0.93	0.80	0.86	0.87
FE+SVM with rbf Kernel	0.89	0.91	0.86	0.88	0.89

4.4.3 Inductive Transfer learning Based AD Classification (EXP 2, 3 and 4)

In EXP2, the source dataset used is GWAS data of animals to train a CNN model to classify the goat into 11 subcontinental breeds. The pretrained model, as in EXP1, adapted to classify the samples of target dataset (Dataset B) by a) only changing and training the top layer, b) fine-tune the model to make them relevant for the target task, c) as a feature extractor. The detailed statistical outcomes of this experiment are shown in Table 4.16. Similar to EXP1 results, the pre-trained model generalised well when fine-tuned and used as feature extractor followed by an SVM with rbf kernel. However, the model shows accuracy of 60.27% when used directly (without fine-tuning the pre-trained model's layers) to predict the class in the target dataset. This is a significant drop in model's performance which clearly indicates the usefulness of fine-tuning of TL for the task of AD classification. Even though a high accuracy of 84% achieved after fine-tuning and utilising the pre-trained model to classify samples in Dataset B, there is clearly a biased performance in terms of precision (93%) and recall (75%) metrics which shows biasedness towards one class. In construction, balanced performance of 87% and 80% for precision and recall was achieved when customising the pretrained model as feature extractor followed by SVM.

In EXP3, the same pre-trained model from EXP2, is used for the TL over Dataset A. The main objective is to investigate if the pre-trained model is able to generalise well for different datasets. Following the same TL strategies, Table 4.17 lists the statistical results from EXP3. Unlike the outcomes from EXP2, the pre-trained model did not perform well in general, however, achieved better recall scores than precision. After fine-tune the model to make it more relevant to Dataset A, 67.5% and 59.37% of accuracy and f1-score were achieved, respectively. These statistical outcomes clearly indicate that employing the pre-trained model as a feature extractor could not help in improving the model performance in this experiment.

Similar to outcomes from EXP1 and EXP2, the rbf kernel outperforms linear kernel which may be due to the non-linear nature of the dataset.

In EXP2 and EXP3, the pre-trained model is reused in target domains of Dataset A and Dataset B individually, to examine the generalisation of pre-trained model on both datasets. Furthermore, the pre-trained model utilized in both EXP2 and EXP3, is

Table 4.16: Results of EXP2 (Transfer from Dataset C to Dataset B)

Model Use	Accuracy	Precision	Recall	F1-score	AUC
classification	0.60	0.60	0.55	0.57	0.6
Fine-tuning	0.84	0.93	0.75	0.83	0.85
FE+RF	0.80	0.84	0.75	0.79	0.81
FE+SVM with linear Kernel	0.76	0.78	0.72	0.75	0.78
FE+SVM with rbf Kernel	0.84	0.87	0.80	0.84	0.85

employed as base model to be fine-tuned over aggregated dataset of A and B. the main intention is to investigate how the pre-trained model will behave in varying settings. Table 4.18 demonstrate the results achieved through experiment (EXP4). The statistical outcomes show that the accuracy dropped to 58% when customising the pre-trained model over the aggregated dataset. This could be due to the effect of Dataset A, as the model in EXP3 did not perform very well. Similar to EXP2, the model achieved highest performance of 69.28% and 64.66% of accuracy and f1-score, respectively, when fine-tuned over the aggregated dataset and used as a feature extractor followed by an SVM.

Table 4.17: Results of EXP3 (Transfer from Dataset C to Dataset A)

Model Use	Accuracy	Precision	Recall	F1-score	AUC
classification	0.58	0.41	0.4642	0.44	0.56
Fine-tuning	0.67	0.52	0.67	0.59	0.68
FE+RF	0.63	0.48	0.5	0.49	0.61
FE+SVM with linear Kernel	0.62	0.47	0.57	0.51	0.61
FE+SVM with rbf Kernel	0.65	0.5	0.53	0.51	0.62

Table 4.18: Results of EXP4 (Transfer from Dataset C to aggregated dataset of Dataset A and dataset B)

Model Use	Accuracy	Precision	Recall	F1-score	AUC
Prediction	0.58	0.57	0.38	0.46	0.57
Fine-tuning	0.66	0.65	0.54	0.59	0.65
FE+RF	0.64	0.63	0.51	0.56	0.63
FE+SVM with linear Kernel	0.66	0.67	0.52	0.59	0.67
FE+SVM with rbf Kernel	0.69	0.68	0.61	0.64	0.69

4.4.4 Benchmark with Related Work

Table 4.19 presents performance comparison between the proposed TL based AD classification approach and related works from the literature. It can be noticed that the approach in the current study outperforms the existing methods in terms of almost all performance metrics with an increase of 5% of accuracy and AUC, and 8% increase in F1-sore comparing to the second best model. In addition, it is very important to note that the proposed approach uses only 60 features as input to ML model as compared to state of the art [49] which uses over 500 features. This caused the proposed model to be less noisy, light weight, and efficient model. Furthermore, identification of fewer most contributing feature to AD might be useful to set a baseline for further analysis and future research direction. The proposed model shows high accuracies, as well as balanced performance in terms all metrics. In contrast, gradient boosted decision tress [54] showed an increase of 11% in terms of AUC comparing to other metrics.

TABLE 4.19: Comparison of related work in the literature

Study	**ML model**	**Dataset**	**Feature No.**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**AUC**
[54]	gradient boosted decision trees	UK-BioBank	145	80%	80%	80%	80%	91%
[117]	1D CNN	ADNI	4000	75%	-	-	-	81%
[48]	Ensemble of several ML algorithms	ADNI	2500	∼70%	-	70%	-	72%
[49]	LASSO	NIA-LOAD	501	84%	-	82%	-	84%
Propsed	Transfer Learning + SVM	ADNI	60	89%	91%	86%	88%	89%

4.4.5 Discussions

The genetics of phenotypes such as AD is of complex nature. Multiple genetic markers play a role in the emergence of complicated human disease. Despite the fact that GWAS were successful in identifying SNPs associated with complex diseases, this strategy lacks

the identification of variants with low influence that might play a significant role when combined with other variants [159]. Additionally, traditional GWAS have only discovered SNPs that can only account for 33% of the estimated 79% [160] of genetic risk related with AD.

ML algorithms have been shown to be more effective in discovering candidate SNPs and predicting complicated genetic diseases [161] [162] [163]. In the last decade, the application of ML-based techniques for genetic-based precision medicine has expanded and it is expected to continue [164].

The results shown in Table 4.15 demonstrate that TL can be used as an effective tool for classifying GWAS data. This is due to the fact that DL models require a large amount of data for training. Given the high dimensionality of GWAS data, training DL models is challenging, and TL from one dataset to another can help to address this issue. However, careful selection of the source dataset for the pre-trained model plays a major role in determining the model's performance when it is transferred to another dataset.

As shown in EXP1 (Table 4.15), TL in a similar task, from Dataset A to Dataset B, yielded the best performance. Despite the slight differences in population type between the two datasets (Dataset A consisted of European participants, whereas Dataset B contained non-Hispanic participants), the pre-trained model generalised well on the target dataset. Since the majority of GWAS data comprises European participants [165], this will pave the road for research of minor population. Particularly, where limited GWAS data exists, proposed approach might be effective to be used. When using a GWAS data from animal population in light of the similarities in biological function among species [166], TL the pre-trained model was effective in classifying participants in Dataset B, but did not perform well on Dataset A (Table 4.17). This may be because of the selection of genotyping platform, as the data is known to be influenced by the selection strategy and number of markers generated by genotyping platforms [147]. The genetic modification of animal models to express human disease-related genes or mutations does not eliminate their fundamental differences from human genomic architecture and biological pathways. The findings become less transferable because of this limitation. The polygenic nature of AD together with its complex traits requires animal models to replicate numerous gene-environment interactions which might not fully match human conditions. Some genetic loci which are linked to human AD risk do not exist in animals or function differently which makes it challenging to study gene-environment interactions and predict treatment responses accurately. This requires more investigation to clarify why the

model was able to perform well on one dataset but not on another similar one. Results shows that the classification accuracy was reduced when the pre-trained model used for aggregated dataset. This suggests that the pre-trained model may have failed to learn GWAS-specific features, and instead, relying on dataset-specific features.

The use of TL allows researchers to use information from a similar domain but it bears from overfitting issues especially when the source and target datasets have different distribution patterns and feature representations as demonstrated in the previous experiments. Overfitting occurs when the transferred features fail to capture the essential characteristics of the target domain.

The datasets high quality helped minimise the risk of overfitting in our analysis. The target dataset contained enough information with training and evaluation subsets that were balanced. The implementation of standard techniques including early stopping helped decrease the possibility of overfitting. The results require careful application to external datasets because population structure and genotyping platform differences may affect their validity.

Three TL customisations were used for the classification of AD. Except for one experiment (EXP3), the other three experiments demonstrated that the pre-trained model, utilised as a feature extractor followed by ML model, outperformed the other customisations. In only one experiment (EXP3), fine-tuning the pre-trained model had better accuracy than the other two strategies. This implies that re-purposing previously learned feature maps (from the source domain) for the target dataset can help achieve better performance with TL.

It was also observed that RF was capable of selecting SNPs that have been previously linked to AD. Therefore, SNP selection based on RF could be a valuable tool for identifying clinically significant risk factors. These findings are consistent with previous studies that have shown that the APOE 4 gene is the primary risk factor for AD [167].

For highly accurate clinical diagnostic, the genetic component alone forms a barrier. Complementing the genetic-based approaches with imaging or clinical data could be one of the possible answers to this challenge. The genetic study might be used to identify subjects who are at a higher risk of acquiring AD and therefore, such subjects can be tracked with imaging technology on regular basis to detect the disease's onset as soon as feasible.

Alongside the proposed study's contributions, small sample size of dataset limits this study; increasing the sample size is expected to increase the forecasting performance of

the deep TL models. As a result, these models are predicted to have a potential for diagnosing AD and other complex diseases.

4.5 Chapter Summary

The outcomes of utilising TL followed by the SVM, to estimate the risk of acquiring LOAD entirely from genetic variation data, were presented in this chapter. The feature selection methodology utilised to decrease the large number of SNPs has the potential to lead to the discovery of new disease-related genetic markers. Based on the preliminary results, the proposed methodology is expected to be a robust tool for the classification of AD. Furthermore, this chapter demonstrates that TL is an effective method for analysing and leveraging a large number of genetic markers that might be utilised for a variety of complicated disorders, such as Alzheimer's. In this study, transductive TL is utilised as a feature extractor, which resulted in the highest classification performance when compared with other settings and customisations of TL. Next chapter will utilise various neural networks architectures for classification of AD cases and NC.

Chapter 5

Wide and Deep Learning Based Approaches for Classification of Alzheimer's Disease Using Genome-Wide Association Studies

The research defined in this chapter has been published in PLoS ONE.

Alatrany AS, Khan W, Hussain A, Al-Jumeily D, for the Alzheimer's Disease Neuroimaging Initiative (2023) Wide and deep learning based approaches for classification of Alzheimer's disease using genome-wide association studies. PLoS ONE 18(5): e0283712. https://doi.org/10.1371/journal.pone.0283712

5.1 Introduction

In recent years, numerous computational approaches have been developed to enhance the diagnosis or uncover novel gene candidates linked to AD. For example, GWAS studies [168] are widely recognised for identifying genomic regions associated with various complex diseases and traits. These studies analyse data from large population samples, examining a vast number of loci (over 100,000 SNPs) across the human genome. Variations at certain loci can result in altered biological functions, potentially leading to disease. These variations are identified by analyzing genotypes from individuals with and without the characteristic of interest [169].
The literature highlights various methods for evaluating SNP susceptibility in GWAS, with each SNP assessed independently [170]. However, it has been observed that only a small fraction of SNPs significantly influence complex disease traits, while most exhibit low individual penetrance [171]. Conversely, many common human diseases are associated with complex interactions among multiple SNPs, known as multi-locus interactions [172].

Beyond traditional methods for GWAS analysis, ML algorithms have been increasingly applied to identify SNPs linked to various diseases. ML techniques have shown exceptional adaptability in handling non-linear problems and high-dimensional datasets, making them well-suited for the complexities of GWAS data analyzed in this study. The literature identifies three primary applications of ML in the context of GWAS [173].

First, ML models have been developed to classify disease cases and healthy controls [174] [48] [52]. Second, ML techniques have been employed to uncover novel genetic markers associated with specific diseases, such as AD [175] [51] [176]. Third, ML has been used to identify interactions between SNPs that contribute to the development of common human diseases [177] [53] [47].

The primary objective of integrating ML in these studies is to create predictive models that achieve optimal classification accuracy between cases and controls. However, a significant challenge persists: managing the computational complexity posed by the vast number of markers in GWAS data relative to the smaller sample sizes (i.e., data records) available [173].

In [178], a study introduced iGnet, a DL model for AD classification that integrates two datasets containing MRI and genetic information. The model employs a computer vision approach to process MRI scans and natural language processing to analyse genetic data. The proposed method was tested on the ADNI dataset, achieving a classification accuracy of 83.78% using MRI data combined with selected SNPs from chromosome 19. In contrast, the proposed novel approach utilises wide and deep learning models to classify NC and AD individuals. The process begins with an association test to identify the most significant SNPs related to the disease, followed by a hybrid feature selection technique to significantly decrease the number of features. The newly introduced method is then employed to select neighboring SNPs, creating a final set of SNPs for training the wide and deep learning classification models for NC and AD subjects. The key contributions of this work in this chapter include:

a. Developing a hybrid dimensionality reduction approach towards identification of the most distinguishing features, leading to robust classification performance.

b. Propose a neighbour SNPs selection approach to test the impact of neighbour SNPs over the classification accuracy.

c. Propose a wide and deep learning models for classification of individuals into NC and AD.

d. Extract human understandable rules from the trained ensemble model, to serve for the ML model's interpretability.

In this chapter the materials and methods of this study will be presented. As well as the experimental design. Finally, the results corresponding to the experimental design along with discussions about the chapter outcomes will be shown.

5.2 Materials and Methods

The proposed method for AD classification integrates data processing, feature selection, and ML algorithms. Initially, quality control is applied to ensure that only high-quality features and samples are included. Then, logistic regression is used to evaluate the association between each feature and AD. The processed dataset is then passed through a hybrid feature selection approach that combines PCA and Boruta algorithms. The

selected features are subsequently used to train machine learning models for AD classification. Figure 5.1 illustrates the overall methodology for the proposed AD classification

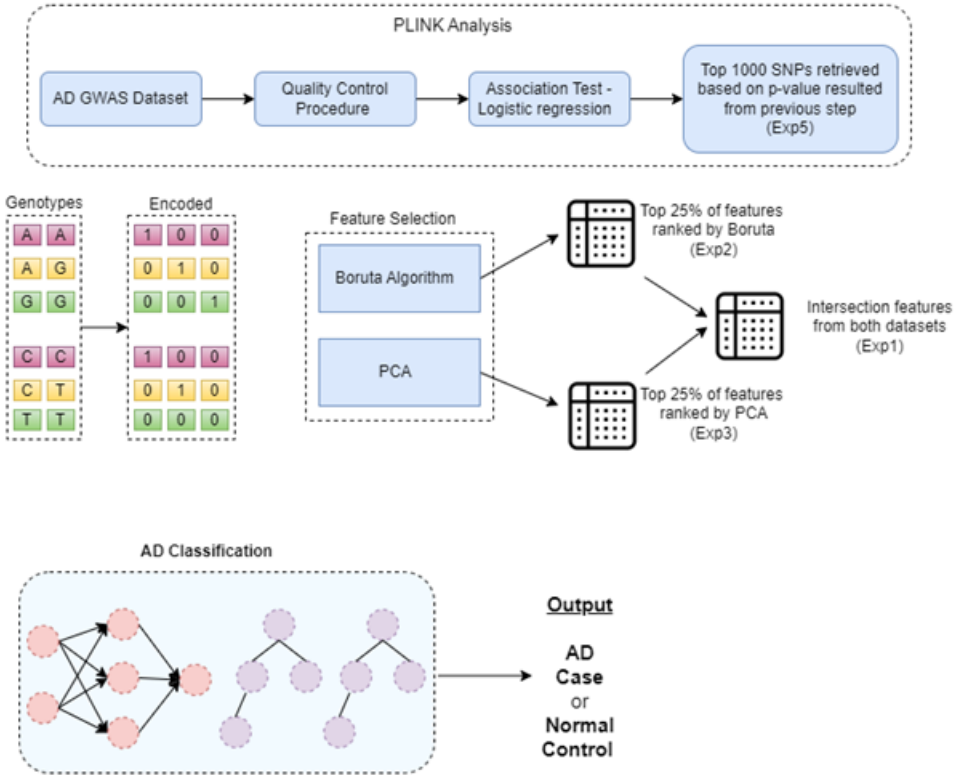


FIGURE 5.1: A graphical representation of proposed approach for AD and NC classification. First block represents the PLINK analysis in which quality control procedure and association test is conducted. Second the genotype data convert into one-hot representation. Third feature selected utilizing Boruta and PCA algorithms. Finally, AD classification is performed using the different feature sets.

5.2.1 ADNI Dataset

The dataset used in this study is obtained from the Alzheimer's Disease Neuroimaging Initiative database. The ADNI was launched in 2003 as a public-private partnership with the primary objective to test whether serial MRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. The dataset is described in Chapter 4 of this thesis.

5.2.2 Quality Control

To eliminate irrelevant information from both genetic markers and samples, various techniques have been explored and applied in genetic data quality control, particularly focusing on SNP data. The methods outlined in Chapter 4 represent best practices for removing individuals and SNPs that could introduce bias or lead to false positive results [136]. The dataset prepared in this study is refined to include a representative set of SNP features and subjects that are more likely to reveal genetic signals associated with the phenotype by removing subjects and SNPs that do not meet the criteria established by these procedures. Initially, there were 620,901 SNPs, which were reduced to 487,037 SNPs following the operations outlined in Table 5.1.

TABLE 5.1: Quality control procedure applied for both samples and genetic markers

Filtering approach	**Threshold Used**
SNPs missingness	0.02 genotyping rate
Individuals' missingness	0.2 genotyping rate
Sex discrepancy	An estimate of the X chromosome homozygosity >0.8 for males and <0.2 for females.
Autosomes Chromosomes	-
Minor allele frequency	0.05 due to sample size.
Hardy–Weinberg equilibrium	SNPs are first filtered out within the controls for HWE p-values of 1e-6, then in cases for HWE with p-value of 1e-10.
Relatedness	employ 0.2 pi-hat threshold.
Population stratification	Only non-Hispanic European participants chosen.

5.2.3 Feature Selection

In the feature selection and dimensionality reduction process of the proposed methodology in this chapter, an association test was performed using logistic regression (as described in section 4.3.3 Association Test) to assess the relationship between each SNP and AD. The top 1000 SNPs, ranked by their significance values (i.e., p-value), were selected for further analysis. These 1000 SNPs were then processed using a combination of feature selection techniques, including PCA [97] and the Boruta algorithm [101], both of which have been applied in various similar fields [99] [100].

5.2.3.1 Principal Component Analysis

The top-ranked 50 features (out of 1000 SNPs) selected by the PCA algorithms (as most important) are shown in Fig. 5.2, including rs12498138 located on gene GOLGB1, rs4072374 located in gene RNASEH1, rs2309772 in TENM3, rs7005164, and gene LOC105375901.

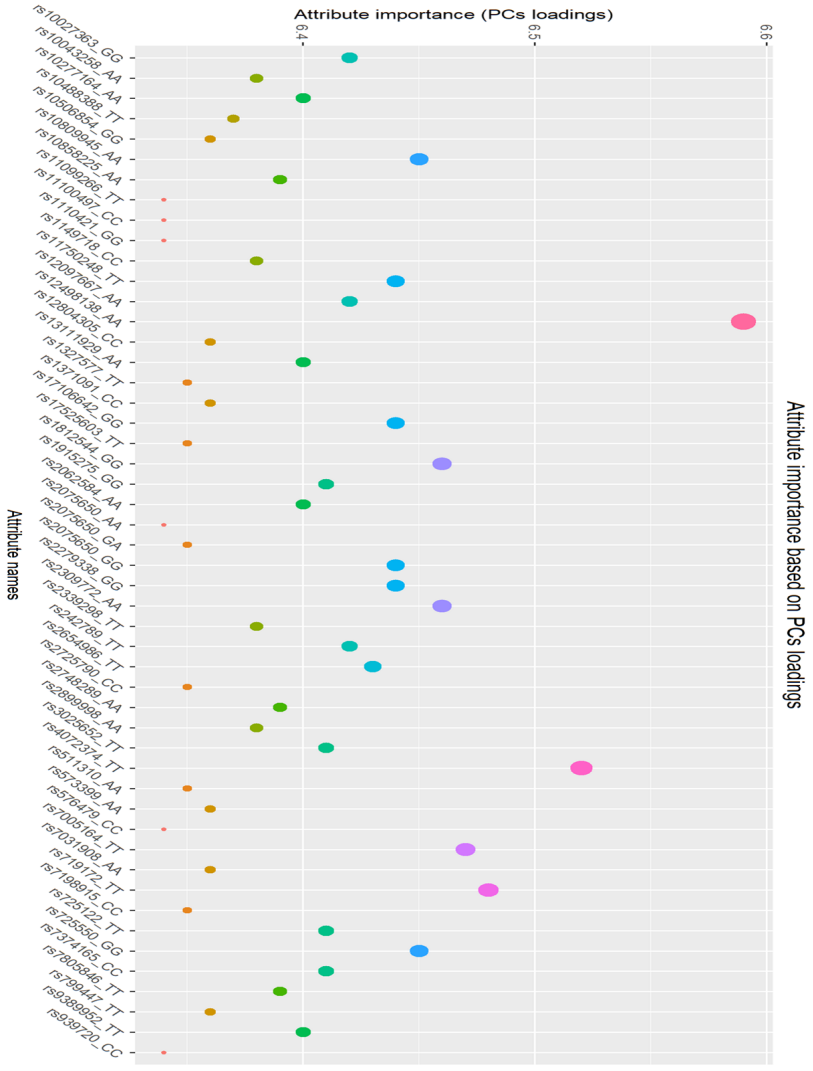


FIGURE 5.2: Top-ranked 50 features (out of 1000 SNPs) selected as important, by the PCA algorithm

5.2.3.2 Boruta Algorithm

SNPs with substantially high scores identified by the Boruta algorithm includes: rs17365991 gene TEF, rs8141950 gene PARVB, rs2654986 gene LUNAR1, and rs2036109 gene ADRA1A. A list of top 50 important features selected by the algorithm is presented in Table 5.2.

Table 5.2: Top 50 features selected by Boruta algorithm

rs6116375_CC	rs11768384_GG	rs6585082_TG	rs11706690_CC	rs10491109_AC
rs17365991_GG	rs2075650_AA	rs6505403_TT	rs327079_TT	rs12670401_CC
rs8141950_CC	rs7342676_CC	rs1927605_AA	rs10790928_TT	rs2208322_AA
rs2654986_TC	rs12822144_AA	rs6585082_GG	rs2654986_CC	rs10871809_TC
rs2036109_TT	rs11253696_GG	rs327079_CT	rs4795895_GG	rs1387089_TT
rs12804305_CT	rs4351677_TC	rs10491109_CC	rs6577539_GA	rs1233651_AA
rs10133989_AA	rs4351677_CC	rs2883782_TT	rs11160481_CC	rs3775770_AA
rs17365991_AG	rs12804305_TT	rs4778636_AG	rs1797779_CT	rs3004297_TT
rs8141950_TC	rs3857224_TT	rs7146951_GG	rs7159863_TT	rs1387089_CT
rs2042599_GG	rs4964453_TT	rs1981542_GT	rs4635275_AA	rs4566279_CT

5.2.3.3 Hybrid Feature Selection

Although both PCA and Boruta algorithms are widely used for feature selection, their underlying mathematical principles differ. By combining the results from both methods, the goal is to eliminate as many irrelevant features as possible while preserving the most important information from the original dataset. A hybrid feature selection strategy was employed, combining the outputs of Boruta and PCA. First, the results from both algorithms were sorted based on feature rankings, reflecting their importance. Then, the top 25% of features identified by both Boruta and PCA were selected, resulting in 121 key features. Boruta uses a wrapper method around random forest to identify features for classification that are based on their predictive power, thus capturing non-linear interactions and features relevant to classification performance. PCA on the other hand ranks features based on variance explained, regardless of the outcome label, thus capturing the informative structure of the data. Thus, the aim to retain both the statistical relevance and the biological signal. Instead of using PCA's transformed components, which can reduce the interpretability, the component loadings were used to assign an importance score to the original features (SNPs). This allow to rank the SNPs directly from PCA, which made them more comparable to Boruta's feature importance scores.

PCA and Boruta can rank features differently, for instance, PCA may select a SNP for high variance and Boruta may not consider it as a relevant feature because of its weak predictive power. However, by cross checking both rankings, to see weather able to identify features that converged SNPs that were important from both structural and predictive viewpoints might improving robustness of the selection.

The choice to retain the top 25% of features from each of the methods was based on

previous empirical benchmarks and the trade-off between dimensionality reduction and classification accuracy. Although the threshold is somewhat arbitrary, it resulted in a reasonable number of features that could be used for downstream analysis. A full list of the commonly selected features is provided in Table 5.3. It is evident that some of the highest-ranked SNPs, such as rs6116375 on the PRNP gene and rs2075650 on the TOMM40 gene, are strongly associated with AD.

TABLE 5.3: List of final feature-set identified as significant using the intersection of selected features from both PCA and Boruta algorithm

rs6116375_CC	rs10176603_TT	rs7747741_GG	rs4290760_CC	rs16864809_TT
rs2654986_TC	rs10031325_CC	rs701880_CC	rs11680332_GG	rs7679260_CC
rs11768384_GG	rs16889565_GA	rs9296691_TC	rs628482_GG	rs9389952_TT
rs2075650_AA	rs2877347_CC	rs4953672_CC	rs518385_TT	rs10804812_CC
rs7342676_CC	rs6114605_GA	rs10068900_GG	rs2577322_CC	rs618236_CC
rs4964453_TT	rs7618348_CC	rs2834714_TT	rs11869174_CT	rs1945624_AA
rs10790928_TT	rs9595108_CC	rs6838005_CC	rs11733633_AA	rs2577322_TT
rs2208322_AA	rs17068548_GG	rs10514486_CC	rs911892_TT	rs7807731_TT
rs7519796_AA	rs13211072_TT	rs7149949_TT	rs3812568_AA	rs2136613_TT
rs10222715_TT	rs6132022_TT	rs2725790_CT	rs799447_GG	rs344783_TT
rs10793982_TT	rs793291_AA	rs11655031_TT	rs17745021_CT	rs1495813_CC
rs775879_GG	rs3771389_CT	rs2833427_CC	rs13245564_GG	rs9410486_GG
rs4837137_AA	rs6695731_CC	rs8007000_TT	rs2305252_AA	rs7096762_AA
rs1789250_AA	rs10044783_CC	rs17430865_CT	rs4472075_AA	rs2309777_GG
rs4868468_AA	rs17345545_CC	rs3815360_CC	rs4793902_TT	rs9515168_GT
rs11752811_TT	rs871049_CC	rs17430865_TT	rs168825_GG	rs6569364_AA
rs2075650_GG	rs4953672_AA	rs11922179_AA	rs6838005_TC	rs12988856_TT
rs2697303_AA	rs2075650_GA	rs1186685_TT	rs775879_AA	rs1891265_GG
rs362584_AA	rs1479884_GG	rs7320494_AA	rs6903956_AA	
rs8000805_GG	rs11253696_AA	rs7206002_GG	rs12480224_AA	
rs10879839_TT	rs13135230_GG	rs367369_TT	rs2339298_TT	
rs2286343_AA	rs10888578_TT	rs1328179_TT	rs7413155_AC	
rs939720_CC	rs7999171_GG	rs4689705_TT	rs9595108_AC	
rs7165661_TT	rs12312628_CC	rs705904_CC	rs6929400_CC	
rs2867922_TT	rs10101666_TT	rs9381936_CC	rs268909_TT	

The aforementioned features (PCA, Boruta, and composite of both) are then used to train and validate the multiple ML models for the task of AD classification over unseen instances.

5.2.4 Proposed Alzheimer's Disease Classification

Once the most relevant features were identified from the original dataset, various well-established classification methods, such as RF and ANN, were applied to classify AD.

To ensure accurate and efficient classification, different combinations of features were tested with the selected models. This approach not only improved the classification of AD but also helped identify the most significant set of features.

5.2.4.1 Random Forest for Proposed AD Classification

RF is an ensemble learning model known for its effective for high-dimensional datasets. Ensemble learning is a powerful technique that combines multiple learning algorithms to enhance overall accuracy. One of the key benefits of ensemble methods is their ability to address the challenge of small sample sizes by averaging and integrating multiple classification models, thus reducing the risk of overfitting the training data. This makes the training dataset more efficient, which is particularly valuable in biological applications where sample sizes are limited. Figure 5.3 illustrates an example of trees that generate multiple results, each using a different subset of features (with bootstrapped data samples) from the proposed RF-based AD classification model.

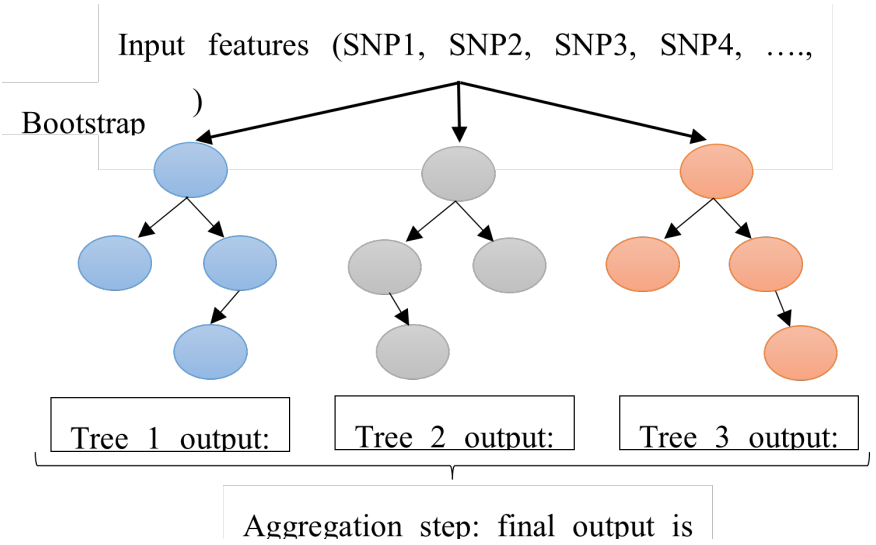


FIGURE 5.3: Random Forest sub-trees for proposed AD classification using GWAS data. The input to the RF is the bootstrapped SNPs features. In the first step (bootstrap step) refers to the process of training each tree in RF on a subset of the training samples. While in the second step (aggregation step) the class with the majority votes from the trees is chosen as the final output (in above example 2/3 votes are in favour of Normal control)

5.2.4.2 Deep Wide Artificial Neural Networks for Proposed AD Classification

Neural networks were applied with gradient descent optimisation using the backpropagation learning method for binary classification tasks. The neural network architecture comprises input, hidden, and output layers, each containing a predefined number of neurons. In this study, different types of neural network structures are used: a Wide Neural Network (WNN), which has a single hidden layer with a large number of neurons, and a Deep Neural Network (DNN), which features multiple hidden layers, each with fewer neurons.

Expanding on the concept of ANNs, a wide and deep neural network (as shown in Fig 5.4) combines a DNN with a linear model based on a limited set of features. This architecture has proven beneficial in similar applications, such as cell type classification [179] and recommender systems [180]. GWAS data characteristics including its high dimensionality and sparsity together with biological linkage disequilibrium (LD) structure influences how ML architectures should be designed for predicting phenotypes or classifying diseases such as AD. The combination of wide and deep neural networks presents a proposed solution for this situation. The proposed architecture benefits from deep learning capabilities for complex pattern detection in large datasets and linear model advantages for feature memorisation and interpretability.

The deep network part utilised in GWAS data because it can analyse neighbouring SNPs to detect nonlinear genomic interactions. The model requires this capability because AD result from polygenic effects and subtle interactions which linear models cannot detect independently. The wide component of the network models a selection of biologically important SNPs which were identified during previous feature selection processes (e.g., those in Table 5.3). These SNPs demonstrate established connections to AD while also functioning as regulatory elements or coding sequences. The wide layer enables the model to focus on important genetic markers which enhances both biological understanding and model clarity. The dual-pathway structure utilises GWAS data heterogeneity to combine strong disease-linked variant detection with the ability to discover new nonlinear patterns in large genomic datasets.

For the proposed AD classification (shown in Fig 5.4), the final set of selected features

(Table 5.3) is passed into the wide component. For each SNP listed in Table 5.3, neighboring SNPs are identified and used as input to the deep component of the network.

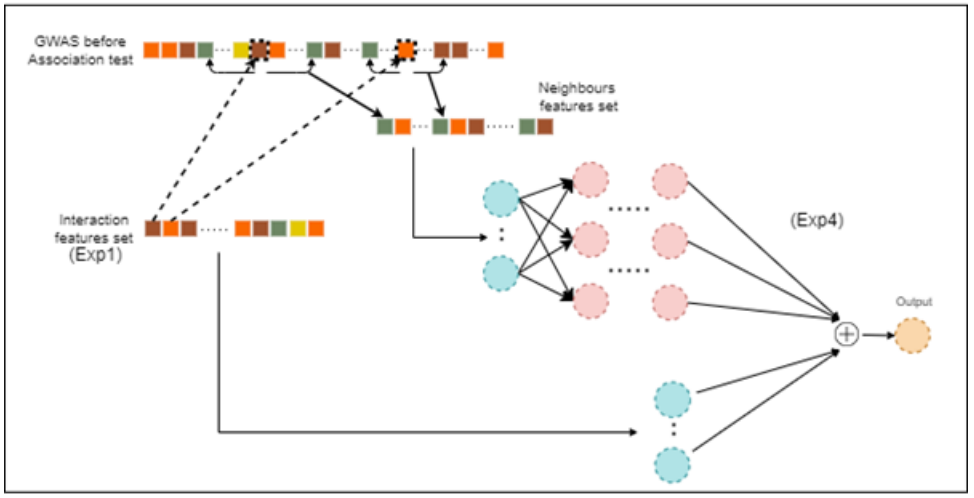


FIGURE 5.4: Proposed Wide and deep NN for AD classification using GWAS Data

5.2.5 Experiment Design

Several experiments are conducted using the features identified through the proposed hybrid feature selection method (refer to Section 5.2.3 Feature Selection) from the ADNI GWAS dataset to train the AD classifiers (RF, WNN, and DNN). The dataset is split into training (70%) and testing (30%) sets. To ensure an accurate and consistent evaluation, 5-fold cross-validation (5-CV) is applied for performance assessment of the AD classifiers. The optimal hyperparameters for all machine learning classifiers are determined through a trial-and-error approach and are outlined in Table 5.4. Quality control and association testing are performed using PLINK software [181], while the ML models are implemented using the Scikit-learn library in Python [154]. The PyPlink library is used to process genotype data in Python [155], and neural network implementation is carried out using Keras and TensorFlow as backends [156]. Based on these configurations and feature sets, the following experiments are conducted in the proposed study:

Experiment 1 (EXP1): The intersection of the top 25% features ranked by both Boruta and PCA algorithms is used as the combined feature set (called Intersection feature set) to train RF, WNN, and DNN classifiers to determine the best-performing AD classifier.

Experiment 2 (EXP2): The top 25% of features ranked by Boruta are selected as the feature set (called Boruta feature set) for AD classification using RF, WNN, and DNN algorithms.

Experiment 3 (EXP3): The top 25% of features ranked by PCA are selected as the feature set (called PCA feature set) for AD classification using RF, WNN, and DNN algorithms.

Experiment 4 (EXP4): To assess the impact of neighboring SNPs, for each SNP in the interaction feature set (from EXP1), the SNP and its six neighboring SNPs (three from each side) are retrieved to create a new feature space called the neighboring features set. The features from EXP1 are used as input to the wide component, and the neighboring features set is used as input to the deep component to train and test the proposed wide and deep model (as illustrated in Fig 5.4).

Experiment 5 (EXP5): The top 25% of features selected by logistic regression are used as the feature set (called original feature set) for AD classification using RF, WNN, and DNN algorithms.

TABLE 5.4: Paramter setting for ML models in experments 1,2,3,4 and 5.

EXP1							
RF	(n_estimators = 500, max_features='auto', max_depth=6, criterion='gini'						
2***WNN**	No of neurons and activation function in hidden layers				No of epochs	Learning rate	optimizer
	121 RELU				15	0.001	Adam
DNN	60 RELU	40 RELU	30 RELU	20 RELU	15	0.001	Adam
EXP2							
RF	n_estimators = 500, max_features='sqrt', max_depth=7, criterion='gini'						
2***WNN**	No of neurons and activation function in hidden layers				No of epochs	Learning rate	optimizer
	747 RELU				15	0.001	Adam
DNN	373 RELU	249 RELU	186 RELU	124 RELU	15	0.001	Adam
EXP3							
RF	n_estimators = 500, max_features='sqrt', max_depth=8, criterion='entropy'						
2***WNN**	No of neurons and activation function in hidden layers				No of epochs	Learning rate	optimizer
	747 RELU				15	0.001	Adam
DNN	373 RELU	249 RELU	186 RELU	124 RELU	15	249 RELU	186 RELU
EXP4							
		No of neurons and activation function in hidden layers			No of epochs	Learning rate	optimizer
2***WDNN**	Wide 100 RELU			15	0.001	Adam	
	Deep 2500 RELU	2000 RELU	1000 RELU	500 RELU	100 RELU		
EXP5							
RF	n_estimators = 600, max_features='sqrt', max_depth=7, criterion='entropy'						
2***WNN**	No of neurons and activation function in hidden layers				No of epochs	Learning rate	optimizer
DNN	1497 RELU	998 RELU	748 RELU	499 RELU	15	0.001	Adam

5.3 Results and Discussion

Using the experimental configurations described above, detailed results and performance metrics were obtained using the testing set. This study successfully identified and extracted a smaller yet highly effective set of features that significantly improved the classification of AD. Among these, several genes were found to be significantly associated with AD, aligning with findings from related literature. These include including rs6116375 on gene PRNP [182], rs2075650 on gene TOMM40 [183], rs10793982 on gene LAMC3 [184], rs2208322 on gene NEURL1 and rs7519796 on gene KAZN [185], This alignment with existing research highlights the effectiveness of the feature selection approach. Additionally, the study identified potential novel (SNPs) significantly associated with AD, including rs2654986 on the LUNAR1 gene and rs2208322 on the NEURL1 gene. A complete list of the significant SNPs identified in this study is provided in Table 5.3.
To evaluate the effectiveness of the feature selection process, a RF classifier and neural networks with varying parameter configurations are employed to classify the AD patients. The performance of the classifiers is presented in Table 5.5 when evaluated over the unseen subjects using features set described in EXP1. It can be noticed that regardless of selected ML model, high performance measures are achieved. WNN indicates an accuracy and f1-score of 94% and 93%, respectively. Followed by a DNN which showed a slightly decline in performance (i.e., 93%) in respect to accuracy. While RF indicate more deteriorations in performance with 89% accuracy and 88% F1 score, which is in line with the existing similar work [186], where higher accuracy is reported using ANN as compared to RF (for preterm birth classification). Oriol et al. [48] employed RF in classification of AD and NC using GWAS data, where they reported accuracy of 67% (significantly lower than proposed approach). Similarly, RF was not the best classifier to discriminate between AD cases and controls as reported in a similar work [54]. It is also important to note that the performance balance from WNN and DNN (in Table 5.5) as compared to RF, which indicates more biasedness towards the precision (96%) as compared to recall (81%).

Table 5.6 summarises outcomes for EXP2 where all classifiers indicated similar performance when trained and tested over the top-ranked (i.e., 1st quartile) features selected by Boruta algorithm. It can be noticed that the overall accuracy of each model is increased specifically, the WNN and DNN which indicate 99% accuracies for unseen

TABLE 5.5: Comparison of ML algorithms for classification of AD and healthy individuals using intersection features selected by Boruta and PCA from the top 25% (Exp 1).

Model	Accuracy	Precision	Recall	F1	AUC
RF	89%	96%	81%	88%	90%
Wide NN	94%	91%	98%	93%	92%
Deep NN	93%	89%	96%	92%	94%

instances. This clearly shows the effectiveness of selected features as well as the model's configurations.

Table 5.7 presents the outcomes for EXP3 where the features identified from PCA algorithm are used to train the ML models. It can be noticed that WNN and DNN models outperformed the RF producing overall 96% and 94% accuracies, respectively, as compared to 84% from RF. Likewise, the performance clearly indicates the balance between recall and precision which is not the case for RF. Overall, the RF demonstrated a notable reduction in performance.

TABLE 5.6: Comparison of ML algorithms for classification of AD and healthy individuals using top 25% features selected by Boruta algorithm (Exp 2).

Model	Accuracy	Precision	Recall	F1	AUC
RF	92%	99%	84%	91%	92%
Wide NN	99%	99%	99%	99%	100%
Deep NN	99%	99%	99%	99%	100%

TABLE 5.7: Comparison of ML algorithms for classification of AD and healthy individuals using top 25% features selected by PCA algorithm (Exp 3).

Model	Accuracy	Precision	Recall	F1	AUC
RF	84%	99%	68%	81%	84%
Wide NN	96%	99%	92%	96%	97%
Deep NN	94%	96%	91%	93%	97%

To assess the impact of the neighbouring SNPs (of the identified most important SNPs) towards the classification of AD, the performance of WDNN classifier was evaluated in EXP4 (Table 5.9). Despite the performance of WDNN is substantially reduced (around 80%) as compared to EXP1-EXP3, it is still inline or outperforms most of the existing related works as shown in Table 5.9, particularly in the domain of GWAS. For the final experiment, we tested the models' performances over the original dataset (EXP5 as illustrated on Figure 5.1) before feature selection (Table 5.8). It can be noticed that the classification performance from each model is nearly as accurate as in EXP2 (Table 5.6). Likewise, the RF indicates a biased performances in terms of precision and recall.

TABLE 5.8: Comparison of ML algorithms for classification of AD and healthy individuals using original features set (Exp 5).

Model	Accuracy	Precision	Recall	F1	AUC
RF	91%	99%	81%	89%	91%
Wide NN	99%	99%	98%	99%	99%
Deep NN	99%	99%	98%	98%	99%

5.3.1 Comparative Analysis

Finally, Table 5.9 compares the performance of proposed method with existing similar approaches, towards the classification of AD based on genome-wide data (SNPs). It is evident that the proposed approach outperforms the Decision tress [54], CNN [117], ensemble models [48], and LASSO [49]. The proposed approach shows stable performance throughout the evaluation metrics including ROC. Whereas, the decision tress utilised in reference [54] showed an increase AUC of 11% comparing to the model's accuracy. Likewise, our work shows the superiority of Boruta algorithm in selecting the optimal number of features and eliminating the redundant SNPs, which reflects the high performance in the classification task. The results indicate that Boruta algorithm is better than other feature selection techniques such as statical techniques applied in [49]. Moreover, the proposed model uses only 121 features as input to the WNN as compared to LASSO [49] which uses over 500 features, and CNN-based approach utilising 400 features [117]. This leads to a less noisy, lighter, and more efficient model. The identification of fewer contributing features to AD may be useful to set a baseline for further analysis and direction in future research.

TABLE 5.9: Comparison of related work from the literature.

Study	ML Model	Dataset	Feature selection	Feature No.	Acc	F score	Recall	Prec	AUC
[54]	Gradient boosted decision trees	ADNI	Previously reported SNPs related to AD from DiaGeNet database.	145	80%	80&	80%	80%	91%
[117]	1D CNN	ADNI	Divided the genome into nonoverlapping fragments, then used CNN to select segments. CNN was run on the selected fragments using a Sliding Window Association Test to identify important SNPs.	4000	75%				81%
[48]	Ensemble of ML models	ADNI	To find significant SNPs, used the statistical summary results from IGAP [23]. The top 2,500 SNPs were then chosen as the final feature set.	2500	~70		70%		72%
[49]	LASSO	NIA-LOAD	Using X2 with kinship correction	501	84%		82%		84%
Proposed Model 1	WNN (EXP1)	ADNI	See section 3.	121	95%	95%	99%	91%	94%
Proposed Model 2	WNN (EXP2)	ADNI	See section 3.	747	99%	99%	99%	99%	100%
Proposed Model 3	WDNN (EXP4)	ADNI	See section 3.	121 for wide component and 4697 for deep component	83%	83%	89%	79%	83%

5.3.2 Discussions

it should be noted that this study is first of its kind to examine GWAS data using a wide and deep neural network approaches as far as the knowledge of the author. Using a relatively small number of identified feature set (only 121 features), the proposed classifying models achieved high performance (Table 5.5), which reveals the robustness of the proposed feature selection methodology. Furthermore, experimental outcomes show that using appropriate classifier can improve the accuracy better than increasing the number of features (See Table 5.7). In addition to performance efficiency, experiments 1,2 and 3 show the strength of neural networks in the existence of complex relations within the dataset. The results demonstrate the effectiveness of the proposed approach (e.g., via the cross validations) which can be easily applied to other chronical disease where larger GWAS datasets are available.

Similar to other related studies, when interpreting the findings, some limitations are also noticed in the proposed work. First, the sample size is relatively small however, this is consisting with other related work that uses the same dataset [48] [187] [188] and other work which use GWAS data with a similar or lower sample size [115] [189]. Second, number of features (SNPs) highly exceeded the number of samples within the original dataset however, this was addressed by substantially reducing the number of features using advanced statistical approaches and highlighted the significant SNPs.

Experiments were conducted to compare the performance of WNN (one hidden layer with a large number of neurons) and DNN (multiple hidden layers with smaller number of neurons in each layer) to explore the implication that architecture selection has in the model performance. The ANNs have variety of parameters to choose from, including the number of hidden layers and neurons per layer. These parameters distinguish the network's architecture and influence how the model performs. It was noticed that in almost all of the experiments, WNN outperforms the DNN that may be because of the size and nature of the dataset.

Furthermore, it can be noticed that the WNN and DNN showed better performance than RF in GWAS domain (Tables 5.5,5.6 and 5.7). However, there is a trade-off between model accuracy and model interpretability. The RF can lead to an interpretable model and extract useful explanation on how the model reached a decision (case or control) which to go beyond simply using a model to get the best possible predictions. The RF model can produce insights which a human expert (e.g., physicians) can use to

understand how the model help in AD diagnosis through genetic data. For this purpose, a list of human understandable rules is extracted from the best performing tree of the RF model as shown in Table 5.10.

From the extracted rules, it can be notice that if a person has the genotype of CC

TABLE 5.10: Rules extracted from best tree of RF model

if (rs705904_CC >0.5) and (rs4953672_CC <= 0.5) and (rs799447_GG >0.5) and (rs701880_CC <= 0.5) then class: Control (proba: 100.0%) — based on 20 samples
if (rs705904_CC <= 0.5) and (rs2075650_AA >0.5) and (rs1789250_AA <= 0.5) and (rs939720_CC <= 0.5) and (rs268909_TT <= 0.5) and (rs7342676_CC <= 0.5) and (rs1479884_GG <= 0.5) then class: Control (proba: 100.0%) — based on 17 samples
if (rs705904_CC <= 0.5) and (rs2075650_AA <= 0.5) and (rs871049_CC >0.5) and (rs2577322_TT <= 0.5) and (rs8000805_GG <= 0.5) then class: Case (proba: 100.0%) — based on 14 samples
if (rs705904_CC <= 0.5) and (rs2075650_AA >0.5) and (rs1789250_AA <= 0.5) and (rs939720_CC <= 0.5) and (rs268909_TT >0.5) and (rs793291_AA <= 0.5) and (rs7342676_CC <= 0.5) and (rs11922179_AA >0.5) and (rs628482_GG <= 0.5) and (rs2577322_CC <= 0.5) and (rs1495813_CC <= 0.5) then class: Control (proba: 100.0%) — based on 10 samples
if (rs705904_CC <= 0.5) and (rs2075650_AA >0.5) and (rs1789250_AA <= 0.5) and (rs939720_CC <= 0.5) and (rs268909_TT >0.5) and (rs793291_AA <= 0.5) and (rs7342676_CC <= 0.5) and (rs11922179_AA >0.5) and (rs628482_GG <= 0.5) and (rs2577322_CC <= 0.5) and (rs1495813_CC >0.5) and (rs11680332_GG <= 0.5) and (rs9296691_TC <= 0.5) and (rs9515168_GT <= 0.5) then class: Control (proba: 100.0%) — based on 9 samples
if (rs705904_CC <= 0.5) and (rs2075650_AA <= 0.5) and (rs871049_CC <= 0.5) and (rs16864809_TT <= 0.5) and (rs1328179_TT <= 0.5) and (rs6116375_CC <= 0.5) and (rs4837137_AA <= 0.5) and (rs3771389_CT <= 0.5) then class: Case (proba: 100.0%) — based on 8 samples
if (rs705904_CC <= 0.5) and (rs2075650_AA <= 0.5) and (rs871049_CC <= 0.5) and (rs16864809_TT <= 0.5) and (rs1328179_TT >0.5) then class: Case (proba: 100.0%) — based on 6 samples
if (rs705904_CC <= 0.5) and (rs2075650_AA <= 0.5) and (rs871049_CC <= 0.5) and (rs16864809_TT <= 0.5) and (rs1328179_TT <= 0.5) and (rs6116375_CC >0.5) then class: Control (proba: 100.0%) — based on 6 samples
if (rs705904_CC <= 0.5) and (rs2075650_AA >0.5) and (rs1789250_AA >0.5) and (rs871049_CC >0.5) then class: Case (proba: 100.0%) — based on 5 samples
if (rs705904_CC <= 0.5) and (rs2075650_AA <= 0.5) and (rs871049_CC <= 0.5) and (rs16864809_TT <= 0.5) and (rs1328179_TT <= 0.5) and (rs6116375_CC <= 0.5) and (rs4837137_AA >0.5) then class: Control (proba: 100.0%) — based on 4 samples
if (rs705904_CC <= 0.5) and (rs2075650_AA <= 0.5) and (rs871049_CC >0.5) and (rs2577322_TT <= 0.5) and (rs8000805_GG >0.5) and (rs799447_GG <= 0.5) then class: Case (proba: 100.0%) — based on 4 samples

for SNP rs705904 and GG for SNP rs799447 or AA for SNP rs11922179, they are less likely to be diagnosed with AD. Furthermore, genotype of AA for SNP rs2075650 is highly associated with controls. On the other hand, a person with genotype AA for SNP rs1789250 or genotype other than AA for SNP rs2075650 is most likely to be a case of AD.

5.4 Chapter Summary

In this chapter, a reliable ML classifier to classify patient with AD and CN was proposed. Both of Boruta and PCA algorithms utilised as feature selectors to reduce the number of features and identify the most promising set of SNPs. detailed experiments were conducted, by training the ML models on different features subsets. Wide and Deep Learning approaches proposed for classifying AD and non-AD subjects. All models achieved high performance; WNN found to be the best classifier with a stable performance of 99% accuracy. The outcomes clearly demonstrate the effectiveness of proposed hybrid feature selection. Although the models used to classify AD patients it can be extended to other chronic disease. Next chapter will address the utilisation of interpretable ML for classification of AD.

Chapter 6

An Explainable Machine Learning Approach for Alzheimer's Disease Classification

The research defined in this chapter has been published in Scientific Reports.

Alatrany, A.S., Khan, W., Hussain, A. et al. An explainable machine learning approach for Alzheimer's disease classification. Sci Rep 14, 2637 (2024). https://doi.org/10.1038/s41598-024-51985-w.

6.1 Introduction

Although ML models have shown impressive results in a variety of medical applications, their black-box nature makes them difficult to use in real-world health situations. As a result, ML techniques in the clinical domain often do not employ sophisticated models, instead opting to use simpler, interpretable statistical models (e.g. linear) that are only capable of achieving limited accuracy [190]. In many studies, researchers have studied complex models and attempted to open the black box of their decision-making processes [191]. In the domain of AD, very few recent researchers have focused on the interpretability and explainability of ML models. To become acceptable and trusted by physicians, these models must be comprehensible, explainable, and traceable. Therefore, these models must explain how a specific medical decision or diagnostic task is achieved. A recent work presented in [192] used ML to investigate some factors reported to have an important impact on the occurrence and progression of AD. Their approach includes training the XGBoost model to discriminate different stages of the disease which reached a f1-score classification of 84%. A SHapley Additive exPlanations (SHAP) model is used on top of the trained ML model to produce both local and global explanations. SHAP model was also used in another study [193] in addition to the RF classifier to classify three classes: normal controls, cognitive impairment, and dementia using cognitive scores as input.

Danso et al. [194] used two tree-based algorithms to build ML models on a dataset from the European population to predict the risk of AD, then transfer learning the best model on another dataset from the UK population. In addition, they apply SHAP to visualize individual and population-level risk factors.

Despite the considerable amount of research conducted, its impact on clinical practice is often limited due to several reasons. First, many studies rely exclusively on a single method of analysis, particularly neuroimaging. This narrow focus may overlook valuable information from other modalities. Second, the emphasis on improving the accuracy of ML models has overshadowed the importance of its interpretability which poses challenges in clinical settings where practitioners may not be familiar with the machine-based complex analysis and decision making. Additionally, ML models often require large amounts of data to achieve accurate predictions, which may pose challenges in real-world applications. To address these limitations, the proposed study presents reliable ML algorithms to classify different cognitive states of a person, with following

contributes:

- Leveraging extensive data: Utilisation of a big dataset of comprising 169408 observations and 1024 features. This extensive dataset provides a robust foundation for our research.
- Accurate Multiclass Classification: classifications of individuals into multiple AD classes, including NC, MCI and AD with high and balanced performance.
- Long-term prediction of cognitive state: Developing a mode capable of predicting the cognitive state of an individual four years after their baseline visit. Predict the cognitive state of a person four years after their baseline visit. This prognostic capability has significant implications for early intervention and personalised treatment strategies.
- Rule extraction in AD classification: This the first time in literature, SIRUS or CAR algorithms have been applied to AD classification. Through these models, we extract human-understandable rules that elucidate the interrelationships between the most significant factors contributing to development of AD.

6.2 Methods and materials

This research focuses on identifying key features strongly linked to the progression of AD by leveraging a combination of feature selection techniques and ML algorithms. Additionally, it explores interpretable ML models to derive human-readable insights from complex data patterns and algorithmic decisions, highlighting potential risk factors for the disease. The proposed methodology, illustrated in Figure 6.1, begins with acquiring data from the National Alzheimer's Coordinating Center (NACC). The dataset undergoes preprocessing to handle outliers, missing values, and format adjustments. Dimensionality reduction is performed using correlation analysis and the Boruta algorithm, followed by implementing ML models for Alzheimer's classification and prediction.

A suite of data analytics methods is then applied to interpret the ML model and pinpoint the most significant features. Beyond achieving dependable classification accuracy, this study provides novel insights into risk factors associated with AD. Moreover, it seeks to

enhance model transparency by generating human-understandable rules, offering findings that could ultimately lead to improved treatment strategies and better patient outcomes.

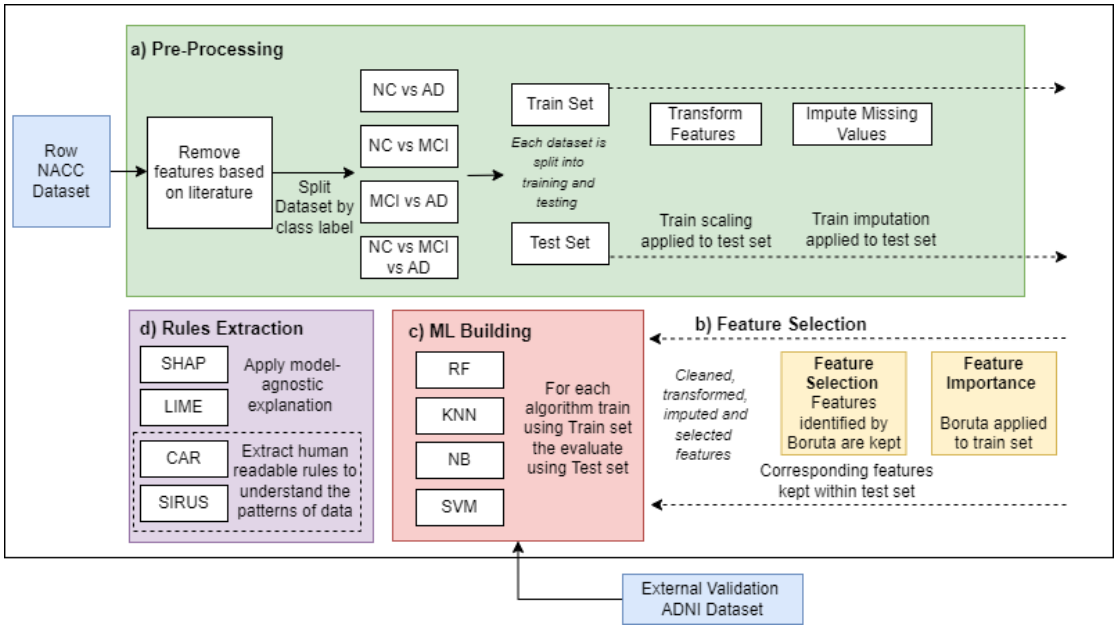


FIGURE 6.1: Workflow Overview of the Proposed Methodology. The process begins with data acquisition from NACC and proceeds through several key stages: (a) Data preprocessing, including the selection of relevant features inspired by existing literature, partition of the dataset based on class labels, division into training and testing subsets, and data transformation and cleansing using the training set as a reference. (b) Feature importance is evaluated using the Boruta algorithm, and only the identified features are retained for subsequent analysis. (c) Construction of four widely recognized ML classifiers to address various tasks related to the classification of cognitive states. External validation of these models is performed using additional data from the ADNI. (d) The final step involves the extraction of human-readable rules from the trained machine learning models, facilitating the interpretation of factors associated with AD.

6.2.1 Dataset

In this chapter, a dataset from the National Alzheimer's Coordinating Centre (NACC) (https://naccdata.org/) was utilized, which was provided to us under a special permission for research purposes. The data collected by NACC researchers from study volunteers was done using a standard set of protocols and procedures to make sure that all the data was collected in an accurate and consistent manner. The data was collected from 37 active Alzheimer's Disease Research Centers which are located in 26 states across the United States. Up to the date of Aug, 2022, NACC has collected a total of 169408 samples belonging to 45923 adults. This study employed data collected from baseline visits and further longitudinal data at successive visits.

A total of 45,923 samples from the baseline visit were divided into four distinct groups based on the diagnosis given during examination. The first group consisted of NC with 18,171 samples. The second group included individuals who showed impairment but did not meet the criteria for MCI, comprising 2,022 samples. The third group comprised 10,043 samples diagnosed with MCI. Finally, the fourth group consisted of 15,687 samples diagnosed with AD and dementia. All criteria and definitions for each group are detailed on the NACC website (https://naccdata.org/). Additional demographic information for these samples can be found in Table 6.1. Data from Only NC, MCI and AD patients were retrieved and utilised in the subsequent sections of this chapter.

The NACC database is an extensive collection of over 1000 variables from various measures and assessments, which were carefully selected to provide a comprehensive overview of each subject's condition and functionality at each visit. These variables include scores from a wide variety of neuropsychological tests and standard questionnaires, both from the participant and study partner. Such tests and questionnaires are commonly used in screening processes to detect memory deficits and behavioural symptoms associated with AD. They are particularly helpful in providing objective information on the progression of the disease. The data categories are explained as follows.

Subject Demographics: This category includes information about the individuals participating in a study, such as age, gender, ethnicity, education level, and socioeconomic background. The demographic data of the subjects provide important contextual information that helps researchers understand the characteristics and diversity of the study population. It allows identification of demographic patterns, risk factors, and possible variations in the manifestation and progression of the condition under investigation.

Physical: The physical data category encompasses objective measurements and assessments related to an individual's physical health. This may include data such as body mass index, blood pressure, cardiovascular health indicators, and other relevant physiological measurements. Physical data provide insight into the general health status of participants, potential comorbidities, and the association between physical health and the condition being studied.

Subject Health History: Subject health history refers to the past medical records and personal health information of participants. It includes details about previous illnesses, medical diagnoses, treatments received, surgeries, family medical history, and lifestyle factors. Subject health history helps researchers identify pre-existing conditions, genetic predispositions, familial links, and other factors that may influence the development,

progression, or management of a particular health condition.

Geriatric Depression Scale (GDS): The GDS is a widely used questionnaire designed to assess the presence and severity of depressive symptoms in older adults. It consists of a series of questions that evaluate mood, feelings of sadness or hopelessness, loss of interest in activities, and other symptoms associated with depression. The scale helps researchers and healthcare professionals identify and measure the presence of depressive symptoms in the target population, which can be important in understanding the impact of depression on overall health and well-being.

Functional Activities Questionnaire (FAQ): The FAQ is a tool used to assess an individual's ability to perform activities of daily living (ADLs). ADLs include tasks such as dressing, bathing, eating, managing finances, and using transportation. The questionnaire provides a structured approach to evaluate functional impairments and limitations in carrying out these essential daily activities. It helps assess the level of functional independence, monitor changes over time, and evaluate the impact of a particular condition or intervention on an individual's ability to perform ADLs.

Neuropsychiatric Inventory Questionnaire (NPIQ): The NPIQ is a comprehensive assessment tool used to evaluate neuropsychiatric symptoms in individuals with cognitive disorders, such as Alzheimer's disease. It covers a range of behavioral and psychological symptoms, including agitation, aggression, anxiety, depression, hallucinations, and sleep disturbances. The questionnaire provides a standardized method to assess and quantify the presence and severity of these symptoms, aiding in the diagnosis, management, and monitoring of neuropsychiatric manifestations in the target population

Global Clinical Dementia Rating (CDR) plus NACC frontotemporal lobar degeneration (FTLD): the global CDR plus NACC FTLD is determined by assessing the severity of impairment across eight specific domains: Memory, Orientation, Judgment and Problem Solving, Community Affairs, Home and Hobbies, Personal Care, Behaviour, and Language. Each domain is individually rated based on standardised criteria, capturing the level of impairment or functional decline in that particular area.

Collecting and analysing data in these categories provides a comprehensive understanding of AD by considering demographic factors, physical health, personal history, depressive symptoms, functional impairments, and neuropsychiatric manifestations. Integrating multiple data categories enhances the evaluation, diagnosis, and management of AD, ultimately contributing to improved patient care and advancing research in the field.

In light of the large number of features and the sparse data set problem, a subset of

features was selected in line with other related studies [60] [195] using the same dataset. Results in selecting a number of features which are informative for the majority of patients, including Subject Demographics, Subject Health History, Physical, GDS, FAQ, NPIQ, CDR Plus NACC FTLD. Table 6.2 lists the features used in this study for further investigation. The dataset size and number of subjects splitting into training and testing sets demonstrated in Figure 6.2. In the prediction tasks concerning NC vs MCI, MCI vs AD, and NC vs MCI vs AD, a downsizing approach was applied to randomly select samples from the NC and AD classes. This selection process aimed to match the size of the MCI class, addressing the issue of class imbalance.

TABLE 6.1: NACC Subjects Demographics by Cognitive Status.

	NC	**Impaired, not MCI**	**MCI**	**AD**
Age (Years)				
<65	3076	543	2541	5394
65-85	7805	844	4084	6032
>85	7181	617	3344	4005
Sex				
Female	11857	1160	5052	8152
Male	6314	862	4991	7535
Education (Years)				
<=12	3076	543	2541	5394
13-16	7805	844	4084	6032
>=17	7181	617	3344	4005
Missing	100	7	61	194
RACE				
White	14349	1439	7901	13077
Black or African American	2855	410	1513	1641
American Indian or Alaska Native	155	23	81	146
Native Hawaiian or Other Pacific Islander	15	3	7	24
Asian	528	53	320	334
Unknown or ambiguous	269	94	221	465

ADNI The ADNI dataset was obtained from the ADNI database (http://adni.loni.usc.edu). Established in 2003 as a collaborative effort between the public and private sectors, ADNI's primary objective is to explore the potential of magnetic resonance imaging, positron emission tomography, biological markers, clinical assessments, and cognitive evaluations for tracking the progression of MCI and early-stage AD. The ADNI dataset serves as an external source for the validation of ML models. The ADNI dataset was pre-processed to align with the NACC dataset, including value mapping and feature

TABLE 6.2: Feature categories and variable name selected from NACC dataset at initial stage of proposed work

NACC Categories	**Variable Name**
Subject Demographics	SEX, HISPANIC , HISPOR, HISPORX, RACE, RACEX, RACESEC, RACESECX, RACETER, RACETERX, PRIMLANG, PRIMLANX, EDUC, MARISTAT, NACCLIVS, INDEPEND, RESIDENC, HANDED,NACCAGE, NACCAGEB, NACCNIHR
Physical	WEIGHT, HEIGHT, NACCBMI, BPSYS, BPDIAS, HRATE, VISION, VISCORR, VISWCORR, HEARING, HEARAID, HEARWAID
Subject Health History	TOBAC30, TOBAC100, SMOKYRS, PACKSPER, ALCOCCAS, QUITSMOK, ALCFREQ, CVHATT, HATTMULT, HATTYEAR, CVAFIB, CVANGIO, CVBYPASS, CVPACDEF, CVPACE, CVCHF, CVANGINA, CVHVALVE, CVOTHR, CVOTHRX, CBSTROKE, STROKMUL, NACCSTYR, ALCOCCAS,ALCFREQ, HATTMULT, CBTIA, TIAMULT,NACCTIYR, PD, PDYR, PDOTHR, PDOTHRYR, SEIZURES, NACCTBI, TBI,TBIBRIEF, TRAUMBRF, TBIEXTEN, TRAUMEXT, TBIWOLOS, TRAUMCHR, TBIYEAR, NCOTHR, NCOTHRX, DIABETES, DIABTYPE, HYPERTEN, HYPERCHO, B12DEF, THYROID, ARTHRIT, ARTHTYPE, ARTHTYPX, ARTHUPEX, ARTHLOEX, ARTHSPIN, ARTHUNK, INCONTU, INCONTF, APNEA, RBD, INSOMN, OTHSLEEP, OTHSLEEX, ALCOHOL, ABUSOTHR, ABUSX, PTSD, BIPOLAR, SCHIZ, DEP2YRS, DEPOTHR, ANXIETY, OCD, NPSYDEV, PSYCDIS, PSYCDISX'
Geriatric Depression Scale (GDS),	NOGDS, SATIS, DROPACT, EMPTY, BORED, SPIRITS, AFRAID, HAPPY, HELPLESS, STAYHOME, MEMPROB, WONDRFUL, WRTHLESS, ENERGY, HOPELESS, BETTER, NACCGDS
Functional Activities Questionnaire (FAQ),	BILLS, TAXES, SHOPPING, GAMES, STOVE, MEALPREP, EVENTS, PAYATTN, REMDATES, TRAVEL
Neuropsychiatric Inventory Questionnaire (NPI-Q)	NPIQINF, NPIQINFX, DEL, DELSEV, HALL, HALLSEV, AGIT, AGITSEV, DEPD, DEPDSEV, ANX, ANXSEV, ELAT, ELATSEV, APA, APASEV, DISN, DISNSEV, IRR, IRRSEV, MOT, MOTSEV, NITE, NITESEV, APP, APPSEV
CDR® Plus NACC FTLD	MEMORY, ORIENT, JUDGMENT, COMMUN, HOMEHOBB, PERSCARE, COMPORT, CDRLANG
Target Class	NACCUDSD

name adjustments, as demonstrated in Tables 6.3 and 6.4. The data size and number of subjects splitting into training and testing sets demonstrated in Figure 6.3.

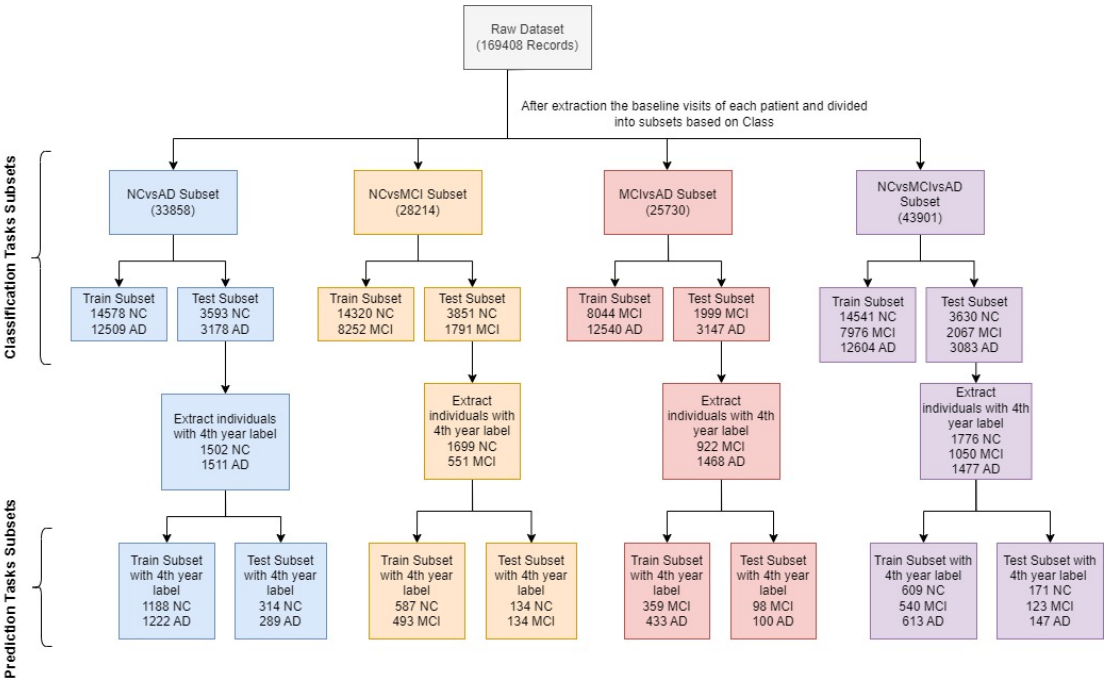


FIGURE 6.2: The sizes of the NACC data subsets for each task. In the prediction tasks concerning NC vs MCI, MCI vs AD, and NC vs MCI vs AD, a downsizing approach was applied to randomly select samples from the NC and AD classes. This selection process aimed to match the size of the MCI class, addressing the issue of class imbalance.

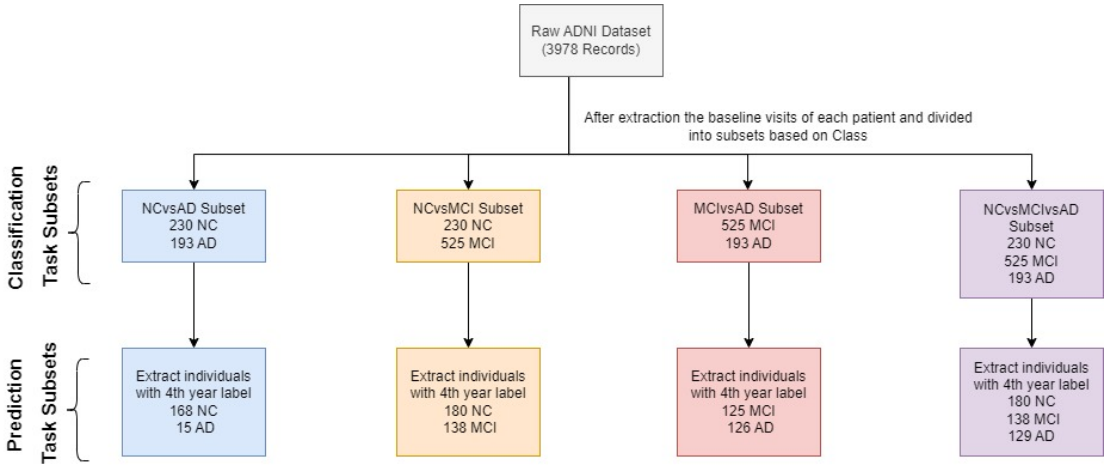


FIGURE 6.3: The sizes of the ADNI data subsets for each task.

6.2.2 Data Pre-processing

In most cases, ML techniques rely on a data set that is supposed to be complete or noise-free. Despite this, real-world data is far from perfect or complete. As part of data pre-processing, techniques are often used for removing noisy data or for imputing (filling in) missing data [196]. Due to the many challenges that incomplete data can pose in both data analysis and building an intelligent model, all subjects and attributes are

Table 6.3: Mapping values of ADNI dataset features to match corresponding values of matching feature of NACC dataset using the ADNI and NACC datasets dictionaries. Since the downloaded data from ADNI has text as values in the features step inmoved mapping this text to corresponding integers.

Feature Names	**Mapping values**
'FAQFINAN','FAQFORM','FAQSHOP', 'FAQGAME','FAQBEVG','FAQMEAL', 'FAQEVENT','FAQTV','FAQTRAVL'	'Normal (0)': 0, 'Never did, but could do now (0)':0, 'Never did, would have difficulty now (1)': 1, 'Has difficulty, but does by self (1)': 1, 'Requires assistance (2)': 2, 'Dependent (3)': 3
'GDMEMORY'	'No(0)': 0, 'Yes(1)': 1
'NPIC','NPIE','NPIG','NPII','NPIJ'	'No': 0, 'Yes': 1
'DX'- Class label	'CN': 0, 'MCI': 1, 'Dementia':2

Table 6.4: Conversion of ADNI feature names to match the corresponding feature names in the NACC dataset to ensure compatibility with ML classifiers.

ANDI Feature Name	**NACC feature Name**
'CDMEMORY'	'MEMORY'
'CDORIENT'	'ORIENT'
'CDJUDGE'	'JUDGMENT'
'CDCOMMUN'	'COMMUN'
'CDHOME'	'HOMEHOBB'
'CDCARE'	'PERSCARE'
'GDMEMORY'	'MEMPROB'
'GDTOTAL'	'NACCGDS'
'NPIC'	'AGIT'
'NPIE'	'ANX'
'NPIG'	'APA'
'NPII'	'IRR'
'NPIJ'	'MOT'
'FAQFINAN'	'BILLS'
'FAQFORM'	'TAXES'
'FAQSHOP'	'SHOPPING'
'FAQGAME'	'GAMES'
'FAQBEVG'	'STOVE'
'FAQMEAL'	'MEALPREP'
'FAQEVENT'	'EVENTS'
'FAQTV'	'PAYATTN'
'FAQTRAVL'	'TRAVEL'

screened to remove incomplete data. The following steps explain the data pre-process techniques employed in the current study.

6.2.2.1 Missing Values and Unmeaningful Features

Missing values pose a challenging problem in data pre-processing, which can be overcome in a variety of ways [197]. Firstly, variables that exhibit the same value in 90% of the participants are removed, this has reduced number of variables from 172 to 118. Secondly, all variables and subjects (i.e., participants) comprising missing values in more than 50 percent of their occurrences are removed. This resulted the number of variables to be further reduced to 67. Likewise, the number of records is reduced from 27087 to 26722 for the training set of the CNvsAD subset. We then impute the missing data of the remaining variables using a simple and widely used imputation technique [60]. For continuous variables, mean of the variable was used while for the categorical variables, mode imputation is used. These processes are first applied to the training set then reflected onto the testing set.

6.2.2.2 Correlation analysis and Data Standardisation

Due to the nature of data collection within NACC, a high degree of correlation among variables is frequently observed. For example, variables like RACE and NACCNIHR, which both relate to a subject's ethnicity, often overlap. Including such closely related variables can impact the reported outcomes. To mitigate this issue, correlation analysis was conducted to identify and remove highly correlated features using the Cramer's V method [198]. This approach is particularly suited for categorical variables.

To handle continuous features, they were discretised into categories based on existing literature. For instance, BMI was categorised into groups such as 'underweight,' 'normal,' 'overweight,' and 'obesity,' with similar transformations applied to other continuous variables. Details of these conversions are presented in Table 6.5. For categorical features, encoding methods were chosen based on the variable type: nominal variables (where order is not important) were one-hot encoded, while ordinal variables (where order matters) were label encoded.

6.2.2.3 Outliers Detection

Outliers are data points that diverge significantly from conventional patterns or are not in accordance with expected normal patterns for the measure under consideration [199].

TABLE 6.5: Discretised continuous values inspired by literature. *Years of education converted into no Bachelor's degree (0), with Bachelor's degree (1), with a postgraduate degree. **Years of smoking converted into bins depending on quantile analysis.

Feature Name	**Categories bins**
NACCAGE	>60, 60 - 75, >75], labels=[0,1,2]
NACCBMI	<18.5, 18.5 - 25, 25 - 30, >30], labels=[0,1, 2,3]
BPSYS	<90, 90 - 140, >140, labels=[0,1, 2]
BPDIAS	<60, 60 - 90, >90, labels=[0,1, 2]
EDU*	<12, 12 – 16 , >16, labels=[0,1, 2]
HRATE	<60, 60 - 100, >100, labels=[0,1,0]
SMOKYRS**	<15, 15 - 30, >30, labels=[0,1, 2]

Despite the importance of this step, several research studies in AD classification ignore this step or may not report it properly. In this study, two approaches were utilised to deal with outliers. Firstly, for categorical features, the percentage of each value in a variable was calculated and then substitute the mode of the variable in all values that have a percentage of less than 3% of the total values. For the numerical features, Inter-Quartile Range (IQR) was used to identify the outliers within each continuous feature. In IQR, the interest falls on the lower quartile (Q1) and the upper quartile (Q3), where IQR is calculated as follows:

$$IQR = Q3 - Q1$$

Outliers are then identified using Eq 2 and 3, representing a decision threshold where the data points falling outside the range are treated as outliers. The decision range is calculated as follows:

$$Lowerbond = (Q1 - 1.5 * IQR)$$

$$Upperbond = (Q1 + 1.5 * IQR)$$

The term outlier in this study refers to data points that fall outside the Lower Bound or that exceed the Upper Bound.

6.2.3 Experiment Design

The present study employs a combination of algorithms to identify the most significant features from the NACC dataset and to derive interpretable rules for AD classification,

enhancing the explainability of machine learning models. In all experiments, the dataset is split into 80% for training and 20% for testing to evaluate performance on unseen data.

Experiment 1 (EXP1):

This experiment assesses the ability of ML models to classify clinical stages of cognitive impairment using the NACC dataset. The process begins with pre-processing steps, followed by training and evaluation of the models, as depicted in Figure 6.1. ML models are trained with 64, 55, 67, and 66 features (before feature selection) for the subsets NC vs AD, NC vs MCI, MCI vs AD, and NC vs MCI vs AD, respectively.

Experiment 2 (EXP2):

This experiment evaluates the effectiveness of the proposed feature selection algorithm. The reduced feature sets include 24 features for NC vs AD, 10 for NC vs MCI, 17 for MCI vs AD, and 18 for NC vs MCI vs AD (as detailed in Table 6.7). ML models, configured as in EXP1, are trained and tested on these reduced subsets to examine their classification performance using the selected key features.

Experiment 3 (EXP3):

Building on EXP2, this experiment uses the same feature sets to train and test ML models but shifts the focus to predicting an individual's cognitive state four years after their baseline visit (i.e. the features used to train the model are taken from the baseline visit, but the labels were obtain from a follow-up visit four years later). This investigation is pivotal for determining whether the identified features (Table 6.7) are effective in forecasting cognitive outcomes over an extended timeframe.

Experiment 4 (EXP4):

The aim of this experiment is to evaluate the generalisability of the best-performing models from EXP2 and EXP3. Using an external dataset (the ADNI dataset), the classifiers are tested for their ability to perform both classification and prediction tasks across different datasets.

Experiment 5 (EXP5):
In this experiment, the CAR and SIURS algorithms are applied to extract interpretable rules that capture significant patterns in the data. These rules provide valuable insights into an individual's cognitive state. To ensure the robustness of the selected features, the results from CAR and SIURS are compared with outputs from SHAP and LIME models, further validating the findings.

6.3 Results and discussions

Detailed results as retrieved from the various experiments (Section 6.2.4 Experiment Design). For each experiment, results are shown from multiple classifiers that include RF, KNN and NB and SVM. For each classifier, detailed metrics are retrieved to compare the classifiers' performances in corresponding experiments that are described as follows.

Data Pre-processing Outcomes

To verify that the pre-processing steps did not bias the dataset, we first examined the effects of data imputation. Figure 6.4 shows the mean and standard deviation of some feature before and after data imputation to ensure the imputation did not affect the statistics of the features. While Table 6.6 shows the number of participants and imputed values for each data subset.

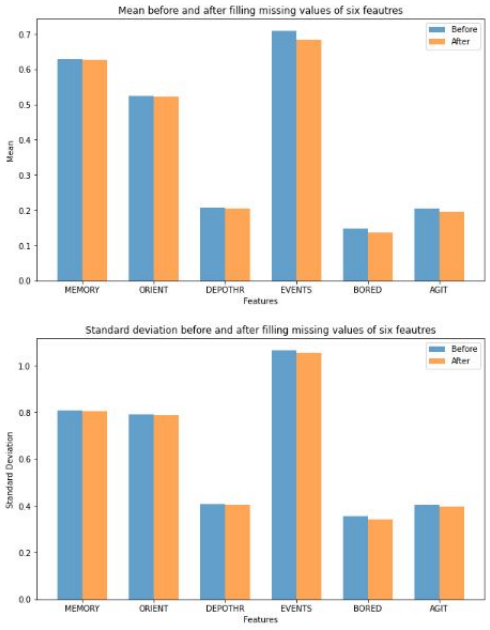


FIGURE 6.4: Mean and Standard deviation for some variables of NC vs AD training subset before and after filling missing values.

TABLE 6.6: Numbers of participants with imputed data and imputed values for each dataset.

Data Subset	**Number of Participants**	**Participants with imputed Data**	**No. of Non-Missing Values**	**No. of Imputed Values**
NC vs AD Training set	27,087	18,667	1,851,437	90,115
NC vs AD Testing set	6,771	4,522	469,606	20,635
NC vs MCI Training set	22,572	13,621	1,348,079	40,875
NC vs MCI Testing set	5,642	9,249	343,093	9,249
MCI vs AD Training set	20,584	15,793	1,494,011	83,520
MCI vs AD Testing set	5,146	3,903	375,781	19,192
NC vs MCI vs AD Training set	35,121	24,351	24,11,249	112,823
NC vs MCI vs AD Testing set	8,780	5,970	606,386	26,309

The influence of outlier handling was then examined. For categorical features, Fig 6.5 shows the distribution of values in 9 categorical features, the variable "CDRLANG" has the value of 3 in very few samples of the dataset. Therefore, these values are substituted with value 0 which is the mode of the variables "CDRLANG" as shown in Fig 6.6. the same process was performed for the remaining categorical variables. While figure 6.7 shows a boxplot for some of the continuous variables in their original form. It can be noticed that the distribution of data points improved after applying IQR-based outlier removal as shown in Figure 6.8.

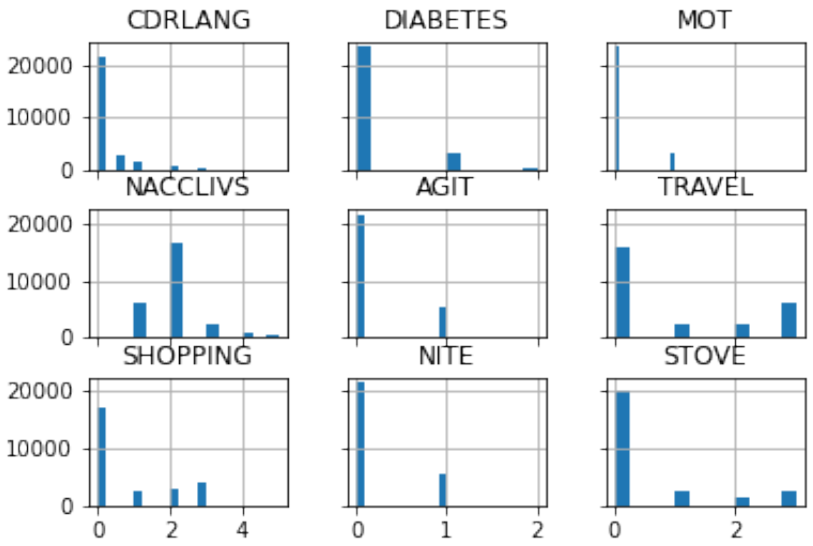


FIGURE 6.5: The distribution of values of some variables of NC vs AD training dataset.

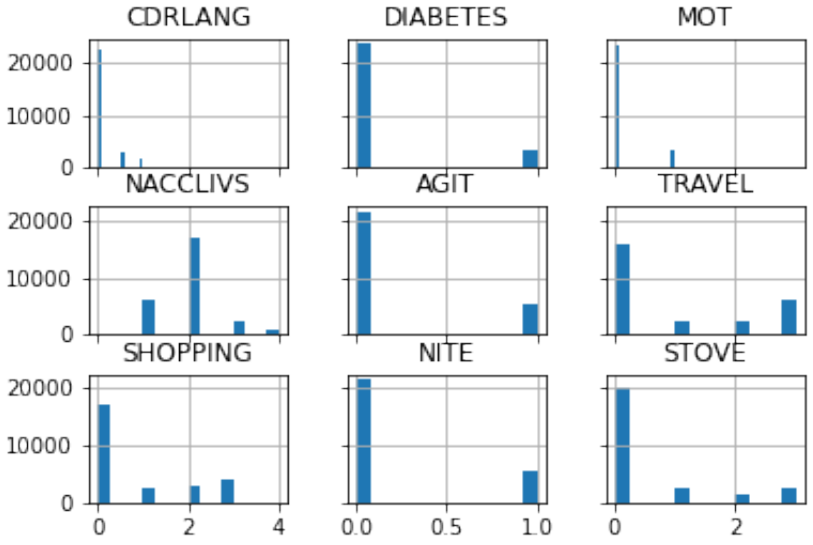


FIGURE 6.6: The distribution of values of some categorical features from NC vs AD training subset after substituting the mode of the feature instead of the values that account of 3% of the feature.

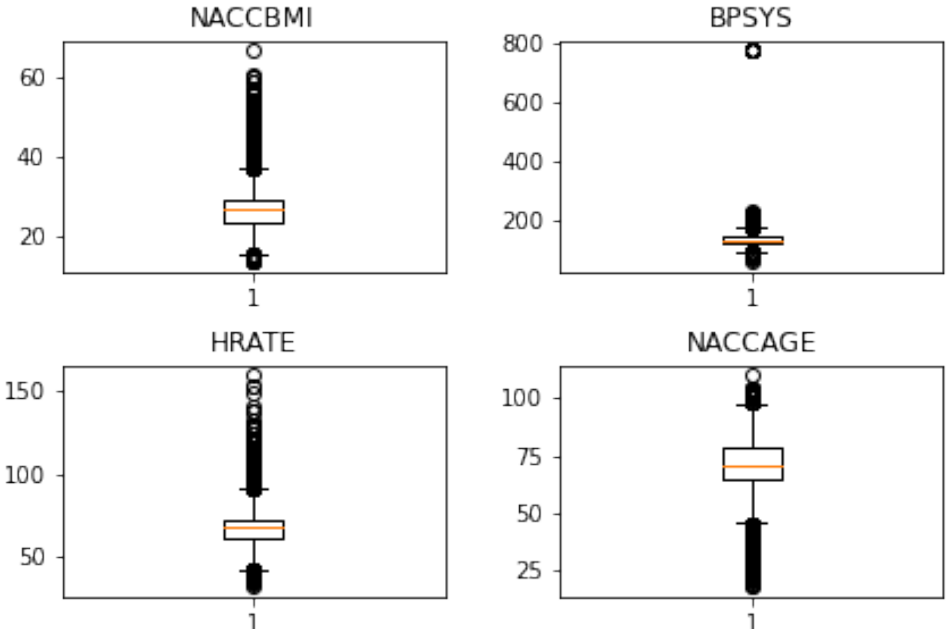


FIGURE 6.7: Boxplot to show the distribution of data points of continuous variables from NC vs AD training subset.

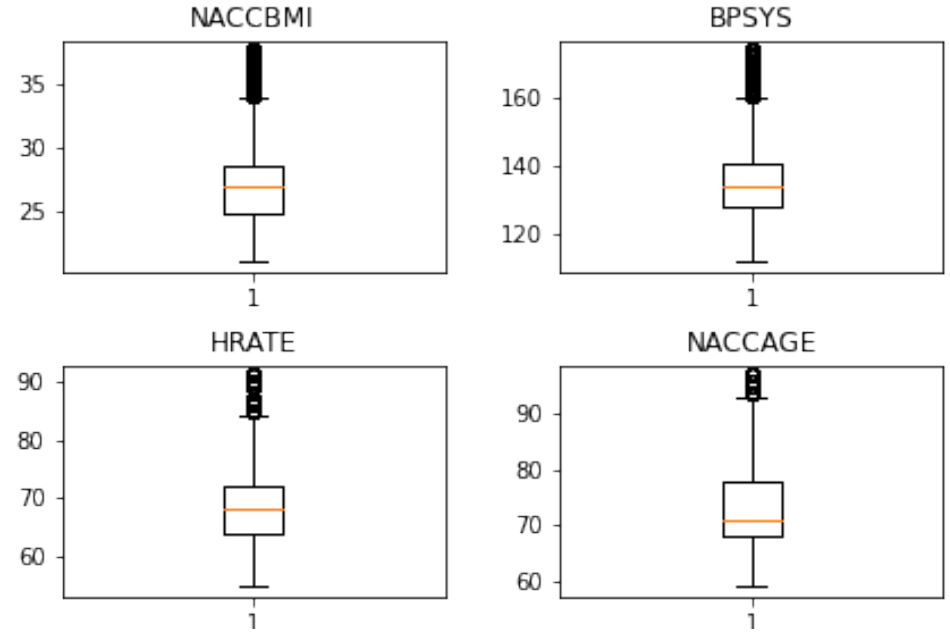


FIGURE 6.8: Boxplot to show the distribution of data points of continuous variables after removing outlier data points from NC vs AD training subset.

Feature Selection Outcomes

Table 6.7 shows the selected features for each data subset identified as most relevant using the Boruta algorithm. It can be noticed that the number of identified features are substantially reduced to 24 only (compared to 64 in original NC vs AD data subset). Furthermore, seven of the selected features belong to CDR which measures the relative severity of dementia by assigning a score between 0 (no impairment) and 3 (severe impairment) [200]. A clinician's clinical judgment and a semi structured interview with the subject and caregiver (informant) determine CDR score. On the other hand, nine features fall with FAQ which measures difficulty with daily living activities and was found to be a valid and reliable measure according to studies in the literature [201]. Five features among the selected ones belong to the NPI-Q which was developed by Cummings [202], to assess behavioural symptoms associated with dementia and found to be an effective tool for the assessment of dementia in different populations [203] [204]. Two feature belongs to GDS and one from subject's Demographics.

In contrast to NC vs AD data subset, Boruta algorithm identified only 10 features as important for NC vs MCI subset. Out of these, six aligns with the CDR, two with GDS and two with FAQ. Similarly, for the MCI vs AD subset, 17 features are identified as informative. Finally, 18 variables were selected for multi-class category NC vs MCI vs AD. The selected variables for each data subset are shown in Table 6.7. Across all classification tasks, we consistently observe a shared set of features, namely MEMORY, ORIENT, JUDGMENT, COMMUN, CDRLANG, MEMPROB, NACCGDS, BILLS, and TAXES. These features consistently demonstrate their significance in distinguishing cognitive states, emphasising their crucial role in AD diagnosis. In addition to these common features, the NC vs. AD classification task incorporates task-specific features such as COMPORT, AGIT, ANX, APA, IRR, and MOT. Notably, the inclusion of features related to behavioural domains (AGIT, ANX, APA, IRR, and MOT) gains importance when classifying NC vs. AD. Additionally, it is noteworthy that the feature HOMEHOBB is shared among all tasks, except in the case of NC vs. AD. This distinctive pattern further emphasises the importance of certain features in differentiating between cognitive states.

To externally validate the ML classifiers, data from ADNI was incorporated. However, it's noteworthy that three features, namely COMPORT, CDRLANG, and INDEPEND, were not present in the ADNI dataset (refer to Figures 6.9, 6.10, 6.11, and 6.12 for

final feature sets). Consequently, we opted to exclude these features. Subsequently, we trained the ML classifiers on the remaining selected features and proceeded with the external evaluation of the classifiers using the ADNI dataset. Although these features could be clinically relevant—behavioural symptoms, language function and functional independence—their removal was necessary to preserve the integrity of external validation. Including them would have required imputation or estimation of completely missing data, which poses several methodological issues. This is not a suitable context for imputation since these features are completely missing for all ADNI participants and not just a subset. In this case, there is no within-dataset information to support reliable imputation and any attempt to estimate values would be highly speculative. Using NACC data to impute missing values in ADNI would introduce data leakage, whereby information from the training dataset is inadvertently used during testing. This would compromise the independence of the external validation process and undermine the primary objective of testing model generalisability across different cohorts. The inclusion of imputed values could also inflate model performance artificially, leading to over-optimistic estimates that would not hold in truly unseen datasets. The decision to exclude these three features was made to ensure that ADNI remains a completely independent external dataset, free from training data influence. This might reduce model performance slightly due to the loss of informative features, but it provides a more rigorous and unbiased assessment of model generalisability.

TABLE 6.7: Informative features selected by Boruta algorithm for each data subset.

Data Subset	Selected features
NC vs AD	MEMORY, ORIENT, JUDGMENT, COMMUN, PERSCARE, COMPORT, CDRLANG, MEMPROB, NACCGDS, AGIT, ANX, APA, IRR, MOT, BILLS, TAXES,SHOPPING, GAMES, STOVE, MEALPREP, EVENTS,PAYATTN, TRAVEL, INDEPEND
NC vs MCI	MEMORY, ORIENT, JUDGMENT, COMMUN, HOMEHOBB, CDRLANG, MEMPROB, NACCGDS, BILLS, TAXES
MCI vs AD	MEMORY, ORIENT, JUDGMENT, COMMUN, HOMEHOBB, PERSCARE, CDRLANG, NACCGDS, BILLS, TAXES, SHOPPING, GAMES, MEALPREP, EVENTS, PAYATTN, TRAVEL, INDEPEND
NC vs MCI vs AD	MEMORY, ORIENT, JUDGMENT, COMMUN, HOMEHOBB, CDRLANG, MEMPROB, NACCGDS, BILLS, TAXES, SHOPPING, GAMES, STOVE, MEALPREP, EVENTS, PAYATTN, TRAVEL, INDEPEND

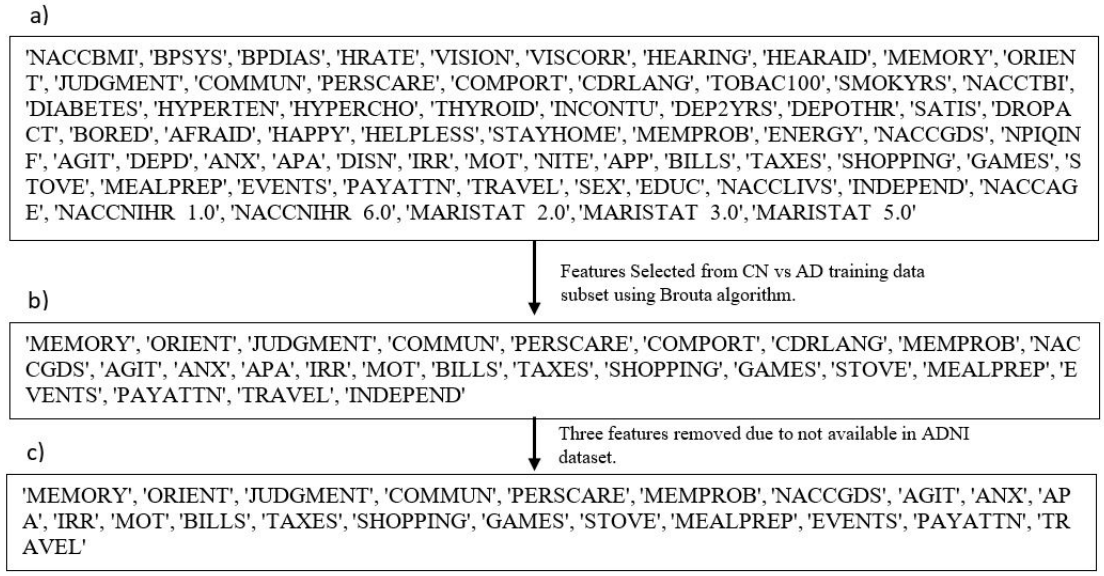


FIGURE 6.9: Features for CN vs AD subset: a) after data pre-processing, b) after feature selection, c) final selected features after remove feature which are not available in ADNI dataset.

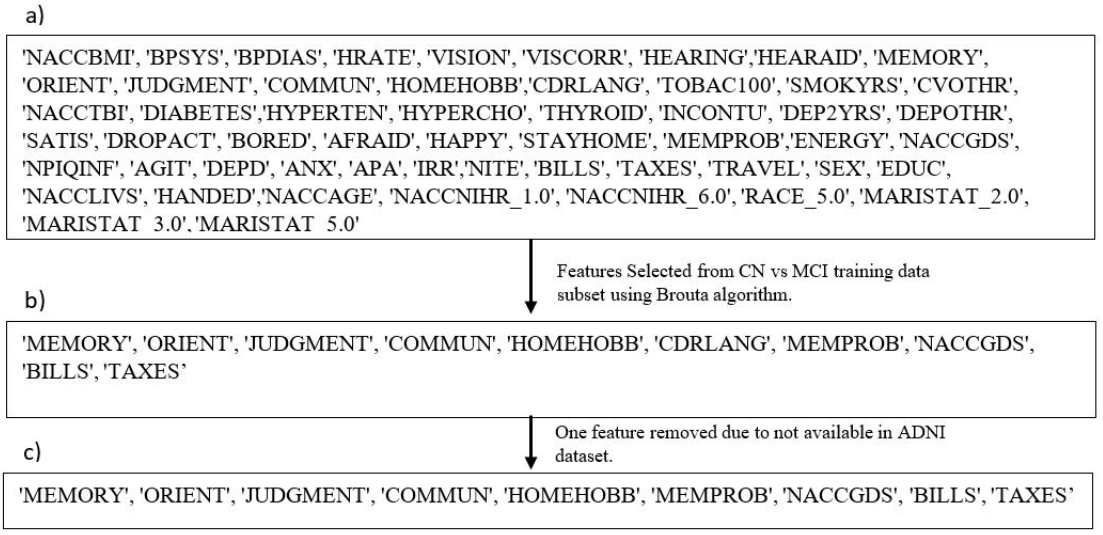


FIGURE 6.10: Features for CN vs MCI subset: a) after data pre-processing, b) after feature selection, c) final selected features after remove feature which are not available in ADNI dataset.

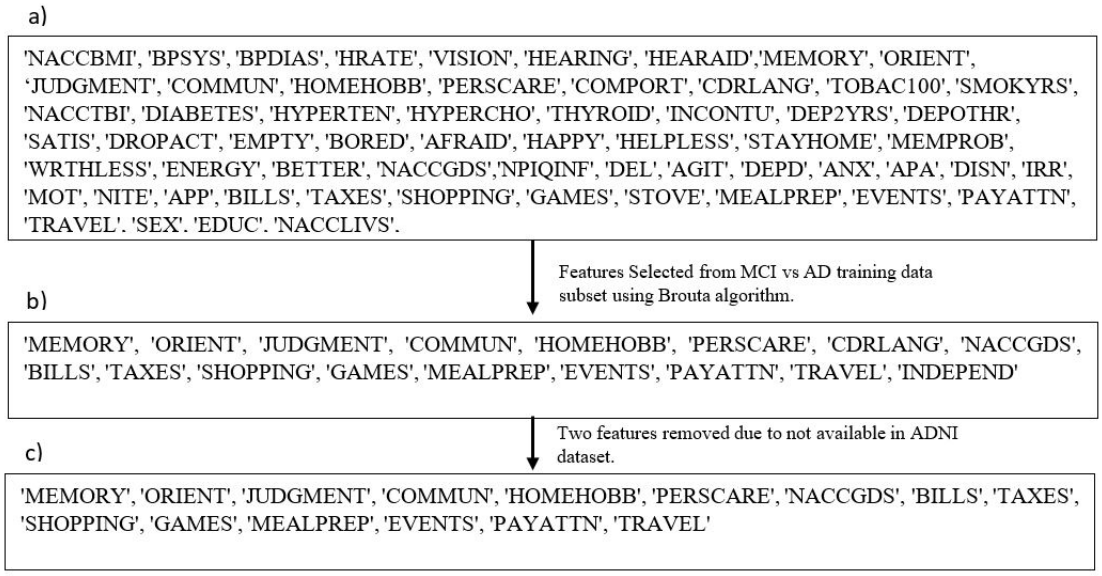


FIGURE 6.11: Features for MCI vs AD subset: a) after data pre-processing, b) after feature selection, c) final selected features after remove feature which are not available in ADNI dataset.

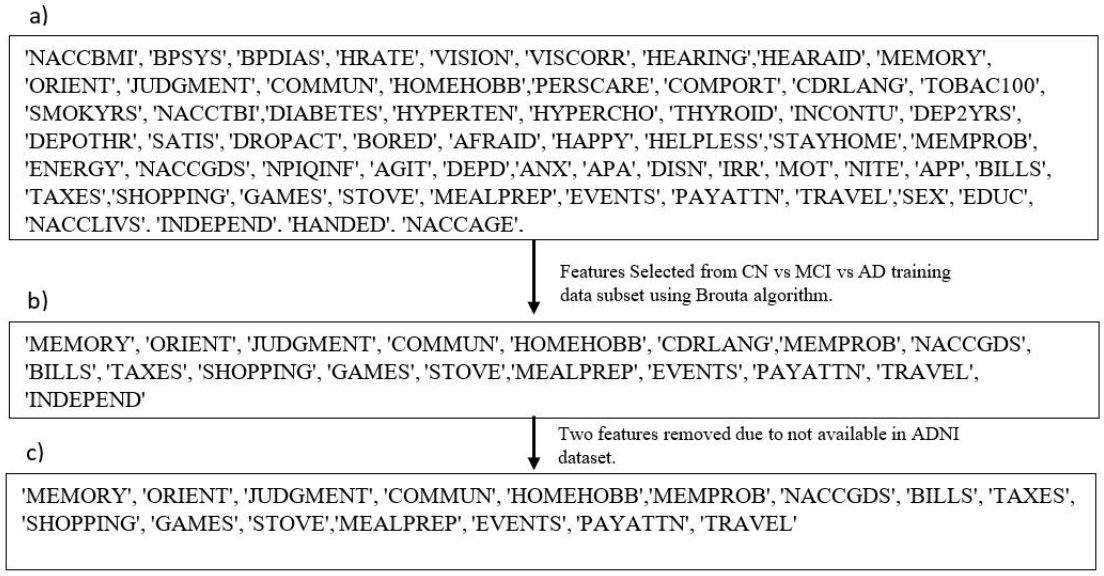


FIGURE 6.12: Features for CN vs MCI vs AD subset: a) after data pre-processing, b) after feature selection, c) final selected features after remove feature which are not available in ADNI dataset.

Results for EXP1

Table 6.8 presents the results of EXP1 which utilises original features (i.e., all features without feature reduction). It can be noticed from Table 6.8.a that the highest accuracy of 97.8% was achieved by the RF algorithm for the classification of NC against AD cases when evaluated over unseen data samples. Furthermore, the RF model indicated robust performance for other metrics such as precision, recall, and F1 (97.2%, 98.1% and 97.6%, respectively), indicating its ability to provide stable and balanced classification with fewer false classifications among both classes. These outcomes suggest the RF model as an effective tool for the classification of NC and AD cases, with a high degree of accuracy and reliability.

Table 6.8.b presents the results of the performance of classifiers in classifying NC and MCI cases. Among the different classifiers, RF achieved the highest accuracy of 88.6%. On the other hand, the KNN classifier indicated poor performance, with an imbalanced precision and recall of 81.2% and 48% respectively. This demonstrates that KNN is not an ideal model for classifying NC and MCI cases.

Table 6.8.c shows the results for classification between MCI and AD cases, indicating RF as outperforming classifier. This is in agreement with the results of Tables 6.8.a and 6.8.b, which also shows that the RF is the best performing classifier. On the other hand, the NB classifier shows comparatively poor performance with an accuracy of 82.4%. Furthermore, the NB is biased in terms of precision and recall of 92.5% and 76.4%, respectively.

Table 6.8.d shows the final results of EXP1, where the classifiers are trained and tested over a multi-classification problem to classify three classes including NC, MCI, and AD. We used one-vs-one strategy [205] where the multi-class classification task is broken up into a series of binary classification problems and was chosen over the alternative strategies as it provides improved performance. It can be noticed that the RF algorithm again outperformed (with 85.2% accuracy) followed by the SVM (85.1%) and KNN with least performance (with accuracy of 75.5%). This is likely due to KNN not being able to capture the complexity of the data of three classes. Additionally, the performance of the RF model was consistent across all metrics, making it a reliable and robust choice for any classification task.

Overall, it can be observed that the classifiers achieved better results when classifying NC vs AD (Table 6.8.a) compared to NC vs MCI (Table 6.8.b) and MCI vs AD (Table

6.8.c). This is not surprising, given that NC cases are closer in terms of characteristics to MCI, and MCI and AD are also similar. However, the classification between NC and AD is easier to carry out due to the significant differences between the two. For example, the cognitive decline in AD is much more pronounced than in NC, making it easier for the classifiers to differentiate between the two.

ML Model	Accuracy%	Precision%	Recall%	F1 score%	Mean%	SD	P-value
a) Results of EXP1 : NC vs AD							
RF	97.8	97.2	98.1	97.6	97.8	0.002	
KNN	94.8	97.8	90.8	94.1	94.2	0.003	P <0.001
NB	96.2	93.8	98.3	96	96.1	0.002	P <0.001
SVM	97.6	97.6	97.2	97.4	97.6	0.003	P = 0.292
b) Results of EXP1 : NC vs MCI							
RF	88.6	81.9	88.6	85.1	85.9	0.003	
KNN	76.8	81.2	48	60.3	59.5	0.006	P <0.001
NB	82.4	76.8	74.7	75.8	76.1	0.005	P <0.001
SVM	88.1	82.1	86.7	84.3	85.3	0.003	P = 0.003
c) Results of EXP1 : MCI vs AD							
RF	87.3	90.2	88.1	89.1	90.5	0.002	
KNN	83.1	89.6	81	85	86.7	0.002	P <0.001
NB	82.4	92.5	76.4	83.7	85.6	0.004	P <0.001
SVM	87.6	90.4	88.5	89.4	90.3	0.003	P = 0.49
d) Results of EXP1 : NC vs MCI vs AD							
RF	85.2	85.6	85.2	85.4	86.3	0.002	
KNN	75.5	74.1	75.5	73.4	73.4	0.005	P <0.001
NB	77.9	78.7	77.9	78	79.1	0.001	P <0.001
SVM	85.1	85.3	85.1	85.2	86	0.004	P = 0.20

TABLE 6.8: Results of EXP1. Performance of ML Models in Classifying: a) NC vs AD, b) NC vs MCI, c) MCI vs AD and d) NC vs MCI vs AD. For each task, we employed five-fold cross-validation on the training data. Four folds were used for training, and the remaining fold was used for testing, resulting in five replicas. Statistics were derived using the F1 score. We conducted a performance comparison between RF and the other models to determine the presence of statistically significant differences. P-values were calculated using a two-sided t-test, and the means and standard deviations are listed in the table. Subsequently, we internally evaluated the model by training it on the entire training dataset and testing it on a hold-out test dataset, with the results reported in the table.

Results for EXP2

As described in the Experiment section , EXP2 evaluates the performance of four ML models in classifying three groups of subjects: NC, MCI, and AD while using the reduced set of features. The results of the classification are presented in Table 6.9.

Table 6.9.a presents the classification results for NC vs AD, where all models achieved

high performance with accuracy above 96%. The RF model performed the best with an accuracy of 97.5%, followed by SVM with an accuracy of 97. 3%. In terms of precision and recall, all models performed almost similar with scores above 94%. Overall, the results suggest that the ML models are capable of accurately distinguishing between NC and AD subjects utilising the reduced features. Table 6.9.b shows the classification results for NC vs MCI, where the RF and SVM models achieved same accuracy rates of 88.1%, while the KNN and NB models should a slightly reduced accuracy rate. RF model achieved highest recall score but should a marginally reduced precision comparing to other classifiers.

Table 6.9.c presents the classification results for MCI vs AD, where all models achieved accuracy rates above 82%. The NB model achieved the highest precision score, while the RF model achieved the highest recall score. Table 6.9.d shows the classification results for multi-class classification of NC vs MCI vs AD, where all models achieved accuracy rates above 78%. The SVM model performed the best, achieving performance rates above 84%. The NB model achieved the lowest accuracy rate among the four models. The SVM model also achieved high precision and recall scores across all classes.

In summary, the ML models indicate reliable performance in classifying NC, MCI, and AD subjects. The RF and SVM models consistently achieved high accuracy rates and precision and recall scores across all classification tasks, suggesting that they are effective models for the task of AD classification.

ML Model	Accuracy%	Precision%	Recall%	F1 score%	Mean%	SD	P-value
a) Results of EXP2 : NC vs AD							
RF	97.5	97	97.6	97.3	97.5	0.002	
KNN	96.4	97.1	95.2	96.1	96.6	0.002	P <0.001
NB	96.1	94.2	97.5	95.8	96.4	0.001	P <0.001
SVM	97.3	97.1	97.1	97.1	97.5	0.002	P = 0.846
b) Results of EXP2 : NC vs MCI							
RF	88.1	81.3	87.7	84.4	89	0.003	
KNN	87.5	81.6	85.4	83.5	88.4	0.006	P = 0.158
NB	82.9	83.6	66.5	74.1	83	0.004	P <0.001
SVM	88.1	82.1	86.7	84.3	89	0.002	P = 0.821
c) Results of EXP2 : MCI vs AD							
RF	86	88.7	87.4	88.1	87	0.003	
KNN	84.4	89.6	83.3	86.3	85.4	0.005	P = 0.001
NB	82.4	93.4	75.7	83.6	84.3	0.006	P <0.001
SVM	86.6	90.4	86.6	88.5	87.6	0.002	P = 0.028
d) Results of EXP2 : NC vs MCI vs AD							
RF	82.6	82.9	82.6	82.7	85.3	0.002	
KNN	82.5	83	82.5	82.7	82.7	0.004	P <0.001
NB	78.2	78.6	78.2	78.1	79.2	0.002	P <0.001
SVM	84.7	85.2	84.7	84.9	85.7	0.004	P = 0.185

TABLE 6.9: Results of EXP2. Performance of ML Models using reduced feature sets in Classifying: a) NC vs AD, b) NC vs MCI, c) MCI vs AD and d) NC vs MCI vs AD. For each task, we employed five-fold cross-validation on the training data. Four folds were used for training, and the remaining fold was used for testing, resulting in five replicas. Statistics were derived using the F1 score. We conducted a performance comparison between RF and the other models to determine the presence of statistically significant differences. P-values were calculated using a two-sided t-test, and the means and standard deviations are listed in the table. Subsequently, we internally evaluated the model by training it on the entire training dataset and testing it on a hold-out test dataset, with the results reported in the table.

Results for EXP3

In Exp 3, we employed ML classifiers to predict an individual's cognitive state four years after their initial visit (i.e. fourth visit). To assess the accuracy of our classifiers, we conducted a series of experiments, the outcomes of which are detailed in Table 6.10. In the binary classification task of distinguishing between NC vs AD, all models achieved notably high accuracy with RF excelled with the highest accuracy of 96.4 while NB exhibited a slightly reduced accuracy of 95.1%.

In the binary classification task of NC vs MCI, all models achieved accuracy rate exceeding 71%, with RF achieving the best accuracy and F1 score, measuring 78.1% and 75.7%, respectively. Furthermore, all ML models demonstrated imbalanced performance in terms of precision and recall. For instance, NB reached precision and recall scores

of 90.4% and 48.5%, respectively. Conversely, RF showed the least biased performance, with precision and recall scores of 85.9% and 67.6% respectively.

In the binary classification of MCI and AD, RF achieved the highest accuracy of 76.7% along with the highest, recall and F1 score, measuring 82% and 78%, respectively. However, NB achieved the highest precision of 90.1% but indicated a comparatively lower recall score of 64%.

In the multi-class classification task encompassing NC, MCI and AD, RF achieved the highest F1 score of 72.6% and maintained stable performance in terms of precision and recall of 72.5% and 73%, respectively. Conversely, NB indicated least F1 score of 67%.

The results presented in Table 6.10 underscore the classifiers' ability to more accurately predict NC and AD classes compared to the MCI class. This outcome aligns with expectations, given that distinguishing between NC and AD classes is typically more straightforward, whereas MCI falls in an intermediate category, presenting a greater challenge.

ML Model	Accuracy%	Precision%	Recall%	F1 score%	Mean%	SD	P-value
a) Results of EXP3: NC vs AD							
RF	96.4	97.5	95.1	96.3	95.2	0.009	
KNN	95.8	97.8	93.4	95.5	90.6	0.017	P = 0.001
NB	95.1	94.8	95.1	94.9	94.1	0.011	P = 0.184
SVM	96.1	97.5	94.4	95.9	92.9	0.014	P = 0.029
b) Results of EXP3: NC vs MCI							
RF	78.1	85.9	67.6	75.7	53.8	0.043	
KNN	72.9	83.1	58	68.3	42.3	0.022	P = 0.001
NB	71.4	90.4	48.5	63.1	53.9	0.029	P = 0.892
SVM	75.9	85.1	63.2	72.5	54.9	0.036	P = 0.729
c) Results of EXP3: MCI vs AD							
RF	76.7	74.5	82	78	89.5	0.009	
KNN	74.2	74.2	75	74.6	87.2	0.008	P = 0.007
NB	78.2	90.1	64	74.8	79	0.021	P <0.001
SVM	76.2	74.7	80	77.2	90.3	0.008	P = 0.249
d) Results of EXP3: NC vs MCI vs AD							
RF	73	72.5	73	72.6	76.5	0.005	
KNN	69.8	69.2	69.8	69.4	72.2	0.006	P <0.001
NB	67.8	68.1	67.8	67	73.1	0.013	P = 0.001
SVM	71.6	71.2	71.6	71.4	73.3	0.012	P = 0.002

TABLE 6.10: Results of EXP3. Performance of ML Models using reduced feature sets in Predicting: a) NC vs AD, b) NC vs MCI, c) MCI vs AD and d) NC vs MCI vs AD. For each task, we employed five-fold cross-validation on the training data. Four folds were used for training, and the remaining fold was used for testing, resulting in five replicas. Statistics were derived using the F1 score. We conducted a performance comparison between RF and the other models to determine the presence of statistically significant differences. P-values were calculated using a two-sided t-test, and the means and standard deviations are listed in the table. Subsequently, we internally evaluated the model by training it on the entire training dataset and testing it on a hold-out test dataset, with the results reported in the table.

Results for EXP4: External Validation

To assess the generalisability of the classifiers, we conducted an external validation using the ADNI Dataset, testing the two top-performing classifiers, RF and SVM across a range of tasks. These tasks encompassed the classification of cognitive states at the baseline visit and the prediction of cognitive states four years later, including CN vs AD, CN vs MCI, MCI vs AD, and CN vs MCI vs AD.

The outcomes of Experiment 4, detailed in Table 6.11, offer insights into the performance of the models. Notably, the classifiers trained for the NC vs AD classification on the NACC dataset exhibited impressive performance when applied to the ADNI dataset. SVM achieved a remarkable 99% accuracy, indicating its superiority, while RF achieved an accuracy of 98.3% (Table 6.11.a). However, RF displayed a degree of bias towards

precision. In a similar vein, when the models trained for the CN vs AD prediction task on the NACC dataset were tested on ADNI data, both SVM and RF showed higher F1 scores, yet both models demonstrated a degree of bias in terms of precision and recall (Table 6.11.b).

SVM proved effective and demonstrated balanced performance in both the classification and prediction of the CN vs MCI subset, as evidenced in Table 6.11.c and 6.11.d, respectively. Notably, SVM exhibited a strong performance in classifying MCI vs AD in the ADNI data, achieving an F1 score of 81% (Table 6.11.e). However, it experienced a drop in performance when tasked with prediction, resulting in an F1 score of 56% (Table 6.11.f).

Finally, SVM demonstrated balanced and high F1 scores, surpassing 90%, for the classification of CN vs MCI vs AD (Table 6.11.g) and maintained a commendable performance in the prediction task, achieving an F1 score of 72.8 (Table 6.11.h). These results underscore the versatility and robustness of SVM across various classification and prediction tasks.

ML Model	Accuracy%	Precision%	Recall%	F1 score%
a) Results of EXP4: NC vs AD Classification				
RF	98.3	100	80	88
SVM	99	994	984	98.9
b) Results of EXP4: NC vs AD Prediction				
RF	97.8	92.3	80	85.7
SVM	98.3	100	80	88
c) Results of EXP4: NC vs MCI Classification				
RF	98.6	99.6	98.4	99
SVM	99.6	99.6	99.8	99.7
d) Results of EXP4: NC vs MCI Prediction				
RF	90.2	86.8	91.3	89
SVM	92.4	91.3	91.3	91.3
e) Results of EXP4: MCI vs AD Classification				
RF	87.1	71.6	86.5	78.4
SVM	88.9	75.4	87.5	81
f) Results of EXP4: MCI vs AD Prediction				
RF	66.1	73	51.5	60.4
SVM	64.9	75.7	44.4	56
g) Results of EXP3: NC vs MCI vs AD Classification				
RF	89	90	89	89.2
SVM	90.5	91.4	90.5	90.7
h) Results of EXP4: NC vs MCI vs AD Prediction				
RF	72.9	73.9	72.9	72.6
SVM	73.6	75.2	73.6	72.8

TABLE 6.11: Results of EXP4: performance of RF and SVM in classification and prediction tasks using external ADNI dataset.

Results for EXP5: ML Explanations and Human understandable Rules Extraction

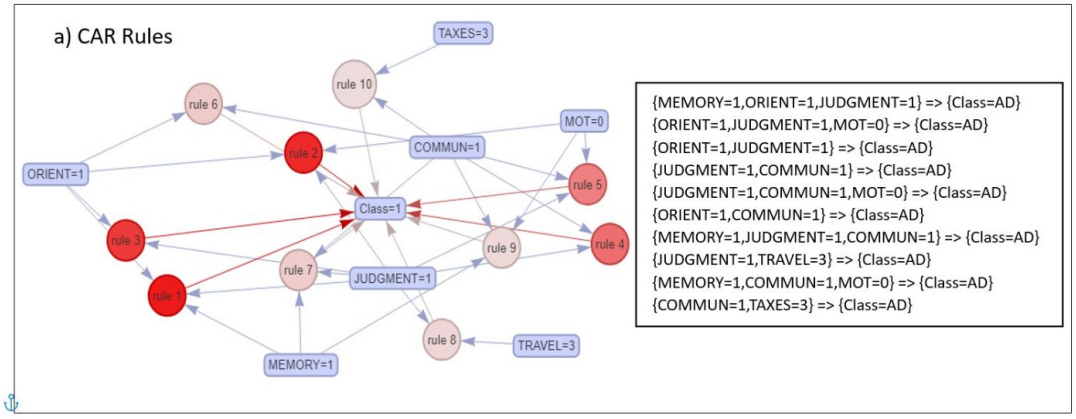
In our pursuit of understanding the intricate patterns with the data and comprehending the behaviour of ML models in classifying AD, CAR algorithm is used in EXP4. Figure 6.13.a illustrates ten representative rules extracted by CAR that are highly associated with AD. The intensity of the red colour of the circles indicates the strength of the rule, evaluated using the lift measure.

Upon analysis of the output rules, it becomes evident that AD is associated with a wide range of factors, including mild impairments in memory (MEMORY =1), orientation (ORIENT =1), judgment and problem-solving (JUDGMENT=1) and impairments in community affairs (COMMUN =1). To elucidate the values of the variables (such as MEMORY, JUDGMENT) are encoded as follows: 0 for no impairment, 0.5 for questionable impairment, 1 for mild impairment, 2 for moderate impairment, and 3 for severe

impairment. The rules shed light on the combinations of these variables with the severity levels of TRAVEL and TAXES, all of which bear a significant connection to AD.

SIRUS algorithm was utilised to extract human readable rules and to compare with the rules extracted from CAR. Figure 6.13.b shows the rules extracted from the NC vs AD subset. the first rule indicates that if the value of the variable 'JUDGMENT' is '0' the classification is likely 'NC' Conversely, if the 'JUDGMENT' value is not '0' the likelihood of 'AD' classification increases significantly. Essentially, the '0' value for the 'JUDGMENT' feature serves as a robust indicator of an individual's AD status. Similarly, another rule indicates that if the value of the 'COMMUN' variable is '0' the individual is most likely classified as 'NC' while other values suggest 'AD' The rules derived from SIRUS also unveil the co-occurrence of higher values in the TRAVEL, ORIENT, and MEMORY variables, which are associated with an elevated risk of AD.

To validate the rules generated by both models and ascertain the informativeness of these variables in the context of ML AD classification, our research venture extended to encompass the application of two model-agnostic explanation methods: SHAP and Local Interpretable Model-Agnostic Explanations (LIME). As visually depicted in Figures 6.13.c and 6.13.d, the variables that SHAP identifies as most informative include MEMORY, COMMUN, JUDGMENT, ORIENT, and BILLS. Concurrently, the insights offered by LIME emphasise the pivotal role of variables such as COMMUN, MEMORY, JUDGMENT, and ORIENT. Table 6.12 presents the informative features selected by each model, along with the common features chosen by all models. Furthermore, Table 6.13 demonstrates the performance of SVM when trained and tested using the common features extracted from Table 6.12. The results of this classifier closely align with the findings of EXP1 and EXP2, underscoring the significance of these features in influencing the model's performance.



b) SIRUS Rules

[1] "if JUDGMENT in {0} then Class=NC else Class=AD
[2] "if COMMUN in {0} then Class=NC else Class=AD
[3] "if JUDGMENT in {0.5, 1, 2, 3} & COMMUN in {0.5, 1, 2, 3} then Class= AD else Class=NC
[4] "if MEMORY in {0} then Class=NC else Class=AD
[5] "if MEMORY in {1, 2, 3} & JUDGMENT in {1, 2, 3} then Class=AD else Class=NC
[6] "if ORIENT in {0} then Class=NC else Class=AD
[7] "if MEMORY in {0, 0.5} & JUDGMENT in {0} then Class=NC else Class=AD
[8] "if ORIENT in {0} & COMMUN in {0} then Class=NC else Class=AD
[9] "if ORIENT in {0.5, 1, 2, 3} & JUDGMENT in {0} then Class=NC else Class=AD
[10] "if TRAVEL in {0} then Class=NC else Class=AD
[11] "if MEMORY in {0, 0.5} & COMMUN in {0} then Class=NC else Class=AD
[12] "if JUDGMENT in {0.5, 1, 2, 3} & TRAVEL in {1, 2, 3} then Class=AD else Class=NC
[13] "if MEMORY in {0} & COMMUN in {0} then Class=NC else Class=AD
[14] "if MEMORY in {0.5, 1, 2, 3} & ORIENT in {0.5, 1, 2, 3} then Class=AD else Class=NC
[15] "if COMMUN in {0.5, 1, 2, 3} & TRAVEL in {1} then Class=AD else Class=NC

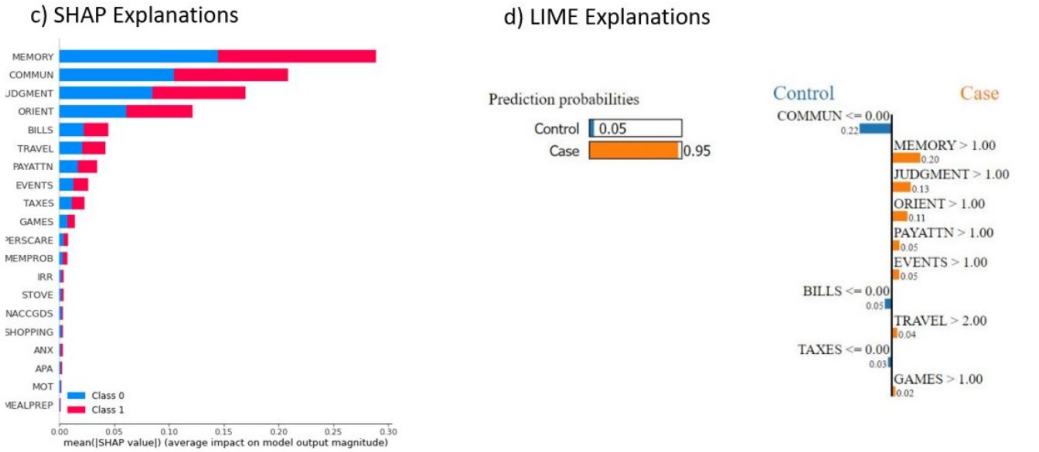


FIGURE 6.13: Explanations and rules extraction for NC vs AD subset: a) Visualisation of representative associations and corresponding written rules between multiple factors and AD in NC vs AD, b) List of rules output by SIRUS model, c) explanation provided by SHAP model, d) explanation provided by LIME model for a single instance of the test set.

Feature selected by CAR	Features selected by SIRUS	Features selected by SHAP	Features selected by LIME	Common features selected by all models
MEMORY	JUDGMENT	MEMORY	COMMUN	MEMORY
ORIENT	COMMUN	COMMUN	MEMORY	COMMUN
JUDGMENT	MEMORY	JUDGMENT	JUDGMENT	JUDGMENT
MOT	ORIENT	ORIENT	ORIENT	ORIENT
COMMUN	TRAVEL	BILLS	PAYATTN	
TRAVEL		TRAVEL	EVENTS	
TAXES				

TABLE 6.12: Features selected from explanations by models for NC vs AD data subset.

Data	Accuracy%	Precision%	Recall%	F1 score%
NC vs AD	97.2	97.7	96.3	97
NC vs MCI	88.8	79.7	87.2	83.3
MCI vs AD	86	90.4	86.3	88.3
NC vs MCI vs AD	83.5	84.2	83.5	83.8

TABLE 6.13: Performance of SVM trained and tested using common features selected by explanation models (from Table 6.12))

In a similar vein, the patterns discerned from the MCI vs AD data subset are systematically extracted using the CAR algorithm, as depicted in Figure 6.14.a. This visualisation encapsulates ten rules of significance in the context of AD. These rules were selected from a comprehensive number of variables based on their discernible influence on AD. Five pivotal variables—ORIENT, MEMORY, COMMUN, BILLS, and TAXES—emerge as the most robust influencers in the realm of AD. The amalgamation of these variables with elevated values strongly correlates with AD, a consistent pattern observed across both the NC vs AD data subset, as presented in Figure 6.13.

Furthermore, the SIRUS algorithm was utilised to extract rules from the MCI vs AD data subset. As elucidated in Figure 6.14.b, the extracted rules unveil that when the feature 'JUDGMENT' assumes a value of either 0 or 0.5, the likelihood of classification as MCI predominates. Conversely, when 'JUDGMENT' adopts any other value, the individual's classification tends toward AD. Similarly, the second rule articulates that when the variable 'MEMORY' manifests values of 0 or 0.5, the probability of MCI classification is accentuated. Intriguingly, a high value associated with 'MEMORY,' signifying moderate or severe memory impairment, distinctly inclines the individual towards an AD diagnosis. These rules cogently imply that combinations of variables with high values generally align with an AD classification, resonating with the outcomes of the CAR algorithm.

Figures 6.14.c and 6.14.d offer insights into the explanations provided by SHAP and LIME, respectively. Both models consistently underscore the pivotal roles of COMMUN, ORIENT, and JUDGMENT as informative variables significantly influencing the AD classification, which is in line with both CAR and SIRUS.

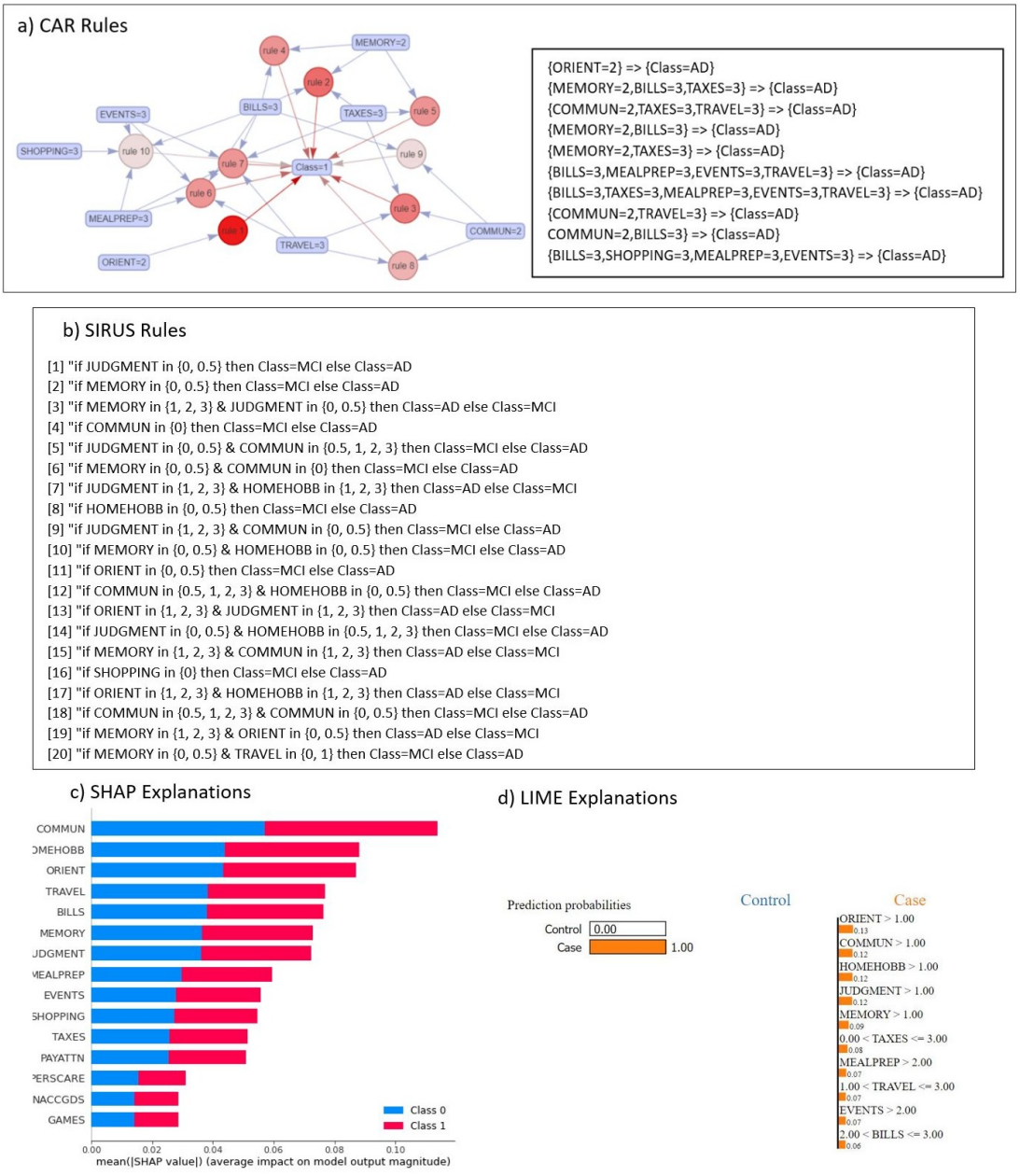


FIGURE 6.14: Explanations and rules extraction for MCI vs AD subset: a) Visualisation of representative associations and corresponding written rules between multiple factors and AD in MCI vs AD, b) List of rules output by SIRUS model, c) explanation provided by SHAP model, d) explanation provided by LIME model for a single instance of the test set.

Discussion

The experimental results for Exp1-Exp4 demonstrates the capabilities of ML models in the classification of AD patients from those with NC or MCI. Out of the four models utilised, RF, SVM, NB and KNN, RF and SVM models consistently achieved the highest accuracy, precision and recall scores across all tasks. These models effectively discriminated between NC and AD subjects, as well as between NC and MCI, MCI and AD subjects. While NB and KNN models also demonstrated considerable accuracy, they generally scored lower than RF and SVM.

This indicates the substantial potential of RF and SVM models for accurate AD diagnosis. It is noteworthy that RF has previously shown high accuracy in the analysis of multi-modal data to predict the conversion of MCI to AD [206]. Additionally, in healthcare domain, RF showed better classification of risk assessment of coronary heart disease than other classifiers [207]. These observations can be attributed to the capacity of RF and SVM models to efficiently process large datasets [208], making them well-suited for large-scale medical diagnoses. Furthermore, these models excel in data generalisation, rendering them more adept at handling the intricacies of medical diagnosis. Consequently, RF and SVM models are better positioned to provide AD diagnosis when compared to NB and KNN classifiers.

ML has increasingly been employed in research to predict the progressions of AD stages. For example, work presented in study [209] developed a hybrid ML framework for the analysis of longitudinal data to predict the prognosis of dementia in patients with MCI. While their model achieved high accuracy of 87.5% using RF, it displayed instability across various performance measures. Notably, the model exhibited a stronger bias towards sensitivity (92.9%) at the expense of specificity, which was only 58.3%. Another study [210] identified and utilised 15 clinical variables predicting MCI converters reporting 71%, 67.7% and 71.7% for accuracy¸ sensitivity and specificity, respectively. In contrast, our ML classifiers demonstrate not only high accuracy, precision and recall scores when applied to classification tasks but also achieved robust outcomes when trained and tested as a predictive tool to estimate the cognitive state of a person four years in the future. Specifically, the classifiers excels at identifying subtle changes in cognitive development over time, thus making it a valuable asset in predicting potential changes in cognitive health. Moreover, our approach is found to be reliable and robust, with a high degree of consistency in its predictions over multiple trials (i.e NC vs AD;

MCI vs AD; and NC vs MCI vs AD). This means that it can be used to reliably forecast a person's cognitive state in the future.

The results section provides convincing evidence of the efficacy of SVM in both classification and prediction tasks. SVM performed well, not only when tested on the NACC testing hold subset but also when evaluated on an external ADNI dataset for various tasks (Table 6.11). This was achieved through the feature selection method, which significantly reduced the number of features from 64 to only 21 features. Despite the substantial reduction in feature space, the results demonstrate that the selected features are highly effective in differentiating AD cases.

It is important to note that the objective of this research is not only to obtain better AD classification but also gaining insight into the influential factors that are important for the classifiers' decision making. To this end, this study conducted a series of experiments to identify the most important features and to understand the underlying relationships that exist between them.

In pursuit of these objectives, we employed two rule extraction methods, CARs and SIRUS, to extract human-readable rules associated with AD. CARs, for instance, utilised seven of the 21 features. While SIRUS used five features to establish its list of most dominant rules. Intriguingly, both algorithms identified common features as shown in Table 6.12 this overlap strongly suggests that the rules generated by these algorithms exhibit a significant degree of similarity, enhancing the confidence in extracted rules accuracy and reliability. The utilisation of two distinct rule extraction methods, with the majority of the rules aligning, underscores the precision and trustworthiness of extracted rules.

Furthermore, the features identified as important by CAR and SUIRS underwent additional validation through SHAP and LIME models, which were utilised to elucidate the decisions made by the top-performing classifier. Notably, both SHAP and LIME consistently identified crucial features that aligned with the rules extracted by CAR and SUIRS (Table 6.12). This alignment in feature selection across diverse models significantly strengthens the overall robustness and reliability of our findings.

It can be noticed that the CAR is more precise than SIRUS in terms of generating the rules. For instance, the first rule extracted by CAR from the NC vs AD dataset (Figure 6.13) specifies that if an individual has the variables MEMERY, JUDGMENT and ORIENT with the value of 1, then it is a case of AD. In contrast, SIRUS, tends to provide generalised predictions. For example, the first rule generated by SIRUS suggest

that if the variable JUDGMENT assume the value of 0, then it's more likely the individual to belong to the class NC. However, if the value of JUDGMENT is not 0 (i.e. 0.5,1,2 or 3) then individual is likely to belong to the class AD. This shows that SIRUS can make broader observations and predictions than CAR, which tends to be more specific in its rules.

The findings highlight the collective significance of the features MEMORY, JUDGMENT, ORIENT, and COMMUN are collectively significant in assessing the risk of developing AD as indicated by all models. These combined features play a crucial role in predicting the likelihood of an individual being diagnosed with AD. Literature supports the Clinical Dementia Rating (CDR) as a valuable tool for detecting MCI and AD. [211] [212]. Research conducted by [213] underscores the significance of considering functional information, namely JUDGMENT, COMMUN, and HOMEHOBB, as assessed by the CDR, when evaluating individuals with MCI. The intact group included individuals with a rating of 0 in all three categories or a rating of 0.5 in one of the three categories. The impaired group comprised individuals with a rating of 0.5 in two or more of the three IADL categories or a rating of 1 in any one of the categories. The results of the experiments have been instrumental in providing key insights into the efficacy of CDR in the prediction of AD.

6.4 Chapter Summary

Diagnosing Alzheimer's disease at an early stage will greatly help so many people in the future. Machine learning models aiming to classify individuals' cognitive state achieved promising performance when trained and tested using only 22 features selected by Boruta algorithm. The models were also found to be effective in predicting the cognitive state after four years. the findings of this research project have established that Random Forest is a powerful tool for predicting the risk of developing AD. Two rule extraction approaches are utilised to find the most influential features on AD. The experiments have shown that MEMERY, JUDGMENT, ORIENT are among the most significant factors in determining the risk of developing AD, and that these factors can be effectively used to predict the chances of a person being diagnosed with the condition. Furthermore, the results of the experiments have provided valuable information about the importance of Clinical Dementia Rating in prediction AD as these variables fall within the domains

that this clinical tool use for grading the relative severity of AD. This research has the potential to revolutionise our understanding of AD and open up new possibilities for researchers looking to utilise explainable ML methods to unlock hidden knowledge in other diseases.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

The thesis aimed at using ML and AI approaches to identify and diagnose AD at its early stages using genetic data. The diagnosis of AD is a challenging task because it is difficult to detect it before the onset of clear clinical symptoms. The work contributes to the development of risk prediction and classification algorithms that can utilise genetic data.

The research was motivated by the ongoing difficulty in diagnosing AD early and the absence of reliable tools that can utilise complex, high-dimensional data to assist in this task. The focus it contributes to the narrower and clearly defined goal of improving diagnosis and prediction through data-driven, AI-based methodologies. The study results show that AI models can accurately classify AD when trained on a subset of GWAS data features. The study conducted feature selection with the Boruta algorithm and found a subset of SNPs including rs6116375 in PRNP and rs2075650 in TOMM40, which were the most useful for prediction. The results obtained show that AI can discover significant genetic markers that might not be evident through standard statistical approaches. Domain adaptation techniques along with TL have been used to enhance the generalisation capabilities of models across various datasets. The results suggest that there is a promising direction for knowledge transfer across domains, for instance, from a population or disease area that has large datasets to another that may have limited data.

A comparison of different architectures showed that wide neural network models produced the best results. The use of a simple wide architecture reached classification accuracy rates of 99% when distinguishing AD from normal controls using GWAS data. It means that it is possible to get good results in genetic classification tasks using lightweight models with simple structures.

The study also explored the classification of AD from a large, multi-source dataset from the NACC. ML models were able to predict AD and forecast future cognitive states over a four-year period using only 22 genetic features. The results obtained from the study demonstrate the capability of these models to be used in cognitive prediction in the future.

To enhance transparency of ML models that considered as ‘black boxes’, the thesis employs explainable AI techniques to extract decision rules from the models. The analysis revealed that MEMORY, JUDGMENT, and ORIENT cognitive domains are crucial for AD classification, which is in line with the Clinical Dementia Rating scale.

However, while the results are encouraging, some important issues still need to be addressed in order to make these solutions useful in practice. A significant barrier to the use of genetic information in clinical settings is the lack of genetic data. GWAS data is not collected as a routine assessment or imaging, which makes it difficult to use these models in practice. Therefore, genetic testing would need to become more integrated into clinical workflows, and this should be supported by user-friendly tools for data interpretation.

Federated learning, a privacy-preserving technique may help make the clinical adoption of these models more possible. These methods enable models to be trained across multiple institutions without the need to share raw data, which helps to ensure confidentiality while increasing the availability of data. Other methods like differential privacy and secure multiparty computation also have a role to play in the safe and ethical deployment of AI in healthcare.

It is important that healthcare professionals are involved in the development and validation of AI models. Their feedback can assist in the model development, enhance user-friendliness, and ensure that the model is relevant to the clinical setting. Real-world validation studies, preferably within prospective clinical trials, are required to assess the performance, usability, and acceptance of models in clinical environments. Through ongoing collaboration with clinicians, these AI tools can evolve into practical systems that can assist medical professionals and enhance the early detection of AD.

This thesis offers a valuable contribution to the use of AI in healthcare by developing highly efficient models for predicting AD risk. It shows the capability of ML to improve the ability to distinguish and forecast cognitive decline and provides a foundation for future work on clinical implementation.

7.2 Implications of Study on Practice

The study of ML in analysing GWAS data and building classification models for AD has significant implications for both research and practical applications. GWAS is a widely used approach to identify genetic variations associated with complex diseases like AD. ML techniques can complement GWAS analysis by extracting valuable insights from large-scale genomic data and enhancing the prediction accuracy of AD classification models.

One of the primary benefits of applying ML to GWAS data analysis is the ability to identify relevant features or genetic markers associated with AD. Traditional statistical methods in GWAS often rely on identifying single genetic variants that have a significant association with the disease. However, the genetic architecture of complex diseases like AD is highly complex, involving multiple genetic variants and interactions between them. ML algorithms, such as RF, SVM, and neural networks, can effectively handle the high-dimensional nature of GWAS data and identify relevant features or combinations of features that contribute to disease risk.

By selecting relevant features, ML models can improve the accuracy of AD classification. Traditional methods often rely on a limited set of genetic markers, which may not capture the full complexity of the disease. ML algorithms, on the other hand, can leverage a broad range of features and identify nonlinear relationships between genetic variations and disease status. This enables the construction of more accurate and robust AD classification models. These models can be used for risk prediction, early diagnosis, and personalised treatment strategies.

In practical applications, the use of ML in analysing GWAS data and building AD classification models holds promise for personalized medicine and precision healthcare. By considering an individual's genetic variations, as well as other clinical and environmental factors, ML models can provide personalised risk assessments for AD. This can enable early interventions, lifestyle modifications, and targeted therapies to delay or prevent disease progression.

Human-readable rules extracted in the study help identify the specific features or variables that contribute significantly to the AD classification. By examining the rules, clinicians and researchers can gain insights into the relative importance and impact of different features in the prediction. This information can guide further investigations and prioritise features for future studies or potential interventions.

7.3 Limitations of the Study

While ML techniques have shown promise in selecting relevant features and building classification models for AD, there are several limitations to consider. These limitations include:

Generalisability across populations: ML models trained on a specific population may not

generalise well to other populations or ethnic groups. Genetic factors can vary among populations, leading to differences in disease manifestations and genetic markers. It is important to validate and evaluate the generalisability of ML models across diverse populations to ensure their applicability in different settings.

Incomplete representation of genetic variation: Focusing only on statistically significant SNPs may result in an incomplete representation of the genetic variation associated with AD. GWAS studies typically employ hypothesis testing to identify SNPs that show a significant association with the disease. However, this approach may overlook SNPs with smaller effect sizes or those involved in complex interactions. By using only statistically significant SNPs, important genetic variations that contribute to AD risk may be missed, leading to an incomplete understanding of the disease.

Transparency and explainability: Many ML models, such as DL algorithms, are often considered black boxes, making it challenging to understand and explain their decisions. Ethical considerations call for transparency and interpretability, enabling clinicians, patients, and stakeholders to understand the rationale behind the model's predictions. Efforts to develop techniques for explaining ML model decisions and promoting transparency are crucial to address these limitations.

7.4 Future work

The research presented in this paper shows how ML and AI methods work for identifying AD using genetic and multi-source data yet various important research directions need to be investigated. A fundamental direction for expansion involves creating unified predictive models that incorporate genetic information along with clinical data. In the current study, these data sources were examined independently. A combined analysis of these data sources would increase model robustness and diagnostic accuracy. Also the addition of MRI imaging or other image modality to the model would increase its capacity to identify brain alterations linked to AD especially during preclinical phases. A vital next research direction should focus on applying multi-view learning or data fusion approaches for processing and uniting these different input types.

Future research should prioritise the development of ML models which demonstrate both broad applicability and inclusivity for different population groups. The current predictive models show limited applicability due to their geographic and demographic

constraints that create performance biases. Domain adaptation techniques should be applied to enhance model robustness by allowing them to learn across populations with different distributions. The development of tools requires training and evaluation on diverse demographic and ethnic datasets to ensure these tools can be used broadly without exacerbating existing health disparities.

Real-world adoption depends heavily on maintaining model transparency and interpretability. The research presented in this thesis used rule extraction and explainable AI techniques, but additional investigations should analyse counterfactual explanations and attention mechanisms in deep models as well as SIRUS models with inherent interpretability. Such methods would enable more straightforward model prediction understanding that helps clinicians comprehend the classification decisions. Interpretability models should derive their rules and insights through clinician validation which may also include small-scale clinical studies to enhance trust and clinical usability.

The development process requires domain expert collaboration to improve model design and choose meaningful clinical features and match tools to diagnostic procedures. Multidisciplinary teams combining data scientists with neurologists and geneticists enable the creation of interfaces and systems which meet both technical requirements and clinical needs. Prospective clinical validation studies should be planned for the assessment of both prediction accuracy and the actual effect of AI tools on clinical diagnostic choices.

The application of federated learning combined with privacy-preserving AI techniques represents an exciting future direction. The training method enables models to use distributed data across multiple institutions without needing to exchange raw data thus resolving privacy issues while providing access to bigger datasets. The implementation of these approaches would improve the development of more robust and generalisable models that fulfil ethical and legal standards.

The methods developed throughout this research can be adjusted to detect other chronic or neurodegenerative diseases. Large AD datasets could provide a starting point for transfer learning to speed up model development and extend the utility of these tools toward rare forms of dementia. Causal inference techniques have the potential to reveal disease progression mechanisms which could reveal therapeutic intervention points and new treatment targets. A long-term goal should aim to develop clinical tools from high-performing experimental models for early diagnosis and personalised treatment planning and long-term cognitive health monitoring.

Bibliography

[1] W. H. Organization, "The global dementia observatory reference guide," report, World Health Organization, 2018.

[2] S. Herrera-Espejo, B. Santos-Zorrozua, P. Álvarez González, E. Lopez-Lopez, and Garcia-Orad, "A systematic review of microrna expression as biomarker of late-onset alzheimer's disease," *Molecular Neurobiology*, vol. 56, no. 12, pp. 8376–8391, 2019.

[3] T. S. Wingo, J. J. Lah, A. I. Levey, and D. J. Cutler, "Autosomal recessive causes likely in early-onset alzheimer disease," *Archives of neurology*, vol. 69, no. 1, pp. 59–64, 2012.

[4] R. Cacace, K. Sleegers, and C. Van Broeckhoven, "Molecular genetics of early-onset alzheimer's disease revisited," *Alzheimer's Dementia*, vol. 12, no. 6, pp. 733–748, 2016.

[5] D. Harman, "Alzheimer's disease pathogenesis: role of aging," *Annals of the New York Academy of Sciences*, vol. 1067, no. 1, pp. 454–460, 2006.

[6] M. Taiana, J. Nascimento, and A. Bernardino, "On the purity of training and testing data for learning: The case of pedestrian detection," *Neurocomputing*, vol. 150, pp. 214–226, 2015.

[7] K. B. Johnson, W. Wei, D. Weeraratne, M. E. Frisse, K. Misulis, K. Rhee, J. Zhao, and J. L. Snowdon, "Precision medicine, ai, and the future of personalized health care," *Clinical and translational science*, vol. 14, no. 1, pp. 86–93, 2021.

[8] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[9] M. Khojaste-Sarakhsi, S. S. Haghighi, S. F. Ghomi, and E. Marchiori, "Deep learning for alzheimer's disease diagnosis: A survey," *Artificial intelligence in medicine*, vol. 130, p. 102332, 2022.

[10] S. Qiu, M. I. Miller, P. S. Joshi, J. C. Lee, C. Xue, Y. Ni, Y. Wang, I. De Anda-Duran, P. H. Hwang, J. A. Cramer, *et al.*, "Multimodal deep learning for alzheimer's disease dementia assessment," *Nature communications*, vol. 13, no. 1, p. 3404, 2022.

[11] H. A. Helaly, M. Badawy, and A. Y. Haikal, "Deep learning approach for early detection of alzheimer's disease," *Cognitive computation*, vol. 14, no. 5, pp. 1711–1727, 2022.

[12] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, "Multimodal deep learning models for early detection of alzheimer's disease stage," *Scientific reports*, vol. 11, no. 1, p. 3254, 2021.

[13] M. Liu, D. Zhang, D. Shen, and A. D. N. Initiative, "Ensemble sparse classification of alzheimer's disease," *NeuroImage*, vol. 60, no. 2, pp. 1106–1116, 2012.

[14] A. Mehmood, S. Yang, Z. Feng, M. Wang, A. S. Ahmad, R. Khan, M. Maqsood, and M. Yaqub, "A transfer learning approach for early diagnosis of alzheimer's disease on mri images," *Neuroscience*, vol. 460, pp. 43–52, 2021.

[15] "Dementia." World Health Organization. https://www.who.int/news-room/fact-sheets/detail/dementia (accessed 26/3/2023).

[16] ©, 2022, IEEE, Reprinted, with, permission, from, A. Alatrany, A. Hussain, J. Mustafina, and D. Al-Jumeily, "Machine learning approaches and applications in genome wide association study for alzheimer's disease: A systematic review," *IEEE Access*, 2022.

[17] J. E. Brush, J. Sherbino, and G. R. Norman, "How expert clinicians intuitively recognize a medical diagnosis," *The American Journal of Medicine*, vol. 130, no. 6, pp. 629–634, 2017.

[18] R. S. Turner, T. Stubbs, D. A. Davies, and B. C. Albensi, "Potential new approaches for diagnosis of alzheimer's disease and related dementias," *Frontiers in neurology*, vol. 11, p. 496, 2020.

[19] S. Amari, *The handbook of brain theory and neural networks*. MIT press, 2003.

[20] "What are the parts of the nervous system?." https://www.nichd.nih.gov/health/topics/neuro/conditioninfo/parts: :text=The%20nervous%20system%20has%20two,all%20parts%20of%20the%20body. Accessed: 10/3/2023.

[21] X. Xia, Y. Wang, Y. Qin, S. Zhao, and J. C. Zheng, "Exosome: a novel neurotransmission modulator or non-canonical neurotransmitter?," *Ageing research reviews*, vol. 74, p. 101558, 2022.

[22] C. R. Noback, D. A. Ruggiero, N. L. Strominger, and R. J. Demarest, *The human nervous system: structure and function*. Springer Science Business Media, 2005.

[23] C. L. Koo, M. J. Liew, M. S. Mohamad, and A. H. Mohamed Salleh, "A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology," *BioMed Research International*, vol. 2013, p. 432375, 2013.

[24] G. S. Bloom, "Amyloid- and tau: the trigger and bullet in alzheimer disease pathogenesis," *JAMA Neurol*, vol. 71, no. 4, pp. 505–8, 2014.

[25] G.-f. Chen, T.-h. Xu, Y. Yan, Y.-r. Zhou, Y. Jiang, K. Melcher, and H. E. Xu, "Amyloid beta: structure, biology and structure-based therapeutic development," *Acta Pharmacologica Sinica*, vol. 38, no. 9, pp. 1205–1235, 2017.

[26] "Brain Tour Part 2." Alzheimer's Association. https://www.alz.org/alzheimers-dementia/what-is-alzheimers/brain$_t our_p art_2 (accessed 11/3/2023)$.

[27] Y. L. Rao, B. Ganaraja, B. V. Murlimanju, T. Joy, A. Krishnamurthy, and A. Agrawal, "Hippocampus and its involvement in alzheimer's disease: a review," *3 Biotech*, vol. 12, no. 2, p. 55, 2022.

[28] B. Klimova, P. Maresova, M. Valis, J. Hort, and K. Kuca, "Alzheimer's disease and language impairments: social intervention and medical treatment," *Clinical interventions in aging*, pp. 1401–1408, 2015.

[29] V. Senanarong, J. Cummings, L. Fairbanks, M. Mega, D. Masterman, S. O'connor, and T. Strickland, "Agitation in alzheimer's disease is a manifestation of frontal lobe dysfunction," *Dementia and geriatric cognitive disorders*, vol. 17, no. 1-2, pp. 14–20, 2004.

[30] M. M. Rahman and C. Lendel, "Extracellular protein components of amyloid plaques and their roles in alzheimer's disease pathology," *Molecular Neurodegeneration*, vol. 16, no. 1, p. 59, 2021.

[31] Y. Hou, X. Dan, M. Babbar, Y. Wei, S. G. Hasselbalch, D. L. Croteau, and V. A. Bohr, "Ageing as a risk factor for neurodegenerative disease," *Nature Reviews Neurology*, vol. 15, no. 10, pp. 565–581, 2019.

[32] C. Van Cauwenberghe, C. Van Broeckhoven, and K. Sleegers, "The genetic landscape of alzheimer disease: clinical implications and perspectives," *Genetics in Medicine*, vol. 18, no. 5, pp. 421–430, 2016.

[33] M. N. Wainaina, Z. Chen, and C. Zhong, "Environmental factors in the development and progression of late-onset alzheimer's disease," *Neuroscience bulletin*, vol. 30, pp. 253–270, 2014.

[34] C. Y. Santos, P. J. Snyder, W.-C. Wu, M. Zhang, A. Echeverria, and J. Alber, "Pathophysiologic relationship between alzheimer's disease, cerebrovascular disease, and cardiovascular risk: a review and synthesis," *Alzheimer's Dementia: Diagnosis, Assessment Disease Monitoring*, vol. 7, pp. 69–87, 2017.

[35] W. S. Bush and J. H. Moore, "Chapter 11: Genome-wide association studies," *PLoS Comput Biol*, vol. 8, no. 12, p. e1002822, 2012.

[36] S. Behjati and P. S. Tarpey, "What is next generation sequencing?," *Archives of disease in childhood - Education amp;amp; practice edition*, vol. 98, no. 6, p. 236, 2013.

[37] X. Zhang, S. Huang, Z. Zhang, and W. Wang, "Chapter 10: Mining genome-wide genetic markers," *PLoS computational biology*, vol. 8, no. 12, p. e1002828, 2012.

[38] S. Qiu, M. I. Miller, P. S. Joshi, J. C. Lee, C. Xue, Y. Ni, Y. Wang, I. De Anda-Duran, P. H. Hwang, J. A. Cramer, B. C. Dwyer, H. Hao, M. C. Kaku, S. Kedar, P. H. Lee, A. Z. Mian, D. L. Murman, S. O'Shea, A. B. Paul, M.-H. Saint-Hilaire, E. Alton Sartor, A. R. Saxena, L. C. Shih, J. E. Small, M. J. Smith, A. Swaminathan, C. E. Takahashi, O. Taraschenko, H. You, J. Yuan, Y. Zhou, S. Zhu, M. L. Alosco, J. Mez, T. D. Stein, K. L. Poston, R. Au, and V. B. Kolachalama, "Multimodal deep learning for alzheimer's disease dementia assessment," *Nature Communications*, vol. 13, no. 1, p. 3404, 2022.

[39] S. Liu, A. V. Masurkar, H. Rusinek, J. Chen, B. Zhang, W. Zhu, C. Fernandez-Granda, and N. Razavian, "Generalizable deep learning model for early alzheimer's disease detection from structural mris," *Scientific Reports*, vol. 12, no. 1, p. 17106, 2022.

[40] J. S. Kim, J. W. Han, J. B. Bae, D. G. Moon, J. Shin, J. E. Kong, H. Lee, H. W. Yang, E. Lim, J. Y. Kim, L. Sunwoo, S. J. Cho, D. Lee, I. Kim, S. W. Ha, M. J. Kang, C. H. Suh, W. H. Shim, S. J. Kim, and K. W. Kim, "Deep learning-based diagnosis of alzheimer's disease using brain magnetic resonance images: an empirical study," *Scientific Reports*, vol. 12, no. 1, p. 18007, 2022.

[41] P. Moore, T. Lyons, J. Gallacher, and A. D. N. Initiative, "Random forest prediction of alzheimer's disease using pairwise selection from time series data," *PloS one*, vol. 14, no. 2, p. e0211558, 2019.

[42] N. Alexander, D. C. Alexander, F. Barkhof, and S. Denaxas, "Identifying and evaluating clinical subtypes of alzheimer's disease in care electronic health records using unsupervised machine learning," *BMC medical informatics and decision making*, vol. 21, no. 1, pp. 1–13, 2021.

[43] G. S. Araújo, M. R. B. Souza, J. R. M. Oliveira, and I. G. Costa, "Random forest and gene networks for association of snps to alzheimer's disease," in *Advances in Bioinformatics and Computational Biology* (J. C. Setubal and N. F. Almeida, eds.), pp. 104–115, Springer International Publishing.

[44] N. Briones and V. Dinu, "Data mining of high density genomic variant data for prediction of alzheimer's disease risk," *BMC Med Genet*, vol. 13, p. 7, 2012.

[45] T.-T. Nguyen, J. Z. Huang, Q. Wu, T. T. Nguyen, and M. J. Li, "Genome-wide association data classification and snps selection using two-stage quality-based random forests," in *BMC genomics*, vol. 16, pp. 1–11, Springer.

[46] M. M. Abd El Hamid, Y. M. Omar, and M. S. Mabrouk, "Identifying genetic biomarkers associated to alzheimer's disease using support vector machine," in *2016 8th Cairo International Biomedical Engineering Conference (CIBEC)*, pp. 5–9, IEEE.

[47] Y.-C. Chang, J.-T. Wu, M.-Y. Hong, Y.-A. Tung, P.-H. Hsieh, S. W. Yee, K. M. Giacomini, Y.-J. Oyang, and C.-Y. Chen, "Genepi: gene-based epistasis discovery using machine learning," *BMC bioinformatics*, vol. 21, pp. 1–13, 2020.

[48] J. De Velasco Oriol, E. E. Vallejo, K. Estrada, and J. G. Tamez Pena, "Benchmarking machine learning models for late-onset alzheimer's disease prediction from genomic data," *BMC bioinformatics*, vol. 20, 2019.

[49] B. L. Romero-Rosales, J. G. Tamez-Pena, H. Nicolini, M. G. Moreno-Treviño, and V. Trevino, "Improving predictive models for alzheimer's disease using gwas data by incorporating misclassified samples modeling," *PLoS One*, vol. 15, no. 4, p. e0232103, 2020.

[50] M. Aflakparast, H. Salimi, A. Gerami, M. Dubé, S. Visweswaran, and A. Masoudi-Nejad, "Cuckoo search epistasis: a new method for exploring significant genetic interactions," *Heredity*, vol. 112, no. 6, pp. 666–674, 2014.

[51] M. E. Stokes, M. M. Barmada, M. I. Kamboh, and S. Visweswaran, "The application of network label propagation to rank biomarkers in genome-wide alzheimer's data," *BMC genomics*, vol. 15, pp. 1–13, 2014.

[52] L. Li, Y. Yang, Q. Zhang, J. Wang, J. Jiang, and A. D. Neuroimaging Initiative, "Use of deep-learning genomics to discriminate healthy individuals from those with alzheimer's disease or mild cognitive impairment," *Behav Neurol*, vol. 2021, p. 3359103, 2021.

[53] J. H. Moore, P. C. Andrews, R. S. Olson, S. E. Carlson, C. R. Larock, M. J. Bulhoes, J. P. O'Connor, E. M. Greytak, and S. L. Armentrout, "Grid-based stochastic search for hierarchical gene-gene interactions in population-based genetic studies of common human diseases," *BioData mining*, vol. 10, no. 1, pp. 1–16, 2017.

[54] M. Arnal Segura, G. Bini, D. Fernandez Orth, E. Samaras, M. Kassis, F. Aisopos, J. Rambla De Argila, G. Paliouras, P. Garrard, and C. Giambartolomei, "Machine learning methods applied to genotyping data capture interactions between single nucleotide variants in late onset alzheimer's disease," *Alzheimer's Dementia: Diagnosis, Assessment Disease Monitoring*, vol. 14, no. 1, p. e12300, 2022.

[55] M. M. Abd El Hamid, M. Shaheen, Y. M. K. Omar, and M. S. Mabrouk, "Discovering epistasis interactions in alzheimer's disease using integrated framework of ensemble learning and multifactor dimensionality reduction (mdr)," *Ain Shams Engineering Journal*, vol. 14, no. 7, p. 101986, 2023.

[56] K. Nho, L. Shen, S. Kim, S. L. Risacher, J. D. West, T. Foroud, C. R. Jack, M. W. Weiner, and A. J. Saykin, "Automatic prediction of conversion from mild cognitive impairment to probable alzheimer's disease using structural magnetic resonance imaging," *AMIA Annu Symp Proc*, vol. 2010, pp. 542–6, 2010.

[57] M. Liu, D. Cheng, K. Wang, Y. Wang, and A. D. N. Initiative, "Multi-modality cascaded convolutional neural networks for alzheimer's disease diagnosis," *Neuroinformatics*, vol. 16, pp. 295–308, 2018.

[58] Y. AbdulAzeem, W. M. Bahgat, and M. Badawy, "A cnn based framework for classification of alzheimer's disease," *Neural Computing and Applications*, vol. 33, pp. 10415–10428, 2021.

[59] F. Zhang, B. Pan, P. Shao, P. Liu, S. Shen, P. Yao, and R. X. Xu, "A single model deep learning approach for alzheimer's disease diagnosis," *Neuroscience*, vol. 491, pp. 200–214, 2022.

[60] T. Wang, R. G. Qiu, and M. Yu, "Predictive modeling of the progression of alzheimer's disease with recurrent neural networks," *Scientific reports*, vol. 8, no. 1, pp. 1–12, 2018.

[61] M. Nguyen, T. He, L. An, D. C. Alexander, J. Feng, B. T. Yeo, and A. D. N. Initiative, "Predicting alzheimer's disease progression using deep recurrent neural networks," *NeuroImage*, vol. 222, p. 117203, 2020.

[62] R. Cui, M. Liu, and A. D. N. Initiative, "Rnn-based longitudinal analysis for diagnosis of alzheimer's disease," *Computerized Medical Imaging and Graphics*, vol. 73, pp. 1–10, 2019.

[63] N. Kulkarni and V. Bairagi, "Extracting salient features for eeg-based diagnosis of alzheimer's disease using support vector machine classifier," *IETE Journal of Research*, vol. 63, no. 1, pp. 11–22, 2017.

[64] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55, no. 3, pp. 856–867, 2011.

[65] M. A. Scelsi, R. R. Khan, M. Lorenzi, L. Christopher, M. D. Greicius, J. M. Schott, S. Ourselin, and A. Altmann, "Genetic study of multimodal imaging alzheimer's disease progression score implicates novel loci," *Brain*, vol. 141, no. 7, pp. 2167–2180, 2018.

[66] J. Wu, Y. Chen, P. Wang, R. J. Caselli, P. M. Thompson, J. Wang, and Y. Wang, "Integrating transcriptomics, genomics, and imaging in alzheimer's disease: A federated model," *Frontiers in radiology*, vol. 1, p. 777030, 2022.

[67] T. M. Mitchell and T. M. Mitchell, *Machine learning*, vol. 1. McGraw-hill New York, 1997.

[68] W. Kratsch, J. Manderscheid, M. Röglinger, and J. Seyfried, "Machine learning in business process monitoring: A comparison of deep learning and classical approaches used for outcome prediction," *Business Information Systems Engineering*, vol. 63, no. 3, pp. 261–276, 2021.

[69] S. Sandeep, S. Ahamad, D. Saxena, K. Srivastava, S. Jaiswal, and A. Bora, "To understand the relationship between machine learning and artificial intelligence in large and diversified business organisations," *Materials Today: Proceedings*, vol. 56, pp. 2082–2086, 2022.

[70] R. S. Bavaresco, L. C. Nesi, J. L. V. Barbosa, R. S. Antunes, R. da Rosa Righi, C. A. da Costa, M. Vanzin, D. Dornelles, C. Gatti, M. Ferreira, *et al.*, "Machine learning-based automation of accounting services: An exploratory case study," *International Journal of Accounting Information Systems*, vol. 49, p. 100618, 2023.

[71] S. Birim, I. Kazancoglu, S. K. Mangla, A. Kahraman, and Y. Kazancoglu, "The derived demand for advertising expenses and implications on sustainability: A comparative study using deep learning and traditional machine learning methods," *Annals of Operations Research*, pp. 1–31, 2022.

[72] J.-A. Choi and K. Lim, "Identifying machine learning techniques for classification of target advertising," *ICT Express*, vol. 6, no. 3, pp. 175–180, 2020.

[73] M. L. Giger, "Machine learning in medical imaging," *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 512–520, 2018.

[74] M. Shehab, L. Abualigah, Q. Shambour, M. A. Abu-Hashem, M. K. Y. Shambour, A. I. Alsalibi, and A. H. Gandomi, "Machine learning in medical applications: A review of state-of-the-art methods," *Computers in Biology and Medicine*, vol. 145, p. 105458, 2022.

[75] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: methodological failures and recommendations for the future," *NPJ digital medicine*, vol. 5, no. 1, p. 48, 2022.

[76] S. Ray, "A quick review of machine learning algorithms," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 35–39.

[77] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128–138, 2017.

[78] A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé, and A. Laio, "Unsupervised learning methods for molecular simulation data," *Chemical Reviews*, vol. 121, no. 16, pp. 9722–9758, 2021.

[79] H. Alashwal, M. El Halaby, J. J. Crouse, A. Abdalla, and A. A. Moustafa, "The application of unsupervised clustering methods to alzheimer's disease," *Frontiers in computational neuroscience*, vol. 13, p. 31, 2019.

[80] L. Chen, Y. Zhai, Q. He, W. Wang, and M. Deng, "Integrating deep supervised, self-supervised and unsupervised learning for single-cell rna-seq clustering and annotation," *Genes*, vol. 11, no. 7, p. 792, 2020.

[81] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, pp. 197–227, 2016.

[82] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[83] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[84] V. Jakkula, "Tutorial on support vector machine (svm)," *School of EECS, Washington State University*, vol. 37, no. 2.5, p. 3, 2006.

[85] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[86] W. Xing and Y. Bei, "Medical health big data classification based on knn classification algorithm," *IEEE Access*, vol. 8, pp. 28808–28819, 2019.

[87] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve bayes algorithm," *Knowledge-Based Systems*, vol. 192, p. 105361, 2020.

[88] S. Kumar, K. Gupta, and M. Gupta, "Naïve bayes classifier model for detecting spam mails," *Annals of Data Science*, pp. 1–11, 2023.

[89] S. I. Gallant, "Perceptron-based learning algorithms," *IEEE Transactions on neural networks*, vol. 1, no. 2, pp. 179–191, 1990.

[90] S. Haykin, *Neural networks and learning machines, 3/E.* Pearson Education India, 2009.

[91] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, pp. 611–629, 2018.

[92] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, p. 53, 2021.

[93] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[94] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[95] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016.

[96] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM computing surveys (CSUR)*, vol. 50, no. 6, pp. 1–45, 2017.

[97] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[98] A. Larumbe, M. Ariz, J. J. Bengoechea, R. Segura, R. Cabeza, and A. Villanueva, "Improved strategies for hpe employing learning-by-synthesis approaches," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 1545–1554, 2017.

[99] L. K. Topham, W. Khan, D. Al-Jumeily, A. Waraich, and A. J. Hussain, "Gait identification using limb joint movement and deep machine learning," *IEEE Access*, vol. 10, pp. 100113–100127, 2022.

[100] W. Khan, K. Crockett, J. O'Shea, A. Hussain, and B. M. Khan, "Deception in the eyes of deceiver: A computer vision and machine learning based automated deception detection," *Expert Systems with Applications*, vol. 169, p. 114341, 2021.

[101] M. B. Kursa and W. R. Rudnicki, "Feature selection with the boruta package," *Journal of statistical software*, vol. 36, pp. 1–13, 2010.

[102] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast discovery of association rules," *Advances in knowledge discovery and data mining*, vol. 12, no. 1, pp. 307–328, 1996.

[103] F. Pohlmeyer, R. Kins, F. Cloppenburg, and T. Gries, "Interpretable failure risk assessment for continuous production processes based on association rule mining," *Advances in Industrial and Manufacturing Engineering*, vol. 5, p. 100095, 2022.

[104] E. C. Gonçalves, I. M. B. Mendes, and A. Plastino, "Mining exceptions in databases," in *AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. Proceedings 17*, pp. 1076–1081, Springer.

[105] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," in *Kdd*, vol. 98, pp. 80–86.

[106] C. Bénard, G. Biau, S. Da Veiga, and E. Scornet, "Sirus: Stable and interpretable rule set for classification," *Electronic Journal of Statistics*, vol. 15, no. 1, pp. 427–505, 2021.

[107] N. Japkowicz, "Classifier evaluation: A need for better education and restructuring," in *Proceedings of the 3rd Workshop on Evaluation Methods for Machine Learning (ICML 2008), Helsinki, Finland*, pp. 5–9, 2008.

[108] ©, 2023, IEEE, Reprinted, with, permission, from, A. Alatrany, W. Khan, A. Hussain, J. Mustafina, and D. Al-Jumeily, "Transfer learning for classification of alzheimer's disease based on genome wide data," *IEEE/ACM transactions on computational biology and bioinformatics*, 2023.

[109] G. D. Rabinovici, "Late-onset alzheimer disease," *Continuum: Lifelong Learning in Neurology*, vol. 25, no. 1, p. 14, 2019.

[110] W. S. Bush, *Genome-Wide Association Studies*, pp. 235–241. Oxford: Academic Press, 2019.

[111] J.-C. Lambert, C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, C. Bellenguez, G. Jun, A. L. DeStefano, J. C. Bis, and G. W. Beecham, "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease," *Nature genetics*, vol. 45, no. 12, pp. 1452–1458, 2013.

[112] I. E. Jansen, J. E. Savage, K. Watanabe, J. Bryois, D. M. Williams, S. Steinberg, J. Sealock, I. K. Karlsson, S. Hägg, and L. Athanasiu, "Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk," *Nature genetics*, vol. 51, no. 3, pp. 404–413, 2019.

[113] R. E. Marioni, S. E. Harris, Q. Zhang, A. F. McRae, S. P. Hagenaars, W. D. Hill, G. Davies, C. W. Ritchie, C. R. Gale, and J. M. Starr, "Gwas on family history of alzheimer's disease," *Translational psychiatry*, vol. 8, no. 1, pp. 1–7, 2018.

[114] B. W. Kunkle, B. Grenier-Boley, R. Sims, J. C. Bis, V. Damotte, A. C. Naj, A. Boland, M. Vronskaya, S. J. Van Der Lee, and A. Amlie-Wolf, "Genetic meta-analysis of diagnosed alzheimer's disease identifies new risk loci and implicates a, tau, immunity and lipid processing," *Nature genetics*, vol. 51, no. 3, pp. 414–430, 2019.

[115] M. Maciukiewicz, V. S. Marshe, A. C. Hauschild, J. A. Foster, S. Rotzinger, J. L. Kennedy, S. H. Kennedy, D. J. Müller, and J. Geraci, "Gwas-based machine learning approach to predict duloxetine response in major depressive disorder," *J Psychiatr Res*, vol. 99, pp. 62–68, 2018.

[116] G. Lee, K. Nho, B. Kang, K.-A. Sohn, and D. Kim, "Predicting alzheimer's disease progression using multi-modal deep learning approach," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[117] T. Jo, K. Nho, P. Bice, and A. J. Saykin, "Deep learning-based identification of genetic variants: application to alzheimer's disease classification," *Brief Bioinform*, vol. 23, no. 2, 2022.

[118] L. Koumakis, "Deep learning models in genomics; are we there yet?," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1466–1473, 2020.

[119] S. R. Dhruba, R. Rahman, K. Matlock, S. Ghosh, and R. Pal, "Application of transfer learning for cancer drug sensitivity prediction," *BMC bioinformatics*, vol. 19, no. 17, pp. 51–63, 2018.

[120] J. Singh, J. Hanson, K. Paliwal, and Y. Zhou, "Rna secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning," *Nature communications*, vol. 10, no. 1, pp. 1–13, 2019.

[121] Z. Zhao, L. G. Fritsche, J. A. Smith, B. Mukherjee, and S. Lee, "The construction of cross-population polygenic risk scores using transfer learning," *The American Journal of Human Genetics*, vol. 109, no. 11, pp. 1998–2008, 2022.

[122] P. Tian, T. H. Chan, Y. F. Wang, W. Yang, G. Yin, and Y. D. Zhang, "Multiethnic polygenic risk prediction in diverse populations through transfer learning," *Front Genet*, vol. 13, p. 906965, 2022.

[123] M. Muneeb, S. Feng, and A. Henschel, "Transfer learning for genotype–phenotype prediction using deep learning models," *BMC Bioinformatics*, vol. 23, no. 1, p. 511, 2022.

[124] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni)," *Alzheimer's Dementia*, vol. 1, no. 1, pp. 55–66, 2005.

[125] J. A. Webster, J. R. Gibbs, J. Clarke, M. Ray, W. Zhang, P. Holmans, K. Rohrer, A. Zhao, L. Marlowe, and M. Kaleem, "Genetic control of human brain transcript expression in alzheimer disease," *The American Journal of Human Genetics*, vol. 84, no. 4, pp. 445–458, 2009.

[126] A. Stella, E. L. Nicolazzi, C. P. Van Tassell, M. F. Rothschild, L. Colli, B. D. Rosen, T. S. Sonstegard, P. Crepaldi, G. Tosser-Klopp, S. Joost, M. Amills, P. Ajmone-Marsan, F. Bertolini, P. Boettcher, R. Boyle Onzima, D. Bradley, D. Buja, M. E. Cano Pereira, A. Carta, G. Catillo, L. Colli, P. Crepaldi, A. Crisà, M. Del Corvo, K. Daly, C. Droegemueller, S. Duruz, A. Elbeltagi, A. Esmailizadeh, O. Faco, T. Figueiredo Cardoso, C. Flury, J. F. Garcia, B. Guldbrandtsen, A. Haile, J. Hallsteinn Hallsson, M. Heaton,

V. Hunnicke Nielsen, H. Huson, S. Joost, J. Kijas, J. A. Lenstra, G. Marras, M. Milanesi, C. Minhui, M. Moaeen-ud Din, R. Morry O'Donnell, O. Moses Danlami, J. Mwacharo, E. L. Nicolazzi, I. Palhière, F. Pilla, M. Poli, J. Reecy, B. A. Rischkowsky, E. Rochat, B. Rosen, M. Rothschild, R. Rupp, B. Sayre, B. Servin, K. Silva, T. Sonstegard, G. Spangler, A. Stella, R. Steri, A. Talenti, F. Tortereau, G. Tosser-Klopp, E. Vajana, C. P. Van Tassell, W. Zhang, and C. the AdaptMap, "Adaptmap: exploring goat diversity and adaptation," *Genetics Selection Evolution*, vol. 50, no. 1, p. 61, 2018.

[127] L. Colli, M. Milanesi, A. Talenti, F. Bertolini, M. Chen, A. Crisà, K. G. Daly, M. Del Corvo, B. Guldbrandtsen, J. A. Lenstra, B. D. Rosen, E. Vajana, G. Catillo, S. Joost, E. L. Nicolazzi, E. Rochat, M. F. Rothschild, B. Servin, T. S. Sonstegard, R. Steri, C. P. Van Tassell, P. Ajmone-Marsan, P. Crepaldi, A. Stella, and C. the AdaptMap, "Genome-wide snp profiling of worldwide goat populations reveals strong partitioning of diversity and highlights post-domestication migration routes," *Genetics Selection Evolution*, vol. 50, no. 1, p. 58, 2018.

[128] C. P. D. Kottaisamy, D. S. Raj, V. Prasanth Kumar, and U. Sankaran, "Experimental animal models for diabetes and its related complications—a review," *Laboratory animal research*, vol. 37, no. 1, p. 23, 2021.

[129] T.-Y. Choi, T.-I. Choi, Y.-R. Lee, S.-K. Choe, and C.-H. Kim, "Zebrafish as an animal model for biomedical research," *Experimental & Molecular Medicine*, vol. 53, no. 3, pp. 310–317, 2021.

[130] P. Dourlen, J. Chapuis, and J.-C. Lambert, "Using high-throughput animal or cell-based models to functionally characterize gwas signals," *Current Genetic Medicine Reports*, vol. 6, pp. 107–115, 2018.

[131] M. Seto, R. L. Weiner, L. Dumitrescu, and T. J. Hohman, "Protective genes and pathways in alzheimer's disease: moving towards precision interventions," *Molecular neurodegeneration*, vol. 16, no. 1, p. 29, 2021.

[132] M. R. Nelson, G. Marnellos, S. Kammerer, C. R. Hoyal, M. M. Shi, C. R. Cantor, and A. Braun, "Large-scale validation of single nucleotide polymorphisms in gene regions," *Genome research*, vol. 14, no. 8, pp. 1664–1668, 2004.

[133] J. van Arensbergen, L. Pagie, V. D. FitzPatrick, M. de Haas, M. P. Baltissen, F. Comoglio, R. H. van der Weide, H. Teunissen, U. Võsa, L. Franke, *et al.*, "High-throughput

identification of human snps affecting regulatory element activity," *Nature genetics*, vol. 51, no. 7, pp. 1160–1169, 2019.

[134] F. Mittag, M. Römer, and A. Zell, "Influence of feature encoding and choice of classifier on disease risk prediction in genome-wide association studies," *PloS one*, vol. 10, no. 8, p. e0135832, 2015.

[135] X. Wan, C. Yang, Q. Yang, H. Xue, X. Fan, N. L. Tang, and W. Yu, "Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies," *The American Journal of Human Genetics*, vol. 87, no. 3, pp. 325–340, 2010.

[136] C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan, "Data quality control in genetic case-control association studies," *Nat Protoc*, vol. 5, no. 9, pp. 1564–73, 2010.

[137] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, and M. J. Daly, "Plink: a tool set for whole-genome association and population-based linkage analyses," *The American journal of human genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[138] V. Q. Truong, J. A. Woerner, T. A. Cherlin, Y. Bradford, A. M. Lucas, C. C. Okeh, M. K. Shivakumar, D. H. Hui, R. Kumar, and M. Pividori, "Quality control procedures for genome-wide association studies," *Current Protocols*, vol. 2, no. 11, p. e603, 2022.

[139] S. Turner, L. L. Armstrong, Y. Bradford, C. S. Carlson, D. C. Crawford, A. T. Crenshaw, M. de Andrade, K. F. Doheny, J. L. Haines, and G. Hayes, "Quality control procedures for genome-wide association studies," *Current protocols in human genetics*, vol. 68, no. 1, pp. 1.19. 1–1.19. 18, 2011.

[140] J. Graffelman and B. Weir, "Testing for hardy–weinberg equilibrium at biallelic genetic markers on the x chromosome," *Heredity*, vol. 116, no. 6, pp. 558–568, 2016.

[141] M. E. Rentería, A. Cortes, and S. E. Medland, "Using plink for genome-wide association studies (gwas) and data analysis," *Genome-wide association studies and genomic prediction*, pp. 193–213, 2013.

[142] G. Montana, "Statistical methods in genetics," *Brief Bioinform*, vol. 7, no. 3, pp. 297–308, 2006.

[143] F. Bertolini, B. Servin, A. Talenti, E. Rochat, E. S. Kim, C. Oget, I. Palhière, A. Crisà, G. Catillo, R. Steri, M. Amills, L. Colli, G. Marras, M. Milanesi, E. Nicolazzi, B. D. Rosen, C. P. Van Tassell, B. Guldbrandtsen, T. S. Sonstegard, G. Tosser-Klopp, A. Stella, M. F. Rothschild, S. Joost, P. Crepaldi, and c. the AdaptMap, "Signatures of selection and environmental adaptation across the goat genome post-domestication," *Genetics Selection Evolution*, vol. 50, no. 1, p. 57, 2018.

[144] Z. Yan, A. Huang, D. Ma, C. Hong, S. Zhang, L. He, H. Rao, and S. Luo, "Atp6ap1 promotes cell proliferation and tamoxifen resistance in luminal breast cancer by inducing autophagy," *Cell Death & Disease*, vol. 16, no. 1, p. 201, 2025.

[145] Y. Kim, T.-Y. Ha, M.-S. Lee, and K.-A. Chang, "Regulatory mechanisms and therapeutic implications of lysosomal dysfunction in alzheimer's disease," *International Journal of Biological Sciences*, vol. 21, no. 3, p. 1014, 2025.

[146] M. Hatano, H. Fukushima, T. Ohto, Y. Ueno, S. Saeki, T. Enokizono, R. Tanaka, M. Tanaka, K. Imagawa, Y. Kanai, *et al.*, "Variants in kif2a cause broad clinical presentation; the computational structural analysis of a novel variant in a patient with a cortical dysplasia, complex, with other brain malformations 3," *American Journal of Medical Genetics Part A*, vol. 185, no. 4, pp. 1113–1119, 2021.

[147] D. Delano, M. Eberle, L. Galver, and C. Rosenow, "Array differences in genomic coverage and data quality impact gwas success," *Illumina*, 2010.

[148] N. Kleanthous, A. Hussain, W. Khan, J. Sneddon, and P. Liatsis, "Deep transfer learning in sheep activity recognition using accelerometer data," *Expert Systems with Applications*, vol. 207, p. 117925, 2022.

[149] B. Yang, F. Liu, C. Ren, Z. Ouyang, Z. Xie, X. Bo, and W. Shu, "Biren: predicting enhancers with a deep-learning-based model using the dna sequence alone," *Bioinformatics*, vol. 33, no. 13, pp. 1930–1936, 2017.

[150] Q. Liao, Y. Ding, Z. L. Jiang, X. Wang, C. Zhang, and Q. Zhang, "Multi-task deep convolutional neural network for cancer diagnosis," *Neurocomputing*, vol. 348, pp. 66–73, 2019.

[151] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[152] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[153] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning–based sequence model," *Nature methods*, vol. 12, no. 10, pp. 931–934, 2015.

[154] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[155]

[156] A. Gulli and S. Pal, *Deep learning with Keras*. Packt Publishing Ltd, 2017.

[157] C. Peng, Y. Zheng, and D. S. Huang, "Capsule network based modeling of multi-omics data for discovery of breast cancer-related genes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 5, pp. 1605–1612, 2020.

[158] Z. Shen, S.-P. Deng, and D.-S. Huang, "Capsule network for predicting rna-protein binding preferences using hybrid feature," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 5, pp. 1483–1492, 2019.

[159] M. R. Robinson, N. R. Wray, and P. M. Visscher, "Explaining additional genetic variation in complex traits," *Trends in Genetics*, vol. 30, no. 4, pp. 124–132, 2014.

[160] N. Raghavan and G. Tosto, "Genetics of alzheimer's disease: the importance of polygenic and epistatic components," *Current neurology and neuroscience reports*, vol. 17, no. 10, pp. 1–10, 2017.

[161] L. Zhu, S. Deng, Z. You, and D. Huang, "Identifying spurious interactions in the protein-protein interaction networks using local similarity preserving embedding," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 2, pp. 345–352, 2017.

[162] W. Lee, D.-S. Huang, and K. Han, "Constructing cancer patient-specific and group-specific gene networks with multi-omics data," *BMC Medical Genomics*, vol. 13, no. 6, p. 81, 2020.

[163] X. Liang, L. Zhu, and D. Huang, "Optimization of gene set annotations using robust trace-norm multitask learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 3, pp. 1016–1021, 2018.

[164] D. S. W. Ho, W. Schierding, M. Wake, R. Saffery, and J. O'Sullivan, "Machine learning snp based prediction for precision medicine," *Frontiers in genetics*, vol. 10, p. 267, 2019.

[165] A. B. Popejoy and S. M. Fullerton, "Genomics is failing on diversity," *Nature*, vol. 538, no. 7624, pp. 161–164, 2016.

[166] R. D. Dowell, "The similarity of gene expression between human and mouse tissues," *Genome Biology*, vol. 12, no. 1, p. 101, 2011.

[167] Y. Shi and D. M. Holtzman, "Interplay between innate immunity and alzheimer disease: Apoe and trem2 in the spotlight," *Nature Reviews Immunology*, vol. 18, no. 12, pp. 759–772, 2018.

[168] B. C. C. J. R. W. . M. W. L. . R. S. M. . S. D. P. . P. M. . 20 and T. . D. C. D. G. . D. D. B. . . N. S. . S. H. E. . W. N. M. . W. B. . . T. J. A. 2, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, no. 7145, pp. 661–678, 2007.

[169] B. A. Goldstein, A. E. Hubbard, A. Cutler, and L. F. Barcellos, "An application of random forests to a genome-wide association dataset: Methodological considerations new findings," *BMC Genetics*, vol. 11, no. 1, p. 49, 2010.

[170] M. H. Wang, H. J. Cordell, and K. Van Steen, "Statistical methods for genome-wide association studies," *Seminars in Cancer Biology*, vol. 55, pp. 53–60, 2019.

[171] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, and A. Chakravarti, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[172] C.-H. Yang, H.-S. Yang, and L.-Y. Chuang, "Pbmdr: A particle swarm optimization-based multifactor dimensionality reduction for the detection of multilocus interactions," *Journal of Theoretical Biology*, vol. 461, pp. 68–75, 2019.

[173] A. S. Alatrany, A. J. Hussain, J. Mustafina, and D. Al-Jumeily, "Machine learning approaches and applications in genome wide association study for alzheimer's disease: A systematic review," *IEEE Access*, vol. 10, pp. 62831–62847, 2022.

[174] A. Alatrany, A. Hussain, J. Mustafina, and D. Al-Jumeily, "A novel hybrid machine learning approach using deep learning for the prediction of alzheimer disease using genome data," in *International Conference on Intelligent Computing*, pp. 253–266, Springer.

[175] H. Xu, X. Li, Y. Yang, Y. Li, J. Pinheiro, K. Sasser, H. Hamadeh, X. Steven, M. Yuan, Initiative, and for the Alzheimer's Disease Neuroimaging, "High-throughput and efficient multilocus genome-wide association study on longitudinal outcomes," *Bioinformatics*, vol. 36, no. 10, pp. 3004–3010, 2020.

[176] L. Zou, Q. Huang, A. Li, and M. Wang, "A genome-wide association study of alzheimer's disease using random forests and enrichment analysis," *Science China Life Sciences*, vol. 55, pp. 618–625, 2012.

[177] H. Wang, T. Yue, J. Yang, W. Wu, and E. P. Xing, "Deep mixed model for marginal epistasis detection and population stratification correction in genome-wide association studies," *BMC Bioinformatics*, vol. 20(Suppl 23), pp. 1–11, 2019.

[178] J. X. Wang, Y. Li, X. Li, and Z. H. Lu, "Alzheimer's disease classification through imaging genetic data with ignet," *Front Neurosci*, vol. 16, p. 846638, 2022.

[179] C. M. Wilson, B. L. Fridley, J. R. Conejo-Garcia, X. Wang, and X. Yu, "Wide and deep learning for automatic cell type identification," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 1052–1062, 2021.

[180] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, and M. Ispir, "Wide deep learning for recommender systems," in *Proceedings of the 1st workshop on deep learning for recommender systems*, pp. 7–10.

[181] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, "Plink: A tool set for whole-genome association and population-based linkage analyses," *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[182] W. Zhang, B. Jiao, T. Xiao, C. Pan, X. Liu, L. Zhou, B. Tang, and L. Shen, "Mutational analysis of prnp in alzheimer's disease and frontotemporal dementia in china," *Scientific reports*, vol. 6, no. 1, pp. 1–7, 2016.

[183] A. D. Roses, M. W. Lutz, H. Amrine-Madsen, A. M. Saunders, D. Crenshaw, S. S. Sundseth, M. Huentelman, K. A. Welsh-Bohmer, and E. Reiman, "A tomm40 variable-length polymorphism predicts the age of late-onset alzheimer's disease," *The pharmacogenomics journal*, vol. 10, no. 5, pp. 375–384, 2010.

[184] D. Patel, J. Mez, B. N. Vardarajan, L. Staley, J. Chung, X. Zhang, J. J. Farrell, M. J. Rynkiewicz, L. A. Cannon-Albright, and C. C. Teerlink, "Association of rare coding mutations with alzheimer disease and other dementias among adults of european ancestry," *JAMA network open*, vol. 2, no. 3, pp. e191350–e191350, 2019.

[185] W. Zhu, S. Xu, C. C. Liu, and Y. Li, "Minimax powerful functional analysis of covariance tests with application to longitudinal genome-wide association studies," *Scandinavian Journal of Statistics*, 2022.

[186] P. Fergus, C. C. Montanez, B. Abdulaimma, P. Lisboa, C. Chalmers, and B. Pineles, "Utilizing deep learning and genome wide association studies for epistatic-driven preterm birth classification in african-american women," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 2, pp. 668–678, 2020.

[187] M. Squillario, G. Abate, F. Tomasi, V. Tozzo, A. Barla, D. Uberti, M. W. Weiner, P. Aisen, R. Petersen, J. R. Clifford, W. Jagust, J. Q. Trojanowki, A. W. Toga, L. Beckett, R. C. Green, A. J. Saykin, J. Morris, L. M. Shaw, Z. Khachaturian, G. Sorensen, M. Carrillo, L. Kuller, M. Raichle, S. Paul, P. Davies, H. Fillit, F. Hefti, D. Holtzman, M. M. Mesulam, W. Potter, P. Snyder, T. Montine, R. G. Thomas, M. Donohue, S. Walter, T. Sather, G. Jiminez, A. B. Balasubramanian, J. Mason, I. Sim, D. Harvey, M. Bernstein, N. Fox, P. Thompson, N. Schuff, C. DeCarli, B. Borowski, J. Gunter, M. Senjem, P. Vemuri, D. Jones, K. Kantarci, C. Ward, R. A. Koeppe, N. Foster, E. M. Reiman, K. Chen, C. Mathis, S. Landau, N. J. Cairns, E. Householder, L. Taylor-Reinwald, V. Lee, M. Korecka, M. Figurski, K. Crawford, S. Neu, T. M. Foroud, S. Potkin, L. Shen, K. Faber, S. Kim, L. Tha, R. Frank, J. Hsiao, J. Kaye, J. Quinn, L. Silbert, B. Lind, R. Carter, S. Dolen, B. Ances, M. Carroll, M. L. Creech, E. Franklin, M. A. Mintun, S. Schneider, A. Oliver, L. S. Schneider, S. Pawluczyk, M. Beccera, L. Teodoro, B. M. Spann, J. Brewer, H. Vanderswag, A. Fleisher, D. Marson, R. Griffith, D. Clark, D. Geldmacher, *et al.*, "A telescope gwas analysis strategy, based on snps-genes-pathways ensamble and on multivariate algorithms, to characterize late onset alzheimer's disease," *Scientific Reports*, vol. 10, no. 1, p. 12063, 2020.

[188] O. Erdoğan, M. Esme, C. Balci, S. Rafatov, M. Cankurtaran, B. B. Yavuz, C. İyigün, and Y. A. Son, "Identification of genomic biomarkers with machine learning for early and differential diagnosis of late-onset alzheimer's disease (load) genetics/omics and systems biology," *Alzheimer's Dementia*, vol. 16, p. e042558, 2020.

[189] Z. Bao, X. Zhao, J. Li, G. Zhang, H. Wu, Y. Ning, M. D. Li, and Z. Yang, "Prediction of repeated-dose intravenous ketamine response in major depressive disorder using the gwas-based machine learning approach," *Journal of Psychiatric Research*, vol. 138, pp. 284–290, 2021.

[190] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, and J. Kim, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature biomedical engineering*, vol. 2, no. 10, pp. 749–760, 2018.

[191] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[192] B. Bogdanovic, T. Eftimov, and M. Simjanoska, "In-depth insights into alzheimer's disease by using explainable machine learning approach," *Scientific Reports*, vol. 12, no. 1, p. 6508, 2022.

[193] A. Lombardi, D. Diacono, N. Amoroso, P. Biecek, A. Monaco, L. Bellantuono, E. Pantaleo, G. Logroscino, R. De Blasi, S. Tangaro, and R. Bellotti, "A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of mild cognitive impairment and alzheimer's disease," *Brain Informatics*, vol. 9, no. 1, p. 17, 2022.

[194] S. O. Danso, Z. Zeng, G. Muniz-Terrera, and C. W. Ritchie, "Developing an explainable machine learning-based personalised dementia risk prediction model: A transfer learning approach with ensemble learning algorithms," *Frontiers in big Data*, vol. 4, p. 21, 2021.

[195] N. An, H. Ding, J. Yang, R. Au, and T. F. A. Ang, "Deep ensemble learning for alzheimer's disease classification," *Journal of Biomedical Informatics*, vol. 105, p. 103411, 2020.

[196] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," *Big Data Analytics*, vol. 1, no. 1, p. 9, 2016.

[197] J. Luengo, S. García, and F. Herrera, "On the choice of the best imputation methods for missing values considering three groups of classification methods," *Knowledge and information systems*, vol. 32, no. 1, pp. 77–108, 2012.

[198] R. K. Prematunga, "Correlational analysis," *Australian Critical Care*, vol. 25, no. 3, pp. 195–199, 2012.

[199] H. Wang, M. J. Bah, and M. Hammad, "Progress in outlier detection techniques: A survey," *IEEE Access*, vol. 7, pp. 107964–108000, 2019.

[200] T. K. Khan, *Chapter 2 - Clinical Diagnosis of Alzheimer's Disease*, pp. 27–48. Academic Press, 2016.

[201] D. A. González, M. M. Gonzales, Z. J. Resch, A. C. Sullivan, and J. R. Soble, "Comprehensive evaluation of the functional activities questionnaire (faq) and its reliability and validity," *Assessment*, vol. 29, no. 4, pp. 748–763, 2022.

[202] J. L. Cummings, "The neuropsychiatric inventory: assessing psychopathology in dementia patients," *Neurology*, vol. 48, no. 5 Suppl 6, pp. 10S–16S, 1997.

[203] Y. Feghali, Y. Fares, and L. Abou Abbas, "Assessment of neuropsychiatric symptoms in dementia: validity and reliability of the lebanese version of the neuropsychiatric inventory questionnaire," *Applied Neuropsychology: Adult*, vol. 28, no. 5, pp. 588–595, 2021.

[204] G. Musa, F. Henríquez, C. Muñoz-Neira, C. Delgado, P. Lillo, and A. Slachevsky, "Utility of the neuropsychiatric inventory questionnaire (npi-q) in the assessment of a sample of patients with alzheimer's disease in chile," *Dement Neuropsychol*, vol. 11, no. 2, pp. 129–136, 2017.

[205] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes," *Pattern Recognition*, vol. 44, no. 8, pp. 1761–1776, 2011.

[206] A. Sarica, A. Cerasa, and A. Quattrone, "Random forest algorithm for the classification of neuroimaging data in alzheimer's disease: a systematic review," *Frontiers in aging neuroscience*, vol. 9, p. 329, 2017.

[207] J. Wang, C. Rao, M. Goh, and X. Xiao, "Risk assessment of coronary heart disease based on cloud-random forest," *Artificial Intelligence Review*, vol. 56, no. 1, pp. 203–232, 2023.

[208] N. Xin, X.-F. Gu, H. Wu, Y.-Z. Hu, and Z.-L. Yang, "Discrimination of raw and processed dipsacus asperoides by near infrared spectroscopy combined with least squares-support vector machine and random forests," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 89, pp. 18–24, 2012.

[209] M. Bucholc, S. Titarenko, X. Ding, C. Canavan, and T. Chen, "A hybrid machine learning approach for prediction of conversion from mild cognitive impairment to dementia," *Expert Systems with Applications*, vol. 217, p. 119541, 2023.

[210] M. Lin, P. Gong, T. Yang, J. Ye, R. L. Albin, and H. H. Dodge, "Big data analytical approaches to the nacc dataset: aiding preclinical trial enrichment," *Alzheimer disease and associated disorders*, vol. 32, no. 1, p. 18, 2018.

[211] H.-C. Huang, Y.-M. Tseng, Y.-C. Chen, P.-Y. Chen, and H.-Y. Chiu, "Diagnostic accuracy of the clinical dementia rating scale for detecting mild cognitive impairment and dementia: A bivariate meta-analysis," *International Journal of Geriatric Psychiatry*, vol. 36, no. 2, pp. 239–251, 2021.

[212] M. L. F. Chaves, A. L. Camozzato, C. Godinho, R. Kochhann, A. Schuh, V. L. De Almeida, and J. Kaye, "Validity of the clinical dementia rating scale for the detection and staging of dementia in brazilian patients," *Alzheimer Disease & Associated Disorders*, vol. 21, no. 3, pp. 210–217, 2007.

[213] Y. L. Chang, M. W. Bondi, L. K. McEvoy, C. Fennema-Notestine, D. P. Salmon, D. Galasko, J. Hagler, D. J., and A. M. Dale, "Global clinical dementia rating of 0.5 in mci masks variability related to level of function," *Neurology*, vol. 76, no. 7, pp. 652–9, 2011.

[214] D. C. Rubinsztein and D. F. Easton, "Apolipoprotein e genetic variation and alzheimer's disease: a meta-analysis," *Dementia and geriatric cognitive disorders*, vol. 10, no. 3, pp. 199–209, 1999.

[215] L. Bertram, M. B. McQueen, K. Mullin, D. Blacker, and R. E. Tanzi, "Systematic meta-analyses of alzheimer disease genetic association studies: the alzgene database," *Nature genetics*, vol. 39, no. 1, pp. 17–23, 2007.

[216] K. E. Rodriguez, H. Herzog, and N. R. Gee, "Variability in human-animal interaction research," *Frontiers in Veterinary Science*, vol. 7, p. 619600, 2021.

[217] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[218] T. Jo, K. Nho, and A. J. Saykin, "Deep learning in alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data," *Frontiers in aging neuroscience*, vol. 11, p. 220, 2019.

[219] E. Lin, C.-H. Lin, and H.-Y. Lane, "Deep learning with neuroimaging and genomics in alzheimer's disease," *International journal of molecular sciences*, vol. 22, no. 15, p. 7911, 2021.

[220] N. Alexander, D. C. Alexander, F. Barkhof, and S. Denaxas, "Identifying and evaluating clinical subtypes of alzheimer's disease in care electronic health records using unsupervised machine learning," *BMC Medical Informatics and Decision Making*, vol. 21, pp. 1–13, 2021.

[221] "2022 alzheimer's disease facts and figures," *Alzheimer's Dementia*, vol. 18, no. 4, pp. 700–789, 2022.

[222] Z. Breijyeh and R. Karaman, "Comprehensive review on alzheimer's disease: causes and treatment," *Molecules*, vol. 25, no. 24, p. 5789, 2020.

[223] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks," *Journal of Big Data*, vol. 7, no. 1, pp. 1–41, 2020.

[224] Y. Huang, X. Sun, H. Jiang, S. Yu, C. Robins, M. J. Armstrong, R. Li, Z. Mei, X. Shi, and E. S. Gerasimov, "A machine learning approach to brain epigenetic analysis reveals kinases associated with alzheimer's disease," *Nature communications*, vol. 12, no. 1, pp. 1–12, 2021.

[225] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

[226] N. Kleanthous, A. Hussain, W. Khan, J. Sneddon, and A. Mason, "Feature extraction and random forest to identify sheep behavior from accelerometer data," in *International Conference on Intelligent Computing*, pp. 408–419, Springer.

[227] E. Lin, C.-H. Lin, and H.-Y. Lane, "Deep learning with neuroimaging and genomics in alzheimer's disease," *International journal of molecular sciences*, vol. 22, no. 15, p. 7911, 2021.

[228] A. A. Moustafa, *Alzheimer's Disease: Understanding Biomarkers, Big Data, and Therapy*. Academic Press, 2021.

[229] A. Ng, "Sparse autoencoder," *CS294A Lecture notes*, vol. 72, no. 2011, pp. 1–19, 2011.

[230] B. Tang, Z. Pan, K. Yin, and A. Khateeb, "Recent advances of deep learning in bioinformatics and computational biology," *Frontiers in genetics*, vol. 10, p. 214, 2019.

[231] P. Yang, Y. Hwa Yang, B. B Zhou, and A. Y Zomaya, "A review of ensemble methods in bioinformatics," *Current Bioinformatics*, vol. 5, no. 4, pp. 296–308, 2010.

[232] Z. Chen, M. Boehnke, X. Wen, and B. Mukherjee, "Revisiting the genome-wide significance threshold for common variant gwas," *G3 Genes—Genomes—Genetics*, vol. 11, no. 2, 2021.

[233] D. Klarin, J. Lynch, K. Aragam, M. Chaffin, T. L. Assimes, J. Huang, K. M. Lee, Q. Shao, J. E. Huffman, and P. Natarajan, "Genome-wide association study of peripheral artery disease in the million veteran program," *Nature medicine*, vol. 25, no. 8, pp. 1274–1279, 2019.

[234] G.-W. Lin, C. Xu, K. Chen, H.-Q. Huang, J. Chen, B. Song, J. K. Chan, W. Li, W. Liu, and L.-Y. Shih, "Genetic risk of extranodal natural killer t-cell lymphoma: a genome-wide association study in multiple populations," *The Lancet Oncology*, vol. 21, no. 2, pp. 306–316, 2020.

[235] E. Uffelmann, Q. Q. Huang, N. S. Munung, J. de Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, and D. Posthuma, "Genome-wide association studies," *Nature Reviews Methods Primers*, vol. 1, no. 1, p. 59, 2021.

[236] S. Weintraub, D. Salmon, N. Mercaldo, S. Ferris, N. R. Graff-Radford, H. Chui, J. Cummings, C. DeCarli, N. L. Foster, D. Galasko, *et al.*, "The alzheimer's disease centers' uniform data set (uds): The neuropsychological test battery," *Alzheimer disease and associated disorders*, vol. 23, no. 2, p. 91, 2009.

[237] D. L. Beekly, E. M. Ramos, G. van Belle, W. Deitrich, A. D. Clark, M. E. Jacka, W. A. Kukull, *et al.*, "The national alzheimer's coordinating center (nacc) database: an

alzheimer disease database," *Alzheimer Disease & Associated Disorders*, vol. 18, no. 4, pp. 270–277, 2004.

[238] V. Margot and G. Luta, "A new method to compare the interpretability of rule-based algorithms," *AI*, vol. 2, no. 4, pp. 621–635, 2021.

[239] J. Wu, C. Liu, L. Xie, X. Li, K. Xiao, G. Xie, and F. Xie, "Early prediction of moderate-to-severe condition of inhalation-induced acute respiratory distress syndrome via interpretable machine learning," *BMC Pulmonary Medicine*, vol. 22, no. 1, p. 193, 2022.

[240] R. Huijzer, F. Blaauw, and R. J. den Hartigh, "Sirus. jl: Interpretable machine learning via rule extraction," *Journal of Open Source Software*, vol. 8, no. 90, p. 5786, 2023.

[241] E. S. Bradley, A. L. Zeamer, V. Bucci, L. Cincotta, M.-C. Salive, P. Dutta, S. Mutaawe, O. Anya, C. Tocci, A. Moormann, *et al.*, "Oropharyngeal microbiome profiled at admission is predictive of the need for respiratory support among covid-19 patients," *Frontiers in Microbiology*, vol. 13, p. 1009440, 2022.

[242] G. Chen, H. Liu, L. Yu, Q. Wei, and X. Zhang, "A new approach to classification based on association rule mining," *Decision Support Systems*, vol. 42, no. 2, pp. 674–689, 2006.

[243] K. Song and K. Lee, "Predictability-based collective class association rule mining," *Expert Systems with Applications*, vol. 79, pp. 1–7, 2017.

[244] A. Telikani, A. H. Gandomi, and A. Shahbahrami, "A survey of evolutionary computation for association rule mining," *Information Sciences*, vol. 524, pp. 318–352, 2020.