University of Reading

Doctor of Philosophy

Enhancing the Performance and Transparency of Machine Learning (ML) using MRI-derived data: Alternative Approaches to ML Interpretability-Explainability

Author: ***Atmakuru Akhila***

Supervisors: ***Prof. Atta Badii, Prof. Giuseppe Di Fatta, Dr. Ferran Espuny-Pujol***

Department of Computer Science

November 2025

Declaration

I confirm that this is my own work, and the use of all material from other sources has been properly and fully acknowledged.

Akhila Atmakuru

Scientific Publications

Atmakuru, A., Di Fatta, G., Nicosia, G., and Badii, A. (2023, September). Improved filter-based feature selection using correlation and clustering techniques. In *International Conference on Machine Learning, Optimization, and Data Science* (pp. 379-389). Cham: Springer Nature Switzerland.

DOI: https://doi.org/10.1007/978-3-031-53969-5_28

Atmakuru, A., Di Fatta, G., Nicosia, G., Varzandian, A., and Badii, A. (2023, September). Sensitivity analysis for feature importance in predicting Alzheimer's Disease. In *International Conference on Machine Learning, Optimization, and Data Science* (pp. 449-465). Cham: Springer Nature Switzerland.

DOI: https://doi.org/10.1007/978-3-031-53966-4_33

Atmakuru, A., Badii, A., and Di Fatta, G. (2024, September). Transfer Learning for the Cognitive Staging Prediction in Alzheimer's Disease. In International Conference on Machine Learning, Optimization, and Data Science (pp. 176-190). Cham: Springer Nature Switzerland.

DOI: https://doi.org/10.1007/978-3-031-82487-6_13

Abstract

This thesis addresses the three fundamental challenges for enhancing the performance of Machine Learning (ML) models. Despite their evolving predictive capabilities, MLs still present significant limitations in generalisability, particularly in high-dimensional settings, interpretability, and high data requirements. These issues require methodologies that reduce input data dimensionality, enhance transparency, and utilise prior knowledge to moderate the scale of data requirements, thereby improving the performance, reliability, and efficiency of machine learning solutions in practical applications.

Accordingly, this thesis introduces three independent methods responsive to the above main limitations that need to be overcome to enhance the performance and transparency of models in complex task domains. First, two filter-based feature selection techniques—one correlation-driven and the other clustering-based—are developed to reduce redundancy and enhance generalisability in high-dimensional data. The correlation-based technique outperforms the state-of-the-art (as represented by ReliefF) in both internal and external validations. Second, an ensemble explainability framework integrates Shapley Additive Explanations (SHAP) values with Sobol indices, combining their rankings to yield stable and interpretable attributions. Third, a multi-stage algorithm couples transfer learning with an autoencoder to minimise labelled data requirements without adversely affecting performance.

All proposed methods yielded quantifiable improvements. The feature selection techniques reduced input dimensionality while enhancing accuracy and generalisability compared to ReliefF. The ensemble explainability framework produced consistent attributions under varying data distributions and reliably identified informative input features. The multi-stage algorithm achieved enhanced classification performance with reduced reliance on labelled data.

Case-Study: The proposed methods were validated in the context of medical diagnosis for early-stage prediction of dementia, utilising a structural Alzheimer's MRI dataset. In this application, optimising the feature selection, as described above, enhanced the cross-cohort accuracy and decreased the data dimensionality. The explainability framework consistently identified clinically relevant regions, such as hippocampal subfields (W. Zhao et al., 2019) and

the temporal horn (Vernooij and van Buchem, 2020), supporting the credibility of feature relevance. The data-efficient multi-stage pipeline achieved an accuracy of 73.26%, exceeding prior baselines (Li et al., 2015; Oh et al., 2019).

This thesis concludes that the proposed correlation and clustering-based feature selection, ensemble explainability combining SHAP and Sobol, and transfer learning with autoencoders have led to enhanced accuracy, robustness, and transparency of the performance of the machine learning models. Although this was validated for the Alzheimer's validation task, these methods are domain-agnostic and provide scalable, reliable, and resource-efficient approaches for high-dimensional, data-limited real-world applications.

List of Abbreviations

AD	Alzheimer’s Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
AI	Artificial Intelligence
AIBL	Australian Imaging Biomarkers and Lifestyle Study of Ageing
ALS	Amyotrophic Lateral Sclerosis
aMCI	Amnestic Mild Cognitive Impairment
ANN	Artificial Neural Network
AutoML	Automated Machine Learning
CAD	Computer-Aided Diagnosis
CN	Cognitive Normal
CNN	Convolutional Neural Networks
CNR	Contrast-to-Noise Ratio
CSF	Cerebro-Spinal Fluid
CV	Computer Vision
DL	Deep Learning
DNN	Deep Neural Networks
ECG	Electrocardiogram
EMCI	Early Mild Cognitive Impairment
FAST	Fourier Amplitude Sensitivity Test
fMRI	Functional Magnetic Resonance Imaging
FS	Feature Selection
FTD	Frontotemporal Dementia
GANs	Generative Adversarial Networks
HATA	Hippocampal-amygdala transition area

HC	Healthy Controls
ICE	Individual Conditional Expectations
IXI	Information extraction from Images lifestyle flagship study of ageing
LMCI	Late Mild Cognitive Impairment
LSTM	Long Short-Term Memory
MCI	Mild Cognitive Impairment
ML	Machine Learning
MMSE	Mini–Mental State Examination
MRI	Magnetic Resonance Imaging
MS	Multiple Sclerosis
NDD	Neuro-Degenerative Diseases
NIFD	Neuroimaging in Frontotemporal Dementia
NLP	Natural Language Processing
NN	Neural Networks
PCA	Principal Component Analysis
PCoA	Principal Co-ordinate Analysis
PDP	Partial Dependence Plot
PET	Positron Emission Tomography
PPMI	Parkinson's Progression Markers Initiative
ResNets	Residual Networks
RFE	Recursive Feature Elimination
RL	Reinforcement Learning
RNN	Recurrent Neural Networks
ROI	Region Of Interest
SA	Sensitivity Analysis

SGD	Stochastic Gradient Descent
SHAP	Shapley Additive exPlanations
sMRI	structural Magnetic Resonance Imaging
SVM	Support Vector Machine
TL	Transfer Learning
XAI	eXplainable Artificial Intelligence

Table of Contents

List of Figures

List of Tables

List of Algorithms

1. INTRODUCTION

Artificial Intelligence (AI) has impacted daily routines, such as work, education, communication and socialising, by imitating human intelligence to perform tasks and make decisions. Sophisticated AI models utilising Deep Learning (DL) can analyse large quantities of varied data, inherently grasp complex nonlinear connections between dependent and independent variables and provide accurate decisions. Therefore, AI models can address numerous real-world issues. Their deployment spans many applications, such as smartphones, autonomous vehicles, and vital services such as banking, healthcare, law enforcement, and the military. AI models excel in speech, image recognition, translation, natural language processing, computer vision, and autonomous driving.

AI is essential in modern industry, as advanced AI models can analyse vast amounts of complex, high-dimensional data. Developing AI-based classifiers facilitates accurate pattern recognition, anomaly detection, and performance forecasting across various domains. AI models assist experts in decision-making, system optimisation, and delivering tailored solutions, demonstrating predictive accuracy comparable to traditional expert-driven methods. However, despite these advancements, the adoption of AI in critical real-world applications remains challenging.

Although AI models perform excellently, industry practitioners often hesitate to deploy these models in operational pipelines. The primary concern is the lack of transparent explanations for model behaviour, primarily due to the black-box nature of DL models. In high-stakes environments, decisions must be explainable, reliable, and trustworthy. Additionally, regulatory frameworks require organisations to provide accountability for decisions made by their algorithms. These challenges have sparked research efforts to advance eXplainable AI (XAI) techniques through Sensitivity Analysis (SA) (Arrieta et al., 2020).

Developing precise and interpretable AI models, particularly for complex classification and decision-making tasks, requires addressing challenges related to data dimensionality, quality, and availability. Feature Selection (FS) is key to improving AI model performance by identifying the most relevant features while minimising irrelevant or noisy data. Many real-world datasets, such as those collected from sensors, logs, or transactional records, are high-dimensional, making FS essential for improving model accuracy,

interpretability, and computational efficiency. Techniques such as Pearson correlation, Recursive Feature Elimination (RFE), and LASSO regression help refine predictive models by selecting the most discriminative features and enhancing pattern recognition.

Despite advancements in AI, a significant challenge in real-world applications is the scarcity of large, labelled datasets, often due to privacy constraints, data sensitivity, or the high cost of expert annotations. Transfer Learning (TL) helps to overcome this issue by utilising pre-trained models on large, publicly available datasets and adapting them to smaller, domain-specific datasets. By fine-tuning DL models on large multi-source datasets, TL enhances predictive accuracy, accelerates training, and reduces overfitting. In monitoring and predictive modelling, TL enables AI systems to efficiently analyse evolving patterns, forecast critical changes, and integrate diverse information sources.

This chapter analyses the relative merits and demerits of various AI methodologies. It also aims to explain key concepts, challenges, and model behaviours in complex, high-dimensional environments, including their broader impact on model reliability and performance. The chapter discusses the integration of AI with real-world applications, with particular emphasis on its use in high-stakes decision-making tasks. This review examines Feature Selection (FS) methods, Sensitivity Analysis (SA) techniques, and Transfer Learning (TL) approaches, providing a foundation for this research. It then sets the stage for an in-depth investigation of these methodologies in the subsequent chapters by explaining how they enhance model accuracy, enhance explainability, and utilise existing knowledge for efficient learning. Additionally, overcoming these challenges is crucial to enhancing the understanding of complex systems, which can lead to the development of effective, trustworthy, and scalable AI applications.

1.1. Artificial Intelligence

Artificial Intelligence (AI) is a branch of computer science that focuses on designing algorithms enabling machines to perform tasks requiring intelligent behaviour. It involves computational models for perception, reasoning, learning, and decision-making. Building on the widespread influence of AI in various sectors, this section explores the fundamental principles, learning methods, and functionalities that define AI systems. The rapid rise in processing capabilities and data accessibility has driven AI to the forefront of its advancement (Duan et al., 2019).

This field of research has undergone a significant transformation, with researchers rigorously evaluating its advantages and current challenges. AI has enhanced several sectors, such as banking, manufacturing, and healthcare, by effectively managing various complex tasks.

Machine Learning (ML) is a subfield of AI that focuses on developing algorithms that enable systems to learn patterns from data. It involves statistical modelling, optimisation, and generalisation to improve performance on tasks without explicit programming. ML has three main learning approaches: supervised, semi-supervised, and unsupervised learning, as shown in Figure 1- 1.

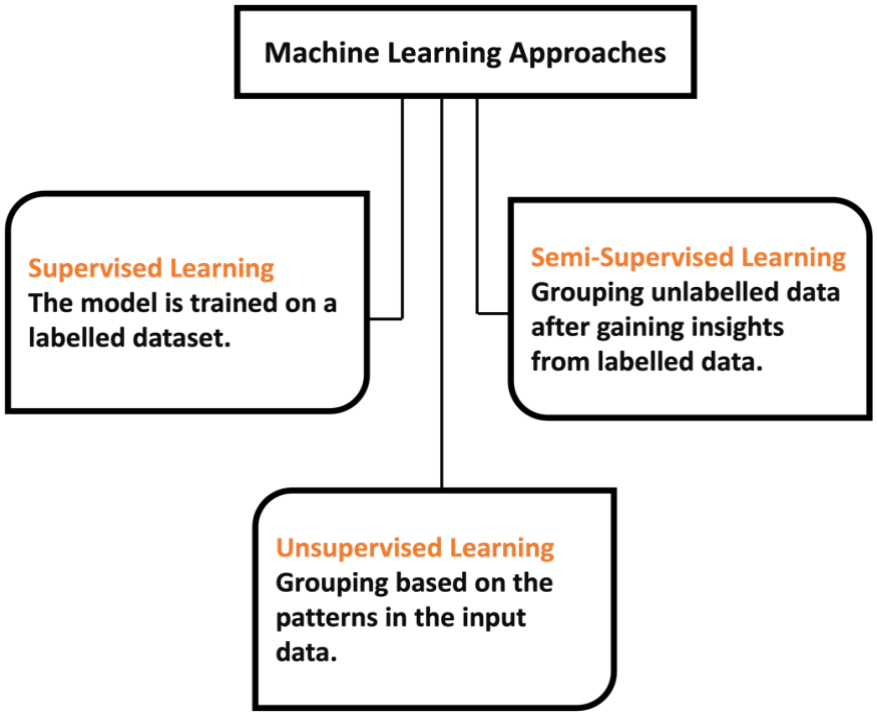

Figure 1- 1 Machine Learning Approaches

Supervised learning is crucial in classification and prediction tasks, as models train on labelled data with known input-output pairs. This approach is effective in scenarios involving historical data, providing precise predictions of results (Cunningham et al., 2008). Semi-supervised learning utilises both labelled and unlabelled data to optimise model performance, particularly in scenarios where acquiring labelled data is resource-intensive. This approach enhances the decision boundary by incorporating the structure of the unlabelled data distribution, improving generalisation and reducing reliance on limited labelled samples (Y. C.A.P.Reddy et al., 2018) .

On the other hand, unsupervised learning models independently discover hidden patterns, correlations, and underlying influences using data without predetermined labels (Barlow, 1989). These approaches provide flexible resources for retrieving valuable data insights, each designed for various industries such as finance, healthcare, and marketing. AI learning methods provide powerful tools for data-driven decision-making, but their real-world impact depends on their strengths and limitations.

1.1.1. Recent Advances and Outstanding Challenges in Machine Learning

AI provides numerous advantages, leading to significant progress in transforming different sectors. An important benefit is its capacity to simplify everyday tasks, improving productivity and enabling the workforce to focus on complex and innovative endeavours. Additionally, different industries, such as manufacturing, healthcare, finance, chatbots, virtual assistants, e-commerce, scientific research, drug development, and climate research, use AI models.

AI-powered robots in manufacturing are capable of carrying out assembly line duties with accuracy and speed, resulting in decreased production expenses and enhanced output quality (Grau et al., 2021). AI analyses large data sets, recognises patterns and insights, and swiftly makes accurate decisions. This ability is crucial in domains such as healthcare, where AI can significantly impact the identification of illnesses, forecasting of patient outcomes, and personalisation of treatment strategies (Panesar, 2021). AI, when applied in the financial sector, can examine market patterns and identify fraudulent activities instantly, leading to increased profitability and security (Hafez et al., 2025). Adam et al. (2021) state that AI-powered chatbots and virtual assistants enhance customer satisfaction by offering immediate assistance and tailored interactions.

Within scientific research, AI accelerates discovery by effectively analysing large datasets, recognising complex patterns, and creating predictive models. These innovations in climate science have resulted in significant advancements in accurately predicting environmental changes (Huntingford et al., 2019). Integrating AI into education enables adaptive systems that model individual cognitive processes and personalise content. Techniques such as reinforcement learning and intelligent tutoring systems support diverse learning styles and promote autonomy. The systems offer real-time feedback and detailed analysis of student engagement and progress (Zhai et al., 2021).

Although AI offers numerous benefits, it also has several significant disadvantages and limitations. Among the many challenges of AI, one of the most pressing issues in critical applications such as finance and healthcare is the 'black box' nature of DL models, which raises concerns about transparency and trustworthiness (Rudin, 2019). Moreover, AI systems often require substantial amounts of data to operate efficiently. Collecting, storing, and analysing this data can require significant resources and potentially raise privacy concerns (Philip Chen and Zhang, 2014). Violating or breaching regulations concerning this information may result in substantial ethical and legal complications, weakening confidence in AI technologies (Stahl, 2021).

A significant drawback of AI is its reliance on the quality of the data it is trained on. When trained on biased or incomplete data, AI can reinforce biases and produce inequitable results. This can significantly affect individuals and communities in sensitive domains such as hiring, law enforcement, and lending (Martin, 2019). It is crucial to acknowledge and address the downsides and limitations of AI, such as privacy issues, biased algorithms, and inadequate transparency. This strategy is vital for ensuring the responsible and beneficial use of AI technology.

1.1.2. Blackbox Behaviour

Despite the significant potential of AI systems, there remains a reluctance to adopt DL and Deep Neural Network (DNN) models in critical sectors such as medical diagnostics, defence, automobile automation, financial prediction, and the justice system. Deep Learning, a subset of machine learning, refers to the use of multilayered artificial neural networks to learn hierarchical feature representations from large-scale data. The black-box nature of DL models poses a significant challenge to their implementation. DNNs often operate using complex, hidden internal mechanisms that are challenging for humans to comprehend, raising concerns about the transparency and reliability of their decision-making. In the healthcare industry, unexplainable systems can cause distrust among clinicians making life-changing decisions, leading to hesitation in trusting or understanding the reasoning of the model (Rosenbacke et al., 2024).

In this context, the concepts of explainability and interpretability become crucial (Nassar et al., 2020). Interpretability focuses on providing clear reasons why an AI model

makes specific predictions or decisions, aiming to make the underlying mechanisms and logic of the model understandable to users. Understanding model decisions is critical in complex models, such as DNNs, where outcomes may not be visible or intuitive. On the other hand, explainability refers to how easily a human can comprehend the relationship between the input and output of the model. Classic models, such as decision trees and linear regression, are inherently explainable because their decision-making processes are straightforward. Nevertheless, as model complexity increases, preserving explainability becomes challenging. Despite their high accuracy, DNNs have limited explainability, hindering their application in real-world settings.

1.1.3. Explainability and Accuracy Trade-Off

Enhancing explainability highlights an inherent challenge in AI development—the trade-off between model accuracy and explainability. When developing and implementing Machine Learning (ML) models, a common performance trade-off arises between precision and explainability (Wanner et al., 2021). Linear regression and decision trees are easy to understand, explain, and validate, even for those with limited AI knowledge (Izza et al., 2020). This simplicity fosters greater trust in these models because their decision-making processes are transparent and understandable. Users can easily trace how these models arrived at a particular decision or prediction, making them particularly appealing when trust and accountability are crucial.

Nevertheless, as research goals become complex, the limitations of these models become evident, frequently requiring the utilisation of advanced DL models. Nonlinear models can manage higher complexity and generate precise outcomes but typically trade off transparency in explanations. One optimal illustration of this compromise is evident in Convolutional Neural Networks (CNNs) (Jung et al., 2021). Although CNNs have shown outstanding results in domains such as image recognition, understanding their internal operations proves challenging, even for experts. The lack of transparency due to this opacity hinders the ability to justify the reasoning behind a particular prediction outcome by the model.

Hence, there is a conflict between models that are understandable yet less precise and those that are accurate yet less interpretable. Basic models are easier to understand, yet

they may not possess the sophistication, i.e., the learning capability needed to achieve maximum predictive precision. Conversely, while yielding enhanced results, complex models often obscure the decision-making process, eroding trust among users, particularly in high-stakes domains. In law, finance, and healthcare, predictive accuracy is of paramount importance for validation. In these domains, the success of the model depends primarily on its ability to produce precise and reliable results. However, despite the emphasis on accuracy, the need for explainability remains critical, particularly from the end user's perspective.

The increased significance of explainability requires AI systems to balance precise predictions and understandable explanations. Despite their exceptional predictive accuracy, complex models such as CNNs and other DNN techniques must provide transparency. This approach has led to the development of XAI methods using SA techniques that focus on making complex models explainable so that experts can trust the decisions of AI models.

In recent years, several diverse domains have embraced the explainability component of AI, prioritising trustworthiness and transparency over pure accuracy. The right balance between accuracy and explainability is crucial for the successful adoption of AI in critical domains. Researchers and domain experts can ensure that DL models provide precise predictions and generate explanations that foster trust and enable informed decision-making by incorporating techniques that enhance explainability, such as XAI frameworks.

1.1.4. Explainability in Artificial Intelligence and Frameworks

To address these obstacles, it is crucial to develop DL models that are both precise and offer clear explanations. This helps experts to understand results, make informed choices, and trust AI-based systems. Realising the full potential of DL in critical domains requires dedicated efforts to develop easily understood and explainable AI (D. Kaur et al., 2023).

Researchers have developed XAI frameworks to address the challenges posed by black-box AI models, thereby enhancing transparency and trust. These frameworks provide insights into how AI generates predictions, making them interpretable for practitioners and researchers. Sensitivity Analysis (SA) is a key technique within XAI that investigates how variations in input features influence model outputs, thereby revealing the internal decision-making process of complex models. By quantifying feature relevance, SA enhances model interpretability and facilitates the identification of inputs that most significantly drive

predictions. SA directly advances the objectives of XAI by enhancing the transparency, interpretability, and trustworthiness of opaque models—such as deep neural networks. XAI frameworks are crucial for generating explanations and enhancing the transparency of DNNs, thereby fostering user confidence in AI-driven decisions.

SHapley Additive exPlanations (SHAP) is a critical framework that utilises game theory, mainly focusing on Shapley values (Lundberg and Lee, 2017). SHAP assigns a weight to each feature based on its contribution to the prediction made by the model by considering all possible feature combinations. This method ensures a fair distribution of feature importance, providing both local and global insights into the model. SHAP is applicable across various ML and DL models, offering accurate and consistent explanations.

Figure 1- 2**Error! Reference source not found.** presents various XAI frameworks. The techniques discussed in this section represent only a small subset of the complete range of available methods.

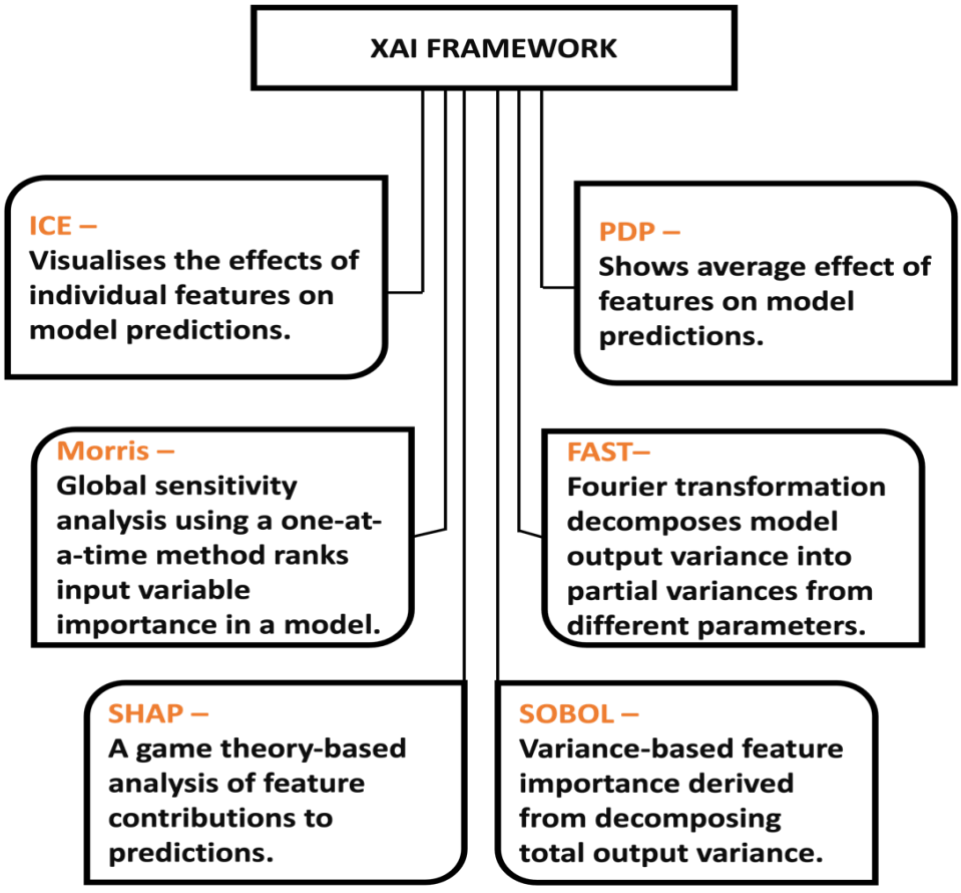

Figure 1- 2 Current eXplainability in AI (XAI) Frameworks

Methods such as Sobol, Morris, and Fourier Amplitude Sensitivity Testing (FAST) are vital for comprehending AI model predictions by assessing feature importance. The Sobol method, a variance-based approach, breaks down the model output variance into contributions from input variables and their interactions, thereby providing a comprehensive global view of feature importance (Sobol, 2001). In contrast, the Morris method is a practical One-At-a-Time (OAT) technique that approximates the elementary effects of input variables through small perturbations and measurement of the resulting output changes. This makes it effective for identifying key features without heavy computational demand (Morris, 1991). The FAST method, which operates in the frequency domain, derives sensitivity indices by transforming input variables and analysing variance across different frequencies, serving as a computationally efficient alternative to the Sobol approach (Saltelli et al., 1999). These XAI frameworks that use SA techniques provide valuable insights into model behaviour by providing global interpretability and fostering trust in AI-driven solutions and decision-making.

For models requiring visual explanations, Individual Conditional Expectations (ICE) extend the Partial Dependence Plot (PDP) method by providing individual-level plots (Friedman, 2001; Goldstein et al., 2015). While PDP provides the average effect of a feature on predictions, ICE generates disaggregated plots for specific data points by altering one feature while keeping others constant. This approach provides a granular understanding of how feature changes affect individual predictions. ICE is model-agnostic and can be applied to many black-box models, making it a valuable tool for local and global interpretability.

XAI frameworks, such as SHAP, Sobol, Morris, FAST, PDP, and ICE, help make AI models interpretable and trustworthy. These frameworks utilise SA to quantify the impact of input features on predications and provide various explanation forms, from visual heatmaps to textual and numerical outputs, suitable for different user needs (Viswan et al., 2024). Whether used for local instance-based explanations or global insights into model-wide behaviour, these tools help bridge the gap between complex AI systems and human understanding, ensuring that AI-driven decisions are transparent and explainable across different domains.

1.2. Neurodegenerative Diseases

As the primary dataset utilised in this research is derived from Alzheimer's Disease (AD) research, this section provides a brief overview of Neurodegenerative Diseases (NDD), with a particular emphasis on AD, to provide contextual understanding for the experimental work. The AD dataset serves as a critical benchmark for validating the proposed methodologies, such as feature selection, sensitivity analysis, and transfer learning techniques. While the core focus of this thesis lies in advancing AI strategies, it is essential to introduce AD to justify its relevance as a complex, high-dimensional, and real-world dataset that presents unique challenges in classification and model interpretability.

NDD is an umbrella term that refers to a range of conditions that involve the progressive loss of neurons in the brain, spinal cord, and central and peripheral nervous systems. Most diseases typically stem from a combination of lifestyle, environmental, and genetic factors. Despite their distinct pathologies, each of the NDDs originates from abnormal protein buildup (Ross and Poirier, 2004). Most NDDs are irreversible; however, proactive management can help mitigate their impact. Treatment aims to control symptoms and slow the progression of the disease. Healthcare providers use various therapies and medications to enhance the patient's quality of life. An ageing global population increases NDD prevalence, presenting a significant public health challenge. Although they have different mechanisms, most NDDs exhibit common traits, such as progressive neurodegeneration and cognitive or motor decline.

1.2.1. Changes in the brain for progression to AD

AD has several clinical forms and is one of the NDDs which progresses in stages, affecting an individual's cognitive abilities and daily living activities (Zhang and Jiang, 2015). AD presents itself in patients at distinct times and with different severities. The symptoms gradually progress in severity and take several years before reaching their peak. Nonetheless, the pace of the disease and the set of symptoms manifesting may vary remarkably from one person to another.

Researchers usually simplify the progression to Alzheimer's Disease (AD) into three phases: the Cognitively Normal (CN) stage, the Mild Cognitive Impairment (MCI) stage, and the Alzheimer's Disease (AD) stage. The literature often uses various designations for subjects

in the preclinical period with no apparent cognitive symptoms, such as the early stage of CN, No Dementia (ND), Normal Condition (NC), and Healthy Controls (HC).

In the Cognitively Normal (CN) or preclinical stage, individuals show no overt symptoms of AD. However, evidence suggests AD symptoms can begin affecting the brain up to 20 years before clinical diagnosis (Rajan et al., 2015). Signs of change, such as beta-amyloid plaques and tau neurofibrillary tangles, are visible in the brain during this stage- both are diagnostic features of Alzheimer's (Paula et al., 2009). These neurotoxic proteins damage neurons and disrupt brain connectivity, even in the absence of clinical symptoms. To detect brain changes, Doctors may diagnose preclinical AD through cerebrospinal fluid (CSF) analysis or advanced imaging techniques, such as MRI and PET scans (Blennow et al., 2015). This condition is commonly caused by excessive lipid accumulation and marks the final stage of AD, where patients show measurable cognitive symptoms.

Figure 1- 3 presents typical MRI images corresponding to different stages of Alzheimer's disease, such as normal, very mild, mild, and moderate stages adapted from Battineni et al. (2021).

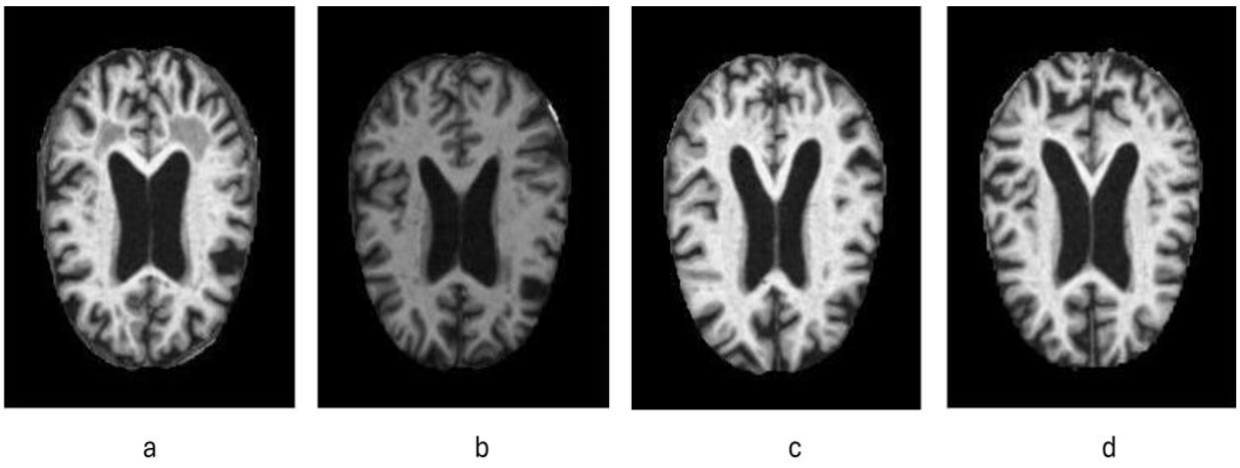

Figure 1- 3 MRI Images presenting different AD Stages. a. non-demented; b. very mild dementia; c mild dementia; d moderate dementia Battineni et al. (2021)

As the pathology progresses to Mild Cognitive Impairment (MCI), noticeable cognitive impairments emerge, but they are not severe enough to significantly disrupt daily activities. MCI often signals AD, linked to memory declines, articulation issues, and executive dysfunction, such as planning or reasoning difficulties, which become evident to individuals and families. The key brain structures affected include the hippocampus, amygdala, and

entorhinal cortex, which are crucial for memory formation and spatial orientation. In this degeneration, the affected regions shrink, while the ventricles enlarge. Early memory deficits, particularly the shrinking of the hippocampus, indicate the onset of these stages (Apostolova et al., 2012).

Figure 1- 4 illustrates a brain cross-sectional image highlighting the differences between a healthy brain and one affected by Alzheimer's disease (Tamanini et al., 2009).

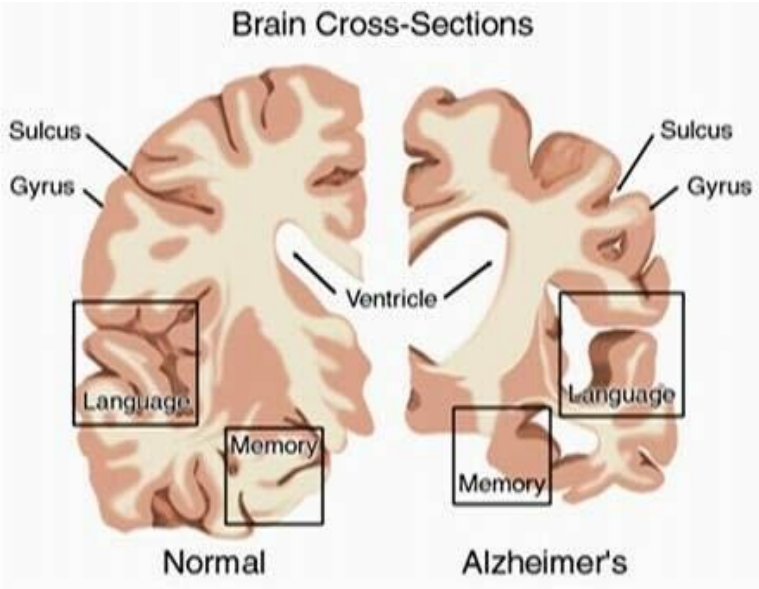

Figure 1- 4 Difference in the structure of the brain between the normal and Alzheimer's (Tamanini et al., 2009)

As the condition progresses to the stage of Alzheimer's Disease (AD), the individual experiences a severe cognitive decline, and their ability to perform typical daily activities significantly deteriorates. At this level, the individual faces extreme memory problems, disorientation, and a failure to recognise people, places, and things that are otherwise familiar. There are still significant structural changes in the brain, in addition to the existing condition, which represents a further advancement in the shrinkage of both the hippocampus and the cerebral cortex, along with an increase in the size of the ventricles. The entorhinal cortex, which plays a role in language, reasoning, and social interaction, undergoes significant atrophy. As a result, many people lose these skills, experience personality changes, and struggle with declining reasoning abilities. The illness progresses through acute exacerbation before being categorised into three stages: mild, moderate, and severe.

AD ultimately causes a total breakdown of cognitive and physical functions, preventing patients from performing basic activities. Its progressive nature makes AD one of

the most challenging NDDs for both patients and caregivers. Progression to the severe stage signifies a decline in cognitive abilities but also creates a heavy emotional burden on caregivers and the healthcare system.

1.2.2. Early diagnosis and its impact

Early diagnosis has enormous potential to enhance the quality of life for those affected by NDDs. In addition to benefiting patients, this proactive approach significantly benefits the economy, society, and families by decreasing social, financial, and emotional burdens. Even though these diseases may not currently have a cure, early diagnosis can make all the difference in managing NDDs. Primarily, it facilitates prompt intervention and access to existing treatments and therapies that have the potential to decelerate the progression of the illness and enhance symptom management.

Furthermore, an early diagnosis enables patients and their families to make plans. It offers a chance to make well-informed decisions about financial arrangements, legal issues, and care while actively enabling the person with NDD to participate in these discussions. By taking this proactive measure, the patient and their family can experience less stress and uncertainty, which enhances their emotional health and facilitates a seamless transition into care arrangements. In addition to these individual advantages, early diagnosis extends to broader financial impacts.

Early diagnosis additionally has significant economic implications. Early detection of NDDs makes outpatient care and support services cost-effective, whereas later diagnosis can often require expensive hospital stays and long-term care. Early diagnosis can lessen the financial stress on individuals, insurance companies, and healthcare systems. It could also enable patients to work longer to contribute significantly to society, reducing the burden on disability support services.

Early diagnosis enables targeted healthcare policies and resource allocation. Government organisations can establish screening programs, fund research, and create infrastructure to meet the needs of NDD patients. These initiatives reflect a commitment to public health and enhance care standards. An early diagnosis also has emotional impacts on families. Caring for someone with an NDD can be exhausting, but early diagnosis enables families to plan and adapt. It offers opportunities to build support networks, seek counselling,

and explore local resources, ultimately empowering families to enhance emotional and medical care, improving their quality of life. While early diagnosis offers many benefits, it also relies heavily on technological advancements. AI methods have shown considerable promise in improving the speed and accuracy of diagnosis for NDDs such as Alzheimer's.

1.2.3 Brief of Stages of AD and MMSE Scores

AD is a severe neurological disorder with global consequences, affecting millions of people and their families. AD and other NDDs create complex health challenges that significantly impact healthcare systems. Individuals diagnosed with AD go through different phases, such as MCI, EMCI, and LMCI. Every phase shows specific clinical signs of deterioration, requiring accurate diagnostic standards and careful observation.

Diagnosis of MCI is vital because it marks the beginning stage of cognitive deterioration and provides a chance for prompt intervention, which could help slow down the progression of the disease. AD progresses from MCI through mild, moderate, and severe stages before reaching the terminal phase. Precise categorisation and forecasting of disease advancement are crucial in creating successful treatment strategies.

The MMSE is commonly used as a screening tool to assess cognitive function and detect any impairment. It evaluates various cognitive domains such as orientation, registration, attention, calculation, recall, language, and visuospatial abilities. A perfect score on the MMSE is 30 points, indicating higher or normal cognitive capability (Joshi et al., 2019).

The MMSE plays a crucial role in differentiating cognitive impairments in AD diagnosis and distinguishing between normal ageing and pathological decline. MMSE scores decrease gradually as AD advances, which is helpful for both diagnosing and monitoring the progression of the disease. MCI can result in minor decreases in scores, whereas substantial drops suggest the early, moderate, or late phases of AD.

The MMSE is crucial for detecting cognitive impairment and tracking the progression of AD. This assists in prompt intervention, customised treatment plans, and assessing treatment results. Observing the different stages of AD results in personalised care plans for individuals, improving their quality of life and advancing the development of treatment for neurodegenerative conditions.

1.2.4 AI methods for AD diagnosis

The healthcare sector is increasingly utilising AI for its exceptional ability to detect hidden patterns within complex and large datasets. This is essential for identifying diseases that display slight changes, such as AD. This type of NDD poses a significant challenge in early detection due to its subtle initial symptoms, such as minor reductions in brain volume. Traditional diagnostic methods, which heavily rely on expert analysis, are often time-consuming and limited by the availability of skilled radiologists. The ability of AI to detect subtle changes in brain structure plays a crucial role in early diagnosis. ML and DL models efficiently diagnose AD by processing large datasets, significantly reducing reliance on skilled human intervention. These technologies are essential for diagnosing AD promptly by identifying subtle alterations in brain structure.

Researchers widely apply ML in diagnosing AD at different stages using various input types, such as MRI scans, cognitive tests, and electronic health records. Early attempts to use AI for AD diagnosis centred on supervised ML techniques, such as decision trees, random forests, SVM, and ANN (Salvatore et al., 2016; Song et al., 2021). These ML models view disease diagnosis, staging, and prognosis as classification problems, where medical experts select discriminative features to achieve adequate disease classification (Moreno-Ibarra et al., 2021). Among the various ML algorithms, ANN algorithms have shown enhanced performance in similar tasks because of their capability to capture complex, nonlinear correlations present in the data.

Recently, ML techniques have been overshadowed by the emergence of DL models. While ML models rely on manually selected features, DL models can automatically extract essential features from complex data sets, offering enhanced performance. In comparison, DL models, specifically DNNs, have become valuable tools for examining high-resolution brain scans using different imaging methods such as structural MRI (sMRI), functional MRI (fMRI), and Positron Emission Tomography (PET) scans. In contrast to traditional ML models, DNNs can automatically discover important features from raw input data, avoiding the necessity of manual feature selection and minimising the risk of human error (LeCun et al., 2015). This is particularly advantageous in cases of AD, as the heterogeneous nature of the disease and the subtle early signs demand precise image analysis.

The ability of DL models to process and understand high-dimensional data is a significant factor in their success in AD detection. For instance, DL models can detect detailed brain atrophy patterns related to the early stages of AD from MRI scans. Additionally, studies have demonstrated that DL models surpass ML algorithms in terms of accuracy and precision, particularly in research comparing the two methods for AD diagnosis (Asl et al., 2018; Sarraf and Tofighi, 2017). Techniques such as stochastic gradient descent (SGD) and dropout are advantageous in improving optimisation processes. Implementing these methods has enhanced the ability of DL models to generalise effectively, enabling them to perform exceptionally well on diverse and complex datasets (Srivastava et al., 2014). These methods help avoid overfitting, ensuring the accurate prediction of AD stages across diverse patient groups.

Although DL models outperform traditional ML methods, critical domains have restricted the use of DL technologies. The primary cause of this hesitancy is the ambiguity surrounding the decision-making process of these models. While ML models are transparent and interpretable, DL models are often perceived as opaque "black boxes," posing challenges for understanding the reasoning behind a particular prediction.

The consequences of this lack of explainability are widespread. Clinicians and patients may be reluctant to trust the predictions of a DL model for essential healthcare decisions if they do not fully understand how the model reaches its conclusions. This reluctance may delay the adoption of advanced tools, limit the use of life-saving technologies, and perpetuate continued dependence on less precise methods.

1.3. Overview of the Focus of Research

The overall objective of this research is to enable the use of AI in critical real-world applications, particularly within MRI-driven healthcare investigative processes such as predicting Alzheimer's Disease (AD) stages onset. Despite advancements in the field, challenges remain, such as complex models that lack explainability and interpretability, dealing with high-dimensional data and the need for substantial amounts of data. This thesis investigates three main disciplines to tackle these problems: Feature selection (FS), sensitivity analysis (SA), transfer learning (TL), and Autoencoder. Choosing the right features enhances model performance by determining the most important features, simplifying complexity, and

boosting explainability and computational efficiency. SA enhances comprehension of factors influencing DNN predictions, which is critical for their explainability and acceptability in real-world applications. TL, in combination with Autoencoder, enables the use of information from larger and related tasks, improving the robustness of the model with a smaller dataset and reducing the time required for training. This research aims to address a crucial gap in explainability and enhance the accuracy of models. Therefore, in line with the central focus of this study, a brief overview of the three topics—feature selection, sensitivity analysis, and transfer learning—is presented in the upcoming sections.

1.3.1. Feature Selection

Feature selection is a crucial preprocessing step in data analysis, particularly significant in MRI-based Alzheimer's disease research, where datasets are high-dimensional. It involves choosing a specific group of important features from all available features to train a model. FS seeks to enhance model performance by decreasing dataset dimensionality and eliminating irrelevant or noisy data that may cause overfitting and diminish generalisation capacity. The main goal is to enhance model precision, reduce computational complexity, and ensure an explainable model.

FS provides various advantages in AI applications. Enhanced model performance by decreasing the number of irrelevant features with FS enhances model accuracy and predictive power. Reducing overfitting is achieved by selecting only the most important features, which prevents the model from learning noise and forces it to focus on the true underlying data patterns. Lower computational costs and quicker training times with fewer features benefit large datasets and real-time applications. Reduced features enhance model interpretability, crucial in scientific domains for understanding feature impacts.

Although FS has its benefits, it also has potential drawbacks, such as the loss of valuable information. This challenge is frequently encountered in Alzheimer's Disease (AD) research due to wide-ranging heterogeneous biomarkers. Removing seemingly irrelevant features may result in discarding interactions that could enhance model performance. Additionally, computationally intensive wrapper methods involve retraining the model for various feature subsets. The indiscriminate inclusion of features, particularly in high-dimensional datasets, can lead to sub-optimal model performance, increasing the risk of

overfitting or underfitting. Applying dimensionality reduction and feature engineering techniques is essential to optimise the feature set, ensuring the model learns meaningful patterns rather than noise. This is particularly critical in domains with naturally imbalanced data distributions, where careful FS directly impacts model reliability and generalisability.

Several studies have explored strategies to identify the most informative features from high-dimensional data, highlighting that robust FS is central to improving MRI-based AD classification. Gallego-Jutglà et al. (2015) proposed a hybrid FS approach using synchrony measures and frequency-relative power derived from EEG signals. This demonstrates that multi-feature classifiers significantly outperform single-feature systems, achieving up to 100% classification accuracy in Mild AD detection. Similarly, Faisal.F.U.R. et al. (2021) introduced a combined FS technique, integrating Principal Component Analysis (PCA) with Recursive Feature Elimination (RFE) to reduce dimensionality while retaining crucial structural MRI features. Their method achieved high classification accuracies (over 95%) across different AD subtypes using Support Vector Machines (SVM). In a broader context, Rado et al. (2019) assessed multiple classification and FS methods across varied datasets, highlighting that optimal feature selection enhances predictive performance and reduces model complexity while improving discriminative efficiency.

The selection of features continues to be a crucial research area, particularly in the realm of explainability in DL, as there is a rising concern about model transparency. Furthermore, it is crucial for enhancing model accuracy, decreasing overfitting, and improving interpretability, particularly in cases involving large datasets. Nevertheless, applying feature engineering techniques should be undertaken cautiously, as the selection of non-informative features could result in either loss of valuable information or underfitting, highlighting the importance of ongoing exploration into resilient techniques.

1.3.2. Sensitivity Analysis

XAI frameworks utilise Sensitivity Analysis (SA) techniques for evaluating the impact of variations in input variables on the resultant output, a crucial requirement when analysing MRI data, where transparency in decision-making is essential. The primary objective of SA is to determine the input variables that significantly influence the predictions made by the

model. This approach will provide a deeper understanding of model performance and help develop robust and reliable models.

SA is essential for interpreting models, particularly in medical applications, where input-output relationships are usually nonlinear and challenging to comprehend. It helps identify the key attributes and assesses the robustness of the model by analysing how slight variations in input data can affect results.

SA provides numerous important advantages, including enhanced model interpretability by understanding how input features influence model predictions. SA can guide the FS procedures by pinpointing the input variables that have the most significant impact. This approach can simplify the model and enhance its ability to generalise. SA enables the assessment of the robustness of a model by examining its sensitivity to minor perturbations in input data. It helps ensure that the model performs consistently across various scenarios. Model assumption validation through SA assists in validating assumptions made in model development. It ensures the model functions correctly with varying inputs and identifies areas for possible improvement.

Although SA has numerous benefits, it has some drawbacks, such as computational complexity, which is a significant consideration when dealing with high-dimensional MRI-derived data. Global Sensitivity Analysis (Global SA) evaluates the impact of input variations on the output across the entire input space, providing a comprehensive measure of feature importance. Specific techniques, such as global SA, require substantial computational resources and may not be feasible for extensive datasets or complex models. Local SA assumes that linear relationships between input features and output exist in local methods, but may not apply to nonlinear models, as small changes in input may not result in proportional changes in output. SA can demonstrate which inputs affect outputs without causality, but it does not provide causal relationships. Researchers may need to examine further, as variations in the output do not always link to a sensitive input. The effectiveness of SA relies on the quality of the underlying model. If there are errors in the model, the conclusions drawn from the analysis could be incorrect.

De Santi et al., (2023) proposed an explainable convolutional neural network using 18f-FDG PET images to enhance early diagnosis while offering insight into the decision-making

process of the model. El-Sappagh et al. (2021) developed a multilayer, multimodal model that integrated 11 data types and utilised random forests alongside SHAP and fuzzy rule-based systems to generate both global and patient-specific explanations. Similarly, Chun et al. (2022) used interpretable ML techniques to predict conversion from amnestic mild cognitive impairment (aMCI) to AD, employing SHAP and ICE to identify key risk factors per individual. Across these studies, SA techniques such as SHAP have proven essential for elucidating the contribution of individual features, supporting model transparency. These approaches demonstrate how interpretability and performance can be jointly optimised to enhance trust and applicability in real-world practice.

SA is a valuable tool for understanding the behaviour of DL models, offering insights into which input features most influence the output. SA plays a critical role in model development and validation by enhancing explainability, guiding feature selection, and evaluating robustness. Despite challenges such as computational complexity and the assumption of linearity in some methods, the field continues to evolve. As SA becomes integrated with complex models such as DNN, its importance in ensuring reliable and interpretable AI systems will only increase.

1.3.3. Transfer learning with autoencoders.

Transfer Learning (TL) involves taking a model trained for one task and adjusting it to carry out a different but related task. This approach is especially advantageous in MRI-based AD classification, where data scarcity is commonly observed.

The concept involves utilising knowledge gained by a model through being trained on the dataset from one domain and implementing the model in another related domain. This approach is particularly valuable when the new task has a small amount of data. TL enables the model to utilise knowledge from a larger, related dataset, enhancing performance with reduced training data.

Autoencoders, an unsupervised technique used in NN, reduce dimensionality by compressing data and extracting features. They consist of two components: an encoder that compresses input data and a decoder that reconstructs it. An autoencoder learns a compact representation of input data while minimising the reconstruction error. Researchers widely use autoencoders in applications such as anomaly detection, image denoising, and

dimensionality reduction. The latent space representations learnt by autoencoders are particularly valuable in transfer learning scenarios, where they can serve as feature extractors for downstream tasks.

Using autoencoders with TL offers numerous advantages. The main advantage of this approach is enhanced generalisation with small datasets, as it utilises representations gained from a larger dataset. This approach can significantly decrease overfitting and enhance model performance. Autoencoders reduce dimensionality by compressing data into a lower-dimensional latent space, simplifying the model, and improving training efficiency. This approach primarily benefits tasks with high-dimensional input data and limited labelled data. The encoder component extracts valuable features that can be utilised in tasks further down the line. The acquired features are insightful than the original data, enhancing the performance of the model in classification, regression, or clustering activities. Autoencoders can undergo unsupervised pretraining, which enables them to train without needing labelled data. This approach enables them to learn from vast quantities of unlabelled data, which is typically easier to acquire than labelled data. These acquired characteristics can support supervised tasks with limited labelled data. When combined with autoencoders, transfer learning is flexible and applicable across various domains, such as computer vision, natural language processing, and medical imaging.

Although TL using autoencoders has its benefits, it also has some drawbacks. TL is most effective when the source and target tasks have a strong connection. If the source data differs vastly from the task, the Autoencoder may not learn transferable features, resulting in poor performance. Training an autoencoder on a large dataset can require many computational resources, mainly when using a complex model architecture. Using pre-trained models or utilising cloud computing resources can reduce this issue. Although TL helps with limited datasets, inadequate fine-tuning can still lead to overfitting. The model might be memorising the limited dataset instead of generalising from the transferred characteristics. It can be challenging to optimise performance on a smaller dataset when fine-tuning the encoder and decoder of an autoencoder for a new task, necessitating meticulous hyperparameter tuning and experimentation.

Nanni et al. (2020) compared TL with traditional ML using structural MRI, finding that while ensemble TL models performed well—achieving an AUC of 90.2% for AD vs CN—classical

methods with careful feature engineering often outperformed them in some classification tasks. However, TL remained competitive in distinguishing MCI converters from non-converters. Gao et al., (2020) introduced AD-NET, a TL-based model incorporating age adjustment as a surrogate biomarker, which significantly enhanced MCI-to-AD conversion prediction across age groups, outperforming eight other models. This approach underscored the value of combining demographic knowledge with feature transfer. Meanwhile, Duc et al., (2020) used resting-state fMRI data and a 3D CNN to classify AD and predict MMSE scores, achieving strong results by combining group ICA features with SVM-RFE. The study demonstrated the potential of TL in enhancing DNNs models, particularly when paired with effective feature selection.

TL combined with autoencoders is a powerful technique for improving model performance on small datasets by utilising the representational power of large datasets. Autoencoders offer a robust method for extracting significant features from data, whereas transfer learning applies these characteristics to new tasks that have limited data. Despite domain mismatch and high computational cost, this approach has proven flexible and successful in various domains, generative modelling, and multimodal learning. As the field advances, sophisticated techniques and applications arise, creating new opportunities for domains with limited data.

1.4. Problem statement and its proposed solution

This section sets out the primary problem statement and motivation for this research, with a particular emphasis on MRI-based applications. It then discusses the existing research gaps in the methods, leading to proposed solutions and research objectives. It provides a comprehensive overview of AI-driven solutions, particularly in real-world applications.

1.4.1. Problem statement

The research addresses the problem of effectively analysing high-dimensional datasets such as MRI data. These datasets often contain a vast number of features, many of which may be irrelevant, making it challenging to extract meaningful insights. Diminishing the size of these data sets is essential in enhancing the interpretability of AI models, enabling them to focus on the most significant features and offering precise, actionable insights. Through sophisticated

algorithms, AI can effectively manage these high-dimensional datasets by highlighting significant features, improving model performance, and ensuring reliable solutions in real-world settings.

Nevertheless, despite enhanced performance, a significant challenge remains in adopting AI models in critical environments, where model accountability is paramount essential. Many professionals hesitate to adopt these technologies due to a lack of explainability, as they need to trust and understand how a model arrives at its predictions. Without this transparency, the potential of AI remains underutilised. Therefore, developing models that reduce complexity and offer clear, explainable outcomes is vital for gaining confidence. The absence of explainability in AI-driven models increases the risk of misinterpretation, which can affect decision-making, particularly in critical domains.

Moreover, limited datasets, a common scenario, compound the challenge. It is necessary to enable models to utilise knowledge from larger, robust datasets to enhance predictive accuracy.

1.4.2. Motivation

The growing complexity and scale of high-dimensional data present significant challenges in ML, particularly in developing models that are efficient, interpretable, and generalisable. As dimensionality increases, computational costs, training time, and the risk of overfitting escalate, making advanced feature selection and explainability techniques essential. These methods not only improve model robustness but also ensure transparency in critical decision-making contexts where black-box models are unacceptable.

Despite the success of DNNs, limited interpretability continues to hinder the broader adoption of complex models, particularly in expert-driven domains where explainability is key to trust and accountability. This thesis is driven by the need to develop computationally efficient, explainable AI solutions capable of handling high-dimensional datasets while delivering clear, reproducible, and trustworthy insights.

An additional motivation is addressing the challenge of small-sample datasets, common in many critical domains. This research integrates transfer learning with

autoencoders to enhance predictive performance in data-scarce scenarios by utilising knowledge from larger datasets.

Although the methods proposed are domain-agnostic, they are validated on Alzheimer's Disease and arrhythmia datasets. The shortage of radiologists exacerbates delays in diagnosing AD, potentially missing opportunities for early intervention, which could drastically enhance patient outcomes (Konstantinidis, 2024). The use of AD datasets offers a relevant test case due to their high dimensionality, limited sample sizes, and the practical need for model interpretability. Addressing these challenges in the AD domain further demonstrates the applicability and impact of the proposed techniques in real-world, high-stakes environments.

1.4.3. Research Gap

This research addresses critical gaps in merging AI and healthcare, focusing on interpretability, explainability, and model performance. Improving model performance and interpretability through feature selection is essential in real-world applications. Effective feature selection strikes a balance between model complexity and predictive power, improving both accuracy and interpretability. While existing literature highlights the need for algorithms that reduce dimensionality while preserving informative features (Jia et al., 2022), a significant research gap exists in validating these algorithms on external datasets to ensure their generalisability.

Explainability is increasingly important due to its influence on decision-making and the necessity of trust in AI systems. Despite the use of sensitivity analysis techniques to assess the impacts of individual features, there is a lack of systematic comparisons to identify commonalities and integrate these methods into a robust, ensemble-based approach. This gap highlights the need for standardised metrics to evaluate explainability and ensure responsible AI deployment.

The combination of transfer learning and autoencoders is also crucial in advancing research in applications involving small datasets. Grasping the trade-offs in interpretability, computational complexity, and generalisability in this context is essential. Transfer learning offers a promising solution by enabling models to utilise knowledge from larger, comprehensive datasets, enhancing performance on smaller, specialised datasets. Despite its

potential, research lacks the best approach to applying transfer learning with autoencoders to maximise accuracy and reliability.

Without addressing these issues, the potential of predictive models to address real-world problems, remains limited. This study aims to fill these gaps by developing and validating advanced feature selection algorithms, systematically comparing sensitivity analysis techniques, and creating a multi-stage algorithm that utilises transfer learning with autoencoders to enhance predictive accuracy in datasets with limited samples. This research is critical for enhancing the interpretability, explainability, and performance of models used in critical application research, ultimately contributing to DNNs adoptability.

1.4.4. Proposed solution

The proposed solution focuses on developing advanced algorithms to tackle key challenges in analysing real-world datasets and is validated on AD datasets to ensure domain relevance. The first component involves designing and developing feature selection algorithms to reduce these dataset dimensionalities effectively. By focusing on the most relevant features, these algorithms not only streamline the data for efficient processing but also enhance the interpretability of the resulting models. This enhanced interpretability will provide precise insights into the underlying data patterns, enabling researchers to understand the factors contributing to the prediction. The generalisability of these algorithms will be validated using an external dataset, ensuring that the solutions are robust across different contexts and not merely tailored to a specific dataset.

The second component focuses on developing and systematically comparing various sensitivity analysis techniques. Understanding the influence of individual features on model output is crucial. Comparing different techniques will help identify commonalities and unique strengths among them. This knowledge will help develop an ensemble approach that combines similar results from various methods, enhancing the reliability of the model outcomes. This method enhances the explainability of the decision-making process of the model, increasing transparency and reliability.

The last component includes designing, developing, and testing a multi-step algorithm that combines transfer learning with autoencoders, particularly relevant for MRI applications constrained by small patient cohorts. This algorithm enhances the performance

of models trained on datasets with fewer samples, a common issue in critical research domains. By transferring knowledge from larger datasets, the algorithm will enhance the predictive capabilities of the model, leading to accurate predictions. This sequential method will provide a detailed comprehension of model development for such scenarios.

1.4.5. Objectives

a) Design and develop feature selection algorithms to reduce the dimensionality of high-dimensional datasets with validation on AD datasets to ensure clinical applicability. These algorithms will be validated on external datasets to demonstrate their generalisability. Effectively reducing the number of features enhances the interpretability of the models, offering precise insights into the underlying data patterns.

b) Develop and systematically compare various XAI frameworks. This comparison will identify commonalities among the techniques, enabling the creation of an ensemble approach. Integrating similar results enhances the robustness and generalisability of the model outputs, enhancing the explainability of the model in the decision-making process.

c) Design, develop, and validate a multi-stage algorithm that uses transfer learning with autoencoders for AD datasets where the sample size is limited. This algorithm enhances model performance on datasets with limited samples by transferring knowledge from larger, relevant datasets.

1.5. Structure of the Thesis

This section presents a meticulously structured framework that encapsulates the core elements of the research. The aim is to enhance the explainability and performance of deep neural networks in healthcare applications with limited data. This research follows a meticulously crafted and structured approach, encompassing six distinct chapters that contribute to a comprehensive understanding of the research.

Chapter 2 presents a comprehensive survey of the related research domains. The first section provides an extensive literature review of feature selection (FS), sensitivity analysis (SA), and transfer learning (TL) in the context of DNNs. Subsequently, the chapter provides an in-depth examination of explainability and interpretability techniques in machine learning,

with particular attention to their significance for AI-driven systems. The discussion systematically categorises key explainable AI (XAI) approaches based on explanation timing, scope, model dependency, and methodological type. This structured review establishes the theoretical foundation for the methodologies proposed in the subsequent chapters and situates the current research within the broader context of contemporary AI advancements.

Chapter 3 explores the dataset used in this research, delving deeply into its sources and elucidating their essential contributions to the research scope. Explore the complex process of dataset preprocessing using the FreeSurfer tool, describing the steps in transforming raw MRI scans into a structured tabular format. Moreover, it provides insights into the post-processing activities applied to the dataset and discusses the resulting refined dataset. This chapter incorporates various visualisation charts to enhance understanding of the dataset, offering valuable insights into its statistics, general trends, and any novel findings.

Chapter 4 explores feature selection techniques applied in the research. The first part presents a comprehensive literature review of existing techniques, providing a comprehensive understanding of their content and methodologies. Next, the methodology section presents two novel feature selection techniques based on correlation and clustering. The method developed was subsequently evaluated using an external dataset and a benchmark algorithm known as ReliefF. The correlation-based method produced a simplified feature set, leading to straightforward interpretability with enhanced accuracy. In contrast, the clustering-based approach produced four features, retaining accuracy similar to the complete feature set. These dimensionality reductions enhanced model robustness and interpretability, ultimately unveiling deeper insights into variable relationships.

Chapter 5 focuses on the SA techniques used in the research. The initial segment of this chapter involves an exhaustive literature review of SA within the domain of DNN. It also offers critical assessments of the current methods and their respective approaches. Subsequently, the thesis sets out to the methodology section, which centres on the novel ensemble SA approach and its design and offers comprehensive insights into its implementation. The chapter culminates in a comprehensive analysis of the results produced through these methodologies, highlighting their significance and discussing their role in enhancing the explainability of the models. The results indicate that the hippocampal sub-regions, fissure/sulcus, and temporal horn of the lateral ventricle can be considered the most

important features in predicting AD. The findings are consistent with earlier results from medical experts, underscoring the impact of the research.

Chapter 6 presents a dedicated examination of the transfer learning and autoencoder models employed in the research. The chapter begins with a comprehensive review of the literature, shedding light on the importance of using transfer learning within the scope of the research and evaluating the existing architectures. The methodology section presents a detailed explanation of the innovative multi-stage algorithm developed. It then delves into discussing its architecture and implementation. The chapter concludes with a rigorous analysis of the results obtained from these methodologies, emphasising their significance within the research context and discussing their contributions to the overall research focus. This approach resulted in around 73.26% accuracy with a standard deviation of 3.92%, with an improvement of approximately an accuracy of 12.18% in comparison to a basic regression model.

Chapter 7, the final chapter of this thesis, presents the conclusion of the summary findings from the previous chapters. This chapter rigorously summarises the conclusions arrived at through each approach discussed. It emphasises the associations between these approaches, highlights their joint contributions to the research goal, and provides a robust approach to support the diagnosis and monitoring of DNN progression in healthcare. This joint viewpoint enables a deeper comprehension of the importance of the research findings. Further, the chapter discusses the limitations and future directions of the research.

2. Related Work

2.1 Literature Review for Feature Selection

Faisal. F.U.R. et al. (2021) explore the early diagnosis of AD using sMRI and traditional ML approaches, focusing on model complexity and feature redundancy challenges. The paper targets the differentiation between AD, Mild Cognitive Impairment (MCI), and Healthy Control (HC) populations using T1-weighted images, a widely used imaging modality in neurodegenerative research.

The study utilises data from the ADNI dataset, comprising 308 subjects with combined subcortical and cortical features. Three binary classification experiments are conducted: AD versus Early MCI, AD versus Late MCI, and AD versus Healthy Cohorts. The authors propose an improvised FS that combines Principal Component Analysis with Recursive Feature Elimination to address the high dimensionality inherent to neuroimaging data. This dual approach serves a twofold purpose: reducing the size of the dataset and selecting the most discriminative features, thereby simplifying the model while maintaining predictive power. The experimental results indicate that the SVM classifier performs best, with impressive accuracies of 97.87% for AD versus EMCI, 95.83% for AD versus LMCI, and 97.83% for AD versus HC. These high classification accuracies demonstrate the potential of the combined FS method in enhancing the diagnostic performance of traditional ML models for AD identification.

A significant advantage of the study lies in its practical solution to the dimensionality challenge frequently encountered in neuroimaging analysis. The approach reduces computational complexity by integrating PCA and RFE while preserving crucial diagnostic information. This is particularly valuable when dealing with limited datasets, as is often the case in clinical studies. However, the modest sample size of 308 subjects may restrict the generalisability of the findings. Additionally, while traditional ML methods such as SVM have shown high performance, the study does not compare their approach with modern DL techniques that have become increasingly prevalent in this domain. Further validation on larger, independent cohorts is also necessary to fully establish the clinical utility of the proposed framework. In summary, the research provides a robust framework for AD diagnosis

using sMRI by effectively addressing feature redundancy and model complexity through a combined PCA-RFE method, yielding high accuracy and promising diagnostic potential.

Farouk and Rady (2020) investigate the potential of unsupervised ML for the early diagnosis of AD, highlighting a key challenge in the field, the frequent lack of or inaccuracy of labelled data in medical datasets. Rather than relying on traditional supervised classification methods, the research employs clustering algorithms to differentiate between stages of brain deterioration using MRI data.

The authors focus on two widely used unsupervised learning techniques, k-means and k-medoids, to cluster subjects based on Voxel-Based Morphometry (VBM) features extracted from structural MRI scans. These features reflect local differences in brain anatomy and are particularly useful for identifying subtle atrophic patterns associated with early-stage AD. A crucial methodological comparison is drawn between two levels of anatomical analysis: whole-brain global features and region-of-interest-based local features. This comparison helps evaluate whether focusing on specific brain regions enhances diagnostic performance. The best-performing approach in the study achieves an accuracy of 76%, demonstrating that even without labelled data, clustering methods can provide meaningful groupings that may align with disease progression. While this accuracy is lower than that reported in supervised models, the value of the research lies in exploring alternative diagnostic methods when labelled data is unreliable or unavailable.

The study is commendable for challenging the conventional classification-based pipeline in AD diagnosis and presenting unsupervised learning as a viable alternative in low-resource or early-stage research settings. However, it also reflects the limitations of clustering in clinical applications, particularly regarding diagnostic precision and interpretability. Overall, this work contributes a novel angle to the literature by advocating for label-free approaches in early AD diagnosis and emphasising the utility of VBM features in distinguishing AD-related brain changes, even in an unsupervised context.

Graña et al. (2011) present a computer-aided diagnosis (CAD) system for AD that utilises features derived from diffusion tensor imaging (DTI), specifically focusing on fractional anisotropy (FA) and mean diffusivity (MD) metrics. The study aims to identify discriminative

features from these scalar measures to train classifiers capable of distinguishing AD patients from healthy controls.

The methodology involves computing correlation using the Pearson method between FA or MD values across subjects and the corresponding class labels at each voxel. Voxels exhibiting high absolute correlation values are selected as features for classification. An SVM classifier, particularly with a linear kernel, is trained and tested using these selected features. The dataset comprises anatomical T1-weighted MRI volumes and DTI data collected from healthy control subjects and AD patients at the Hospital de Santiago Apostol.

The results demonstrate that using FA features with a linear SVM classifier achieves perfect accuracy, sensitivity, and specificity in several cross-validation studies. This underscores the potential of DTI-derived features as effective imaging biomarkers for AD and supports the feasibility of developing CAD systems based on these metrics. This study contributes to the field by highlighting the efficacy of combining DTI-derived features with ML techniques for early and accurate diagnosis of AD. The approach offers a promising avenue for enhancing diagnostic tools and potentially aiding clinicians in assessing AD.

Karegowda et al. (2010) explore the importance of feature subset selection in data mining, particularly for high-dimensional data, which makes training and testing classification models challenging. The paper compares two FS methods: Gain Ratio and Correlation-based FS (CFS). These methods are used to identify the most relevant features for classifying the Pima Indian Diabetes Dataset (PIDD), which is commonly used for evaluating ML algorithms in the medical domain.

The paper uses the decision tree algorithm with Gain Ratio to split the data and select the most informative features. Additionally, a Genetic Algorithm (GA) is employed as a search method, with CFS being used as the evaluation mechanism for feature subsets. The resulting feature subsets are then tested with two supervised classification methods: the Backpropagation Neural Network (BPNN) and the Radial Basis Function Network (RBFN). These classifiers were chosen for their ability to model non-linear relationships in the data, making them suitable for a wide range of classification tasks.

The experimental results demonstrate that the CFS method, which uses correlation to assess feature relevance, significantly enhances classification accuracy as compared to the

Gain Ratio method. Both methods reduce the features needed for effective classification, but CFS yields an accurate classification model for both BPNN and RBFN classifiers. These findings highlight the importance of feature subset selection in improving the performance of classification algorithms, particularly when dealing with high-dimensional data. The paper also suggests that using a Genetic Algorithm for a feature subset search enhances the FS process, leading to classification accuracy.

The study concludes that CFS is effective than the gain ratio method in selecting feature subsets that enhance classification accuracy. It also demonstrates the significance of combining FS techniques with search algorithms such as Genetic Algorithms to enhance ML models. The results underline the potential of feature subset selection in simplifying complex datasets and improving the performance of ML models, particularly for tasks such as medical diagnosis, where FS can play a crucial role in obtaining accurate predictions.

Chormunge and Jena, (2018) address the dimensionality problem in data mining tasks, particularly focusing on FS, a critical technique for handling high-dimensional data. Traditional FS algorithms often struggle to scale efficiently when dealing with large datasets, leading to the need for effective methods.

The authors propose a novel approach integrating clustering with correlation-based FS to enhance feature subset selection. The method works in two key stages. First, irrelevant features are eliminated using the k-means clustering algorithm. This clustering approach groups the features based on similarity, enabling the algorithm to identify and remove those features that do not contribute significantly to the classification task. Once the irrelevant features are eliminated, the next step involves selecting relevant features within each cluster using a correlation measure. This step ensures that only the most informative features are retained, minimising redundancy and improving the efficiency of the model. To evaluate the effectiveness of the proposed method, the authors test it on Microarray and Text datasets, which are commonly used in ML research. The performance of the method is compared with several well-known FS techniques, and the Naïve Bayes classifier is employed to assess the classification accuracy. A percentage-wise criterion is used to measure the accuracy of the proposed method across different numbers of relevant features, enabling an objective comparison.

The experimental results show that the proposed method significantly outperforms traditional FS methods in terms of efficiency and accuracy. By effectively reducing dimensionality and selecting the most relevant features, the approach enhances the performance of classification tasks, particularly when dealing with high-dimensional data. Combining clustering with correlation-based selection enables the model to handle large datasets efficiently, a common challenge in many real-world applications. In conclusion, the paper demonstrates that combining clustering techniques with correlation-based FS effectively solves the dimensionality problem in data mining. The proposed method proves to be a robust approach to identifying and selecting relevant as distinct from non-redundant features, thereby improving the efficiency and accuracy of classification models.

Yu and Liu (2003) propose a novel approach for FS in high-dimensional data. They introduce the "predominant correlation" concept to identify relevant features and reduce duplication among them. This method aims to overcome the limitations of traditional FS approaches, which rely on pairwise correlation analysis, making the process slower and less scalable.

The paper presents a fast correlation-based filter method that identifies relevant features and removes redundancy without the computational overhead of pairwise correlation analysis. The technique is designed to efficiently handle datasets with large numbers of features, making it particularly suited to high-dimensional data where traditional methods struggle. The authors demonstrate the efficacy of their method through extensive comparisons with other FS techniques. The proposed method outperforms existing methods in both speed and accuracy, utilising real-world high-dimensional datasets. This approach significantly reduces computational time while maintaining or improving the accuracy of feature selection, making it a viable solution for large-scale ML tasks.

The proposed fast correlation-based filter method effectively addresses the challenges of high-dimensional data in FS. By introducing the concept of predominant correlation, the method enables efficient FS that reduces redundancy and enhances model performance, offering a valuable tool for ML tasks involving large datasets.

Trambaiolli et al. (2017) explored the role of FS in improving the performance of electroencephalogram (EEG)-based classification systems for diagnosing AD. In decision

support systems, irrelevant features in the data can lead to model complexity and decrease classification accuracy. This is particularly crucial in AD diagnosis, where EEG spectral features often contain relevant and irrelevant information. Therefore, effective FS is essential to enhance the performance of the model by identifying the most informative features while eliminating noise and redundancy.

The paper investigates eight FS algorithms for EEG spectral data collected from 22 AD patients and 12 healthy age-matched controls. The authors focus on determining how these FS algorithms affect the accuracy of SVM classifiers. SVM is known for its robust performance in high-dimensional data classification, making it an ideal choice for this study. The authors use a leave-one-subject-out cross-validation strategy to assess the FS methods. This helps reduce the potential bias from small sample sizes and provides a generalisable model evaluation of the performance.

The results indicate that the Filtered Subset Evaluator method produced the best performance improvements. This method achieved an impressive accuracy of 91.18% on a per-patient basis and 85.29 ± 21.62% on a per-epoch basis, demonstrating the positive impact of FS on model performance. Furthermore, applying FS led to a substantial reduction in the number of features—88.76 ± 1.12% of the original features were removed—without compromising the accuracy of the classification task. This reduction in feature space can significantly enhance the computational efficiency of the diagnostic system, making it feasible for clinical settings.

A key finding was that all FS algorithms recognised alpha and beta frequency bands as crucial for distinguishing AD patients from healthy controls. This concurs with prior clinical studies, emphasising these frequency bands in AD diagnosis. Alpha and beta waves are known to be impacted in NDD, such as Alzheimer's, which displays changes in brain activity, particularly in the prefrontal and temporal regions. The ability of the research to replicate these findings further confirms the relevance of these frequency bands in EEG-based AD diagnostic systems.

In summary, this paper highlights the significance of FS as a pre-processing step in EEG-based AD diagnosis. By applying FS techniques, the researchers were able to enhance the classification accuracy, reduce computational complexity, and enhance the interpretability of

the model. The study demonstrates that biologically relevant EEG data, when combined with effective FS methods, can significantly boost the performance of diagnostic systems. This could pave the way for accurate, efficient, and interpretable AD detection systems, ultimately contributing to clinical decision-making in the early stages of the disease.

Sadiq et al. (2021) propose a novel approach for distinguishing AD patients from healthy controls using resting-state functional magnetic resonance imaging (rs-fMRI) data, focusing on brain connectivity patterns. The study underscores the significance of understanding the functional organisation of the brain, particularly in NDD, such as AD. Since rs-fMRI captures spontaneous brain activity during rest, it has become a valuable tool in assessing intrinsic functional connectivity and alterations associated with neurological conditions.

The authors combine Pearson correlation connectivity (PCC) and the ReliefF FS algorithm to enhance classification accuracy. PCC is a well-established statistical method used to quantify the degree of linear correlation between different brain regions, effectively creating a functional connectivity matrix that serves as a high-dimensional feature set. However, due to the large number of features typically generated from such matrices, FS becomes crucial to mitigate the curse of dimensionality and reduce model overfitting.

The study employs ReliefF, a popular algorithm for its robustness in identifying relevant features in high-dimensional datasets, to address this. ReliefF evaluates the importance of features based on how well their values differentiate between instances near each other, thus identifying informative attributes. The integration of PCC with ReliefF enables the extraction of connectivity features that are statistically meaningful and diagnostically relevant. For classification, a K-Nearest Neighbour (KNN) algorithm is used. KNN, being a non-parametric and instance-based learning technique, classifies new data based on the majority label of its closest neighbours in the training set. Despite its simplicity, KNN is particularly effective when combined with well-selected features, as in this study.

The proposed method achieves a classification accuracy of 93.5%, which indicates a strong potential for this combined approach in clinical AD diagnosis. The high performance also highlights the effectiveness of combining a connectivity-based feature extraction method (PCC) with a robust FS mechanism (ReliefF) to reduce dimensionality and retain informative

biomarkers. The study contributes to the growing body of research on ML applications in neuroimaging by offering a method that effectively utilises functional connectivity and intelligent FS. The work demonstrates that targeted use of statistical and algorithmic tools can result in high diagnostic accuracy, potentially aiding the development of early detection systems for AD. This is particularly important given the progressive nature of AD and the clinical emphasis on early intervention.

In conclusion, FS techniques remain central to addressing high-dimensionality challenges in neuroimaging and medical datasets. Filter-based methods are particularly prevalent due to their efficiency and ability to identify relevant features before model training. Among these, methods such as Gain Ratio, CFS, and ReliefF are frequently employed as they serve as a baseline for comparison. However, existing approaches often fail to effectively capture deeper inter-feature dependencies, particularly when relying solely on pairwise correlations. To address this limitation, this research chapter introduces two complementary FS methods that integrate correlation analysis with clustering principles, offering a structured approach to reducing feature redundancy and improving model interpretability in AD classification.

2.2 Literature Review for Sensitivity Analysis

El-Sappagh et al. (2021) introduced a multilayer, multimodal model designed for both the early diagnosis and progression prediction of AD, emphasising explainability. The paper targets key shortcomings in the existing literature, such as the over-reliance on unimodal data, separation of diagnosis and progression tasks, and the general lack of model transparency. The proposed model aims to bridge the gap between high-performance AI systems and clinical usability.

The study uses data from the ADNI, incorporating 11 modalities across 1,048 subjects. The cohort includes CN individuals, stable MCI (sMCI), progressive MCI (pMCI), and AD patients. The model is structured in two layers: the first performs multi-class classification (CN, sMCI, pMCI, AD), while the second focuses on the binary classification to predict MCI-to-AD conversion within three years. A random forest classifier is employed in both layers, with FS tailored to optimise performance. Importantly, the model integrates explainability at global and instance levels using SHAP, complemented by 22 additional explanation modules based

on decision trees and fuzzy rule-based systems. These explanations are also translated into natural language, enhancing interpretability for clinical users. The model achieves high-performance metrics, with a cross-validation accuracy of 93.95% and an F1-score of 93.94% in the diagnosis layer and 87.08% accuracy and 87.09% F1-score in the progression layer. Strengths of the work include its multimodal approach, unified handling of diagnosis and progression, and comprehensive commitment to explainability.

Despite these merits, certain limitations remain. The use of a random forest, while interpretable, may not fully exploit the temporal and spatial complexities present in neuroimaging data. Additionally, while 11 modalities are integrated, the generalisability of the model and scalability in real-world clinical settings are not explicitly validated across independent cohorts or sites. Furthermore, potential biases introduced during feature selection and explainer design are not critically addressed. Nonetheless, the study represents a significant step toward clinically viable AI in AD, offering diagnostic precision and trust-enhancing interpretability. The approach aligns well with current calls for transparent, actionable, and patient-centred medical AI systems.

Chun et al. (2022) present an interpretable ML approach to predict the conversion of patients with aMCI to AD. The study addresses a clinically pressing need, as not all individuals with aMCI progress to AD, and accurate risk stratification could significantly enhance early intervention efforts. Traditional parametric models, such as logistic regression, often fall short in capturing complex, non-linear relationships among predictors; this research aimed to overcome such limitations by integrating modern ML algorithms with interpretability techniques.

The study prospectively analysed a cohort of 705 aMCI patients from the Samsung Medical Center, with a minimum of three years of follow-up data. Key features included neuropsychological test results and an apolipoprotein E (APOE) genotype. The dataset was split into a model-building set (n=565) and a validation set (n=140). Four algorithms were evaluated: logistic regression, random forest, support vector machine, and XGBoost. The XGBoost model achieved the highest performance with an AUC of 0.852 and an accuracy of 0.807. Crucially, the study enhances model transparency through global and local interpretability methods. SHAP and ICE were used to identify the most influential features per

patient. Key predictors included age, education level, memory and visuospatial scores, Clinical Dementia Rating (CDR) sum of boxes, MMSE, and APOE status.

Strengths of the study include the use of a relatively large, well-characterised prospective cohort, the combination of high-performance modelling with explainability tools, and the focus on individualised risk interpretation. However, the model relies primarily on neuropsychological and genetic data, excluding neuroimaging and biomarkers that may further enhance predictive accuracy. Additionally, external validation using diverse populations is lacking, which limits generalisability. Overall, the study demonstrates a practical and interpretable framework for predicting dementia conversion in aMCI patients. By balancing predictive power with clinical interpretability, the proposed model supports informed, patient-specific decision-making and offers a template for future applications in cognitive decline prediction.

De Santi et al. (2023) introduce a 3D CNN framework intended for the early diagnosis of AD through volumetric 18F-FDG PET scans. The model effectively addresses a notable challenge in neuroimaging-based AI diagnostics, specifically the lack of transparency inherent in black-box DL models. To enhance interpretability, the authors integrate two post hoc explanation techniques, Sensitivity Map (SM) and Layer-wise Relevance Propagation (LRP) to visualise the significance of various brain regions in the classification process.

The study uses a large dataset of 2552 PET scans sourced from the ADNI, representing three diagnostic classes: CN, MCI, and AD. A 3D CNN is trained for multiclass classification, with the model achieving Area Under the Curve (AUC) scores of 0.81 for CN, 0.63 for MCI, and 0.77 for AD. The relatively lower performance on MCI classification highlights the ongoing difficulty in detecting this transitional phase.

A significant strength of this study is the integration of explainability tools, particularly the application of LRP, which has been demonstrated to generate heatmaps with greater anatomical relevance than SM. The authors further enhance their analytical approach by aligning these heatmaps with the Talairach brain atlas, facilitating region-specific quantitative evaluations. Nevertheless, the study indicates an absence of a definitive correlation between the explanation provided by the heatmaps and the intensity of the PET signal. This observation suggests a possible disconnection between the attention mechanisms

of the model and its biological plausibility. Although the model exhibits performance comparable to that reported in the existing literature, several limitations are noted. The relatively modest AUC for MCI diminishes the clinical applicability of the model in the context of early intervention.

Furthermore, the absence of external validation utilising independent datasets restricts the generalisability of the findings. Although visually informative, the explainability methods lack clinical validation or expert review to verify their compatibility with established neuroanatomical biomarkers of AD. In conclusion, this study presents a methodologically rigorous and explainability-oriented approach to AD classification through PET imaging. It significantly contributes to initiatives promoting transparent AI in the context of AD research. Nevertheless, it would benefit from an extensive clinical evaluation, enhanced detection of prodromal stages, and a stronger connection between model outputs and biological interpretation.

Bogdanovic et al. (2022) present a comprehensive application of explainable ML to investigate AD, utilising an extensive dataset of over 12,000 individuals. In contrast to numerous studies focusing on prediction, this research emphasises the extraction of clinically significant insights and the validation of existing hypotheses concerning the risk and diagnosis of AD through the interpretability of models.

The dataset includes various features encompassing medical, cognitive, and lifestyle variables. The study applies a meticulous preprocessing pipeline, addressing missing data, feature redundancy, data imbalance, and inter-feature correlations. After this rigorous data preparation, the authors employ the XGBoost algorithm, a gradient-boosted decision tree ensemble known for its robustness and performance. The model achieves an F1-score of 0.84, placing it among the top-performing models in the domain. However, the authors frame this metric as secondary to their central aim: deriving interpretable, clinically actionable insights.

The SHAP framework generates both global and local interpretations of feature importance. Notably, the study presents a unified influence scheme that illustrates the directionality, positive or negative, of each significant effect of a feature on AD diagnosis. This scheme functions as evidence-based guidance for clinicians, potentially aiding in the interpretation of individual patient profiles.

A key strength of the study is its emphasis on hypothesis testing through interpretability rather than treating ML as purely predictive. The scale of the dataset also lends credibility to the derived conclusions. However, the study is limited by the absence of detailed information regarding the external validation of the model across distinct populations or clinical settings. Additionally, while SHAP enhances transparency, it remains sensitive to the training data and model structure, which may influence the consistency of the interpretations. The paper contributes meaningfully to explainable AI in NDD research. Prioritising insight over accuracy showcases a paradigm shift towards interpretable, hypothesis-driven ML applications. The results ensure to enhance early diagnosis and reshape how complex clinical data is utilised in uncovering patterns behind AD.

Varghese et al. (2023) propose a transparent diagnostic framework for AD classification using XAI. The primary goal is to bridge the gap between model performance and clinical interpretability by embedding explanation mechanisms into a non-linear neural network model, specifically focusing on early detection through MCI classification.

The study recognises a key barrier in AD diagnosis—delayed identification due to subtle early-stage symptoms and reliance on non-transparent, high-performing models that lack clinical trust. To address this, the authors develop an NN classifier that differentiates between demented and non-demented individuals. The novelty lies in enhancing this black-box model with local post hoc explanation techniques SHAP and LIME to transform it into a glass-box system. These XAI tools enable interpretability by highlighting feature contributions to individual predictions. Model evaluation indicates that CDR, age, and Atlas Scaling Factor (ASF) are strongly positively associated with dementia prediction, aligning with established clinical understanding. Conversely, features such as normalised Whole Brain Volume (nWBV), MMSE, and estimated Total Intracranial Volume (eTIV) contributed towards identifying non-demented individuals. The dual application of LIME and SHAP adds robustness by offering individual and global insights into model behaviour.

A key strength of this study is its focus on building clinician trust through interpretability without compromising classification accuracy. Including features clinically relevant to AD enhances the practical utility of the system. Additionally, two complementary XAI techniques provide a richer and reliable interpretative framework. However, whether the system was validated across external or independent cohorts remains unclear, which is crucial

for assessing generalisability. The study contributes to AD research by combining performance with interpretability. Incorporating XAI techniques fosters transparency, trust, and potential for clinical adoption in early-stage AD diagnosis and monitoring.

Alatrany et al. (2024) present an explainable ML approach tailored to address the challenges of AD classification, focusing on predictive performance and model interpretability. Recognising the complexity and high dimensionality of AD datasets, the study utilises data from the National Alzheimer's Coordinating Center, encompassing 169,408 records and 1024 features—a notably large dataset in the AD research domain.

The central aim is to enhance classification performance and extract interpretable rules to support clinical understanding. The researchers implement dimensionality reduction techniques to manage data complexity and employ SVM for classification tasks. SVMs are evaluated on external validation data and demonstrate strong performance, with an F1-score of 98.9% in binary classification (Normal Control vs AD) and 90.7% in multiclass settings. Additionally, the model predicts AD progression over a four-year period, achieving F1-scores of 88% (binary) and 72.8% (multiclass), highlighting its temporal predictive capability. To address the explainability challenge, the authors incorporate two rule-extraction methods, class Rule Mining and a Stable and Interpretable Rule Set approach. These generate transparent, human-readable decision rules, offering insight into the most influential features for classification. Key predictors identified include MEMORY, JUDGMENT, COMMUN (communication abilities), ORIENT (orientation), and the CDR score. These features were further validated using SHAP and LIME, ensuring consistency between rule-based and feature attribution perspectives.

A notable strength of the study is the integration of high-performing predictive models with interpretation mechanisms, offering both accuracy and clinical transparency. Using a large, real-world dataset enhances generalisability, while validation on external data sets supports robustness. However, while multiple explainability tools were used, the clinical utility of the extracted rules in a real-world diagnostic workflow remains underexplored. In conclusion, the study makes a meaningful contribution by combining high accuracy with interpretable outputs. It demonstrates how explainable ML can support early AD diagnosis and risk stratification in ways that align with clinical reasoning and evidence.

Amoroso et al. (2023) present a novel XAI approach for understanding how AD impacts brain connectivity, utilising both graph theory and interpretability methods. Their work addresses a central challenge in clinical AI applications: the black-box nature of many high-performing models, which impedes clinical adoption and trust. The study uses structural brain data from the ADNI, encompassing 432 T1-weighted MRI scans: 92 from AD patients, 126 from CN individuals, and 214 from those with MCI. Graph-based models are constructed to represent brain connectivity networks, enabling for topological analysis of structural brain changes across diagnostic groups.

ML models trained on these graph features successfully distinguish between AD, MCI, and NC groups. Crucially, the study integrates Shapley values to provide insight into the contribution of individual brain regions to the classification decisions. This enhances interpretability by quantifying the influence of specific nodes in the connectivity network. The interpretability results are biologically and clinically meaningful. The hippocampus and amygdala are shown to be highly relevant in AD classification—a finding that aligns well with established neurodegenerative patterns. For MCI subjects, the posterior cingulate and precuneus emerged as important, supporting the hypothesis that these regions are early markers of disruption. Interestingly, putamen and temporal gyri were highlighted as playing a role across the spectrum.

Strengths of the paper lie in its methodological innovation, combining graph theory with XAI tools, and its clinical relevance. By pinpointing disease-relevant regions in an interpretable way, the approach supports diagnosis, disease tracking, and intervention design. The study could explore whether such graph-based explainable frameworks can generalise across imaging modalities or cohorts. In conclusion, the paper contributes a compelling framework that links structural connectivity with explainability, bridging the gap between computational neuroscience and clinical neurology.

Raghupathy et al. (2025) present an ensemble-based ML approach combined with explainability techniques for the accurate diagnosis of AD. Their work emphasises the growing importance of XAI, particularly in clinical settings where model transparency is as essential as accuracy.

The study focuses on boosting ensemble classifiers, specifically XGBoost, LightGBM, and Gradient Boosting, which are known for their robustness and efficiency, particularly with structured data. A key contribution is the integration of SHAP into the modelling pipeline—not only to interpret the model outputs but also to guide feature selection. This dual role of SHAP enhances the transparency and the performance of the system. The authors report an accuracy exceeding 94%, achieved with a reduced set of features, demonstrating the effectiveness of using SHAP for dimensionality reduction without compromising performance. This study aligns with current trends in medical AI research, prioritising interpretable, high-performing models that can be used in real-world diagnostic settings. While boosting methods are already well-regarded for their predictive power, their black-box nature often limits their clinical use. By using SHAP, the authors enable local and global explanations, helping identify which features most influence individual and overall predictions.

Strengths of the paper include its focus on ensemble model robustness, efficiency with minimal feature sets, and the explicit use of XAI for explainability and trust-building. It contributes to the growing literature advocating for hybrid pipelines that merge model performance with interpretability. This work effectively demonstrates the power of combining boosting ensembles with SHAP for accurate and explainable AD classification, supporting clinical utility and trust in ML-based diagnostics.

(Jahan et al., 2023) propose an explainable ML framework for predicting and managing AD using a multimodal dataset. The paper responds to two key limitations in current AD prediction research: the over-reliance on neuroimaging alone and the lack of transparency in ML models that inhibits trust among end-users, particularly clinicians.

The authors use a data fusion strategy that combines clinical, MRI segmentation, and psychological data, enabling a holistic understanding of the disease. This multimodal integration is applied to a five-class classification task, distinguishing between AD, cognitively normal individuals, non-Alzheimer's dementia, uncertain dementia, and others. This goes beyond the common binary or three-class classification approaches in much of the literature. Nine machine learning models were evaluated, with RF emerging as the top performer. It achieved a cross-validated accuracy of 98.81%, suggesting strong discriminative capability and robustness. The choice of models, ranging from classifiers such as logistic regression and decision trees to complex ones such as MLP and ensemble techniques, adds credibility to the

comparative aspect of their methodology. Explainability is addressed through SHAP, which interprets model predictions and identifies influential features. A unique contribution of this study is also the inclusion of a proposed patient management architecture.

The main strength of the study lies in its novel use of multimodal data for five-way classification, which aligns well with the heterogeneity of dementia presentations in clinical reality. Using OASIS-3, a well-regarded open-access dataset, adds reproducibility value to the work. Additionally, incorporating explainability directly into the pipeline is a necessary step toward trusted, clinically applicable AI systems. The paper would benefit from discussion around the clinical validity of the most predictive features. Furthermore, while the proposed management architecture is a forward-thinking addition, its practical utility must be assessed in real-world deployment scenarios. In summary, this study makes a valuable contribution by demonstrating how integrating diverse data types and explainability methods can enhance performance and transparency in AD diagnosis and management.

P..A.Menon and R.Gunasundari (2024) presents an explainable ML framework for early AD classification that balances accuracy and interpretability. The key novelty lies in integrating SHAP for feature selection and model explainability and utilising PyCaret, a low-code automated ML tool, for rapid model evaluation and deployment.

The study uses the OASIS dataset and explores various classifiers for AD prediction. Among them, Naive Bayes achieves the highest classification accuracy of 96%. While this model is relatively basic, its performance suggests that with appropriate feature selection and preprocessing, even basic models can perform competitively. SHAP is used post hoc for interpretation and to identify and retain the most impactful features, effectively acting as a filter for feature importance. PyCaret simplifies model comparison and tuning, which may appeal to researchers or clinicians with limited coding experience. This low-code approach also aligns well with the push for democratising AI in healthcare. The use of SHAP adds transparency by showing which features drive predictions. However, there is limited discussion about class imbalance, model robustness, or external validation, which would be crucial for deployment in real-world settings.

In terms of contribution, this work shows the potential of combining automated ML tools with explainable AI to build interpretable and clinically relevant prediction models

efficiently. It demonstrates how methodological simplicity and strong interpretability coexist, particularly in early diagnostic tasks. discussion on the clinical meaning of selected features and a comparison with complex models, such as XGBoost or ensemble learners, would have been valuable to strengthen the study. Still, the approach provides a strong, practical foundation for building transparent AD classification systems that are easier to understand and implement.

2.3 Literature Review for Transfer Learning

Choe et al. (2020) aim to evaluate sub-scores from the MMSE to predict the progression from MCI to AD. The research used data from 306 people with MCI obtained from the ADNI database, including various standardised clinical and neuropsychological tests conducted at baseline and a two-year follow-up.

The researchers employed logistic regression analysis to investigate how MMSE total and subscale scores were related to the risk of developing AD. The analysed MMSE subscale scores comprised memory, orientation, construction, attention, and language. The research also accounted for possible factors, such as demographic and clinical variables, ensuring a robust data analysis. The results indicated a greater likelihood of developing AD, which was linked to decreased MMSE scores in memory, orientation, and construction subscales. In particular, the delayed memory recall section and the time aspect of the orientation section (specifically focusing on the week and day) were significant indicators of disease progression. However, the relationship between the attention and language subscales and AD conversion was not statistically significant.

This research utilises the MMSE cognitive evaluation tool and examines subscale results, guiding healthcare professionals. Using a large cohort from the ADNI database enhances the reliability and generalisability of the results. However, focusing on a single assessment tool may be a drawback since other biomarkers may be overlooked alongside neuroimaging data that could enhance predictive accuracy. Overall, the study underscores the significance of MMSE subscale scores, particularly in memory and orientation, as early indicators of AD progression in MCI patients. It advocates integrating accessible clinical evaluations into routine assessments to identify high-risk individuals.

Nanni et al. (2020) investigate the effectiveness of TL versus traditional models in diagnosing and predicting AD through structural MRI scans. The goal was to identify which method performed best in differentiating stages of cognitive decline. The study involved over 600 participants from the ADNI database, including individuals with AD, MCI converters to AD (MCIc), MCI non-converters (MCInc), and CN patients. Three methods were evaluated: an ensemble of five DL models fine-tuned for MRI tasks, training a 3D CNN model from scratch, and combining two conventional ML models with feature extraction and SVM. Performance was assessed in binary classifications: AD vs. CN, MCIc vs. CN, and MCIc vs. MCInc.

The ensemble TL model achieved an AUC of 90.2% for AD vs CN, 83.2% for MCIc vs CN, and 70.6% for MCIc vs MCInc. Traditional ML techniques outperformed TL in AD vs CN and MCI vs CN comparisons, with AUCs of 93.1% and 89.6%. However, MCIc and MCInc results varied, with AUCs from 69.1% to 73.3%. The CNN trained from scratch underperformed due to the small dataset. Using an ensemble of pre-trained models is a notable innovation in TL. A limitation is the small dataset, which may hinder DL model performance, particularly for CNNs that require large data volumes

This research highlights TL in medical diagnosis without extensive labelled datasets. Even with training on non-medical images, TL models achieved impressive results, indicating room for further investigation. The findings show that traditional ML techniques, combined with careful feature development, can still compete with or surpass DL methods, emphasising the ongoing relevance of classic approaches in medical contexts. The study provides insights into the effectiveness of TL compared to traditional methods for early detection and prediction of AD, stressing the importance of ensemble methods and interdisciplinary strategies in advancing medical diagnostics.

Jha and Kwon (2017) introduce a technique using sparse autoencoders (SAE) to identify AD in its early stages. This technique combines scale conjugate gradient (SCG) optimisation with a SoftMax output layer for patient categorisation. The main objective of this study was to develop an accurate and efficient algorithm for distinguishing AD patients from individuals with normal cognitive function.

The research utilised OASIS neuroimaging data, employing a sparse autoencoder to extract key input features. This was combined with SCG, an optimisation algorithm that

efficiently fine-tuned NN by minimising loss functions than traditional methods. The model included a stacked autoencoder with a SoftMax layer for classification, converting outputs into probability distributions. It was refined to enhance accuracy, sensitivity, and specificity, addressing overfitting and feature redundancy. The model achieved 91.6% accuracy, 98.09% sensitivity, and 84.09% specificity, demonstrating reliable detection of both positive and negative AD cases for early diagnosis.

Using sparse autoencoders helps the model focus on the most important characteristics, decreasing the likelihood of overfitting. The SCG optimisation enhances the ability of the model to learn from the data. In general, this research provides significant knowledge on employing autoencoders for detecting AD, particularly through the unique integration of sparse autoencoders and SCG.

Bhatkoti and Paul (2016) present a novel DL method for AD detection that uses brain MRI scan data, CSF, and PET images. They develop a framework that employs a modified k-sparse autoencoder and a multi-class classification model. The model distinguishes between various stages of AD, including MCI and advanced AD.

The approach uses a k-sparse autoencoder to enhance feature extraction from input MRI data. This model is combined with a DL classifier for multi-class classification. The autoencoder incorporates a sparsity constraint, activating only a limited number of neurones and improving feature learning. The classifier uses these features to distinguish between healthy, MCI, and AD phases, significantly improving accuracy. The research highlights that the k-sparse autoencoder boosts model resilience and precision, which is crucial for early-stage MCI detection and vital for prompt intervention.

The main advantage is its use of the k-sparse autoencoder, improving feature extraction and classification accuracy. Multiple classes provide a breakdown of Alzheimer's stages, enhancing diagnostics with nuanced information. However, relying on a single data set may limit the applicability of the results. Additionally, significant modifications and computational resources are required due to the complexity of the model, including the sparsity constraint. The study offers a promising approach for detecting early Alzheimer's through advanced ML methods and neuroimaging data.

Mehmood et al. (2021) explore a new method to enhance early AD detection by applying TL methods to MRI images from ADNI. This approach focuses on using CNNs pre-trained on large image datasets, specifically the VGG19 architecture, fine-tuning them for MRI scans of AD patients to differentiate stages of the disease. Data augmentation was essential in artificially expanding the training set, preventing overfitting and improving model generalisability. The study also incorporated batch normalisation and dropout layers for enhanced performance, adjusting only the final layers of pre-trained networks for the AD classification task while retaining the general features from earlier layers.

The TL models accurately distinguished normal controls from MCI and AD patients. The VGG-based model outperformed traditional techniques reliant on manually selected features regarding accuracy. This suggests that pre-trained CNNs, tailored for specific medical imaging, can greatly enhance diagnostic accuracy. The innovative use of TL in addressing complex medical issues highlights the promise of DL in healthcare. However, the research is limited to one imaging technique (MRI) and lacks additional data such as genetic information or clinical evaluations for a comprehensive diagnostic approach. However, it significantly contributes to AD diagnosis, showcasing the effectiveness of advanced ML methods in medical imaging.

C. Wu et al. (2018) aim to develop a CNN model for accurate classification of MCI and prediction of its progression to AD, addressing the need for early and reliable diagnostic tools. The study employed MRI data from the ADNI dataset to differentiate between MCI and normal cognitive function and predict the progression from MCI to AD. The dataset included structural MRI and clinical information, ensuring a comprehensive approach to training and validating the model.

The research used a 3D CNN model to analyse MRI data, extracting spatial features to recognise patterns linked to cognitive states. The model integrated TL and data augmentation methods to enhance effectiveness and generalisation. It also explored various DL structures and optimisation techniques to enhance accuracy in classifying MCI and predicting progression to AD. The CNN model achieved high accuracy in both classification and prediction, indicating its potential to predict the transition from MCI to AD and providing insights into cognitive decline progression.

The strength is its advanced DL structure that captures specific MRI spatial details. Using TL enables the model to utilise pre-trained networks, enhancing accuracy and robustness. However, a limitation is its reliance on a single imaging modality, potentially overlooking important biomarkers for AD progression. The research contributes by applying CNNs to MCI classification and conversion prediction tasks. While results show promise, dependence on MRI data limits usefulness in contexts where this imaging is not feasible.

Spasov et al. (2019) seek to develop a DL model that is efficient in terms of parameters to predict the transition of individuals from MCI to AD. The study employed data from the ADNI dataset, including MRI images, demographic data, genetic details, and cognitive evaluations. The dataset contained stable MCI patients as well as individuals who progressed to AD during a specified follow-up period.

The study employed an efficient CNN method that reduced parameters while maintaining accuracy. It introduced a new 3D CNN model integrating spatial and temporal data, addressing high computational demands. This structure effectively captured spatial aspects of brain atrophy related to AD progression. A feature extraction technique identified the most predictive biomarkers, enabling accurate MCI to AD transition predictions, achieving 86% accuracy. The model effectively distinguishes between sMCI and pMCI, serving as a valuable tool for early detection.

This research emphasises an efficient model, enhancing applicability in resource-limited clinical settings. Combining multi-modal data with advanced feature extraction boosts robustness and predictive power. However, its reliance on a single dataset (ADNI) limits generalisability across diverse populations. Although the model demonstrates a reasonable level of accuracy, there exists significant potential for enhancement, particularly regarding the specificity of predictions. This work significantly advances neuroimaging and predictive modelling for AD, addressing common challenges in deep learning model deployment by utilising a model with fewer parameters.

Fouladvand et al. (2019) focused on creating a DL algorithm for forecasting the progression from MCI to AD using the Mayo Clinic Study on Ageing (MCSA). The emphasis was on utilising longitudinal data to enhance predictive accuracy, enabling earlier and precise interventions for high-risk patients. The dataset contained 558 electronic health records (EHR)

of individuals with MCI, including details on personal information, clinical notes, diagnoses, lab results, medications, and cognitive scores.

The research used a DL system that integrated various EHR data. It focused on temporal aspects, recognising that changes in clinical and cognitive markers over time are crucial for forecasting disease progression. The RNN architecture used was the LSTM network, which handled the sequential nature of EHR data. The LSTM was trained to predict the progression of MCI to AD in patients over a period of time. The results were promising, demonstrating a robust predictive capability for identifying patients with MCI who are likely to progress to AD. The study found LSTMs enhanced to random forests in F1 scores and compared the DL approach with traditional ML methods. The DL model excelled in addressing the complexity and temporal dynamics of the data.

One of the primary strengths of this research is its detailed long-term dataset from EHR, providing a complete view of patient well-being over time. LSTM networks enable the model to capture essential temporal dependencies for forecasting disease progression. Additionally, integrating various EHR data types, such as clinical notes and cognitive scores, offers a comprehensive approach. However, the performance of the model relies on the quality of EHR data, which can vary. Incorporating clinical biomarkers could enhance predictive accuracy. This study represents a significant advancement in using DL to forecast progression from MCI to AD, highlighting EHR data in clinical prediction models for early detection in a scalable, non-intrusive manner.

D. Zhang and Shen (2012) seek to enhance the forecasting of AD progression among individuals with MCI. The authors employ longitudinal data, various biomarkers, such as MRI and FDG-PET imaging, and clinical scores such as MMSE and ADAS-Cog. The research aims to predict both qualitative changes (such as transitioning from MCI to AD) and quantitative changes (such as fluctuations in cognitive scores) in MCI patients over a period of time. Accurately predicting these changes is essential for the early diagnosis and tracking of AD progression. The researchers applied data from the ADNI, which involved 88 MCI participants monitored at various intervals.

The approach included conducting a longitudinal selection process to identify relevant and important brain regions over time for each type of modality. This was

accomplished using a sparse linear regression model that incorporated 'group regularisation' to group the weights related to the same brain region over multiple time points. This method identifies brain areas based on cumulative evidence from multiple time points. The longitudinal features obtained from the selected areas were then combined with a multi-kernel SVM to predict future clinical outcomes. The model demonstrated results with 78.4% accuracy compared to traditional techniques. In particular, the research accurately predicted cognitive scores (MMSE and ADAS-Cog) after 24 months and the progression from MCI to AD with high accuracy, employing data from at least half a year prior to the progression from MCI to AD.

The main advantage of this research is its comprehensive method, which employs various types of data and multiple time points to enhance predictive accuracy. Nevertheless, the complexity of the model and the requirement for extensive longitudinal data may restrict its practical application in clinical settings, where obtaining such detailed data may not always be feasible. The research effectively showcases the potential of combining advanced ML methods with multimodal biomarkers for predicting AD outcomes. However, further studies must focus on making these models straightforward for broader clinical applications and ensuring their effectiveness in various patient demographics.

Oh et al. (2019) investigate innovative methods for diagnosing AD using MRI data. The research aims to enhance the accuracy and interpretability of AD diagnosis by utilising a convolutional NN (CNN) trained on volumetric data in conjunction with transfer learning. This method uses pre-trained DL models adjusted to a dataset to identify stages such as AD, progressive pMCI, stable sMCI, and NC.

The research employs VCNNs and convolutional autoencoders (CAE) to analyse MRI data, offering a comprehensive analysis than standard 2D approaches. Pre-trained models were modified using TL and an inception module-based CAE to reduce the specialised training data required for AD. The models receive MRI data from ADNI to ensure robust training. The combination of VCNN and TL methods successfully attained high accuracy, varying from 60% to 86% in distinguishing various presentations of diseases. The model captured subtle characteristics of disease progression by utilising volumetrically labelled data. The research also highlighted the ability of the model to visualise essential features that impact classification, improving the understandability of the outcomes.

The main advantage of this research is its utilisation of volumetrically labelled data, which gives an extensive dataset for analysis and enhances diagnostic accuracy. Moreover, TL lowers the requirement for extensive, disease-specific datasets, resulting in efficient and accessible model training. One possible drawback is the dependency on TL, which could result in biases from the pre-trained models if they do not align well with the specific features of AD. The research significantly impacts the field by merging sophisticated ML methods with medical imaging. It emphasises the promise of VCNNs, CAEs, and TL in improving AD diagnosis, but research is needed to perfect these techniques and ensure their effectiveness in various settings.

Aderghal et al. (2020) address the classification of AD stages by utilising different MRI modalities in conjunction with TL methods. The primary focus of this study was to enhance the accuracy of classifying Alzheimer's stages (normal cognition, MCI, and AD) using DL models. The research aimed to enhance the accuracy of diagnosis and early detection of AD progression by analysing MRI data sourced from the ADNI database. The dataset consisted of 306 individuals, 133 classified as having MCI, 58 with AD, and 115 as normal controls. The MRI images were segmented before being used to train and evaluate the models.

The research utilised TL, specifically employing pre-trained LeNet-like CNN architectures trained on MNIST. Features were derived from the MRI and DTI data, which were specifically created for classifying multiple stages of Alzheimer's. The research investigated different DL architectures, utilising methods such as weighted cross-entropy loss to tackle imbalanced class problems. The results showed that the fusion method produced the highest performance, showing varying accuracy rates in different projections but demonstrating strong classification abilities.

One key advantage of this study is its all-encompassing strategy, which combines various MRI methods and utilises advanced DL methods to address the complex issue of classifying Alzheimer's Disease. Utilising TL enabled efficient management of inadequately labelled data by capitalising on insights from similar or MNIST data. Nevertheless, the research encountered obstacles such as addressing class disparities and the risk of overfitting because of the complex model structures employed. Moreover, although the results showed potential, validation is needed to determine if the findings can be applied to different datasets or clinical contexts. The research shows promising possibilities for utilising TL to detect Alzheimer's

cognitive stages early and accurately, providing a strong framework that can be enhanced and evaluated in clinical settings.

Mehmood et al. (2024) aim to enhance the diagnostic accuracy of AD across various stages by utilising advanced DL methods such as TL and CNNs. The main emphasis is on utilising the Siamese NN structure, particularly the 4D-AlzNet model, which consists of four parallel CNNs, to analyse MRI data. This method is significant as it looks at both spatial features and temporal variations, rigorously examining the structural changes to the brain as AD progresses.

The research utilises TL, where the model is pre-trained on VGG-19, VGG-16, and customised AlexNet, which is trained on a comprehensive dataset. The model is then further trained on a task-related dataset, specifically the ADNI dataset. This technique enhances the capability of the model to identify minor distinctions in MRI images that suggest different stages of AD. The Siamese network excels at comparing pairs of images, which is essential for differentiating between the early and late stages of MCI and AD. The results are promising, as the Siamese 4D-AlzNet demonstrates a high accuracy of 95.05% in distinguishing AD stages. This is particularly important for quickly detecting and tracking the development of diseases, as this is vital for prompt intervention. Utilising four data types in the model introduces a new aspect to the analysis, which may result in precise forecasts than conventional 2D or 3D imaging methods.

The strength of the research lies in its creative utilisation of sophisticated NN structures and TL, which collectively enhance the capacity of the model to generalise and accurately classify complex data. However, the research also has limitations, such as the computational resources needed to develop and use these complex models, potentially hindering their use in real-world clinical settings. Moreover, despite the high accuracy of the model, additional validation on various populations is necessary to verify its overall applicability. This study makes a substantial contribution to the neuroimaging field and the diagnosis of NDD by introducing an innovative method that integrates sophisticated DL strategies with extensive, multiple types of data. This may lead to accurate and quicker AD detection, possibly enhancing patient results with earlier and focused treatments.

Qiu et al. (2018) explore the combined use of multiple diagnostic modalities to enhance the detection of MCI. The research aims to determine whether combining MRI scans, MMSE, and logical memory (LM) tests can enhance the accuracy of diagnosing MCI, a precursor to AD.

The data for DL models were MRI data from the National Alzheimer's Coordinating Centre database, consisting of 386 individuals with normal cognition, or MCI. The research methodology fine-tuned the VGG-11 model, which was pre-trained on large datasets, for classifying cognitive status for MRI scans. Modifications involved batch normalisation, dropout layers, and additional fully connected layers. The research compared the accuracy of various models in identifying MCI through different data types. The MRI model reached an accuracy of 83.1%, the MMSE model 84.3%, and the LM model 89.1%. The accuracy was greatly enhanced to 90.9% by blending these models using a majority voting technique. Combining various data sources using a multimodal strategy enhances the dependability of MCI identification, providing accuracy. Various predictions from different sources, such as MRI, MMSE, and LM tests, were integrated through majority voting to make the final model and diagnosis. This method of multimodal fusion sought to capitalise on the advantages of each data type to enhance diagnostic robustness. This suggests that merging different data sources can greatly enhance diagnostic accuracy compared to using just one data type.

The strength of the research is in its original utilisation of diverse multimodal and comprehensive DL methods, offering a rigorous approach to diagnosing MCI. However, the research is constrained by its retrospective design and the risk of overfitting because of the relatively small sample size. Furthermore, the model has not been tested for its generalisability to other populations, potentially restricting its broader applicability. This research contributes substantially to the field by showing how combining various neuroimaging and neuropsychological tests with DL techniques can lead to a promising method for detecting cognitive impairments at an early stage.

Duc et al. (2020) investigate a novel method for AD diagnosis and MMSE score prediction using resting-state functional MRI data. The main goal of this research is twofold: first, to create a DL model to classify AD, and second, to predict cognitive impairment levels based on MMSE scores.

The research used rs-fMRI data from 331 participants in South Korea, consisting of individuals with AD and those who were healthy. The researchers obtained 3D independent component spatial maps from fMRI scans and used them as features in a 3D CNN for the classification task. Multiple regression models, such as linear least squares regression (LLSR), support vector regression, and ensemble techniques, were evaluated for MMSE score prediction. Techniques such as LASSO and SVM-RFE were used for feature optimisation. The findings were encouraging, as the CNN obtained an average balanced accuracy of 85.27% in differentiating between AD patients and healthy controls. Moreover, the research found that networks such as the medial visual, default mode, dorsal attention, executive, and auditory-related networks strongly correlate with AD. The best results for MMSE score prediction were achieved by combining gICA features with SVM-RFE, resulting in an R square value of 0.63 and an RMSE of 3.27.

The strengths of the study are its comprehensive feature extraction from rs-fMRI data and the use of state-of-the-art DL techniques, which increase the accuracy of AD detection and cognitive decline prediction. Nevertheless, the research is constrained by its concentration on a particular group, potentially impacting the applicability of the findings. Moreover, although helpful, rs-fMRI can be susceptible to motion artefacts and other factors that may affect the reliability of the results. The paper greatly impacts the field by showing how combining neuroimaging data with DL can be valuable for diagnosing NDDs and MMSE scores jointly. Still, testing with diverse populations is needed to verify its broader applicability.

Gao et al., (2020) propose an innovative approach that enhances the prediction accuracy for converting MCI to AD. The study aims to address a critical challenge in NDD research by identifying which MCI patients are at higher risk of progressing to AD. This prediction is vital for early intervention and improving patient outcomes. The research uses a dataset from the ADNI, including neuroimaging data, demographic information, and cognitive assessments, to build a robust predictive model.

The authors introduce a novel DL model named AD-NET (age-adjust NN), which utilises TL to maximise the utility of limited medical imaging data. The model uniquely incorporates an age-adjusted component, recognising age as a significant factor in the progression to AD. This is achieved by transferring knowledge from a pre-trained model

trained on healthy subjects to the AD-NET for feature extraction and utilising age-related information as a surrogate biomarker. The dual purpose of the TL approach in AD-NET sets it apart from other methods, enhancing the ability of the model to predict conversion accurately across different age groups. The experimental results demonstrate that AD-NET significantly outperforms eight other classification models in predicting the conversion from MCI to AD, particularly highlighting its effectiveness in young cohorts. The performance of the model was validated using metrics such as accuracy and AUC, which gave rise to enhanced results. This success underscores the capability of the model to integrate both feature extraction and demographic information effectively.

However, the research does have its limitations. One significant drawback is the lack of generalisability of the model, as it was only trained on a particular dataset. This could restrict its applicability to a broader range of diseases and purposes. Moreover, the interpretability of the NN is also a challenge due to its complexity, which is vital for clinical use and decision-making. In conclusion, AD-NET has made significant advancements in neuroimaging and Alzheimer's research, particularly in addressing data scarcity and the role of age as a predictive factor.

2.4 Comprehensive Survey of Explainability and Interpretability Techniques

2.4.1 Brief overview of XAI and explainability in ML/AI

Over the past decade, the rapid expansion of AI has led to significant improvement in predictive accuracy, optimisation capabilities, and real-world deployment. State-of-the-art models, particularly DNNs, ensemble learning systems, and complex generative architectures, have demonstrated exceptional performance across a diverse range of applications, such as computer vision (Krizhevsky et al., 2012), natural language processing (Vaswani et al., 2017), healthcare diagnostics (Esteva et al., 2017), finance (Heaton et al., 2017; Lipton, 2018), and autonomous systems (Bojarski et al., 2017). However, this increase in model complexity has also resulted in significant challenges related to interpretability, transparency, and accountability.

Frequently, these complex models are referred to as “black boxes” (Lipton, 2018), indicating that their internal decision-making processes are either inaccessible or

incomprehensible to users. As AI systems increasingly influence critical decisions with social, ethical, and legal consequences, the demand for explainability of the reasoning behind decisions, also known as XAI, has grown substantially. Explainability refers to the extent to which an end user can understand, trust, and verify the output of an AI system. While traditional, classic models, such as decision trees or linear regression, naturally provide interpretable structures, most modern ML algorithms sacrifice explainability in pursuit of improved predictive performance.

The field of XAI has thus emerged to address the trade-off between explainability and performance. It encompasses a diverse set of methods designed to generate human-comprehensible explanations for complex model outputs. This need for explainability is driven by multiple stakeholders, including domain experts seeking to validate AI recommendations, developers aiming to debug and improve models, regulators requiring transparency for compliance, and end-users who need to trust system outputs in high-stakes settings (Doshi-Velez and Kim, 2017; Rudin, 2019). Regulatory pressures such as the European Union's General Data Protection Regulation (GDPR) (Goodman and Flaxman, 2017) and the EU AI Act have further accelerated interest in developing explainable AI systems.

Furthermore, explainability is increasingly linked with other critical dimensions of responsible AI, including fairness, bias mitigation, robustness, and trustworthiness (Gilpin et al., 2018; Mittelstadt, 2019). Explanations facilitate the identification of biased correlations, highlight spurious features, and reveal vulnerabilities to adversarial examples. As a result, explainability not only increases transparency but also acts as a vital diagnostic tool for enhancing model integrity.

Despite the growing emphasis on research in this area, a universally accepted definition or standardised approach to achieving explainability in AI remains elusive. Instead, a variety of techniques have been developed, encompassing both inherently interpretable models and post-hoc explanation frameworks (Carvalho et al., 2019; Molnar et al., 2020). These methodologies vary in several aspects, such as timing, scope, dependency on model architecture, and the underlying mechanisms employed. Each approach is characterised by distinct strengths, limitations, and suitability, which are contingent upon the specific domain and use case.

2.4.2 Key questions addressed by the literature review

This survey seeks to systematically explore the landscape of explainable AI by addressing several interrelated research questions that have emerged within the field.

a) What are the major categories and taxonomies used to classify explainability methods in AI?

The review aims to synthesise the diverse taxonomic frameworks that have been proposed, considering multiple dimensions such as timing of explanation, scope, model dependence, and technique type.

b) Which specific methods and algorithms are currently most influential for generating explanations?

This encompasses both classical and contemporary methodologies, such as decision trees, rule lists (Ustun and Rudin, 2016), surrogate models (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), gradient-based approaches (Selvaraju et al., 2020; Simonyan et al., 2014), example-based strategies (Kim et al., 2014; Koh and Liang, 2017), among others.

c) How do different explanation techniques perform in terms of faithfulness, fidelity, stability, and human interpretability?

The review will explore the strengths and weaknesses of competing methods across these evaluation dimensions, drawing on comparative studies and benchmarking efforts (Doshi-Velez and Kim, 2017; Vilone and Longo, 2021).

d) What are the emerging research challenges and future directions in XAI?

The review will highlight unresolved problems, including the absence of standardised benchmarks, the risk of misleading or incomplete explanations, the interaction between explainability and fairness, and the prospects for causal and interactive explainability frameworks (Ghorbani et al., 2019; Lipton, 2018; Rudin, 2019).

Through the systematic examination of these inquiries, this literature review aims to provide a comprehensive and contemporary synthesis of the field, presenting both a broad overview and a critical analysis of XAI methodologies.

2.5 Conceptual Foundations

2.5.1 Definitions and Terminology

In the XAI literature, the terms "explainability" and "interpretability" are frequently utilised interchangeably; however, nuanced distinctions between them have been proposed. Lipton (2018) asserts that interpretability pertains to the extent to which an individual can comprehend the internal mechanisms of a system without the aid of external tools, while explainability refers to the degree to which a system can generate external artefacts or reasoning to substantiate its decisions. Interpretability typically refers to the inherent transparency of a model, such as linear regression or decision trees. In contrast, explainability may encompass post-hoc techniques applied to models that are otherwise opaque.

Doshi-Velez and Kim (2017) argue that interpretability is a sub-component of explainability, where explanations should be comprehensible to humans and should support specific goals, such as debugging, trust-building, or regulatory compliance. Rudin (2019) adopts a radical stance, asserting that inherently interpretable models should be prioritised over black-box models, as post-hoc explanations may prove to be approximate and potentially misleading. Despite the lack of consensus, it is broadly accepted that both interpretability and explainability serve the common purpose of making AI systems transparent, trustworthy, and aligned with human understanding (Carvalho et al., 2019; Miller, 2019).

In addition to interpretability and explainability, several other related terms frequently appear in XAI discussions, such as Transparency, which refers to the visibility of the internal structure and functioning of an AI model. Highly transparent models are naturally interpretable. Trust is another term that relates to confidence in predictions of the system and its willingness to rely on its outputs (Gunning et al., 2019). Trust may not directly correlate with technical interpretability, as humans may trust systems for reasons unrelated to their true reliability (Miller, 2019). Another term, Causality, involves understanding not only correlations but also the underlying causal mechanisms that drive predictions (Pearl, 2009). Causal explanations are often considered robust, as they reflect actual data-generating processes rather than superficial patterns. The growing attention to these concepts reflects the multidisciplinary nature of XAI, which draws from ML, human-computer interaction, cognitive psychology, philosophy, and law (Mittelstadt, 2019).

2.5.2 Importance of Explainability

The necessity for explainability in AI systems emanates from various, frequently intersecting dimensions: ethical, technical, regulatory, and domain-specific. One of the foremost ethical arguments for explainability focuses on accountability. When AI systems are engaged in high-stakes decisions, such as medical diagnoses, loan approvals, or legal sentencing, it is essential that affected individuals and decision-makers possess the capability to comprehend and contest the rationale underpinning those decisions (Wachter et al., 2017). In the absence of sufficient explanations, individuals are deprived of the opportunity for recourse or informed consent, thus raising significant ethical concerns.

From a technical perspective, the concept of explainability significantly contributes to the processes of model debugging, validation, and enhancement. Explanations serve as valuable tools for researchers, enabling the identification of issues such as data leakage, spurious correlations, and overfitting (Hooker et al., 2019; A. S. Ross et al., 2017). Furthermore, they facilitate feature engineering, enabling practitioners to identify critical variables and their interrelationships. In addition, explanations have the potential to reveal vulnerabilities associated with adversarial examples and failures in robustness (Ghorbani et al., 2019).

From a regulatory perspective, emerging legal frameworks have introduced formal obligations of explainability. The GDPR policy of the EU includes the "right to explanation" (Goodman and Flaxman, 2017). The proposed EU AI Act suggests stringent requirements concerning transparency, risk management, and accountability in AI systems, particularly those classified as high-risk.

Explainability is also closely linked to fairness and the mitigation of bias. Transparent explanations can help identify and rectify systematic biases against particular groups, ensuring that AI systems do not perpetuate or amplify existing inequalities (Barocas et al., 2021; Mehrabi et al., 2021). Consequently, numerous scholars regard explainability as a fundamental element of comprehensive frameworks for responsible and trustworthy AI (Cowls et al., 2019; Jobin et al., 2019).

While explainability holds universal significance, its importance becomes particularly pronounced within certain application domains. In the realm of clinical decision-making, both

physicians and patients must possess the capability to comprehend and validate AI-generated recommendations. In the absence of interpretability, the level of trust in AI-assisted diagnostics and treatments remains considerably restricted. (Caruana et al., 2015; Holzinger et al., 2017). For instance, models that predict disease risk must provide clear rationales based on medically meaningful features to support clinical adoption (Tonekaboni et al., 2019).

In the financial services industry, AI models are utilised for credit scoring, fraud detection, and informed investment decisions (Ryman-Tubb et al., 2018). Regulatory bodies often require clear documentation of model decisions to ensure fairness, prevent discrimination, and maintain market integrity. In legal and judicial systems, predictive models employed within the realm of criminal justice necessitate transparency to prevent opaque decision-making that could unjustly impact individuals (Dressel and Farid, 2018; Surden, 2021). In safety-critical applications such as autonomous vehicles and robotics, real-time explanations can aid system monitoring, safety validation, and post-incident analysis (Amodei et al., 2016).

In these domains, the absence of explainability can significantly impede adoption, diminish trust, and heighten public apprehensions regarding the implementation of AI. Consequently, explainability is progressively regarded not merely as a desirable attribute but as an essential requirement for the ethical utilisation of AI.

2.6 Taxonomy of Explainability Techniques

As the field of XAI continues to evolve, various frameworks have been proposed to categorise explanation methods according to their objectives, characteristics, and underlying mechanisms. A systematic taxonomy facilitates the organisation of this expanding field of research and assists in the selection of appropriate techniques, tailored to specific use cases, model types, and interpretability requirements. This section presents a structured taxonomy of XAI techniques, organised along four core dimensions: timing of explanation, scope, model dependency, and technique type.

2.6.1 By Time of Explanation

2.6.1.1 Intrinsic Interpretability

Intrinsic interpretability pertains to models that are inherently transparent by design. Such models enable direct human comprehension of their internal logic and decision-making processes, thereby avoiding the necessity for post-hoc analysis. Decision Trees (L. Breiman et al., 2017; Quinlan, 2014), Rule-Based Systems (Rivest, 1987; Ustun and Rudin, 2016), and Generalised Additive Models (GAMs) (Caruana et al., 2015; Hastie and Tibshirani, 1986) serve as exemplars of this methodology. Each mechanism produces comprehensible outputs, whether through decision-making pathways, established rule sets, or the effects of additive features, thus enabling immediate analysis.

These models provide high interpretability and a minimal cognitive burden, particularly in domains that require transparency, such as finance, healthcare, and law. However, their expressiveness is limited. They frequently underperform on high-dimensional, unstructured, or highly non-linear data, where model complexity may be essential for achieving predictive accuracy. Nonetheless, their alignment with human reasoning continues to sustain their relevance in safety-critical and regulatory settings.

2.6.1.2 Post-hoc Explainability

Post-hoc explainability encompasses techniques applied after model training to interpret otherwise opaque black-box systems. These methods enable interpretability without modifying the underlying model and are particularly crucial for explaining DNN, ensemble methods, and kernel methods. While highly versatile, these techniques are often subject to trade-offs between fidelity and interpretability. A wide range of tools fall under this category:

SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) is a unified framework for interpreting model predictions based on Shapley values from cooperative game theory. Each feature is a “player” in a coalition, with the prediction of the model as the “payout” distributed based on their contribution. SHAP attributes feature importance by computing the marginal contribution of each feature across all possible subsets of features. Unlike heuristic-based methods, SHAP provides strong theoretical guarantees, satisfying properties such as local accuracy, consistency, and robustness against missing data. Its formal

mathematical grounding has made it a prominent tool in the XAI landscape, particularly in high-stakes domains where interpretability must be both rigorous and actionable.

Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016) is a flexible, post-hoc technique designed to enhance transparency in black-box models. LIME generates perturbed samples around a prediction instance and fits an interpretable model, usually linear, to approximate the complex decision in a local neighbourhood. The strength of LIME lies in its model-agnostic nature, enabling it to be applied across various domains, including image classification, text processing, and tabular data. Its early popularity was driven by its intuitive conceptual framework and ease of integration with any classifier, making it a cornerstone technique in the formative years of XAI.

Saliency-based methods (Simonyan et al., 2014) identify input regions that influence outputs, primarily in image tasks. Despite being computationally efficient, their reliability has been scrutinised, suggesting that explanations may not always accurately represent the actual decision-making process (Adebayo et al., 2018).

Grad-CAM (Selvaraju et al., 2020) enhances saliency maps for CNNs by incorporating intermediate activations to produce spatially coherent heatmaps. It finds extensive application in vision tasks; however, it demonstrates limitations in generalisability beyond CNNs.

Counterfactual explanations (Wachter et al., 2017) generate minimal alterations to input variables that influence predictions, thereby addressing inquiries of a hypothetical nature. Their inherent actionability renders them suitable for domains that necessitate user-centred interpretability; however, the issue of feasibility in high-dimensional spaces remains an unresolved matter.

While post-hoc methods expand interpretability across diverse domains and model types, concerns persist regarding their faithfulness and robustness. Krishna et al. (2025) argue that post-hoc explanation methods, such as SHAP and LIME, are prone to spurious feature attribution, wherein irrelevant features are assigned elevated significance due to correlated noise or dataset artefacts. This predicament is exacerbated when models are trained on biased or imbalanced data, resulting in misleading explanations that mirror data artefacts rather than the actual reasoning of the model. They assert that this undermines their

applicability in fairness-sensitive contexts and promotes a transition towards intrinsically interpretable models. These limitations highlight a persistent challenge in post-hoc XAI: the necessity to produce reliable, stable, and truly reflective explanations of model behaviour.

2.6.2 By Scope

2.6.2.1 Global Explanations

Global explanations aim to elucidate the behaviour of the model across the whole input space, providing a macroscopic perspective on how features impact predictions on average. These methodologies facilitate the identification of overarching trends and feature significance throughout the dataset; however, they may obscure heterogeneity in localised contexts.

Partial Dependence Plots (PDPs) (Friedman, 2001; Greenwell et al., 2018) serve to estimate marginal effects by averaging model predictions across the distribution of all other features while systematically varying one or more target features. This methodology aids in visualising overarching relationships, such as monotonicity or threshold effects. Nevertheless, PDPs are predicated on the assumption of feature independence, which often fails in real-world datasets, thereby leading to potentially misleading conclusions when significant feature interactions or correlations are present (Apley and Zhu, 2020).

Feature importance scores, such as those derived from permutation tests (Breiman, 2001), quantify the degree to which model accuracy diminishes when feature values are randomly shuffled. Although beneficial for ranking features, these scores provide no directional insight and may prove unreliable in the context of multicollinearity (Molnar et al., 2020).

SHAP, previously introduced, also supports global interpretability by aggregating local Shapley values across the dataset. This yields a consistent global importance measure that reflects both the direction and magnitude of feature effects. However, the aggregation process may obscure local nuances, and additive assumption of SHAP remains a limiting factor in highly non-linear or interaction-heavy models (Kumar et al., 2020).

2.6.2.2 Local Explanations

In contrast, local explanations focus on individual predictions, revealing which features influenced a specific output. This degree of granularity holds particular significance in domains where justifications at the individual instance level are essential, notably within clinical or legal contexts.

Local LIME provides local explanations by training a surrogate model around a given input instance. Its value lies in offering case-specific rationales; however, concerns remain regarding the stability and faithfulness of its approximations in non-linear regions (Alvarez-Melis and Jaakkola, 2018).

Similarly, local SHAP values represent the contribution of each feature to a single prediction based on cooperative game theory. They offer theoretically grounded, instance-level attribution. As previously indicated, these values are computationally intensive and are constrained by the assumptions of the additive model, which may overly simplify interactions (Frye et al., 2021).

While global methodologies contribute to the comprehension of overarching model trends and the significance of features, local methodologies complement these by offering actionable insights at the individual level. Collectively, they establish a dual perspective for interpretability: global explanations facilitate transparency, whereas local explanations enhance accountability.

2.6.3 By Model Dependency

2.6.3.1 Model-Agnostic Methods

Model-agnostic methods operate independently of model internals, relying solely on input–output behaviour. This black-box approach offers broad applicability across diverse architectures, ranging from tree ensembles to DNN, though often at the cost of faithfulness and computational efficiency.

LIME exemplifies this paradigm by constructing local surrogate models through perturbation-based sampling techniques. Its strength resides in its versatility; however, it

remains sensitive to perturbation design and may yield unstable results in non-linear regions (Alvarez-Melis and Jaakkola, 2018).

Permutation Feature Importance (Breiman, 2001), although initially developed for Random Forests, generalises across various models. It quantifies feature importance by assessing the decline in performance when feature values are permuted. Nevertheless, its ability to accurately represent importance may be compromised in the presence of multicollinearity, as permutation can disrupt joint distributions (Molnar et al., 2020; Strobl et al., 2008).

PDPs and Individual Conditional Expectation (ICE) plots (Goldstein et al., 2015) effectively offer visual representations of both global and individual-level effects by marginalising or conditioning predictions across features. PDPs elucidate average effects, while ICE delineates per-instance trajectories. Nevertheless, both methodologies are predicated on independence assumptions and may obscure interaction effects (Apley and Zhu, 2020).

Anchors (Ribeiro et al., 2018) offer localised if-then rules that serve as sufficient conditions for predictions. Their objective is to maximise precision amid sampling-based perturbations, thereby enhancing the discreteness and interpretability of explanations. Nevertheless, the generation of informative anchors may prove to be computationally intensive and is dependent on the dataset utilised.

Despite their inherent flexibility, model-agnostic methods typically depend on approximations, whether through perturbation, marginalisation, or local surrogates. Consequently, this reliance introduces a potential divergence between the explanations provided and the actual behaviour of the model. This discrepancy raises concerns in domains that demand high fidelity and accountability (Rudin, 2019).

2.6.3.2 Model-Specific Methods

Model-specific approaches utilise the internal architecture of the model, utilising gradients, structural elements, or attention mechanisms to directly trace the influence of features. These approaches often yield higher-fidelity explanations, but they are closely tied to the model type, thereby limiting their transferability and applicability.

For neural networks, Saliency Maps use gradients to estimate feature relevance. Although easy to compute, they are susceptible to noise and sensitive to model parameters, with evidence showing similar outputs for randomised models (Adebayo et al., 2018).

Grad-CAM enhances saliency by utilising gradients related to intermediate feature maps, resulting in class-discriminative spatial heatmaps. This method is effective for CNNs in visual tasks; however, it is not generalisable to non-spatial models.

In tree-based models, TreeSHAP (Lundberg et al., 2020) offers an efficient and precise computation of Shapley values, utilising the tree structure to facilitate rapid and accurate attributions. Although it is highly effective for gradient-boosted ensembles, its design remains model-specific and non-transferable.

Explanations based on the approach of Saabas (Ando Saabas, 2021) offer rapid and heuristic approximations by assigning predictive changes along a singular decision pathway. Although these methods are efficient, they do not possess the axiomatic guarantees found in methodologies such as SHAP. They inadvertently neglect contributions from features that are not included in the decision-making process path.

In transformer architectures, attention weights are often visualised as proxies for feature importance (Vig, 2019). However, their interpretability is debated: attention can be manipulated without altering outputs (Jain and Wallace, 2019). To improve robustness, newer methods, such as those by (Chefer et al. (2021), propagate relevance scores through attention blocks; however, these remain tightly tied to transformer internals and are challenging to validate.

In summary, model-specific methods enable profound insight into internal computations but sacrifice generalisability. They are most effectively utilised when the architecture of the model is transparent and accessible.

2.6.4 By Technique Type

Contemporary explainability methods can be classified according to their methodological foundations. This section examines seven principal categories: surrogate models, perturbation-based methods, gradient-based techniques, decomposition-based approaches, attention-based strategies, feature attribution methods, and counterfactual or example-

based techniques. Although there are overlaps, many methods traverse multiple categories; this taxonomy elucidates their fundamental mechanisms and underlying assumptions.

2.6.4.1 Surrogate Models

Surrogate models are designed to replicate the behaviour of complex, often opaque models through the utilisation of interpretable models such as linear regressions, decision trees, or rule lists. These models can be trained either locally, focusing on a specific prediction, or globally, encompassing the entire dataset.

Key examples of such models include LIME, Anchors, and Explainable Boosting Machines (EBMs). Notably, EBMs (Nori et al., 2019) utilise Generalised Additive Models featuring pairwise interactions to achieve an equilibrium between interpretability and performance. Nevertheless, they may fail to capture profound, non-linear dependencies adequately.

2.6.4.2 Perturbation-Based Methods

Perturbation-based techniques explain model behaviour by systematically altering input features and observing changes in the output, without needing access to the internal architecture of the model. They estimate feature importance based on how predictions vary with input perturbations.

Permutation Feature Importance, PDP and ICE plots are prime examples. LIME and SHAP, although often considered surrogate methods, also fall under this category due to their use of input perturbation. A shared limitation across all perturbation-based approaches is their computational inefficiency, particularly for large models, and sensitivity to perturbation schemes, particularly in high-dimensional or structured domains.

2.6.4.3 Gradient-Based Methods

Gradient-based methods involve computing the partial derivatives of the output with respect to its input features. By utilising differentiability, these techniques enable the assessment of local feature sensitivity and are primarily employed in the context of neural networks.

Prominent examples include Saliency Maps, Grad-CAM, and Integrated Gradients (Sundararajan et al., 2017). Integrated Gradients address the issue of gradient saturation by

integrating gradients along a straight-line path from a baseline input to the actual input. Although these methods are theoretically grounded, their outputs can vary significantly depending on the choice of baseline, as highlighted by (Kindermans et al., 2018), which raises concerns about their reliability in specific applications.

2.6.4.4 Decomposition-Based Methods

Decomposition methods attribute model output by breaking it down into additive contributions from input features, often utilising cooperative game theory principles, such as SHAP.

Layer-wise Relevance Propagation (LRP) (Bach et al., 2015) redistributes output scores through neural network layers. While effective for computer vision, interpretability relies on tuning propagation rules and may not generalise across architectures.

DeepLIFT (Shrikumar et al., 2017) compares neuron activations to a reference input, bypassing local gradient issues. It shares baseline sensitivity concerns with Integrated Gradients but offers stable attribution.

Although decomposition methods offer structured, axiomatic attribution, their reliance on additive assumptions can obscure interaction effects or feature dependencies intrinsic to DNN.

2.6.4.5 Attention-Based Methods

In attention-based architectures such as transformers, attention weights are often interpreted as indicators of feature importance. These visualisations, while intuitive, are not inherently explanatory. (Jain and Wallace, 2019) demonstrated that attention distributions can be adversarial and have a negligible impact on output, and attention alone lacks the causal criteria necessary to serve as faithful explanations.

Robust variants include attention rollout (Abnar and Zuidema, 2020) and gradient-weighted attention (Chefer et al., 2021), which attempt to trace relevance across layers. These approaches improve attribution fidelity but remain model-specific and do not generalise well beyond attention-based frameworks.

2.6.4.6 Feature Attribution Methods

Feature attribution encompasses a broad class of methods that assign numeric importance scores to input features for a specific prediction. Techniques such as SHAP, Integrated Gradients, DeepLIFT, and LIME are unified under this paradigm.

Despite methodological differences, these methods share common challenges, including sensitivity to baseline choice, instability under input perturbation, and limited capacity to reflect complex feature interactions. SHAP remains the most theoretically rigorous, while gradient-based variants offer computational tractability with architectural access.

2.6.4.7 Counterfactual and Example-Based Methods

Counterfactual and example-based methods provide contrastive explanations by identifying the minimal changes to an input that would alter the decision of the model. Rather than attributing prediction to features, they answer "what if" scenarios, such as Counterfactual explanations.

Influence functions (Koh and Liang, 2017) trace the effect of training points on a given prediction by approximating the impact of removing or upweighting instances. Though theoretically appealing, they rely on convexity assumptions and become intractable in DNN.

Prototypes and criticisms (Kim et al., 2014) aim to summarise the dataset by identifying representative and outlier examples, facilitating intuitive understanding. However, these methods struggle with scalability and maintaining semantic relevance in large, heterogeneous datasets.

While these approaches align closely with human reasoning, their dependence on suitable data distributions and their computational demands limit their general utility. A summary of the taxonomy, including its techniques, is presented in Table 2- 1 below.

Table 2- 1 Taxonomy of techniques

Taxonomy Dimension	Category	Representative Methods
Time of Explanation	Intrinsic Interpretability	Decision Trees, Rule Lists, Linear Models, GAMs
	Post-hoc Explainability	LIME, SHAP, Grad-CAM, Counterfactuals
Scope of Explanation	Global Explanations	PDP, Feature Importance, SHAP (Global)
	Local Explanations	LIME, SHAP (local), Counterfactuals
Model Dependency	Model-Agnostic	LIME, Anchors, PDP, Permutation Importance
	Model-Specific	Saliency Maps, Grad-CAM, Integrated Gradients, Tree-Explainer
Technique Type	Surrogate Models	LIME, Anchors, EBMs
	Perturbation-Based	Permutation Importance, PDP, ICE
	Gradient-Based	Saliency Maps, Integrated Gradients, Grad-CAM
	Decomposition-Based	Deep-LIFT, LRP, SHAP
	Attention-Based	Attention Visualisation, Rollout
	Feature Attribution	SHAP, LIME, Integrated Gradients, DeepLIFT
	Example-Based	Counterfactuals, Prototypes, Influence Functions

2.7 Literature review of explainability techniques

As the deployment of ML models becomes increasingly widespread across high-stakes domains, the demand for explainability has intensified. This literature review examines a range of explainability techniques that have contributed to recent advances in interpretable AI, focusing on their methodological design, empirical performance, and practical utility across various model architectures and application contexts.

By critically examining how these techniques are applied in contemporary research, the review identifies their strengths, limitations, and underlying assumptions. The goal is to provide a clear understanding of their interpretability and reliability, as well as the challenges they pose in terms of scalability, consistency, and trustworthiness.

2.7.1 SHAP (SHapley Additive exPlanations)

SHAP is widely used due to its game-theoretic formulation, which decomposes the prediction of the model into additive contributions from individual features. The theoretical appeal of this approach arises from its adherence to axioms such as local accuracy, missingness, and consistency, attributes that furnish a compelling rationale for its implementation. However, its practical reliability has come under scrutiny in recent literature, with concerns around computational cost, instability, and questionable alignment with real-world interpretability needs.

Bitton et al. (2022) proposed Latent SHAP, which improves human interpretability by shifting the operation of SHAP to a low-dimensional latent space learned via autoencoders. This facilitates the attribution to semantically meaningful concepts as opposed to raw inputs such as pixels. Although this approach is intuitive, it is significantly reliant on the quality and fidelity of the learned latent space, which introduces an additional layer of complexity and potential distortion in the explanations.

To address the computational inefficiency associated with SHAP, Kelodjou et al. (2024) have introduced a neighbourhood-based approximation for KernelSHAP. This methodology samples from structured local regions of the input space, aiming to enhance stability and mitigate runtime variance. Nonetheless, this approach may lead to an oversimplification of interactions by concentrating exclusively on local contexts, which often fails to capture significant global dependencies among features adequately.

Critiques of foundational assumptions have gained increasing prominence. Huang and Marques-Silva, (2023) demonstrated that SHAP can yield misleading rankings even when applied within regression models, thereby calling into question the trustworthiness of its feature importance scores. They contend that reliance on additive decompositions of SHAP does not align adequately with intuitive or causal attributions. Reflecting these concerns, Letoffe et al. (2024) performed stress tests under controlled and idealised conditions, discovering that explanations of SHAP diverge from the ground truth even when models exhibit smoothness and continuity, thereby underscoring a discrepancy between theoretical guarantees and practical outcomes.

Muschalik et al. (2024) introduced TreeSHAP-IQ, which extends SHAP to facilitate counterfactual-style queries in decision tree models. This innovative method enables users to examine "what-if" scenarios, thereby aligning SHAP closely with causal reasoning. Nonetheless, it presupposes a causal interaction of features and is limited to tree-based architectures, which restricts its broader applicability.

In conclusion, SHAP remains a benchmark attribution method due to its robust theoretical foundations and extensive practical application. However, recent literature has identified significant limitations, particularly in terms of robustness and causal interpretability. While it remains valuable, particularly for benchmarking and comparison, SHAP is increasingly supplemented by methodologies that emphasise stability, causality, or domain-aligned representations.

2.7.2 LIME (Local Interpretable Model-Agnostic Explanations)

LIME constitutes a fundamental approach in explainable AI, offering post-hoc local explanations through the fitting of sparse linear surrogate models surrounding individual predictions. This method perturbs the input data and utilises the resultant outputs to develop an interpretable model within the vicinity of a query instance. Despite its widespread adoption, the foundational assumptions of LIME, particularly those of local linearity and neighbourhood sampling, have been scrutinised in recent studies due to their limitations when applied to complex models.

Anchor LIME (Ribeiro et al., 2018) extends the original approach by generating high-precision rules ("anchors") that explain model decisions over subregions of the input space. These rule-based explanations provide actionable insights than local regressions, particularly in classification tasks. However, anchor generation relies on sampling and heuristic coverage thresholds, which can reduce interpretability when rules are either too specific or too sparse to generalise. Moreover, the method struggles with capturing nuanced nonlinearities beyond its anchored region.

Slack et al. (2020) utilised LIME to elucidate vulnerabilities within XAI pipelines, demonstrating that adversarial models possess the capability to manipulate LIME to obscure biased behaviours while still generating seemingly plausible explanations. This research suggests that LIME, due to its model-agnostic nature, is susceptible to deception and lacks

assurances regarding causal or ethical alignment, thereby raising significant concerns about its deployment in sensitive domains.

The stability of LIME remains a prominent criticism, as discussed by Thibault Laugel (2020), where the explanations can differ markedly between adjacent inputs, even when the model outputs are nearly equivalent. This inconsistency is attributed to both the stochastic characteristics of perturbations and the complex decision boundary of the model. The authors proposed decision-boundary-aware sampling as a solution, which enhanced consistency, though at the expense of high computational demands. These findings raise concerns regarding the reliability of LIME in high-stakes or audit-intensive contexts.

The use of random perturbation in LIME can lead to unstable explanations for the exact prediction, which is a problem in domains such as medical diagnosis, where consistency is crucial. Zafar and Khan (2019) present a stable alternative called Deterministic LIME (DLIME), which utilises hierarchical clustering and K-Nearest Neighbours to select relevant data. Experimental results show that DLIME offers stable explanations compared to LIME.

Despite its influence, LIME continues to face fundamental challenges. The assumption of local linearity often breaks down in DNN or ensemble models. Its perturbation-based sampling can yield misleading attributions if the sampled neighbourhood is not representative. Moreover, LIME usually struggles with the fidelity-interpretability trade-off; linear surrogates may oversimplify to be faithful or overly complex to remain interpretable.

In summary, LIME remains a pivotal method in XAI, particularly for its model-agnostic framework and simplicity. However, its reliability, stability, and vulnerability to misuse limit its utility in isolation. It is best viewed as an introductory or complementary tool to be used in conjunction with robust and domain-aligned explanation techniques.

2.7.3 Counterfactual Explanations

Counterfactual explanations (CFE) provide a direct and accessible means of interpretability by identifying minimal alterations to an input that would modify the prediction of the model. Formally introduced by Wachter et al. (2017), CFE frames this as an optimisation problem, seeking the nearest input (according to a specified distance metric) that results in a differing outcome. This approach aligns effectively with human reasoning and is notably beneficial in

sectors demanding transparency and accountability, such as credit scoring, medical decision-making, or legal adjudication.

Karimi et al. (2021) advanced this line of inquiry by incorporating causal reasoning. They posited that numerous CFEs derived solely from data may be implausible or misleading; for instance, altering the income of an individual without impacting other causally interconnected features. Their methodology employs Structural Causal Models (SCMs) to ensure that the generated counterfactuals are both feasible and actionable within the causal framework of the domain. Although this approach is principled, it necessitates access to dependable causal graphs, which are frequently challenging to construct or estimate.

Dandl et al. (2020) addressed the multi-objective nature of CFE by balancing proximity, sparsity, diversity, and plausibility and proposed Multi-Objective Counterfactuals (MOC). This gradient-based method efficiently generates diverse counterfactuals using differentiable objectives. This method addressed an essential limitation in earlier works that yielded single or redundant explanations, thus enhancing user trust and flexibility.

Russell (2019) advanced an integer programming method for generating plausible, sparse CFEs in tabular data. His approach constrains counterfactuals within the convex hull of training data, ensuring realism. Despite its effectiveness, the method is computationally intensive and poorly scales with feature dimensionality, and convexity assumptions limit its ability to capture the full complexity of real-world datasets.

Kommiya Mothilal et al. (2021) evaluated CFE for fairness auditing, proposing CF-Fairness to quantify fairness violations based on changes in model output due to sensitive attributes (e.g., race or gender). Their results showed that many top models demonstrate counterfactually unfair behaviour, despite appearing fair by group-level metrics. This highlights the diagnostic power of CFE while raising ethical concerns about the deployment of sensitive biases.

In summary, CFE excels in human-aligned reasoning, actionability, and questioning decision boundaries in a model-agnostic manner. They are crucial for fostering recourse, transparency, or fairness. However, these benefits rely on generating counterfactuals that are mathematically valid, semantically meaningful, and causally grounded. Without these safeguards, CFEs can mislead or harm, particularly in sensitive areas. Counterfactuals are not

stand-alone explanations; they are part of a broader interpretability toolkit, ideally used with feature attribution, sensitivity analysis, and causal diagnostics. They show promise but require careful design, domain adaptation, and critical interpretation.

2.7.4 Layer-wise Relevance Propagation (LRP)

LRP is a decomposition-based technique for interpreting predictions of layered neural networks. It redistributes the prediction score backwards to input features based on their contribution to the final decision, using relevance conservation rules to maintain the score across layers. This creates a heatmap over the input that reflects the importance of each feature.

In Montavon et al. (2017), the authors extended LRP to deep convolutional networks, demonstrating its effectiveness in visual classification. Unlike gradient-based methods, which can be affected by saturation or noise, LRP offers stable and class-discriminative explanations. The paper presents relevance rules, such as the z-rule and ε-rule, each with unique propagation assumptions, providing flexibility for various architectures. A noted limitation of LRP is sensitivity to the chosen propagation rule, which may not generalise well across tasks without careful tuning.

Samek et al. (2019) applied LRP to medical imaging, particularly for tumour classification in MRI scans. They compared LRP with Grad-CAM and saliency maps, concluding that LRP offered finer-grained, spatially localised explanations, which radiologists found actionable. However, they noted that the usefulness of the method declined in architectures with non-standard layers or residual connections, where the conservation principle was harder to apply rigorously.

In Lapuschkin et al. (2019), LRP was used for model debugging, revealing “Clever Hans” predictors, models that relied on spurious correlations instead of semantically meaningful features. For example, an image classifier for horses relied on copyright watermarks in the training set. This demonstrated auditing model behaviour and validated training pipelines. However, LRP required manual inspection and domain knowledge to interpret heatmaps meaningfully, an inherent limitation at scale.

LRP is a unique tool in the XAI toolkit. Model-specific and efficient, it effectively works with deep feedforward or convolutional architectures. Its strength lies in attribution faithfulness, redistributing relevance while respecting the internal computation graph, thus avoiding some pitfalls of model-agnostic methods, such as LIME. LRP excels in debugging and auditing tasks, making it a favourite for researchers validating model integrity. However, its internal propagation rules limit flexibility across architectures, particularly for models that use dynamic routing, attention, or non-standard modules. It also lacks intuitive interpretability outside image domains. When applied to text or tabular data, visualisation and relevance semantics can be obscured. Additionally, LRP assumes linear additivity of relevance, which may not apply to how non-linear transformations distribute semantic meaning.

2.7.5 Graph Neural Networks with Causal Structural Models

The integration of Causal Structural Models (CSMs) into Graph Neural Networks (GNNs) represents a novel direction in explainable AI, aiming to ground explanations in counterfactual and interventional semantics within graph-structured data. Unlike traditional post-hoc methods, such as GNNExplainer, which focus on saliency or feature attribution, CSM-enhanced GNNs embed causal reasoning through either explicit causal graphs or learned causal structures in the latent space. These approaches are particularly beneficial in domains such as molecular property prediction, recommendation systems, and social network analysis, where understanding causality across subgraphs can aid generalisation.

However, these benefits come with trade-offs. Most causal GNN methods require strong assumptions, such as causal sufficiency and faithfulness, which are often untestable from observational data (Peter Spirtes et al., 2000). Wu et al. (2023) highlight the challenge of generating counterfactual graphs that are both plausible and interpretable, particularly without incorporating domain constraints. These challenges are compounded by computational burdens: estimating interventional distributions in large, sparse graphs is expensive and sensitive to noise.

Critically, the absence of standardised evaluation metrics for causal explanations within graph models also constrains the reproducibility and reliability of existing results. Although GNNExplainer has established initial foundations for interpreting GNN predictions, it is devoid of causal grounding. In contrast, methodologies such as CF-GNNExplainer (Lucic et

al., 2022) seek to address this discrepancy by generating counterfactual subgraphs, the inclusion or exclusion of which alters predictions, thereby offering actionable insights. Nonetheless, the challenges of scalability and generalisability persist.

Overall, while still in its nascent stage, GNNs with CSMs offer a promising framework for providing faithful, intervention-aware explanations in structured domains. CSMs in particular is said to offer support for characterisation of the causal reasoning rationale of the model in emergent contexts. Their ability to reason causally makes them a valuable addition to the XAI landscape, particularly when interpretability and robustness under distribution shift are critical. However, widespread adoption hinges on formalising evaluation protocols, easing computational demands, and reducing the domain-specific expertise required to build valid causal graph priors.

This literature review has critically evaluated key explainability methods in AI, each offering distinct strengths—from model-agnostic local approximations to gradient- and decomposition-based insights. While widely adopted, these techniques face persistent limitations, including sensitivity to baselines, assumptions of feature independence, and poor scalability. Despite this, methods such as SHAP remain popular due to their balance of generality and interpretability, while counterfactuals offer actionable insights with added complexity. Overall, no singular method is adequate across all scenarios; a hybrid, context-aware approach is imperative.

2.8 Evaluation of Explainability Methods

2.8.1 Metrics and Benchmarks

Evaluating the effectiveness and reliability of explainability methods in AI systems is significantly complex than evaluating conventional performance metrics such as accuracy, precision, or recall. A singular value cannot adequately encapsulate the quality of an explanation; it is intrinsically multidimensional, varying according to context, application domain, user expertise, and regulatory prerequisites. Researchers have proposed several dimensions across which explanations should be evaluated. These dimensions include fidelity, sparsity (or parsimony), stability, human simulatability, and consistency. Each dimension addresses a specific aspect of interpretability, and there are often trade-offs between them.

Comprehending and critically assessing these dimensions is imperative for both researchers and practitioners who aim to implement explainable models in practical settings.

2.8.1.1 Fidelity

Fidelity is one of the most fundamental metrics in explainability research. It refers to the degree to which an explanation method accurately reflects the internal mechanisms or decision boundaries of the underlying model. High fidelity implies that the explanation reveals the actual reasoning process of the model, rather than providing a simplified or heuristic summary (Doshi-Velez and Kim, 2017).

Several methods have been developed to quantify fidelity. One commonly used approach involves constructing a surrogate model, typically a straightforward and interpretable model such as a decision tree or linear regressor, that approximates the behaviour of the complex black-box model (Ribeiro et al., 2016). The agreement between the predictions of the surrogate and the original model can be measured using accuracy or R-squared values, depending on the task. Another approach involves faithfulness scores, which quantify the impact of removing or masking top-ranked features identified by the explanation method. For instance, if removing these features leads to a significant degradation in prediction quality, the explanation is considered faithful.

Despite these advancements, high fidelity does not inherently translate into human usability. A highly accurate surrogate model may itself be complex and opaque to human understanding, thus defeating the purpose of explainability. Therefore, fidelity is necessary but not sufficient; it must be coupled with additional properties, such as simplicity and clarity.

2.8.1.2 Sparsity

Sparsity, also referred to as parsimony, is another critical dimension in explanation evaluation. The principle underlying sparsity is grounded in the cognitive limitations of humans to comprehend straightforward explanations. An explanation that highlights a smaller number of relevant features or rules is likely to be comprehensible to end users. Sparsity is typically quantified by counting the number of features involved in the explanation or measuring the depth and length of decision rules in tree-based or rule-based models.

For example, in LIME, the local surrogate coefficients of the model can be inspected to determine which features are considered most important, and only a few top-ranked ones are typically shown to the user. Similarly, decision rule extraction methods evaluate the number and complexity of rules employed to make a prediction. However, this preference for simplicity introduces a significant trade-off.

Overly sparse explanations might fail to capture the complex dependencies between features, thereby omitting critical information and potentially misleading users. The challenge lies in achieving a balance between simplicity and completeness, where the explanation is concise enough for human consumption yet still accurately represents the reasoning process of the model (Guidotti et al., 2018).

2.8.1.3 Stability

Stability or robustness is an important but often overlooked property of explanation methods. It refers to the consistency of explanations in response to small, usually imperceptible changes in input data. Ideally, similar inputs should produce similar explanations. This property is particularly vital for applications in safety-critical domains, where reliability and consistency are paramount.

Quantitative evaluation of stability typically involves computing the similarity between explanations for slightly perturbed inputs. For instance, in saliency-based methods, researchers measure the overlap or cosine similarity between saliency maps generated from original and perturbed data samples. In feature attribution techniques, the similarity of feature importance vectors across different runs or input variations can be examined.

Alvarez-Melis & Jaakkola (2018) highlighted that many popular methods, including LIME and gradient-based saliency maps, often exhibit poor robustness. Small perturbations in input can result in disproportionately large changes in the generated explanations. This instability undermines the credibility of the explanation and erodes user trust. Additionally, randomness in the explanation algorithm itself, such as stochastic sampling in LIME, can further exacerbate this instability. Consequently, improving robustness remains a key research direction in the field.

2.8.1.4 Human simulatability

Human simulatability provides a user-centric approach to evaluating explanations. It concerns the ability of humans to replicate or predict decisions by a model based solely on the provided explanations. This metric focuses less on technical correctness and on practical utility, as it measures how well an explanation aids human understanding and decision-making.

Empirical assessment of simulatability is typically performed through controlled user studies, where participants are asked to simulate model outputs based solely on the input data and its accompanying explanation. Their success in these tasks, often measured through accuracy or task completion time, indicates the quality of the explanation. Research by Poursabzi-Sangdeh et al. (2021) demonstrated that even explanations perceived as intuitive can fail to improve human decision-making if they do not align with user mental models.

Simulatability becomes particularly critical in high-stakes domains such as healthcare, legal reasoning, and financial services, where explanations are often scrutinised by domain experts and regulators. The major challenge, however, lies in the cost and scalability of conducting rigorous user studies, which are typically resource-intensive and domain-specific. Nonetheless, they remain among the most reliable methods for assessing the real-world impact of explainability.

2.8.1.5 Consistency

Consistency is another vital criterion, particularly in the context of fairness and regulatory compliance. It demands that similar models, or even the same model making similar predictions, should generate similar explanations. Inconsistent explanations can lead to confusion and distrust, particularly in settings where accountability and transparency are required.

Evaluating consistency involves measuring the similarity of explanations across different model instances trained on similar data or across different inputs leading to the same output. Ribeiro et al. (2016) argued that explanation methods must offer a degree of invariance to model data permutations. However, achieving this is challenging due to factors such as model stochasticity, feature correlation, and algorithmic randomness.

In neural networks, for example, different initialisations or training paths can result in different internal representations even when predictive performance remains unchanged. This variance can cascade into the explanation layer, producing divergent rationalisations for identical outcomes. Addressing these issues requires both algorithmic innovation and rigorous evaluation protocols that go beyond superficial consistency checks.

2.8.2 Limitations of Current Evaluation Metrics

Despite the development of multiple evaluation metrics, the field of explainable AI remains constrained by several foundational challenges that hinder robust assessment of explanation quality.

One of the most pressing issues is the lack of standardisation. The field lacks universally accepted definitions, taxonomies, or benchmarks, making it challenging to compare different explainability methods fairly. Terminological inconsistencies further compound the problem. Terms such as “explanation” and “interpretation” are often used interchangeably in the literature, despite referring to distinct concepts (Doshi-Velez and Kim, 2017; Lipton, 2018). This lack of clarity hampers meaningful communication and cross-comparison between studies. Initiatives such as OpenXAI and the DARPA XAI program have pushed towards consistent evaluation frameworks, but widespread adoption remains limited.

Subjectivity is another significant limitation in evaluating explainability. Unlike performance metrics, which are objective and reproducible, the effectiveness of an explanation often depends on user perception, background knowledge, and task context. As a result, many studies rely on subjective human evaluations, such as Likert scale ratings (Joshi et al., 2015) or user satisfaction surveys. While these metrics provide valuable insights, they are inherently biased and often lack generalisability. Furthermore, such evaluations tend to be domain-specific. The integration of human factors into explanation evaluation is essential but fraught with methodological and practical challenges, including participant recruitment, experiment design, and ethical considerations (Miller, 2019).

A particularly contentious issue is the distinction between explanation and Interpretability. Post-hoc explanation methods, such as LIME or SHAP, are designed to provide insights into the behaviour of the model after it has been trained. However, they often act as rationalisers rather than actual reflectors of the internal decision-making process. This

distinction is particularly problematic in contexts that demand a high degree of transparency and accountability, such as judicial decision-making or algorithmic lending. (Rudin, 2019) has strongly argued for the use of inherently interpretable models in such high-stakes domains, noting that post-hoc methods are often incapable of providing faithful and verifiable explanations.

Another major limitation lies in the datasets used for benchmarking explainability methods. Many widely used datasets, such as MNIST, CIFAR-10, or tasks from the UCI repository, are simplistic and fail to capture the complexities of real-world applications. These datasets are often insufficient to stress-test the nuanced behaviour of explainability techniques. As Arrieta et al. (2020) and Vilone and Longo (2021) note, the lack of diversity and complexity in benchmark datasets can lead to overfitting to specific tasks or explanation styles, thereby limiting generalisability. There is a growing consensus that richer, contextually grounded datasets, such as those involving electronic health records, legal documents, or financial transactions, are necessary to advance the state of the art in XAI evaluation.

Ultimately, it is essential to recognise the broader epistemological challenge of explanation in AI. Unlike traditional software systems, where the logic is explicitly encoded and traceable, ML models often operate through distributed representations and non-linear interactions that are not easily decomposed into human-understandable components. This fundamental mismatch between how models represent knowledge and how humans understand reasoning processes complicates all efforts at explainability. Addressing this gap demands collaborative efforts across disciplines to create evaluation frameworks that draw on insights from computer science and human-computer interaction, combining rigour with practical applicability.

In summary, while significant strides have been made in defining and quantifying various aspects of explanation quality, the evaluation of XAI methods remains an open and evolving area of research. Fidelity, sparsity, stability, human simulatability, and consistency each provide valuable but incomplete views of explanation quality. Moreover, limitations such as lack of standardisation, subjectivity, reliance on post-hoc rationalisation, and benchmark bias continue to undermine the robustness of current evaluation practices. Addressing these issues will be essential for the development of trustworthy AI systems that can be deployed responsibly in real-world contexts.

2.8.3 Summary of the Literature Survey

Table 2- 2 presented below provides a consolidated overview of seminal contributions from key authors who have significantly shaped the field of Explainable Artificial Intelligence. It highlights foundational methods across various interpretability paradigms, including decision trees, rule-based systems, GAMs, model-agnostic explanation frameworks such as LIME and SHAP, and saliency-based techniques. This curated collection of literature captures the diversity and evolution of XAI methodologies, offering a meta-perspective on the landscape of interpretable models and the critical advancements that enable transparent, reliable, and user-centred AI systems.

Table 2- 2 Summary of Literature Survey

Authors	**Title**	**Year of Publication**	**Summary of the Publication**
Quinlan J. R.	C4.5: programs for machine learning.	2014	Authors proposed the C4.5 algorithm for generating decision trees for classification tasks.
Breiman, L. et al.	Classification and regression trees.	2017	Authors proposed CART, a decision tree algorithm for classification and regression.
Rivest R. L.	Learning decision lists.	1987	Authors proposed decision lists as a simple, rule-based classification method.
Ustun, B., & Rudin, C.	Supersparse linear integer models for optimized medical scoring systems.	2016	Authors proposed Supersparse Linear Integer Models (SLIM), a sparse, interpretable linear model for scoring systems.
Hastie, T., & Tibshirani, R.	Generalized additive models.	1986	Authors proposed GAMs to model non-linear relationships while maintaining interpretability.
Caruana, R., et al.	Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission.	2015	Authors demonstrated the application of interpretable GAMs for healthcare predictions.
Lundberg, S. M., & Lee, S. I.	A unified approach to interpreting model predictions.	2017	Authors proposed SHAP, a unified framework for feature attribution using Shapley values.

Authors	Title	Year of Publication	Summary of the Publication
Ribeiro, M. T., et al.	"Why should I trust you?": Explaining the predictions of any classifier.	2016	Authors proposed LIME, a local, model-agnostic explanation method.
Simonyan, K., et al.	Deep inside convolutional networks: Visualising image classification models and saliency maps.	2014	Authors proposed saliency maps to visualise pixel importance in CNNs.
Selvaraju, R. R., et al.	Grad-CAM: Visual explanations from deep networks via gradient-based localisation.	2020	Authors proposed Grad-CAM for visual explanations using class-discriminative heatmaps.
Wachter, S., et al.	Counterfactual explanations without opening the black box: Automated decisions and the GDPR.	2017	Authors proposed generating counterfactual explanations without accessing the internal model.
Friedman, J. H.	Greedy function approximation: A gradient boosting machine.	2001	Authors proposed gradient boosting and introduced PDPs for interpreting model predictions.
Greenwell, B. M., et al.	A simple and effective model-based variable importance measure.	2018	Authors proposed an approach using PDPs for variable importance in complex models.
Breiman, L.	Random forests.	2001	Authors proposed permutation feature importance as part of the Random Forest framework.
Goldstein, A., et al.	Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation.	2015	Authors proposed ICE plots to visualise individual feature effects.

Authors	Title	Year of Publication	Summary of the Publication
Ribeiro, M. T., et al.	Anchors: High-precision model-agnostic explanations.	2018	Authors proposed Anchors, a high-precision local explanation method using if-then rules.
Lundberg, S. M., et al.	From local explanations to global understanding with explainable AI for trees.	2020	Authors proposed TreeSHAP for consistent, efficient explanations in tree ensembles.
Saabas, A.	Interpreting random forests.	2014	Authors proposed a local feature contribution method specific to decision trees.
Vig, J.	A Multiscale Visualization of Attention in the Transformer Model.	2019	Authors proposed attention visualisation techniques for Transformer models.
Jain, S., & Wallace, B. C.	Attention is not Explanation.	2019	Authors argued that attention weights are not reliable explanations.
Nori, H., et al.	InterpretML: A unified framework for machine learning interpretability.	2019	Authors proposed the Explainable Boosting Machine (EBM), a GAM-like, interpretable model.
Sundararajan, M., et al.	Axiomatic attribution for deep networks.	2017	Authors proposed Integrated Gradients, an attribution method for deep networks.
Bach, S., et al.	On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.	2015	Authors proposed Layer-wise Relevance Propagation (LRP) for pixel-level explanations.
Shrikumar, A., et al.	Learning important features through propagating activation differences.	2017	Authors proposed DeepLIFT, an efficient backpropagation-based attribution method.
Abnar, S., & Zuidema, W.	Quantifying Attention Flow in Transformers.	2020	Authors proposed attention rollout for robust attention flow quantification in Transformers.

Authors	Title	Year of Publication	Summary of the Publication
Chefer, H., et al.	Transformer interpretability beyond attention visualization.	2021	Authors proposed gradient-based Transformer interpretability beyond attention flow.
Koh, P. W., & Liang, P.	Understanding black-box predictions via influence functions.	2017	Authors proposed using influence functions to trace training data influence on predictions.
Kim, B., et al.	The Bayesian Case Model: A generative approach for case-based reasoning and prototype classification.	2014	Authors proposed a generative case-based reasoning framework using prototypes.
Bitton, R., et al.	Latent SHAP: Toward practical human-interpretable explanations.	2022	Authors proposed Latent SHAP for interpretable explanations in latent feature spaces.
Kelodjou, G., et al.	Shaping up SHAP: Enhancing stability through layer-wise neighbour selection.	2024	Authors proposed a neighbourhood-based KernelSHAP approximation to improve stability.
Muschalik, M., et al.	Beyond treeSHAP: Efficient computation of any-order SHAPley interactions for tree ensembles.	2024	Authors proposed TreeSHAP-IQ for computing higher-order SHAP interactions efficiently.
Zafar, M. R., & Khan, N. M.	DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems.	2019	Authors proposed DLIME, a deterministic variant of LIME for reliable local explanations.

Authors	Title	Year of Publication	Summary of the Publication
Karimi, A. H., et al.	Algorithmic recourse: From counterfactual explanations to interventions.	2021	Authors proposed integrating causal reasoning into counterfactual explanations for actionable recourse.
Dandl, S., et al.	Multi-objective counterfactual explanations.	2020	Authors proposed a multi-objective optimisation framework for generating counterfactuals.
Russell, C.	Efficient search for diverse coherent explanations.	2019	Authors proposed using integer programming for generating diverse, plausible counterfactuals.
Kommiya Mothilal, R., et al.	Towards unifying feature attribution and counterfactual explanations: Different means to the same end.	2021	Authors proposed aligning feature attributions with counterfactuals to support fairness audits.
Montavon, G., et al.	Explaining nonlinear classification decisions with deep Taylor decomposition.	2017	Authors proposed extending LRP with deep Taylor decomposition for CNNs.
Samek, W., et al.	Explainable AI: Interpreting, explaining and visualizing deep learning.	2019	Authors applied LRP-based methods to interpret medical imaging models.
Lapuschkin, S., et al.	Unmasking Clever Hans predictors and assessing what machines really learn.	2019	Authors used LRP to expose spurious correlations in machine learning models.
Lucic, A., et al.	CF-GNNExplainer: Counterfactual explanations for graph neural networks.	2022	Authors proposed CF-GNNExplainer for generating counterfactual explanations in GNNs.
Doshi-Velez, F., & Kim, B.	Towards a rigorous science of interpretable machine learning.	2017	Authors advocated for formalising interpretability with high-fidelity explanations.

Authors	Title	Year of Publication	Summary of the Publication
Guidotti, R., et al.	A survey of methods for explaining black box models.	2018	Authors provided a comprehensive survey on explainable machine learning methods.
Poursabzi-Sangdeh, F., et al.	Manipulating and measuring model interpretability.	2021	Authors demonstrated the gap between intuitive explanations and effective human decision-making.
Lipton, Z. C.	The mythos of model interpretability.	2018	Authors critically analysed the ambiguous use of "interpretability" in machine learning.
Joshi, A., et al.	Likert scale: Explored and explained.	2015	Authors explained the design and application of Likert scale ratings in surveys.
Miller, T.	Explanation in artificial intelligence: Insights from the social sciences.	2019	Authors proposed integrating social science principles to improve explanation design in AI.

2.9 Challenges in Explainability Research

While a growing body of work has focused on the development and evaluation of explainability methods, a range of systemic challenges persist in the field. These issues span methodological limitations, epistemic inconsistencies, and the practical realities of deploying XAI in real-world contexts. Collectively, they reflect the crudeness of the domain and the need for principled and theoretically grounded approaches. The following sections provide a detailed examination of these challenges.

2.9.1 Methodological Limitations

Beyond conceptual ambiguity, the specific methods employed in XAI often suffer from severe technical limitations. Surrogate model techniques, such as LIME and Anchors, approximate local decision boundaries using interpretable models, including linear classifiers and decision trees. However, these approximations can have low fidelity in non-linear or high-dimensional spaces, misrepresenting the original decision logic of the model (Thibault Laugel, 2020).

Gradient-based saliency maps, such as those generated via Grad-CAM, are widely used in image and text models; however, they are sensitive to input noise and adversarial perturbations. Empirical studies have shown that randomly initialised networks can produce saliency maps similar to those of trained models, casting doubt on their utility as explanations (Adebayo et al., 2018; Kindermans et al., 2018).

Feature attribution methods, including SHAP and Integrated Gradients, assume additive feature contributions; however, they are often undermined by feature collinearity or causal ambiguity. These approaches risk attributing importance to features that are merely correlated with causal drivers, thus providing misleading insights (Hooker et al., 2019).

Counterfactual explanation methods aim to identify the minimal changes required to alter model decisions. While promising, they frequently generate unrealistic or infeasible examples that lie off the data manifold, which diminishes their practical interpretability.

Finally, attention-based methods, while intuitive, have been criticised for combining attention weights with explanatory relevance, despite evidence that attention does not always correlate with model outputs (Jain and Wallace, 2019; Serrano and Smith, 2020).

2.9.2 Performance–Explainability Trade-off

A fundamental obstacle in XAI is the apparent trade-off between model explainability and predictive performance. High-performing models, particularly those based on deep learning, typically rely on complex, non-linear representations that defy intuitive understanding. In contrast, models such as logistic regression or decision trees are inherently interpretable but often lack the representational capacity required for tasks in vision, language, and genomics.

This trade-off has been particularly salient in high-stakes domains such as medicine and finance. For example, CNN have achieved high accuracy in radiological diagnosis but is opaque to clinicians, undermining trust and regulatory compliance (Caruana et al., 2015). Similarly, credit-scoring algorithms based on ensemble methods may outperform traditional scorecards, but they also raise concerns about fairness and explainability.

Emerging research seeks to mitigate this trade-off by developing models that strike a balance between explainability and performance. GAMs and EBM use additive structure and monotonic constraints to ensure transparency while capturing non-linear effects. Neural-

symbolic systems (J. Zhang et al., 2021) integrate DL with logical reasoning, enabling traceable inference processes. However, these approaches are still developing and often require domain-specific tuning.

2.9.3 Faithfulness vs. Plausibility

Another pervasive issue in explainability is the tension between generating faithful explanations —those that accurately reflect internal model computations and plausible ones that are intelligible and satisfying to human users. Faithfulness is crucial for technical transparency, yet explanations optimised for human comprehension often sacrifice this in favour of simplicity or coherence (Jacovi and Goldberg, 2020).

An example is the use of saliency maps in image classification. While these heatmaps may visually highlight regions of interest, they are frequently unfaithful to the actual reasoning of the model. Adebayo et al. (2018) demonstrated that saliency methods produce virtually identical outputs even for untrained or randomised networks, indicating that these explanations are artefacts of input structure rather than actual model reasoning. This phenomenon is not limited to vision; in natural language processing, attention heatmaps often fail to align accurately with attention heads that influence outputs (Wiegreffe and Pinter, 2019).

This disconnect poses significant epistemic and practical risks. Users may accept plausible yet incorrect explanations, leading to misplaced trust or faulty decisions. To address this, some researchers advocate hybrid approaches that explicitly balance faithfulness and interpretability, such as using concept bottlenecks (Koh et al., 2020) or causal constraints (Pearl, 2009).

2.9.4 Bias Amplification and Adversarial Explanations

The deployment of XAI systems in real-world applications has surfaced concerns about the amplification of biases and the vulnerability of explanation mechanisms to adversarial manipulation. Explanations often reflect underlying data patterns, and biased training data can result in explanations that rationalise discriminatory or unfair decisions. Mehrabi et al. (2021) have demonstrated that models trained on biased datasets not only propagate harmful

stereotypes but also generate explanations that obscure or justify these patterns, thereby compounding the problem.

Ghorbani et al. (2019) introduced the concept of adversarial explanations, where small perturbations to input data can cause significant shifts in the resulting explanations, without affecting model predictions. This raises the possibility of intentionally manipulating explanations to conceal biases, introduce misleading rationales, or fabricate a false sense of fairness. Slack et al. (2020) demonstrated that models could be trained to appear fair in explanations while being discriminatory in operation, which is a serious concern for auditability and compliance.

Addressing these issues requires robust training paradigms, fairness-aware explanation methods, and detection mechanisms for identifying and mitigating adversarial manipulations. Approaches such as invariant risk minimisation (Arjovsky et al., 2020) represent promising directions but are computationally demanding and not yet conventional methods.

2.9.5 The Rashomon Effect

Derived from statistical learning theory, the Rashomon Effect describes how multiple, equally plausible yet structurally distinct explanations can explain the same model prediction (Breiman, 2001; Rudin, 2019). This multiplicity complicates the landscape of interpretability, particularly in parameterised models where many decision paths can lead to the same output.

From a practical perspective, this multiplicity undermines the trustworthiness of explanations. Users are left uncertain as to which explanation, if any, reflects the "true" rationale behind the decision of the model. In high-stakes scenarios, such as legal adjudication or autonomous vehicle operations, this ambiguity poses significant risks.

Efforts to mitigate the Rashomon Effect include the use of causal inference frameworks to constrain the space of valid explanations (Pearl, 2009) and ensemble explanation strategies that aggregate across multiple models or runs (Thibault Laugel, 2020). However, these solutions are not universal. Causal models require strong assumptions and domain expertise, whereas ensemble methods may introduce additional complexity and compromise interpretability.

Overall, these challenges underscore that explainability is not a mere add-on to ML pipelines, but a foundational requirement that intersects with model design, training data, evaluation, and human factors. Advancing the field will require multidisciplinary collaboration, principled frameworks, and emphasis on empirical validation across real-world settings.

2.10 Summary of the Key Findings

The expedited adoption of AI in critical decision-making contexts, including healthcare, finance, autonomous systems, and public policy, has precipitated an urgent demand for transparent, accountable, and interpretable AI. This chapter undertakes a comprehensive examination of XAI, encompassing its conceptual foundations, evolving taxonomy, prevailing methodological paradigms, evaluation frameworks, and enduring challenges. From this comprehensive synthesis, it becomes evident that explainability is not merely an ancillary feature of ML systems, but rather a fundamental prerequisite, epistemologically, ethically, and practically, for establishing trustworthy and responsible AI.

One of the central tensions in XAI resides in the ambiguity surrounding its foundational terminology. Terms such as “interpretability” and “explanation” are frequently employed interchangeably, despite possessing distinct semantic and operational implications. This semantic fluidity obstructs the process of consensus-building, reproducibility, and standardised benchmarking. The lack of a unifying framework constitutes a significant impediment to the advancement of the field. Efforts to address this through taxonomies and benchmarking initiatives show promise; however, their acceptance across various domains and use cases remains limited. In the absence of shared standards, comparisons between methodologies become anecdotal rather than principled, thereby hindering the cumulative advancement of XAI.

The complexity is further compounded by the methodological limitations inherent in contemporary XAI approaches. Although local surrogate models, gradient-based visualisations, attribution methods, counterfactuals, and attention mechanisms each offer valuable insights into model behaviour, none can be deemed universally reliable or robust. Many of these methods fail to generalise beyond narrow experimental frameworks, exhibit fragility when subjected to adversarial perturbations, or render explanations that, while plausible, ultimately do not faithfully represent the underlying model logic. This inconsistency

not only raises doubts regarding their utility but also creates opportunities for manipulation and adversarial misuse, thereby challenging the notion that explanations are inherently stabilising or trustworthy.

Moreover, the longstanding trade-off between explainability and predictive performance remains a structural dilemma. DNN achieve state-of-the-art results across domains, but often at the cost of human comprehensibility. While interpretable-by-design models such as GAMs or EBMs offer a middle ground, their scope and applicability remain domain-bound and data-sensitive. Emerging neural-symbolic hybrids and concept-based models promise greater integration of logic and learning, yet they too demand rigorous empirical validation and precise theoretical articulation.

Most concerning are the socio-technical risks posed by explainability methods themselves. Explanations can be gamed, manipulated, or weaponised to mask algorithmic bias, justify unfair decisions, or simulate regulatory compliance. The dual-use nature of XAI implies that it must be developed with security, fairness, and adversarial resilience in mind. This further raises concerns about the assumption that explanations are always beneficial. Explanations must be accurate, faithful, and useful to the target audience, often requiring a delicate balance between technical transparency and cognitive plausibility.

The Rashomon Effect underscores this complexity, revealing that many models admit multiple, equally valid explanations for a single prediction. This undermines any simplistic notion of a single "true" explanation and highlights the epistemic uncertainty inherent in high-capacity models. While causal inference and ensemble explanation approaches offer pathways to manage this multiplicity, they require deeper engagement with domain knowledge, experimental design, and philosophical perspectives on causality and inference.

Taken together, these insights suggest that explainability cannot be retrofitted into AI systems as an afterthought. Instead, it must be integral to model architecture, data curation, evaluation protocols, and deployment pipelines. This necessitates a rethinking of the entire AI lifecycle, from data collection and feature engineering to training, inference, and human–AI interaction. Crucially, explainability research must embrace interdisciplinary collaboration, drawing on computer science, human–computer interaction, legal theory, and

ethics. Only by embedding explainability within these broader epistemic and societal contexts can the field realise its promise.

Future research must thus aim at (i) formalising and standardising explanation goals across use-cases, (ii) creating robust explanation methods, and (iii) grounding evaluations in human-centred design and real-world deployment feedback. Explainability is not simply about making models understandable; it is about enabling accountable, fair, and informed decision-making in an increasingly opaque algorithmic system. The path forward lies not in searching for a universal solution, but in building a pluralistic, context-sensitive ecosystem of techniques, metrics, and theoretical frameworks that together constitute a robust science of explanation in AI.

3 Dataset

This chapter aims to comprehensively analyse the data utilised for the proposed research, specifically focusing on neuroimaging techniques. Identifying even the most minor changes in brain atrophy is essential to conducting an in-depth examination and developing models. Selecting the correct type of neuroimaging is essential. It plays a key role in early AD detection and diagnosis, with various techniques offering unique insights into brain structure and function.

Few of these methods help identify early signs and stages of AD. Magnetic Resonance Imaging (MRI) employs strong magnets and radio waves to generate highly detailed brain images. MRI is particularly effective in identifying structural abnormalities such as brain atrophy, stroke damage, tumours, and fluid accumulation. In the context of Alzheimer's research, MRI is invaluable for quantitatively characterising the disease progression (Mcevoy et al., 2009). Quantitative measurements from MRI images can track the subtle changes in brain structures over time, providing a clear picture of Alzheimer's disease and its impacts on the brain.

There are two types of MRI techniques. Functional MRI, or fMRI, is a primary technique used to study brain function by measuring changes in blood flow. fMRI scans of individuals with AD frequently reveal reduced brain activity in certain regions, suggesting a deterioration in neuronal function (Dennis and Thompson, 2014). Resting-state fMRI has emerged as a potential diagnostic tool for identifying functional brain alterations in the initial phases of AD.

Secondly, structural MRI (sMRI) is essential for this study as it enables the detailed visualisation of brain regions impacted by Alzheimer's disease. By capturing changes in the brain structure and shape, sMRI can help in the early detection of atrophy, which is crucial for diagnosing and understanding the progression of Alzheimer's (Vemuri and Jack, 2010). Moreover, advances in DL have enabled the classification of MRI images at a level comparable with the performance of expert radiologists. Advancements in DL enhance the diagnostic value of MRI, making it a powerful tool in both clinical and research settings.

Since the primary aim of this thesis is to detect patterns arising from alterations in brain structure and morphology, sMRI is the optimal neuroimaging modality for this research.

Its capacity to offer complex brain morphology images makes it perfect for examining the slight alterations linked with AD. The information gathered from sMRI will serve as the basis for the models created in this study, enabling precise categorisation and forecasting of Alzheimer's advancement.

3.1 Sources of the data

A meticulously chosen and diverse collection of publicly accessible datasets, each offering a distinctive perspective on the research framework, has been thoughtfully put together. These datasets have been meticulously chosen based on their relevance to research objectives and comprehensive representation of patients with varying cognitive conditions and demographics.

The Alzheimer's Disease Neuroimaging Initiative (ADNI) plays a crucial role in neuroimaging research (ADNI Database, 2021). ADNI is highly valued for its extensive collection of MRI scans, clinical assessments, and genetic data from individuals diagnosed with AD and Cognitively Normal (CN) controls. The primary benefit of ADNI lies in its longitudinal design, enabling the observation and examination of the progression of cognitive decline in the early stages of AD.

The AIBL study, known as the Australian Imaging, Biomarker, and Lifestyle Flagship Study of Ageing, offers a vast and invaluable data collection (AIBL Database, 2021). This data includes information from individuals who have undergone cognitive assessments, MRI scans, and biomarker measurements. The dataset is a comprehensive resource that encompasses individuals at various stages of cognitive deterioration and those who remain cognitively healthy. Consequently, it is a crucial framework for studying the development of AD-associated changes and the various factors that influence cognitive ageing. Adding AIBL to the research plan significantly enhances the comprehension of the complicated mechanisms linked to AD and cognitive deterioration.

In contrast, the Information eXtraction from Images (IXI) dataset provides a unique perspective on the field of research (IXI Database, 2021). This dataset contains a variety of MRI scans, from a range of Cognitively Normal (CN) individuals of different ages. Encompassing diverse subjects in the dataset boosts the research by enabling evaluations

between AD and CN. This analysis reveals the distinctive patterns and characteristics associated with each scenario.

Research objectives have guided the careful and strategic selection of datasets to investigate the patterns and biomarkers associated with AD. The ADNI, IXI, and AIBL datasets provide a comprehensive and holistic approach to navigating neuroimaging data. This deliberate selection makes it possible to examine various sources of information, offering a comprehensive and multidimensional understanding of AD and its correlation to other NDDs. Employing this extensive approach can help discover ground-breaking insights and make substantial contributions to neurodegenerative research.

3.2 Further information regarding the MRI scans.

In neuroimaging, acquiring sMRI scans is crucial for understanding the human brain. These scans provide a detailed view of the brain anatomy, helping to identify and characterise various structures. The choice of MRI weights, particularly T1 and T2, is a crucial component of sMRI scanning. This decision is far from arbitrary, as it fundamentally affects the information derived from the scans and shapes the goals of subsequent analyses. This thesis focuses on examining and interpreting brain structures, making the choice of T1 weighting of essential significance.

T1 and T2 are two distinct MRI weights that represent various aspects of the composition of the brain. T1-weighted images mainly highlight the existence of fat in tissues, particularly in the context of brain structures. They excel at emphasising differences in various brain tissues, predominantly white and grey matter. On the other hand, T2-weighted images enhance the detection of fat and water content in tissues, such as Cerebro-Spinal Fluid (CSF) (MacKay et al., 2006). T2-weighting is often employed to investigate fluid-related issues in the brain, such as identifying lesions, tumours, or abnormalities in CSF flow. Figure 3- 1**Error! Reference source not found.** shows two sequences in sMRI scans, modified from (Atia et al., 2022).

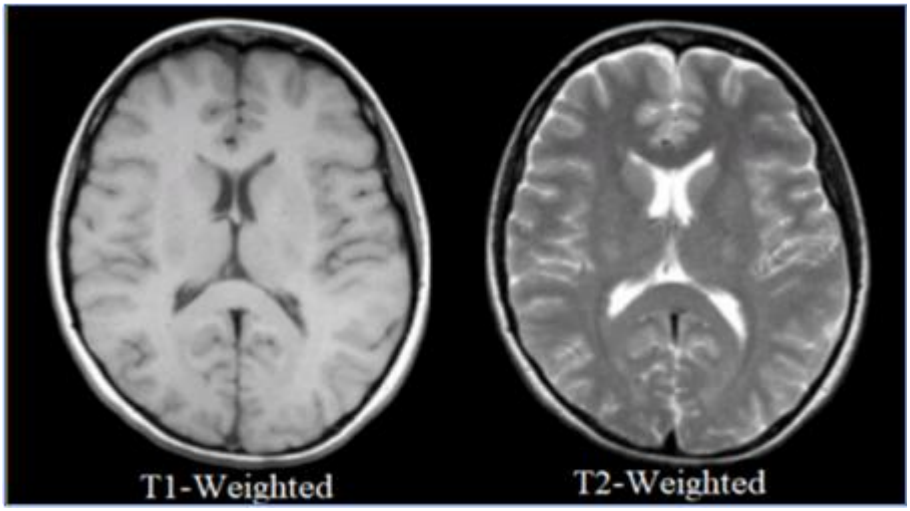

Figure 3- 1 Two sequences in sMRI scans, modified from (Atia et al., 2022)

White and grey matter are two fundamental constituents of the brain, each possessing distinct functions and characteristics. White matter consists of axons responsible for transmitting signals between different brain regions, while grey matter primarily comprises cell bodies and is integral to various cognitive functions (Mercadante and Tadi, 2020). Accurately examining these structures is essential for understanding neurological conditions, cognitive functions, and the overall operation of the brain.

T1-weighted sMRI scans outline the borders between white and grey matter, delivering exceptional detail and contrast. This increased awareness of fat levels in tissues is particularly beneficial for evaluating the structural integrity of the brain and identifying subtle changes that may occur in diseases such as Alzheimer's or multiple sclerosis (Marcisz and Polanska, 2023). By using T1 weighting, this thesis aligns with a specific focus on white and grey matter analysis, enabling a rigorous examination of the structural properties of the brain as well as any potential alterations or abnormalities.

In sMRI scanning, choosing the MRI weighting is crucial for guiding future research. This thesis strategically selects T1 weighting over T2 for a detailed analysis of brain structures, primarily focusing on white and grey matter. This choice highlights the need for accuracy and sensitivity in understanding the complexities of the human brain and provides insights into neurological function and dysfunction.

3.3 FreeSurfer and Its Processing

This thesis examines the complex field of structural neuroimaging, incorporating quantitative measurements as a fundamental aspect of analysis. These measurements, such as thickness, volume, and area, are essential for understanding the human brain. Researchers conduct a comprehensive examination using high-resolution three-dimensional brain scans, also known as sMRI data. The well-known neuroimaging program FreeSurfer v.6.0 (Fischl, 2012) facilitates an advanced processing pipeline for retrieving these numerical measurements.

The investigation aims to enhance understanding of brain structure and its links to neurological disorders and cognitive functions. Using advanced quantitative methods, it seeks to gain deeper insights into brain morphology and its impact on well-being and cognition. The data provides a comprehensive view of the internal structure through three-dimensional scans, but the raw data requires careful preprocessing to enable meaningful numerical measurements.

FreeSurfer version 6.0 serves as the foundation for the preprocessing pipeline. This software package provides tools and algorithms particularly designed for neuroimaging data analysis. The initial stage is image registration, which aligns scans to a standard reference system (Wyawahare et al., 2009). This alignment ensures consistency and compatibility across scans, enabling accurate comparisons of brain structures over time.

After registration, the software proceeds with skull stripping, an essential step that removes non-brain tissues from the images (Kalavathi and Prasath, 2016). This rigorous procedure ensures that all future analyses are exclusively concentrated on the structural components of the brain. After the skull stripping process, clean brain images are ready for further analysis.

Brain segmentation and parcellation are crucial in preprocessing, as they enable the identification and labelling of distinct brain regions with unique anatomical boundaries (Backhausen et al., 2016). FreeSurfer excels at accurately labelling various structures, ranging from the cortex to subcortical areas. This automated parcellation step enables the extraction of region-specific measurements for detailed examination of brain diversity components.

A notable feature of Freesurfer is its ability to estimate cortical measurements, such as surface area, volume, and thickness. These measurements are vital in research, providing

insights into cortical structure and variations. Cortical thickness offers crucial information regarding brain health and developmental changes. Quantifying these metrics enables statistical evaluations, thereby enhancing the depth of analysis.

The extracted numerical measurements are crucial for research. These measurements encompass a broad range of brain regions and characteristics, providing a diverse perspective on brain structure. For example, one may focus on quantifying the thickness of specific regions, such as the right HATA (Hippocampal-amygdala transition area), or the volume of important structures, such as the right Hippocampus. Each measurement represents a numerical value, which not only enables quantitative comparisons but also enables sophisticated statistical analyses.

This study benefits from a comprehensive viewpoint, covering the entire brain and specific areas. This perspective ensures the research captures the full range of brain structure and its implications. It recognises the complexity of the brain and aims to accurately represent this intricacy.

Figure 3- 2 represents the overview of the FreeSurfer steps in extracting the necessary numerical values (Grossner et al., 2018).

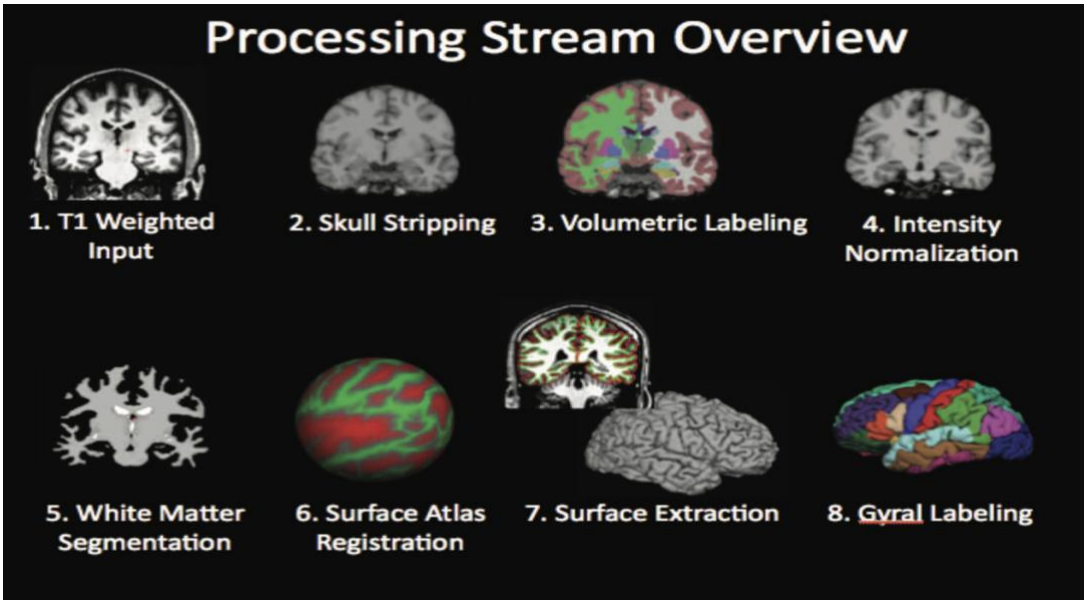

Figure 3- 2 Processing overview of the FreeSurfer program used to extract grey matter volumes (Grossner et al., 2018)

This thesis focuses on carefully extracting and examining numerical measurements from sMRI data. By utilising the advanced preprocessing capabilities of FreeSurfer v.6.0, the

software ensures that these measurements are accurate and reliable. These numerical attributes, such as thickness, volume, and area, play a crucial role in research by providing insight into the intricacies of brain shape and its relationship to neurological well-being, cognitive abilities, and other key factors. The aim is to enhance the growing understanding of the human brain by merging advanced technology with in-depth analysis.

3.4 Post-processing

This thesis develops a critical data selection process utilising three datasets to enhance comprehension and diagnostic capabilities in AD. These datasets provide valuable information, including multiple scans per subject taken at various disease stages, such as screening, baseline, or follow-ups.

The data selection is vital as it significantly influences the thesis goals and results. The research aims to enhance early AD detection, emphasising the importance of identifying the disease at an early stage. This aligns with the healthcare goals of early intervention and treating NDD, potentially improving patient outcomes and quality of life.

Selecting the earliest scan for each subject is essential to represent the initial stages of AD accurately. Clinicians typically conduct these scans when symptoms are mild or even before they appear. By focusing on early scans, the goal is to maximise the detection of subtle brain changes occurring before noticeable clinical signs arise. This approach underscores the importance of early detection and intervention in AD, which is crucial for enhancing patient care outcomes.

An additional criterion is used to enhance data selection during the same stage of gathering information (for example, multiple scans taken on the same date). Clinicians prioritise the scan with the highest Contrast-to-Noise Ratio (CNR), a valuable measure that assesses image quality and clarity, ensuring the most accurate depiction of the complexity of the brain.

After meticulously implementing these selection criteria across all datasets, the next step was to curate a subset of data comprising 3,974 sMRI scans successfully. Each scan corresponds to an individual subject, creating a dataset uniquely tailored to the specific goals and objectives of this thesis.

After undergoing pre-processing, every sMRI scan from subjects provides a comprehensive collection of 446 attributes, which represent numerical assessments of various brain areas. These characteristics comprehensively examine the structural qualities of the brain, including measurements such as thickness, volume, and area. Nevertheless, similar to any data, it is crucial to ensure the quality and precision of the information.

The data cleaning phase removed 45 features from the initial set of 446. These excluded features were either duplicates of other measurements or contained errors. Factors such as head movement during scanning or complexities in the pre-processing steps of FreeSurfer could cause these errors. Removing these irrelevant features is crucial to prevent any distortions or biases in the subsequent analysis, ultimately enhancing the reliability of the dataset.

This data-cleaning process resulted in a well-organised and enhanced dataset using Knime software (Berthold et al., 2009; Sarica et al., 2014). It was presented as a table of 3,974 rows, each representing an individual subject's scan. The table also contained 404 columns representing the features extracted from these scans. These 404 columns comprised 401 unique brain features generated by FreeSurfer v.6.0, in addition to age, gender, and research group, which indicated the subject's disease. These characteristics provided valuable insights into various structural aspects of the brain, such as cortical thickness, subcortical volumes, and other essential measurements. Moreover, the dataset contained age and gender data for each subject, which was crucial for examining the potential impact of these demographic variables on brain structure and the diagnosis of AD.

Research indicates that the ADNI dataset is of great value due to its large number of Alzheimer's disease subjects. Consistency across datasets, particularly in age range, is essential for valid findings and reliable comparisons. The ADNI study includes participants ranging in age from 55 to 90 years. To maintain consistency with ADNI and uphold the integrity of the research, any additional datasets used in this study must also fall within this age range. For example, the IXI dataset, which covers a broader age range from 19 to 90, has been narrowed down to only include individuals aged 55 and above. This careful selection ensures that the dataset used in this thesis remains consistent in terms of age, which is an important factor considering the significant impact of age on brain size and structure.

The post-processing and cleaning stages of the dataset are crucial. Removing duplicate or incorrect features enhances the quality and reliability of the dataset. The resulting dataset, which consists of 3,974 subjects and 404 features, is a valuable resource for investigating the complex structure of the brain and its implications for AD diagnosis. Additionally, the focus on maintaining consistent age ranges across datasets ensures that the research is built on a strong foundation and can provide accurate and meaningful insights into the relationship between brain structure and AD within the specified age range. The flow chart depicted in Figure 3- 3**Error! Reference source not found.** illustrates the procedure for acquiring data sources, processing using Free Surfer, and implementing post-processing steps.

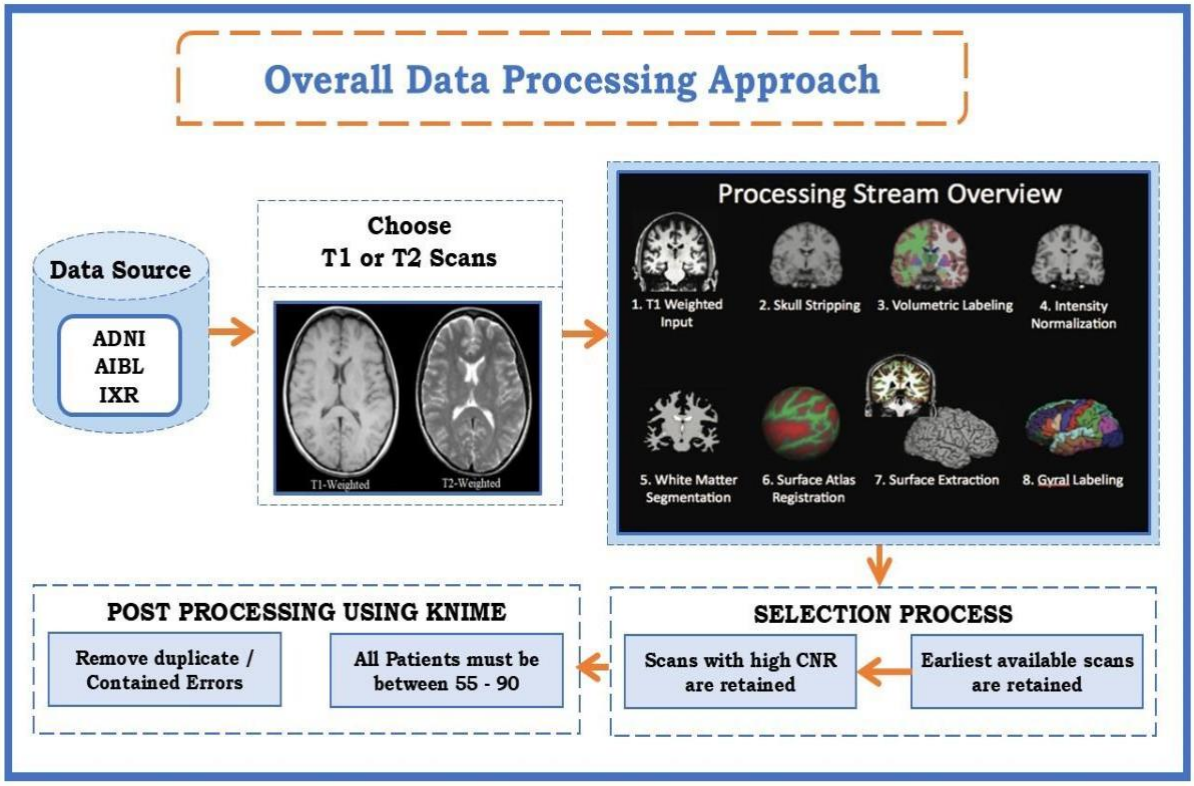

Figure 3- 3 Pipeline depicting the overall data processing steps

3.5 General overview and statistics of the dataset

The data visualisation component of this research is a crucial element aimed at visually representing and analysing the extensive and complex data associated with neurodevelopmental disorders (NDDs). This section uses various graphical and descriptive methods to provide a clear overview of the dataset, enhancing our understanding of the connections, trends, and patterns within the data. These visualisations simplify the

interpretation of the results and enhance the overall coherence and understanding of this study.

The visualisations are organised into three main sections. First, they will provide an overview of the dataset by focusing on general statistics, including the distributions and the ranges of key features. Second, the visualisations will compare each condition, examining their distributions and the ranges of select features. Finally, the visualisations will emphasise insights that can be illustrated through plots, highlighting trends and relationships between variables to enable a clear visual interpretation of data patterns.

3.5.1 Contributions of each data source

The pie chart in Figure 3- 4**Error! Reference source not found.** below clearly represents the distribution of data sources used throughout this research. The ADNI data source is the most significant contributor, accounting for 77.4% of the overall dataset. This dominance reflects a substantial portion of the data, comprising approximately 3,080 patient records. Following ADNI, the second-largest contributor is the AIBL data source, which accounts for 16.73% of the dataset, equating to roughly 670 patient records. At the other end of the spectrum is the IXI data source, contributing the smallest share at 5.86%, with just 230 records.

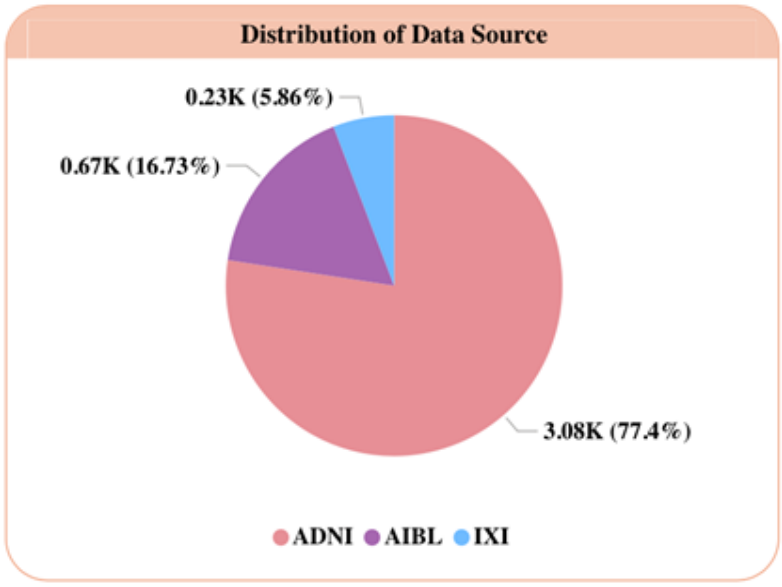

Figure 3- 4 Distribution of the data sources

Despite its smaller size, the IXI dataset is significant because it contains the highest proportion of healthy records, which are essential for further comparative studies. This detailed breakdown provides a comprehensive view of the distribution and significance of each data source, highlighting not only their sizes but also their unique contributions to the overall analysis.

3.5.2 Distribution of different genders among each of the data sources

The pie charts in Figure 3- 5 below clearly visualise the gender distribution across the different data sources. In the most extensive dataset, ADNI, males comprise 50.99%, corresponding to approximately 1,570 records. Females account for 49.01%, representing around 1,510 records. In the next largest dataset, AIBL, the gender distribution shifts slightly, with females comprising 55.79% (approximately 371 records) and males comprising 44.21% (roughly 294 records). The smallest dataset, IXI, also shows a higher proportion of females, with 61.73%, around 143 records, while males account for 38.63% or roughly 90 records.

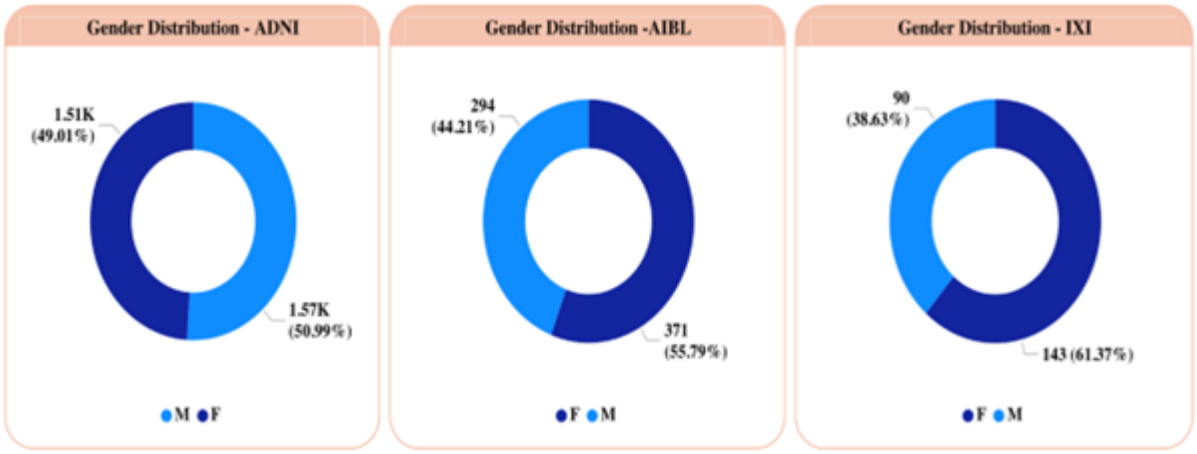

Figure 3- 5 Gender distribution for each different data source

Overall, the gender distribution is generally adequate across the datasets, although some variations exist, particularly in AIBL and IXI, where females slightly outnumber males. Sampling techniques are necessary to ensure a balanced representation in future studies that focus on gender-specific factors. However, since this research primarily focuses on disease analysis rather than gender, the current gender distribution remains sufficient and does not require further adjustments.

3.5.3 Distribution of healthy and multiple diseases within each data source

The tree graphs in Figure 3- 6 below depict the distribution of various diseases across the different data sources. For the ADNI dataset, CN patients form the largest category, comprising approximately 1,360 records. Following that, AD represents about 840 records, while MCI contributes around 360 records. Close behind is the EMCI category, with about 340 records. The smallest proportion within the ADNI dataset is LMCI, which has around 180 records. In the AIBL dataset, CN patients comprise the most significant proportion, with 484 records. The next largest group is MCI, contributing 102 records; finally, AD accounts for 79 records. The IXI dataset is distinct because it exclusively contains CN patients, with 233 records, and does not represent neurodegenerative diseases such as AD, MCI, Early MCI, or Late MCI.

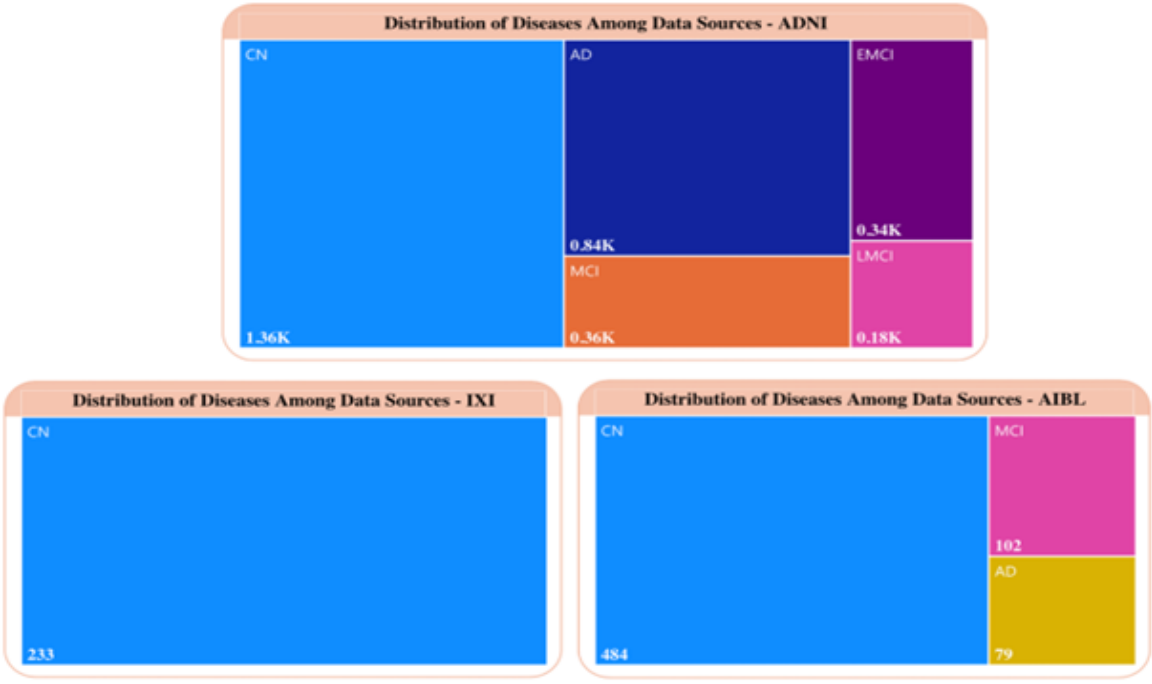

Figure 3- 6 Distribution of healthy and various diseases within each data source

Overall, the distribution across the datasets shows a relatively balanced representation between cognitively normal individuals and those affected by various stages of neurodegenerative diseases, except the IXI dataset, which only includes healthy participants. This distribution offers a valuable overview of disease and cognitive state representation, helping to shape the research scope and focus based on contributions from each dataset.

3.5.4 The average age of data-subjects included in the data source

Figure 3- 7 Visualises the mean age across the various data sources, providing valuable demographic context for this research. Starting with the AIBL dataset, the average age is 75 years, reflecting a relatively older population, which is significant when considering age-related factors in neurodegenerative diseases. In the ADNI cohort, the average age is slightly younger, at 74 years. However, it still represents an older demographic that is typical for studies focused on conditions such as Alzheimer's disease and other age-associated disorders. In contrast, the IXI cohort presents a notably younger average age of 65 years, suggesting a relatively youthful group of participants compared to the other datasets.

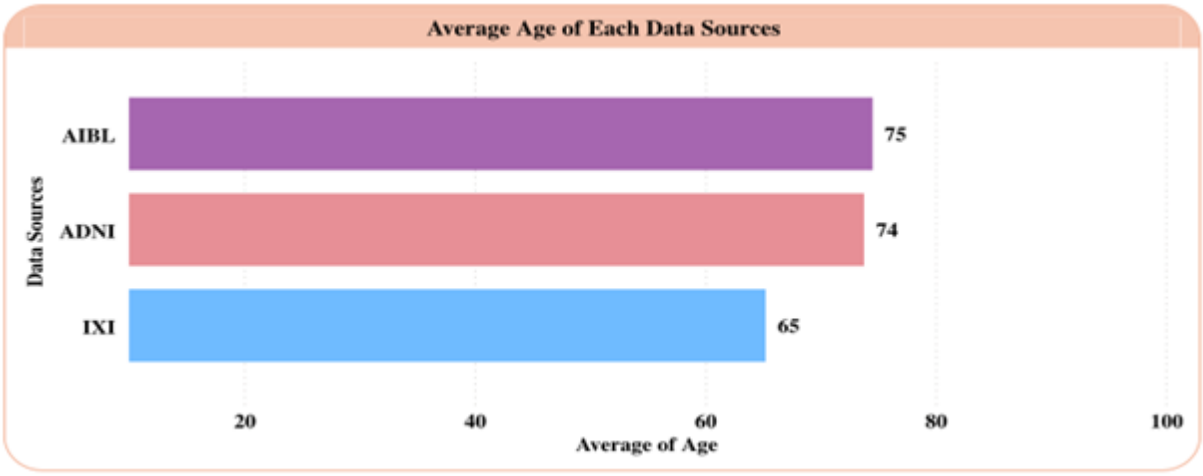

Figure 3- 7 Average age of each data source

This younger demographic could be particularly useful for comparative studies, particularly when examining early-stage disease markers or drawing comparisons between younger, healthier individuals and older populations prone to cognitive decline. This demographic breakdown underscores the importance of factoring in age when conducting research across these datasets. Age plays a critical role in the onset and progression of neurodegenerative diseases, and the variation in average ages across these cohorts may have meaningful implications for both the analysis and interpretation of the results. By considering these age differences, researchers can understand the role of ageing in the data, potentially leading to tailored and accurate conclusions.

3.5.5 Distribution of all the diseases

The bar graph in Figure 3- 8 provides a detailed visualisation of the distribution of neurodegenerative diseases and the healthy control group within the dataset under examination. The CN category holds the largest share, with 2,198 patient records, making up approximately 48.25% of the dataset. The high proportion of healthy individuals in the data is a critical reference point for comparative analyses with neurodegenerative conditions. The second-largest category is AD, comprising 921 patient records, which accounts for roughly 20.22% of the dataset. The dataset distribution emphasises the prominence of Alzheimer's patients, focusing on studying neurodegenerative diseases. MCI is represented by 461 patients, making up around 10.12% of the dataset. This classification reflects individuals experiencing cognitive decline that does not yet meet the threshold for a diagnosis of Alzheimer's disease or another severe neurodegenerative condition. The dataset includes 335 records under the EMCI category, representing an earlier stage of cognitive impairment, which is crucial for tracking disease progression. On the other hand, the LMCI group forms the most minor proportion, with 179 records, accounting for 3.39% of the dataset. LMCI typically signifies an advanced stage of impairment, often preceding full-blown Alzheimer's disease or other severe conditions.

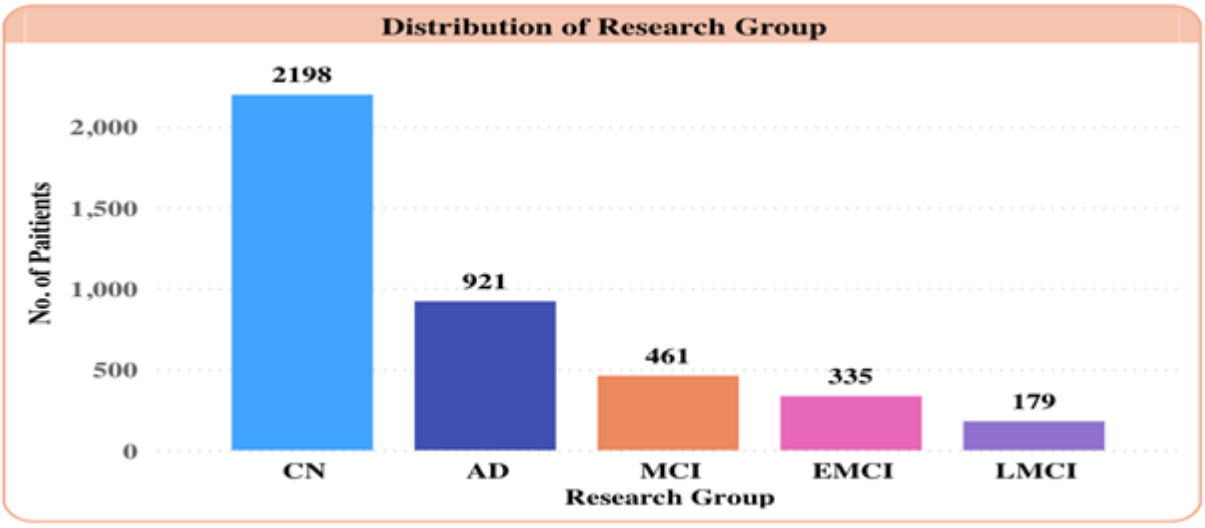

Figure 3- 8 Distribution of Diseases within the whole dataset

This comprehensive distribution clearly explains the prevalence of different neurodegenerative stages within the dataset and offers insight into the balance between the healthy control group and the various cognitive states. Such a breakdown is essential for

researchers, as it enables nuanced analysis of disease progression and comparisons between healthy and affected individuals, ultimately enhancing the depth and precision of the research findings.

3.5.6 Gender distribution among the diseases

The investigation into the gender distribution across various categories of neurodegenerative diseases reveals a relatively balanced representation between males and females, with some variation across different conditions, as presented in Figure 3- 9**Error! Reference source not found.**. For the CN category, females comprise 54.64% of the group, with approximately 1,200 patient records, while males account for 45.36%, with around 1,000 records. This slight female dominance in the healthy control group is notable. In the AD category, the gender distribution shifts slightly, with males representing 51.79% of 477 records, while females make up 48.21% of roughly 444 records. This minor discrepancy suggests a near-equal representation of genders among Alzheimer's patients. The MCI category exhibits a pronounced gender difference, with males comprising 60.74% of approximately 280 records, while females account for 39.26% of around 181 records. The data indicate a higher representation of males in the MCI stage of cognitive decline. In the EMCI category, males constitute 54.33%, about 182 records, while females are slightly fewer, at 45.67%, around 153 records. Similarly, for LMCI, males represent 55.62%, approximately 99 records, while females account for 44.38%, roughly 79.

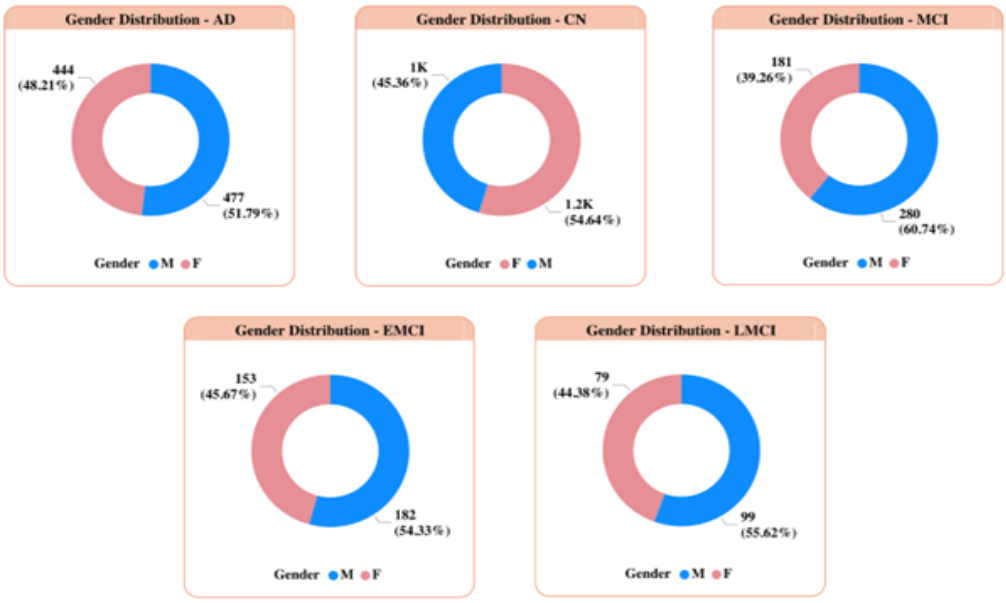

Figure 3- 9 Gender distribution within each disease

This detailed breakdown of gender distribution across neurodegenerative disease categories provides valuable insights into the subtle discrepancies between male and female involvement. Understanding these variations is crucial for interpreting demographic trends within the dataset, which can significantly influence research outcomes, particularly in studies examining gender-specific risk factors or disease progression pathways.

3.5.7 The average age of each instance of disease places the progression stages of AD

The line plot in Figure 3- 10 illustrates the average age across cognitively normal individuals and those affected by various neurodegenerative diseases, offering crucial demographic insights for the research. The data shows the youngest average age is in the EMCI category, with an average age of 71.2 years. Following that, the LMCI group has an average age of 72.3 years, slightly higher than EMCI. The CN group sits just above that, with an average age of 72.5 years. The MCI group shows a higher average age of 74.7 years, while the AD category shows the highest average age of 75.0 years. This gradual increase in age from EMCI through to AD reflects a logical and real-life progression of cognitive decline. Alzheimer's, typically diagnosed in individuals over the age of 60, often takes several years to fully manifest severe symptoms, while earlier stages, such as EMCI, can appear much sooner.

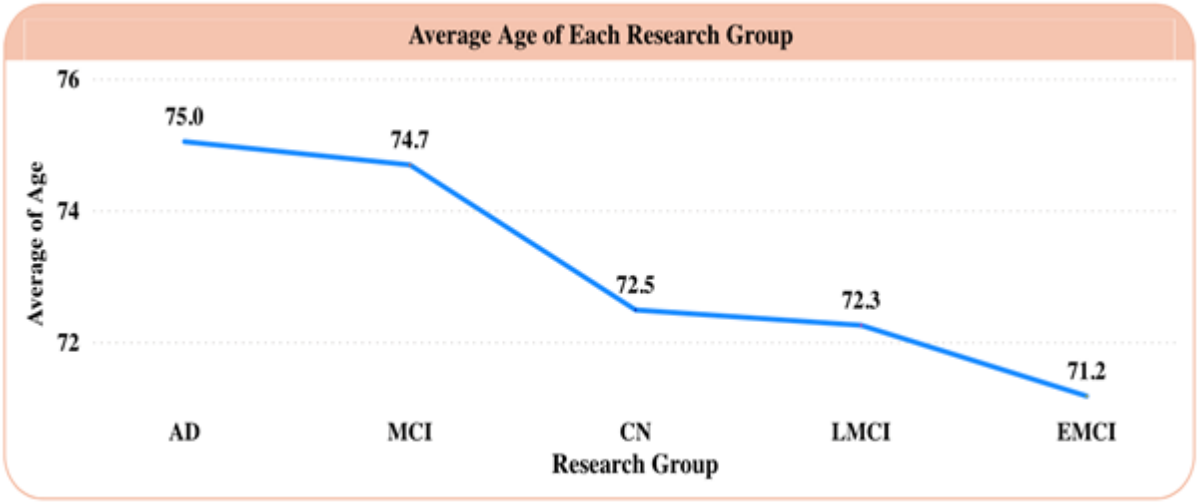

Figure 3- 10 Average age of each disease

This demographic breakdown is essential because it highlights the significance of age-related factors in the progression of neurodegenerative diseases. By accounting for these age differences, the research can accurately assess disease development, symptom onset, and progression patterns, ensuring that age-related trends are factored into the analysis of the study cohorts.

3.5.8 Types of data attributes

The doughnut chart in Figure 3- 11 visually shows the allocation of variables in the dataset, highlighting their distribution. "Area" is slightly dominant, accounting for 69 attributes, underscoring its significance for early AD diagnosis research analysis. Following closely behind is "Meancurv", making up 68% of the dataset, indicating its importance in analysis. Volume, Thickness, and Thickness standard deviation each contribute 68%, emphasising their critical roles. The remaining variables, Volume and Misc., account for 60% of the dataset. Although smaller, their contribution helps to provide a comprehensive understanding of the dataset.

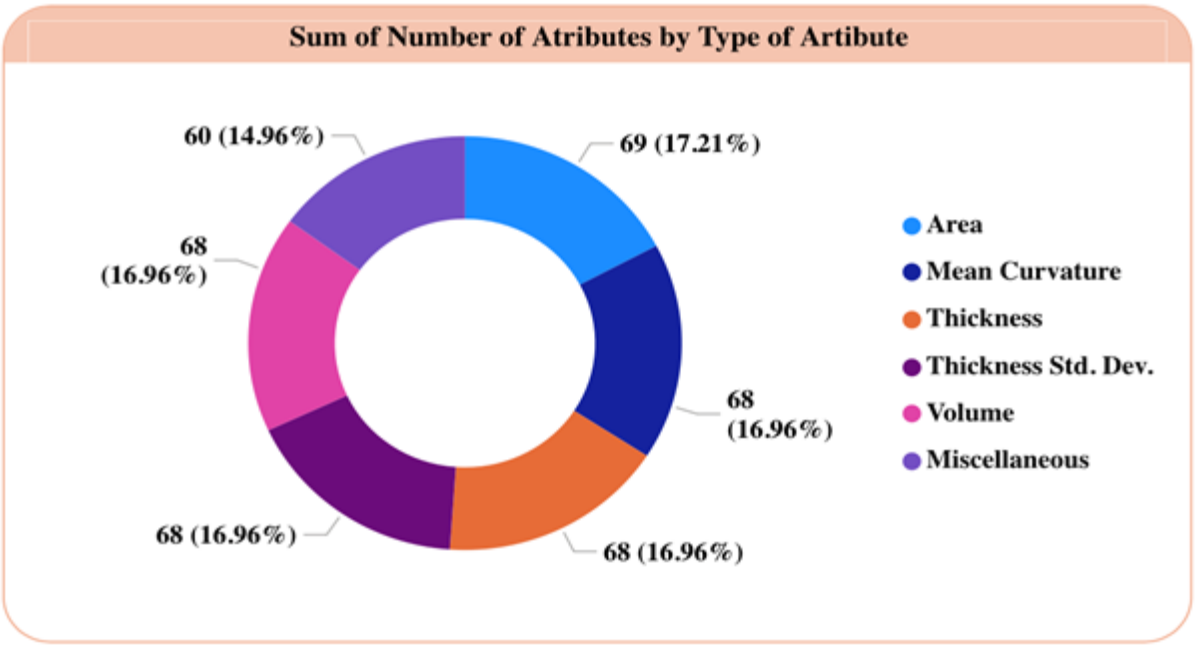

Figure 3- 11 Distribution of type of attributes within the dataset

This doughnut chart clearly and immediately visualises how different variables are distributed and meticulously breaks down their significance within the dataset. Highlighting the prevalence of key variables such as area, mean curvature, thickness, and volume enables researchers to quickly grasp the relative importance of each factor, facilitating a focused and informed analysis in the research process.

3.5.9 Cortex Volume

The line graph in Figure 3- 12 illustrates changes in cortex volume by age and gender across four neurodegenerative conditions: AD, MCI, EMCI, and LMCI. A key observation is the consistent downward trend in cortex volume for both genders, indicating reduced brain volume as the diseases progress. This decline is most evident in EMCI and LMCI, where the drop is visually striking. The MCI plot shows a significant reduction with a less sharp slope, while the AD plot reveals a gradual decline over time, though still substantial. Another notable feature is the difference in male and female trajectories, with females consistently having lower cortex volumes. The data suggest that females, on average, start with a smaller cortex volume than males, whose starting points are higher in each disease category, possibly reflecting gender-related differences in brain anatomy or progression rates.

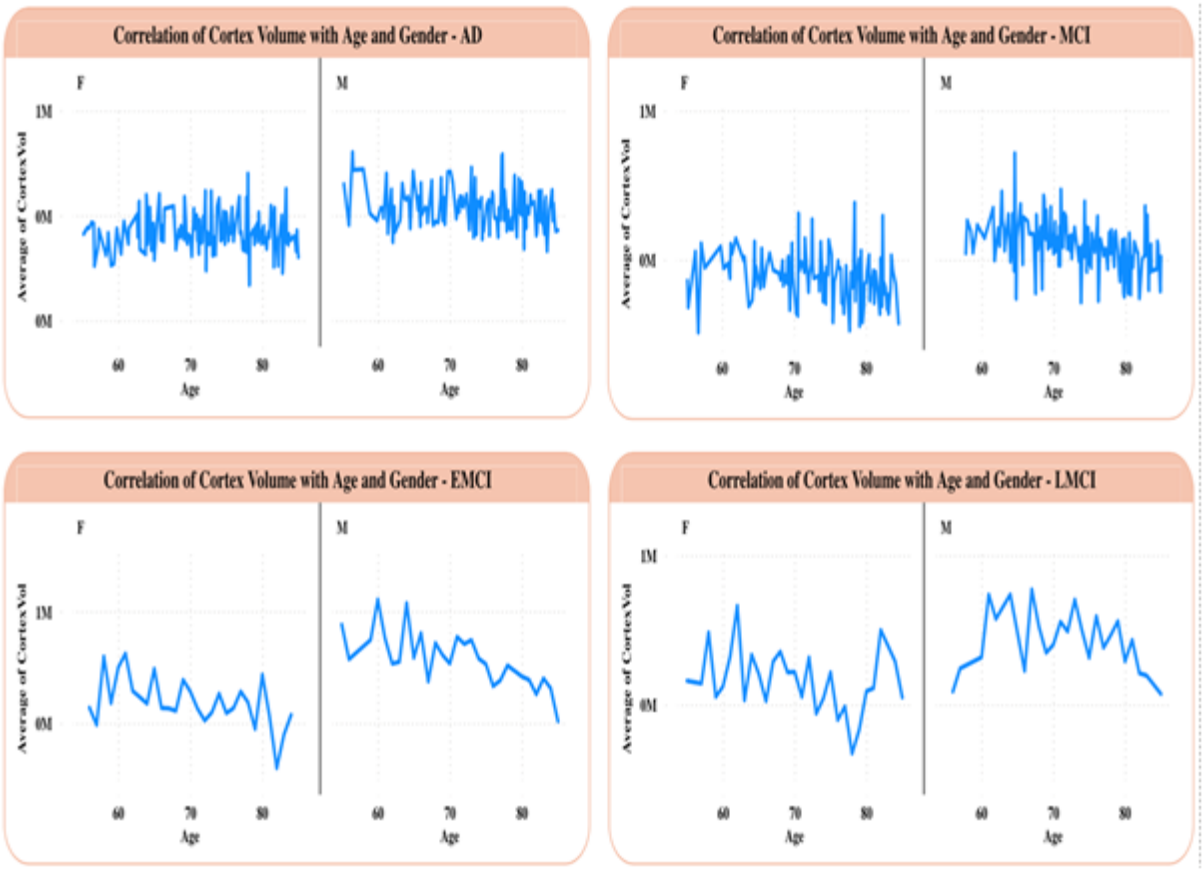

Figure 3- 12 Correlation of Cortex Volume of the brain with Age and Genders for four different Diseases

This visualisation aligns with real-world observations of NDD progression. Cortex volume tends to shrink as cognitive decline advances, and in the case of Alzheimer's Disease, cortical atrophy, particularly in regions responsible for spatial reasoning and visual processing, can profoundly affect an individual's ability to interpret spatial and visual information. The decline in cortex volume observed in these graphs underscores the importance of monitoring cortical changes in patients as a critical marker of disease progression, and it highlights the potential gender differences in how these diseases impact the brain.

3.5.10 Amygdala

The line graphs in Figure 3- 13 below present the changes in the left and right amygdala as a correlation of age and gender across four neurodegenerative conditions: AD, MCI, EMCI, and LMCI. A key observation from the graph is the consistent downward trend in the amygdala, both right and left, for all genders across all four conditions, indicating a reduction in this brain region as the diseases progress.

The decline is most evident in the EMCI and LMCI plots, where the sharp drop in amygdala size indicates significant loss of brain tissue early in disease progression. The MCI plot shows a gradual reduction, reflecting ongoing atrophy at this stage. In contrast, the AD plot demonstrates a milder decline, particularly in females, while males experience a pronounced loss of amygdala volume. The data suggests the most substantial loss of amygdala volume occurs early in Alzheimer's disease rather than post-diagnosis. Another feature is the difference in male and female trajectories. In all conditions, female amygdala volume is slightly lower than that of males, consistent with earlier observations of gender differences in brain anatomy. Male patients exhibit a notable difference between the left and right amygdalae, particularly in the EMCI and LMCI stages. This data suggests a potential gender-based asymmetry in brain degeneration, with differing effects on the right and left sides in males.

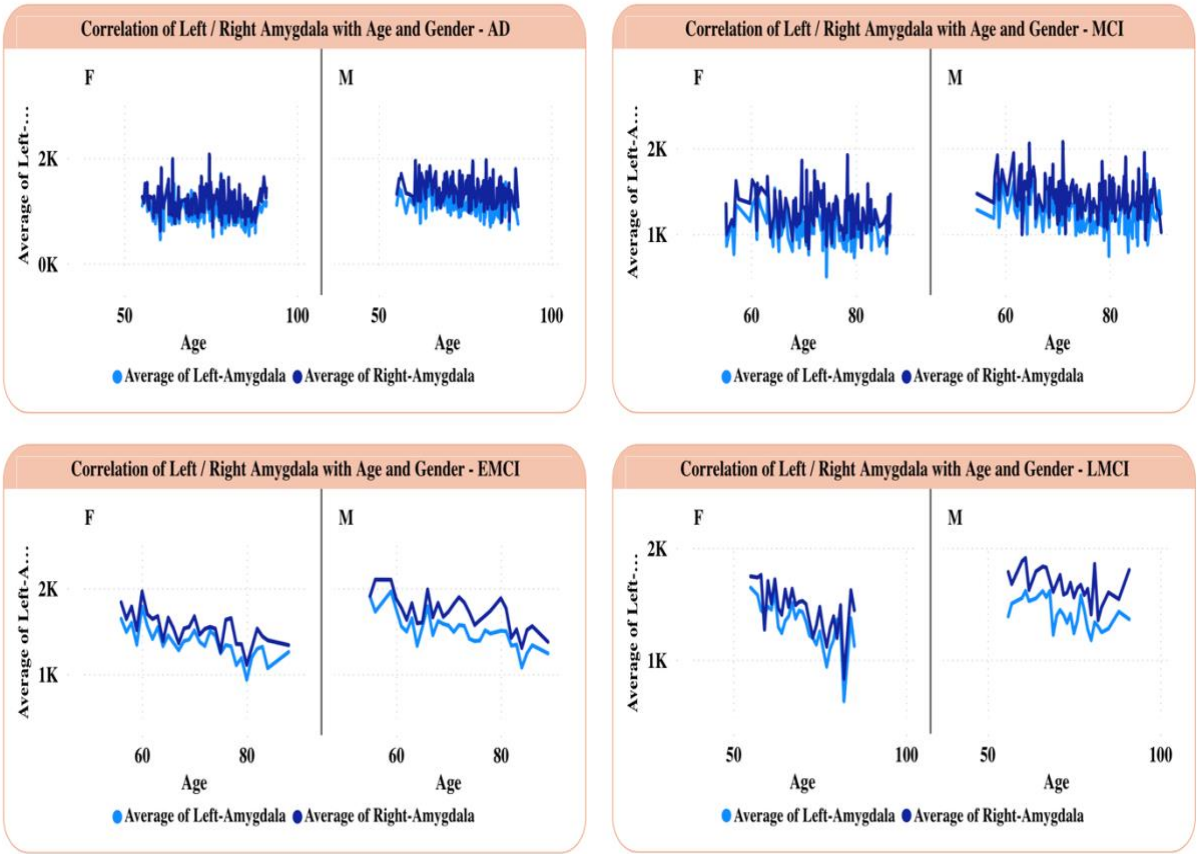

Figure 3- 13 Correlation of Left and Right Amygdala of the brain with Age and Genders for four different Diseases

This visualisation depicts real-world patterns of NDD progression. The amygdala, a structure located in the temporal lobe, plays a critical role in emotional processing, memory,

and responses to stimuli. In patients with Alzheimer's and related conditions, the amygdala tends to shrink as cognitive decline advances. The line graphs reveal that much of the amygdala volume loss occurs during the early stages, particularly in EMCI and LMCI, emphasising the importance of early detection and monitoring. This data underscores the necessity of tracking amygdala changes as a marker of disease progression while also highlighting potential gender differences in how these neurodegenerative conditions impact brain anatomy.

3.5.11 Whole Hippocampus

The graphs in Figure 3- 14 below present the changes in the left and right hippocampus related to age and gender across four neurodegenerative conditions: AD, MCI, EMCI, and LMCI. A key observation from these graphs is the consistent downward trend in hippocampal volume, both right and left, across all genders and conditions, signalling a reduction in this critical brain region as the diseases progress.

The EMCI and LMCI stages show significant declines in hippocampal size, indicating their importance in understanding disease progression. The MCI plot reveals noticeable hippocampal shrinkage, with male patients experiencing a gradual decline. In contrast, female MCI patients show a sharper reduction, indicating a faster atrophy rate. The AD plot shows a slower decline in hippocampal volume, particularly in males, while females exhibit a noticeable decrease. Gender differences in hippocampal atrophy are evident in advanced stages such as MCI and AD, but less so in earlier EMCI and LMCI stages.

The hippocampal decline reflects the progression of real-world NDD. However, the hippocampus shrinks with age, and conditions such as Alzheimer's speed up this volume loss. Neuron loss correlates closely with hippocampal atrophy than with tau protein accumulation or other markers. Line graphs show significant shrinkage during the early stages, particularly EMCI and LMCI, underscoring the need for early detection and ongoing monitoring to understand and intervene in disease progression.

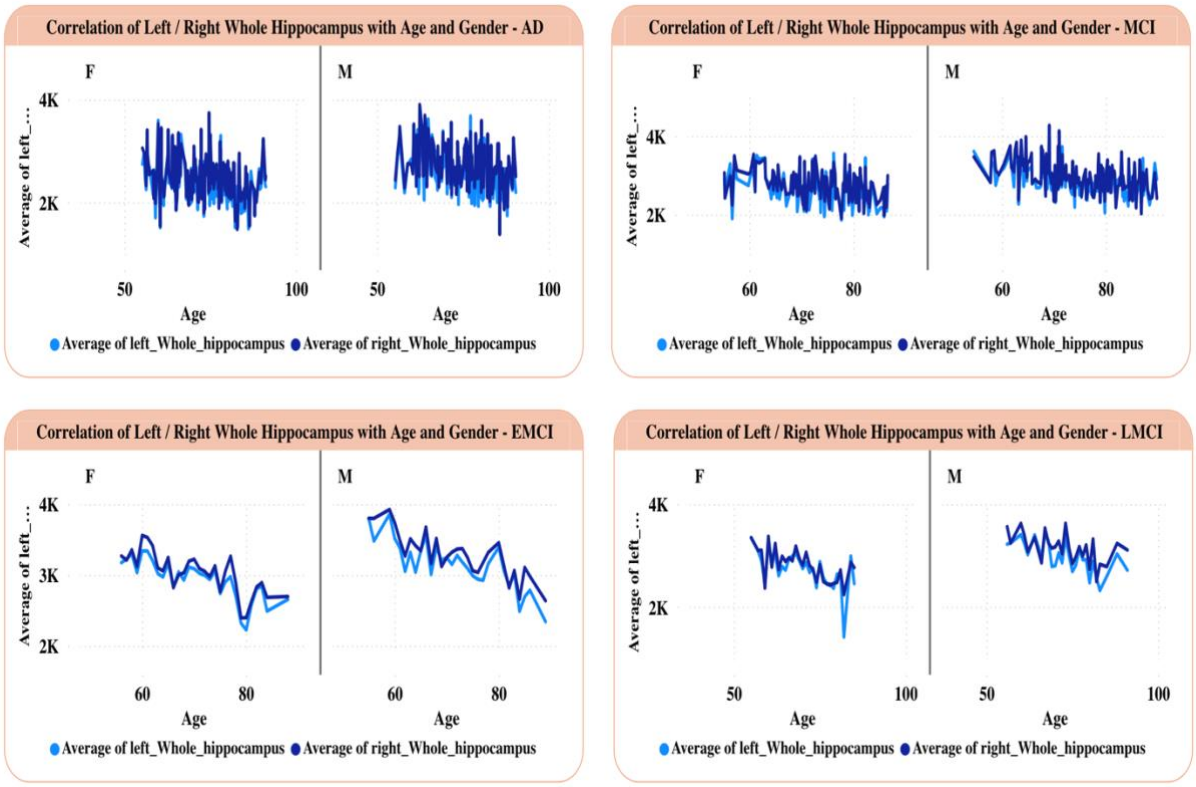

Figure 3- 14 Correlation of Left and Right Whole Hippocampus of the brain with Age and Genders for four different Diseases

This data also highlights the importance of tracking hippocampal changes as a key indicator of disease progression. It highlights potential gender differences in how neurodegenerative diseases such as Alzheimer's impact the brain, suggesting that men and women may experience these conditions differently in terms of brain volume loss and the pace of cognitive decline. Understanding these nuances is critical for tailoring early interventions and treatments that account for these gender-specific differences.

3.5.12 Ventricle

The line graphs in Figure 3- 15 below display the changes in the left and right Lateral Ventricles as they relate to age and gender across four neurodegenerative conditions: AD, MCI, EMCI, and LMCI. A significant observation from the graphs is the consistent upward trend in lateral ventricle size for all genders and conditions on both the right and left sides. This expansion of the lateral ventricles indicates brain matter shrinkage as the disease progresses, a hallmark of neurodegenerative decline.

One key feature is the clear jump in the line graphs as patients age, further emphasising that ageing plays a significant role in the enlargement of lateral ventricles. The ageing process accelerates ventricle growth, corresponding with the loss of brain tissue over time. This phenomenon is most prominent in the EMCI and LMCI stages, where the rapid increase in lateral ventricle size is visually striking, suggesting that significant brain atrophy occurs early in the progression of the disease. In AD, although the ventricles continue to expand, the rate of increase is slower and stable, reflecting ongoing but moderate atrophy in the later stages of the disease.

In contrast, the MCI plot shows a relatively milder increase in lateral ventricle size. Interestingly, female patients with MCI exhibit a gradual rise in ventricle size, while male patients experience a slightly pronounced expansion. These findings suggest that, although ventricle enlargement remains consistent, its rate may vary by gender in certain conditions.

A key feature is the distinction between male and female trajectories. Unlike past findings of gender differences in brain atrophy, changes in lateral ventricle size are uniform across genders. This suggests ventricle enlargement in NDD may not differ significantly by gender. The left and right lateral ventricles expand at similar rates in male and female patients, contradicting earlier observations of gender-based anatomical variations in brain degeneration.

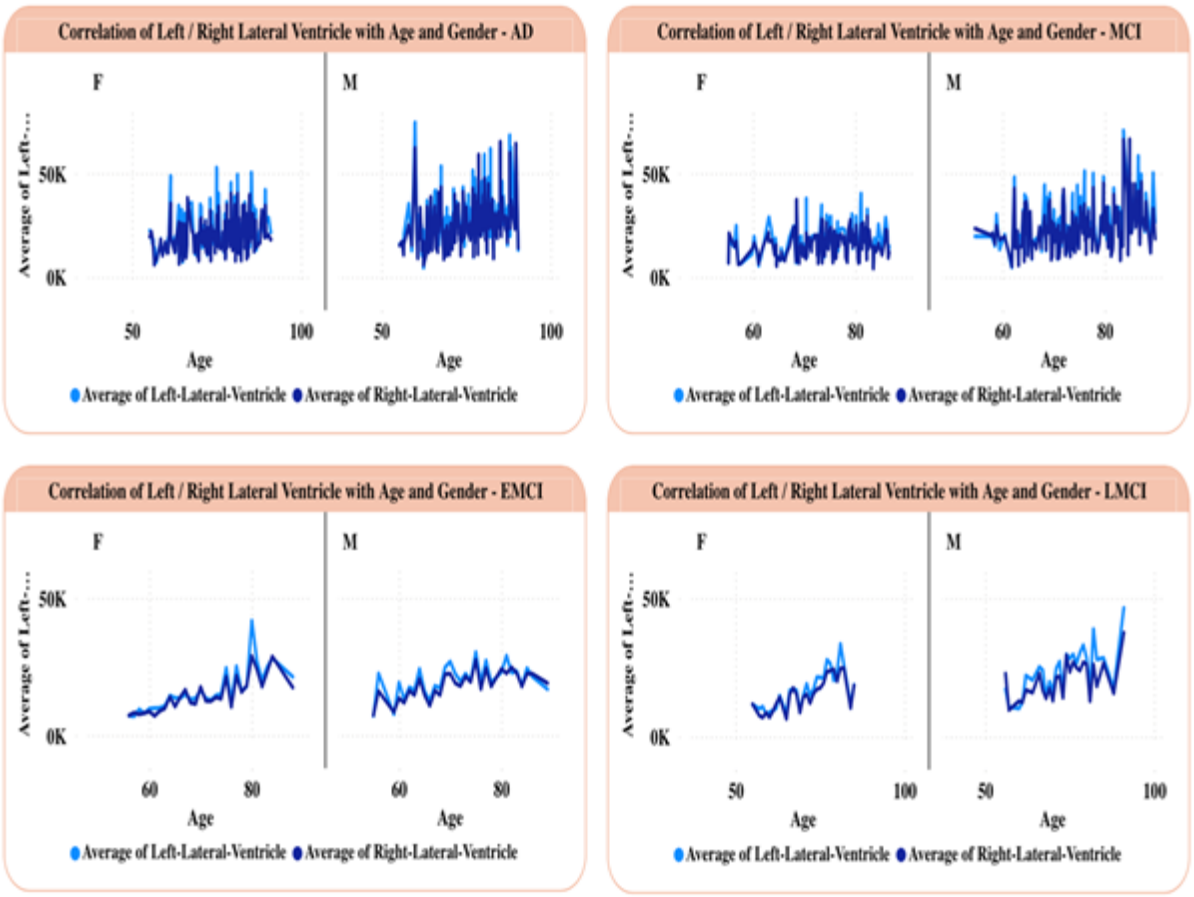

Figure 3- 15 Correlation of Left and Right Lateral Ventricles of the Brain with Age and Gender for four different Diseases

This visualisation shows patterns of ventricular enlargement in neurodegenerative diseases. As brain matter shrinks, particularly in the cortex, ventricles expand to fill the space. This ventricular enlargement is a marker of brain atrophy, often seen in conditions such as AD. A sharp increase in the early stages (EMCI and LMCI) emphasises the importance of early detection. The lateral ventricles are crucial for tracking disease progression, providing insights into the severity of brain tissue loss severity over time. Notably, the absence of significant gender differences suggests that ventricular expansion is a consistent marker of neurodegeneration for both men and women, aiding researchers and clinicians in developing standardised approaches for diagnosis and monitoring across diverse patient populations.

3.6 Experimental Setup

3.6.1 Dataset for Feature Selection

Only two targets, AD and CN, have been selected for the FS project in which experiments will be conducted. Data for classes AD and CN with all metrics includes 404 features in Dataset set 01. Dataset set 02 includes 268 traits for classes AD and CN, encompassing essential metrics such as the volume, area, and thickness of various brain regions.

For external validation of the proposed techniques, Dataset 03 is an arrhythmia dataset (Guvenir et al., 1997). The UCI Arrhythmia Dataset is a medical dataset for classifying cardiac arrhythmias, differentiating between normal heart function and various arrhythmic conditions. It includes categorical, integer, and real-valued features derived from electrocardiogram (ECG) recordings and patient information. The primary goal of the dataset is to classify instances into one of 16 categories, encompassing normal heartbeats and different types of arrhythmias. For research purposes, all arrhythmias can be combined into a single “arrhythmia” class, while normal cases remain “normal,” simplifying the classification task. This dataset makes it valuable for ML medical diagnosis and predictive modelling research.

This chapter established a strong foundation for developing robust FS techniques using the datasets presented in Table 3- 1. These techniques enhance model performance, enhance explainability, and reduce computational costs.

Table 3- 1 Dataset and its Number of features

Dataset Name	**Target**	**Number of features**
Dataset 01: Full Set of MRI Features	AD/CN	401
Dataset: Reduced Set of MRI Features	AD/CN	268
Dataset 03: Arrythmia	Arrhythmia/Normal	279

3.6.2 Dataset for Sensitivity Analysis

Only two targets, Alzheimer's Disease (AD) and Cognitively Normal (CN), were selected for the Sensitivity analysis experiments. Data for classes AD and CN with all metrics such as volume, thickness, standard deviation of thickness, mean curvature, and area, totalling 404 features in Dataset set 01.

Dataset set 02 includes 268 traits for classes AD and CN, encompassing essential metrics such as the volume, area, and thickness of various brain regions. The standard deviation of thickness and mean curvature were excluded from Dataset 02, as these are derivative features that may introduce irrelevant variability. The focus was placed on primary structural features such as volume, area, and thickness, which are directly interpretable and typically hold stronger discriminative power in classification tasks.

This chapter builds a solid foundation for developing effective SA techniques using the datasets presented below in Table 3- 2. These techniques enhance model explainability, which is crucial for incorporating AI into real-world applications.

Table 3- 2 Datasets utilised in the sensitivity analysis

Dataset Name	Target	Number of Features
Dataset 01: Full Set of MRI Features	AD/CN	401
Dataset 02: Reduced Set of MRI Features	AD/CN	268

3.6.3 Dataset for Transfer Learning

Only targets such as Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), Early MCI (EMCI), and Late MCI (LMCI) were selected for the transfer learning process experiments. The study used two distinct datasets to examine and develop predictive models for cognitive deterioration in individuals with AD.

The first dataset comprised records from patients with different stages of MCI, including Early MCI and Late MCI. This dataset consisted of 975 entries, broken down into 335 records from EMCI patients, 461 from MCI patients, and 179 from LMCI patients. Each entry included 404 features related to cognitive and morphological data, along with demographic information such as age, gender, and research group classification. The numerous features provided a comprehensive dataset, enabling a detailed exploration of the various factors contributing to cognitive impairment in patients at different stages of MCI.

Designated values were assigned based on MMSE score ranges to prepare the second dataset for the primary research goal, i.e., evaluating the degree of cognitive impairment in AD patients (Joshi et al., 2019). These ranges helped categorise the severity of cognitive decline in patients. An MMSE score of 21–30 was assigned a value of 3, indicating mild cognitive impairment. A score between 15-20 was assigned a value of 2, representing moderate cognitive deterioration. Finally, a score in the range of 0–14 was assigned a value of 1, signifying moderate to severe cognitive impairment. This scoring system helped simplify the categorisation of patients for further analysis.

The second dataset, focused on AD, contained 2,110 entries. Each entry included morphological characteristics, MMSE scores, and corresponding disease stages derived from the MMSE scores. However, the analysis in this chapter focused solely on the mild and moderate cognitive stages, as data on patients with severe cognitive impairment were limited. Due to the inherent class imbalance, where mild cases outnumbered moderate ones, random undersampling was employed to balance the dataset. After under-sampling, the dataset consisted of 230 records, comprising 115 entries for mild cognitive impairment and 115 for moderate impairment. The summary of all the datasets utilised in this study is presented in Table 3- 3 below. This balanced dataset was then used for two key tasks: regression analysis to predict MMSE scores and classification of cognitive impairment severity.

Table 3- 3 Description of the datasets

Dataset Name	**Diseases Included**	**Target**	**Number of Records**
Dataset01: MCI dataset	MCI, LMCI, EMCI	Age	975
Dataset01: MCI dataset	MCI, LMCI, EMCI	NA	975
Dataset02: AD Dataset	AD	MMSE Scores	230

4. Improved Filter-Based Feature Selection Techniques Based on Correlation and Clustering Techniques

This chapter explores FS methods that enhance model performance and interpretability in high-dimensional medical datasets. FS is vital in ML research, where datasets often contain numerous interdependent variables that introduce irrelevancy and noise, degrading model efficiency. Correlation-based and clustering-based techniques effectively identify informative

features while reducing dimensionality. These approaches enhance model generalisability and robustness by retaining discriminative patterns relevant to the target task while eliminating irrelevant or redundant inputs (Hall, 2000).

Three primary motivations drive these innovative techniques. Firstly, they address the "curse of dimensionality" prevalent in high-dimensional datasets, where the presence of many features can lead to overfitting and reduce predictive performance (Debie and Shafi, 2019). Correlation-based methods, such as Minimum Redundancy Maximum Relevance (MRMR) (Peng et al., 2005; Radovic et al., 2017), select the most informative features by focusing on those highly relevant to the target variable while reducing redundancy. Instance-based approaches, such as ReliefF, evaluate feature importance by assessing how well individual features differentiate between instances of the same and different classes based on nearest-neighbour distances (Kononenko, 1994). Clustering methods, such as hierarchical and spectral clustering, group similar features to enhance representative selection, enhancing classification accuracy in AI models. These techniques enhance model efficiency by streamlining data while retaining essential predictive information, thereby enhancing classification and prediction performance in complex tasks.

Second, FS techniques enhance the explainability of models, which is essential for real-world applications. By reducing the number of features, these methods make it easier to interpret the relationships between input variables and model predictions. Transparent and interpretable models enable developers and domain experts to verify the basis of predictions, ensuring that AI systems integrate with task-specific knowledge and integrate effectively into operational workflows. Additionally, removing irrelevant features reduces the risk of spurious correlations, resulting in a robust and trustworthy model recommendation.

Finally, High-dimensional datasets present considerable computational challenges, as training and executing ML models on such data can demand extensive resources. With an increasing number of features, model complexity escalates, resulting in prolonged training times, higher memory needs, and a heightened possibility of overfitting. This is particularly relevant in domains such as image analysis, sensor data, and text processing, where datasets often contain thousands of variables, making model optimisation and hyperparameter tuning computationally expensive (Guyon and Elisseeff, 2003). Utilising FS to reduce dimensionality

can alleviate these problems by decreasing computational expenses while preserving crucial predictive information, ultimately enhancing model efficiency and scalability.

This chapter aims to provide a framework for enhancing the performance and interpretability of ML models using correlation and clustering-based FS techniques. These methods support the development of AI-driven tools that are accurate, transparent, and adaptable for complex, real-world applications.

4.1 Methodology:

This section introduces a structured and systematic approach to FS techniques developed to enhance the performance and interpretability of ML models. Implemented using Python 3.9.13 and relevant scientific libraries, the techniques use correlation-based and clustering-based strategies to identify the most informative features from datasets.

The process involved implementing the methodology on two carefully curated subsets of the original dataset and an external dataset for validation. These datasets captured different data dimensions, such as demographic variability and MRI imaging-derived features, facilitating a robust evaluation of the FS techniques across multiple datasets.

Multiple ML and DL models were employed to validate the effectiveness of the selected features. This approach enabled the comparison of FS impact across a spectrum of model complexities commonly used in predictive data analysis. Metrics were used to benchmark performance, including accuracy.

Cross-validation strategies were integrated into the pipeline to mitigate overfitting and ensure the generalisability of the results. The reliability of the FS methods was rigorously assessed by systematically partitioning the datasets into training and test splits.

The following sub-sections will explain each stage of the FS framework in detail, including the rationale behind the techniques and the observed impact on downstream model performance.

This chapter presents a novel FS approach that enhances model accuracy while preserving interpretability, which is crucial in real-world settings. By identifying robust and significant features, the method supports the development of reliable predictive tools.

4.1.1 CGN-FS: Correlation-based Greedy Neighbourhood Feature Selection

This section introduces the Correlation-based Greedy Neighbourhood FS (CGN-FS) methodology, a novel and systematic approach for identifying and retaining the most informative features from high-dimensional datasets. CGN-FS combines correlation analysis with threshold-based filtering and evaluation metrics to reduce redundancy, enhance model interpretability, and enhance predictive performance. A pseudo code for the algorithm is given in Algorithm 4- 1 table below.

Algorithm 01: Correlation-based Greedy Neighbourhood Feature Selection Method (CGN-FS)

Input: High Dimensional Dataset

Threshold: To be chosen after thorough analysis

Output: Subset of features

Procedure:

1. Calculate the feature correlation using Pearson Method.
2. Obtain the absolute values of the feature correlation matrix.
3. Calculate the sum of each feature's correlation values w.r.t all other features.
4. For each feature i, count the number of correlation values above threshold w.r.t to all other features and identify these features as neighbours(i). The number of neighbours is 'count' value.
5. Sort by decreasing order of count (primary) and sum (secondary).
6. Initialize flag as 'Keep' for all features.
7. For each feature i, if Flag(i) is "Keep" mark the features in the neighbours(i) as 'Removed".
8. Return the features with flag "Keep".

Algorithm 4- 1 CGN-FS: Correlation-based Greedy Neighbourhood Feature Selection

Step 1: Correlation Matrix Generation

The initial stage of CGN-FS involves constructing a correlation matrix to capture the linear relationships between all pairs of features within the dataset. Pearsons correlation coefficient is used for this purpose. It offers a well-established measure of linear association, ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation). A coefficient close to zero indicates a minimal or no linear relationship.

The resulting matrix is square, with rows and columns corresponding to the input features. Diagonal entries, representing the correlation of each feature with itself, are always

equal to 1 and are excluded from further analysis. This exclusion is essential for computational efficiency and relevance, as self-correlation does not provide helpful information for selection.

The absolute values of the correlation coefficients are computed to standardise the interpretation of correlation strength. This ensures that strong positive and negative relationships are treated equally, enabling a holistic understanding of inter-feature interactions. Features that exhibit high absolute correlations with others are considered potentially irrelevant, setting the foundation for the subsequent selection steps.

Step 2: Computation of Evaluation Metrics

Once the absolute correlation matrix is generated, two key metrics—sum and count — are computed for each feature. The schematic of the Sum variable is presented in the Figure 4- 1 below.

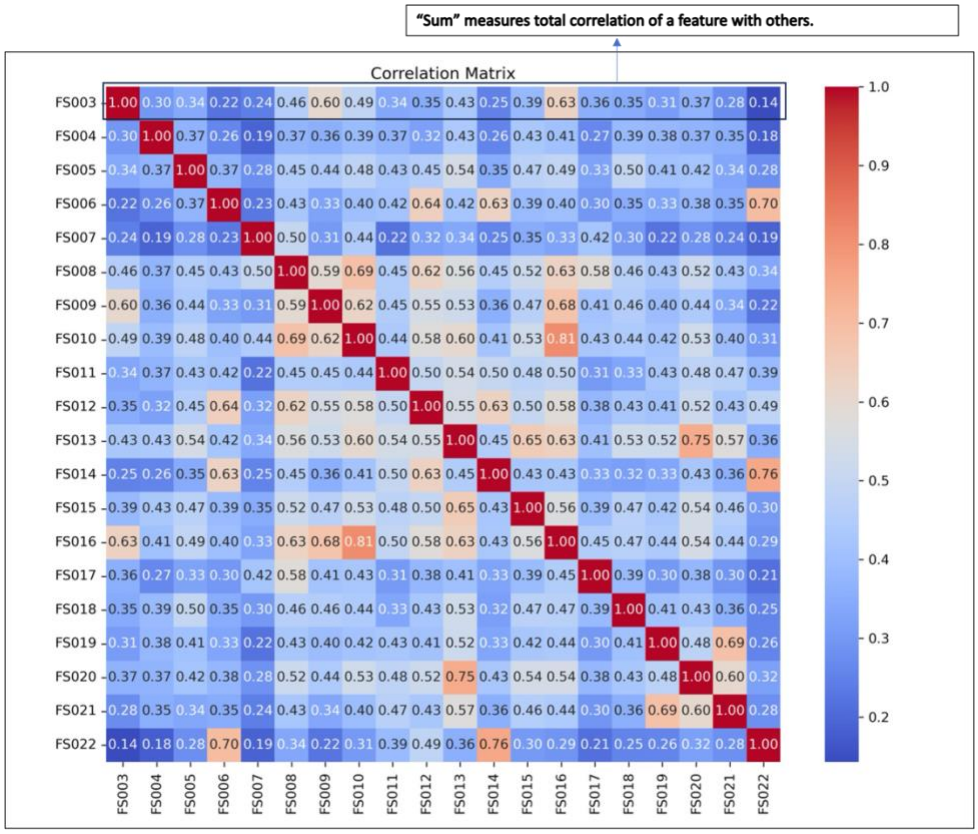

Figure 4- 1 Sample of 'SUM' attribute calculation

The Sum quantifies the total absolute correlation of a given feature with all other features. A high Sum indicates that a feature is generally well-connected within the feature space, often implying redundancy or high similarity with other variables.

The schematic of the count variable is presented in the Figure 4- 2 below. In this example, a user-defined threshold is applied, where a sample is considered valid if its value exceeds 0.60.

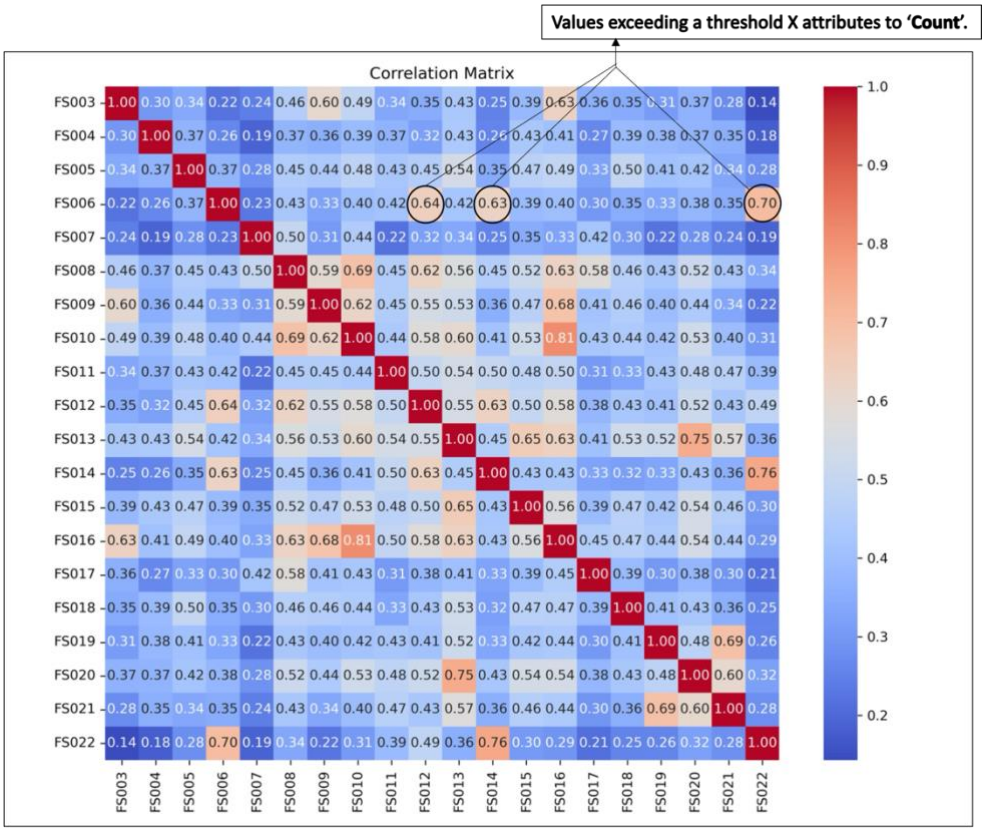

Figure 4- 2 Sample of calculation of the 'Count' Attribute

The count represents the number of features with which a given feature shares a strong correlation, defined by a user-specified threshold (e.g., > 0.75). Features exceeding this threshold are labelled Neighbours, highlighting their dense connectivity within the dataset. The methodology is evaluated across a spectrum of correlation thresholds to assess the robustness of CGN-FS. Starting from 0.50, thresholds are incremented by 0.05 until they reach 1.00.

These metrics serve complementary roles—Sum provides a global view of correlation strength, while count identifies local clusters of highly related features.

Step 3: Sorting and Filtering Features

After calculating the metrics, features are sorted in descending order based on their Count and Sum values. This prioritisation highlights features with strong and widespread

correlations. Each feature is assigned a flag “Keep” and its respective neighbours are assigned a flag “Remove”. The schematic of the CGN-FS methodology is presented in the Figure 4- 3 below.

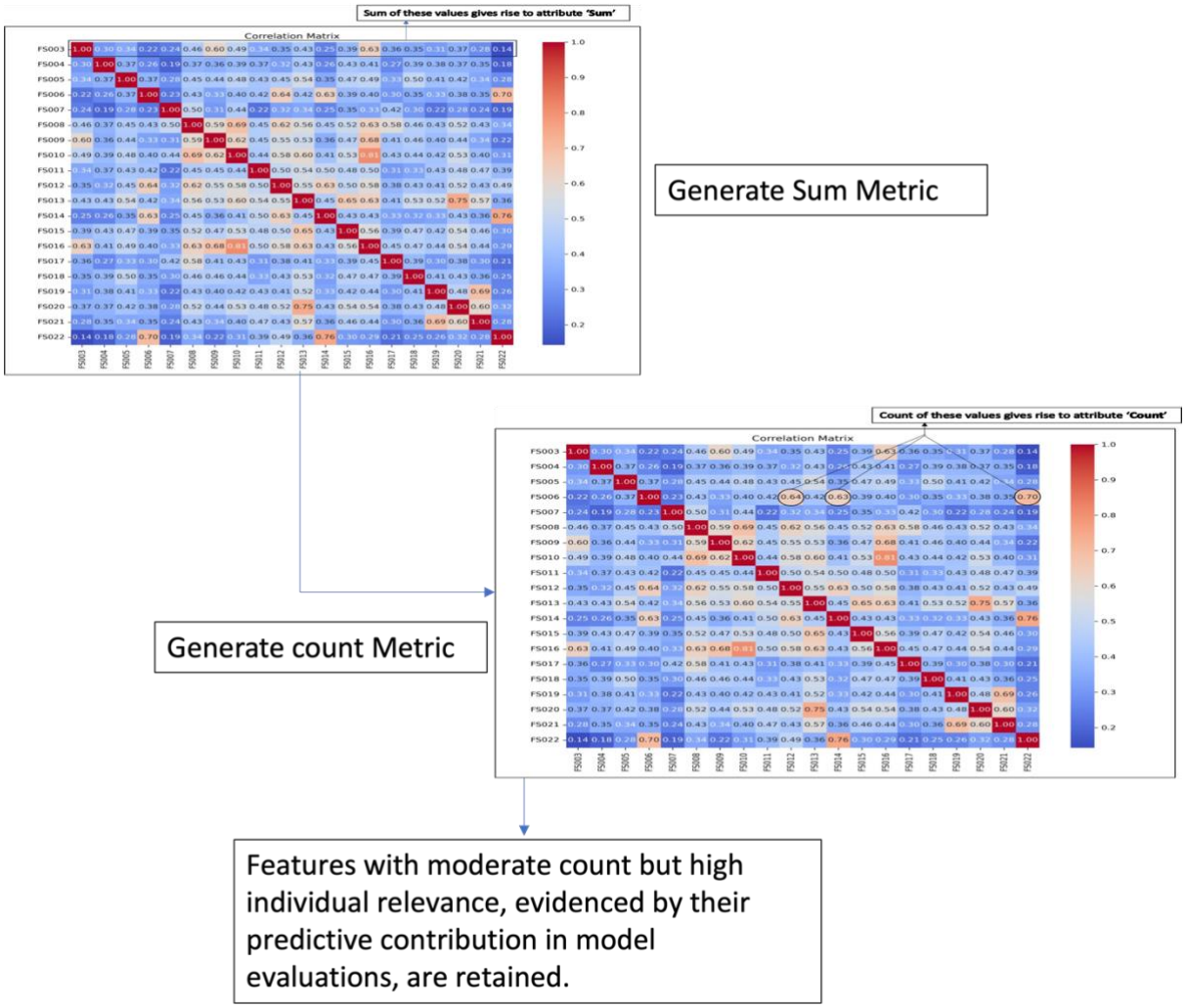

Figure 4- 3 Schematic of the CGN-FS methodology

Step 4: Final Feature Selection and Retrieval

The final subset of features consists of those flagged as “Keep.” These represent the most informative features from the original dataset. This list forms the input for downstream predictive modelling tasks. The CGN-FS algorithm is flexible and iterative, enabling it to be reapplied to various datasets and making it appropriate for numerous applications.

Step 5: Evaluation Across Multiple Thresholds

At each threshold level, subsets of features are selected and evaluated using multiple ML classifiers, such as Random Forests, SVM, and shallow neural networks. These classifiers represent diverse modelling strategies suitable for handling complex high-dimensional data.

Each classifier is trained and evaluated using accuracy, precision, recall, and F1-score performance metrics. Additionally, the standard deviation of accuracy scores across repeated trials is computed to assess the stability of the selected feature subsets. The objective is to maximise accuracy while minimising variance, ensuring that selected features generalise well across different model configurations and do not introduce instability.

In conclusion, the CGN-FS method provides a robust, transparent, and practical approach to reducing dimensionality in high-dimensional datasets. The method efficiently identifies the most relevant features by utilising correlation analysis and strategic thresholding while mitigating multicollinearity and redundancy. This approach enhances the performance and stability of ML models and enhances interpretability, an essential requirement in decision-making contexts. Through its application, CGN-FS demonstrates potential in enhancing model efficiency and reliability in complex, high-dimensional tasks, ultimately contributing to accurate predictions and informed decision-making across diverse domains.

4.1.2 RCH-FSC: Region and Clustering-based Heuristic Feature Selection with Clustering Analysis

This section presents the Region and Clustering-based Heuristic Feature Selection with Clustering Analysis (RCH-FSC), an advanced technique designed to address the challenges of high-dimensional feature spaces. RCH-FSC offers a structured, data-driven method that combines correlation analysis with clustering techniques to identify a compact, interpretable subset of representative features. The primary goal of this technique is to enhance the efficiency, accuracy, and interpretability of downstream predictive models without compromising essential information. A pseudo code for the algorithm is given in Algorithm 4-2 table below.

Algorithm 02: Region and Correlation based Heuristic Feature Selection with Clustering Analysis Method (RCH-FSC)

Input: High Dimensional Dataset

Output: Subset of features

Procedure:

1. Calculate the feature correlation using the Pearson Method.
2. Obtain the distance matrix from the correlation matrix.
3. Perform Principal Co-Ordinate Analysis with Multi Dimensional Scaling.
4. Perform K-Medoids Clustering analysis.
5. Identify one feature to represent each and entire cluster.
6. Identified features are the final subset.

Algorithm 4- 2 RCH-FSC: Region and Clustering-based Heuristic Feature Selection with Clustering Analysis

Step 1: Input and Initial Setup

The RCH-FSC process begins with a high-dimensional input dataset, typically comprising numerous features. These features often exhibit complex interdependencies, warranting a systematic approach to reduce dimensionality while preserving essential information.

Step 2: Correlation Matrix Generation

The first analytical step involves generating a correlation matrix to capture the pairwise relationships among all features. Pearsons correlation coefficient measures the linear dependency between features. As with CGN-FS, the diagonal elements of this matrix, representing self-correlations, are excluded, given their lack of contribution to inter-feature relationship analysis.

To facilitate uniform assessment, the absolute values of correlation coefficients are calculated, ensuring that strong positive and negative correlations are treated equivalently. This enables a comprehensive understanding of redundancy and similarity among features.

Step 3: Correlation Distance Calculation and Normalisation

Next, the absolute correlation matrix is transformed into a correlation distance matrix, which quantifies dissimilarity between features. This conversion is essential for the subsequent

clustering process, as clustering algorithms generally operate on distance metrics rather than similarity measures.

The distance matrix is normalised to prevent any individual or cluster of features from disproportionately influencing the outcome. This standardisation step ensures that all features contribute equitably to the clustering process and promote balanced cluster formation.

Step 4: Dimensionality Reduction via Principal Coordinate Analysis

Following normalisation, the correlation distance data undergoes Principal Coordinate Analysis (PCoA), a technique that projects high-dimensional data into a lower-dimensional space. This step preserves the relative distances between features, thereby maintaining the integrity of feature relationships while making them tractable for visualisation and clustering.

The dimensionality reduction enhances the clarity of inter-feature patterns, preparing the data for robust clustering by highlighting the underlying structure and separability.

Step 5: K-medoids Clustering and Feature Selection

With the lower-dimensional representation of features, the K-medoid clustering algorithm is applied to group similar features based on their proximity in the transformed space. In contrast to K-means, K-medoids select actual data points as cluster centres (medoids), offering greater resilience to outliers and noise, which are common attributes in high-dimensional datasets.

- The optimal number of clusters is determined using techniques such as the elbow method and silhouette score analysis:
- The elbow method identifies the point at which increasing the number of clusters yields diminishing improvements in cluster cohesion.
- The silhouette score assesses cluster consistency, with higher scores indicating better-defined groupings.

The medoid, the most central and representative feature, is selected for each cluster formed. These medoids constitute the final subset of features, encapsulating the diversity of the whole dataset while reducing redundancy.

Step 6: Final feature subset

The selected medoids represent the most informative and distinct features within the original dataset. This approach significantly reduces dimensionality while preserving, and in some cases enhancing, the predictive power of the models. The resulting feature subset enhances computational efficiency and facilitates model interpretability.

In summary, the RCH-FSC methodology introduces a robust, clustering-based heuristic approach to FS which is presented in Figure 4- 4 below.

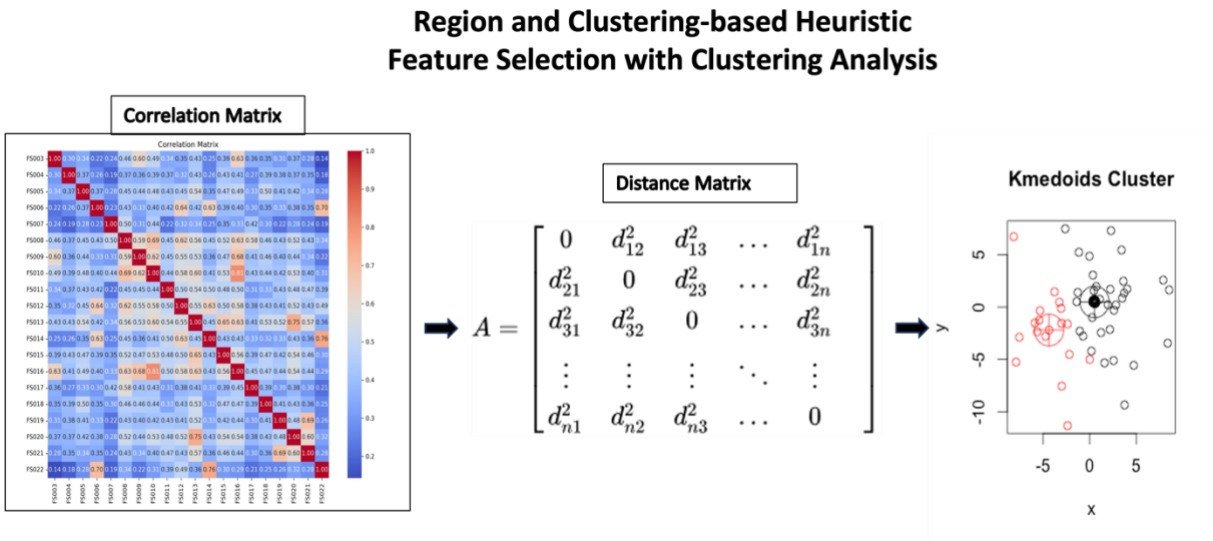

Figure 4- 4 Schematic diagram of RCH-FSC

The method provides a principled means of reducing feature space in high-dimensional neuroimaging datasets by integrating correlation analysis, distance normalisation, dimensionality reduction, and K-medoid clustering. The selected representative features enable efficient modelling, enhanced interpretability, and reliable performance, key considerations for ML applications.

4.2 Results and Discussions:

4.2.1 Quantitative Analysis:

In this section, the quantitative results obtained are discussed by applying the proposed techniques to the AD and Arrhythmia datasets. Although validated on these datasets, the

proposed methods are generalisable and suitable for application across other domains with high-dimensional data.

Feature selection was performed on two AD datasets, one comprising 401 features (Dataset 1) and the other 265 features (Dataset 2). To ensure robustness, external validation was carried out using the Arrhythmia dataset. The CGN-FS method was further validated using the ReliefF algorithm, demonstrating consistency and reliability across diverse datasets.

CGN-FS Method:

The developed models were evaluated using repeated 10-fold stratified cross-validation across thresholds from 0.1 to 0.95 and presented in Figure 4- 5 below.

Three classifiers were evaluated using the AD Dataset 1 with 401 features for FS. Logistic Regression achieved an accuracy of 88.91% ± 3.03, while the SVM classifier reached 87.66% ± 2.70. The Shallow NN outperformed both, achieving the highest accuracy of 97.29% ± 0.94, demonstrating its effectiveness on this feature set. Of the three, the Shallow NN was the most accurate and consistent, likely due to its ability to capture complex, non-linear patterns within the data. In contrast, though comparable, Logistic Regression and SVM showed significant variability and lower performance. However, the increased interpretability of these traditional models, particularly Logistic Regression, could still make them valuable in scenarios where transparency is crucial.

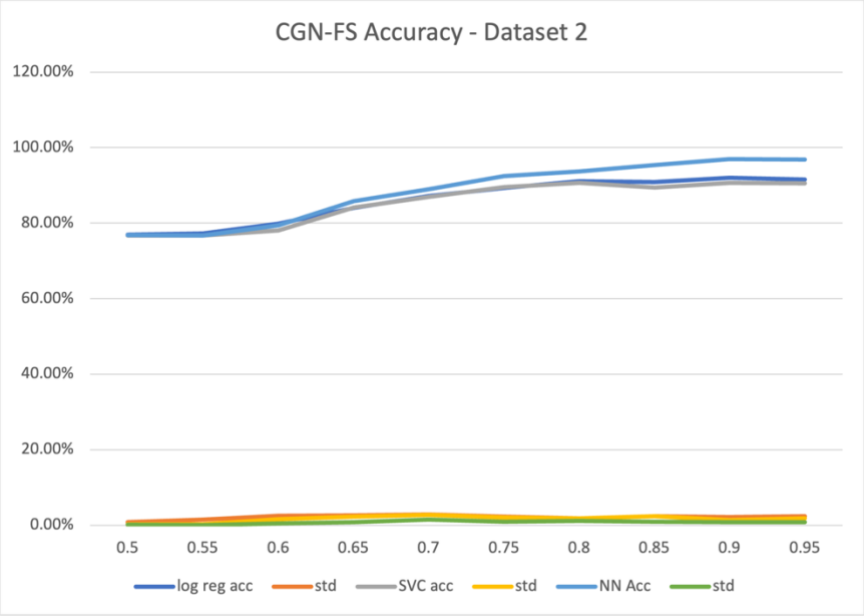

Figure 4- 5 Accuracy plot for CGN-FS

Using the AD Dataset 2 with 265 features, three classifiers were assessed. Logistic Regression achieved an accuracy of 91.06% ± 1.76, while the SVM classifier performed slightly lower at 90.66% ± 1.83. The Shallow Neural Network again showed enhanced performance with an accuracy of 92.49% ± 0.95. Although the margin was narrower than that of dataset 2, the neural network remained the most accurate and consistent model, benefiting from its ability to learn complex patterns. Nonetheless, Logistic Regression offered strong performance with interpretability, making it a competitive choice for applications requiring model transparency.

The summary of the results obtained using the CGN-FS methodology, validated on different datasets along with their respective accuracies, is presented in Table 4- 1 below.

Table 4- 1 Performance summary of CGN-FS methodologies and their respective accuracy

Method	**Data**	**Number of Features**	**Model**	**Mean Accuracy**	**ReliefF**
CGN-FS	AD Dataset 1 (401 features)	182	Logistic Regression	88.91 ± 3.03	90.30
		297	SVM Classifier	87.66 ± 2.70	76.72
		217	Shallow Neural network	97.29 ± 0.94	-
	AD Dataset 2 (265 features)	95	Logistic Regression	91.06 ± 1.76	90.40
		95	SVM Classifier	90.66 ± 1.83	78.06
		67	Shallow Neural network	92.49 ± 0.95	-
	Arrhythmia Dataset	101	Logistic Regression	68.57 ± 3.87	70.58
		101	SVM Classifier	67.26 ± 2.70	73.91
		200	Shallow Neural network	92.10 ± 2.77	-

On the Arrhythmia dataset, Logistic Regression achieved an accuracy of 68.57% ± 3.87, while the SVM classifier followed closely with 67.26% ± 2.70. The Shallow NN significantly outperformed both, achieving 92.10% ± 2.77. Notably, the NN utilised 200

features—nearly double the number used by the traditional classifiers—highlighting its ability to utilise higher-dimensional representations effectively. While this resulted in a substantial performance boost, it also comes with increased computational cost and reduced interpretability, essential considerations for deployment in real-world settings.

The ReliefF method was also implemented for comparison using 95 features and 20 random neighbours. Using logistic regression, the ReliefF approach achieved an accuracy of 90%, demonstrating that the CGN-FS methodology provided enhanced FS and model performance.

RCH-FSC method:

Two distinct AD/CN datasets were evaluated using the clustering-based FS method for RCH-FSC. To determine the number of clusters, the elbow method and silhouette score were utilised, and their results are presented in Figure 4- 6 below. The first dataset, containing 401 features, yielded four optimal features: rh_paracentral_area, lh_middletemporal_meancurv, rh_parsorbitalis_volume, and rh_pericalcarine_thickness. SVM on this feature set achieved the best accuracy of 77.42%, with a standard deviation of 0.80.

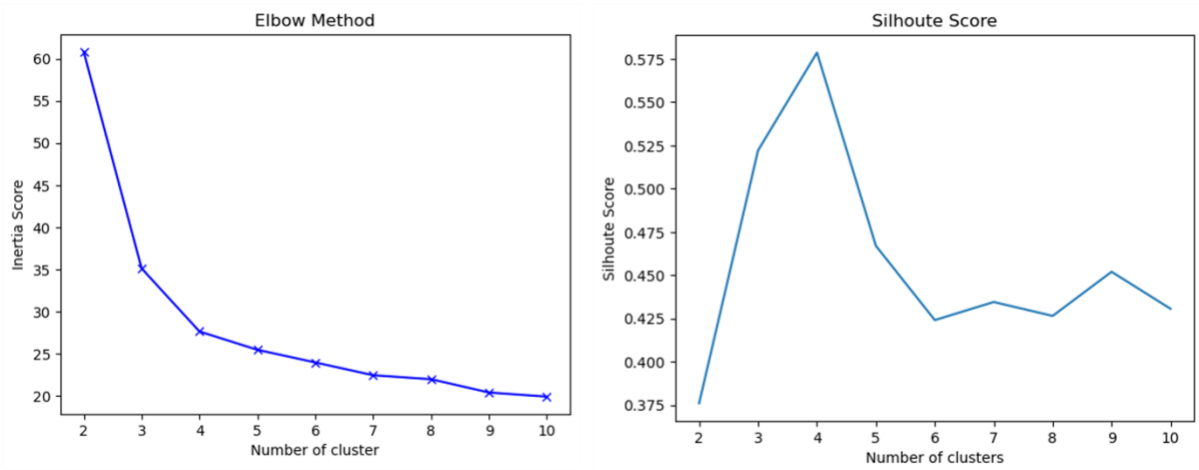

Figure 4- 6 Number of Cluster Analysis

In contrast, the second dataset, with 265 features, identified four key features: lh_bankssts_thickness, lh_pericalcarine_thickness, rh_parsorbitalis_area, and lh_cuneus_volume. Logistic regression performed best on these four features, achieving an accuracy of 80.41% and a standard deviation of 1.6.

The summary of the results obtained using the RCH-FSC methodology, validated on different datasets along with their respective accuracies, is presented in Table 4- 2 below.

Table 4- 2 Performance summary of RCH-FSC methodologies and their respective accuracy

Method	Data	Number of Features	Model	Mean Accuracy
RCH-FSC	AD Dataset 1 (401 features)	4	Logistic Regression	76.92 ± 0.60
			SVM Classifier	77.42 ± 0.80
			Shallow Neural network	76.94 ± 0.16
	AD Dataset 2 (265 Features)	4	Logistic Regression	80.41 ± 1.60
			SVM Classifier	80.10 ± 1.90
			Shallow Neural network	78.42 ± 0.42
	Arrhythmia Dataset	17	Logistic Regression	68.82 ± 8.04
			SVM Classifier	73.45 ± 5.30
			Shallow Neural network	71.36 ± 2.01

For external validation, the Arrhythmia dataset, with 279 features, identified 17 key features. SVM performed best on these features, achieving an accuracy of 73.45% and a standard deviation of 5.30.

In both cases, the FS methodology demonstrated its capacity to significantly reduce the number of features while maintaining high model accuracy. These results underscore the robustness of the selected feature sets in differentiating between AD and CN classes, particularly in comparison to models trained on random or complete feature sets.

4.2.2 Discussion

This research explored and compared two distinct FS methodologies, CGN-FS and RCH-FSC, validated on internal (AD datasets) and external (Arrhythmia dataset) data. The objective was to assess the effectiveness of each method in minimising dimensionality while maintaining or enhancing model performance and interpretability, a factor that is particularly paramount when managing high-dimensional datasets.

The CGN-FS method demonstrated strong performance across all datasets and classifiers. Applied to the AD Dataset 1 (401 features), CGN-FS significantly reduced the feature space while maintaining high classification accuracy. Notably, the Shallow NN achieved 97.29% ± 0.94 accuracy using 217 selected features, outperforming Logistic Regression (88.91% ± 3.03) and the SVM classifier (87.66% ± 2.70). This trend continued with the 265-feature AD Dataset 2, where the NN again showed the highest performance (92.49% ± 0.95), though the performance gap between classifiers narrowed. The consistency of the deep learning model results in a relatively low standard deviation in both cases supports its robustness and suitability for complex, non-linear patterns common in real-time data. However, this performance comes at the cost of interpretability and computational complexity, particularly when the number of retained features is considerably higher than for traditional classifiers.

The CGN-FS method was further validated through the ReliefF algorithm. Across datasets, ReliefF-supported feature subsets showed slightly lower performance, suggesting that CGN-FS captured strongly predictive features and managed redundancy effectively. For instance, in AD Dataset 1, ReliefF with Logistic Regression yielded an accuracy of 90.30%, while CGN-FS yielded a slightly lower 88.91%, though the difference was not substantial. Interestingly, in some cases, ReliefF produced comparable performance (e.g., AD Dataset 2 with Logistic Regression at 90.40%), reinforcing the reliability of both techniques but slightly favouring CGN-FS for nuanced real-world datasets.

The external validation on the Arrhythmia dataset added an essential dimension to this analysis. Using CGN-FS, the Shallow NN achieved an impressive 92.10% ± 2.77 accuracy using 200 features—nearly double the number used by Logistic Regression and SVM (each using 101 features). This result underscores the capacity of the NN to harness high-dimensional representations effectively. However, the increased feature count also brings potential overfitting risks and interpretability challenges, which should be cautiously addressed in real-world applications. Meanwhile, Logistic Regression and SVM, while less accurate (68.57% and 67.26% respectively), offered interpretable models and required significantly fewer features, which could be preferable in resource-constrained or regulatory-sensitive environments.

In contrast, the RCH-FSC method took a fundamentally different approach by identifying a compact and highly representative feature subset using clustering. This method reduced AD datasets to four features while achieving reasonable classification accuracy. On AD Dataset 1, SVM achieved the highest accuracy (77.42% ± 0.80), followed closely by the Shallow NN and Logistic Regression. In Dataset 2, Logistic Regression performed best (80.41% ± 1.60), slightly outperforming SVM and NN. Although the absolute accuracies were lower than those achieved through CGN-FS, the significant reduction in feature count, down to approximately 1% of the original dimensionality, highlights the strength of RCH-FSC in generating lightweight, interpretable models. This is particularly important where reducing complexity can lead to practical and explainable AI tools.

The Arrhythmia dataset, used for external validation of RCH-FSC, further demonstrated the utility of this clustering-based approach. With only 17 selected features, SVM achieved the highest accuracy (73.45% ± 5.30), surpassing both Logistic Regression (68.82% ± 8.04) and Shallow NN (71.36% ± 2.01). Despite a moderate drop in accuracy compared to CGN-FS, RCH-FSC still provided strong generalisability, showing that it can effectively reduce dimensionality without a significant loss in performance. Additionally, the interpretable nature of RCH-FSC-selected features (e.g., specific brain regions) could make this method particularly appealing for applications that require in depth insights or logical reasoning.

Comparatively, CGN-FS produces high-performing models by maintaining features, particularly when model performance is the priority. Meanwhile, RCH-FSC prioritises compactness and interpretability, showing strength when the goal is to identify a small set of meaningful features or when computational efficiency is essential. Together, these results address the trade-off between performance and interpretability, where the specific demands of the critical task should guide the choice of FS method. CGN-FS combined with a Shallow NN is most effective for predictive accuracy. However, for model simplicity and transparency, which are crucial in real-world deployment, RCH-FSC offers a balanced and explainable solution.

Comparison with recent feature selection methods

Recent research on feature selection has been predominantly characterised by approaches driven by deep learning, such as attention-enhanced Convolutional Neural Networks (CNNs), Vision Transformers, hybrid deep-feature pipelines, sparse or embedded methods, and a wide range of meta-heuristic optimisation strategies. These methods have shown improvements, especially when implemented on extensive MRI image datasets where attention mechanisms or global token interactions can enhance representation learning. Hybrid pipelines that combine deep features with classical FS (e.g., LASSO, PSO, WOA) remain favourable because of their flexibility and generally robust performance. However, despite their strengths, many of these techniques rely on intensive computation, large sample sizes, unstable attention mechanisms, or heuristic search procedures that can limit interpretability, reproducibility, and applicability to smaller ROI-based MRI datasets. The following Table 4- 3 summarises such examples in contrast to the proposed FS techniques.

Table 4- 3 Comparison with Recent Feature Selection Methods

Method category	**Recent FS approaches**	**Strengths**	**Limitations**	**How CGN-FS / RCH-FSC improve on this**
Attention-based Deep FS	3D CNN + attention, ROI-wise 3D-ViT approaches. (Saoud & AlMarzouqi, 2024; Zhou et al., 2025)	Learns task-specific importance, highlights image regions, often improves accuracy.	Requires relatively large data, attention maps can be unstable or hard to interpret as a global feature ranking.	CGN-FS/RCH-FSC are **model-agnostic** and produce stable, global smaller MRI-tabular datasets.
Transformer-derived / ViT methods	Vision-Transformer and hybrid ViT+CNN models for MRI. (Mahmud Joy et al., 2025; Z. Zhao et al., 2024)	Capture long-range/global spatial relationships in images; strong performance on large image sets.	High compute; attention ≠ formal feature selection; not directly suited to ROI-tabular features.	FS operates on ROI/tabular features (lightweight) and yields interpretable subsets without heavy compute.
Embedded / Sparse methods (LASSO, sparse AE)	Sparse autoencoders / stacked sparse AEs. (Alorf & Khan, 2022; Helaly et al., 2021)	Built-in regularisation; directly enforces sparsity; simple to implement.	Sensitive to hyperparameters; may be instability across folds; not necessarily redundancy-aware (group correlated features).	CGN-FS reduces correlated groups and provides more reproducible ranking; RCH-FSC yields ultra-compact interpretable sets (useful when 1–4 ROI features are needed).
Traditional filter / wrapper (ReliefF, mRMR, RFE)	ReliefF and mRMR hybrid CNN. (Eroglu et al., 2022; Sadiq et al., 2021a)	Fast, explainable, widely used; works well as a first filter.	Struggle with very high-dimensional / highly correlated data; ignore higher-order interactions.	CGN-FS addresses redundancy and neighbourhood structure; RCH-FSC provides cluster-level selection (reduces

Method category	Recent FS approaches	Strengths	Limitations	How CGN-FS / RCH-FSC improve on this
				correlated ROI duplication).
Metaheuristic optimisation (WOA/PSO/GA)	WOA, PSO, GA for feature selection in AD pipelines. (Cao et al., 2024; S. Kaur et al., 2022; Mohammad & Al Ahmadi, 2023)	Robust global search; can find small high-accuracy subsets from huge feature pools.	Often heuristic, heavy compute, less interpretable why features chosen; potential overfitting.	CGN-FS / RCH-FSC target interpretability/ stability first (not pure search), validated across external dataset (Arrhythmia) to reduce overfitting risk.

As the above comparison shows, CGN-FS and RCH-FSC directly target several limitations common to recent deep and hybrid FS approaches. Instead of relying on attention weights, image-based token structures, or a meta-heuristic search, both proposed methods operate as model-agnostic, prioritise stability, redundancy-handling, and interpretability. CGN-FS explicitly incorporates correlation and neighbourhood information, producing consistent feature rankings even when datasets are small or highly structured, where transformers or attention-based models typically struggle. RCH-FSC complements this by producing extremely compact, clinically interpretable feature subsets without sacrificing generalisability, as demonstrated through external validation on the Arrhythmia dataset. Together, these two methods address the performance–interpretability trade-offs present in many of the recent research studies and offer a lightweight, transparent alternative that is consistent with real-world clinical deployment and the nature of MRI-derived tabular data.

4.3 Summary of the Key Findings

In this study, two straightforward FS methods—correlation-based and clustering-based—were implemented and validated on the internal AD/CN dataset and the arrhythmia dataset for external validation. These methods aim to enhance model performance by reducing the feature space while maintaining or improving accuracy and interpretability.

The correlation-based method selects features with low inter-feature correlation, ensuring that irrelevant or highly correlated features are excluded. This process reduces the feature set, streamlining the modelling process without compromising and often improving

accuracy. The correlation-based approach outperformed the ReliefF method, demonstrating increased robustness and accuracy in both datasets. Reducing the number of features also enhanced the interpretability of the model, providing clearer insights into the relationship between the selected variables and their impact on classification.

The clustering-based method was equally effective in selecting a reduced set of relevant features by analysing the pattern of data points in a 2D space and determining cluster centroids based on their distances. This method successfully identified key features, significantly reducing dimensionality without significant trade-offs in accuracy. For instance, using the arrhythmia dataset, this approach reduced the feature space to 15 features while maintaining accuracy within a 1.5% margin compared to the complete feature set. This highlights the capability of the algorithm to capture essential information for classification while improving computational efficiency and model interpretability.

In conclusion, applying correlation-based and clustering-based FS methods significantly enhanced model performance, accuracy, and interpretability across the datasets. The correlation-based approach effectively reduced the feature set while maintaining or improving accuracy, particularly in datasets where features were highly correlated. Meanwhile, the clustering-based approach provided a compact and efficient feature set that captured the core patterns in the data, thereby improving model performance with minimal trade-offs.

These methods offer a powerful toolkit for developing efficient ML models, particularly in critical domains where high-dimensional datasets are standard. Streamlining features while retaining important information enhances the practicality of these models in real-world applications. However, attention must be given to the specific characteristics of each dataset to ensure that the selected method aligns with the underlying data structure.

As FS techniques evolve, integrating these methods into broader ML frameworks holds promise for optimising models across various domains, leading to robust, interpretable, and efficient solutions in data-driven decision-making.

4.3.1 Advantages and Challenges of the proposed techniques

FS methods demonstrated increased accuracy and robustness compared to existing methods such as ReliefF.

This is particularly notable in the AD/CN dataset, where feature reduction through correlation analysis enhanced model performance without sacrificing accuracy. These methods significantly reduce the number of input features. For example, the clustering approach selected only 15 features for the arrhythmia dataset, and the correlation method selected 95 features for AD/CN, making the models efficient and faster to train.

Reducing feature space contributes to interpretability. Models trained on fewer but relevant features provide clearer insights into the relationships between the input data and the classification outcomes.

The successful application of these techniques to two distinct datasets highlights the generalisability of the methods. This indicates potential for use in other high-dimensional medical datasets, further broadening the scope of these methodologies.

While the correlation and clustering methods are effective, their performance is still sensitive to the underlying characteristics of the dataset. For instance, the correlation method might underperform if the features are not strongly correlated or have complex interdependencies, as seen in the arrhythmia dataset, where NN performed than logistic regression.

In some scenarios, these methods may not drastically reduce dimensionality, particularly if the dataset has a high degree of feature variability, as observed in the arrhythmia dataset, where the final number of features remained relatively high after clustering (15 features).

4.3.2 Clinical Relevance

The research has significant clinical implications, as the proposed techniques have been validated on the AD dataset, particularly in the early diagnosis and treatment monitoring of AD. By applying advanced FS techniques such as the correlation-based CGN-FS and clustering-based RCH-FSC methods, the model identifies key biomarkers from high-dimensional MRI data that contribute to accurate AD/CN classification. The reduced feature sets, which focus

on specific brain regions (such as the entorhinal cortex and hippocampus), are highly relevant in detecting early cognitive decline and tracking disease progression.

Moreover, these methods enhance the interpretability of ML models, enabling clinicians and researchers to understand the relationships between selected brain structures and AD. This transparency is critical in medical decision-making, as it enables clinicians to base their diagnoses on interpretable, biologically meaningful features rather than black-box models. The external validation using the arrhythmia dataset further underscores the potential for these methodologies to be generalised and applied to other medical domains, improving diagnostic accuracy in areas such as cardiovascular disease.

By streamlining the number of features, these methods also pave the way for efficient and cost-effective diagnostic tools. Reducing the computational burden without sacrificing accuracy could lead to faster, real-time clinical decision support systems, helping practitioners in hospitals and clinics.

4.4.3 Future Work

While this study demonstrates the effectiveness of correlation and clustering FS methods, several avenues remain open for future exploration:

1) Integration with Longitudinal Data: A logical next step would be integrating longitudinal data into the model. By tracking over time, FS methods could identify patterns of disease progression, enabling predictive modelling of when symptoms might emerge or worsen.
2) Cross-Domain Application: The successful validation of these FS methods on the arrhythmia dataset underscores their potential applicability across various domains. Future work could test these methods on other high-dimensional datasets, such as cancer genomics, cardiovascular imaging, or wearable health data, to further validate their effectiveness.
3) Explainability and Interpretability: To enhance transparency, further enhancement of explainability techniques could be integrated with these FS methods. For instance, applying XAI methods such as SHAP or SOBOL could offer deeper insights into why certain features are selected and how they influence model predictions. This would further boost trust in AI models.

4) Model Generalisation and Transfer Learning: Investigating how these models generalise across different datasets and populations is another key area for future work. Utilising transfer learning techniques could help adapt models trained on large datasets to smaller, less well-represented datasets, enhancing their applicability in under-resourced settings.

In summary, its potential lies in enhancing accuracy and its flexibility for adaptation across other domains. Future research should focus on expanding the capabilities of the model, enhancing its explainability, and integrating it into real-world applications for broader impact.

5 Sensitivity Analysis for Feature Importance in Predicting Alzheimer's Disease

This chapter explores XAI frameworks that enhance the transparency and trustworthiness of ML models applied to high-dimensional datasets. As AI is increasingly embedded in critical workflows, understanding how input features influence model predictions is essential, particularly in high-stakes domains where decision reliability is crucial. XAI frameworks utilise SA techniques enabling the interpretation of complex DL models by quantifying the effect of individual input variables on model output, thus contributing to the development of explainable AI systems for real-world applications (Razavi et al., 2021).

Three key motivations underpin the application of XAI in the critical domains. Firstly, XAI directly addresses the black-box nature of modern ML and DL models, which often lack transparency despite their high performance (Bloch and Friedrich, 2022). In real-world settings, the opacity of such models can limit trust and adoption. SA techniques, such as input perturbation, gradient-based saliency maps, and layer-wise relevance propagation, reveal how and why models make specific predictions by attributing importance to input features. This interpretability is essential for experts who must understand the rationale behind AI-assisted decisions, particularly in high-stakes scenarios where clarity and traceability of decisions are critical.

Secondly, XAI plays a dual role in enhancing domain relevance and validating the acceptability of the decision-making process by the model. It identifies which input features significantly influence classification outcomes and ensures they align with known, theoretically or empirically relevant patterns in the dataset. This alignment ensures that predictions are both statistically robust and practically meaningful. The approach becomes particularly vital when models are applied to rare, edge-case, or low-representation scenarios, where early and accurate identification supports timely actions and reliable decision-making. Furthermore, sensitivity-driven insights can guide the discovery of novel, underexplored feature interactions, advancing research in complex, high-dimensional datasets.

Lastly, SA techniques employed provide a mechanism for model refinement and robustness assessment. By revealing the features on which the model relies, researchers can

identify potential overfitting to noise in the dataset. This approach facilitates iterative data preprocessing, feature engineering, and model design improvements. Moreover, SA methods help evaluate the consistency of model behaviour across different data subgroups, ensuring generalisability and fairness, crucial attributes for real-world adoption.

For this research, SHAP and Sobol global sensitivity analysis methods were chosen over alternatives such as LIME, Grad-CAM, Integrated Gradients, and permutation-based importance due to their robustness and precision for high-dimensional MRI-derived datasets. Local methods such as LIME and Grad-CAM struggle with the highly correlated, non-spatial tabular MRI features utilised in Alzheimer's disease research. Integrated Gradients and permutation methods frequently fail to capture nonlinear interactions or score inconsistently when feature distributions are imbalanced. Conversely, SHAP provides theoretically resilient, model-agnostic attributions, and Sobol provides robust global sensitivity measures capable of revealing complex interactions, thereby enhancing their reliability and clinical relevance for Alzheimer's disease prediction tasks.

This chapter utilise SA techniques tailored to high-dimensional datasets, highlighting their role in enhancing model interpretability, domain relevance, and reliability. These strategies contribute to creating AI systems that are powerful, accurate, transparent, explainable, and aligned with real-world deployment needs.

5.1 Methodology:

This section outlines the methodologies employed to conduct SA on a DNN model developed for a classification task on high-dimensional dataset. SA is critical for understanding the internal workings of ML models, particularly DL models, which are often considered black boxes. By scrutinising the impact of input features on predictions, SA facilitates the enhancement of model interpretability, a critical prerequisite for implementing AI models in high-stakes environment.

In this research, multiple SA techniques have been evaluated to compute the importance of features derived from a DNN model classifying AD/CN. Features were obtained from MRI dataset, processed using FreeSurfer to extract various neuroimaging measures, such as cortical thickness, volume, and surface area.

The analysis uses local and global explanation methods, offering a comprehensive approach to feature relevance assessment. Local methods offer valuable insights into specific predictions, whereas global methods provide a comprehensive understanding of the behaviour of the model as a whole dataset. The techniques used in this study include SHAP and the SALib library, which implement global and local SA methods, such as Sobol, Morris, and FAST. These methods provide diverse perspectives on feature importance, enables a robust and reliable evaluation of which features most strongly influence model predictions.

By performing SA on two datasets—Dataset 1 (401 features) and Dataset 2 (268 features)—this study seeks to identify a consistent set of important features for developed DNN model. The findings across methods are compared to ensure the robustness and stability of the results, further advancing the interpretability and trustworthiness of the DNN model in practical applications. Ultimately, this methodology aims to provide valuable insights into the decision-making process and contribute to the development of transparent and explainable AI models. Figure 5- 1 shows the architecture of Model 1 for Dataset 1.

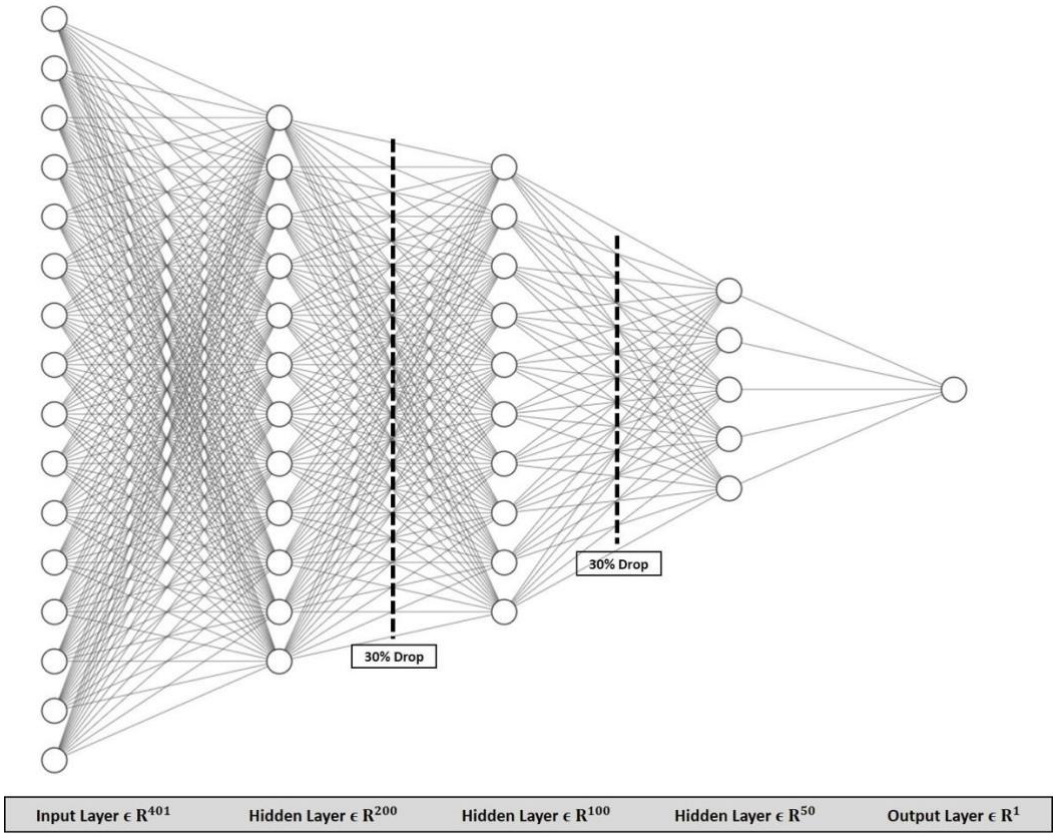

Figure 5- 1 Architecture of Model 1 for dataset 1

The two datasets are input into two distinct DNN models. Each model is similarly structured but adapted to manage different input feature sets. DNN Model 1 has been trained utilising Dataset 1. In contrast, DNN Model 2 has been trained using Dataset 2, which encompasses fewer features but emphasises a refined subset that may still provide robust predictive capabilities.

Each model uses an NN with multiple layers, including an input layer corresponding to the number of features, three hidden layers with ReLU activation functions to introduce non-linearity and an output layer with a sigmoid activation function for binary classification. The models are trained using the Adam optimiser and binary cross-entropy loss with accuracy as the evaluation metric. Figure 5- 2 shows the architecture of the DNN model 2 using Dataset 2.

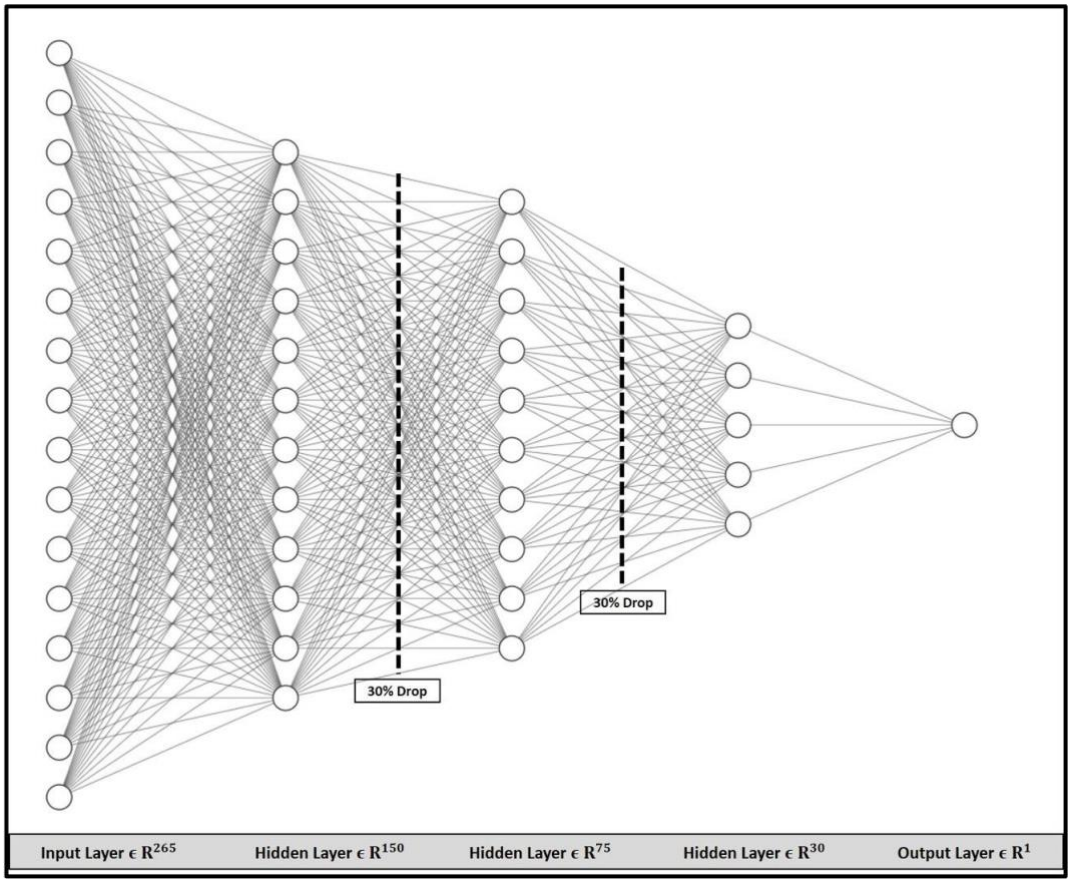

Figure 5- 2 Architecture of the DNN model 2 using dataset 2

These DNN models have been utilised in both the SALib and SHAP methodologies to ensure consistency in the training of models. Both models are designed explicitly for respective datasets and are employed to predict the classification tasks of AD versus CN individuals.

5.1.1 Methodology using SALib

This section outlines the methodology used to perform SA for DNN models validated on the AD dataset. SA uses the SALib Python library, implementing three key methods: Sobol, Morris, and FAST. These techniques assess the importance of various features in the two respective datasets. The approach consists of four main steps: data preparation, augmented data generation, model training, and application of SA techniques.

Step 1: Data Preparation and SA technique Initialisation.

The first step in the methodology involves preparing the datasets for model training. The datasets consist of various features, each representing a distinct characteristic related to the structure of the brain. Dataset 1 contains 401 features, and Dataset 2 includes 268 features, a reduced subset derived from the original dataset.

The training features and target labels for each dataset are defined, followed by scaling and normalisation to ensure the features are on a comparable scale for model input.

Step 2: Augmented Data Generation in SALib Using Mean and Standard Deviation

In the SALib framework, SA methods—namely, Sobol, Morris, and FAST—necessitate a precisely defined input space from which to sample to assess the influence of individual features on model outputs. When the input features are assumed to adhere to a normal distribution, the mean and standard deviation for each feature must be specified. Using these parameters and the defined distribution type, SALib generates a representative and augmented dataset by sampling from the specified distributions. Small perturbations, in the form of controlled noise, are then introduced to the original dataset to create varied samples that reflect plausible variations in the input space.

These perturbations are method-specific: the Sobol method employs Monte Carlo sampling to estimate first-order, total-order, and interaction sensitivity indices; the Morris method generates multiple trajectories through the input space to identify both main effects

and higher-order interactions; and the FAST technique applies Fourier analysis to decompose the variance in model output attributable to each input feature. This enhanced sampling strategy enables systematic exploration of input variability and its effect on model behaviour. By evaluating the response of the model across these variations, SALib quantifies the relative importance and sensitivity of each feature. Such analysis is particularly valuable in complex models, where understanding the contribution of individual features supports interpretability, model optimisation, and informed decision-making.

Step 3. Sensitivity Analysis Using SALib

Once the models are trained, a generated sample dataset (with perturbations) is utilised as input into the trained DNN models, making predictions based on the perturbed data.

The SA techniques are applied to the predictions using the 'analyse' function from the SALib library; the analysis computes the feature importance scores for each feature based on the degree to which perturbations influence the output of the model.

Multiple iterations of training, prediction, and SA are performed to account for the stochastic nature of DNN training (due to random initialisation of weights and biases). This approach ensures robust and reliable results by averaging across several runs to mitigate the effects of randomness.

Step 4. Results Processing

Once the analysis is complete, the output from the Sobol, Morris, and FAST methods is processed to identify the most important features of each method. The results from each method are compared to evaluate the consistency of feature importance across the different techniques. This comparison helps determine which features consistently influence AD prediction across various models and methods. Figure 5- 3 visualises the schematic flowchart for Sobol, Morris and FAST techniques.

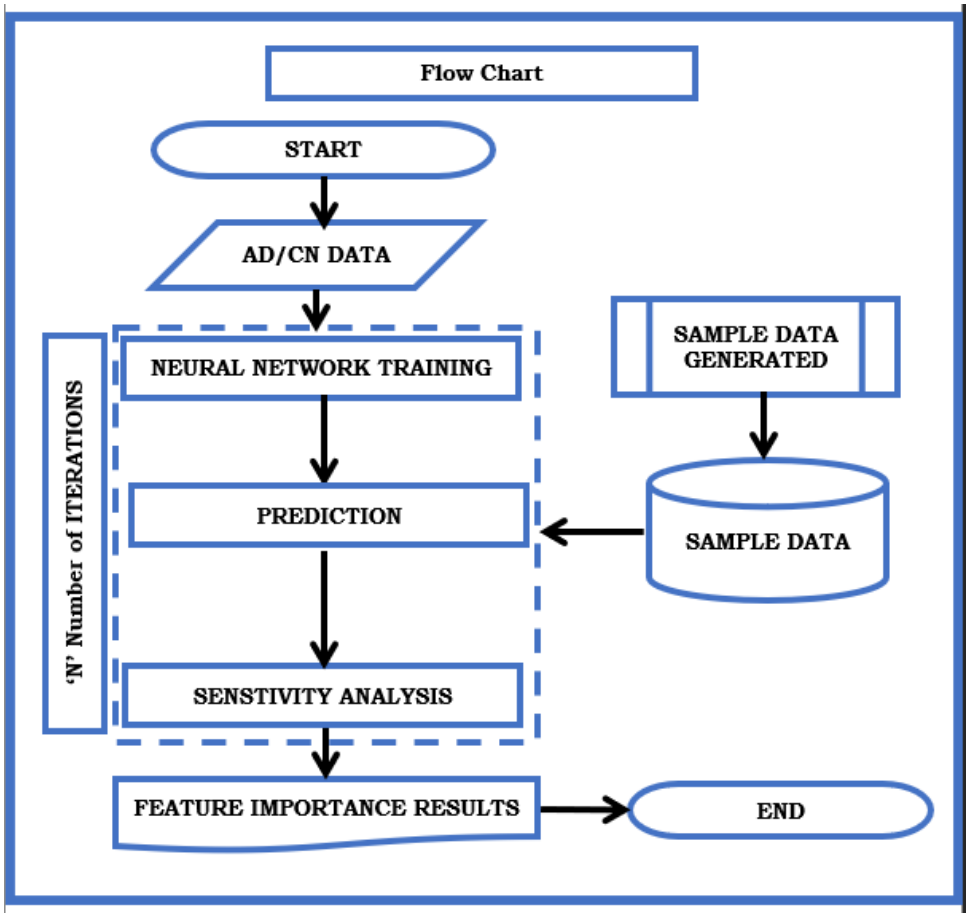

Figure 5- 3 Schematic Flowchart for Sobol, Morris and FAST techniques

5.1.2 Methodology using SHAP

This section outlines the methodology used to perform Shapley value-based SA on the DNN models, as described in the previous section. In this approach, SHAP (SHapley Additive exPlanations) using the SHAP Python library is implemented to identify the most important features in predicting AD using two distinct datasets: Dataset 1 and Dataset 2. This methodology focuses on performing SA on DNN Model 1 and DNN Model 2 and extracting the relevant feature importance scores. Figure 5- 4 visualises the Schematic flowchart for SHAP technique.

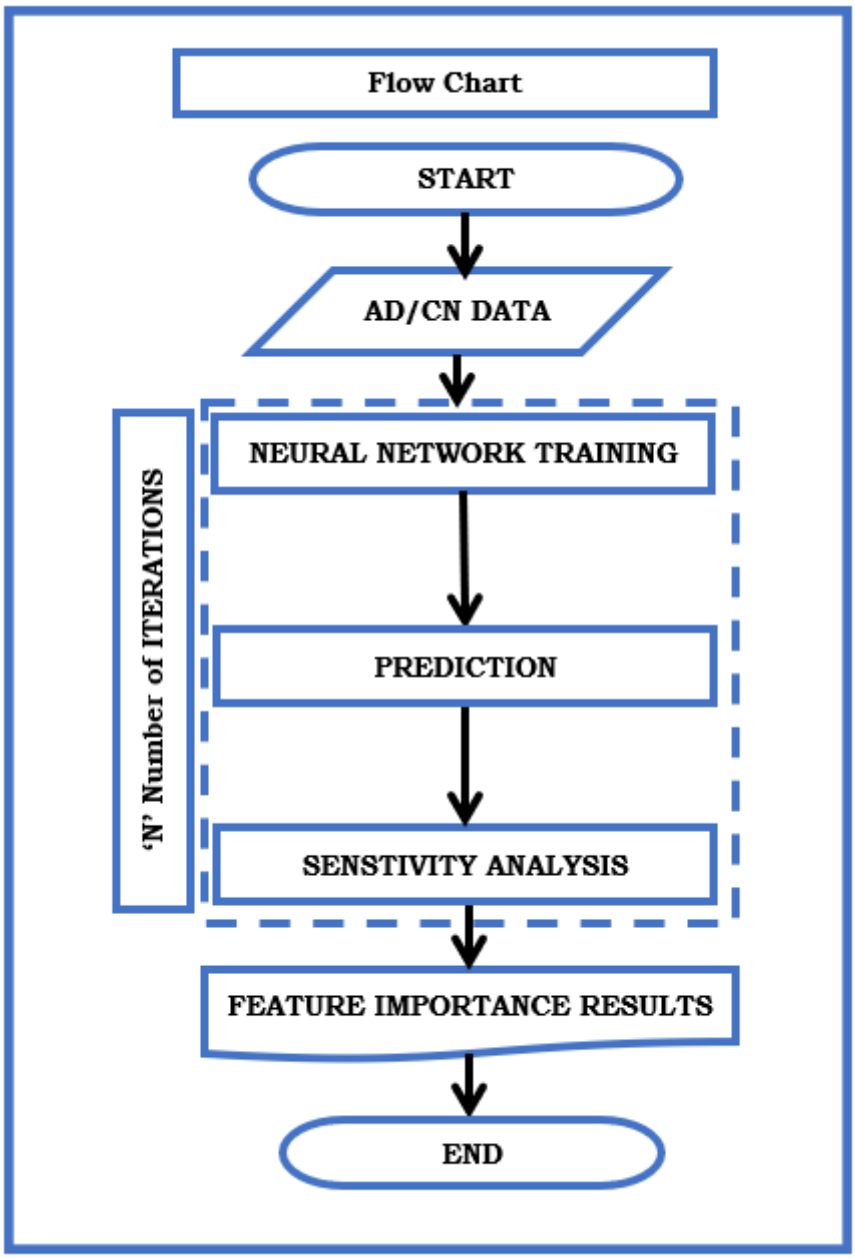

Figure 5- 4 Schematic Flowchart for SHAP technique

The DNN models are trained to utilise the respective datasets, and a SHAP analysis is performed to determine the feature importance scores. The procedure commences with the initialisation of the SHAP DeepExplainer function, which is specifically designed to elucidate the output generated by DNN models.

Subsequently, the SHAP explainer is initialised, utilising the trained DNN model and the scaled training dataset. The SHAP DeepExplainer function computes the Shapley values, thereby quantifying the contribution of each feature to the predictions of the model. These

values represent the impact each feature has on the output of the model, enabling the identification of the most influential features. Features with higher Shapley values are considered important, as they significantly impact the output of the model.

To accommodate the inherent randomness of the DNN training process—arising from factors such as random weight initialisation, mini-batch sampling, and stochastic optimisation—multiple iterations of training, prediction, and SHAP analysis are conducted. This approach ensures robust results that are not biased by arbitrary initialisations, with averaged outcomes yielding reliable feature importance scores.

The output from the SHAP analysis is processed to identify the most important features of each dataset and model. These results are post-processed to identify which features consistently influence model predictions. The final feature importance scores could be visualised using SHAP plots, such as bar and summary plots, to represent the most influential features for the DNN model.

5.2 Results and Discussion:

5.2.1 Quantitative results:

In this section, the quantitative results obtained are discussed by applying the proposed techniques to the AD datasets. Although validated on these datasets, the proposed methods are generalisable and suitable for application across other domains with high-dimensional data.

This study developed two DNN models (Model 1 and Model 2) for detecting AD using two distinct datasets: dataset 1, containing 401 features, and Dataset 2, containing 265 features. The explainability of these models was assessed through SA methods, specifically SHAP and SALib, which include Sobol, Morris, and FAST methods.

The SA was performed using both SHAP and SALib methods. Each method was executed 500 times for Dataset 1 and 300 times for Dataset 2 to account for any fluctuations in the results. These methods provided feature importance scores, which were then analysed to determine the most essential features for AD detection.

Ranking of Features

Each list was converted into a corresponding ranking pattern to compare the feature importance scores obtained from the different methods. The similarity between the rankings was determined by calculating the absolute difference between the rankings using the following equation:

$$\frac{abs(A - B)}{SNSF}$$

If A ≤ SNSF or B ≤ SNSF - Eq. (1)

Where:

- (A) is the rank from Rank list A,

- (B) is the rank from Rank list B, and

- (SNSF) is the number of selected features specified.

This equation calculates the relative discrepancy between the ranks of two lists, normalising by the number of features to facilitate comparison across different datasets. Utilising this methodology, the outcomes from the four techniques (SHAP, Sobol, Morris, and FAST) were evaluated. The average rank differences were computed to assess central tendency, thereby summarising the collective similarity across all methods.

Similarity Analysis

Upon conducting a comparative analysis, the SHAP and Sobol methodologies demonstrated a significant degree of similarity in the ranking of feature importance. The discrepancies in rankings between SHAP and Sobol were consistently minimal when contrasted with those observed between alternative methodologies. This observation culminated in the conclusion that these two methods yielded the most consistent and dependable results in identifying crucial features. To ascertain the robustness and reliability of the analysis, only the results garnered from SHAP and Sobol were selected for further scrutiny.

The results of the similarity analysis, as shown in Figure 5- 5, demonstrate the comparison of ranking patterns across different methods for the 401 features dataset and indicate the high degree of consistency between SHAP and Sobol

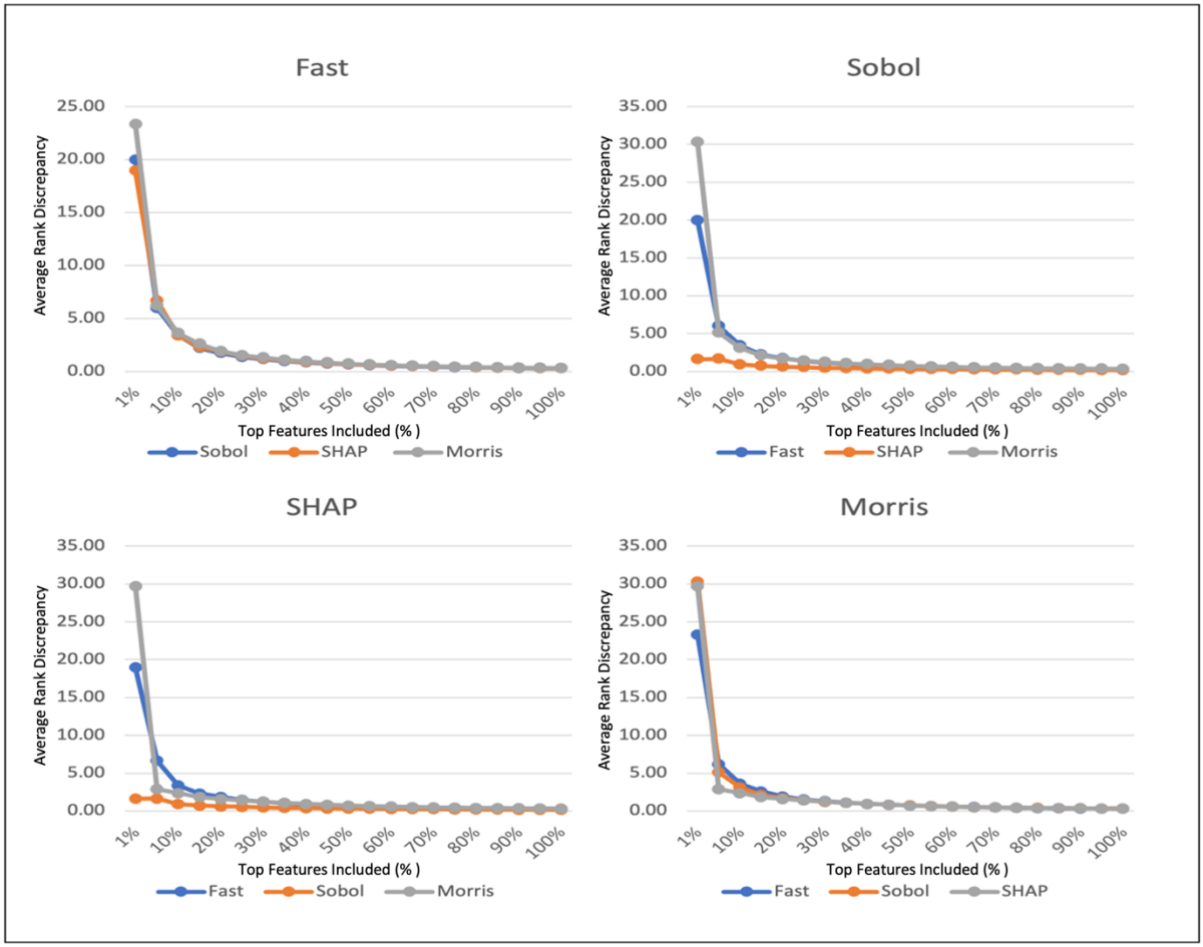

Figure 5- 5 Similarity analysis for four different approaches and 401 features dataset

Final Feature Importance Ranking

To derive the final feature importance ranking, the results from SHAP and Sobol were combined using the Rank Position Method (reciprocal rank method) (Nuray & Can, 2006), as described in equation (2):

$$r(d_j) = \frac{1}{\sum j \frac{1}{position(d_{ij})}} \quad \text{- Eq. (2)}$$

This method computes a rank score for each feature based on its position across all retrieval systems, where (n) represents the total number of features. The resulting rank scores were used to sort the features in non-decreasing order, forming the final feature importance ranking.

The final rankings were then used to identify the top 20 most important features for AD detection. These features were selected based on their position in the final combined ranking from SHAP and Sobol.

Feature Importance results in Tables

The findings outlined in this section derive from proposed SA techniques utilising the AD dataset, which serves as the validation platform for the methods. Although the methods developed are domain-agnostic and adaptable across multiple disciplines, the AD dataset was explicitly chosen to exemplify the efficacy of the approach on a complex, high-dimensional dataset.

Table 5- 1 lists the 20 most important features identified from Model 1, which was trained on Dataset 1 with 401 features. The table provides the feature names, their corresponding medical terminology, and references to relevant medical literature. These features are crucial in early AD detection, as indicated by their strong importance in the feature ranking analysis.

Table 5- 1 Feature Importance for Dataset 1

Feature Name	Medical Names	Medical Reference
Left-Inf-Lat-Vent	Temporal horn of left lateral ventricle	Vernooij and van Buchem, 2020
Right-Inf-Lat-Vent	Temporal horn of right lateral ventricle	Vernooij and van Buchem, 2020
Left_Hippocampal_tail	Hippocampal tail	Zhao et al., 2019
left_presubiculum	Pre subiculum	Carlesimo et al., 2015
left_Whole_hippocampus	Hippocampus	Rao et al., 2022
left_molecular_layer_HP	Molecular Layer Hippocampus	Scheff et al., 1996
left_subiculum	Subiculum	Carlesimo et al., 2015
right_Hippocampal_tail	Hippocampal tail	Zhao et al., 2019
lh_bankssts_volume	Banks of Superior Temporal Sulcus	Sacchi et al., 2023
lh_bankssts_thicknessstd	Banks of Superior Temporal Sulcus	Sacchi et al., 2023
lh_parahippocampal_thickness	Para Hippocampal	Van Hoesen et al., 2000
rh_paracentral_thicknessstd	Paracentral	Yang et al., 2019
right_subiculum	Subiculum	Carlesimo et al., 2015
rh_inferiorparietal_thickness	Inferior Parietal	Jacobs et al., 2012
lh_transversetemporal_meancurv	Transverse Temporal	Peters et al., 2009

Left-Amygdala	Amygdala	Poulin et al., 2011
left_hippocampal fissure	Hippocampal Sulcus	De Bastos-Leite et al., 2006
left_GC-ML-DG	Granule Cell (GC) and Molecular Layer (ML) of the Dentate Gyrus (DG)	Ohm, 2007
Right-Amygdala	Amygdala	Poulin et al., 2011
rh_inferiortemporal_volume	Inferior Temporal	Scheff et al., 2011

Table 5- 2 presents the 20 most important features identified from Model 2, which was trained on Dataset 2 with 265 features. Interestingly, there is a 60% overlap between the features identified in **Error! Reference source not found.** and **Error! Reference source not found.**, highlighting the consistency of these brain regions across distinct datasets and models. This overlap further reinforces the reliability of the proposed SA methods, demonstrating their ability to consistently identify the most informative features that drive accurate model predictions.

Table 5- 2 Feature Importance for Dataset 2

Feature Name	**Medical Names**	**Medical Reference**
Left-Inf-Lat-Vent	Temporal horn of left lateral ventricle	Vernooij & van Buchem, 2020
Right-Inf-Lat-Vent	Temporal horn of right lateral ventricle	Vernooij and van Buchem, 2020
right_Hippocampal_tail	Hippocampal tail	Zhao et al., 2019
left_presubiculum	Presubiculum	Carlesimo et al., 2015
left_subiculum	Subiculum	Carlesimo et al., 2015
left_Hippocampal_tail	Hippocampal tail	Zhao et al., 2019
left_hippocampal-fissure	Hippocampal Sulcus	De Bastos-Leite et al., 2006
lh_parahippocampal_thickness	Para Hippocampal	Van Hoesen et al., 2000
left_molecular_layer_HP	Molecular Layer Hippocampus	Scheff et al., 1996
rh_entorhinal_thickness	Entorhinal	van Hoesen et al., 1991
rh_rostralmiddlefrontal_thickness	Rostral Middle Frontal	Vasconcelos et al., 2014
rh_inferiorparietal_thickness	Inferior Parietal	Greene and Killiany, 2010
left_Whole_hippocampus	Hippocampus	Rao et al., 2022
lh_precuneus_thickness	Precuneus	Koch et al., 2022
Left-Amygdala	Amygdala	Poulin et al., 2011
Optic-Chiasm	Optic-Chiasm	Sadun & Bassi, 1990

Feature Name	Medical Names	Medical Reference
Right-Pallidum	Pallidum	Miklossy, 2011
rh_entorhinal_volume	Entorhinal	van Hoesen et al., 1991
right_presubiculum	Pre-Subiculum	Carlesimo et al., 2015
Left-Pallidum	Pallidum	Miklossy, 2011

The comparison between the two tables reinforces the importance of specific brain regions, such as the hippocampus, amygdala, and various subiculum regions, in Alzheimer's diagnosis. These findings are consistent with current medical literature that highlights the role of these structures in AD pathology. In conclusion, the SA demonstrated that the SHAP and Sobol methods yielded the most consistent feature importance rankings.

5.2.2 Discussions:

The present analysis focuses on improving the explainability of DNN classifiers by utilising various SA techniques. Several recent studies have investigated explainable deep learning approaches for Alzheimer's Disease (AD) classification using MRI-derived features. AbdelAziz et al. (2024) proposed a hybrid SECNN-RF framework that combines a Squeeze-and-Excitation CNN with a Random Forest classifier, using attention weights and saliency maps to highlight important features. While their method achieves high classification accuracy, its interpretability is relatively coarse and primarily local, lacking global feature sensitivity analysis. Kang et al. (2023) integrated CNN feature extraction with an Explainable Boosting Machine to obtain interpretable rankings of brain regions. This approach provides glass-box interpretability, but the feature importance is derived from the boosting model rather than a direct sensitivity analysis of the DNN, potentially missing higher-order interactions among features.

Table 5- 3 Comparison with recent SA techniques

Study	Model / Method	Explainability / Feature Importance	Limitation vs Proposed Ensemble Method
AbdelAziz et al. (2024)	SECNN + RF	Saliency maps / attention	Local SA only; no global sensitivity analysis

Study	Model / Method	Explainability / Feature Importance	Limitation vs Proposed Ensemble Method
Kang et al. (2023)	CNN + EBM	Feature ranking via boosting	No direct DNN sensitivity analysis.
(Jumaili & Sonuç, 2025)	Attention-CNN + Grad-CAM/LIME	Local / instance-level explanation	Not global; no MRI feature ranking.
(Junior et al., 2024)	CNN + Grad-CAM	instance-level explanation	Instance/local; no stability analysis.

Other studies have focused on local or visual interpretability of deep networks. Jumaili and Sonuc (2025) deployed an attention-based CNN with Grad-CAM and LIME for instance-level explanations, while Junior et al., (2024) used Grad-CAM to generate instance-level explanations for multi-stage AD classification. Although these methods highlight key regions in MRI images, they do not provide a systematic global ranking of tabular MRI features or assess feature stability across multiple model execution instances..

In comparison, the proposed research applies a combination of SHAP and variance-based sensitivity analysis methods (Sobol, Morris, FAST) to generate stable, global feature importance rankings for DNN classifiers trained on high-dimensional MRI datasets. This approach captures both main and interaction effects among features, is repeatable across execution instances , and is directly applicable to high-dimensional tabular MRI features, making it more robust and interpretable than prior methods. Notably, the resulting top-ranked features are consistent across two distinct datasets, demonstrating the reliability and generalisability of the proposed methods. By providing a consensus-based ranking with iterative stability analysis, this framework surpasses prior approaches in interpretability and robustness, while remaining directly applicable to tabular MRI-derived features commonly used in clinical and research settings.

5.3 Summary of the Key Findings

This chapter details the application of DNN models in critical domains, focusing on integrating SA techniques to assess the explainability of these models. The research used SHAP and SALib-based Sobol, Morris, and FAST methods to comprehensively understand the importance of features in the given dataset. Two distinct DNN models processed datasets of different sizes, offering insights into how varying amounts of data influence model performance and explainability.

The methodological approach employed in this study provides a significant contribution to the XAI field by developing advanced AI models and addressing the crucial challenge of model explainability. The research presented an integrated framework that combines DNN models with SA techniques to enhance the interpretability of complex models, which is often a challenge in DL. By utilising SHAP and SALib, this study has provided valuable insights into which features most significantly impact the predictive capabilities of DNN models. This approach sheds light on the factors that influence model outputs and provides a straightforward interpretation of the decision-making process of the AI model, which is critical in a high-stakes setting where transparency is essential.

In addition to providing valuable insights into the interpretability of AI models, the study offers a comparative analysis of SHAP and SALib techniques, offering a comprehensive evaluation of their performance and suitability for feature importance assessment. By comparing these methods, this research provides an understanding of their strengths and weaknesses, guiding future researchers in selecting the most appropriate SA technique for their specific needs.

Furthermore, this study bridges the gap between computer science and real-world deployment, demonstrating how AI methodologies can effectively address challenges that demand high levels of transparency and accountability in decision-making processes.

In conclusion, this research presents a substantial contribution to advancing AI methodologies for high-dimensional data analysis, with a particular focus on improving model explainability and reliability. By integrating DNN with rigorous sensitivity analysis techniques, this work delivers a scalable and transparent framework for developing interpretable AI systems. The methods proposed not only optimise predictive performance but also address one of the most critical challenges in modern AI: balancing model accuracy with explainability.

While the experimental validation was conducted using Alzheimer's Disease datasets, the approaches developed are broadly applicable across domains where data complexity, limited sample sizes, and decision accountability are critical. This research paves the way for future advancements in building trustworthy, generalisable, and transparent AI solutions suitable for real-world deployment.

5.3.1 Clinical Relevance

The sensitivity analysis techniques explored in this study—including SALib-based Sobol, Morris, FAST, and SHAP—offer significant computational insights into feature relevance for AD prediction using MRI scan data. These methodologies provide a robust, model-agnostic framework for identifying the most influential features in complex datasets, demonstrating how XAI methods can systematically improve the interpretability of DNN models. This combination of SA and DNN contributes to the growing need for transparent and trustworthy AI systems, particularly in domains where decisions must be explainable and verifiable.

The analysis consistently highlighted several key brain structures across both datasets, including the hippocampus, subiculum, presubiculum, amygdala, and the temporal horns of the lateral ventricles. These regions are well-established in the literature as being involved in the progression of AD. For example, hippocampal atrophy is widely recognised as an early biomarker, supported by numerous neuroimaging studies (De Bastos-Leite et al., 2006; Van Hoesen et al., 2000; W. Zhao et al., 2019). The identification of the subiculum and presubiculum aligns with Braak's neuropathological staging, which places these regions among the earliest to exhibit tau pathology (C. Macedo et al., 2023).

The inclusion of lateral ventricular structures—the left and right inferior lateral ventricles—further validates existing findings, as ventricular enlargement is often observed in neurodegenerative conditions and serves as an indirect indicator of tissue loss (Vernooij and van Buchem, 2020). Additional features such as the molecular layer of the hippocampus, hippocampal tail, and hippocampal fissure provide further anatomical precision, corroborating established volumetric studies (Scheff et al., 1996).

In a clinical context, comprehending which brain regions exert the most significant influence in predicting AD can assist healthcare professionals in concentrating their diagnostic efforts on the most pertinent biomarkers. By identifying and prioritising these essential

features, clinicians can implement targeted interventions, enhancing diagnostic accuracy and early treatment efficacy.

Furthermore, using DNN models in this study represents a significant advancement toward creating sophisticated and precise diagnostic tools for NDD. As AI advances, the potential for integrating complex datasets, including MRI scans, genetic information, and cognitive scores, is becoming increasingly paramount. The amalgamation of these data sources with rigorous analytical methodologies has the potential to enhance prediction accuracy and facilitate the development of personalised treatment strategies, which ultimately contribute to enhanced patient outcomes.

5.3.2 Future Work

Although the methodologies outlined in this study offer significant insights in XAI techniques, numerous promising avenues for future research may further refine and expand upon the obtained results here.

1) Enhanced Model Interpretability: Although SHAP offers valuable interpretability for DNN models, further research could focus on enhancing the explainability of complex models. Various techniques could be explored to enhance transparency, such as incorporating attention mechanisms or layer-wise relevance propagation. Additionally, developing methods that quantify the contribution of individual neurons or layers in an NN could provide a deeper understanding of the inner workings of DNN models and offer experts insight into how the model arrives at its predictions.
2) Model Optimisation: Although the current models are effective, there is room for optimisation. Future studies could explore hyperparameter tuning, transfer learning techniques or alternative NN architectures to enhance model performance, particularly for smaller datasets. Applying ensemble methods or hybrid models that combine multiple algorithms could enhance prediction accuracy and model robustness, particularly in cases with limited data.

In conclusion, while this study makes significant strides in applying SA to deep learning models, several areas for future research could further enhance the performance of the model and applicability. By incorporating additional model interpretability techniques and model optimisations, the potential for AI-driven tools in practical settings continues to expand.

6. Transfer Learning for Predicting Cognitive Staging in Alzheimer's Disease

This chapter focuses on developing strategies to improve classification performance in complex, high-dimensional datasets. It investigates the integration of transfer learning and autoencoder-based techniques to enhance predictive accuracy in scenarios where labelled data is limited.

The chapter addresses two key challenges. First, it addresses the challenge of data scarcity by utilising knowledge from related, larger datasets through transfer learning, which improves model robustness and stability when training samples are limited. Second, it demonstrates the importance of learning compact, non-linear feature representations using autoencoders. This enables the models to extract meaningful, abstract patterns from high-dimensional inputs that conventional algorithms often fail to capture, ultimately improving both accuracy and generalisability in challenging, real-world applications.

By undertaking these initiatives, the chapter contributes to advancing the development of efficient DL solutions for complex, high-dimensional classification tasks. It highlights the potential of innovative, cutting-edge ML techniques—such as transfer learning and autoencoders—to improve model accuracy, generalisability, and robustness in scenarios where data is limited, noisy, or challenging to obtain. These methods offer scalable and adaptable solutions that are broadly applicable across domains where data scarcity presents a significant modelling challenge.

6.1 Methodology

This section introduces the proposed novel multi-stage algorithm that employs advanced ML methods such as autoencoders and TL. Utilising the advanced techniques could help in scenarios with limited data samples. To evaluate this methodology, the AD dataset has been utilised to enhance the accuracy and prediction of MMSE scores and different stages of AD using MCI data. The data are sourced from the ADNI dataset, which includes MRI images, demographic data, genetic details, and cognitive evaluations. The dataset consisted of MCI, LMCI, EMCI, and AD, which were employed in this research.

Python version 3.9.13 was used to develop the methodology, utilising its libraries for scientific computing and data analysis. Keras version 2.9, a high-level NN API, was employed for advanced ML models in TL and autoencoders. This was supported by TensorFlow version 2.9.2 as a backend, enabling efficient computation and model deployment for NN development and fine-tuning.

Figure 6- 1 illustrates the comprehensive approach to predicting and classifying MMSE scores and stages of AD. The process starts by training a regression model to predict the ages of patients diagnosed with MCI, EMCI, and LMCI. This regression model acts as a pre-trained model for subsequent stages of the algorithm, facilitating knowledge transfer. The initial layer of the model is used to learn representation of the knowledge learned by the pre-trained model, which captures generalised features necessary for pattern analysis within the MCI dataset. These layers capture important, broad-spectrum features across the three stages of MCI.

Next, an autoencoder is trained using a combined input comprising data from MCI, EMCI, and LMCI alongside the knowledge extracted from the pre-trained regression model. Autoencoders are employed to identify the most essential features in the dataset, efficiently compressing the data to generate a reduced representation that preserves essential information.

In the final phase, an AD dataset trains a regression model that predicts MMSE scores. This model uses the AD dataset and the features extracted by the autoencoder to predict the MMSE score. The predicted MMSE scores are subsequently employed to categorise the patients into two distinct stages of AD: mild and moderate cognitive impairment. This classification reflects the progression of cognitive decline in patients. The following sub-sections will comprehensively explain each phase to facilitate replication and documentation.

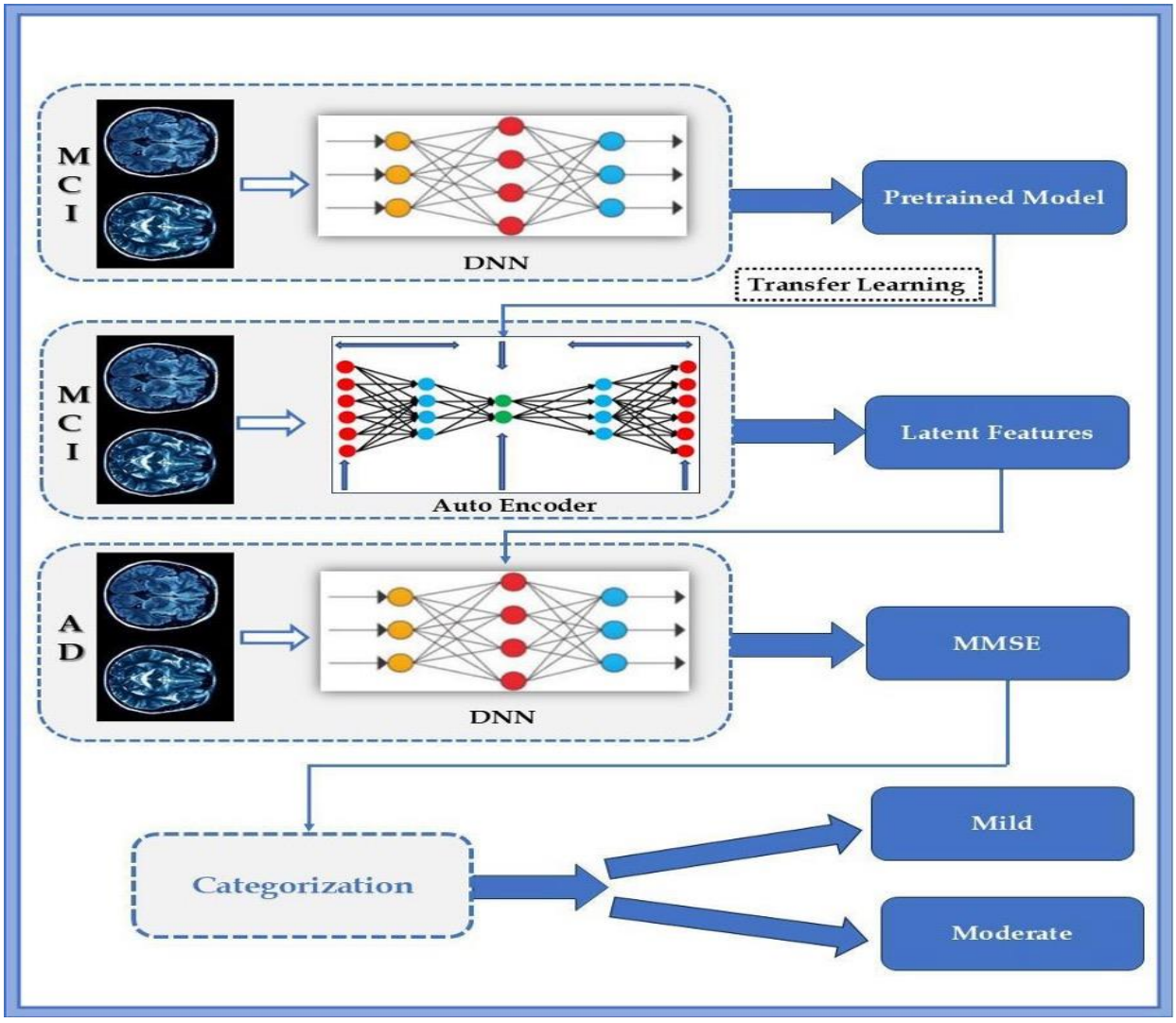

Figure 6- 1 Overall approach for using transfer learning and autoencoders to predict MMSE scores and Cognitive stages of AD

Overall, this multi-stage algorithm uses autoencoders and TL to enhance the understanding and tracking of cognitive decline in AD patients. This is achieved by improving the accuracy of MMSE score prediction and AD stage classification.

6.1.1 Regression analysis to predict the age of the MCI patients

The initial step of the multi-stage algorithm involves designing and developing a regression model to estimate the age of patients with MCI. This model employs a feed-forward NN due to its ability to capture complex, non-linear relationships within the data. The regression neural network architecture discussed in this section is represented in Figure 6-2 below.

The optimal number of hidden layers was determined using the GridSearchCV module from scikit-learn, which systematically tests various configurations to find the best

model structure. Algorithm 6- 1 summarises the optimisation process. This procedure involves training the model on the MCI dataset with various input and output layers. Table 6- 1 presents the Grid Search Results for an Age-Regression Model with 3-Fold cross-validation.

Algorithm 6- 1 Algorithm for Grid Search for Regression Model

Algorithm for Grid Search for Regression Model

Input:

```
X_train (features), y_train (age labels)
L = {2, 3, 4}                # candidate number of layers
F = {0.5, 0.6}               # units reduction factors
D = {0.2, 0.3}               # dropout rates
optimizer = RMSprop
epochs = 1000
batch_size = 32
```

Output: Best model architecture θ*

Procedure:

```
1: Initialise results_list = []
2: For each l in L do
3:     For each f in F do
4:         For each d in D do
5:             # Construct model
6:             model ← BuildNetwork(num_layers=l,
units_factor=f,dropout=d)
7:             Compile model using MSE loss and RMSprop
optimizer
8:             Perform 3-fold cross-validation on (X_train,
y_train)
9:             Compute mean_MAE and std_MAE across folds
10:            Append (l, f, d, mean_MAE, std_MAE) to
results_list
11:        End For
12:    End For
13: End For
14: Select configuration θ* with minimum mean_MAE
15: Retrain final model using θ* on full training dataset with
early stopping
16: Return θ*
```

Table 6- 1 Grid Search Results for Age-Regression Model (3-Fold CV)

Rank	Layers	Units Factor	Dropout	Mean MAE	Std MAE
1	3	0.5	0.3	**52.76**	1.1804
2	2	0.5	0.3	54.04	2.9567
3	3	0.6	0.3	54.27	0.8689
4	4	0.5	0.2	54.8	1.9178
5	4	0.6	0.2	55.12	1.6212
6	2	0.6	0.2	55.87	1.1886
7	2	0.6	0.3	55.88	1.4879
8	2	0.5	0.2	55.92	2.2143
9	3	0.6	0.2	56.45	2.6597
10	3	0.5	0.2	58.08	1.062
11	4	0.5	0.3	58.11	3.4402
12	4	0.6	0.3	59.39	1.3468

The best-performing configuration was the 3-layer network with a 0.5 reduction factor and 0.3 dropout. The resultant regression model design architecture from grid search is used to forecast the age of patients in the MCI dataset as shown in Figure 6- 2. The final design had an input layer with 401 neurones and a ReLU activation function; the number of neurones matches the number of the dataset features. Three hidden layers then follow the input layers. Every hidden layer employ ReLU for activation and includes a dropout layer with a 30% dropout rate. The initial hidden layer consists of 200 neurons, half the size of the previous layer. The next set of hidden layers has 100 neurons, and the last hidden layer contains 50 neurons. The ReLU activation function is used to introduce non-linearity and aid in identifying complex data patterns. A dropout rate of 0.3 was added after every hidden layer to enhance the generalisation of the model by reducing overfitting.

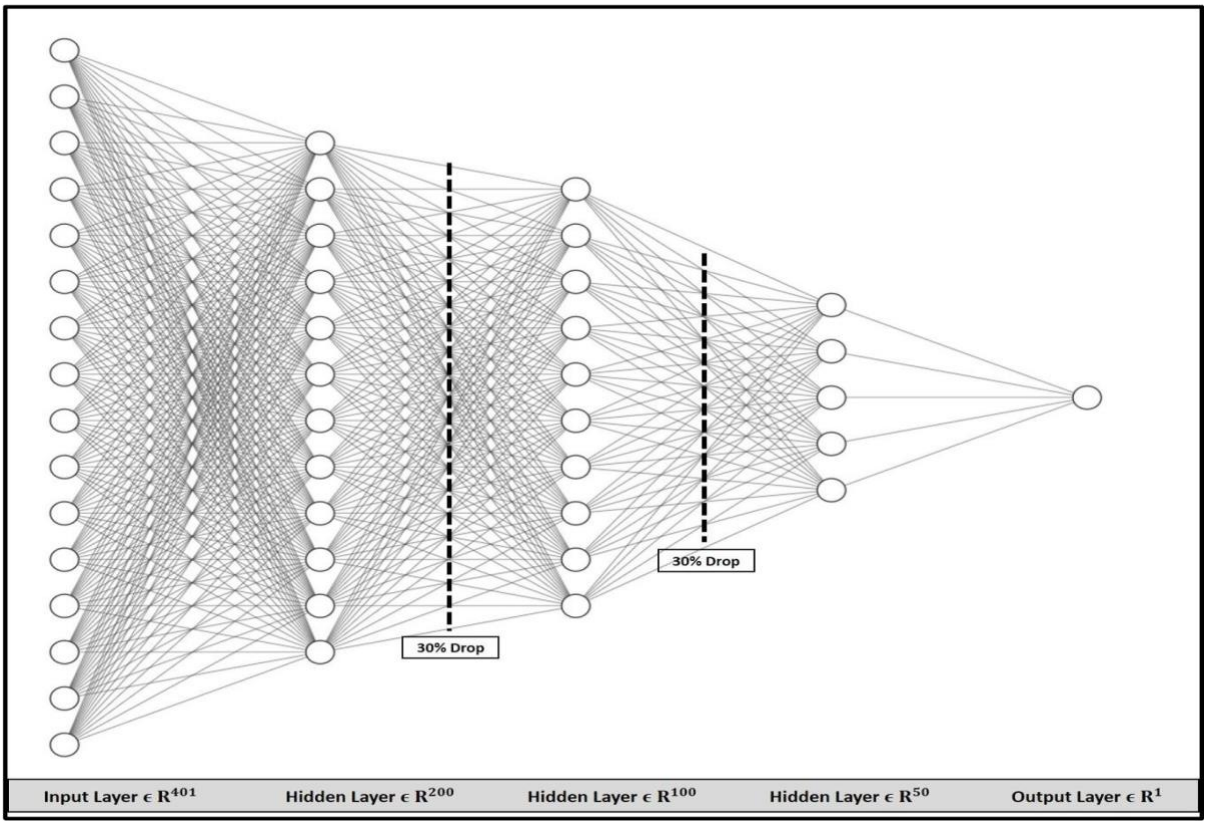

Figure 6- 2 Regression model to predict the Age of the MCI patients

Before the output layer, a dense layer with 10 neurons was included to reduce the dimensionality of the features gradually. The output layer, consisting of 1 neuron with ReLU, was used to predict the age of the patients. The mean squared error loss function was utilised to train the model, which was further optimised using the RMSprop optimiser during fine-tuning. To prevent overfitting and enable early stopping, a callback method was implemented with a tolerance of three epochs and a setback to the best weights. The model performance on new and unseen data was observed throughout the training process by dividing 30% of the data for validation across 1000 epochs. Following training, the model was evaluated on the test dataset, and its performance was assessed based on the MAE metric.

6.1.2 Transfer Learning

In the multi-stage algorithm, the following phase involves utilising the TL approach. The encoder section of a pre-trained model for predicting age was isolated and used as a feature extractor. The encoder, which includes the first layers of the model, is responsible for transforming the input data into a lower-dimensional, abstract representation that captures essential patterns and relevant features for the task. This step is vital for employing the

acquired features from a model pre-trained on a similar task, thus decreasing the need for extensive training data and computational resources.

The extraction process included developing a new model, `encoder model`, with the original model as the input and a specific chosen layer from the initial layers as the output. By choosing this particular layer, the encoder can be sure to extract a sufficiently abstract and high-level feature representation. The number of layers of the encoder was decided based on the required abstraction level and task needs.

To maintain the accuracy of the learnt representations and avoid any modifications to the encoder while fine-tuning, all layers within the `encoder model` were kept frozen. This was achieved by setting the `trainable` attribute of every layer to `False`. Keeping the layers frozen ensures that the encoder weights and biases stay consistent throughout training, preserving the robustness of the features extracted. This step is essential in TL, as it enables knowledge transfer between different domains, potentially enhancing performance and convergence speed in the new task despite having limited data. Using the pre-trained encoder, the model gains an advantage from the knowledge embedded into the original network, establishing a strong foundation for additional training and fine-tuning on the target dataset.

6.1.3 Autoencoder:

The next stage in the development process included designing and improving an autoencoder architecture that can learn and extract important features from the MCI dataset.

Autoencoders were chosen over recent architectures such as Transformers due to their ability to learn compact, low-dimensional representations from limited, highly structured datasets. In contrast to Transformer-based models, which typically require large-scale training datasets and significant computational resources, autoencoders provide stable, data-efficient feature extraction that aligns with the constraints of medical and MRI-derived datasets. They also provide controllable latent-space behaviour, enabling structured regularisation and reconstruction-based constraints, which are particularly valuable when preserving subtle anatomical patterns.

TL was used to enhance the ability of the encoder to extract features even further. TL is a robust method where a model, previously trained on one task, is adjusted and used for a similar task. In this development stage, an encoder pre-trained and derived from a fine-tuned regression model was incorporated into the autoencoder. Precisely, a layer of 50 neurons was linked to the autoencoder encoder output, integrating information from the pre-trained model. Combining the capabilities of the autoencoder and the pre-trained encoder enabled the autoencoder to use previously learnt patterns and representations to enhance the feature extraction process.

To enhance the model stability and accelerate convergence during training, batch normalisation layers were included in both the encoder and decoder. These layers standardised the output of every dense layer, ensuring the resilience of the model to fluctuations in the input data distribution. This not only enhanced the speed of training but also aided in preventing overfitting by controlling the learning process of the model. Batch normalisation enabled increased learning rates by addressing the issue of disappearing or amplifying gradients, leading to enhanced learning efficiency for the model.

The autoencoder was trained using the RMSprop optimiser, which is well-suited for tasks involving large datasets and complex models. The learning rate was set to 0.0001, a conservative value that ensured slow, steady enhancements in the model weights to minimise the Mean Squared Error (MSE) loss. MSE was used as the primary loss function, directly measuring the difference between the original input data and the reconstructed output. The training was conducted over 1000 epochs, with a batch size of 64, effectively balancing computational efficiency and memory utilisation. Additionally, early stopping was employed with a patience of 5 epochs, ensuring that the model would halt training if no significant improvement in the loss metric was observed, preventing overfitting and reducing computational resources.

A grid search strategy was used to find the best autoencoder architecture. Grid search involves systematically testing a pre-defined set of hyperparameters to find the optimal model configuration. Algorithm 6- 2, below, sets out the procedure.

Algorithm 6- 2 Algorithm for Grid Search for Autoencoder

Algorithm for Grid Search for Autoencoder

Input:
`X_train, hyperparameter sets (layers, neurons, batch size, learning rate)`

Output: `Best autoencoder configuration θ*`

Procedure:
1: Initialize results_list = []
2: For each combination of layers and neurons do
3: Build autoencoder with current configuration
4: Compile model with MSE loss and RMSprop
5: Perform k-fold cross-validation on X_train
6: Compute mean_score and std_score across folds
7: Append (layers, neurons, batch_size, mean_score, std_score) to results_list
8: End For
9: Select θ* with minimum mean_score
10: Retrain autoencoder with θ* on full dataset with early stopping
11: Return θ*

Each configuration generated by the grid search was evaluated using k-fold cross-validation to ensure robust performance across different subsets of the MCI dataset. The mean and standard deviation of the reconstruction score (MSE-based) were recorded for every combination of encoder and decoder architectures. The resulting scores provide insight into the trade-off between model complexity and reconstruction accuracy, enabling the selection of an optimal autoencoder configuration that balances efficient feature extraction with generalisability. Table 6- 2, below, sets out the results of the grid search for the autoencoder.

Table 6- 2 Grid Search Results for Autoencoder Model (3-Fold CV)

Rank	Encoder Architecture (Neurons)	Decoder Architecture (Neurons)	Mean Score	Std Dev of Score
1	[200, 100]	[100, 200]	0.7089	0.0121
2	[100, 50]	[100, 200]	0.7109	0.0134
3	[150, 50]	[100, 200]	0.7148	0.0138
4	[100, 50]	[50, 150]	0.7269	0.0132
5	[200, 100]	[50, 150]	0.7305	0.0161
6	[150, 50]	[50, 150]	0.7315	0.0176
7	[200, 100]	[50, 100]	0.7356	0.0185
8	[150, 50]	[50, 100]	0.7389	0.0231
9	[100, 50]	[50, 100]	0.7404	0.0187

Figure 6- 3 illustrates the final autoencoder structure, which combines the optimal encoder/decoder design with transfer learning. This configuration efficiently captures latent patterns from the MCI dataset, providing features for downstream MMSE prediction and cognitive stage classification.

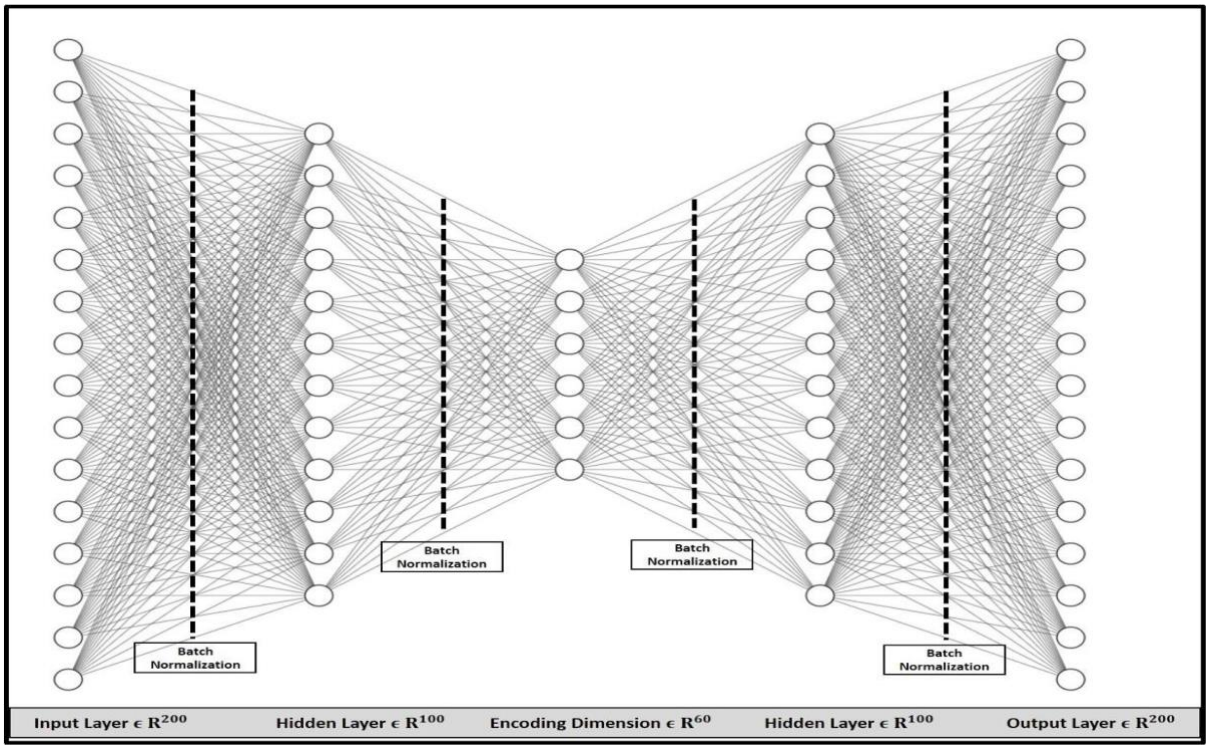

Figure 6- 3 Architecture of the autoencoder

Ultimately, the optimised autoencoder structure achieved an excellent balance between model complexity and computational efficiency. By extracting and encoding critical features from the MCI dataset, the autoencoder demonstrated an ability to capture latent patterns that were not readily apparent in the raw input data. This robust feature extraction capability, combined with the use of TL, significantly enhanced the performance of the model in downstream predictive tasks.

The utilisation of TL and the grid search-driven optimisation technique resulted in a robust autoencoder structure that played a crucial role in the complex ML process created in this research. The autoencoder-acquired characteristics were then used further in a predictive model to approximate MMSE scores and categorise cognitive conditions in AD patients. Being able to reliably forecast cognitive decline and classify patients according to their cognitive status is clinically significant. The autoencoder-based feature extraction method contributes to improving the accuracy and reliability of these forecasts.

The resultant architecture of the autoencoder also included TL and grid search optimisation. The capability to derive significant and coherent attributes from complex

datasets is anticipated to enhance significantly the accuracy and dependability of the models, ultimately leading to results.

6.1.4 Regression followed by categorisation:

The final phase of this multi-stage algorithm aimed to predict MMSE scores, a critical measure of cognitive function, in AD patients. This was achieved by utilising the output of feature representations from an autoencoder to perform regression analysis and subsequently classify these MMSE scores into two distinct cognitive impairment categories: mild and moderate. The accurate prediction and classification of MMSE scores is vital, as it can significantly contribute to the timely and effective management of cognitive decline in AD patients.

Algorithm 6-3, below, sets out the procedure of the grid search applied to the regression model used to predict the MMSE score. This is followed by Table 6-3, which presents the experimental results for various configurations of the model.

Algorithm 6- 3 Algorithm for Grid Search for Regression Model with MMSE Score

Algorithm for Grid Search for Regression Model with MMSE scores

Input:

```
X_train, y_train (features and MMSE scores)
L = {2, 3, 4}            # candidate number of hidden layers
U0 = {300, 400}        # initial units in first hidden layer
F = {0.5, 0.6} # unit reduction factors for subsequent layers
optimizer = RMSprop
epochs = 500
batch_size = 64
```

Output:

Best regression model configuration θ*

Procedure:

```
1: Initialize results_list = []
2: For each l in L do
3:     For each u0 in U0 do
4:         For each f in F do
5:             # Construct model
6:                     model ← BuildNN(num_layers=l, initial_units=u0,
reduction_factor=f, dropout=0.3)
7:             Compile model with MSE loss and RMSprop optimizer
8:             Perform 3-fold cross-validation on (X_train, y_train)
9:             Compute mean_score and std_score across folds
10:            Append (l, u0, f, mean_score, std_score) to results_list
11:        End For
12:    End For
13: End For
14: Select configuration θ* with minimum mean_score
15: Retrain final model using θ* on full training dataset with early stopping
16: Return θ*
```

Table 6- 3 Grid Search Results for Regression followed by categorisation (3-Fold CV)

Rank	Layers	Initial Units (U0)	Unit Reduction Factor	Mean Score	Std Dev of Score
1	2	400	0.6	4.325	0.5073
2	2	400	0.5	4.459	0.4923
3	2	300	0.5	4.522	0.4514
4	2	300	0.6	4.555	0.5115
5	3	300	0.6	4.636	0.489

Rank	Layers	Initial Units (U0)	Unit Reduction Factor	Mean Score	Std Dev of Score
6	3	400	0.5	4.638	0.4756
7	3	400	0.6	4.726	0.6418
8	3	300	0.5	4.733	0.4368
9	4	400	0.6	4.855	0.2438
10	4	300	0.6	4.873	0.8139
11	4	300	0.5	5.072	0.14
12	4	400	0.5	5.092	0.063

The optimisation of these parameters enabled the model to capture the non-linear relationships inherent in the disease progression, leading to enhanced accuracy and stability in MMSE score predictions. Figure 6- 4 illustrates the NN architecture for regression model.

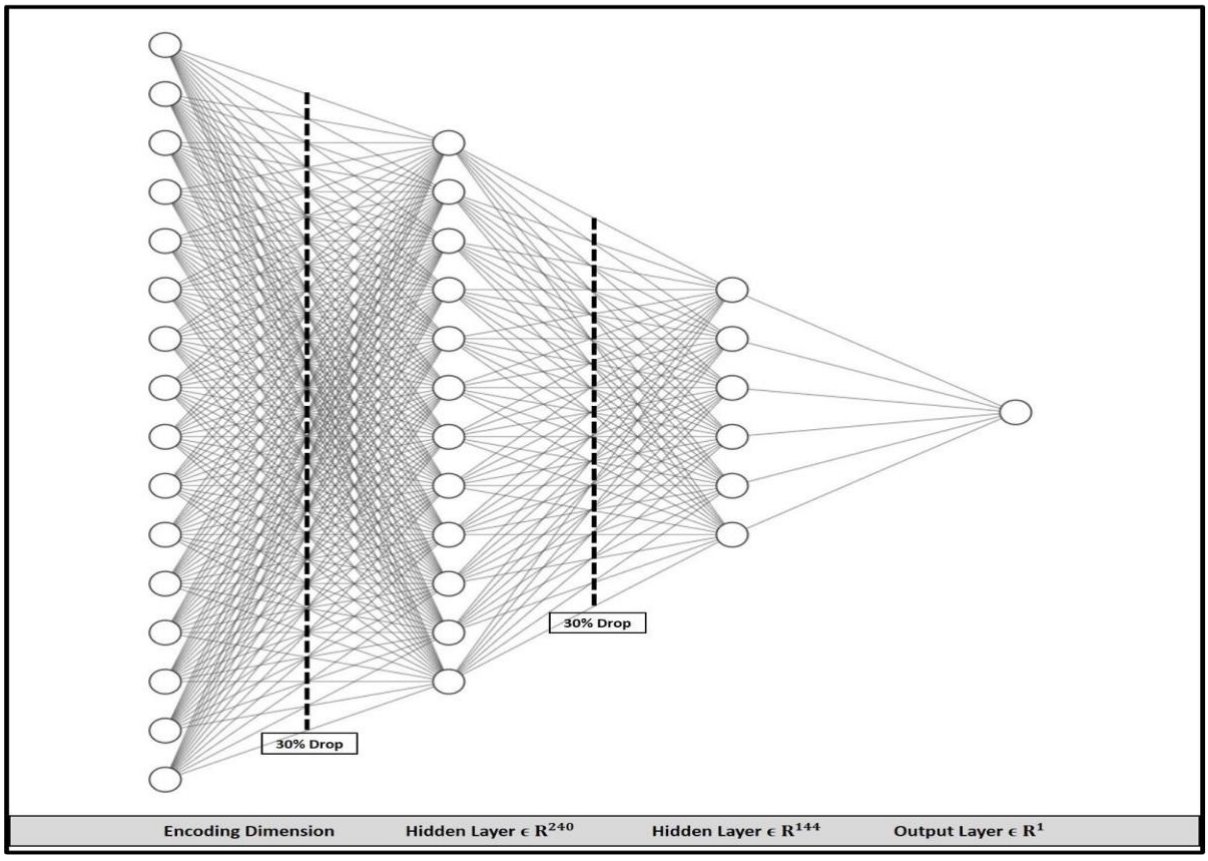

Figure 6- 4 NN architecture for Regression model

To further guard against overfitting, early stopping was implemented with a patience of 10 epochs. Early stopping is another regularisation technique that halts training if the performance of the model on a validation set stops improving for a specified number of epochs or iterations. By doing this, the model avoids overfitting to the training data while retaining weights that produce the best generalisation performance on unseen data. The patience parameter was set to 10, enabling the model enough time to explore potential improvements in performance while preventing overtraining. The model was trained for up to 500 epochs, with a batch size of 64, which balanced training performance and memory utilisation. Using mini-batches during training enabled the model to update its weights frequently, leading to faster convergence.

After training, the predictive performance of the model was evaluated on a separate test dataset. The evaluation metrics included Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and MSE. These metrics are commonly used in regression tasks to quantify the difference between predicted and actual values. MAE provides a straightforward error measure by calculating the average absolute difference between predictions and actual values, while RMSE and MSE assign weight to penalise significant errors heavily. This combination of metrics ensured a comprehensive evaluation of the accuracy and reliability of the model in predicting MMSE scores.

To classify the predicted MMSE scores into cognitive impairment categories, the regression predictions of the model were rounded to the nearest whole number. The rationale was to map the continuous MMSE predictions into discrete categories that reflect clinically meaningful levels of cognitive impairment. Based on established MMSE thresholds, the rounded scores were classified into two categories: mild cognitive impairment and moderate cognitive impairment. These categories are clinically significant because they represent different stages of cognitive decline, with MCI often being an early indicator of AD progression.

The classification accuracy of the model was evaluated by comparing the predicted cognitive categories to the actual cognitive status labels in the test data. A confusion matrix was generated to assess the ability of the model to distinguish between mild and moderate cognitive impairment correctly. The confusion matrix provided a detailed breakdown of the classification performance of the model, highlighting not only the accuracy but also any errors

made in predicting the cognitive status of patients. This analysis was particularly valuable in assessing the clinical utility of the model, as the ability to classify cognitive impairment levels accurately could directly impact patient care and treatment decisions.

The final step in the multi-stage process was to compute average performance metrics, including MAE, RMSE, MSE, and classification accuracy, over multiple model execution instances. This step was crucial for ensuring the robustness and reproducibility of the results. By averaging the metrics over multiple iterations, any variability in the model performance due to random initialisation or other factors could be mitigated, leading to reliable conclusions about the predictive capabilities of the model.

In summary, the last stage of the multi-stage method successfully integrated regression analysis and classification to predict MMSE scores and categorise them into mild or moderate cognitive impairment. Using autoencoder-generated feature representations and a well-optimised NN architecture enabled the model to capture complex, non-linear patterns in the AD data. The grid search optimisation, regularisation techniques such as dropout and early stopping, and careful tuning of hyperparameters contributed to the strong performance of the model. The evaluation metrics demonstrated that the model could reliably predict MMSE scores and classify patients into relevant cognitive categories. It is a valuable tool for clinical decision-making in managing AD. By advancing the state-of-the-art predictive modelling for neurodegenerative diseases, this approach holds promise for improving patient care and supporting clinicians in the early detection and treatment of cognitive decline.

6.2 Results and Discussion:

6.2.1 Quantitative results

In this section, the quantitative results obtained are discussed by applying the proposed techniques to the NDD datasets. Although validated on these datasets, the proposed methods are generalisable and suitable for application across other domains.

The initial regression model was designed to predict the age of patients with MCI, EMCI, and LMCI, and it performed well. During training, the model achieved a Mean Absolute Error (MAE) between 6 and 7 years, demonstrating reasonable accuracy. Even though the

testing MAE increased to 12 years, this result is still considered good, particularly given the complexity of the data and the challenge of predicting the age of cognitively declining patients. The ability of the model to perform reasonably well on unseen data highlights its robustness and potential for utilising it for the TL process.

In the TL approach, the second layer from the top was selected for its ability to capture the most generalised features from the MCI dataset. These features, abstracted from deeper layers of the model, are highly transferable and valuable for related tasks. For this purpose, 50 neurons were used to represent the features learned from the model, providing an optimal balance between complexity and generalisation. This enabled the model to adapt effectively in predicting outcomes such as cognitive decline with minimal fine-tuning, enhancing its overall performance.

For the autoencoder, the input data was a combination of MCI data and the 50 neurons derived from TL. The autoencoder was trained with early stopping-to-halt training when no further loss improvement was observed. A reconstruction error distribution histogram was plotted to assess the autoencoder performance. On this histogram presented in Figure 6- 5**Error! Reference source not found.**, the X-axis represents the error, while the Y-axis shows the frequency of errors. As illustrated, most errors are minimal, clustering around the 0.25 to 0.50 range, with very few values exceeding 1, indicating effective reconstruction performance.

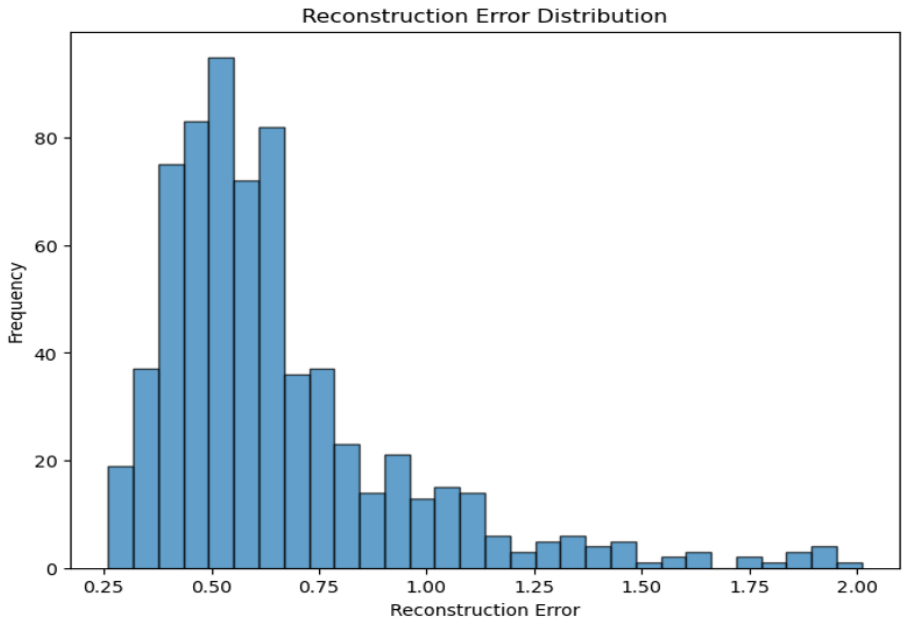

Figure 6- 5 Reconstruction Error Distribution for the Autoencoder

In the autoencoder, the reconstruction error represents the difference between the original input and its reconstruction after passing through the encoder-decoder process. The reconstruction error distribution histogram helps to visualise how well the autoencoder performs by plotting error values on the X-axis and their frequency on the Y-axis. A low reconstruction error suggests that the autoencoder has effectively learned the essential features from the data. In this case, most errors fall between 0.25 and 0.50, indicating strong model performance, with only a few values exceeding 1, suggesting minimal outliers or poor reconstructions. This distribution confirms that the autoencoder has successfully captured the important patterns in the MCI data.

The regression model for predicting MMSE scores showed strong performance, using early stopping and training over 10 iterations to prevent overfitting. The model achieved an average MAE of 3.51 with a standard deviation of 0.12, demonstrating consistent predictions. Additionally, the RMSE was 4.53, and the MSE was 20.54, reflecting accurate predictions of cognitive decline. Following the regression, the predicted MMSE scores were classified into mild and moderate categories, achieving an overall accuracy of 73.26% with a standard deviation of 3.93 across 10 iterations. The performance consistency of the model over multiple runs was visualised using a line graph in Figure 6- 6, illustrating reliable outcomes across all stages of the process. This combination of regression and classification highlights the robustness of the model in predicting and categorising disease severity.

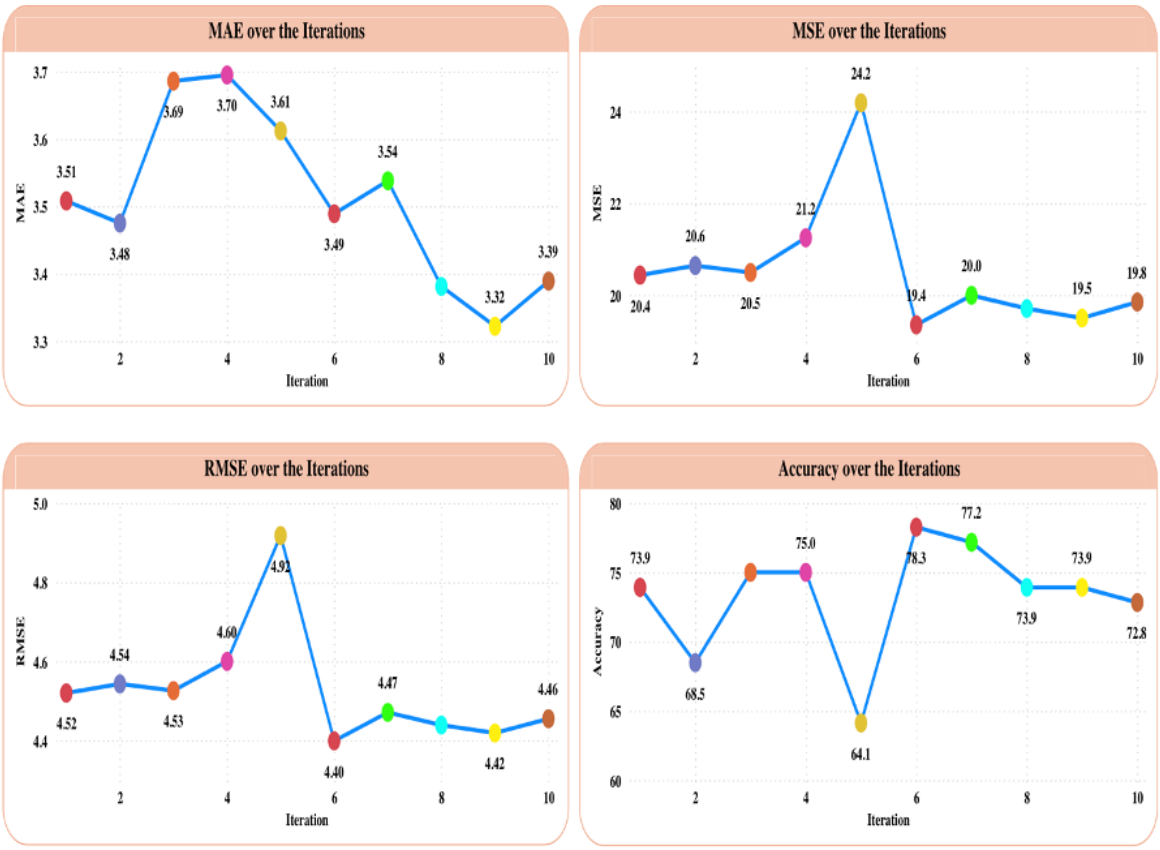

Figure 6- 6 Performance of the last stage over 10 iterations

Further, in the analysis of diagnostic accuracy for distinguishing between mild and moderate cognitive decline, an average of all 10 confusion matrices was derived and plotted below in Figure 6- 7. This provides a comprehensive overview of the performance of the model.

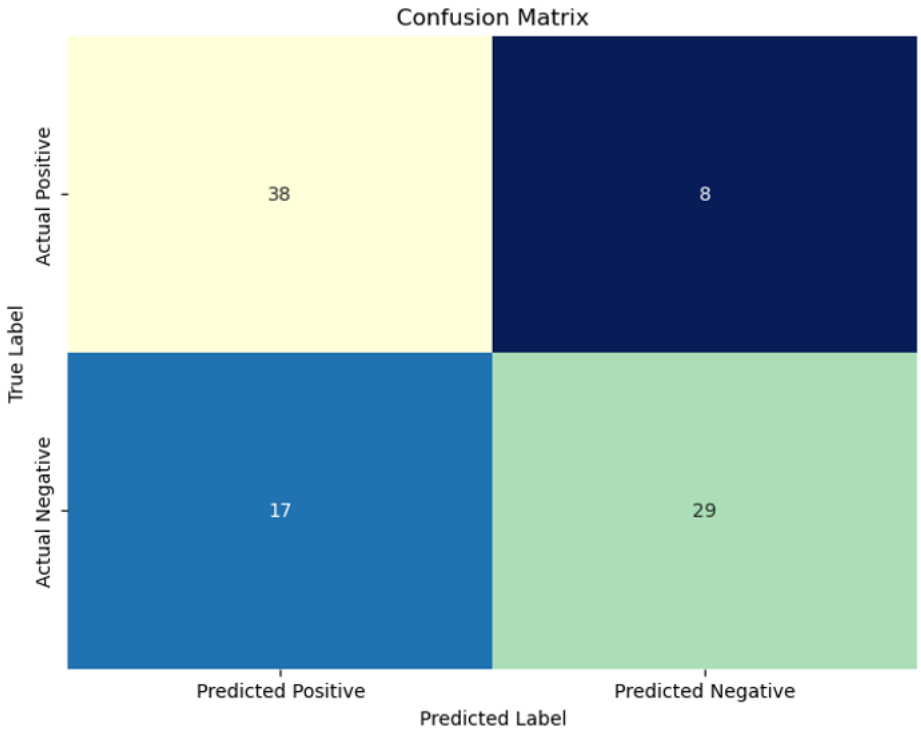

Figure 6- 7 Average value of Confusion Matrix for the Categorisation of MMSE Scores over 10 iterations

The true negative rate (TN) was recorded at 38.4, indicating a substantial number of correctly identified mild cognitive decline. The model also exhibited 28.7 true positives (TP), accurately identifying individuals with moderate cognitive decline. However, there were 7.6 false positives (FP), representing instances where individuals were incorrectly classified as having the condition, and 17.3 false negatives (FN), highlighting cases where moderate conditions were missed and misclassified as mild. These values underscore the importance of balancing sensitivity and specificity in diagnostic evaluations.

6.2.2 Discussions

A comprehensive comparison of model performance was conducted to evaluate the effectiveness of the proposed multi-stage algorithm in predicting cognitive stages of AD progression, particularly regarding several baseline and state-of-the-art models. This evaluation was critical to understanding how well the new approach, which integrates advanced techniques such as TL and autoencoders, performs compared to traditional models that rely solely on structural MRI data or different combinations of NN architectures. By benchmarking the proposed algorithm against established models, the study aims to highlight

its strengths in improving accuracy, reducing error margins, and offering reliable predictions, ultimately contributing to the broader field of transfer learning modelling.

As a first step, a baseline regression model was constructed using the same architecture as the final regression model from the proposed approach. However, this baseline model was exclusively trained and evaluated on the AD dataset with MMSE scores as the target without pretraining or enhancement techniques. The architecture of the baseline model consisted of two hidden layers, each incorporating a dropout layer with a dropout rate of 0.3 to reduce the risk of overfitting. The first hidden layer comprised 401 neurons, matching the input dimensionality of the AD dataset. Subsequent layers followed a reduction strategy, where each had 60% of the neurons of the preceding layer. Unlike the proposed multi-stage algorithm, which utilised TL and autoencoders for enhanced generalisation and performance, this baseline model used only AD data, devoid of advanced pretraining methods.

After completing the regression task in the baseline model, a classification step was introduced to provide a comprehensive model evaluation of the performance. The results were averaged over 10 runs to capture variability. The regression task yielded an MAE of 4.9, with a standard deviation of 0.1919, demonstrating a consistent performance across runs. In terms of classification, the model achieved a mean accuracy of 61.08%, with a standard deviation of 2.21. These results serve as a reference for gauging the performance improvements in sophisticated models.

In contrast, the multi-stage algorithm proposed in this research demonstrated a significant performance improvement over the baseline model. This model also utilised structural MRI and age data sourced from ADNI, AIBL, and IXI datasets, but with the added benefit of TL and autoencoders. These techniques helped enhance diagnostic accuracy, reduce training time, and lower the dependence on large amounts of training data. The model was evaluated on the AD dataset, predicting MMSE scores, key indicators of cognitive decline and disease severity in AD patients. By incorporating data from patients with MCI, the model effectively predicted MMSE scores, offering an accurate assessment of disease progression.

The integration of TL enabled feature extraction by using pre-trained models. At the same time, autoencoders facilitated efficient data representation, resulting in a notable accuracy of 73.26% with a standard deviation of 3.92. This marks a substantial improvement

over the baseline model, underscoring the advantages of combining these advanced techniques. The enhanced performance of the proposed model highlights the ability of TL and autoencoders to capture complex patterns in the data, ultimately leading to precise predictions of AD progression and severity.

These results underscore the gap in performance between the baseline model and the primary approach, reaffirming the benefits of using TL and autoencoders to enhance prediction accuracy and minimise error margins in AD cognitive progression forecasting.

Li et al. (2015) developed a DL model to classify AD and MCI patients using MRI data, combining DL with a stability selection method to enhance feature extraction. This approach enabled the model to handle the variability and noise typical in medical imaging data. The process began with PCA, which captured unsupervised latent feature representations from the MRI scans, providing a detailed understanding of brain structure. These features were further refined using the stability selection method, which applied Lasso regularisation to optimise feature selection by minimising the cost function. The refined features were then used in a multi-task DL model with dropout, incorporating additional labels such as MMSE and ADAS-Cog scores to enhance prediction accuracy. Finally, the outputs of the multi-task model were passed to an SVM classifier for the final classification of AD and MCI patients. This combined approach achieved a classification accuracy of 70.1% with a standard deviation of 2.3, demonstrating a significant improvement over baseline models by utilising deep feature learning and stability selection.

In Oh et al. (2019), the model approach involved two DL architectures: a Convolutional Autoencoder (CAE)-based model and an Inception CAE (ICAE)-based model aimed at classifying progressive MCI (pMCI) vs. stable MCI (sMCI). The CAE model consisted of 3D convolutional and fully connected layers to extract meaningful features from MRI data. The ICAE model incorporated an inception module, which used multi-scale convolutional kernels to capture visual representations at different levels. Both models utilised TL by initialising their convolutional layers with pre-trained weights from an AD vs NC classification task, followed by supervised fine-tuning for pMCI vs MCI classification.

In the CAE model, the architecture had three 3×3×3 convolutional layers with ReLU activations, max-pooling, and Gaussian dropout to prevent overfitting. It included fully

connected layers with 32 and 16 nodes and a final output layer with two nodes for classification. The ICAE-based model, on the other hand, included two convolutional layers followed by an inception module that combined different kernel sizes (1×1×1 and 3×3×3) to enhance feature extraction at multiple scales. The output from the inception module was directly fed into the classifier without fully connected layers, optimising performance for the pMCI vs. sMCI task. Both models applied supervised TL to utilise shared features between the AD vs. NC task and the pMCI vs. sMCI task. The use of pre-trained weights enabled the network to attain enhanced generalisation capability despite the smaller dataset size for the pMCI vs sMCI classification, resulting in enhanced accuracy—73.23% for the CAE model and 73.95% for the ICAE model. This demonstrated that TL could effectively bridge performance gaps when dataset size is limited and the classification task is inherently challenging.

Methods	**Model**	**Data**	**Accuracy (in %)**
Baseline Model	DL	Structural MRI	61.08 ± 2.2
(Li et al., 2015)	DL	Structural MRI	70.1% ± 2.3
CAE (Oh et al., 2019)	DL	Structural MRI	73.23 ± 4.21
Multi-Stage Algorithm	DL	Structural MRI	**73.26 ± 3.93**
ICAE (Oh et al., 2019)	DL	Structural MRI	73.95 ± 4.82

Table 6- 4 highlights the accuracy of each model when trained on similar datasets, predominantly using structural MRI data, with all models incorporating TL. These results highlight the competitiveness of the proposed model, which demonstrates a marked improvement over baseline methods and performs comparably to existing advanced models. This demonstrates the efficacy of TL techniques in scenarios with limited datasets.

Table 6- 4 Comparison of results between current and existing models

Methods	**Model**	**Data**	**Accuracy (in %)**
Baseline Model	DL	Structural MRI	61.08 ± 2.2
(Li et al., 2015)	DL	Structural MRI	70.1% ± 2.3
CAE (Oh et al., 2019)	DL	Structural MRI	73.23 ± 4.21
Multi-Stage Algorithm	DL	Structural MRI	**73.26 ± 3.93**
ICAE (Oh et al., 2019)	DL	Structural MRI	73.95 ± 4.82

Although a range of DL models has been applied to structural MRI data to predict Alzheimer's Disease, only a small subset of existing work has focused specifically on predicting

MMSE scores using transfer learning or autoencoder-based feature representations. Consequently, the proposed multi-stage pipeline occupies a distinct methodological space: it combines a regression-based MMSE prediction model with transfer-learning and an autoencoder-derived latent feature space, providing a more generalisable representation under limited sample availability. This design positions the model as one of the few AD-focused frameworks to utilise stage-informed feature transfer for cognitive-score estimation explicitly.

A broader context of recent literature presented in Table 6-5, below, has examined transfer learning in prognostic tasks, particularly in predicting conversion from MCI to AD, which is closely related to cognitive decline and strongly correlated with MMSE progression. Models such as those proposed by Dhinagar et al., (2022; S.-C. Huang et al., (2023); Khan et al., (2022) apply transfer learning, 3D CNNs, or autoencoder-driven latent spaces to capture structural markers of disease progression. While these methods demonstrate a strong performance in binary conversion prediction, they do not estimate clinical scores directly and thus address a fundamentally different problem. Nevertheless, their success highlights the underlying rationale of the approach proposed in this research study, in that feature distributions learned from MCI subjects encoding a cognitively meaningful variation that can be transferred to improve downstream MMSE prediction in AD.

Table 6- 5 TL-based MCI to AD Conversion

Study	**Model**	**Data**	**Accuracy / Key Results**
Dhinagar et al., (2022)	Transfer learning (pretrained CNNs, fine-tuned)	ADNI + independent tests (AIBL, MIRIAD, OASIS)	91.3% with CV; 94.2% / 87.9% on external datasets
Khan et al., (2022)	VGG-based TL on grey-matter slices	ADNI (NC, EMCI, LMCI, AD)	~97.9% multiclass accuracy
S.-C. Huang et al., (2023)	Transformer (pretrained + fine-tuned)	ADNI / AIBL	99.6% (ADNI), 94.0% (AIBL)

These works collectively demonstrate rapid progress in MRI-based transfer learning for disease-stage classification but differ substantially from clinical-score prediction tasks, underscoring the relative scarcity of TL-based MMSE regression pipelines.

More recent studies presented in Table 6-6, below, have predicted MMSE or cognitive impairment scores from neuro-imaging and multimodal data. However, these pipelines generally lack transfer-learning, integrate an autoencoder-derived latent space, or target broader multimodal contexts rather than MRI-derived structural representations. As a result, they are methodologically related but not directly comparable to the multi-stage approach proposed in this research.

Table 6- 6 Non-TL MMSE Prediction Models

Study	Method	Data	Output Type
Dong et al., (2020)	Patch-based CNN + multi-task learning	MRI	MMSE regression
Liu et al., (2024)	Multi-task network	MRI	MMSE + diagnosis
Ilias and Askounis, (2022)	Multimodal regression model (speech/text/vision)	Non-MRI	MMSE regression
Bass et al., (2023)	ICAM-reg (VAE-GAN)	MRI	MMSE regression

Taken together, these findings emphasise that while MMSE prediction and disease-stage classification have both been explored in the literature, very few models combine transfer learning, autoencoder-derived representations, and regression within a unified multi-stage framework. Existing approaches either rely solely on multi-task learning, focus exclusively on MCI-to-AD conversion, or use alternative modalities such as speech and text. The method proposed in this thesis therefore contributes a distinct architectural strategy—one that leverages MCI-informed latent representations to enhance MMSE regression performance in AD patients, addressing a gap in current transfer-learning research for clinical-score prediction.

6.3 Summary of the Key Findings

The proposed multi-stage algorithm offers several advantages that enhance its performance in predicting and classification tasks. One key benefit is enhanced predictive accuracy, achieved by integrating advanced techniques such as TL and autoencoders. These methods enable the model to extract meaningful features from complex datasets, resulting in a high accuracy rate of 73.26%, which exceeds that of baseline models. This enhanced precision is particularly valuable for applications involving nuanced class boundaries and complex data structures, where reliable predictions can further optimise real-time decision-making processes.

Another significant advantage is efficient feature extraction. The algorithm uses a pre-trained regression model to capture a generalised representation of the features from a larger dataset, which are then transferred to subsequent stages. This reduces the need for manual feature selection and extensive data preprocessing, streamlining the entire modelling process. Additionally, the autoencoder further enhances this by enabling dimensionality reduction without sacrificing critical information. This reduces computational costs and helps the model generalise, improving its ability to work effectively even with smaller datasets. The algorithm also demonstrates robustness and flexibility by utilising information from various datasets.

Moreover, its consistency across multiple runs, as indicated by low standard deviations in accuracy and error metrics, highlights its reliability in real-world applications. This robustness is further supported by techniques such as early stopping, which prevent overfitting by halting training when no further improvement is observed, ensuring the model remains generalised. Additionally, TL reduces the data required for training and decreases computational time while maintaining strong performance. The combination of TL and autoencoders also leads to scalability, enabling the model to be easily expanded to incorporate other data modalities without requiring a complete redesign.

Finally, the algorithm demonstrates promising results for the early diagnosis of Alzheimer's disease. By accurately predicting MMSE scores and classifying different stages of cognitive impairment, the model can assist in the timely detection of AD progression. This early identification is essential for initiating appropriate interventions that may slow the disease progression and enhance patient outcomes.

While the proposed multi-stage algorithm offers numerous advantages, there are minor disadvantages. One potential issue is increased model complexity. Integrating multiple advanced techniques such as TL and autoencoders requires careful tuning and design, which can be challenging to implement than models. This can make the approach harder to interpret, particularly for experts who may not be familiar with DL methods.

Another drawback is the dependence on high-quality data. While TL helps mitigate the need for large datasets, the model performance still heavily relies on the quality of the input data. Any noise or inaccuracies in this data could reduce the effectiveness of the algorithm, particularly in real-world environments where data may not always be perfect.

6.3.1 Clinical relevance

The proposed multi-stage algorithm has significant clinical relevance, as it was evaluated using an AD dataset. Advanced techniques such as TL and autoencoders enable the model to extract key patterns from complex datasets, such as MRI scans and demographic data, making it possible to detect subtle changes in brain structure that might not be obvious in standard assessments. This can lead to earlier diagnosis than traditional methods, which often rely on observable symptoms that may appear in the later stages of the disease. As a result, healthcare professionals could use this algorithm to identify at-risk individuals before a significant cognitive decline occurs, opening opportunities for preventive care or early therapeutic interventions.

Additionally, by accurately predicting MMSE scores and classifying stages of cognitive decline, the algorithm provides an effective tool for identifying individuals at varying stages of impairment, including mild and moderate cognitive impairment. Early and precise detection of cognitive decline is critical for clinicians, as it enables timely intervention strategies that could slow disease progression, enhance quality of life, and offer long-term outcomes for patients.

Furthermore, the ability of the algorithm to work with diverse types of data, including MRI images and cognitive scores, reflects its versatility in real-world clinical settings. It can be integrated into existing diagnostic workflows, providing clinicians with an additional tool to enhance decision-making. By automating the feature extraction and prediction process, the algorithm reduces the cognitive load on clinicians, enabling faster and consistent diagnostic

evaluations. This could be particularly beneficial in clinical environments with limited access to specialised neurological expertise, enabling accurate assessments in a broader range of healthcare settings.

In summary, the proposed approach has clear clinical implications. It offers a precise, automated, and scalable method for diagnosing and tracking the progression of AD, ultimately leading to patient management and enhanced clinical outcomes.

6.3.2 Future work

Future directions for the proposed multi-stage algorithm involve several avenues for improvement and broader application.

One important direction is enhancing the interpretability of the model. While the algorithm performs well in terms of prediction, its complexity can make it challenging to adopt without understanding the underlying factors driving its decisions. Future work could focus on developing explainability tools, such as attention mechanisms or feature attribution methods, to offer clearer insights into which features are most strongly associated with the target. This would make the model transparent and user-friendly, aiding its acceptance and adoption.

Finally, another important direction is the optimisation of the algorithm for scalability and efficiency, enabling its deployment on large-scale and high-dimensional datasets typical in real-world scenarios. This could involve investigating efficient architectures, pruning strategies, or utilising distributed computing frameworks to reduce computational cost without sacrificing accuracy.

7. Conclusion

This thesis has examined a range of AI methodologies to enhance predictive performance and explainability for complex real-world applications, with a particular emphasis on Alzheimer's disease (AD) prediction, wherein model reliability is crucial. Various methodologies were investigated to enhance the accuracy, explainability and data limitations of AI models. To address the primary issues, novel methodologies for feature selection, sensitivity analysis, and transfer learning have been proposed.

The initial research introduced two novel filter-based FS methods designed to address the challenges of high-dimensional and noisy datasets. Based on correlation and clustering, these techniques significantly reduced the number of input features while maintaining or even improving predictive accuracy. The validation of these methodologies against an external arrhythmia dataset has indicated their potential to generalise and enhance the efficiency of model training and enhance explainability, constituting a significant contribution to the field of ML. This research focused on reducing dimensionality and developing efficient, interpretable models for practical applications.

The subsequent research focused on model explainability, representing one of the most significant challenges in implementing DNNs within high-stakes domains. The research evaluated the feature importance scores of a DNN model using SA techniques such as SHAP and Sobol. The research contributed to the understanding of the decision-making patterns of complex models, minimising the gap between black box AI models and their practical applicability. This technique provides critical assistance in the continuous endeavours concerning AI models, which is vital for the predictive models developed for domains requiring high interpretability.

The final research introduced a novel multi-step algorithm designed to mitigate the challenges associated with limited data availability in critical domains, thereby enabling data-efficient and precise predictive modelling. This research significantly enhanced prediction accuracy by combining TL and autoencoder methods. Using pre-trained models and applying feature extraction techniques through TL and autoencoders has proven to be an effective strategy, yielding accurate and reliable predictions. This study illustrates the potential to enhance AI applications for predicting outcomes in data-constrained scenarios, underscoring its contribution to advancing the technical capabilities of AI in real-world applications. This methodology was employed to predict the cognitive stages of AD patients based on their MMSE scores.

The three research studies have made considerable contributions to advancing AI within real-world applications. As evidenced by the thesis, advancements in feature selection, sensitivity analysis, and transfer learning provide a solid foundation for enhancing the accuracy, transparency, and reliability of AI models. This research has laid a robust foundation for future initiatives to develop effective AI systems by addressing critical challenges such as

performance, explainability and data constraints. In conclusion, this research holds the potential to transform the utilisation of AI in high-stakes healthcare applications such as predicting AD. This transformation could lead to significant advancements in the accuracy and efficacy of AI applications within the high-stakes domains.

7.1. Feature Selection Summary

Chapter 4 presented two novel filter-based FS techniques to enhance model performance, minimise overfitting, and enhance interpretability in high-dimensional datasets. Feature selection plays a vital role when dealing with noisy or irrelevant features within datasets, enabling models to concentrate on the most pertinent inputs. The three predominant FS approaches, filter, wrapper, and embedded methods are notably characterised by their computational efficiency and generalisability in effectiveness. Both proposed techniques in this research were filter-based approaches.

The first technique developed used a correlation-based approach referred to as CGN-FS, where features with correlation values above a threshold were selected to represent the broader dataset. This method identified features with low inter-feature correlation, reducing the feature set without sacrificing accuracy. The second technique employed was clustering analysis, referred to as RCH-FSC, where clusters were formed based on data correlations. The centroids of these clusters identified by the K-Means algorithm were utilised to represent the entirety of the cluster, thereby creating the most relevant subset of features. This clustering methodology has successfully identified a concise set of features that preserve critical information pertinent to classification tasks while simultaneously reducing the dimensionality of the dataset.

The correlation-based methodology yielded a feature set with fewer features, resulting in straightforward models to interpret. The model demonstrates a negligible impact on overall accuracy. The technique was validated on the arrhythmia dataset to illustrate the generalisability of the established method. For comparison, the ReliefF algorithm, a traditional FS technique, demonstrated marginally lower classification performance in the SVM model and comparable performance in linear regression models, suggesting that CGN-FS captured strongly predictive features and effectively reduced the input feature space.

The clustering-based approach yielded a feature set comprising four features while maintaining comparable accuracy to the model that utilises the complete feature set. These reductions in dimensionality fostered the development of robust and interpretable models, ultimately unveiling deeper insights into the relationships among variables.

Both developed techniques exhibited an observable enhancement in accuracy. This observation underscores their efficacy in identifying the most pertinent features and emphasises their straightforward implementation. The correlation and clustering-based feature selection represents a robust methodology for optimising ML models, particularly in high-dimensional datasets.

The proposed FS techniques enhance the predictive accuracy and interpretability of models, particularly in AD prediction, where concise biomarker sets are necessary for clinical decision-making. By significantly minimising the feature space, these techniques facilitate transparent and reliable decision-making processes, which are crucial for real-world applications. Furthermore, these methods establish a robust framework for managing complex datasets, thereby paving the way for subsequent innovations in FS methodologies.

7.2. Sensitivity Analysis Summary

Chapter 5 presents the DNN models, which were evaluated using SA techniques to assess the degree of their explainability. However, DNNs are frequently regarded as black-box models due to their lack of transparency. This raises concerns about their trustworthiness in critical applications, which led to increased demand for the explainability of AI models, to understand and trust their decision-making processes.

This research used a high-dimensional dataset categorised into two groups to train the classification model. Two G-SA libraries, SHAP and SALib, which comprise Sobol, Morris, and FAST methods, were utilised to compute feature importance scores, thereby understanding the features that most significantly influenced the predictions of the DNN model.

The feature importance scores derived from SHAP and SALib libraries were compared and combined based on their similarities, ensuring the robustness of the findings. This

comprehensive analysis of the importance of features helps enhance the interpretability of the DNN model, making its predictions transparent and aligned with established knowledge.

The ensemble approach of SA was implemented on the Alzheimer's Disease dataset and their neuro-anatomical findings correspond to established AD biomarkers, thereby enhancing its clinical validity. Ensemble-based SA approach identified several significant brain regions strongly associated with AD, including the temporal horn of the lateral ventricles, the hippocampus, the hippocampal tail, the subiculum, and other related areas. Experts in the medical field rigorously evaluated and correlated these features with established neuroimaging biomarkers, structural alterations, and clinical indicators of AD. Understanding these significant features contributes to a deeper comprehension of Alzheimer’s progression and underlying mechanisms, providing valuable insights for both AI model development and medical research.

This study focused on enhancing AI models designed to predict and, significantly, address the critical issue of model explainability. By employing model-agnostic explainability techniques such as SHAP and SALib, this research identifies the most influential features driving model predictions. This contributes to enhancing interpretability in ML pipelines and provides a structured framework for future AI research focused on complex, high-dimensional datasets. The comparative analysis of these methodologies provides valuable insights into their efficacy, helping researchers select the most suitable approaches for evaluating feature importance within AI models.

Ultimately, this work advances the integration of AI methodologies with real-world applications, contributing to the development of robust and interpretable AI systems. The outcomes of this research are expected to support the design of reliable, transparent AI models applicable to high-stakes decision-making environments, while providing a foundation for further exploration in explainable AI and algorithmic reliability.

7.3 Transfer Learning with Autoencoder Summary

Chapter 6 presents a novel multi-stage algorithm developed to enhance the accuracy of predictive modelling in data-constrained environments, particularly in AD prediction, where labelled training data is often limited. DNN-based techniques are increasingly utilised to

address complex prediction and classification tasks across various domains. This research integrates regression and classification models to utilise insights derived from multi-class datasets, enhancing predictive performance in sequential learning tasks.

Initially, a regression model was created to predict the ages of the patients using a combined MCI dataset, which included individuals with EMCI, MCI, and LMCI. The model was designed with multiple hidden layers and ReLU activation functions while incorporating dropout regularisation and RMSprop optimisation. Hyperparameter tuning was performed using GridSearchCV to optimise performance. TL techniques were then applied to transfer knowledge from the regression model into an autoencoder. The autoencoder effectively extracted key features from the MCI dataset, generating encoded representations to predict MMSE scores. These scores were then employed to classify patients into two cognitive categories: Mild and Moderate.

The proposed multi-stage algorithm yielded promising results, achieving an accuracy of approximately 73.26% with a standard deviation of 3.92%. In contrast, a regression model that did not employ transfer learning or autoencoders achieved only 61.08% accuracy with a 2.21% standard deviation. This 12.18% improvement in accuracy highlights the significant contribution of the TL and autoencoder techniques to the performance of the model. The comparison underscores the effectiveness of the proposed methodology.

The proposed algorithm was assessed against various published DNN-based methodologies utilising structural MRI data. When compared to the performance metrics of Li et al. (2015) (70.1 ± 2.3%) and the CAE model developed by Oh et al. (2019) (73.23 ± 4.21%), as well as their ICAE model (73.95 ± 4.82%), the proposed algorithm demonstrates competitive performance. Although the enhanced accuracy variations range from 0.03% to 3.2%, the consistent performance underscores the robustness of this approach in delivering reliable and interpretable classification without dependence on additional estimated variables.

Overall, this study demonstrates the efficacy of transfer learning and autoencoder-based feature extraction for improving predictive modelling in data-limited scenarios. The proposed multi-step algorithm has proven effective in enhancing both classification and regression tasks, offering a scalable framework for future AI applications. The research

contributes to the advancement of DNN-based methodologies and establishes a strong foundation for subsequent developments in predictive modelling across diverse, high-impact domains. The multi-step algorithm has also proven effective in enhancing cognitive stage classification and predicting MMSE scores, providing valuable insights for clinical applications.

7.4 Overall Conclusions

This thesis has explored various AI techniques to enhance predictive performance and explainability of DNN models, particularly in AD prediction, where accuracy and interpretability are essential for clinical contexts. The research has focused on three primary aspects: Feature Selection, Sensitivity Analysis, and Transfer Learning with autoencoders. These methodologies have significantly enhanced accuracy, explicability, and data efficiency. Integrating these approaches can lead to robust and interpretable AI frameworks, advancing the field of AI and facilitating the integration of DNN models into critical domains.

A particularly noteworthy finding of this thesis is that the features selected by the clustering-based FS technique, approximately four features, were also prioritised by SA models. This convergence suggests that the chosen features are statistically significant and medically relevant. This alignment strengthens the reliability of these features in AD diagnosis and progression prediction.

Beyond the specific methodologies explored in this thesis, these findings highlight broader implications for AI in critical domains. Integrating FS, SA, TL, and autoencoders aligns with key objectives in AI-driven domains, including improving transparency, reducing data requirements, and increasing the accuracy of predictive models. This research has demonstrated that lowering dimensionality while preserving essential information is possible and beneficial for improving model generalisation. Additionally, by utilising pre-trained models and autoencoders, AI models can be developed with less reliance on extensive labelled datasets, addressing one of the significant challenges in healthcare.

An effective integration combines feature selection, transfer learning, and autoencoders to address high-dimensional, noisy datasets. The correlation-based FS method proposed in this thesis efficiently reduces input space while maintaining or improving predictive accuracy. Applying FS prior to TL refines the dataset, lowering computational

complexity and improving model generalisation. Autoencoders further optimise feature representations by compressing irrelevant information and extracting essential patterns. This synergistic approach enables the construction of efficient, scalable, and interpretable AI models for complex tasks.

Additionally, sensitivity analysis can be integrated with TL and autoencoders to enhance feature prioritisation. Techniques such as SHAP and SALib, as employed in this research, guide the selection of influential features from pre-trained models, improving the reliability and focus of TL-based architectures. Autoencoders can further refine these features by eliminating redundancy while preserving critical information, supporting the development of explainable and computationally efficient AI pipelines.

Alternatively, SA can be applied after the TL process to evaluate and interpret feature importance in the transferred model. These evaluations are particularly valuable for AD models, as they ensure the transferred features maintain clinical relevance. Using diverse SA techniques such as SHAP, Permutation Feature Importance (PFI), and DiCE offers complementary perspectives—local, global, and counterfactual—on model behaviour. This post-hoc sensitivity evaluation enhances interpretability by validating which features remain critical after knowledge transfer, further supporting model transparency and robustness.

Furthermore, the potential impact of these integrated approaches extends beyond the specific application domain explored in this study. The methodologies developed in this thesis can be applied to other complex, high-dimensional datasets and broader AI tasks where data complexity, interpretability, and limited labelled data remain significant challenges. The use of feature selection, sensitivity analysis, and transfer learning with autoencoders offers a scalable blueprint for designing AI models that are not only highly accurate but also aligned with real-world computational and interpretability requirements.

In conclusion, this thesis has laid the groundwork for a comprehensive AI framework capable of producing accurate, reliable, and interpretable models. Utilising feature selection, sensitivity analysis, transfer learning, and autoencoders presents a robust approach to addressing key challenges in explainability, data efficiency, and predictive performance. Future research should focus on refining and validating these methods using larger, heterogeneous datasets and exploring their deployment within practical, real-world AI

systems. Ultimately, this research significantly contributes to advancing the development of transparent, efficient, and scalable AI solutions for complex, high-stakes decision-making environments across various domains.

7.5 Limitations of the Study

While there are significant advancements in AI applications, particularly in FS, model explainability and TL, it is imperative to point out the limitations experienced during the analysis.

7.5.1 Limited Dataset Size:

A prominent limitation identified in the existing literature is the limited number of datasets employed, particularly specific conditions such as AD and supplement evaluations. Although the gathered data has proven advantageous, it is imperative for researchers to obtain additional data to enhance the generalisability and robustness of their models. With a broader and diverse array of patient data, the findings could achieve excellent representation and applicability within broader healthcare contexts. It is recommended that future research emphasises the augmentation of dataset sizes through partnerships with larger institutions to ensure results that are both reliable and broadly applicable.

7.5.2 Generalisability of MRI-Based Models:

The use of MRI scans as the primary data source presents another limitation in terms of generalisability. Although the MRI scans provide valuable structural insight, the models built within this research study may not fully capture the complexity of the development of AD over different populations or within different stages of the progression of the disease. Training on MRI data alone may not adequately support model generalisability across different types of patients or NDDs. Including other forms of data, such as clinical notes, medical evaluations or data on other biomarkers, would alleviate this issue and enhance the applicability of the models in various medical settings.

In summary, the few demerits of this research are the small size of the available dataset, and over-dependence on MRI scans. Addressing these challenges with volumetric

data and diverse diseases would be imperative in increasing the robustness and applicability of AI models in the medical field.

7.6 Future Directions

7.6.1 Enhancing AI with Integrated Methods

One avenue for future research is to further explore the integration of feature selection, transfer learning, autoencoders, and sensitivity analysis to enhance model efficiency, interpretability, and scalability. Applying sensitivity analysis either before or after transfer learning, using techniques such as SHAP, PFI, and DiCE, can offer diverse insights into feature importance and enhance model transparency in complex, high-dimensional tasks. This combined approach can reduce computational costs, enable effective generalisation to unseen datasets, and support the development of AI models that are both accurate and explainable. Such models would be particularly valuable in real-world applications where decision traceability, reliability, and data efficiency are critical. Ultimately, this framework could serve as a foundation for building robust AI solutions capable of addressing complex challenges across various domains.

7.6.2 Integrating Real-Time Data Streams into AI Models

One area with potential for further exploration and advancement is the integration of real-time data into AI systems. As technology advances, the utilisation of real-time data, for instance, through health trackers or monitoring patients in a hospital, enables the possibility of deploying AI models that could evolve with re-training on data as the disease progresses. This approach would provide much-needed support in clinical settings where care is time-bound and data-driven decisions are made rapidly. However, further research is necessary to address the real-time AI issues that arise from the practical application of these models, while also preventing risks to patients and breaches of data confidentiality. One promising area for further exploration is the integration of real-time data into AI systems. As technology advances, utilising continuous data streams from sensors, IoT devices, or dynamic environments presents opportunities to deploy AI models capable of evolving through incremental or online learning. This would enable AI systems to adapt to changing conditions

and support rapid, data-driven decision-making in time-sensitive applications. However, further research is required to address the practical challenges of real-time AI, including system reliability, computational efficiency, and data privacy concerns in dynamic deployment scenarios.

References

AbdelAziz, N. M., Said, W., AbdelHafeez, M. M., & Ali, A. H. (2024). Advanced interpretable diagnosis of Alzheimer's disease using SECNN-RF framework with explainable AI. *Frontiers in Artificial Intelligence*, *7*, 1456069. https://doi.org/10.3389/frai.2024.1456069

Abnar, S., & Zuidema, W. (2020). Quantifying attention flow in transformers. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.385

Adam, M., Wessel, M., & Benlian, A. (2021). AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, *31*(2). https://doi.org/10.1007/s12525-020-00414-7

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, *2018-December*.

Aderghal, K., Afdel, K., Benois-Pineau, J., & Catheline, G. (2020). Improving Alzheimer's stage categorization with Convolutional Neural Network using transfer learning and different magnetic resonance imaging modalities. *Heliyon*, *6*(12). https://doi.org/10.1016/j.heliyon.2020.e05652

ADNI Database . (2021). Alzheimer's Disease Neuroimaging Initiative.

AIBL Database. (2021). In *Australian Imaging, Biomarkers & Lifestyle Study of Ageing*.

Alatrany, A. S., Khan, W., Hussain, A., Kolivand, H., & Al-Jumeily, D. (2024). An explainable machine learning approach for Alzheimer's disease classification. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-51985-w

Alvarez-Melis, D., & Jaakkola, T. S. (2018). *On the Robustness of Interpretability Methods*.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety*.

Amoroso, N., Quarto, S., La Rocca, M., Tangaro, S., Monaco, A., & Bellotti, R. (2023). An eXplainability Artificial Intelligence approach to brain connectivity in Alzheimer's disease. *Frontiers in Aging Neuroscience*, *15*. https://doi.org/10.3389/fnagi.2023.1238065

Ando Saabas. (2021). *treeinterpreter*. GitHub.

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *82*(4). https://doi.org/10.1111/rssb.12377

Apostolova, L. G., Green, A. E., Babakchanian, S., Hwang, K. S., Chou, Y. Y., Toga, A. W., & Thompson, P. M. (2012). Hippocampal atrophy and ventricular enlargement in normal aging, mild cognitive impairment (MCI), and Alzheimer disease. *Alzheimer Disease and Associated Disorders*, *26*(1). https://doi.org/10.1097/WAD.0b013e3182163b62

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2020). *Invariant Risk Minimization*.

Asl, E. H., Ghazal, M., Mahmoud, A., Aslantas, A., Shalaby, A., Casanova, M., Barnes, G., Gimel'farb, G., Keynton, R., & Baz, A. El. (2018). Alzheimer's disease diagnostics by a 3D deeply supervised adaptable convolutional network. *Frontiers in Bioscience - Landmark*, *23*(3). https://doi.org/10.2741/4606

Atia, N., Benzaoui, A., Jacques, S., Hamiane, M., Kourd, K. El, Bouakaz, A., & Ouahabi, A. (2022). Particle Swarm Optimization and Two-Way Fixed-Effects Analysis of Variance for Efficient Brain Tumor Segmentation. *Cancers*, *14*(18). https://doi.org/10.3390/cancers14184399

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, *10*(7). https://doi.org/10.1371/journal.pone.0130140

Backhausen, L. L., Herting, M. M., Buse, J., Roessner, V., Smolka, M. N., & Vetter, N. C. (2016). Quality control of structural MRI images applied using FreeSurfer-a hands-on workflow to rate motion artifacts. *Frontiers in Neuroscience*, *10*(DEC). https://doi.org/10.3389/fnins.2016.00558

Barlow, H. B. (1989). Unsupervised Learning. *Neural Computation*, *1*(3). https://doi.org/10.1162/neco.1989.1.3.295

Bass, C., Silva, M. da, Sudre, C., Williams, L. Z. J., Sousa, H. S., Tudosiu, P.-D., Alfaro-Almagro, F., Fitzgibbon, S. P., Glasser, M. F., Smith, S. M., & Robinson, E. C. (2023). ICAM-Reg: Interpretable Classification and Regression with Feature Attribution for Mapping Neurological Phenotypes in Individual Scans. *IEEE Transactions on Medical Imaging*, *42*(4), 959–970. https://doi.org/10.1109/TMI.2022.3221890

Battineni, G., Chintalapudi, N., Amenta, F., & Traini, E. (2021). Deep learning type convolution neural network architecture for multiclass classification of Alzheimer's disease. *BIOIMAGING 2021 - 8th International Conference on Bioimaging; Part of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2021*, 209–215. https://doi.org/10.5220/0010378602090215

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., & Wiswedel, B. (2009). KNIME-the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD Explorations Newsletter*, *11*(1).

Bhatkoti, P., & Paul, M. (2016). Early diagnosis of Alzheimer's disease: A multi-class deep learning framework with modified k-sparse autoencoder classification. *International Conference Image and Vision Computing New Zealand*, *0*. https://doi.org/10.1109/IVCNZ.2016.7804459

Bitton, R., Malach, A., Meiseles, A., Momiyama, S., Araki, T., Furukawa, J., Elovici, Y., & Shabtai, A. (2022). *Latent SHAP: Toward Practical Human-Interpretable Explanations*.

Blennow, K., Dubois, B., Fagan, A. M., Lewczuk, P., De Leon, M. J., & Hampel, H. (2015). Clinical utility of cerebrospinal fluid biomarkers in the diagnosis of early Alzheimer's disease. In *Alzheimer's and Dementia* (Vol. 11, Issue 1). https://doi.org/10.1016/j.jalz.2014.02.004

Bloch, L., & Friedrich, C. M. (2022). Machine Learning Workflow to Explain Black-Box Models for Early Alzheimer's Disease Classification Evaluated for Multiple Datasets. *SN Computer Science*, *3*(6). https://doi.org/10.1007/s42979-022-01371-y

Bogdanovic, B., Eftimov, T., & Simjanoska, M. (2022). In-depth insights into Alzheimer's disease by using explainable machine learning approach. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-10202-2

Bojarski, M., Jackel, L., Firner, B., & Muller, U. (2017). Explaining How End-to-End Deep Learning Steers a Self-Driving Car. *Nvidia*.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

C. Macedo, A., Tissot, C., Therriault, J., Servaes, S., Wang, Y. T., Fernandez-Arias, J., Rahmouni, N., Lussier, F. Z., Vermeiren, M., Bezgin, G., Vitali, P., Ng, K. P., Zimmer, E. R., Guiot, M. C., Pascoal, T. A., Gauthier, S., & Rosa-Neto, P. (2023). The Use of Tau PET to Stage Alzheimer Disease According to the Braak Staging Framework. *Journal of Nuclear Medicine*, *64*(8). https://doi.org/10.2967/jnumed.122.265200

Carlesimo, G. A., Piras, F., Orfei, M. D., Iorio, M., Caltagirone, C., & Spalletta, G. (2015). Atrophy of presubiculum and subiculum is the earliest hippocampal anatomical marker of Alzheimer's disease. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, *1*(1). https://doi.org/10.1016/j.dadm.2014.12.001

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *2015-August*. https://doi.org/10.1145/2783258.2788613

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. In *Electronics (Switzerland)* (Vol. 8, Issue 8). https://doi.org/10.3390/electronics8080832

Chefer, H., Gur, S., & Wolf, L. (2021). Transformer Interpretability Beyond Attention Visualization. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR46437.2021.00084

Choe, Y. M., Lee, B. C., Choi, I. G., Suh, G. H., Lee, D. Y., & Kim, J. W. (2020). Mmse subscale scores as useful predictors of ad conversion in mild cognitive impairment. *Neuropsychiatric Disease and Treatment*, *16*. https://doi.org/10.2147/NDT.S263702

Chormunge, S., & Jena, S. (2018). Correlation based feature selection with clustering for high dimensional data. *Journal of Electrical Systems and Information Technology*, *5*(3), 542–549. https://doi.org/10.1016/J.JESIT.2017.06.004

Chun, M. Y., Park, C. J., Kim, J., Jeong, J. H., Jang, H., Kim, K., & Seo, S. W. (2022). Prediction of conversion to dementia using interpretable machine learning in patients with amnestic mild cognitive impairment. *Frontiers in Aging Neuroscience*, *14*. https://doi.org/10.3389/fnagi.2022.898940

Cowls, J., King, T., Taddeo, M., & Floridi, L. (2019). Designing AI for Social Good: Seven Essential Factors. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3388669

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised Learning. In *Machine Learning Techniques for Multimedia* (pp. 21–49). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_2

Dandl, S., Molnar, C., Binder, M., & Bischl, B. (2020). Multi-objective counterfactual explanations. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12269 LNCS*. https://doi.org/10.1007/978-3-030-58112-1_31

De Bastos-Leite, A. J., Van Waesberghe, J. H., Oen, A. L., Van Der Flier, W. M., Scheltens, P., & Barkhof, F. (2006). Hippocampal sulcus width and cavities: Comparison between patients with Alzheimer disease and nondemented elderly subjects. *American Journal of Neuroradiology*, *27*(10).

de Paula, V. de J. R., Guimarães, F. M., Diniz, B. S., & Forlenza, O. V. (2009). Neurobiological pathways to Alzheimer's disease: Amyloid-beta, TAU protein or both? In *Dementia e Neuropsychologia* (Vol. 3, Issue 3). https://doi.org/10.1590/s1980-57642009dn30300003

De Santi, L. A., Pasini, E., Santarelli, M. F., Genovesi, D., & Positano, V. (2023). An Explainable Convolutional Neural Network for the Early Diagnosis of Alzheimer's Disease from 18F-FDG PET. *Journal of Digital Imaging*, *36*(1). https://doi.org/10.1007/s10278-022-00719-3

Debie, E., & Shafi, K. (2019). Implications of the curse of dimensionality for supervised learning classifier systems: theoretical and empirical analyses. *Pattern Analysis and Applications*, *22*(2). https://doi.org/10.1007/s10044-017-0649-0

Dennis, E. L., & Thompson, P. M. (2014). Functional brain connectivity using fMRI in aging and Alzheimer's disease. In *Neuropsychology Review* (Vol. 24, Issue 1). https://doi.org/10.1007/s11065-014-9249-6

Dhinagar, N. J., Thomopoulos, S. I., Rajagopalan, P., Stripelis, D., Ambite, J. L., Ver Steeg, G., & Thompson, P. M. (2022). *Evaluation of Transfer Learning Methods for Detecting Alzheimer's Disease with Brain MRI*. https://doi.org/10.1101/2022.08.23.505030

Dong, Q., Zhang, J., Li, Q., Wang, J., Leporé, N., Thompson, P. M., Caselli, R. J., Ye, J., & Wang, Y. (2020). Integrating Convolutional Neural Networks and Multi-Task Dictionary Learning for Cognitive Decline Prediction with Longitudinal Images. *Journal of Alzheimer's Disease*, *75*(3), 971–992. https://doi.org/10.3233/JAD-190973

Doshi-Velez, F., & Kim, B. (2017). A Roadmap for a Rigorous Science of Interpretability. *ArXiv Preprint ArXiv:1702.08608v1*.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1). https://doi.org/10.1126/sciadv.aao5580

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management*, *48*. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

Duc, N. T., Ryu, S., Qureshi, M. N. I., Choi, M., Lee, K. H., & Lee, B. (2020). 3D-Deep Learning Based Automatic Diagnosis of Alzheimer's Disease with Joint MMSE Prediction Using Resting-State fMRI. *Neuroinformatics*, *18*(1). https://doi.org/10.1007/s12021-019-09419-w

El-Sappagh, S., Alonso, J. M., Islam, S. M. R., Sultan, A. M., & Kwak, K. S. (2021). A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Scientific Reports*, *11*(1), 2660. https://doi.org/10.1038/s41598-021-82098-3

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639). https://doi.org/10.1038/nature21056

Faisal.F.U.R., U. Khatri, & G. R. Kwon. (2021). Diagnosis of Alzheimer's disease using combined feature selection method. *Journal of Korea Multimedia Society*, *24(5)*, 667–675.

Farouk, Y., & Rady, S. (2020). Early Diagnosis of Alzheimer's Disease using Unsupervised Clustering. *International Journal of Intelligent Computing and Information Sciences*, *20*(2). https://doi.org/10.21608/ijicis.2021.51180.1044

Fischl, B. (2012). FreeSurfer. *Neuroimage*, *62(2)*, 774–781.

Fouladvand, S., Mielke, M. M., Vassilaki, M., St Sauver, J., Petersen, R. C., & Sohn, S. (2019). Deep Learning Prediction of Mild Cognitive Impairment using Electronic Health Records. *Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019*. https://doi.org/10.1109/BIBM47256.2019.8982955

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Frye, C., de Mijolla, D., Begley, T., Cowton, L., Stanley, M., & Feige, I. (2021). SHAPLEY EXPLAINABILITY ON THE DATA MANIFOLD. *ICLR 2021 - 9th International Conference on Learning Representations*.

Gallego-Jutglà, E., Solé-Casals, J., Vialatte, F. B., Elgendi, M., Cichocki, A., & Dauwels, J. (2015). A hybrid feature selection approach for the early diagnosis of Alzheimer's disease. *Journal of Neural Engineering*, *12*(1). https://doi.org/10.1088/1741-2560/12/1/016018

Gao, F., Yoon, H., Xu, Y., Goradia, D., Luo, J., Wu, T., & Su, Y. (2020). AD-NET: Age-adjust neural network for improved MCI to AD conversion prediction. *NeuroImage: Clinical*, *27*. https://doi.org/10.1016/j.nicl.2020.102290

Ghorbani, A., Wexler, J., Zou, J., & Kim, B. (2019). Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, *32*.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *Proceedings - 2018*

IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018. https://doi.org/10.1109/DSAA.2018.00018

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, *24*(1). https://doi.org/10.1080/10618600.2014.907095

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision making and a "right to explanation." *AI Magazine*, *38*(3). https://doi.org/10.1609/aimag.v38i3.2741

Graña, M., Termenon, M., Savio, A., Gonzalez-Pinto, A., Echeveste, J., Pérez, J. M., & Besga, A. (2011). Computer Aided Diagnosis system for Alzheimer Disease using brain Diffusion Tensor Imaging features selected by Pearson's correlation. *Neuroscience Letters*, *502*(3). https://doi.org/10.1016/j.neulet.2011.07.049

Grau, A., Indri, M., Lo Bello, L., & Sauter, T. (2021). Robots in Industry: The Past, Present, and Future of a Growing Collaboration with Humans. *IEEE Industrial Electronics Magazine*, *15*(1). https://doi.org/10.1109/MIE.2020.3008136

Greene, S. J., & Killiany, R. J. (2010). Subregions of the inferior parietal lobule are affected in the progression to Alzheimer's disease. *Neurobiology of Aging*, *31*(8). https://doi.org/10.1016/j.neurobiolaging.2010.04.026

Greenwell, B. M., Boehmke, B. C., & McCarthy, A. J. (2018). *A Simple and Effective Model-Based Variable Importance Measure*.

Grossner, E. C., Bernier, R. A., Brenner, E. K., Chiou, K. S., & Hillary, F. G. (2018). Prefrontal gray matter volume predicts metacognitive accuracy following traumatic brain injury. *Neuropsychology*, *32*(4). https://doi.org/10.1037/neu0000446

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5). https://doi.org/10.1145/3236009

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37). https://doi.org/10.1126/scirobotics.aay7120

Guvenir, H. A., Acar, B., Demiroz, G., & Cekin, A. (1997). A supervised machine learning algorithm for arrhythmia analysis. *Computers in Cardiology 1997*, 433–436. https://doi.org/10.1109/CIC.1997.647926

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hafez, I. Y., Hafez, A. Y., Saleh, A., Abd El-Mageed, A. A., & Abohany, A. A. (2025). A systematic review of AI-enhanced techniques in credit card fraud detection. *Journal of Big Data*, *12*(1), 6. https://doi.org/10.1186/s40537-024-01048-8

Hall, M. A. (2000). *Correlation-based feature selection of discrete and numeric class machine learning.*

Hastie, T. ;, & Tibshirani, R. (1986). Generalized additive models (with discussion). *Statistical Science*, *1*.

Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: deep portfolios. In *Applied Stochastic Models in Business and Industry* (Vol. 33, Issue 1). https://doi.org/10.1002/asmb.2209

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). *What do we need to build explainable AI systems for the medical domain?*

Hooker, S., Erhan, D., Kindermans, P. J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, *32*.

Huang, S.-C., Pareek, A., Jensen, M., Lungren, M. P., Yeung, S., & Chaudhari, A. S. (2023). Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *Npj Digital Medicine*, *6*(1), 74. https://doi.org/10.1038/s41746-023-00811-0

Huang, X., & Marques-Silva, J. (2023). *Refutation of Shapley Values for XAI -- Additional Evidence*.

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., & Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, *14*(12). https://doi.org/10.1088/1748-9326/ab4e55

Ilias, L., & Askounis, D. (2022). Multimodal Deep Learning Models for Detecting Dementia from Speech and Transcripts. *Frontiers in Aging Neuroscience*, *14*. https://doi.org/10.3389/fnagi.2022.830943

IXI Database. (2021). In *Information eXtraction from Images* .

Izza, Y., Ignatiev, A., & Marques-Silva, J. (2020). *On Explaining Decision Trees*.

Jacobs, H. I. L., Van Boxtel, M. P. J., Jolles, J., Verhey, F. R. J., & Uylings, H. B. M. (2012). Parietal cortex matters in Alzheimer's disease: An overview of structural, functional and metabolic findings. In *Neuroscience and Biobehavioral Reviews* (Vol. 36, Issue 1). https://doi.org/10.1016/j.neubiorev.2011.06.009

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.386

Jahan, S., Taher, K. A., Kaiser, M. S., Mahmud, M., Rahman, M. S., Hosen, A. S. M. S., & Ra, I. H. (2023). Explainable AI-based Alzheimer's prediction and management using multimodal data. *PLoS ONE*, *18*(11 November). https://doi.org/10.1371/journal.pone.0294253

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*.

Jha, D., & Kwon, G. R. (2017). Alzheimer's disease detection using sparse autoencoder, scale conjugate gradient and softmax output layer with fine tuning. *International Journal of Machine Learning and Computing*, *7*(1). https://doi.org/10.18178/ijmlc.2017.7.1.612

Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: a review. *Complex and Intelligent Systems*, *8*(3). https://doi.org/10.1007/s40747-021-00637-x

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9). https://doi.org/10.1038/s42256-019-0088-2

Joshi, A., Kale, S., Chandel, S., & Pal, D. (2015). Likert Scale: Explored and Explained. *British Journal of Applied Science & Technology*, *7*(4). https://doi.org/10.9734/bjast/2015/14975

Joshi, A., Upadhyay, H., Parekh, N., Shah, S., & Parekh, K. (2019). Drug prescription patterns in patients with Alzheimer's disease in an urban neuro-specialty clinic in Western India. *National Journal of Physiology, Pharmacy and Pharmacology*, *9*(10), 1. https://doi.org/10.5455/njppp.2019.9.0828114082019

Jumaili, M. L. F., & Sonuç, E. (2025). An Attention-Based CNN Framework for Alzheimer's Disease Staging with Multi-Technique XAI Visualization. *Computers, Materials & Continua*, *83*(2), 2947–2969. https://doi.org/10.32604/cmc.2025.062719

Jung, Y. J., Han, S. H., & Choi, H. J. (2021). Explaining CNN and RNN Using Selective Layer-Wise Relevance Propagation. *IEEE Access*, *9*. https://doi.org/10.1109/ACCESS.2021.3051171

Junior, K. J., Carole, K. S., Theodore Armand, T. P., Kim, H.-C., & The Alzheimer's Disease Neuroimaging Initiative, T. A. D. N. I. (2024). Alzheimer's Multiclassification Using Explainable AI Techniques. *Applied Sciences*, *14*(18), 8287. https://doi.org/10.3390/app14188287

Kalavathi, P., & Prasath, V. B. S. (2016). Methods on Skull Stripping of MRI Head Scan Images—a Review. In *Journal of Digital Imaging* (Vol. 29, Issue 3). https://doi.org/10.1007/s10278-015-9847-8

Kang, W., Li, B., Papma, J. M., Jiskoot, L. C., Deyn, P. P. De, Biessels, G. J., Claassen, J. A. H. R., Middelkoop, H. A. M., Flier, W. M. van der, Ramakers, I. H. G. B., Klein, S., & Bron, E. E. (2023). An Interpretable Machine Learning Model with Deep Learning-Based Imaging Biomarkers for Diagnosis of Alzheimer's Disease. *Springer*, 69–78. https://doi.org/10.1007/978-3-031-47401-9_7

Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection. *International Journal of Information Technology and Knowledge Management*, *2*(2).

Karimi, A. H., Schölkopf, B., & Valera, I. (2021). Algorithmic recourse: From counterfactual explanations to interventions. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3442188.3445899

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2023). Trustworthy Artificial Intelligence: A Review. In *ACM Computing Surveys* (Vol. 55, Issue 2). https://doi.org/10.1145/3491209

Kelodjou, G., Rozé, L., Masson, V., Galárraga, L., Gaudel, R., Tchuente, M., & Termier, A. (2024). Shaping Up SHAP: Enhancing Stability through Layer-Wise Neighbor Selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(12), 13094–13103. https://doi.org/10.1609/aaai.v38i12.29208

Khan, R., Akbar, S., Mehmood, A., Shahid, F., Munir, K., Ilyas, N., Asif, M., & Zheng, Z. (2022). A transfer learning approach for multiclass classification of Alzheimer's disease using MRI images. *Frontiers in Neuroscience*, *16*, 1050777. https://doi.org/10.3389/fnins.2022.1050777

Kim, B., Rudin, C., & Shah, J. (2014). The Bayesian case model: A generative approach for case-based reasoning and prototype classification. *Advances in Neural Information Processing Systems*, *3*(January).

Kindermans, P. J., Schütt, K. T., Alber, M., Müller, K. R., Erhan, D., Kim, B., & Dähne, S. (2018). Learning how to explain neural networks: Patternnet and Patternattribution. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*.

Koch, G., Casula, E. P., Bonnì, S., Borghi, I., Assogna, M., Minei, M., Pellicciari, M. C., Motta, C., D'Acunto, A., Porrazzini, F., Maiella, M., Ferrari, C., Caltagirone, C., Santarnecchi, E., Bozzali, M., & Martorana, A. (2022). Precuneus magnetic stimulation for Alzheimer's disease: a randomized, sham-controlled trial. *Brain*, *145*(11). https://doi.org/10.1093/brain/awac285

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. *34th International Conference on Machine Learning, ICML 2017*, *4*.

Koh, P. W., Nguye, T., Tang, Y. S., Mussmann, S., Pierso, E., Kim, B., & Liang, P. (2020). Concept Bottleneck Models. *37th International Conference on Machine Learning, ICML 2020*, *PartF168147-7*.

Kommiya Mothilal, R., Mahajan, D., Tan, C., & Sharma, A. (2021). Towards Unifying Feature Attribution and Counterfactual Explanations: Different Means to the Same End. *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3461702.3462597

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *784 LNCS*. https://doi.org/10.1007/3-540-57868-4_57

Konstantinidis, K. (2024). The shortage of radiographers: A global crisis in healthcare. In *Journal of Medical Imaging and Radiation Sciences* (Vol. 55, Issue 4). https://doi.org/10.1016/j.jmir.2023.10.001

Krishna, S., Han, T., Gu, A., Wu, S., Jabbari, S., & Lakkaraju, H. (2025). *The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *2*.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. A. (2020). Problems with Shapley-value-based explanations as feature importance measures. *37th International Conference on Machine Learning, ICML 2020*, *PartF168147-8*.

L. Breiman, J. Friedman, R. A. Olshen, & C. J. Stone. (2017). *Classification and regression trees*.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-08987-4

LeCun, Y., Hinton, G., & Bengio, Y. (2015). Deep learning (2015), Y. LeCun, Y. Bengio and G. Hinton. *Nature*, *521*.

Letoffe, O., Huang, X., & Marques-Silva, J. (2024). *SHAP scores fail pervasively even when Lipschitz succeeds*.

Li, F., Tran, L., Thung, K. H., Ji, S., Shen, D., & Li, J. (2015). A Robust Deep Model for Improved Classification of AD/MCI Patients. *IEEE Journal of Biomedical and Health Informatics*, *19*(5). https://doi.org/10.1109/JBHI.2015.2429556

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3).

Liu, J., Tian, X., Lin, H., Li, H. D., & Pan, Y. (2024). Multi-Task Learning for Alzheimer's Disease Diagnosis and Mini-Mental State Examination Score Prediction. *Big Data Mining and Analytics*, *7*(3), 828–842. https://doi.org/10.26599/BDMA.2024.9020025

Lucic, A., ter Hoeve, M., Tolomei, G., de Rijke, M., & Silvestri, F. (2022). CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks. *Proceedings of Machine Learning Research*, *151*.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1). https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *2017-December*.

MacKay, A., Laule, C., Vavasour, I., Bjarnason, T., Kolind, S., & Mädler, B. (2006). Insights into brain microstructure from the T2 distribution. In *Magnetic Resonance Imaging* (Vol. 24, Issue 4). https://doi.org/10.1016/j.mri.2005.12.037

Marcisz, A., & Polanska, J. (2023). Can T1-Weighted Magnetic Resonance Imaging Significantly Improve Mini-Mental State Examination-Based Distinguishing Between Mild Cognitive Impairment and Early-Stage Alzheimer's Disease? *Journal of Alzheimer's Disease*, *92*(3). https://doi.org/10.3233/JAD-220806

Martin, N. (2019). The Major Concerns Around Facial Recognition Technology. *Forbes*, *03*.

Mcevoy, L. K., Fennema-notestine, C., Roddey, J. C., Holland, D., Pung, C. J., Brewer, J. B., & Dale, A. M. (2009). Alzheimer Disease: Quantitative Structural Neuroimaging for Detection and Prediction of Clinical and Purpose: Methods : Results : Conclusion : *Radiology*, *251*(1).

Mehmood, A., Shahid, F., Khan, R., Ibrahim, M. M., & Zheng, Z. (2024). Utilizing Siamese 4D-AlzNet and Transfer Learning to Identify Stages of Alzheimer's Disease. *Neuroscience*, *545*. https://doi.org/10.1016/j.neuroscience.2024.03.007

Mehmood, A., yang, S., feng, Z., wang, M., Ahmad, A. S., khan, R., Maqsood, M., & Yaqub, M. (2021). A Transfer Learning Approach for Early Diagnosis of Alzheimer's Disease on MRI Images. *Neuroscience*, *460*. https://doi.org/10.1016/j.neuroscience.2021.01.002

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. In *ACM Computing Surveys* (Vol. 54, Issue 6). https://doi.org/10.1145/3457607

Mercadante, A. A., & Tadi, P. (2020). Neuroanatomy, Gray Matter. In *StatPearls*.

Miklossy, J. (2011). Alzheimer's disease - a neurospirochetosis. Analysis of the evidence following Koch's and Hill's criteria. In *Journal of Neuroinflammation* (Vol. 8). https://doi.org/10.1186/1742-2094-8-90

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. In *Artificial Intelligence* (Vol. 267). https://doi.org/10.1016/j.artint.2018.07.007

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11). https://doi.org/10.1038/s42256-019-0114-4

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. *Communications in Computer and Information Science*, *1323*. https://doi.org/10.1007/978-3-030-65965-3_28

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K. R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, *65*. https://doi.org/10.1016/j.patcog.2016.11.008

Moreno-Ibarra, M. A., Villuendas-Rey, Y., Lytras, M. D., Yáñez-Márquez, C., & Salgado-Ramírez, J. C. (2021). Classification of diseases using machine learning algorithms: A comparative study. *Mathematics*, *9*(15). https://doi.org/10.3390/math9151817

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, *33*(2). https://doi.org/10.1080/00401706.1991.10484804

Muschalik, M., Fumagalli, F., Hammer, B., & Hüllermeier, E. (2024). Beyond TreeSHAP: Efficient Computation of Any-Order Shapley Interactions for Tree Ensembles. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(13), 14388–14396. https://doi.org/10.1609/aaai.v38i13.29352

Nanni, L., Interlenghi, M., Brahnam, S., Salvatore, C., Papa, S., Nemni, R., & Castiglioni, I. (2020). Comparison of Transfer Learning and Conventional Machine Learning Applied to Structural Brain MRI for the Early Diagnosis and Prognosis of Alzheimer's Disease. *Frontiers in Neurology*, *11*. https://doi.org/10.3389/fneur.2020.576194

Nassar, M., Salah, K., ur Rehman, M. H., & Svetinovic, D. (2020). Blockchain for explainable and trustworthy artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(1). https://doi.org/10.1002/widm.1340

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). *InterpretML: A Unified Framework for Machine Learning Interpretability*.

Nuray, R., & Can, F. (2006). Automatic ranking of information retrieval systems using data fusion. *Information Processing and Management*, *42*(3). https://doi.org/10.1016/j.ipm.2005.03.023

Oh, K., Chung, Y. C., Kim, K. W., Kim, W. S., & Oh, I. S. (2019). Classification and Visualization of Alzheimer's Disease using Volumetric Convolutional Neural Network and Transfer Learning. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-54548-6

Ohm, T. G. (2007). The dentate gyrus in Alzheimer's disease. In *Progress in Brain Research* (Vol. 163). https://doi.org/10.1016/S0079-6123(07)63039-8

P..A.Menon, & R.Gunasundari. (2024). SHAP-based Feature Selection and Explainable Machine Learning Classification of Alzheimer's Disease. *Journal of Computational Analysis & Applications*, *33(6)*, 798.

Panesar, A. (2021). Machine Learning and AI for Healthcare. In *Machine Learning and AI for Healthcare*. https://doi.org/10.1007/978-1-4842-6537-6

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*. https://doi.org/10.1214/09-SS057

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8). https://doi.org/10.1109/TPAMI.2005.159

Peter Spirtes, Glymour C N, & Scheines R. (2000). *Causation, prediction, and search*. MIT Press.

Peters, F., Collette, F., Degueldre, C., Sterpenich, V., Majerus, S., & Salmon, E. (2009). The neural correlates of verbal short-term memory in Alzheimer's disease: an fMRI study. *Brain*, *132*(7). https://doi.org/10.1093/brain/awp075

Philip Chen, C. L., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*. https://doi.org/10.1016/j.ins.2014.01.015

Poulin, S. P., Dautoff, R., Morris, J. C., Barrett, L. F., & Dickerson, B. C. (2011). Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. *Psychiatry Research - Neuroimaging*, *194*(1). https://doi.org/10.1016/j.pscychresns.2011.06.014

Poursabzi-Sangdeh, F., Goldstein, D. G., & Hofman, J. M. (2021). Manipulating and measuring model interpretability. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3411764.3445315

Qiu, S., Chang, G. H., Panagia, M., Gopal, D. M., Au, R., & Kolachalama, V. B. (2018). Fusion of deep learning models of MRI scans, Mini–Mental State Examination, and logical memory test enhances diagnosis of mild cognitive impairment. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, *10*. https://doi.org/10.1016/j.dadm.2018.08.013

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*.

Rado, O., Ali, N., Sani, H. M., Idris, A., & Neagu, D. (2019). Performance Analysis of Feature Selection Methods for Classification of Healthcare Datasets. *Advances in Intelligent Systems and Computing*, *997*. https://doi.org/10.1007/978-3-030-22871-2_66

Radovic, M., Ghalwash, M., Filipovic, N., & Obradovic, Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics*, *18*(1). https://doi.org/10.1186/s12859-016-1423-9

Raghupathy, B. K., Reddy, M. R., Prasad Theeda, Balasubramanian, E., Namachivayam, R. K., & Ganesan, M. (2025). Harnessing Explainable Artificial Intelligence (XAI) based SHAPLEY Values and Ensemble Techniques for Accurate Alzheimer's Disease Diagnosis. *Engineering, Technology & Applied Science Research*, *15*(2), 20743–20747. https://doi.org/10.48084/etasr.9619

Rajan, K. B., Wilson, R. S., Weuve, J., Barnes, L. L., & Evans, D. A. (2015). Cognitive impairment 18 years before clinical diagnosis of Alzheimer disease dementia. *Neurology*, *85*(10). https://doi.org/10.1212/WNL.0000000000001774

Rao, Y. L., Ganaraja, B., Murlimanju, B. V., Joy, T., Krishnamurthy, A., & Agrawal, A. (2022). Hippocampus and its involvement in Alzheimer's disease: a review. In *3 Biotech* (Vol. 12, Issue 2). https://doi.org/10.1007/s13205-022-03123-4

Razavi, S., Jakeman, A., Saltelli, A., Prieur, C., Iooss, B., Borgonovo, E., Plischke, E., Lo Piano, S., Iwanaga, T., Becker, W., Tarantola, S., Guillaume, J. H. A., Jakeman, J., Gupta, H., Melillo, N., Rabitti, G., Chabridon, V., Duan, Q., Sun, X., … Maier, H. R. (2021). The Future of Sensitivity Analysis: An essential discipline for systems modeling and policy support. *Environmental Modelling and Software*, *137*. https://doi.org/10.1016/j.envsoft.2020.104954

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" explaining the predictions of any classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*. https://doi.org/10.18653/v1/n16-3020

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*. https://doi.org/10.1609/aaai.v32i1.11491

Rivest, R. L. (1987). Learning decision lists. *Machine Learning*, *2*(3), 229–246. https://doi.org/10.1007/BF00058680

Rosenbacke, R., Melhus, Å., McKee, M., & Stuckler, D. (2024). How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review. *JMIR AI*, *3*, e53207. https://doi.org/10.2196/53207

Ross, A. S., Hughes, M. C., & Doshi-Velez, F. (2017). Right for the right reasons: Training differentiable models by constraining their explanations. *IJCAI International Joint Conference on Artificial Intelligence*, *0*. https://doi.org/10.24963/ijcai.2017/371

Ross, C. A., & Poirier, M. A. (2004). Protein aggregation and neurodegenerative disease. *Nature Medicine*, *10*(7). https://doi.org/10.1038/nm1066

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. In *Nature Machine Intelligence* (Vol. 1, Issue 5). https://doi.org/10.1038/s42256-019-0048-x

Russell, C. (2019). Efficient search for diverse coherent explanations. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3287560.3287569

Ryman-Tubb, N. F., Krause, P., & Garn, W. (2018). How Artificial Intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark. In *Engineering Applications of Artificial Intelligence* (Vol. 76). https://doi.org/10.1016/j.engappai.2018.07.008

Sacchi, L., Contarino, V. E., Siggillino, S., Carandini, T., Fumagalli, G. G., Pietroboni, A. M., Arcaro, M., Fenoglio, C., Orunesu, E., Castellani, M., Casale, S., Conte, G., Liu, C., Triulzi, F., Galimberti, D., Scarpini, E., & Arighi, A. (2023). Banks of the Superior Temporal Sulcus in Alzheimer's Disease: A Pilot Quantitative Susceptibility Mapping Study. *Journal of Alzheimer's Disease*, *93*(3). https://doi.org/10.3233/JAD-230095

Sadiq, A., Yahya, N., & Tang, T. B. (2021). Diagnosis of Alzheimer's Disease Using Pearson's Correlation and ReliefF Feature Selection Approach. *2021 International Conference on Decision Aid Sciences and Application, DASA 2021*. https://doi.org/10.1109/DASA53625.2021.9682409

Sadun, A. A., & Bassi, C. J. (1990). Optic Nerve Damage in Alzheimer's Disease. *Ophthalmology*, *97*(1). https://doi.org/10.1016/S0161-6420(90)32621-0

Saltelli, A., Tarantola, S., & Chan, K. P. S. (1999). A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, *41*(1). https://doi.org/10.1080/00401706.1999.10485594

Salvatore, C., Battista, P., & Castiglioni, I. (2016). Frontiers for the Early Diagnosis of AD by Means of MRI Brain Imaging and Support Vector Machines. *Current Alzheimer Research*, *13*(5). https://doi.org/10.2174/1567205013666151116141705

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Muller, K.-R. (2019). Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. *Lecture Notes in Computer Science (LNCS)*, *11700*.

Sarica, A., Di Fatta, G., & Cannataro, M. (2014). K-Surfer: A KNIME extension for the management and analysis of human brain MRI FreeSurfer/FSL data. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8609 LNAI*. https://doi.org/10.1007/978-3-319-09891-3_44

Sarraf, S., & Tofighi, G. (2017). Deep learning-based pipeline to recognize Alzheimer's disease using fMRI data. *FTC 2016 - Proceedings of Future Technologies Conference*. https://doi.org/10.1109/FTC.2016.7821697

Scheff, S. W., Price, D. A., Schmitt, F. A., Scheff, M. A., & Mufson, E. J. (2011). Synaptic loss in the inferior temporal gyrus in mild cognitive impairment and Alzheimer's disease. *Journal of Alzheimer's Disease*, *24*(3). https://doi.org/10.3233/JAD-2011-101782

Scheff, S. W., Sparks, D. L., & Price, D. A. (1996). Quantitative assessment of synaptic density in the outer molecular layer of the hippocampal dentate gyrus in alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, *7*(4). https://doi.org/10.1159/000106884

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*(2). https://doi.org/10.1007/s11263-019-01228-7

Serrano, S., & Smith, N. A. (2020). Is attention interpretable? *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. https://doi.org/10.18653/v1/p19-1282

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *34th International Conference on Machine Learning, ICML 2017*, *7*.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings*.

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3375627.3375830

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, *55*(1–3). https://doi.org/10.1016/S0378-4754(00)00270-6

Song, M., Jung, H., Lee, S., Kim, D., & Ahn, M. (2021). Diagnostic classification and biomarker identification of alzheimer's disease with random forest algorithm. *Brain Sciences*, *11*(4). https://doi.org/10.3390/brainsci11040453

Spasov, S., Passamonti, L., Duggento, A., Liò, P., & Toschi, N. (2019). A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *NeuroImage*, *189*. https://doi.org/10.1016/j.neuroimage.2019.01.031

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*.

Stahl, B. C. (2021). *Ethical Issues of AI*. https://doi.org/10.1007/978-3-030-69978-9_4

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*. https://doi.org/10.1186/1471-2105-9-307

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *34th International Conference on Machine Learning, ICML 2017*, *7*.

Surden, H. (2021). Machine learning and law: An overview. In *Research Handbook on Big Data Law*. Edward Elgar Publishing. https://doi.org/10.4337/9781788972826.00014

Tamanini, I., De Castro, A. K., Pinheiro, P. R., & Pinheiro, M. C. D. (2009). Applied neuroimaging to the diagnosis of alzheimer's disease: A multicriteria model. *Communications in Computer and Information Science*, *49*. https://doi.org/10.1007/978-3-642-04757-2_57

Thibault Laugel. (2020). *Local post-hoc interpretability for black-box classifiers.* Sorbonne Université.

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *Proceedings of Machine Learning Research*, *106*.

Trambaiolli, L. R., Spolaôr, N., Lorena, A. C., Anghinah, R., & Sato, J. R. (2017). Feature selection before EEG classification supports the diagnosis of Alzheimer's disease. *Clinical Neurophysiology*, *128*(10). https://doi.org/10.1016/j.clinph.2017.06.251

Ustun, B., & Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, *102*(3). https://doi.org/10.1007/s10994-015-5528-6

Van Hoesen, G. W., Augustinack, J. C., Dierking, J., Redman, S. J., & Thangavel, R. (2000). The parahippocampal gyrus in Alzheimer's disease. Clinical and preclinical neuroanatomical correlates. *Annals of the New York Academy of Sciences*, *911*. https://doi.org/10.1111/j.1749-6632.2000.tb06731.x

van Hoesen, G. W., Hyman, B. T., & Damasio, A. R. (1991). Entorhinal cortex pathology in Alzheimer's disease. In *Hippocampus* (Vol. 1, Issue 1). https://doi.org/10.1002/hipo.450010102

Varghese, A., George, B., Sherimon, V., & Al Shuaily, H. S. (2023). Enhancing Trust in Alzheimer's Disease Classification using Explainable Artificial Intelligence: Incorporating Local Post Hoc Explanations for a Glass-box Model. *Bahrain Medical Bulletin*, *45*(2).

Vasconcelos, L. de G., Jackowski, A. P., de Oliveira, M. O., Ribeiro Flor, Y. M., Lino Souza, A. A., Amodeo Bueno, O. F., & Dozzi Brucki, S. M. (2014). The thickness of posterior cortical areas is related to executive dysfunction in Alzheimer's disease. *Clinics*, *69*(1). https://doi.org/10.6061/clinics/2014(01)05

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-December*.

Vemuri, P., & Jack, C. R. (2010). Role of structural MRI in Alzheimer's disease. In *Alzheimer's Research and Therapy* (Vol. 2, Issue 4). https://doi.org/10.1186/alzrt47

Vernooij, M. W., & van Buchem, M. A. (2020). *Neuroimaging in Dementia*.

Vig, J. (2019). A multiscale visualization of attention in the transformer model. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations*. https://doi.org/10.18653/v1/p19-3007

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*. https://doi.org/10.1016/j.inffus.2021.05.009

Viswan, V., Shaffi, N., Mahmud, M., Subramanian, K., & Hajamohideen, F. (2024). Explainable Artificial Intelligence in Alzheimer's Disease Classification: A Systematic Review. In *Cognitive Computation* (Vol. 16, Issue 1). https://doi.org/10.1007/s12559-023-10192-x

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3063289

Wanner, J., Herm, L. V., Heinrich, K., & Janiesch, C. (2021). Stop Ordering Machine Learning Algorithms by Their Explainability! An Empirical Investigation of the Tradeoff Between Performance and Explainability. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12896 LNCS*. https://doi.org/10.1007/978-3-030-85447-8_22

Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.18653/v1/d19-1002

Wu, C., Guo, S., Hong, Y., Xiao, B., Wu, Y., & Zhang, Q. (2018). Discrimination and conversion prediction of mild cognitive impairment using convolutional neural networks.

Quantitative Imaging in Medicine and Surgery, *8*(10). https://doi.org/10.21037/qims.2018.10.17

Wu, L., Chen, Y., Shen, K., Guo, X., Gao, H., Li, S., Pei, J., & Long, B. (2023). Graph Neural Networks for Natural Language Processing: A Survey. *Foundations and Trends in Machine Learning*, *16*(2). https://doi.org/10.1561/2200000096

Wyawahare, M. V, Patil, P. M., & Abhyankar, H. K. (2009). Image Registration Techniques: An overview. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, *2*(3).

Y. C.A.P.Reddy, Viswanath, P., & Eswara Reddy, B. (2018). Semi-supervised learning: a brief review. *International Journal of Engineering & Technology*, *7*(1.8), 81. https://doi.org/10.14419/ijet.v7i1.8.9977

Yang, H., Xu, H., Li, Q., Jin, Y., Jiang, W., Wang, J., Wu, Y., Li, W., Yang, C., Li, X., Xiao, S., Shi, F., & Wang, T. (2019). Study of brain morphology change in Alzheimer's disease and amnestic mild cognitive impairment compared with normal controls. *General Psychiatry*, *32*(2). https://doi.org/10.1136/gpsych-2018-100005

Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *Proceedings, Twentieth International Conference on Machine Learning*, *2*.

Zafar, M. R., & Khan, N. M. (2019). *DLIME: A Deterministic Local Interpretable Model-Agnostic Explanations Approach for Computer-Aided Diagnosis Systems*.

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J. B., Yuan, J., & Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. In *Complexity* (Vol. 2021). https://doi.org/10.1155/2021/8812542

Zhang, D., & Shen, D. (2012). Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *PLoS ONE*, *7*(3). https://doi.org/10.1371/journal.pone.0033182

Zhang, F., & Jiang, L. (2015). Neuroinflammation in Alzheimer's disease. *Neuropsychiatric Disease and Treatment*, 243. https://doi.org/10.2147/NDT.S75546

Zhang, J., Chen, B., Zhang, L., Ke, X., & Ding, H. (2021). Neural, symbolic and neural-symbolic reasoning on knowledge graphs. In *AI Open* (Vol. 2). https://doi.org/10.1016/j.aiopen.2021.03.001

Zhao, W., Wang, X., Yin, C., He, M., Li, S., & Han, Y. (2019). Trajectories of the hippocampal subfields atrophy in the Alzheimer's disease: A structural imaging study. *Frontiers in Neuroinformatics*, *13*. https://doi.org/10.3389/fninf.2019.00013

Appendix A

MRI-Derived Dataset, Access and Reproducibility

The MRI and cognitive assessment datasets used in this thesis were sourced from established research repositories. Due to licensing and data-use restrictions, these datasets cannot be redistributed directly as part of this thesis or any accompanying repository. To support full reproducibility, the official access points, approval requirements, and download instructions are provided below.

1. ADNI
    - Access portal: ADNI Website
    - Access requires registration, acceptance of the Data Use Agreement, and approval from the data managers of the repository.
    - Imaging scans (T1-weighted MRI) and associated clinical/cognitive metadata can be downloaded in standard formats.
2. AIBL
    - Access portal: AIBL website
    - Access is granted upon free registration, agreement to the Data Use Policy, and approval from the data administrators.
    - Imaging scans (T1-weighted MRI) and associated clinical/cognitive metadata can be downloaded in standard formats.
3. IXI
    - Access portal: IXI website
    - Access requires free registration, acceptance of the Data Use Agreement, and approval from the data administrators.
    - Imaging scans (T1-weighted MRI) and associated clinical/cognitive metadata can be downloaded in standard formats.

Researchers intending to reproduce the experiments should obtain the datasets directly from the corresponding repositories, follow the preprocessing pipelines and apply the feature extraction procedures described in Chapter 3. This appendix ensures that the entire experimental workflow can be replicated without violating any data-sharing policies.

All structural MRI features were generated using FreeSurfer and grouped into cortical morphometry, subcortical volumes, ventricular measures, and hippocampal subfields.

Annexure A Table - 1 Overview of Feature Types and Counts

Anatomical Category	Area	Thickness	Thickness SD	Mean Curvature	Volume	Total Features
Left Hemisphere Cortex	34	34	34	34	34	170
Right Hemisphere Cortex	34	34	34	34	34	170
Subcortical and Ventricular Structures	-	-	-	-	29	29
Corpus Callosum Regions	-	-	-	-	6	6
Hippocampal Subfields (L/R)	-	-	-	-	26	26
Total	68	68	68	68	129	401

Raw datasets (ADNI, AIBL, IXI) cannot be redistributed, but the full list of feature names used in this thesis is provided below for transparency and reproducibility.

(A) Features from Left Hemisphere Cortical Regions

Each region includes: area, meancurv, thickness, thicknessstd, volume.

lh_bankssts
lh_caudalanteriorcingulate
lh_caudalmiddlefrontal
lh_cuneus
lh_entorhinal
lh_fusiform
lh_inferiorparietal
lh_inferiortemporal
lh_isthmuscingulate
lh_lateraloccipital
lh_lateralorbitofrontal
lh_lingual
lh_medialorbitofrontal
lh_middletemporal
lh_parahippocampal
lh_paracentral
lh_parsopercularis
lh_parsorbitalis

lh_parstriangularis
lh_pericalcarine
lh_postcentral
lh_posteriorcingulate
lh_precentral
lh_precuneus
lh_rostralanteriorcingulate
lh_rostralmiddlefrontal
lh_superiorfrontal
lh_superiorparietal
lh_superiortemporal
lh_supramarginal
lh_frontalpole
lh_temporalpole
lh_transversetemporal
lh_insula

(B) Features from Right Hemisphere Cortical Regions

Each region includes: area, meancurv, thickness, thicknessstd, volume.

rh_bankssts
rh_caudalanteriorcingulate
rh_caudalmiddlefrontal
rh_cuneus
rh_entorhinal
rh_fusiform
rh_inferiorparietal
rh_inferiortemporal
rh_isthmuscingulate
rh_lateraloccipital
rh_lateralorbitofrontal
rh_lingual
rh_medialorbitofrontal
rh_middletemporal
rh_parahippocampal
rh_paracentral
rh_parsopercularis
rh_parsorbitalis
rh_parstriangularis
rh_pericalcarine
rh_postcentral
rh_posteriorcingulate

rh_precentral
rh_precuneus
rh_rostralanteriorcingulate
rh_rostralmiddlefrontal
rh_superiorfrontal
rh_superiorparietal
rh_superiortemporal
rh_supramarginal
rh_frontalpole
rh_temporalpole
rh_transversetemporal
rh_insula

(C) Features from Subcortical & Ventricular Structures

Left-Lateral-Ventricle
Left-Inf-Lat-Vent
Left-Cerebellum-White-Matter
Left-Cerebellum-Cortex
Left-Thalamus-Proper
Left-Caudate
Left-Putamen
Left-Pallidum
Left-Amygdala
Left-Accumbens-area
Left-VentralDC
Left-choroid-plexus
Right-Lateral-Ventricle
Right-Inf-Lat-Vent
Right-Cerebellum-White-Matter
Right-Cerebellum-Cortex
Right-Thalamus-Proper
Right-Caudate
Right-Putamen
Right-Pallidum
Right-Amygdala

Right-Accumbens-area
Right-VentralDC
Right-choroid-plexus
3rd-Ventricle
4th-Ventricle
Brain-Stem
Optic-Chiasm

(D) Features from Corpus Callosum Regions & Supratentorial Volume

CC_Posterior
CC_Mid_Posterior
CC_Central
CC_Mid_Anterior
CC_Anterior
SupraTentorialVolNotVent

(E) Features from Hippocampal Subfields (Left & Right)

left_Hippocampal_tail
left_subiculum
left_CA1
left_hippocampal-fissure
left_presubiculum
left_parasubiculum
left_molecular_layer_HP
left_GC-ML-DG
left_CA3
left_CA4
left_fimbria
left_HATA
left_Whole_hippocampus
right_Hippocampal_tail
right_subiculum
right_CA1

right_hippocampal-fissure
right_presubiculum
right_parasubiculum
right_molecular_layer_HP
right_GC-ML-DG
right_CA3
right_CA4
right_fimbria
right_HATA
right_Whole_hippocampus

Appendix B

Reproducibility Resources

To support reproducibility and transparency of the experiments conducted in this thesis, the full source code used for feature selection, sensitivity analysis, and transfer-learning modules has been made publicly available on GitHub.

Source Code Repository

The complete implementation developed for this thesis is accessible at:

GitHub: https://github.com/akhilatmakuru/Research_Papers

The repository includes:

- Python scripts for all proposed algorithms
- A README file with instructions to run the code.

'requirements.txt' listing all dependencies and versions.