University of Sheffield

# Deep Learning-Based Interpretable Multimodal Data Fusion for Early Diagnosis of Alzheimer's Disease


Shunbao Li

*Supervisor:* Dr. Po Yang

A thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy in Computer Science

*in the*

School of Computer Science

August 11, 2026

# Declaration

All sentences or passages quoted in this document from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure.

Name: Shunbao Li

Signature:

Date:

# Abstract

As the global population ages, Alzheimer's disease (AD) poses an increasing threat to public health. Early diagnosis is critical to enable timely intervention, monitoring, and trial recruitment before severe neurodegeneration occurs. Machine learning (ML) can assist clinicians with diagnostic decision-making and aid researchers in identifying biologically meaningful biomarkers. However, current diagnostic tools face major challenges: high-dimensional, small-sample gene expression data; limited interpretability of deep learning models; and the loss of complementary information during heterogeneous multimodal data fusion.

This thesis addresses these challenges through three main contributions. First, a stacked sparse autoencoder (SSAE) is developed for gene expression dimensionality reduction. By incorporating sparsity constraints and stacking hidden layers, the SSAE effectively models non-linear relationships, mitigates overfitting, and outperforms traditional dimensionality reduction algorithms in early AD classification.

Second, an interpretable feature selection framework is proposed, combining a shallow sparse autoencoder (AE) with XGBoost-based feature importance ranking. This approach captures complex non-linear data structures while retaining direct mapping between hidden nodes and input probes. By providing transparent feature selection, it addresses the black-box limitation of conventional deep learning and identifies gene expression probes with validated biological relevance through enrichment analysis.

Third, a novel heterogeneous multimodal data fusion method is introduced to integrate neuroimaging and gene expression data. Unlike existing fusion techniques that focus primarily on cross-modal consistency, this framework preserves modality-unique complementary information to construct a richer fused representation, outperforming both unimodal baselines and consistency-only multimodal methods.

In summary, this thesis presents an interpretable deep learning framework for multimodal data fusion, improving both diagnostic accuracy and biological interpretability in early AD diagnosis. Future work will focus on validating the framework across larger multi-cohort datasets, incorporating additional omics modalities, and enhancing computational efficiency and model robustness for clinical decision support.

# Glossary and Acronyms

**AD** Alzheimer's disease.

**ADAS-Cog** Alzheimer's Disease Assessment Scale-Cognitive Subscale.

**ADNI** Alzheimer's Disease Neuroimaging Initiative.

**AE** Autoencoder.

**ANM** AddNeuroMed.

**AUC** Area under the receiver operating characteristic curve.

**CCA** Canonical correlation analysis.

**CDR** Clinical Dementia Rating.

**CNN** Convolutional neural network.

**CSF** Cerebrospinal fluid.

**DAE** Denoising autoencoder.

**DCCA** Deep canonical correlation analysis.

**DEG** Differentially expressed gene.

**DNN** Deep neural network.

**DTI** Diffusion tensor imaging.

**EEG** Electroencephalography.

**EHR** Electronic health record.

**EMCI** Early mild cognitive impairment.

**FDG-PET** Fluorodeoxyglucose positron emission tomography.

**fMRI** Functional magnetic resonance imaging.

**GCN** Graph convolutional network.

**GBM** Gradient boosting machine.

**GO** Gene Ontology.

**ICA** Independent component analysis.

**ISOMAP** Isometric feature mapping.

**KEGG** Kyoto Encyclopedia of Genes and Genomes.

**KPCA** Kernel principal component analysis.

**LASSO** Least absolute shrinkage and selection operator.

**LLE** Locally linear embedding.

**MCI** Mild cognitive impairment.

**ML** Machine learning.

**MML** Multimodal metric learning.

**MMSE** Mini-Mental State Examination.

**MDML** Multimodal deep metric learning.

**MRI** Magnetic resonance imaging.

**NC** Normal control.

**PCA** Principal component analysis.

**PET** Positron emission tomography.

**PhFT** Phonemic fluency test.

**PRE** Precision.

**pMCI** Progressive mild cognitive impairment.

**PRS** Polygenic risk score.

**RF** Random forest.

**RFE** Recursive feature elimination.

**RAVLT** Rey Auditory Verbal Learning Test.

**ROC** Receiver operating characteristic.

**RR** Ridge regression.

**RMA** Robust Multi-array Average.

**ROI** Region of interest.

**SDAE** Stacked denoising autoencoder.

**SEN** Sensitivity.

**SFT** Semantic fluency test.

**sMCI** Stable mild cognitive impairment.

**sMRI** Structural magnetic resonance imaging.

**SPE** Specificity.

**SNP** Single nucleotide polymorphism.

**SSAE** Stacked sparse autoencoder.

**SVM** Support vector machine.

**TMT** Trail-making test.

**XGBoost** Extreme Gradient Boosting.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Alzheimer's disease (AD) is a common neurodegenerative disease that occurs mostly in older adults. With the ageing of the global population, AD has increasingly become a major threat to human health. AD is a progressive, irreversible, and fatal disease. The current therapies are ineffective for patients whose condition has become advanced. However, early diagnosis enables the implementation of supportive treatments that can mitigate the progression of the disease (Fathi et al., 2022). This thesis will investigate interpretable machine learning (ML) and deep learning techniques for supporting effective early diagnosis of AD patients. This chapter provides an overview of the background and motivation for this thesis, its stated aims and objectives, and the knowledge contributions it makes.

## 1.1 Background

AD is the most common cause of dementia, accounting for about 60%-80% of cases (Alzheimer's Association, 2024). The number of people affected by dementia and AD has increased as populations age. For example, Wittenberg et al. (2019) estimate that about 850,000 people in the UK were living with dementia in 2019, with an approximate prevalence of 7.1% among older people. The risk of AD also increases sharply with age: around 10% of adults may suffer from AD at the age of 65, rising to about 50% after the age of 85 (Kumar et al., 2018, World Health Organization, 2022).

AD was first discovered by German doctor Alois Alzheimer as a neurodegenerative disease of the central nervous system (Goedert and Spillantini, 2006). Many recent studies have shown that AD pathology involves the accumulation of $\beta$-amyloid plaques outside neurons in the brain and twisted strands of $\tau$ tangles inside neurons, accompanied by neuronal death and brain tissue damage (Song et al., 2022). Previous studies of AD also have revealed that such brain changes can begin 20 years or more before symptoms appear (Qin et al., 2022). Due to the death of specific functional neurons, patients with AD gradually lose the ability to live independently over time. This pathological process is called the AD continuum (Gustavsson et al., 2023). In terms of AD progression, researchers have divided the AD continuum into three phases (see Figure 1.1) (Alzheimer's Association, 2024):


Figure 1.1: Progression of AD from preclinical changes through MCI and dementia, illustrating the clinical continuum used in this thesis.

- **Preclinical AD:** Minor brain changes occur but the brain compensates for them, enabling individuals to function as normal.
- **MCI due to AD:** Measurable biomarkers of brain changes occur. These changes exceed the brain's compensation capacity and cause mild cognitive problems that may be noticed by people close to the individual.
- **Dementia due to AD:** Noticeable cognitive symptoms occur when the individual's condition enters this phase. The severity of the symptoms is related to the degree of damage to the patient's brain.

AD is a progressive, irreversible, and ultimately fatal condition. The progression of AD is associated with the extent of neuronal damage in the brain. Currently, no pharmacological or non-pharmacological treatments have been proven effective

in curing or preventing AD (Alzheimer's Association, 2024). However, emerging evidence suggests that certain interventions can improve life expectancy and quality of life in patients before they transition to the AD phase (Fessel, 2019, Lissek and Suchan, 2021, Zanetti et al., 2009, Li et al., 2011). Consequently, early-stage intervention is crucial, making early diagnosis essential (Prince et al., 2014).

Limited by the current understanding of AD, early diagnosis is mainly aimed at patients in the MCI stage rather than the preclinical AD stage. The current diagnostic techniques of preclinical AD depend on decreased metabolism of glucose and abnormal levels of amyloid-$\beta$42. However, only some research institutions have the ability to recognise these brain changes that cause AD, and most clinical settings do not yet have this ability. Furthermore, in an analysis of neuropathology of cognitively normal elderly people, Knopman et al. (2021) found that very few subjects with these AD-related brain changes had converted to AD. In contrast to the low incidence of AD in individuals with preclinical AD, previous studies have observed that MCI patients have a cumulative conversion rate of 38.2% within 5 years or more (Mitchell and Shiri-Feshki, 2009). Moreover, it has been established that a series of testing methods – including cerebrospinal fluid tests, neuropsychological tests, neuroimaging tests, genetic tests, and molecular probe tests – are widely available in clinical settings and can be used to diagnose MCI. Therefore, recent years have seen increased interest in MCI diagnosis as a route to early AD diagnosis (Alzheimer's Association, 2024).

## 1.2 Motivation

In recent years, medical doctors and computer scientists have attempted to achieve early AD diagnosis using a variety of biomarkers. Depending on the modality, biomarkers can be classified as molecular, neuropsychological, neuroimaging, and genetic biomarkers. Cerebrospinal fluid (CSF) analysis is an effective method of detecting molecule biomarkers. It can be used to detect levels of Amyloid-$\beta$42, Total $\tau$ protein and Phosphorylated $\tau$ protein in cerebrospinal fluid (Wattmo et al., 2020). However, CSF testing is an invasive test that causes some damage to the

human body, with contraindications and risk of infection, it is generally difficult for patients to accept, despite its diagnostic reliability.

Neuroimaging techniques, including positron emission tomography, magnetic resonance imaging, diffusion tensor imaging, and functional MRI, allow non-invasive detection of structural and functional changes in brain tissue. Many studies have shown that the combination of neuroimaging biomarkers such as MRI and PET is an effective multimodal neuroimaging modality (Khagi and Kwon, 2020, Mila-Aloma et al., 2021, Mayblyum et al., 2021). Song et al. (2021) developed a new fusion modality called 'GM-PET' by aligning and mask coding grey matter (GM) tissue regions from MRI and FDG-PET images, resulting in a synthetic image that highlights the GM region essential for AD diagnosis while maintaining the contours and metabolic features of the brain, achieving 94.11% accuracy for AD/NC classification, though the accuracy decreased to 71.52% for the AD/MCI/NC classification task. Lu et al. (2018) proposed a deep learning-based framework for diagnosing AD by integrating multimodal neuroimaging data from MRI and FDG-PET, which allow for in vivo assessment of brain structure and glucose metabolism; the proposed multiscale DNN achieved 82.4% accuracy in identifying patients with MCI. However, the features it extracted for classification lacked pathophysiologically meaningful interpretations. Rallabandi and Seetharaman (2023) developed an Inception-ResNet wrapper model to distinguish healthy controls, MCI, and AD using joint MRI and PET. The accuracy was 95.5%, 94.1%, and 95.9% in classifying healthy controls vs MCI, MCI vs AD, and AD vs healthy controls, respectively.

However, despite their utility, neuroimaging techniques have several limitations. First, they are expensive and resource-intensive, making them impractical for large-scale population screening. Second, while neuroimaging biomarkers provide valuable information about macroscopic changes in brain structure and function, they do not capture the molecular and cellular processes underlying these changes. This gap is particularly significant in the context of early diagnosis, as neuroimaging may fail to detect subtle pathological changes that precede observable structural abnormalities.

In contrast, gene expression data offers a complementary perspective by provid-

ing insights into the molecular mechanisms driving AD progression. Gene expression data, which measures mRNA abundance in cells, can reveal dysregulation in specific pathways and genes associated with AD. For instance, Karlsson Linnér et al. (2019) demonstrated that gene expression profiles in peripheral blood are partially correlated with those in brain tissue, suggesting that blood-based gene expression data can serve as a surrogate marker for brain pathology. Similarly, Horvath and Raj (2018) identified a common set of age-related co-methylated genes in human brain and blood that are involved in functions such as embryonic morphogenesis, transcriptional regulation, and neuronal differentiation.

The potential of gene expression data for AD diagnosis is further supported by its advantages over traditional biomarkers. Unlike cerebrospinal fluid (CSF) testing, which is invasive and associated with risks of infection, blood-based gene expression tests are minimally invasive and more acceptable to patients. Additionally, gene expression testing is relatively inexpensive, making it suitable for longitudinal studies and large-scale population screening. Furthermore, gene microarray technology has matured to the point where it can measure the expression levels of thousands of genes in a single experiment, enabling high-throughput analysis. This high throughput, however, is both a strength and a challenge, as the large number of genes and the complexity of their interactions necessitate advanced analytical methods.

One of the primary challenges in analysing gene expression data is its high dimensionality and the non-linear nature of its relationships. Gene expression datasets typically contain tens of thousands of features (genes) but only a limited number of samples, leading to the "curse of dimensionality" and increasing the risk of overfitting. Moreover, the intricate, non-linear interactions between genes further complicate the analysis, as traditional ML methods, such as PCA and clustering, are primarily designed to capture linear relationships and often fail to model these complex patterns effectively. To address these challenges, recent studies have turned to deep learning methods, particularly AEs, which are well-suited for handling high-dimensional, non-linear data. AEs are neural networks designed for unsupervised learning, capable of reducing dimensionality while preserving non-linear dependencies in the data. For example, Lee et al. (2019) demonstrated that AEs can extract

meaningful, low-dimensional representations from high-dimensional gene expression data, while Zhang and Bom (2021) applied AEs to AD gene expression datasets, achieving superior classification performance compared to traditional methods by capturing the underlying non-linear patterns.

Despite their promise, deep learning methods face a critical challenge in the form of interpretability. The features learned by AEs and other deep learning models are often abstract and difficult to map to specific biological processes or pathways. This lack of transparency limits their utility in clinical settings, where interpretability is essential for gaining the trust of medical professionals. Developing interpretable feature selection algorithms that combine the strengths of deep learning with the transparency of traditional ML is therefore a crucial research direction. Such algorithms could ensure that the features extracted from gene expression data are not only informative but also biologically meaningful, facilitating their integration into diagnostic frameworks.

Another major challenge is the integration of gene expression data with neuroimaging data. These two modalities are heterogeneous in scale, format, and biological meaning: neuroimaging biomarkers primarily reflect macroscopic structural or functional abnormalities, whereas gene expression data captures molecular processes that may precede or accompany those abnormalities. Fusion does not solve the cost of neuroimaging, but it can help ensure that expensive imaging measurements are interpreted alongside lower-cost molecular evidence when both data types are available. Effective multimodal fusion methods must therefore preserve complementary information, avoid collapsing one modality into the other, and remain interpretable enough to support biological analysis.

Existing multimodal ML approaches have important gaps in this setting. Some methods emphasise consistency between modalities, which can suppress modality-specific signals that are clinically informative. Other methods concatenate features or learn a single shared representation without explicitly modelling the different geometry and scale of gene expression and neuroimaging features. Metric learning algorithms, which learn task-specific distances or embeddings, offer a promising route because they can align heterogeneous data while preserving relationships

within each modality (Venugopalan et al., 2021). This thesis therefore investigates whether metric learning can support both improved classification performance and more interpretable multimodal analysis.

In summary, the motivation of this research is to address the following three research questions:

1. How can an AE be employed to improve the classification of AD gene expression datasets, especially when dealing with high-dimensional and non-linear data?

2. How can an interpretable feature selection algorithm be developed to combine the feature extraction capabilities of deep learning with the interpretability of traditional ML to ensure the selection of biologically meaningful probes in AD?

3. How can a heterogeneous multimodal fusion approach effectively integrate gene expression and neuroimaging data, preserving the complementary information from each modality, to improve the performance of AD classification?

## 1.3 Research Aims and Objectives

The aim of this research is to enhance the accuracy and interpretability of early AD diagnosis by developing deep learning, feature selection, and multimodal fusion methods for gene expression and neuroimaging data.

The preceding review of the problem identifies three gaps in existing ML approaches. First, many dimensionality reduction methods used for gene expression data are linear or are not designed for small-sample, high-dimensional biomedical data. Second, deep learning models can capture non-linear gene-expression patterns, but their latent features are often difficult to interpret biologically. Third, common multimodal fusion strategies tend to favour simple concatenation or consistency learning, which may lose complementary molecular and imaging information. These gaps motivate the following objectives:

The work of this thesis consists of the following objectives:

- To review common biomarkers for AD diagnosis, dimensionality reduction methods, feature selection methods, and multimodal data fusion methods, with emphasis on their strengths and limitations for early diagnosis.
- To develop a methodology for using gene expression data and neuroimaging data to support early AD diagnosis effectively.
- To develop an SSAE model that enhances classification performance for gene expression datasets by addressing high dimensionality, small sample size, and non-linear correlations. The stacked sparse architecture is motivated by the need to learn hierarchical non-linear representations while controlling overfitting through sparsity.
- To propose an interpretable feature selection framework that combines the feature extraction capability of a shallow sparse AE with the interpretability of model-based feature ranking, enabling the identification of biologically significant gene expression probes.
- To design a heterogeneous multimodal data fusion method that integrates neuroimaging data with gene expression data, emphasising complementarity between modalities and creating a comprehensive representation to improve AD classification.
- To validate the biological relevance of the selected gene expression probes through enrichment analysis and assess the performance of the proposed models against existing unimodal and multimodal approaches.

## 1.4 Knowledge Contribution

To summarise, the main knowledge contributions are:

1. An SSAE has been proposed to solve the problem that mainstream dimensionality reduction algorithms in the medical field are unable to effectively perform feature reduction on high-dimensional gene expression data with small samples. The proposed method improves the generalisation ability of the model

by introducing sparsity constraints to help the AE prevent overfitting. Meanwhile, it maps high-dimensional data to a lower-dimensional space by stacking hidden layers, retains important features, removes redundant information during dimensionality reduction, and substantially outperforms traditional feature dimensionality reduction algorithms in terms of early AD classification accuracy.

2. An interpretable deep learning feature selection framework is proposed integrating a shallow sparse AE for dimensionality reduction, with hidden layer nodes serving as features. Feature importance is ranked using XGBoost to filter high-weighted gene expression probes. The proposed framework outperforms the traditional conventional linear classification algorithms in the task of early diagnosis of AD due to the characteristic of the shallow sparse AE that can capture the non-linear relationship of the data. Meanwhile, the proposed framework solves the problem that traditional deep learning methods cannot extract features with interpretability due to the characteristic of transparency of shallow sparse AEs. The biological significance of the extracted gene expression probes is also verified by enrichment analysis experiments.

3. A heterogeneous multimodal data fusion method is proposed to fuse neuroimaging data with gene expression data for AD classification. In contrast to traditional multimodal fusion algorithms based on the consistency of different modalities, the proposed algorithm fuses heterogeneous modalities based on their complementarities, preserving the complementary information of each modality in order to obtain a comprehensive representation after data fusion. The proposed method outperforms unimodal methods that use only neuroimaging or gene expression data, and also outperforms multimodal fusion methods that consider only consistency in the task of fusing neuroimaging data with gene expression data.

## 1.5 Thesis Organisation

The thesis is divided into seven main chapters. Chapter 1 introduces the motivation, research questions, objectives, and knowledge contributions of this work. Chapter 2 reviews AD biomarkers, dimensionality reduction, feature selection, interpretability, and multimodal data fusion methods, identifying the gaps that motivate the contribution chapters. Chapter 3 summarises the common theory, datasets, preprocessing protocols, and evaluation metrics used across the thesis. Chapter 4 proposes and evaluates stacked AE-based dimensionality reduction for gene expression data. Chapter 5 presents an interpretable deep learning framework for selecting high-weight AD gene expression probes and validates the biological relevance of the selected probes through enrichment analysis. Chapter 6 investigates a heterogeneous data fusion method for integrating gene expression and neuroimaging data based on complementarity. Chapter 7 concludes the thesis and discusses limitations and future research directions.

## 1.6 List of Publications

I have published one conference paper related to this thesis, and one journal paper based on this thesis is in the major revision stage. They are listed below in chronological order.

- S. Li, P. Yang, and V. Lanfranchi, 'Examine and Evaluating Dimension Reduction Algorithms for Classifying Alzheimer's Diseases using Gene Expression Data', in 2021 17th International Conference on Mobility, Sensing and Networking (MSN), Dec. 2021, pp. 687–693. doi: 10.1109/MSN53354.2021.00106.
- S. Li, K. Liu, and P. Yang, 'An Interpretable Deep Learning Approach for the Diagnosis of Alzheimer's Disease Using Gene Expression Data' IEEE/ACM Transactions on Computational Biology and Bioinformatics (under major revision)

The next chapter moves from the problem definition and research objectives to

the literature base. It reviews the biomarkers and ML methods that inform the technical choices made in the subsequent methodology and contribution chapters.

# Chapter 2

# Literature Review

## 2.1 AD and MCI Diagnosis Biomarkers

In the past century, researchers have attempted to explain the pathogenesis of AD from different angles, and a variety of early diagnostic markers have been developed. The early diagnostic biomarkers that have been confirmed mainly include biochemical markers, neuropsychological markers, neuroimaging markers, and genetic markers.

### 2.1.1 Neuropsychological Biomarkers

In 1984, the National Institute of Neurological and Communicative Disorders and Stroke (NINDS) and the Alzheimer's Disease and Related Disorders Association (ADRDA) jointly formulated clinical diagnostic criteria for AD. Among the most commonly used indicators are the Mini-Mental State Examination (MMSE) and Clinical Dementia Rating (CDR) (Folstein et al., 1975, Morris, 1993). MMSE evaluates general cognitive status across abilities such as memory, orientation, comprehension, attention, reading, writing, and learning. It is simple to administer and is widely used as a preliminary examination of cognitive impairment. CDR focuses on six domains: memory, orientation, home and hobbies, personal care, judgement and problem solving, and community affairs. MMSE and CDR are therefore useful clinical indicators for distinguishing MCI and AD, although they should be interpreted

alongside other clinical and biological evidence. In addition, cognitive assessment scales include ADAS-Cog, RAVLT, PhFT, SFT, and TMT (Rosen et al., 1984, Schmidt, 1996).

### 2.1.2 Neuroimaging-Based Biomarkers

Neuroimaging techniques, such as PET and MRI, are used to detect changes in the structure and function of the brain (Alzheimer's Association, 2024). PET is a functional imaging method. By injecting radiotracers into the human body and observing metabolic activity, PET can reflect the activity of tissues or organs. It has been used in AD diagnosis because it can reveal pathological information related to glucose metabolism or amyloid deposition. However, this method requires radioactive tracers, which may be difficult for some patients to accept despite the clinical value of the information obtained. MRI is a commonly used neuroimaging method for assisting AD diagnosis because it can reflect brain structure non-invasively. sMRI can show regional atrophy associated with AD; bilateral hippocampal atrophy is particularly important, and the amygdala, prefrontal lobe, and temporal lobe can also be affected as the disease progresses (Liu, Pan, Li, Chen, Tang, Lu and Wang, 2018). Through the analysis of MRI images, researchers can extract grey matter structure and other image-derived features, although the high dimensionality of voxel data requires careful feature extraction and validation.

### 2.1.3 Gene Expression Biomarkers

Gene expression data in blood can also be used for the diagnosis of AD. Gene expression is the process by which genes are transcribed and translated to produce an active protein. Gene expression data, also known as transcriptomic data, refers to data reflecting mRNA abundance based on DNA microarray experiments. Sullivan et al. (2006*a*) showed that gene expression in peripheral blood partially correlated with gene expression in brain tissue and that differentially expressed genes in blood could be a surrogate marker for brain tissue lesions. Another study further found that gene co-expression modules of brain tissue are significantly conserved in blood

(Lunnon et al., 2013). Similarly, Horvath et al. (2012) found common age-related co-methylated genes in human brain and blood. In this context, co-methylated genes are genes whose DNA methylation patterns vary together across samples or tissues; such patterns may indicate shared regulatory processes relevant to ageing and neurodegeneration. From these study reports, the use of gene expression in blood instead of brain tissue is a viable alternative. In addition, the use of gene expression data for diagnosis has the following advantages over biochemical markers and neuroimaging markers.

1. Gene expression data reflects the abundance of mRNA in cells, which can reflect the current physiological state of cells and can reflect the changing status of the disease process earlier than the protein level detected by biochemical markers and the structural changes in brain reflected by neuroimaging markers. It is also more helpful to decipher the pathological mechanism of AD at the genetic level.

2. Compared with biochemical and imaging tests, gene expression tests are inexpensive and can meet the need for continuous monitoring.

3. Gene microarray is a mature technology and can measure the expression levels of thousands of genes simultaneously in one experiment, which has the feature of high throughput.

4. The high throughput of gene expression data is both an advantage and a challenge. The large number of genes, intrinsic expression patterns that are not yet fully understood, and the spatial-temporal specificity of gene expression make the analysis of gene expression data a more challenging task for a variety of reasons. This is one of the reasons why gene expression data are rarely used for diagnosis. However, with the development of analysis technology, this difficulty should be overcome gradually.

## 2.2 Dimensionality Reduction of High-Dimensional Gene Expression Data

AD is a common neurodegenerative disease that occurs mostly in older adults. With the ageing of the global population, AD has increasingly become a major threat to human health. AD is a progressive, irreversible, and fatal disease and the current therapies are not effective for patients whose condition has become advanced (Castellani et al., 2010). In terms of AD progression, researchers have divided the AD continuum into three phases: NC, MCI, and dementia due to AD (Alzheimer's Association, 2024). The progression of AD depends on the extent of neuronal damage in the brain. At present, no pharmacological treatments or non-pharmacological therapies have been proved to be effective in the treatment and prevention of AD (Alzheimer's Association, 2024). However, recent evidence suggests that some interventional treatments have a positive effect on the life expectancy and quality of life of patients whose condition has not entered the AD phase (Fessel, 2019, Lissek and Suchan, 2021, Zanetti et al., 2009, Li et al., 2011). Therefore, recent studies have begun to focus on early AD diagnosis (Prince et al., 2014).

To achieve early diagnosis of AD, scientists have experimented with data from multiple modalities. Decreased levels of $A\beta 42$ and increased concentrations of T-tau and P-tau in the CSF are now recognised as AD-specific biomarkers (Gatz et al., 2006). However, these biomarkers are mainly obtained from solid brain tissue. This method is damaging to the body, has contraindications and risks of infection, and although diagnostic reliability is high, it is not acceptable to all patients. Neuroimaging uses PET, MRI, DTI, and fMRI to detect structural and functional changes in brain nerves to classify NC and AD&MCI (Liu, Pan, Li, Chen, Tang, Lu and Wang, 2018). However, this method is expensive and therefore not all hospitals are equipped to perform the test and not all patients can afford it. Compared to these methods, extracting human peripheral blood is a less invasive and more affordable method. Therefore, researchers are working to find alternative diagnostic markers in the peripheral blood. Sullivan et al. (2006*b*) showed that gene expression in peripheral blood correlates in part with gene expression in brain

tissue, and that DEGs in blood can be a surrogate marker for brain tissue lesions. This has made it possible to use gene expression data from peripheral blood for AD classification.

Although gene expression data has the advantages of high throughput, affordability and low invasiveness, researchers have used it less often for AD classification studies because gene expression datasets are high-dimensional and usually have small sample sizes (Espezua et al., 2015). High dimensionality can provide a detailed description of the problem, but it also introduces redundant, irrelevant, and noisy features. Due to these characteristics, high-dimensional small-sample data can easily cause the "curse of dimensionality", where the data space becomes sparse and computational complexity increases rapidly. In addition, because the number of samples is small, the effectiveness of traditional classification learning methods is reduced and overfitting can easily occur. Dimensionality reduction is therefore important for mitigating this problem while improving classifier generalisation.

### 2.2.1 PCA

PCA is a linear mapping. It transforms the data into a new coordinate system such that the first greatest variance of any data projection is at the first coordinate (called the first principal component), the second greatest variance is at the second coordinate (the second principal component), and so on (Jolliffe and Cadima, 2016). PCA is often used to reduce the dimensionality of a dataset while maintaining the features of the dataset that contribute the most to the variance. The main steps of PCA are as follows.

1. Using each feature of the original feature set as a column vector, construct a sample matrix X of N*M, where X is the feature length and M is the number of features

2. Calculating the mean vector of the samples: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

3. Calculating the residuals between vectors: $\tau_i = x_i - \bar{x}$

4. Constructing the matrix $A = \delta\left[\tau_1, \tau_2, \cdots, \tau_m\right]$

5. Calculating the covariance matrix:$C = \frac{1}{N}\sum_{i=1}^{N}\tau_i\tau_i^T = AA^T$,where C is an m×m dimensional matrix to represent the scatter of the data

6. Calculating the eigenvalues and eigenvectors of the covariance matrix C. Sort the eigenvalues from greatest to smallest, such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$, and adjust the positions of the eigenvectors accordingly to obtain the eigenvector matrix.

7. The base formed by the eigenvectors corresponding to the first $d(d \ll m)$ eigenvalues is selected as the feature space and N samples are projected into the feature space to obtain $V_i = U^T x_i, i = 1, 2, *, N$, thus we reduce the samples from m dimensions to d dimensions. d values are usually set in advance by the user and can be determined by setting a threshold for the cumulative contribution rate (CR) (usually above 70%): $CR = \frac{\sum_{i=1}^{d}\lambda_i}{\sum_{i=1}^{m}\lambda_i}$.

### 2.2.2 KPCA

KPCA is a non-linear extension of PCA obtained by applying the kernel method to PCA. For linearly indistinguishable datasets, KPCA first maps the dataset to a higher dimensional space and then performs standard PCA (Schölkopf et al., 1997). In our study, we use the Gaussian kernel as the kernel function:

$$K\big(x_i, x_j\big) = \exp\big(\frac{-\|x_i - x_j\|^2}{\sigma^2}\big) \tag{2.1}$$

### 2.2.3 ISOMAP

In 2000, Tenenbaum et al. (2000) proposed the ISOMAP algorithm. The algorithm is based on multidimensional scaling (MDS). Its main idea is to minimise the change in relative distance between data points in the original space and the corresponding low-dimensional representation.

In the traditional MDS algorithm, the distance between data points is calculated using Euclidean distance, which is only applicable to linear space data. ISOMAP uses the geodesic distance to calculate the distance between data points, which can

reflect the internal structure of the data in the non-linear space of the manifold (Tenenbaum et al., 2000).

The main steps of ISOMAP are as follows.

1. Construct a weighted adjacency graph G. Use $\varepsilon$ neighbourhood or k-nearest neighbour to determine the edge of the adjacent graph, and the weight of the edge is the Euclidean distance between the data points

2. Calculate the shortest path of every two points in the graph G, the distance matrix is: $D_G = \{d_G(i, j)\}$

3. Use MDS to calculate low-dimensional embedded coordinates

$$coordinates : \begin{cases} H = (H_{ij}) = \left(\delta_{ij} - \frac{1}{N}\right) \\ S = (S_{ij}) = (D_{ij}^2) \end{cases} \tag{2.2}$$

4. The obtained d-dimensional low-dimensional embedding coordinates are the eigenvectors corresponding to the largest d eigenvalues of $\tau(D)$

### 2.2.4 LLE

LLE is an unsupervised manifold learning algorithm (Roweis and Saul, 2000). It is effective for non-linear data that lie on a manifold structure. The basic principle of the LLE algorithm is to assume that each sample and its neighbouring data points are locally linear or close to linear, so that the non-linear manifold space can be approximated by local linear relationships and a local reconstruction weight matrix can be calculated to obtain a low-dimensional mapping.

The main steps are as follows:

1. Calculate the distance between each sample point $x_i$ and its close points in the original high-dimensional data and select $k(k < n)$ points close to each other as the adjacent points.

2. Calculate local reconstruction weight matrix: Define the error function:

$$\min \varepsilon(w) = \sum_{i=1}^{N} \left| x_i - \sum_{j=1}^{k} w_{ij} x_{ij} \right|^2 \tag{2.3}$$

where $x_{ij}(j = 1, 2, \cdots, k)$ are the $k$ neighbouring points of $x_i$, $w_{ij}$ is the weight and $\sum_{i=1}^{k} w_{ij} = 1$, the error for any point $x_i$ is:

$$\varepsilon = \sum_{j=1}^{k} \sum_{m=1}^{k} w_{ij} w_{im} Q_{jm}^{i} \tag{2.4}$$

where $Q_{jm}^{i} = (x_i - x_j)^T (x_i - x_m)$. Using the Lagrange multiplier method, the local reconstruction weight matrix can be obtained:

$$w_{ij} = \frac{\sum_{m=1}^{k} (Q_{jm}^{i})_{pq}^{-1}}{\sum_{p=1}^{k} \sum_{q=1}^{k} \Sigma_{q=1}^{k} (Q_{jm}^{i})_{pq}^{-1}} \tag{2.5}$$

3. Use the local neighbourhood structure learned in the high-dimensional space to project the data into the low-dimensional space. Since the local structure of the high-dimensional space needs to be preserved, the weight matrix remains unchanged, and the mapping objective function is:

$$min\varepsilon(Y) = \sum_{i=1}^{N} \left| y_i - \sum_{j=1}^{k} w_{ij} y_{ij} \right|^2 = \sum_{j=1}^{N} \sum_{m=1}^{N} M_{ij} y_i^T y_j \tag{2.6}$$

And meet the condition:

$$\begin{cases} \sum_{i=1}^{N} y_i = 0 \\ \frac{1}{N} \sum_{i=1}^{N} y_i^T y_j = I \end{cases} \tag{2.7}$$

In the objective function, $M$ is an $n \times n$ symmetric matrix: $M = (I - W)^T (I - W)$, using LaGrange multiplier method can get:

$$MY^T = \lambda Y^T \tag{2.8}$$

To minimize the objective function, $Y$ takes the eigenvector corresponding to

the smallest eigenvalue of $M$, and arranges the eigenvectors of $M$ in descending order of eigenvalues, we usually take the eigenvectors corresponding to $2 \sim (d+1)$ as the dimension of the low-dimensional space.

### 2.2.5 AE-Based Dimensionality Reduction

AEs have increasingly been used for dimensionality reduction in high-dimensional biological data because they can learn non-linear latent representations. For example, Xie et al. (2017) proposed a deep AE model for gene expression prediction, demonstrating that neural representation learning can capture complex gene expression patterns. Wang and Gu (2018) proposed VASC, a variational AE for dimension reduction and visualisation of single-cell RNA sequencing data, showing that probabilistic AE architectures can be useful when biological expression data are noisy and sparse. These studies indicate the potential of AEs for gene expression analysis, but they also highlight limitations relevant to this thesis: deeper representations can improve predictive performance while making it harder to trace latent features back to biological probes, and many published studies focus on general expression representation rather than early AD diagnosis. This motivates the comparison of stacked AEs for classification in Chapter 4 and the use of a shallow sparse architecture for interpretability in Chapter 5.

## 2.3 Feature Selection for Gene Expression Data

Feature selection, which aims to optimise the performance of an ML algorithm by discovering the smallest possible subset of features, is a challenging combinatorial optimisation problem: firstly, the search space grows exponentially with the number of features, and secondly, there are complex interactions between the features. For example, two correlated features containing information of a similar class can lead to redundancy. On the contrary, two weakly correlated traits can jointly provide complementary information about the classification task and are called complementary traits; thirdly, trait selection is essentially a multiple-objective optimisation problem, and its two main objectives are to maximise the classification performance

and to minimise the number of selected traits. In today's society, high-dimensional data are widely available in many practical applications. In the task of disease classification, biological data usually have high-dimensional features, including a large number of irrelevant, noisy or redundant features, which severely affect the classification performance. Specifically, irrelevant or noise features often degrade classification performance due to their misleading information. Redundant features do not improve the classification performance and result in longer computation time. High-dimensional data also significantly increases storage space and processing time, and makes ML models difficult to interpret. In addition, due to the 'curse of dimensionality', ML methods require a large number of training samples to obtain reliable results, otherwise they are prone to overfitting. Feature selection improves classification performance by selecting some important features to reduce data dimensions, reduce storage space usage, improve computational efficiency, and facilitate data visualisation. At the same time, it preserves the original semantics of the data, making it an interpretable and easy-to-understand method.

The rapid development of information technology in the current era has generated a huge amount of high-dimensional data in the fields of biomedicine and bioinformatics. In biological data, gene expression data is the activity of genes at transcriptional levels under specific conditions, which can reveal how genes are stimulated or repressed in specific environments, at specific time points, or in specific cell types. The analysis and interpretation of gene expression data are of great significance to the research of gene expression data as they can help in the interpretation and discovery of gene functions, understanding of disease mechanisms, drug development and personalised medicine. The development of bioinformatics has greatly facilitated the analysis and application of gene expression data, such as single-cell RNA sequencing, microarrays, and other technologies that reflect the expression of genes through different technological means, providing abundant data support for the analysis and research of gene expression data.

At the same time, the growth of data has created demands for effective and efficient data management. Data mining provides an effective approach to data management. Feature selection plays a crucial role in the field of data mining,

which essentially means selecting some features from the original set of features to form a new subset of features without significantly losing the potential information of the original set of data, and is commonly used in ML, data mining, and bioinformatics (Wang et al., 2018, Li et al., 2017, Nag and Pal, 2015, Li et al., 2016). The feature selection process can reduce the data dimensions and computational burden, and further improve the performance and accuracy of ML algorithms, as well as facilitate the interpretation and rationalisation of models (Zhou, Zhang, Zhang and Liu, 2019). In the field of biomedicine, signature selection can be applied to biological data for cancer detection, drug discovery and medical diagnosis. In high-dimensional biological data, such as gene expression data, irrelevant features may interfere with the true features, introduce heterogeneity in the data, and create dependencies between features. At the same time, if there are dependencies among the features, statistical analyses will no longer play an important role (Khaire and Dhanalakshmi, 2022). Therefore, it is important to select key genes to improve the accuracy of cancer diagnosis and the efficiency of drug discovery.

Existing feature selection methods can be divided into two categories depending on whether they rely on a classifier. Classifier-dependent methods include wrapper and embedded methods, while classifier-independent methods are mainly filter methods.

### 2.3.1 Wrapper methods

Wrapper methods for feature selection are a class of techniques that evaluate subsets of features based on their ability to improve the performance of a predictive model. These methods are highly effective as they directly assess the model's performance during the feature selection process, which allows for the identification of relevant features tailored to the specific learning algorithm. In contrast to filter methods, which evaluate individual features independently of any model, wrapper methods account for feature dependencies and interactions, thus providing potentially better feature subsets.

The process in a wrapper method involves selecting subsets of features and evaluating them by training a model on the selected subset and measuring its perfor-

mance, typically using a criterion like accuracy, F1 score, or another suitable metric. This approach, however, comes with high computational cost due to the need to train and evaluate the model multiple times for different feature subsets. Despite this, the advantage of wrapper methods lies in their ability to consider complex relationships between features and to optimise feature subsets in conjunction with model training, making them highly suitable for tasks where feature interactions play a crucial role.

One of the most well-known wrapper methods is RFE. RFE works by recursively removing features, one at a time, starting from the most irrelevant ones, based on model performance. At each step, the model is trained, and the importance of each feature is determined. Features that contribute the least to the model's performance are removed, and the process is repeated until the optimal feature subset is achieved (Guyon et al., 2002*a*). RFE is particularly popular in SVM and linear regression models, where feature importance is easily computed. However, RFE's efficiency decreases as the dimensionality of the feature set increases, making it computationally expensive for datasets with a large number of features (Saeys et al., 2007).

Forward selection is another widely-used wrapper method. In this method, feature selection begins with an empty feature set and iteratively adds features that lead to the best improvement in model performance, according to a predefined criterion. The process continues until adding new features no longer improves the model's accuracy. Forward selection is often preferred for smaller datasets, as it is more efficient than RFE in such contexts, but it can still suffer from overfitting if the model is overly complex or if the dataset is noisy (Whitley, 1991). It is also susceptible to the "greedy" nature of the search process, which can sometimes lead to suboptimal solutions if the global optimal feature subset is not reached.

Backward elimination, in contrast, begins with all available features and removes the least important features one by one based on their contribution to model performance. While backward elimination can handle more complex models and larger feature sets better than forward selection, it is also computationally demanding due to the necessity of evaluating all subsets at each step (Mandel et al., 2015). Like

forward selection, backward elimination is vulnerable to overfitting and can be less efficient when the number of features is large.

Stepwise selection is a hybrid of forward selection and backward elimination. It adds features to the model one at a time, but it also considers the removal of features that no longer contribute meaningfully to model performance. This hybrid approach helps to balance the advantages of both forward and backward methods, but it is also prone to overfitting and computational inefficiencies, particularly when the dataset is large (Hocking, 2013).

More advanced wrapper methods, such as those using genetic algorithms (GAs), offer a solution to the computational challenges posed by large feature spaces. GAs are inspired by natural evolutionary processes and use mechanisms such as selection, crossover, and mutation to generate new feature subsets. The algorithm evaluates the fitness of each subset based on model performance, selects the best subsets, and iterates through the evolutionary process until convergence. GA-based feature selection has been successfully applied in various domains, including bioinformatics (Vafaie and DeJong, 1995), pattern recognition (Michalewicz, 1995), and ML (Kennedy and Eberhart, 2015). Although GAs can handle large feature sets effectively, they are computationally expensive and require careful tuning to avoid premature convergence and ensure the algorithm does not get trapped in local optima.

Simulated annealing (SA) is another optimisation technique that has been applied to feature selection. SA mimics the process of cooling metal to find the most stable configuration. In feature selection, SA explores the feature space probabilistically, accepting worse solutions with decreasing likelihood as the process continues, in order to escape local minima. This probabilistic approach helps to explore a larger search space, but like genetic algorithms, SA can be computationally intensive (Kirkpatrick et al., 1983). SA has been found useful in high-dimensional feature selection tasks in bioinformatics, where feature interactions are complex and high-quality subsets are critical.

The computational complexity of wrapper methods remains a major drawback, particularly when dealing with high-dimensional datasets where the number of pos-

sible feature subsets grows exponentially. Various heuristic and metaheuristic methods have been proposed to mitigate this issue. Greedy algorithms, for example, make local optimisations, selecting or removing features based on immediate gains in model performance. While greedy methods are efficient, they are susceptible to local minima and may fail to find the optimal subset (Liu et al., 2005).

Moreover, wrapper methods are often prone to overfitting, especially in small datasets. This occurs because wrapper methods optimise feature subsets based on model performance on the training set, which can result in a feature set that is too closely tailored to the training data, thus losing generalisability. To reduce the risk of overfitting, cross-validation techniques are commonly used, where the data is split into multiple training and validation sets to assess model performance. While cross-validation helps ensure that the selected feature subset generalises well to new data, wrapper methods can still suffer from overfitting, particularly when the model is complex or the dataset is noisy (Kohavi, 1995).

Despite these challenges, wrapper methods have been successfully applied in various domains. In bioinformatics, wrapper methods have been used to select gene subsets for the diagnosis of diseases such as AD (Liu, Zhang, Li and Wang, 2018). In medical image analysis, they have been used for feature selection in tasks like tumour detection and organ segmentation (Tosic et al., 2019). In these applications, wrapper methods have been shown to improve the accuracy of predictions by identifying feature subsets that are highly relevant to the problem at hand, often outperforming simpler filter-based methods.

Furthermore, wrapper methods are frequently combined with dimensionality reduction techniques like PCA. PCA reduces the dimensionality of the feature space while retaining the most relevant information. By applying PCA prior to feature selection, wrapper methods can operate more efficiently, as they focus on a reduced set of features, maintaining the computational feasibility of the process while still capturing the essential information (Deng et al., 2016).

In conclusion, wrapper methods are powerful tools for feature selection that can effectively optimise feature subsets based on model performance. Although they can be computationally expensive and prone to overfitting, they are particularly

useful when feature interactions are important and when fine-tuning is required to achieve optimal performance. As computational resources continue to improve and new optimisation techniques are developed, wrapper methods are likely to remain a popular choice for feature selection in complex, high-dimensional datasets.

### 2.3.2 Embedded Methods

Embedded methods are feature selection techniques that perform feature selection during the process of model training. Unlike filter methods, which independently evaluate features, or wrapper methods, which evaluate feature subsets by training a model multiple times, embedded methods integrate feature selection directly into the model training process. This integration allows them to take into account feature interactions, providing an efficient way to identify relevant features while also ensuring that the selected features are optimally suited for the model at hand.

One of the most popular embedded methods is regularisation, which incorporates a penalty term into the objective function of the model to constrain the number of features used. Regularisation techniques such as Lasso (Least Absolute Shrinkage and Selection Operator) and Ridge Regression are widely used for linear models. Lasso, in particular, encourages sparsity by shrinking the coefficients of less important features to zero, effectively performing feature selection (Tibshirani, 1996). The regularisation parameter in Lasso controls the strength of this shrinkage, and by tuning this parameter, it is possible to select the most relevant features. Ridge regression, on the other hand, penalises large coefficients but does not typically shrink them to zero, making it more suitable for situations where all features are expected to contribute to the model (Hoerl and Kennard, 1970).

Elastic Net is a regularisation technique that combines both Lasso and Ridge regression. It was introduced to address the limitations of Lasso when there are highly correlated features in the dataset. While Lasso can perform poorly when features are highly correlated, Elastic Net balances the advantages of both Lasso and Ridge, making it more effective for datasets with complex feature relationships (Zou and Hastie, 2005). By combining the L1 (Lasso) and L2 (Ridge) penalties, Elastic Net can select relevant features while maintaining regularisation.

Another well-known embedded method is decision tree-based algorithms, such as RF and GBM. Decision trees inherently perform feature selection during their construction process. In decision trees, features are selected at each node based on criteria that maximise information gain or minimise impurity. In RF, an ensemble method that aggregates multiple decision trees, feature selection is performed implicitly through the random selection of features at each node. The importance of each feature can be determined by measuring how much it improves the performance of the model, usually using metrics like Gini impurity or information gain (Breiman, 2001). RF provides a robust method for feature selection, particularly when dealing with datasets that have a large number of features.

Gradient Boosting Machines (GBM), another ensemble method, build decision trees sequentially, with each tree attempting to correct the errors of the previous one. GBMs also perform feature selection implicitly by focusing on the most important features in the model's residuals, effectively selecting relevant features over iterations. Feature importance in GBMs is typically measured using metrics like Mean Decrease Impurity (MDI) or Mean Decrease Accuracy (MDA) (Friedman, 2001).

SVM with a linear kernel can also be considered an embedded method for feature selection. In the case of linear SVM, the features are weighted by their importance in defining the optimal hyperplane. Features with low weights contribute little to the model and can be discarded. SVM-based feature selection can be seen as an inherent part of the training process, making it an efficient embedded method (Guyon et al., 2002*a*). When combined with techniques like RFE, SVM can be used to select the most relevant features in high-dimensional datasets, particularly in domains such as bioinformatics and image classification.

Embedded methods offer several advantages over filter and wrapper methods. One key benefit is their efficiency. Because feature selection is performed during the model training process, there is no need to train multiple models or evaluate subsets of features independently, as in wrapper methods. This reduces computational cost and allows embedded methods to scale better to large datasets. Additionally, embedded methods often provide better generalisation by selecting features that are

specifically suited to the learning algorithm, thus reducing the risk of overfitting.

Despite these advantages, embedded methods have their limitations. One challenge is that they are often model-specific, meaning that feature selection is tied to the particular ML algorithm being used. For example, Lasso is only applicable to linear regression models, and decision tree-based methods are more effective for datasets with hierarchical feature interactions. This makes embedded methods less flexible than filter and wrapper methods, which can be applied to any model.

Moreover, embedded methods may not perform well when there are highly correlated features in the dataset. While techniques like Elastic Net and RF can handle correlated features to some extent, other methods may struggle to select a minimal set of features when correlations exist (Zou and Hastie, 2005). In these cases, additional preprocessing, such as dimensionality reduction or feature decorrelation, may be required to improve the performance of embedded methods.

In conclusion, embedded methods are powerful tools for feature selection, offering a balance between efficiency and model-specific optimisation. By selecting features during the model training process, these methods can effectively identify relevant features that improve model performance while reducing computational cost. Regularisation-based techniques like Lasso, Elastic Net, and RR, as well as decision tree-based methods like RF and gradient boosting, are among the most commonly used embedded methods. Although they are model-dependent and can struggle with correlated features, embedded methods remain a popular choice in many ML tasks, especially for high-dimensional datasets where computational efficiency is crucial.

### 2.3.3 Filter Methods

Filter methods are a family of feature selection techniques that evaluate the relevance of features independently of any ML algorithm. Unlike wrapper and embedded methods, which assess feature subsets based on model performance, filter methods rely on statistical techniques to rank and select features before the model is trained. This characteristic makes filter methods computationally efficient, particularly in high-dimensional datasets, as they avoid the need to repeatedly train

and evaluate models on feature subsets. The primary advantage of filter methods is their speed and scalability, as they can handle datasets with large numbers of features without significant computational cost.

Filter methods typically evaluate features using statistical measures of relevance, such as correlation, mutual information, or chi-squared tests. These measures assess the relationship between each feature and the target variable, enabling the identification of features that are strongly associated with the output. Based on these evaluations, features that are deemed irrelevant or redundant are removed, leaving a subset of features that are most likely to improve model performance.

One of the most commonly used filter methods is the Pearson correlation coefficient, which measures the linear relationship between two variables. In the context of feature selection, Pearson correlation is used to identify features that are highly correlated with the target variable. Features that have low correlation with the target are considered less informative and are removed. Pearson correlation can also be used to identify redundant features by evaluating the pairwise correlation between features. Features that are highly correlated with each other are likely to provide redundant information, and one of the correlated features can be removed to reduce dimensionality without sacrificing much information (Hall, 1999).

Another widely used filter method is mutual information (MI), which measures the amount of information that one feature provides about another. MI is a non-parametric method that does not assume any specific distribution for the data, making it suitable for both linear and nonlinear relationships. By calculating the mutual information between each feature and the target variable, MI can identify features that provide the most relevant information for predicting the output. Features that have low mutual information with the target variable can be discarded. MI-based methods are particularly useful when dealing with categorical data or when the relationships between features and the target are complex (Kullback and Leibler, 1959).

The chi-squared test is another statistical measure commonly used in filter-based feature selection, especially when the data is categorical. The chi-squared test evaluates the independence between each feature and the target variable by comparing

the observed frequencies of feature-target combinations with the expected frequencies under the assumption of independence. Features that have a high chi-squared statistic are considered to be significantly related to the target variable and are retained, while features with low chi-squared values are discarded (Duda et al., 2001). The chi-squared test is particularly effective in classification tasks with categorical data, as it identifies features that are most relevant for distinguishing between different classes.

ReliefF is an extension of the Relief algorithm, which is designed to select relevant features by evaluating their ability to distinguish between instances that are near each other in the feature space. Unlike methods like correlation or mutual information, which evaluate the relevance of each feature in isolation, ReliefF takes into account feature interactions by considering the local neighbourhood of instances. The algorithm assigns weights to features based on how well they differentiate between similar instances from different classes. Features with low weights are deemed irrelevant and are discarded. ReliefF is particularly useful in high-dimensional and noisy datasets, where feature interactions are complex (Kira and Rendell, 1992).

Another popular filter method is the Fisher score, which ranks features based on the ratio of between-class variance to within-class variance. The Fisher score is widely used in classification tasks, where it evaluates how well each feature separates the different classes in the target variable. Features with high Fisher scores are those that provide the most information about the class labels, while features with low scores are less informative and can be removed (Fisher, 1936). The Fisher score is simple to compute and can be applied to both continuous and categorical data.

While filter methods are computationally efficient and scalable, they have several limitations. One of the main drawbacks is that they do not consider feature interactions or dependencies. Since filter methods evaluate each feature independently, they may overlook interactions between features that are important for the model. For example, two features that individually have weak relationships with the target variable may, when combined, provide strong predictive power. Filter methods may fail to capture such interactions, leading to suboptimal feature subsets.

Furthermore, filter methods do not always account for the specific characteristics

of the learning algorithm being used. Since feature selection is performed independently of the model, the selected feature subset may not be the most suitable for a particular ML algorithm. This is in contrast to wrapper and embedded methods, which take into account the model's performance when selecting features, ensuring that the selected features are optimally suited for the algorithm.

Despite these limitations, filter methods are often used as a preprocessing step in ML pipelines. They provide a quick and efficient way to reduce the dimensionality of the dataset before applying more computationally intensive methods, such as wrapper or embedded methods. By removing irrelevant or redundant features, filter methods can help improve the overall performance of the model and reduce the risk of overfitting.

In conclusion, filter methods are a fast and efficient approach to feature selection, particularly suited for high-dimensional datasets. They rely on statistical measures of relevance, such as correlation, mutual information, chi-squared tests, and Fisher scores, to identify features that are strongly associated with the target variable. While filter methods have the advantage of being computationally efficient and scalable, they do not consider feature interactions and may not always select the most suitable features for a given learning algorithm. Nevertheless, filter methods remain an important tool in feature selection, particularly when combined with other methods in a hybrid approach.

### 2.3.4 Interpretability in Feature Selection

Interpretability is particularly important in biomedical feature selection because a useful diagnostic model should not only classify samples accurately but also support biological understanding. Model-agnostic explanation methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) can explain individual predictions or estimate feature contributions, but they do not by themselves reduce a high-dimensional gene expression input into a compact, experimentally tractable probe set. Attention mechanisms can also highlight influential inputs in neural models, but their weights are not always equivalent to causal or biological importance. For this reason, Chapter 5 focuses on a shallow sparse AE whose first

hidden layer remains directly connected to input probes, combined with XGBoost feature ranking and enrichment analysis. The one-layer design is less expressive than a deep network, but it improves traceability from latent features back to gene expression probes, which is central to the interpretability contribution of this thesis.

## 2.4 Feature Selection Applications on Gene Expression Data

Gene expression data, typically generated through high-throughput technologies like microarrays and RNA sequencing, contain thousands of features (genes) with relatively few samples. This high-dimensional nature of gene expression data exacerbates the curse of dimensionality, increasing the likelihood of overfitting and reducing generalisation in ML models. Therefore, feature selection is an essential step in the analysis of gene expression data, helping to reduce dimensionality, improve model accuracy, and ensure interpretability. Feature selection methods aim to identify a subset of genes that are most relevant for the target task, such as classification, disease diagnosis, or biomarker discovery.

One of the most prominent applications of feature selection in gene expression data is in disease diagnosis, particularly cancer classification. For example, in breast cancer classification, gene expression data are used to distinguish between benign and malignant tumours or to predict cancer subtypes. Several studies have demonstrated the effectiveness of feature selection in improving the accuracy of classification models. Recent approaches have employed methods like Lasso regression, RF, and SVM to identify key genes that are predictive of cancer outcomes. Feature selection not only improves the model's performance but also helps reduce the complexity of the data, which is particularly beneficial when dealing with large-scale datasets with a high number of features (Xia et al., 2019).

In the context of lung cancer, feature selection methods have been employed to identify biomarkers that can predict patient survival or response to treatment. A study by Yang et al. (2020) used a combination of filter-based methods and ML

classifiers to identify a small set of genes that could effectively classify lung cancer subtypes. The authors found that the selected genes were highly informative for predicting survival, and their model outperformed traditional classifiers in terms of predictive accuracy. Similar studies have demonstrated the potential of feature selection for identifying biomarkers for colorectal cancer, where the ability to select relevant genes has led to the discovery of novel biomarkers for early diagnosis (Wang, Li and Sun, 2021).

Another important application of feature selection in gene expression data is in the field of AD research. AD is a complex neurodegenerative disorder with a multifactorial aetiology, where gene expression data can provide valuable insights into the genetic underpinnings of the disease. However, as with cancer data, the high dimensionality of gene expression data in AD research poses challenges. Recent advancements in feature selection have facilitated the identification of genes that are significantly associated with AD progression and diagnosis. Methods such as Lasso and Elastic Net have been used to identify genes whose expression levels are predictive of cognitive decline in AD patients (Liu et al., 2022). These studies highlight the growing importance of ML and feature selection in precision medicine, enabling more accurate predictions and facilitating early diagnosis of AD.

Recent advancements in ML-based feature selection methods, such as deep learning and ensemble techniques, have also shown promise in gene expression analysis. AEs have been applied to gene expression data for dimensionality reduction and feature selection. By training an AE to reconstruct gene expression data, researchers can identify the most relevant features (genes) that contribute to the data reconstruction. This method allows for the selection of non-linear relationships between genes, which traditional methods like correlation or mutual information may miss (Chen et al., 2021). Additionally, ensemble methods such as RF and GBM have been employed to rank genes based on their importance in classification tasks. These methods have proven particularly useful for selecting genes that interact in complex, non-linear ways, making them valuable for disease classification tasks (Zhou et al., 2023).

A particularly innovative approach has been the integration of feature selection

with multi-omics data, where gene expression data is combined with other types of omics data, such as proteomics or metabolomics. Multi-omics analysis aims to provide a more comprehensive understanding of biological systems by incorporating various types of molecular data. Feature selection techniques, such as RFE and mutual information-based methods, have been successfully applied to identify key features across multiple omics layers. In one study, feature selection techniques were used to integrate gene expression data with DNA methylation and proteomics data to build predictive models for cancer prognosis (Li et al., 2022). By selecting features across different omics layers, the model was able to provide more robust and accurate predictions than those based on a single data source.

Despite the advances in feature selection for gene expression data, several challenges remain. One of the main difficulties is the inherent noise in gene expression data, which can lead to the selection of irrelevant or redundant features. Additionally, the interactions between genes are often complex and non-linear, making it difficult to identify a truly relevant subset of genes using traditional linear methods. Recent research has focused on developing more sophisticated feature selection algorithms, such as hybrid methods that combine filter and wrapper approaches, or deep learning-based methods that can capture non-linear relationships between genes (Zhou et al., 2023).

In conclusion, feature selection plays a crucial role in the analysis of gene expression data, particularly in the context of disease diagnosis, biomarker discovery, and precision medicine. Recent advancements in feature selection techniques, particularly those based on ML and deep learning, have improved the accuracy and efficiency of models built from gene expression data. These methods enable the identification of key biomarkers for cancer, AD, and other complex diseases. As new technologies and algorithms continue to emerge, the role of feature selection in gene expression analysis will become increasingly important in facilitating the discovery of novel biomarkers and improving the precision of disease diagnosis.

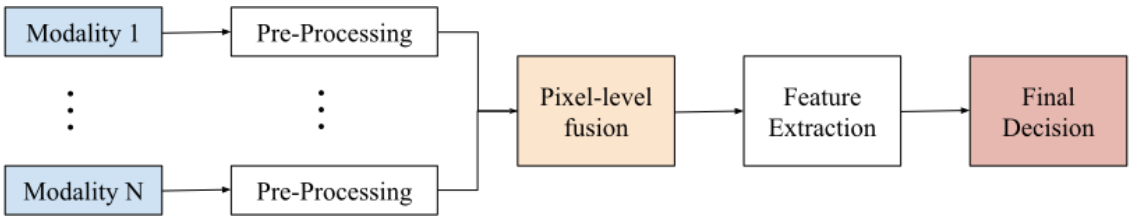



Figure 2.1: Pixel-level data fusion, where raw or minimally processed data from different modalities are fused before feature extraction.

## 2.5 Multimodal Data Fusion for AD Diagnosis

Brain image genetics has brought unprecedented opportunities for in-depth analyses of brain function and pathogenesis of brain diseases, and has greatly facilitated the research of brain diseases. From the perspective of the hierarchy of multimodal data fusion, brain image genetics analysis methods based on multimodal data fusion can be divided into three levels in the diagnosis of brain diseases and identification of pathogenic mechanisms: pixel level fusion, feature level fusion, and decision level fusion.

### 2.5.1 Pixel level fusion

Pixel-level fusion methods (as shown in Figure 2.1) usually directly fuse raw data from different modalities to alleviate inconsistency across modalities, and typical methods include data-driven approaches such as ICA and CCA. For example, Du et al. (Du, Liu, Yao, Risacher, Han, Saykin, Guo and Shen, 2020) proposed a structured sparse CCA (SCCA) method for detecting the association between genetic variants and brain phenotypes of AD patients. This method represents genetic data and brain imaging features as matrices and explores their potential relationships through SCCA, which can reveal underlying pathological mechanisms of AD. Qi et al. (Qi et al., 2022) proposed a risk assessment method combining multimodal brain imaging and PRS, and identified a PRS-associated brain pattern associated with reduced grey matter volume and functional activity in the anterior temporal cortex. Zhang et al. (Wang, Shao, Hao and Zhang, 2021) performed a fusion analysis of brain image and genetic data through an auto-expression network to identify complex imaging genetic patterns. These studies demonstrate the potential of data-

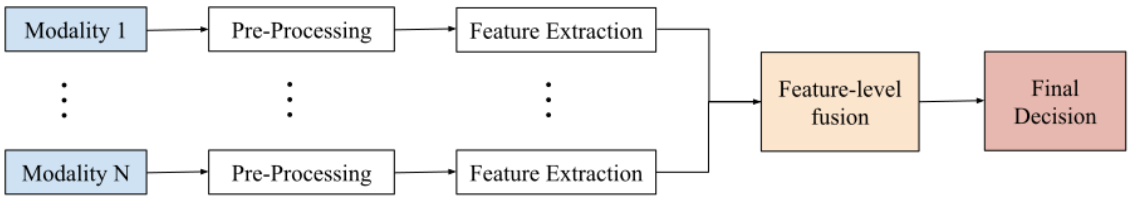



Figure 2.2: Feature-level data fusion, where modality-specific features are extracted or selected before being combined into a joint representation.

layer fusion in the diagnosis and identification of pathogenic mechanisms of brain diseases, but raw-data fusion often requires strong assumptions about feature correspondence and may be sensitive to preprocessing choices.

### 2.5.2 Feature level fusion

Feature-level fusion methods (as shown in Figure 2.2) are usually achieved by feature selection, feature extraction, or feature transformation. These methods can extract common features between different modalities and generate a new joint feature space to represent their correlations. For example, Zhou et al. (Zhou, Liu, Thung and Shen, 2019) proposed a method based on latent co-expression space learning to extract common and complementary features in complete and incomplete multimodal data. Du et al. (Du, Liu, Liu, Yao, Risacher, Han, Saykin and Shen, 2020) proposed a multitask learning method called Dirty Multi-Task Learning (DMTL) to correlate multimodal brain image representations with genetic risk factors. Ghosal et al. (Ghosal, Chen, Pergola, Goldman, Ulrich, Berman, Blasi, Fazio, Rampino, Bertolino et al., 2021) proposed a DNN architecture for combining neuroimaging and genetic data, and Ghosal et al. (Ghosal, Chen, Pergola, Goldman, Ulrich, Weinberger and Venkataraman, 2021) further proposed an end-to-end deep learning framework based on whole-brain and whole-genome fusion. These methods improve diagnostic performance and biomarker identification, but they can still introduce redundant features and may under-represent modality-specific information if the shared feature space is too restrictive.

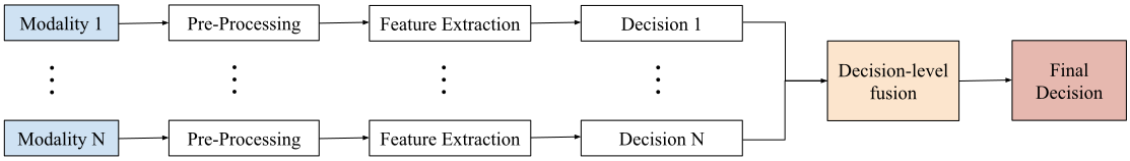



Figure 2.3: Decision-level data fusion, where separate models are trained for each modality and their predictions are combined.

### 2.5.3 Decision level fusion

Decision-level fusion methods (as shown in Figure 2.3) refer to the strategy of using different models for each modality and fusing the resulting decisions. The decision-layer fusion model has an independent semantic space, which provides more flexibility. Lyu et al. (Lyu et al., 2021) proposed a deep learning-based decision-layer fusion model that automatically learns the weighting ratio of image and gene features, allowing the model to analyse brain diseases from different perspectives. The experimental results show that the classification performance of the fusion model is better than that of a single data source. The results suggest that combining information from multiple data sources can improve MCI classification performance, which can help to detect cognitive impairment earlier.

These three fusion strategies have their own advantages and disadvantages. Compared with data- and feature-level fusion, decision-level fusion can be more flexible when modality correlation is low but complementary, and it can be more robust when some modalities are absent. However, decision-layer fusion ignores direct correlations between modality features and cannot provide a comprehensive view of interactions between disease-causing factors in different modalities. Data-layer fusion can perform well when redundancy between modalities is high, but it is often dependent on handcrafted features driven by specific hypotheses. Feature-layer fusion can represent correlations between multimodal data through explicit representation learning, but it often introduces redundant features and may not fully preserve complementarity between modalities.

## 2.6 Metric Learning for Heterogeneous Multimodal Fusion

Metric learning is an ML approach where the goal is to learn a distance metric that reflects the similarity between data points in a feature space. The learned metric can then be used in classification, clustering, and retrieval tasks. It is included here because Chapter 6 uses metric learning as the technical basis for heterogeneous multimodal fusion: rather than simply concatenating gene expression and neuroimaging features, metric learning can align modalities in a task-specific space while preserving discriminative relationships.

### 2.6.1 Metric Learning for Unimodal Data

In unimodal data, the objective of metric learning is to learn a similarity measure that is effective for the given single modality, such as images, text, or gene expression data. The goal is to adjust the feature space so that similar data points are closer together, while dissimilar points are farther apart. Commonly used techniques in metric learning for unimodal data include contrastive loss, triplet loss, and large margin nearest neighbour loss.

One of the earliest and most widely used approaches in metric learning for unimodal data is the Contrastive Loss (Chopra et al., 2005). Contrastive loss is designed for learning embeddings that preserve the similarity between pairs of points. The model is trained to bring similar data points closer and push dissimilar points further apart in the learned feature space. This approach has been applied to various tasks, such as face verification (Schroff et al., 2015), where the goal is to learn an embedding that accurately reflects the similarity between images of the same person.

Another popular method is the Triplet Loss (Schultz and Joachims, 2003), which uses triplets of samples: an anchor, a positive sample (similar to the anchor), and a negative sample (dissimilar to the anchor). The triplet loss function is designed to ensure that the anchor is closer to the positive sample than to the negative sample

by a predefined margin. This technique has been successfully applied in image retrieval and face recognition tasks, where the goal is to learn an embedding space that reflects the similarity between images of the same class.

These methods have also been extended to the analysis of gene expression data, where the goal is to learn a metric that captures the similarities between gene expression profiles associated with disease states. For instance, in cancer classification, metric learning methods have been used to learn a distance metric that improves the accuracy of classifying tumour subtypes based on gene expression data (Xia et al., 2021). By learning an appropriate metric, the model can better differentiate between similar and dissimilar samples, thus improving the performance of downstream classification algorithms.

### 2.6.2 Metric Learning for Multimodal Data

Multimodal data refers to data that comes from multiple sources or modalities, such as combining imaging data (e.g., MRI or PET scans) with genetic data, clinical data, or other forms of biosignals. The goal of metric learning for multimodal data is to learn a shared space where data from different modalities can be directly compared, and the relationships between them can be better understood. This allows the integration of complementary information from multiple sources, leading to more robust models that can provide richer insights.

Learning a distance metric for multimodal data is more complex than for unimodal data because the data from each modality may lie in a different feature space. Approaches for multimodal metric learning often involve joint embedding learning, where the features from different modalities are mapped into a shared space using methods like CCA, DCCA (Andrew et al., 2013), and Multimodal Neural Networks (Ngiam et al., 2011).

DCCA is a popular approach that aims to maximise the correlation between representations of multimodal data in a common space. DCCA is based on neural networks, and it learns the projection of each modality's feature space into a shared space in a way that preserves the correlations between them. This approach has been successfully used in multimodal data integration, such as combining visual and

textual information for multimedia applications (Ngiam et al., 2011), or combining neuroimaging and genetic data in neuroscience.

In multimodal metric learning, the challenge is to learn a shared representation that effectively captures the relationships across modalities, especially when the modalities are heterogeneous or the feature distributions differ significantly. Some approaches, like MDML (Vinyals et al., 2015), have focused on learning a distance metric that aligns the representations from multiple modalities while preserving their individual characteristics.

Multimodal metric learning has been particularly useful in healthcare applications, where data from different sources (e.g., neuroimaging, genetic, and clinical data) must be combined for improved diagnosis and prognosis. For example, the integration of neuroimaging and genomic data for AD classification can benefit from a shared metric space that brings together the complementary information from both modalities (Liu et al., 2019). This approach helps improve the diagnostic accuracy of AD, as it allows for a more comprehensive representation of the patient's condition by combining the genetic and neuroimaging features in a common metric space.

### 2.6.3 Metric Learning for AD Classification

AD is a neurodegenerative disorder that leads to cognitive decline, and early diagnosis is critical for improving patient outcomes. AD is complex, with multiple contributing factors, including genetic, environmental, and neurobiological elements. As a result, there is often a need to integrate data from multiple modalities, such as gene expression, neuroimaging (MRI and PET), and clinical data, to improve diagnostic accuracy.

A promising approach is the use of Deep Metric Learning to integrate functional MRI (fMRI) data with clinical assessments. Liu et al. (2020) used a deep learning-based metric learning method to combine fMRI data and clinical scores for AD classification. By learning an embedding space where both neuroimaging features and clinical scores are aligned, the model was able to improve the predictive power for AD detection, especially in the early stages of the disease, where clinical

symptoms are minimal.

In addition to neuroimaging, gene expression data have been integrated with neuroimaging data for AD classification using metric learning. Wang, Shao, Hao and Zhang (2021) applied a metric learning framework to combine gene expression profiles with sMRI data to classify AD patients from healthy controls. The learned distance metric was able to capture the complex relationships between genetic factors and brain structure, leading to improved diagnostic accuracy.

In conclusion, metric learning has shown significant promise in the classification of AD, especially when combining data from different modalities. The ability to learn a shared distance metric between multimodal data sources enables more accurate and interpretable models that can assist in early diagnosis and prognosis prediction for AD. As the field continues to evolve, the integration of additional modalities, such as blood biomarkers or EEG, may further enhance the power of metric learning models in AD classification.

## 2.7 Review Findings and Research Gaps

This review identifies three research directions for the thesis. First, blood-based gene expression data are attractive for early AD diagnosis because they are less invasive and can reflect molecular processes, but they are high-dimensional, noisy, and usually available with limited sample size. This creates a need for dimensionality reduction methods that can capture non-linear structure without overfitting. Second, feature selection methods for gene expression data must balance predictive performance with biological interpretability. Existing wrapper, embedded, and filter methods each have strengths, but deep learning-based representations often lack direct biological traceability. Third, multimodal AD diagnosis can benefit from integrating gene expression and neuroimaging data, but conventional fusion strategies may either lose modality-specific information or fail to model heterogeneous feature spaces. These gaps motivate the methods developed in Chapters 4, 5, and 6.

## 2.8 Summary

The literature review explores various biomarkers used for the early diagnosis of AD and MCI. These biomarkers include neuropsychological markers like MMSE and CDR, neuroimaging markers such as PET and MRI, and gene expression markers from blood samples. Each type of biomarker has its advantages and limitations in terms of accuracy, invasiveness, and cost.

The chapter also discusses dimensionality reduction techniques for high-dimensional gene expression data, such as PCA, KPCA, ISOMAP, and LLE. These techniques help in reducing the complexity of the data while retaining essential features.

Additionally, the review covers feature selection methods for gene expression data, highlighting the importance of selecting relevant features to improve diagnostic accuracy. Various studies and methods, including ML approaches, are examined to identify the most effective techniques for AD diagnosis.

Finally, the chapter delves into multimodal data fusion for biomedical data, emphasising the need to combine different types of data to enhance diagnostic performance. The review highlights various metric learning algorithms and their applications in multimodal classification, which are crucial for integrating diverse data sources effectively.

# Chapter 3

# Common Methodological Framework

Chapter 2 reviewed the biomarkers and ML methods relevant to early AD diagnosis. This chapter now provides the common methodological foundation used by the three contribution chapters. It summarises the shared theory and techniques, the datasets and preprocessing protocols, and the evaluation metrics used to assess diagnostic performance. The purpose is to avoid repeating background material in Chapters 4–6 and to make the experimental assumptions explicit before the individual studies are presented.

## 3.1 Common Theory and Techniques

### 3.1.1 AEs for Gene Expression Representation

An AE is a neural network trained to reconstruct its input. Given an input vector $x$, the encoder maps it into a latent representation $h$, and the decoder reconstructs $\hat{x}$:

$$h = f(Wx + b), \qquad \hat{x} = g(W'h + b'). \tag{3.1}$$

For gene expression data, where the number of probes is much larger than the number of samples, the latent representation can be used as a lower-dimensional feature set for downstream classification. The reconstruction loss encourages the

latent features to preserve information from the original expression profile, while the reduced dimensionality helps mitigate overfitting.

This thesis uses sparse and denoising variants of the AE. A sparse AE adds a penalty that encourages most hidden units to remain inactive, forcing the model to identify a compact set of informative patterns. A DAE corrupts the input and learns to reconstruct the original version, encouraging robust features. Stacked AEs combine multiple hidden layers to learn hierarchical representations; this is useful for classification performance, but deeper layers are harder to map back to individual probes. For this reason, Chapter 4 evaluates stacked architectures for dimensionality reduction, while Chapter 5 uses a shallow sparse architecture to preserve interpretability.

### 3.1.2 Feature Ranking and RFE

Feature selection is used to reduce a large feature set to a smaller subset that is informative for classification and easier to interpret. RFE repeatedly trains a model, ranks features by importance, removes the least important features, and retrains until a target subset is reached. This strategy can be effective for small-sample biomedical data, but standard RFE is computationally expensive when tens of thousands of probes are present.

Chapter 5 therefore uses XGBoost to rank features and a faster RFE-style selection procedure to reduce the hidden-node and probe sets. XGBoost is suitable for this role because it handles structured tabular data, includes regularisation, and provides gain-based feature importance. Gain measures the improvement in the model objective attributable to a feature split, which makes it useful for identifying features that contribute strongly to classification.

### 3.1.3 Metric Learning for Heterogeneous Fusion

Metric learning aims to learn a task-specific distance function so that samples from the same class are closer together and samples from different classes are farther apart. For heterogeneous multimodal data, this is useful because gene expression

and neuroimaging features differ in scale, dimensionality, and biological meaning. Rather than simply concatenating the modalities, the method in Chapter 6 learns modality-specific transformations and a shared metric space in which complementary information can be preserved.

In this thesis, MML is used to fuse gene expression features with neuroimaging features and to classify pMCI and sMCI. Here, pMCI denotes MCI cases that progress to AD during follow-up, while sMCI denotes MCI cases that remain stable during follow-up.

## 3.2 Datasets and Preprocessing

### 3.2.1 ADNI Gene Expression Dataset

The primary gene expression dataset used in Chapters 4 and 5 was obtained from ADNI. ADNI was established to develop clinical, imaging, genetic, and biochemical biomarkers for early detection and tracking of AD. The gene expression data were obtained from peripheral blood and measured using Affymetrix Human Genome U219 arrays. The chapter-specific experiments use 744 ADNI samples: 246 NC, 382 MCI samples, and 116 AD samples.

The gene expression profiles were normalised using RMA, which performs background correction, quantile normalisation, and summarisation for microarray probe intensities. Chapter 4 uses the ADNI gene expression data to evaluate dimensionality reduction and classification. Chapter 5 further harmonises the probe set with ANM datasets for cross-dataset validation.

### 3.2.2 ADNI, ANM1, and ANM2 Cross-Dataset Validation

Chapter 5 evaluates generalisability using ADNI, ANM1, and ANM2. ADNI uses Affymetrix Human Genome U219 arrays, whereas ANM1 and ANM2 use Illumina HumanHT-12 Expression BeadChip platforms. Because these platforms do not contain identical probes, the preprocessing retains probes that can be matched across the three datasets. This harmonised probe set is then used for model training

and validation so that performance is not driven by platform-specific probes.

### 3.2.3 Multimodal ADNI Dataset

Chapter 6 uses a multimodal ADNI subset containing complete neuroimaging and genetic information. The dataset contains 263 samples labelled as pMCI or sMCI according to longitudinal disease progression. sMRI, fMRI, PET-derived features, and genetic features are preprocessed before fusion. Neuroimaging features are represented through ROIs, while genetic information is represented through selected genes and SNPs. This multimodal subset supports the evaluation of whether molecular and imaging information provide complementary diagnostic evidence.

## 3.3 Evaluation Protocol

Across the contribution chapters, model evaluation is designed to reduce overfitting and to separate model selection from performance estimation. Where hyperparameters are selected, the selection is carried out within the training portion of the data. Test folds or held-out test sets are excluded from parameter optimisation. For cross-validation experiments, folds are stratified where possible so that diagnostic classes remain approximately balanced across folds.

The 10-fold cross-validation protocol is used to estimate classification performance on small-sample datasets. For each fold, the training split is used for preprocessing steps that require fitting, dimensionality reduction model training, classifier training, and hyperparameter selection. The held-out fold is then transformed using the fitted training parameters and evaluated once. This protocol is especially important for gene expression data, because leakage from test folds during feature selection or normalisation can inflate performance estimates.

## 3.4 Evaluation Metrics

Classification performance is evaluated using ACC, PRE, SEN, SPE, F1 score, and AUC. For binary classification, true positives (TP), true negatives (TN), false

positives (FP), and false negatives (FN) are defined with respect to the positive class. The metrics are:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3.2}$$

$$PRE = \frac{TP}{TP + FP}, \tag{3.3}$$

$$SEN = \frac{TP}{TP + FN}, \tag{3.4}$$

$$SPE = \frac{TN}{TN + FP}, \tag{3.5}$$

$$F1 = 2 \times \frac{PRE \times SEN}{PRE + SEN}. \tag{3.6}$$

SEN is also known as the true positive rate. SPE is the proportion of true negatives correctly identified among all actual negative samples. AUC summarises the ROC curve, which plots SEN against the false positive rate (1 − SPE). For multi-class AD classification, metrics are computed using one-vs-rest comparisons and then averaged across classes.

## 3.5 Summary

This chapter has centralised the common theory, datasets, preprocessing assumptions, and evaluation metrics used in the thesis. Chapter 4 applies this foundation to stacked AE-based dimensionality reduction for gene expression data. Chapter 5 extends the gene expression analysis toward interpretable feature selection and biological validation. Chapter 6 then uses the selected molecular information with neuroimaging features in a heterogeneous multimodal fusion framework.

# Chapter 4

# Stacked AE-Based Dimensionality Reduction for AD Gene Expression Classification

This chapter contributes an evaluation of stacked AE-based dimensionality reduction for high-dimensional, non-linear gene expression data in AD classification. The contribution is not merely an empirical comparison of existing dimensionality reduction methods; it investigates whether sparse and denoising AE representations can preserve diagnostically useful expression patterns better than conventional linear and manifold-learning baselines, and whether the resulting latent dimensions improve downstream classification.

## 4.1 Introduction

The rapid progression of high-throughput technologies in genomics has led to an explosion of gene expression data, providing unprecedented opportunities to uncover the genetic underpinnings of complex diseases such as AD. However, the curse of dimensionality inherent in gene expression data presents significant challenges to traditional ML methods, particularly for classification tasks. This chapter addresses the first research question introduced in Chapter 1 and supported by the methodological background in Section 3.1.1: how can an AE be employed to im-

prove the classification of AD gene expression datasets, especially in the context of high-dimensional and non-linear data?

To tackle this question, we explore the potential of deep learning-based dimensionality reduction techniques, focusing specifically on AEs. Two AE methods are investigated: SDAE and SSAE. These methods are applied to gene expression datasets derived from the ADNI database to extract meaningful patterns associated with AD. The extracted features are subsequently evaluated for their effectiveness in classification tasks.

For comparative purposes, this study also considers one linear dimensionality reduction algorithm, PCA, and three non-linear algorithms: KPCA, ISOMAP, and LLE. These algorithms provide a benchmark for assessing the performance of AEs in feature extraction.

The choice of linear SVM as the classifier is motivated by its robustness and effectiveness in high-dimensional spaces, where it can yield strong classification results with relatively simple hyperparameter tuning. By employing linear SVMs to classify features extracted by different dimensionality reduction methods, this study provides a systematic comparison of their performance, highlighting the advantages of AE-based approaches.

The structure of the present chapter is as follows: Section 4.2 provides a description of the dataset, the implementation of the AEs and the baseline algorithms; Section 4.3 presents the experimental results; and Section 4.4 discusses the findings in the context of existing literature.

## 4.2 Data and Methods

### 4.2.1 Data and preprocessing

The gene expression data used in this chapter were obtained from the ADNI database, described centrally in Section 3.2. ADNI was initiated to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and follow-up of AD. The project recruits volunteers aged 55–90 years from research centres in the

United States and Canada, including cognitively normal older adults, people with MCI, and people with AD.

The ADNI data were collected in three phases. In the first phase (ADNI-1), more than 800 volunteers aged 50-90 years from more than 50 research institutions in the United States and Canada were classified into a healthy control group (NC), a group of patients with MCI, and a group of patients with AD. ADNI-1 followed up the data of each group at a determined time (e.g., every 6 or 12 months), including neuroimaging (MRI, PET), blood and cerebrospinal fluid markers (ApoE, A$\beta$, tau protein), gene expression, and so on. Neuroimaging (MRI, PET), blood and cerebrospinal fluid markers (ApoE, A$\beta$, tau protein), and gene expression data were collected from each group at a defined time (e.g., 6 months or 12 months). In this phase, the transition of MCI patients to AD was objectively counted and the diagnostic features of the transition from MCI to AD were analysed and obtained, including structural changes in MRI, $\beta$-like protein deposition in PET, and abnormalities in biological or genetic markers such as tau and ApoE. ADNI-2 is the third phase of ADNI. In addition to the follow-up of NC and MCI subjects in ADNI-1, 150 NC controls, 250 MCI patients, and 150 AD patients were recruited in ADNI-2, which subdivided MCI subjects into EMCI and advanced MCI. The data collection process of the three phases of the ADNI database is basically the same. Since its establishment, the ADNI database has provided an open database with a large sample size and longitudinal tracking pattern for the study of AD-related diseases, and has provided important data support for the study of early diagnosis of AD, and thousands of colleges and research institutes all over the world have carried out a large number of AD-related researches using its data.

A gene expression dataset consisting of 744 samples was selected from the ADNI database because these participants had complete peripheral blood microarray data and diagnostic labels suitable for the NC versus AD/MCI classification task. Table 4.1 shows the detailed information of the selected data, including 246 NC samples and 498 samples from AD and MCI patients. The gene expression data were obtained from human peripheral blood and measured using Affymetrix U219 gene chips, each containing 49,293 probes. The microarray data were normalised by

| Class | Number of sample | Male/Female | Age | MMSE | CDR |
|---|---|---|---|---|---|
| NC | 246 | 117/129 | 76.20+6.49 | 29.06+1.23 | 0.07+0.28 |
| MCI | 382 | 216/166 | 72.96+7.96 | 28.03+1.69 | 1.45+1.01 |
| AD | 116 | 75/41 | 77.31+7.65 | 21.26+4.42 | 5.78+1.01 |

Table 4.1: ADNI gene expression dataset used for the dimensionality reduction experiments, including diagnostic group sizes and clinical score summaries.

RMA (Irizarry et al., 2003) to ensure comparable probe intensity values.

### 4.2.2 Methods

The main objective of this study was to investigate the performance of different dimensionality reduction algorithms on ADNI gene expression data for the AD classification task. Dimensionality reduction maps data points from the original high-dimensional space to a lower-dimensional space; this mapping may be linear or non-linear. In this study, we used one linear algorithm (PCA), three non-linear algorithms (KPCA, ISOMAP, and LLE), and two deep learning algorithms (SDAE and SSAE) for comparison. We then classified the features obtained after dimensionality reduction using linear SVMs and radial basis function kernel SVMs. For every classifier and dimensionality reduction setting reported in this chapter, we used random stratified 10-fold cross-validation. In each fold, the dimensionality reduction model and SVM classifier were fitted on the training split only. Hyper-parameters were selected within the training portion of that fold, and the held-out fold was used only for final performance estimation. The detailed pipeline of the experiments is shown in Figure 4.1.

#### Dimensionality reduction and classification workflow

The theoretical background for PCA, KPCA, ISOMAP, LLE, sparse AEs, and denoising AEs is reviewed in Chapter 2, and the common AE rationale used across the thesis is summarised in Section 3.1.1. This chapter therefore treats these methods as experimental feature extractors rather than repeating their full derivations.

The baseline methods were selected to cover both linear and non-linear dimensionality reduction. PCA provides a linear variance-preserving baseline; KPCA

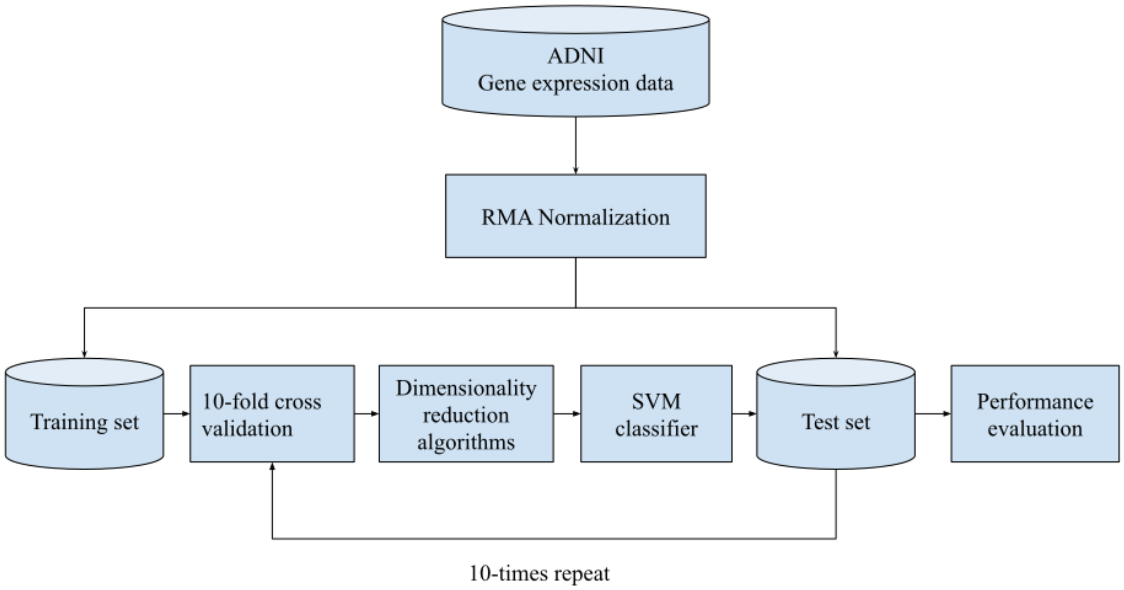



Figure 4.1: Pipeline of the dimensionality reduction experiments. Gene expression data are normalised, reduced using baseline and AE methods, and evaluated using SVM classifiers under cross-validation.

introduces a kernel-based non-linear extension; ISOMAP evaluates global manifold structure; and LLE evaluates local neighbourhood preservation. The proposed SDAE and SSAE are compared against these baselines because they can learn non-linear latent representations from high-dimensional gene expression data. All dimensionality reduction models were fitted within the training split of each cross-validation fold, and the fitted transformations were then applied to the held-out fold.

The reduced features were classified using SVMs rather than a direct DNN classifier. The reason is that the chapter focuses on whether dimensionality reduction improves representation quality in a small-sample, high-dimensional gene expression setting. Training a supervised deep classifier directly on 744 samples with 49,293 probes would substantially increase the risk of overfitting and would make it harder to separate the value of the learned representation from the capacity of the classifier. A linear SVM provides a more controlled downstream classifier for comparing feature representations.

The complete evaluation metric definitions used in the thesis are provided in Section 3.4. In this chapter, the dimensionality sweep tables report mean 10-fold

cross-validation accuracy as the primary summary statistic because the purpose is to compare representation quality across many reduced dimensions and algorithms. PRE, SEN, SPE, F1 score, and AUC are used as complementary diagnostic metrics in the subsequent contribution chapters where full classifier outputs are analysed in more detail.

## 4.3 Experiments

In the benchmark, we applied different dimensionality reduction algorithms to the ADNI gene expression data for dimensionality reduction, and the obtained features were fed into a linear SVM classifier to classify NC and AD&MCI patients.

The input to the SDAE is a vector of elements with values [0,1]. Therefore, before constructing the SDAE, it is necessary to normalise the expression data by probes 0-1.

There are many parameters that need to be adjusted in SDAE, such as batch size, number of training sessions, learning rate, damage level, number of hidden layers, and number of nodes per layer. The selection of parameters for neural networks is a difficult problem, and a commonly used method is to use cross-validation experiments to select relatively better parameters, and at the same time, cross-validation can also prevent the model from overfitting to the training data. In order to obtain the appropriate parameters, 10-fold cross-validation was used to train the SDAE model with various parameter combinations, and the parameter combination that minimised the average reconstruction error across the training folds was selected.

We constructed a three-layer SDAE, and for the selection of parameters in each layer, a layer-by-layer greedy strategy was used to minimise the reconstruction error in each layer. A smaller batch size can obtain more training gradients, so the batch size is set to 10; the learning rate affects the convergence effect, and a larger learning rate is prone to oscillations, so the learning rate is set to 0.001 to ensure convergence; for the SDAE model, the reconstruction error starts to stabilise after 800 trainings, so the number of trainings is set to 1,000; the damage level is set to 0-0.4; the loss level is set to 0.4; in SDAE, the reconstruction error in each layer is set to 1,000; in

| Reconstruction Error | | Corruption Level | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 0.05 | **0.1** | 0.2 | 0.4 |
| Number of Nodes | 50 | 34443.84 | 33257.56 | 33110.55 | 33801.98 | 34055.69 |
| | 500 | 33994.43 | 33483.30 | 32707.64 | 32773.95 | 33116.87 |
| | 2000 | 32573.32 | 32200.38 | 32074.02 | 32057.40 | 32145.62 |
| | **5000** | 31724.02 | 30973.59 | **30575.53** | 31247.60 | 32607.75 |
| | 10000 | 33030.60 | 32461.72 | 31724.02 | 32350.77 | 33030.60 |
| | 20000 | 34086.04 | 33466.92 | 33056.21 | 32855.24 | 33827.02 |

Table 4.2: Average reconstruction error for the first SDAE layer under different hidden-node counts and corruption levels. Corruption level denotes the proportion of input values randomly masked or perturbed during DAE training.

| Reconstruction Error | | Corruption Level | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 0.05 | 0.1 | **0.2** | 0.4 |
| Number of Nodes | 50 | 18.12 | 17.40 | 15.34 | 16.08 | 22.80 |
| | 100 | 16.08 | 13.79 | 11.18 | 11.75 | 18.15 |
| | 200 | 11.54 | 9.90 | 8.77 | 9.63 | 15.75 |
| | **500** | 11.95 | 10.45 | 8.07 | **5.89** | 13.33 |
| | 1000 | 10.54 | 11.13 | 9.59 | 7.79 | 14.16 |
| | 2000 | 13.13 | 10.68 | 9.86 | 7.95 | 14.08 |

Table 4.3: Average reconstruction error for the second SDAE layer under different hidden-node counts and corruption levels.

SDAE, the reconstruction error in each layer is set to 2.5; and the loss level is set to 0.4. In SDAE, since the input nodes of the current layer are the hidden nodes of the previous layer, for the number of nodes, the number of nodes in the first layer should be determined first, and then the nodes of the second layer should be selected on the basis of this. The third layer node selection depends on the number of nodes in the second layer.

Tables 4.2 to 4.4 list the average reconstruction errors of the three hidden layers under each combination, in which the underlined values are the smallest average reconstruction errors and correspond to the optimal number of nodes and damage level. After the 10-fold cross-validation experiments, the number of nodes in the first layer was set to 5000 with a damage level of 0.1; the number of nodes in the second layer was set to 500 with a damage level of 0.2; and the number of nodes in the third layer was set to 50 with a damage level of 0.1.

Tables 4.2–4.4 show how reconstruction error changes with hidden-layer size

| Reconstruction Error | | Corruption Level | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 0.05 | **0.1** | 0.2 | 0.4 |
| Number of Nodes | 20 | 198.02 | 195.30 | 193.29 | 193.32 | 200.14 |
| | **50** | 195.11 | 193.15 | **190.21** | 192.52 | 194.64 |
| | 100 | 196.90 | 195.37 | 192.20 | 194.13 | 198.04 |
| | 200 | 197.22 | 194.98 | 192.92 | 195.30 | 199.17 |

Table 4.4: Average reconstruction error for the third SDAE layer under different hidden-node counts and corruption levels.

| No. of hidden Layers | No. of nodes in each layer | Learning rate | Epochs | Batch size | Sparsity parameter |
|---|---|---|---|---|---|
| 1 | 20000 | 0.001 | 1000 | 10 | 0.0005 |
| 2 | 20000<br>5000 | 0.001 | 1000 | 10 | 0.001 |
| 3 | 20000<br>5000<br>2000 | 0.001 | 1000 | 10 | 0.05 |

Table 4.5: Parameter settings used for the SSAE models with one, two, and three hidden layers.

and corruption level. In this context, lower reconstruction error indicates that the reduced representation preserves more information from the original gene expression profile after denoising. However, reconstruction error alone does not guarantee better classification, because a representation can reconstruct noise or redundant variation. The classification experiments below therefore evaluate whether the reduced features are also discriminative for AD-related labels.

It is worth mentioning that before comparing the results of these dimensionality reduction algorithms, we first experimented with each of the two stacked AEs to find the most appropriate number of hidden layers. The specific parameter settings for the SSAE are shown in Table 4.5, and those for the SDAE are shown in Table 4.6.

Next, we performed dimensionality reduction on the gene expression data using these two types of AE with different numbers of hidden layers, and then identified the optimal number of layers for each AE after classification using a linear SVM. All accuracy values in Table 4.7 and Table 4.8 are mean held-out-fold accuracies from the 10-fold cross-validation protocol described in Section 4.2.

| No. of hidden Layers | No. of nodes in each layer | Learning rate | Epochs | Batch size | Corruption level |
|---|---|---|---|---|---|
| 1 | 20000 | 0.001 | 1000 | 10 | 0.1 |
| 2 | 20000<br>5000 | 0.001 | 1000 | 10 | 0.2 |
| 3 | 20000<br>5000<br>50 | 0.001 | 1000 | 10 | 0.1 |

Table 4.6: Parameter settings used for the SDAE models with one, two, and three hidden layers.

| Dimensionality | 1 hidden layer | 2 hidden layers | 3 hidden layers |
|---|---|---|---|
| 5000 | 0.9154 | 0.9415 | 0.9425 |
| 2000 | 0.9331 | 0.9405 | 0.9413 |
| 500 | 0.8748 | 0.9440 | 0.9485 |
| 50 | 0.8255 | 0.8924 | 0.9245 |

Table 4.7: Mean 10-fold cross-validation classification accuracy of SSAEs with different numbers of hidden layers and reduced dimensions.

| Dimensionality | 1 hidden layer | 2 hidden layers | 3 hidden layers |
|---|---|---|---|
| 5000 | 0.8832 | 0.9341 | 0.9047 |
| 2000 | 0.8415 | 0.8951 | 0.9104 |
| 500 | 0.7941 | 0.9081 | 0.8171 |
| 50 | 0.7605 | 0.8405 | 0.7102 |

Table 4.8: Mean 10-fold cross-validation classification accuracy of SDAEs with different numbers of hidden layers and reduced dimensions.

From the experimental results, with the increase of the number of training layers, the indexes of SDAE gradually decrease, especially the third layer, whose classification accuracy is only 0.71. The reconstruction error of the third layer is much higher than that of the second layer in the model construction, which may result in a larger distortion of the features during the increase of the number of layers, leading to the decrease of the classification effect. During the experiment, we examined the classification results of SDAE three-layer features in depth and found that samples that were false-positive in one classifier were not necessarily false-positive in the other classifiers. This indicates that there are differences in the features extracted from each layer. Deep learning theory suggests that as the depth increases, the deep network extracts features at different levels of abstraction.

The SSAE achieves the highest accuracy in all four dimensionality reduction levels when the number of hidden layers is 3. For the SDAE, the best classification accuracy is achieved when the number of hidden layers is 2, except when reducing dimensionality to 2000. Therefore, in the following experiments we used a 3-layer SSAE and a 2-layer SDAE to compare with other dimensionality reduction algorithms.

## 4.4 Discussions

Table 4.9 and Figure 4.2 demonstrate the performance of six dimensionality reduction algorithms in AD classification tasks under the same 10-fold cross-validation protocol. We also use SVM to directly classify the unreduced data (abbreviated as RAW in the figure and table) as a comparison. The performance of most algorithms is greatly reduced when the dimensionality is reduced to 50. A likely reason is that excessive dimensionality reduction removes information relevant to AD. The only exception is SDAE-2, which maintains a classification accuracy of nearly 90% when dimensionality is reduced to 50, reflecting the algorithm's robustness. The overall performance of PCA is worse than other algorithms in this classification task, while KPCA has good classification accuracy, confirming that the relationship between gene expression data and pathogenesis is non-linear. ISOMAP performs worse in

| Algorithm | 50 Dim | 500 Dim | 1000 Dim | 2000 Dim | 3000 Dim | 4000 Dim | 5000 Dim |
|---|---|---|---|---|---|---|---|
| PCA | 0.7462 | 0.8242 | 0.8354 | 0.8475 | 0.8921 | 0.8722 | 0.8964 |
| KPCA | 0.7954 | 0.9243 | 0.8956 | 0.8943 | 0.897 | 0.9016 | 0.891 |
| ISOMAP | 0.8284 | 0.9347 | 0.9247 | 0.895 | 0.8756 | 0.8594 | 0.8146 |
| LLE | 0.8462 | 0.8563 | 0.915 | 0.9099 | 0.9202 | 0.9129 | 0.9052 |
| SDAE-2 | 0.8405 | 0.9031 | 0.9004 | 0.9132 | 0.9101 | 0.9582 | 0.9341 |
| SSAE-3 | 0.8924 | 0.9329 | 0.9252 | 0.9501 | 0.927 | 0.9453 | 0.9415 |
| RAW | | | | 0.882 | | | |

Table 4.9: Mean 10-fold cross-validation NC versus AD/MCI classification accuracy obtained after applying different dimensionality reduction algorithms across target dimensions. RAW denotes classification without dimensionality reduction.

higher-dimensional classification tasks, which may indicate that ADNI gene expression data do not lie on an internally flat low-dimensional manifold. In contrast, LLE has good overall performance, suggesting that local neighbourhood information is useful for this dataset. Finally, SDAE is not as robust as SSAE across all settings, but when the dimensionality is reduced to 4000 it achieves the highest classification accuracy in this experiment. This suggests that there is a "sweet spot" in dimensionality reduction: too much or too little reduction can lead to performance loss.

## 4.5 Summary

This chapter investigated six dimensionality reduction methods for ADNI gene expression classification, using SVM classifiers to evaluate the discriminative value of the extracted features. We constructed SDAEs and SSAEs to extract AD-related gene expression representations from 246 NC samples and 498 MCI/AD samples. After 10-fold cross-validation experiments, a three-layer SSAE and a two-layer SDAE showed the strongest performance among the AE configurations tested. The results indicate that sparse and denoising AE representations can improve classification over several mainstream dimensionality reduction methods, especially when the reduced dimension is chosen carefully.

These findings also motivate the next chapter. Although stacked AEs improve classification performance, their deeper latent features are difficult to interpret bi-

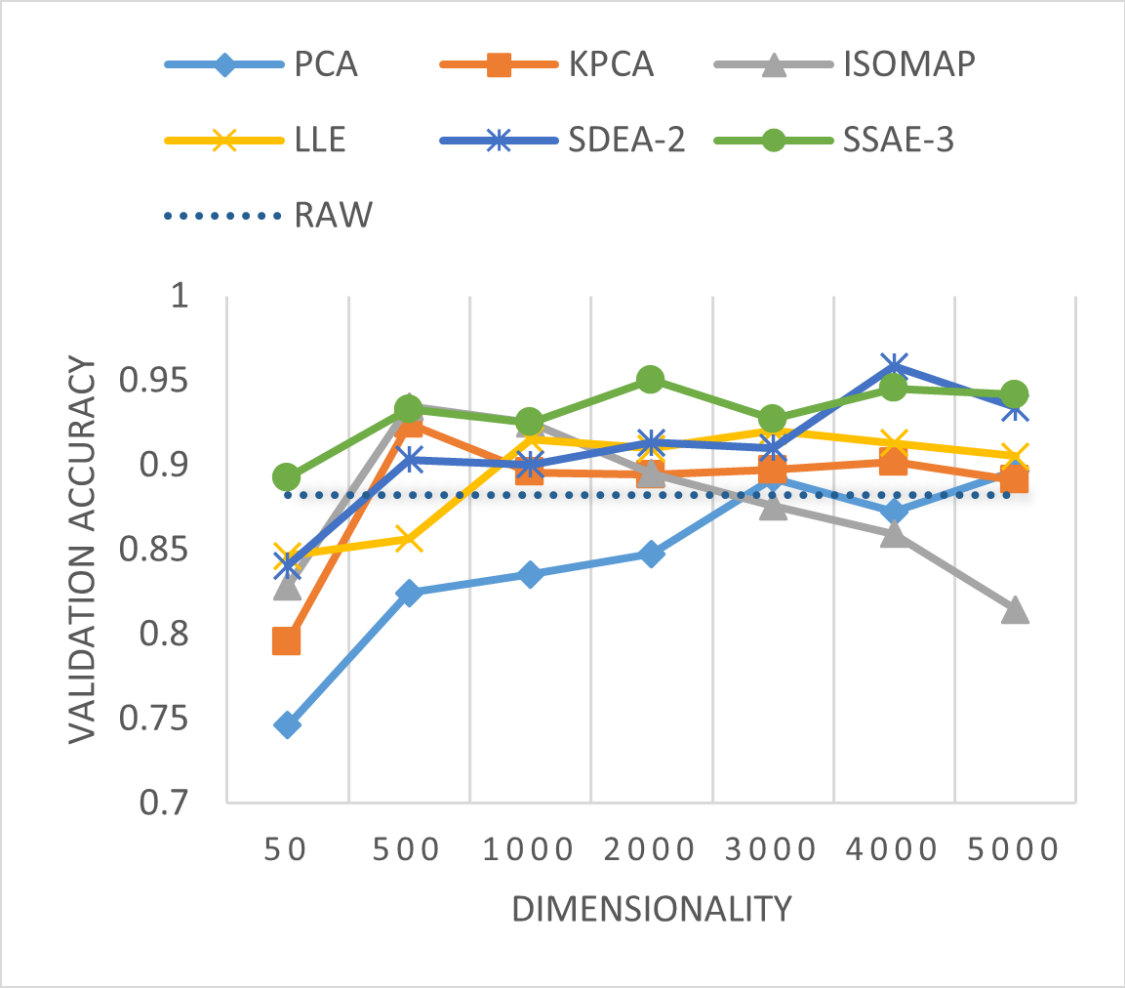



Figure 4.2: Mean 10-fold cross-validation NC versus AD/MCI classification accuracy trends for the dimensionality reduction algorithms compared in Table 4.9.

ologically. Chapter 5 therefore focuses on a shallow sparse AE and feature ranking framework, preserving a direct link between input probes and hidden nodes so that selected features can be analysed through biological enrichment.

# Chapter 5

# Feature Selection and Interpretability Study for Gene Expression Data for Early AD Diagnosis

Chapter 4 examines the feasibility of the proposed AEs for processing gene expression data, and validates their effectiveness in dealing with non-linear high-dimensional data through comparative experiments. The focus of this chapter is to further address the issue of interpretability in the utilisation of gene expression data for AD early diagnosis tasks.

## 5.1 Introduction

Advancements in deep learning have provided powerful tools for feature extraction from complex datasets. However, the interpretability of these features often falls short, particularly in applications such as genomics, where biological insights are paramount. This chapter addresses the second research question introduced in Chapter 1: how can an interpretable feature selection algorithm combine the feature extraction capabilities of deep learning with the interpretability of traditional ML to ensure the selection of biologically meaningful probes in AD?

Building upon the sparse AE framework introduced in Chapter 4, this study proposes a novel feature selection method based on a shallow sparse AE. Sparse AEs are particularly suited for dimensionality reduction in gene expression data due to their ability to enforce sparsity constraints, which encourage the model to identify a smaller subset of informative features. This characteristic makes sparse AEs ideal for addressing the high-dimensional and noisy nature of gene expression data while maintaining interpretability. Unlike deeper layers of AEs, where nodes represent high-dimensional mappings of previous layers, the first layer nodes are directly connected to the input probes, making them more interpretable in the context of biological data. Experimental results from Section 4.3 demonstrate that the classification accuracy of the first layer of the SSAE reaches 96%, indicating its strong discriminative capabilities. Thus, this chapter focuses on analysing the first layer nodes to identify high-contribution features that are effective for classification.

The first layer of the SSAE contains 5000 nodes, a dimensionality that remains too large for meaningful interpretability. To address this, a feature selection process is required to pinpoint the most influential nodes for classification. Additionally, due to variations in probe designs across different gene expression microarrays, reducing the probe set further improves the transferability of selected features to other platforms. Inspired by SVM-RFE, we develop a feature selection algorithm, Fast-RFE, which leverages XGBoost. XGBoost is chosen as the classifier for its ability to handle high-dimensional datasets efficiently, its robustness to noise, and its strong performance in tabular ML tasks. Furthermore, XGBoost provides feature importance metrics such as "gain", which measures the improvement in model objective brought by a particular feature. Gain is suitable for RFE in this study because it captures the contribution of each feature in a hierarchical tree ensemble, aligning well with the goal of selecting biologically meaningful probes.

Alternative explainable AI methods were considered. SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016) can explain model predictions, and attention mechanisms can highlight influential inputs in neural models. However, the aim of this chapter is not only to explain a trained classifier, but also to reduce thousands of gene expression probes to a smaller biologically interpretable set. The shallow

sparse AE plus XGBoost-RFE design therefore provides a more direct route from latent features back to input probes, while enrichment analysis provides biological validation of the selected probe set.

To evaluate the biological relevance of the selected probes, we perform GO and KEGG pathway enrichment analyses, uncovering their potential biological significance in the context of AD.

The structure of this chapter is as follows. Section 5.2 outlines data processing, the proposed methods, and their implementation. Section 5.3 presents the experimental results, comparative analyses, and biological interpretation of the findings.

## 5.2 Data and Methods

The proposed method is divided into four steps: preprocessing the data, reducing the dimensionality of the sparse AE, XGBoost classification, and interpretability analysis. In the data pre-processing, we constructed a dataset based on ADNI by selecting probes common to the ANM1 and ANM2 datasets. The vectorised probe data were then fed into the sparse AE as input for feature selection. At the same time, dimensionality reduction is achieved by limiting the number of nodes in the hidden layer of the sparse AE. Hence, the nodes in the hidden layer of the sparse AE were used as selected features and input to the XGboost classifier, which was trained and performs the classification task. Finally, we performed an interpretability analysis. We first ranked the features in terms of importance using XGBoost, and then filtered out the high-contributing nodes and high-contributing probes, which were used for enrichment analysis to verify the interpretability of the extracted probes. The general framework is shown in Figure 5.1.

### 5.2.1 Data Pre-Processing

The experiments in this study used peripheral blood gene expression data. As described in Section 3.2, we used the ADNI gene expression dataset to train and validate the feature selection and classification model. In addition to ADNI, we used gene expression data from ANM1 and ANM2 to validate model generalisability

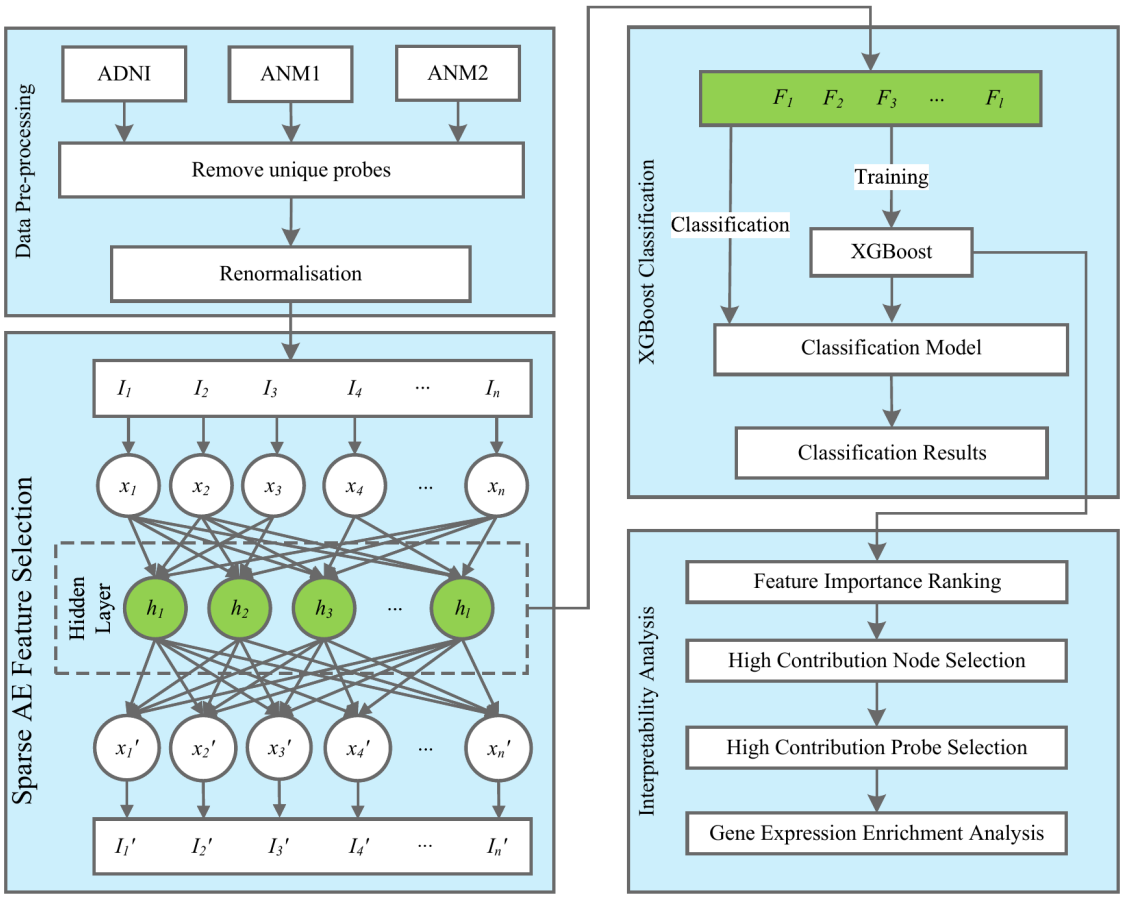



Figure 5.1: Architecture of the proposed Sparse AE-XGBoost model for AD classification. Common probes are selected across datasets, sparse AE hidden nodes are used as features, XGBoost classifies the hidden features, and high-contribution probes are selected for enrichment analysis.

| | ADNI | | | ANM1 | | | ANM2 | | |
|---|---|---|---|---|---|---|---|---|---|
| | NC | AD | MCI | NC | AD | MCI | NC | AD | MCI |
| Number of Cases | 246 | 116 | 382 | 104 | 145 | 80 | 134 | 139 | 109 |
| Gender, % Males | 47.6% | 63.8% | 56.5% | 37.1% | 30.0% | 47.3% | 31.5% | 36.4% | 32.8% |
| Age | 76.2±6.5 | 77.3±7.7 | 72.9±8.0 | 73.7±7.5 | 76.0±6.6 | 74.9±5.3 | 75.7±6.1 | 78.6±5.4 | 77.2±3.2 |
| MMSE | 29.1±1.2 | 21.3±4.4 | 28.1±1.7 | 29.1±1.1 | 20.9±4.7 | 26.8±1.7 | 28.4±1.7 | 19.9±4.6 | 28.1±1.1 |

Table 5.1: Details of participants from ADNI, ANM1, and ANM2 used for cross-dataset validation of the feature selection framework.

across databases. Participants in the three databases were classified using diagnostic labels supported by Mini-Mental State Examination (MMSE) scores (Folstein et al., 1975). In this study we included 744 samples from ADNI (246 NC, 382 MCI and 116 AD), 329 samples from ANM1 (104 NC, 80 MCI and 145 AD), and 382 samples from ANM2 (134 NC, 109 MCI and 139 AD). Details of the dataset are given in Table 5.1.

Regarding the collection chip of gene expression data, ADNI uses Affymetrix Human Genome U219, while ANM1 and ANM2 use Illumina HumanHT-12 Expression BeadChip v3 and v4, respectively. As the genes and probes targeted by the Affymetrix Human Genome U219 and the Illumina HumanHT-12 Expression BeadChip are not identical, we performed data pre-processing on the three datasets to enable controlled experiments between the datasets. The first step was to remove probes that were not common to the three datasets: we removed genes that were unique to each of the three datasets, leaving 14,498 genes, which reduced the number of probes from 49,386 to 38,947 for ADNI, from 48,804 to 29,485 for ANM1, and from 47,231 to 20,177 for ANM2. We selected the probes common to all three datasets, so the final number of probes selected was 16,482. The second step is to normalise the three datasets. Although all three datasets provided normalised probe data, the median RNA expression values for ADNI, ANM1 and ANM2 were calculated to be 3.897, 7.584 and 6.154 respectively. This indicated that the gene expression intensity of ADNI dataset is significantly lower than that of ANM1 and ANM2. To ensure the quality of the cross-dataset experiments by mitigating the influence of batch effects and differences between the datasets, we renormalised the three datasets by performing RMA (Irizarry et al., 2003).

### 5.2.2 Shallow Sparse AE for Dimensionality Reduction

An AE is an unsupervised learning model based on artificial neural networks (Rumelhart et al., 1986), which aims to capture the hidden features of the input data so as to complete the efficient reconstruction of the original data. Its advantage lies in the fact that the AE can learn both linear and non-linear information in the original data, and by using a DNN structure and non-linear activation function for non-linear mapping, it enables the AE to better capture the complex features of the data. In gene expression data, genes are associated with each other in a complex non-linear relationship through their transcription factors, their pathway membership, and other biological properties. Traditional linear dimensionality reduction methods (e.g., PCA) are often used to achieve dimensionality reduction by finding suitable transformation matrices for linear mapping, which often fails to capture the association distribution information of gene expression data. In contrast to PCA, which attempts to discover low-dimensional hyperplanes describing the original data, AEs are capable of learning and describing non-linear manifolds to better handle the association complexity of gene expression data.

The AE network can be divided into an encoding stage and a decoding stage. These two phases can be selected with different activation functions depending on the goal of the task. The process of encoding can be represented as follows:

$$h(x) = s(W_1 x + b_1) \tag{5.1}$$

$x$ and $W_1$ are the input vector and the activation values of the nodes in the hidden layer, respectively. $s$ is the sigmoid, a common activation function for encoders. $W_1$ is the weight matrix. $b_1$ is the bias vector of the hidden layer.

The process of decoding can be represented as follow:

$$x^{'} = s(W_2 x + b_2) \tag{5.2}$$

where $x^{'}$ is the output vector of the network and $W_2$ is the weight matrix connecting the hidden and output layers. $b_2$ is the bias vector of the output layer.

When training AE, a loss function is needed to judge the training results, including a reconstruction error term, i.e., the mean square error between the reconstructed values and the actual values of the input vectors, and an L2 regular penalty term for the weights (to avoid training overfitting) as shown in the following equation:

$$J(w,b) = \frac{1}{N}\sum_{n=1}^{N}(x_n - \widehat{x}_n)^2 + \frac{\lambda}{2}\sum_{l=1}^{L}\sum_{j=1}^{N}\sum_{i=1}^{k}\left(w_{ji}^{(l)}\right)^2 \tag{5.3}$$

Where,$L$, $N$ and $k$ are the number of hidden layers, samples and variables in the training dataset, respectively, and $w_{ji}^{(l)}$ is the weight of the hidden layer.

Sparse AEs are based on the AE by adding a limit on the number of activated neurons in the hidden layer (Ng et al., 2011), thus obtaining stronger feature extraction and noise immunity compared with a general stacked AE by forcing the neural network to focus on the main features in the data and ignoring secondary signals.

Assuming input $x$ is given, $h_j(x)$ denotes the activation value of the hidden neuron $j$, the average activation value of a hidden neuron $j$ is denoted by:

$$\widehat{\rho_i} = \frac{1}{n}\sum_{j=1}^{n} h_i(x_j) \tag{5.4}$$

where $n$ is the number of samples in the training dataset, $x_j$ is the $j$-th training data sample. This function of the average output activation value of a neuron is denoted as the regularisation factor. The sparsity of the AE is improved by adding this regularisation factor to the loss function of the network structure thereby creating the SSAE. The sparsity regularisation term is defined as:

$$\sum_{j=1}^{D^{(1)}} KL(\rho \parallel \hat{\rho}_j) = \sum_{j=1}^{D^{(1)}}(\rho \log\frac{\rho}{\hat{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j}) \tag{5.5}$$

where $KL$ scatter is a measure of the difference between the two distributions. $D^{(1)}$ denotes the number of neurons in the hidden layer, and the above equation shows that when $\rho$ and $\hat{\rho}_j$ are close to each other, the sparsity penalty is close to 0, and when $\rho$ and $\hat{\rho}_j$ deviate from each other, the sparsity penalty becomes larger. In

the process of minimising the loss function, the sparsity regularisation term will be very small, so $\rho$ and $\hat{\rho}_j$ will be close to each other eventually. Bringing the sparsity regularisation term into the loss function $J(w, b)$ in Eq. 5.3 yields the loss function for the sparse AE:

$$J_{SAE}(w, b) = J(w, b) + \beta \sum_{i=1}^{D^{(1)}} KL(\rho \parallel \hat{\rho}_i) \tag{5.6}$$

where $\beta$ controls the weighting of sparsity penalty factors. The sparsity regularisation term can make the model simpler and have better generalisation performance. When the average activation value $\hat{\rho}_i$ of a neuron $i$ is not close to its expected value $\rho$, the value of the sparsity regularisation term in the loss function will increase and the goal of sparsity will be achieved.

By constructing a stacked AE with multiple hidden layers, a network model with deep structure can be obtained, and more representative high-level features can be extracted from the original data to achieve better training results (Romeu et al., 2015, Vincent et al., 2010). However, since only the first hidden layer is directly connected to the original probe, in order to carry out interpretable analysis, this study uses a shallow sparse AE with only one hidden layer for feature extraction and dimensionality reduction.

### 5.2.3 XGBoost for Multi-Class Classification

XGBoost is an ML system based on boosted trees proposed by Chen and Guestrin (2016) based on the work of gradient boosting algorithm (GBDT). The algorithm consists of a collection of iterative residual trees, i.e., the Nth decision tree learns the residuals of the previous N-1 trees, and the predicted outputs of each tree are summed up to be the final output of the sample. At the same time, the splitting strategy adopted by XGBoost in constructing residual trees can be used to evaluate the importance of features by metrics. It has been proved that XGBoost achieves excellent results compared with other classifiers in classifying small samples and unbalanced data.

Assume that in the sample dataset $D = (x_i, y_i)$, $x_i$ is the feature data of the $i$-th sample, and $y_i$ is the label output value. XGBoost consists of $K$ CART trees, which assign scores to each leaf node, and finally the predicted scores of each CART are summed up to obtain the final total score, which is evaluated by $K$ additive functions as shown in the following equation:

$$\widehat{y}_\iota = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{5.7}$$

$f_k$ denotes the independent tree structure with leaf node weights, and $f_k(x_i)$ denotes the weight value of the i-th sample $x_i$ that falls on a leaf node in the $k$-th tree. $F$ is the overall space of the K trees. To optimise the function objective $Obj(\theta)$ is:

$$Obj(\theta) = \sum_{i}^{n} l(y_i, \widehat{y}_\iota) + \sum_{k}^{K} \Omega(f_k) \tag{5.8}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{5.9}$$

$\iota(\cdot)$ is the differential loss function,which measures the error between the true value and the predicted value of the model. $\Omega(\cdot)$ is the regularisation term, which represents the complexity of each CART tree, $T$ represents the number of leaf nodes in each CART tree, and $w_j$ represents the fraction of each leaf node, and in this way it is used to constrain the objective function to prevent overfitting. $\gamma$ and $\lambda$ are the constants that control the degree of regularisation of the constants. Since the algorithm uses additive training to generate trees one by one, then for round $t$ the predicted value $\widehat{y}_\iota$ and the loss function $Obj(\theta)$ can be expressed as:

$$\widehat{y}_\iota^{(t)} = \Sigma_{k=1}^{t} f_k(x_\iota) = \widehat{y}_\iota^{(t-1)} + f_\iota(x_\iota) \tag{5.10}$$

$$Obj(\theta)^{(t)} = \Sigma_i^n \, l(y_i, \widehat{y}_l^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{5.11}$$

Gradient boosting decision tree using first order derivative information in optimisation. While XGBoost transforms $\iota(\cdot)$ using second order Taylor's formula for

faster convergence of the objective function. The second order Taylor expansion is given as:

$$f(x + \Delta x) \cong f(x) + f'(x)\Delta x + \tfrac{1}{2}f''(x)\Delta x^2 \tag{5.12}$$

Let $f_t(x_i)$ be $\Delta x$ in Taylor's formula, and a Taylor second order expansion of the loss function $Obj(\theta)^{(t)}$ has:

$$\begin{aligned} Obj(\theta)^{(t)} \cong \Sigma_i^n (l(y_i, \widehat{y}_i^{(t-1)}) + g_i f_t(x_i) \\ + h_i f_t^2(x_i)) + \Omega(f_t) \end{aligned} \tag{5.13}$$

$g_i$and $h_i$ denote the first and second order derivatives of $l(y_i, \widehat{y}_i^{(t-1)})$ with respect to $\widehat{y}_i^{(t-1)}$. The function $l(y_i, \widehat{y}_i^{(t-1)})$ in round $t$ can be treated as a constant term. Hence, substituting Eq. 5.9 into Eq. 5.13, the following equation is obtained:

$$\begin{aligned} Obj(\theta)^{(t)} = \Sigma_i^n \left(g_i f_t(x_i) + h_i f_t^2(x_i)\right) \\ + \ \gamma T + \frac{1}{2}\lambda \Sigma_{j=1}^T w_j^2 \end{aligned} \tag{5.14}$$

Let $w \in R^T$ , $w$ be the sequence of weights of the leaf nodes, $q : R^d \to \{1, 2, ...T\}$, $q$ be the tree structure. Therefore $q(x)$ is denoted as the position of the sample $x$ falling in the leaf node. $f_t(x_i)$ can be expressed by the following equation:

$$f_t(x_i) = w_{q(x)}, w \in R^T, q : R^d \to \{1, 2, ...T\} \tag{5.15}$$

Convert the loss function for traversing the sample data to a loss function for traversing the leaf nodes:

$$Obj(\theta)^{(t)} = \Sigma_{j=1}^T \left(G_j w_j + \tfrac{1}{2}(H_j + \lambda)w_j^2\right) + \gamma T \tag{5.16}$$

$I_j$ is the set of samples belonging to the leaf node $j$,and subsequently its derivative on $Obj(\theta)^{(t)}$ yields the extreme points with extreme values of:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{5.17}$$

$$Obj^* = -\frac{1}{2}\Sigma_{j=1}^{T}\frac{{c_j}^2}{H_j + \lambda} + \gamma T \tag{5.18}$$

XGboost uses Eqs. 5.19 to evaluate whether or not a node splits, and ultimately determine the structure of the tree:

$$\begin{aligned} Gain = \frac{1}{2}\left[\frac{{G_L}^2}{H_L + \lambda} + \frac{{G_R}^2}{H_R + \lambda} - \frac{(G_R + G_L)^2}{H_R + H_L + \lambda}\right] \\ - \gamma \end{aligned} \tag{5.19}$$

The number of times a feature acts as a split node throughout the construction of the model is *weight* and the average gain of the feature as a split node is *gain*:

$$gain = \sum_{Gain} /weight \tag{5.20}$$

### 5.2.4 Fast RFE for Feature Selection

Medical tasks place more emphasis on interpretability than traditional ML tasks. This requires not only generating classification results, but also interpreting these results in a biologically meaningful way. Since the sparse AE has only the first layer of nodes directly connected to the probes, and the deeper nodes represent a high-dimensional mapping of the upper features, which poses a challenge in recognising the relationship between the representations and the probes, the dimensionality of the features was not reduced in this study by using a stacked AE. Nonetheless, the dimensionality of the features after dimensionality reduction using a sparse AE with a single hidden layer, while allowing the classifier to achieve optimal classification, is still too high for interpretable biological analysis. Therefore, to achieve interpretable analyses, we need to perform further feature selection on these features.

One advantage of using the gradient boosting algorithm is that after the boosting tree has been created, the importance score of each feature can be obtained for effective feature selection. The importance score measures the value of features in boosting decision tree construction. The more often a feature is used for node segmentation in the model, the higher its relative importance. Feature importance is

obtained by calculating and ranking each feature in the sample dataset. In decision-tree models, split-point quality is commonly measured using impurity criteria such as the Gini index (Breiman, 2001). The larger a feature's performance measure for split-point improvement, i.e., the closer it is to the root node, the larger its importance weight is. Finally, the results of a feature in all boosting trees are weighted, summed, and averaged to obtain the importance score.

Inspired by the RFE feature selection algorithm proposed by Guyon et al. (2002*b*), we proposed an improved version of RFE optimised for high-dimensional data called Fast-RFE. RFE is based on the importance of model features, eliminating a number of low-ranked features at each iteration based on the feature importance ranking, and using the dataset containing the retained features as the training samples for the next round until the features are reduced to a certain dimension. RFE is effective for small-sample classification tasks, but its computational complexity and costs are high when the feature dimensions are high.

We proposed Fast-RFE to accommodate the high-dimensional nature of the gene expression data classification task. The XGBoost model can be used to rank feature importance by the feature importance metric, and in this work, the normalised weight score of *gain* in Eq. 5.20 is used as the metric for feature importance ranking. The Fast-RFE algorithm, as presented in Algorithm 1, employs a two-phase approach to identify significant features efficiently. Initially, features are sorted based on their gain values, and the mean ($\mu$) and standard deviation ($\sigma$) of these gains are calculated. The algorithm begins with an initial threshold $\delta = \mu$ and iteratively refines it. In the first phase, features with gain exceeding $\delta$ are selected to construct an XGBoost classifier. If the resulting model's accuracy decrease is less than 0.01 compared to the original model, $\delta$ is incremented by $\sigma$. This process continues until a significant accuracy drop is observed, establishing an interval $[a, b]$ for further refinement. The second phase employs a binary search within $[a, b]$ to optimise the feature selection. At each iteration, features with $|w| > \delta$, where $\delta = (a + b)/2$, are chosen to train an XGBoost classifier. Based on the model's performance, either $a$ or $b$ is updated to $\delta$. This binary search persists until the interval $[a, b]$ contains only one feature, at which point features with $|w| > a$ are

selected as the optimal feature subset.

**Algorithm 1** Fast-RFE

Sort features by gain values
Calculate $\mu$ and $\sigma$ of all features' gain values
Set initial threshold $\delta = \mu$
$a \leftarrow 0$, $b \leftarrow \infty$
**while** true **do**
    Select features with gain $> \delta$
    Build XGBoost classifier with selected features
    Calculate classification accuracy
    **if** accuracy drop $< 0.01$ compared to original model **then**
        $a \leftarrow \delta$
        $\delta \leftarrow \delta + \sigma$
    **else**
        $b \leftarrow \delta$
        **break**
    **end if**
**end while**
**while** $a < b$ **do**
    $\delta \leftarrow (a + b)/2$
    Select nodes with $|w| > \delta$
    Train XGBoost classifier with selected features
    Calculate classification accuracy
    **if** accuracy drop $< 0.01$ compared to original model **then**
        $a \leftarrow \delta$
    **else**
        $b \leftarrow \delta$
    **end if**
    **if** only 1 node satisfies $a < |w| < b$ **then**
        Select features with $|w| > a$ to construct the optimal feature subset
        **break**
    **end if**
**end while**

## 5.3 Experimental Results and analysis

To validate the effectiveness of our proposed approach in classifying AD and to verify the biological significance of the extracted features. First, we used feature extraction algorithms and classifiers that have previously been shown to be effective for AD classification in the literature as a control group to demonstrate the effectiveness of our proposed method. The detailed flow of the control experiment is shown

in the Figure 5.2. We introduced feature extraction algorithms and classifiers that have previously been shown to be effective in the literature as a control group to demonstrate the effectiveness of our proposed method. Three feature extraction algorithms are included: PCA, LASSO regression, and DEG analysis. Four classifiers are included: SVM, RF, L1 regularisation Logistic Regression (L1-LR), and DNN. Then, using the feature selection method proposed in Section 5.2.4, we selected the high contributing nodes from the hidden layer of the sparse AE. Further, we filtered out the high contributing gene expression probes from the high contributing nodes. Last but not least, GO biological enrichment analysis and KEGG pathway enrichment analysis uncovered the biological significance of the high contributing probes.

### 5.3.1 Evaluation Metrics for Classification

Classification performance is evaluated using the metrics defined in Section 3.4: ACC, PRE, SEN, SPE, ROC curves, and AUC. SPE is the proportion of true negatives correctly identified among all actual negative samples, not the false positive rate. The ROC curve plots SEN against the false positive rate $(1 - \text{SPE})$. Because this experiment evaluates multi-class classification among AD, MCI, and NC, ROC curves are calculated using one-vs-rest comparisons and then averaged across the three classes.

### 5.3.2 Optimal Hyperparameter Selection for the Models

The experimental environment of this study is: Intel i9-13900k, Nvidia RTX 3090, 64gb Ram, windows 11. In order to obtain appropriate hyperparameters for SSAE, we use 10-fold cross-validation to train each combination of hyperparameters of the model, while avoiding over-training that leads to over-fitting. Finally, the hyperparameter combination that minimises the average reconstruction error is selected. For the XGBoost classifier, we randomly divided the ADNI dataset into training and testing sets with a ratio of 7:3. Then we optimise the model hyperparameters by learning curve and grid search. The model hyperparameters are also adjusted

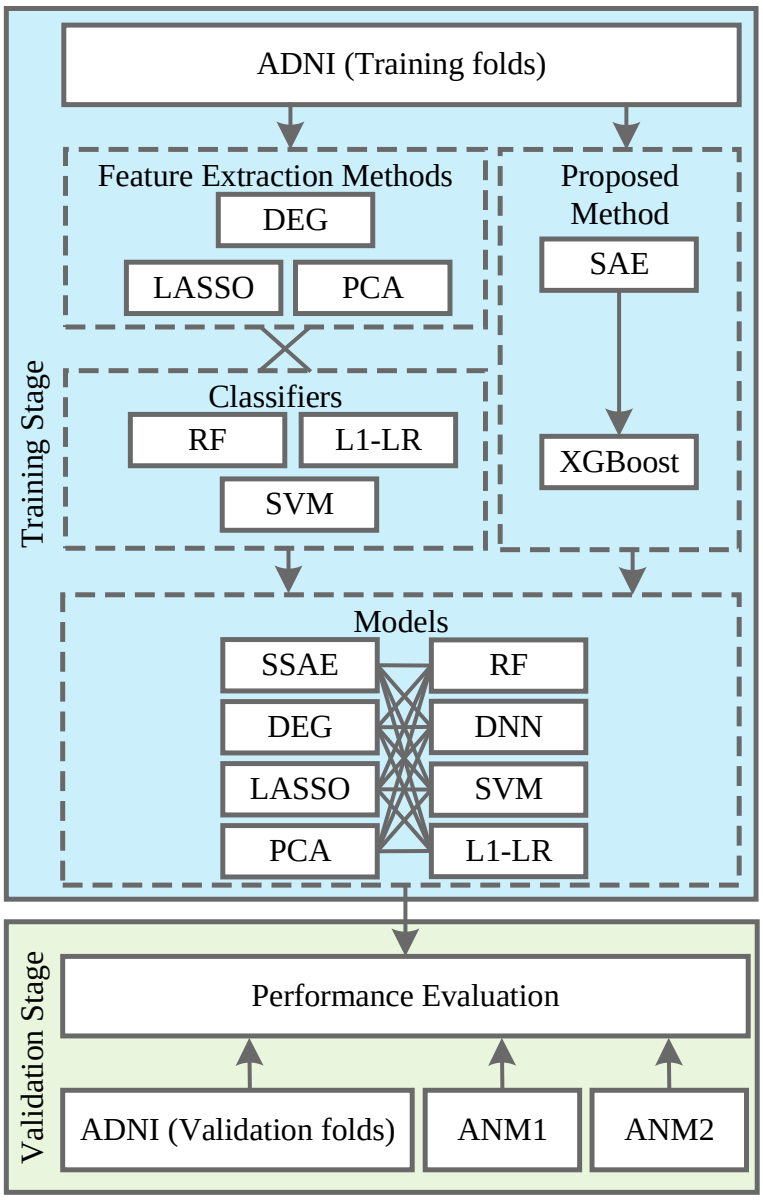



Figure 5.2: The pipeline of the control experiment. We utilise three of the most commonly employed feature selection methods in the field of gene expression research, namely DEG, LASSO, and PCA, in conjunction with four commonly used classifiers, namely RF, SVM, and L1-LR. The various feature selection algorithms were combined with the classifiers one by one in order to train the model, and their performance was compared with that of the proposed algorithms using SSAE as the feature selection method and XGBoost as the classifier. The performance on three datasets of each model was then compared with that of the proposed algorithm using SSAE as the feature selection method and XGBoost as the classifier.

Table 5.2: Parameters of the sparse AE

| Parameter | Value |
|---|---|
| Active Function | Sigmoid |
| Batch Size | 256 |
| Decay | 0.99 |
| Sparsity Parameter | 0.001 |
| Penalty factor | 1.00 |

Table 5.3: Parameters of XGBoost

| Parameter | Value |
|---|---|
| Booster | gbtree |
| eta | 0.003 |
| max depth | 6 |
| gamma | 0.3 |
| objective | multi:softprob |
| subsample | 1 |
| num class | 3 |

to avoid overfitting by 10-fold cross-validation on the training set. The parameter settings for the sparse AE and XGBoost are shown in Table 5.2 and Table 5.3 respectively.

### 5.3.3 Comparison of Different Models

We use the preprocessed ADNI dataset to train SSAE and XGBoost for classification tasks on AD, MCI and NC. The experiments are compared with three dimensionality reduction algorithms and four classifiers commonly used in the gene expression field. Figure 5.3 shows the comparison of Average Area Under the Receiver Operating Characteristic (AUROC) curve. As shown in the figure 5.3, the proposed method outperforms other comparative methods.

### 5.3.4 Analysis of Bioinformatics Interpretability

In order to obtain biologically significant gene expression probes, we firstly used the proposed Fast-RFE algorithm to extract high weight nodes from the high weight features obtained by SSAE, and then we used the algorithm to extract high weight

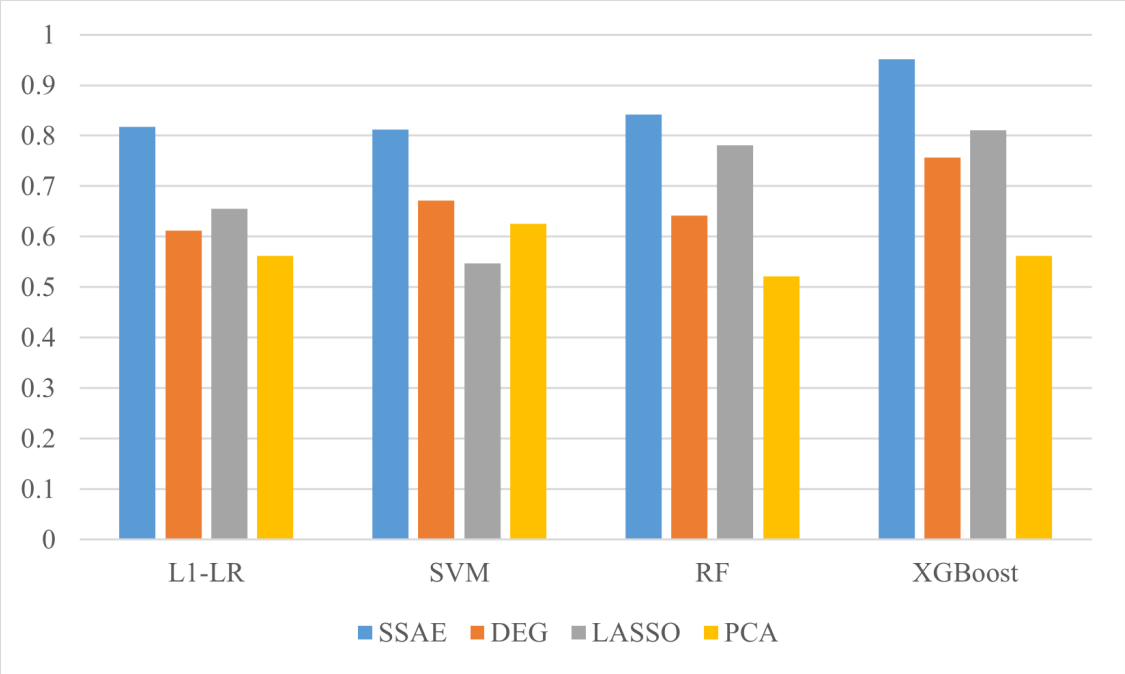



Figure 5.3: Average AUROC comparison between the proposed SSAE-XGBoost model and baseline feature selection and classification combinations.

Table 5.4: Evaluation of model performance using selected hidden nodes and selected gene expression probes for classifying CN, AD, and MCI.

| Features | ACC | SPE | SEN | PRE |
|---|---|---|---|---|
| Nodes | **93.23** | **94.51** | **93.12** | **90.51** |
| Probes | 85.18 | 86.10 | 85.12 | 81.48 |

probes from the high weight nodes.

As shown in Figure 5.4, the gain value of the nodes is heavily clustered around 0, and the contribution of these nodes to the classification is negligible. Further, as shown in Figure 5.5, we plot the weight density curve of the nodes, which shows that the weights of all probes on a single node approximately follow a normal distribution with mean 0. In other words, in a node, the weights of all probes on a single node are distributed in the same way as the weights of all probes on a single node. That is to say, in a node, there are always a large number of probes whose weights are enriched around 0, and the influence of these probes on the value of the node is very small, while the probes distributed at the two ends of the weight density curve affect the value of the node to a great extent, and these probes with larger weights are called high-weight probes.

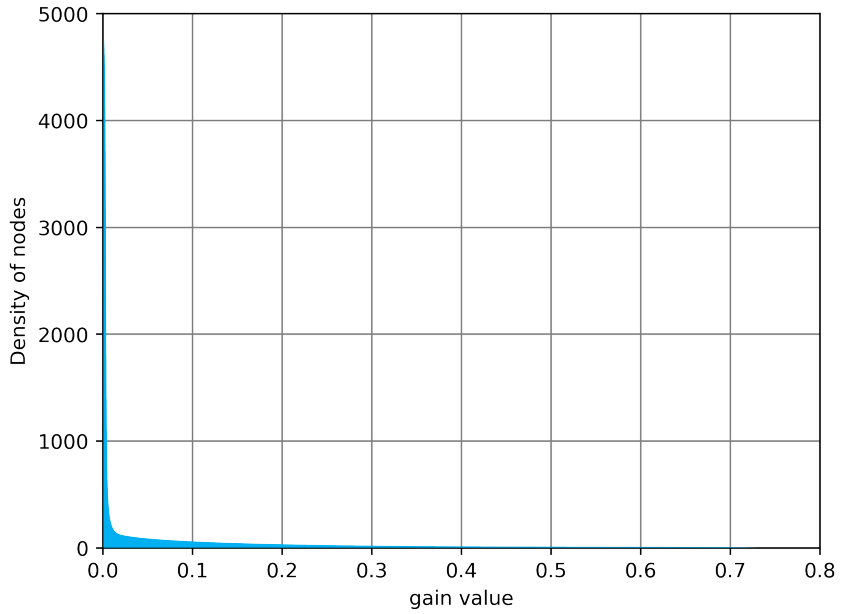



Figure 5.4: Histogram of normalised XGBoost gain scores for sparse AE hidden nodes, using discrete bins to show the concentration of low-contribution features.

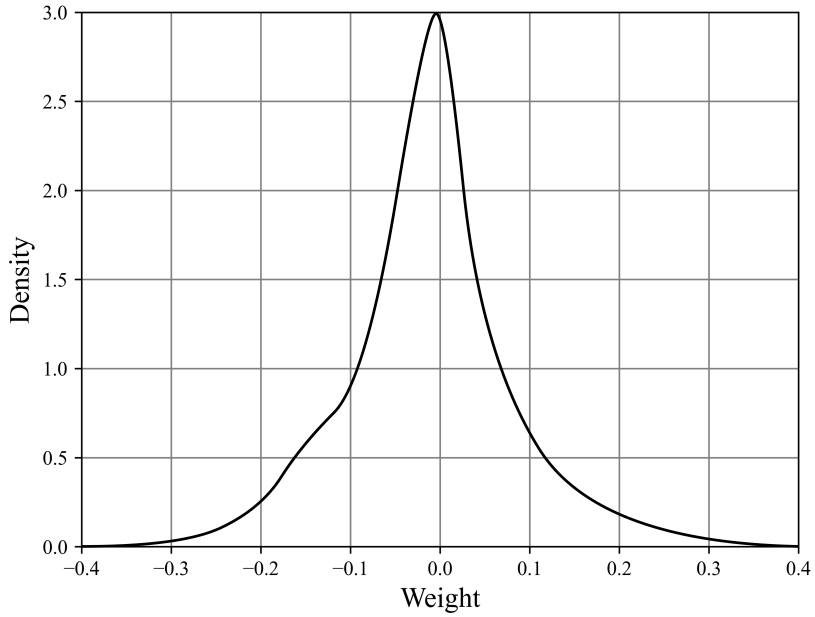



Figure 5.5: Distribution of probe weights inside node 2, illustrating how high-weight probes are selected from the tails of the node-weight distribution.

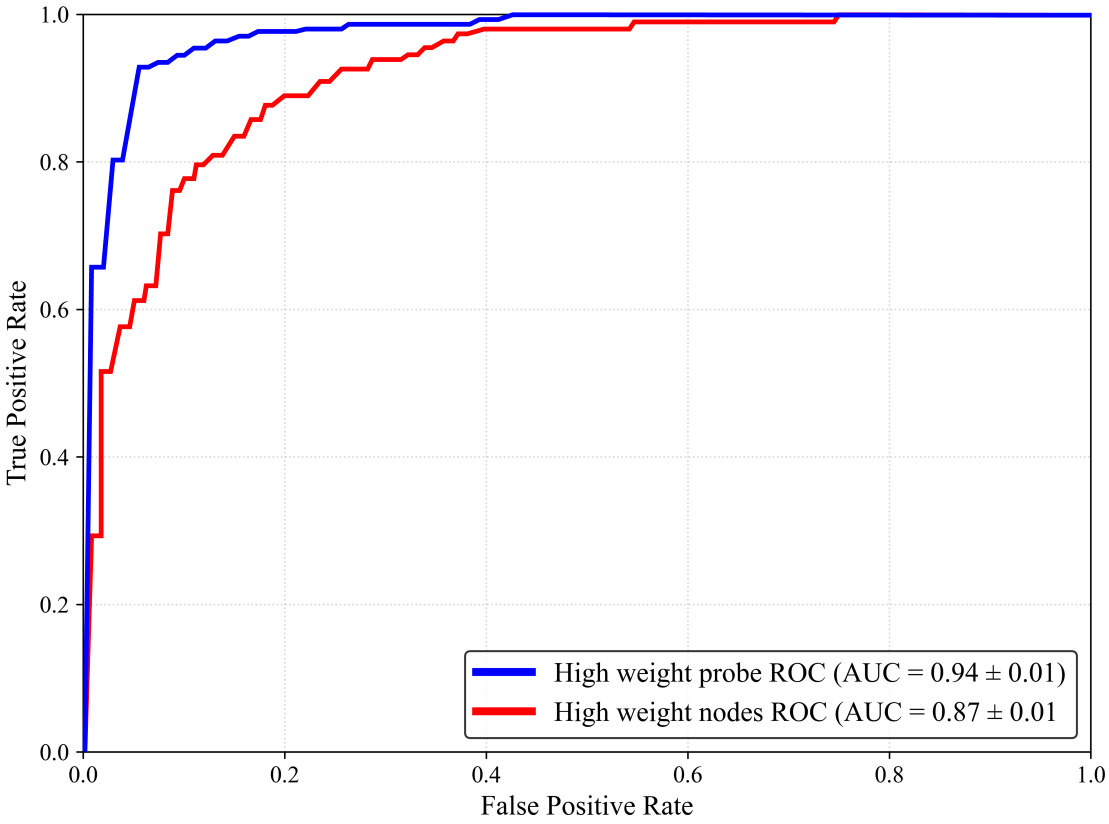



Figure 5.6: ROC curves comparing classification performance using high-contribution hidden nodes and reconstructed high-contribution probes.

As shown in Figure 5.5, we plotted the weight density curve of a node, which shows that the weights of all probes on a single node approximately follow a normal distribution with mean 0 . That is to say, in a node, there are always a large number of probes whose weights are enriched around 0, and the influence of these probes on the value of the node is very small, while the probes distributed at the two ends of the weight density curve affect the value of the node to a large extent, and these probes with larger weights are called high-weight probes.

For a normal distribution $N(\mu, \sigma^2)$, the probability that the weights fall outside $(\mu - 3\sigma, \mu + 3\sigma)$ is less than three in a thousand. Therefore, we select the nodes whose weights take values outside $(\mu - 3\sigma, \mu + 3\sigma)$ as high-weight probes.

High-weight probes have a great influence on nodes, but whether they are effective for AD classification remains to be verified. We investigated two options: firstly, the expression values of the high-weight probes are not processed in any way and directly used for classification; secondly, the high-weight probes are recombined into high-contribution nodes, and the generated nodes are used for classification. The experimental procedure is as follows:

1. Extracting the expression matrix of 5467 high-weight probes from the raw probe expression values $X_f$;

2. Extract the weights corresponding to the high-weight probes and high-contributing nodes from the weight matrix of the first layer of SSAE, and form a new feature weight matrix $W_f$; from the bias vector of the first layer of SSAE, take out the elements corresponding to the high-contributing nodes, and form a new feature bias vector $b_f$;

3. Calculate the new feature nodes: $Y_f = sigmoid(W_f X_f) + b_f$;

4. Constructing classifiers for $X_f$ and $Y_f$ and comparing their performance.

We finally selected 37 high contributing nodes from 5000 nodes, and then filtered 4790 high weight probes from these 37 high weight nodes. Table 5.5 shows the statistics of high-weight probes in each node. We constructed XGBoost classifiers using high weighted nodes and high weighted probes respectively. Table 5.4 shows

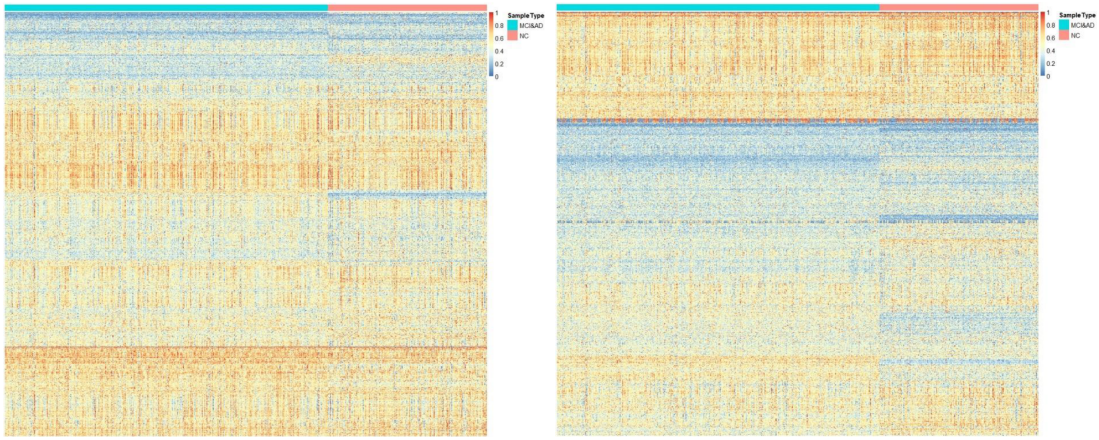
| Average | Standard deviation | Maximum | Minimum | Sum |
|---|---|---|---|---|
| 200 | 82 | 361 | 30 | 4790 |

Table 5.5: Statistical information on high-weight probes

the performance metrics of the two classifiers after cross-validation, and Figure 5.6 plots their ROC curves. Combining the classification performance of Table 5.4 and Figure 5.6, the effect of the classifier constructed by the high weight probe is only slightly decreased compared with that of the high weight node, which indicates that the feature nodes have largely retained the information of the high-contributing nodes, and further proves that the selection of the high weight probe is effective. At the same time, it can be seen that the classifier constructed by the feature nodes is stronger than the high weight probe in every aspect, which may be due to a series of nonlinear transformations performed by the feature nodes on the high weight probe to improve its feature expressiveness, and thus it is more suitable for the AD classification problem. Compared with the original 5000 nodes classification, the accuracy decreases less than 2%, which shows that the feature nodes can represent the original nodes well. In conclusion, from the ADNI dataset alone, the feature nodes we constructed can significantly enhance the AD classification effect.

Taking the intersection of high-weighted probes and 2045 differentially expressed probes with false discovery rate (FDR) below 0.05, only 298 probes existed in both categories. In other words, only 220 of the high-weighted probes were differentially expressed in NC and MCI&AD, suggesting that the pattern represented by the high-weighted probes is very different from that of differential expression. The expression heatmaps of the high-weighted probes and the differentially expressed probes were plotted separately to visually compare their distributions as shown in Figure 5.7, with the heatmaps representing the probes in the horizontal direction and the samples in the vertical direction. The comparison indicates that the expression pattern of the high-weighted probes does not simply reproduce differential expression, so the biological significance of the high-weighted probes is further examined through enrichment analysis.

Firstly, we selected 28 significant AD-related genes from the OMIM and AlzGene



(a) High-weighted probe heat map (b) Differential expression probe heat map

Figure 5.7: Comparison of heat maps for high-weighted probes selected by the proposed framework and differentially expressed probes selected by statistical testing.

databases, and examined their distribution in the high weighted (HW) probes and differential expression (DE) probes. Table 5.6 lists the names and descriptions of these 28 genes, and the genes marked with * indicate that the gene is present in the corresponding probes. Of the 28 AD-related genes, 14 were found in the high-weighted probes, while only 5 were found in the differentially expressed probes.

Comparing the GO enrichment results in Tables 5.7 and 5.8, neither the high-weighted probes nor the differentially expressed probes were directly enriched in the biological processes related to AD. However, from the KEGG enrichment results in Tables 5.9 and 5.10, it can be seen that the high weight probes are significantly enriched in three metabolic pathways, namely AD, Parkinson's disease and Huntington's disease, but no similar pathway enrichment can be seen in the differential expression probes. Figure 5.8 demonstrates the enrichment of high weighted probes in KEGG and GO. From the perspective of metabolic pathways, the high weight probes may be more representative of the expression pattern in AD.

At the same time, DAVID provides KEGG CLUSTER function to cluster the enriched pathways with similar genes. Table 5.11 shows the other pathways related to AD, and we can see that there is a strong correlation between AD, Parkinson's disease and Huntington's disease, which are neurological disorders and may share similar metabolic processes. In addition, we found that NAFLD and oxidative phosphorylation pathways are also strongly associated with AD, which is consistent

| Gene | Description | HW probe | DE probe |
|---|---|---|---|
| A2M | Alpha-2-macroglobulin | | |
| ABCA2 | ATP-binding cassette sub-family A (ABC1) member 2 | | |
| ABCA7 | ATP-binding cassette sub-family A (ABC1) member 7 | | |
| APOE | Apolipoprotein E | | |
| APP | amyloid beta (A4) precursor protein | * | * |
| ATP5G2 | ATP synthase H+ transporting mitochondrial Fo complex subunit C2 (subunit 9) | | |
| BIN1 | Bridging integrator 1 | * | |
| CD2AP | CD2-associated protein | * | |
| CD33 | CD33 molecule | | |
| CLU | [CLU] Clusterin | | |
| CR1 | complement component (3b/4b) receptor 1 (Knops blood group) | * | |
| HFE | Hemochromatosis | | |
| LRP1 | [LRP1] | * | * |
| MPO | Myeloperoxidase | * | |
| MS4A4E | Membrane-spanning 4-domains subfamily A member 4E | | |
| MS4A6A | Membrane-spanning 4-domains subfamily A member 6A | * | |
| NOS3 | nitric oxide synthase 3 (endothelial cell) | | |
| PICALM | Phosphatidylinositol binding clathrin assembly protein | * | * |
| PLAU | Plasminogen activator urokinase | * | |
| PSEN1 | Presenilin 1 | * | * |
| PSEN2 | Presenilin 2 (Alzheimer disease 4) | | |
| SHROOM2 | shroom family member 2 | | |
| SLC9A7 | solute carrier family 9 subfamily A (NHE7 cation proton antiporter 7) member 7 | | |
| TF | Transferrin | * | |
| TNF | Tumor necrosis factor | | |
| VEGF | Vascular endothelial growth factor A | * | * |
| VPS13 | Vacuolar protein sorting 13 homolog A (S. cerevisiae) | * | |
| ZNF673 | Zinc finger protein 673 | * | |

Table 5.6: AD-related genes in the database

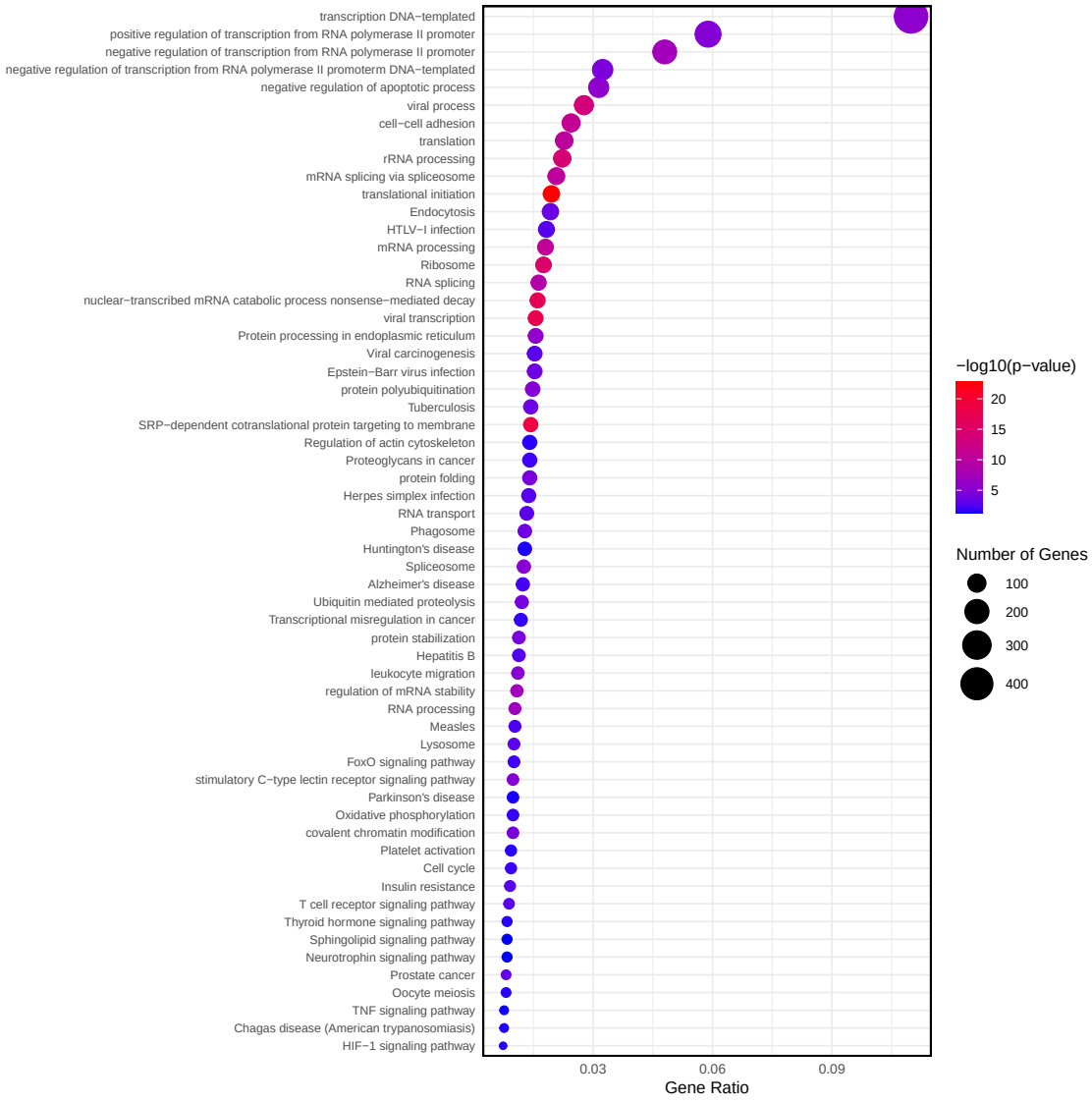
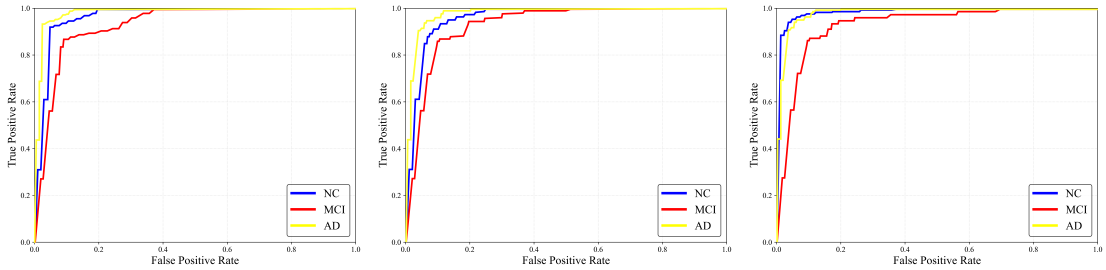



Figure 5.8: Enriched Pathways dot plot. This dot plot illustrates the results of the gene enrichment pathway analysis, with each dot representing an enriched pathway. The horizontal axis displays the gene ratio, calculated as the number of genes in a pathway divided by the total number of genes analysed, where a higher ratio indicates a greater proportion of genes from the dataset present in that pathway. The size of each dot corresponds to the number of genes from the dataset found in the pathway, with larger dots signifying pathways containing more genes. The colour of the dots represents the statistical significance of the enrichment, shown as -log10(p-value), transitioning from blue (less significant) to red (more significant).




(a) ADNI (b) ANM1 (c) ANM2

Figure 5.9: ROC curves for model generalisability evaluation across ADNI, ANM1, and ANM2 datasets.

with some of the current findings. Liver disease has been reported to be a risk factor for cognitive decline in the elderly (Yilmaz and Ozdogan, 2009), and the study by Seo et al. (2016) showed that NAFLD was directly associated with cognitive decline, and that NAFLD accelerated the emergence of AD pathology in a rat model of AD (Kim et al., 2016). From these results, it is clear that high weight probes have better biological performance in AD.

### 5.3.5 Evaluation of Model Generalisability Performance

In order to validate the generalisability of the proposed method and to test whether it can be generalised to other AD gene expression datasets, we carried out this work using the gene expression datasets of ANM1 and ANM2. We first preprocessed all the datasets using the method as described in the previous chapter, and then used SSAE and XGBoost to extract and classify features from the data of ANM1 and ANM2, and plotted the ROC curves respectively. From the results, it can be seen that the classification performance of the proposed method on ANM1 and ANM2 datasets is only slightly degraded compared with ADNI, and the feature nodes have proved to be effective for the AD classification problem, considering that the high weight probes on ANM1 and ANM2 datasets are partially missing.

| GO_Term | Number of Genes | p-value |
|---|---|---|
| Translational initiation | 79 | 1.8E-23 |
| SRP-dependent cotranslational protein targeting to membrane | 58 | 2.4E-19 |
| Viral transcription | 63 | 3.9E-18 |
| Nuclear-transcribed mRNA catabolic process, nonsense-mediated decay | 65 | 8.1E-18 |
| rRNA processing | 90 | 6.8E-15 |
| Viral process | 112 | 4.3E-14 |
| Cell-cell adhesion | 99 | 8.0E-12 |
| mRNA processing | 73 | 1.7E-11 |
| mRNA splicing, via spliceosome | 84 | 4.5E-11 |
| Translation | 92 | 6.2E-11 |
| RNA splicing | 66 | 6.2E-10 |
| Negative regulation of transcription from RNA polymerase II promoter | 194 | 3.8E-8 |
| Regulation of mRNA stability | 44 | 5.5E-8 |
| RNA processing | 42 | 7.3E-8 |
| T cell receptor signaling pathway | 55 | 2.7E-7 |
| Negative regulation of apoptotic process | 127 | 1.5E-6 |
| Transcription, DNA-templated | 444 | 2.2E-6 |
| Leukocyte migration | 45 | 4.7E-6 |
| Stimulatory C-type lectin receptor signaling pathway | 40 | 7.2E-6 |
| Positive regulation of transcription from RNA polymerase II promoter | 238 | 8.4E-6 |
| Protein polyubiquitination | 60 | 9.8E-6 |
| Negative regulation of transcription, DNA-templated | 131 | 2.9E-5 |
| Protein folding | 57 | 4.3E-5 |
| Protein stabilization | 46 | 4.6E-5 |
| Covalent chromatin modification | 40 | 5.0E-5 |

Table 5.7: High weight probe GO-enriched pathway

| **GO_Term** | **Number of Genes** | **P-Value** |
|---|---|---|
| Homophilic cell adhesion via plasma membrane adhesion | 35 | 1.9E-7 |
| Positive regulation of endothelial cell proliferation | 15 | 1.4E-3 |
| Peptidyl-serine phosphorylation | 22 | 1.4E-3 |
| Phosphorylation | 18 | 3.4E-3 |
| Activation of GTPase activity | 15 | 5.1E-3 |
| Ephrin receptor signaling pathway | 15 | 1.1E-2 |
| Negative regulation of transcription from RNA polymerase II promoter | 78 | 1.2E-2 |
| Positive regulation of apoptotic process | 36 | 2.5E-2 |
| Spermatogenesis | 44 | 2.7E-2 |
| Positive regulation of angiogenesis | 17 | 2.8E-2 |
| Peptidyl-tyrosine phosphorylation | 21 | 2.8E-2 |
| Cellular response to DNA damage stimulus | 26 | 3.8E-2 |
| Endocytosis | 19 | 3.8E-2 |
| Single organismal cell-cell adhesion | 15 | 3.9E-2 |
| Transcription from RNA polymerase II promoter | 55 | 4.0E-2 |
| Negative regulation of cell proliferation | 44 | 4.1E-2 |
| Positive regulation of I-kappaB kinase/NF-kappaB signaling | 21 | 4.4E-2 |

Table 5.8: Differentiate expression probe GO-enriched pathway

| **KEGG_Term** | **Number of gene** | **P-Value** |
|---|---|---|
| Ribosome | 71 | 1.1E-15 |
| Protein processing in endoplasmic reticulum | 63 | 1.0E-6 |
| Spliceosome | 51 | 5.0E-6 |
| Ubiquitin mediated proteolysis | 49 | 6.5E-5 |
| Epstein-Barr virus infection | 62 | 1.4E-4 |
| Phagosome | 52 | 1.7E-4 |
| Tuberculosis | 58 | 2.0E-4 |
| Endocytosis | 78 | 2.8E-4 |
| Prostate cancer | 33 | 4.9E-4 |
| Lysosome | 41 | 9.8E-4 |
| RNA transport | 54 | 1.1E-3 |
| T cell receptor signaling pathway | 36 | 1.2E-3 |
| Viral carcinogenesis | 62 | 1.3E-3 |
| Insulin resistance | 37 | 1.5E-3 |
| HTLV-I infection | 74 | 1.6E-3 |
| Herpes simplex infection | 56 | 1.7E-3 |
| Hepatitis B | 46 | 2.1E-3 |
| Measles | 42 | 3.7E-3 |
| **Alzheimer's disease** | 50 | 5.6E-3 |
| FoxO signaling pathway | 41 | 7.6E-3 |
| Proteoglycans in cancer | 57 | 8.1E-3 |
| Cell cycle | 38 | 1.0E-2 |
| Oxidative phosphorylation | 40 | 1.1E-2 |
| Transcriptional misregulation in cancer | 48 | 1.5E-2 |
| Oocyte meiosis | 33 | 2.0E-2 |
| Platelet activation | 38 | 2.2E-2 |
| Thyroid hormone signaling pathway | 34 | 2.3E-2 |
| HIF-1 signaling pathway | 30 | 2.4E-2 |
| Regulation of actin cytoskeleton | 57 | 2.4E-2 |
| **Huntington's disease** | 52 | 3.0E-2 |
| Chagas disease (American trypanosomiasis) | 31 | 3.1E-2 |
| **Parkinson's disease** | 40 | 3.2E-2 |
| TNF signaling pathway | 31 | 3.9E-2 |
| Neurotrophin signaling pathway | 34 | 4.6E-2 |
| Sphingolipid signaling pathway | 34 | 4.6E-2 |

Table 5.9: High weight probe KEGG-enriched pathway

| KEGG_Term | Number of gene | P-Value |
|---|---|---|
| Proteoglycans in cancer | 31 | 1.5E-3 |
| FoxO signaling pathway | 23 | 1.9E-3 |
| Toxoplasmosis | 21 | 2.1E-3 |
| Leishmaniasis | 15 | 2.3E-3 |
| Dilated cardiomyopathy | 16 | 4.4E-3 |
| Hypertrophic cardiomyopathy (HCM) | 15 | 5.5E-3 |
| Prostate cancer | 16 | 6.8E-3 |
| TNF signaling pathway | 18 | 7.7E-3 |
| Chagas disease (American trypanosomiasis) | 17 | 1.4E-2 |
| Pathways in cancer | 47 | 1.5E-2 |
| Hepatitis B | 21 | 2.1E-2 |
| Oxytocin signaling pathway | 22 | 2.6E-2 |
| Inflammatory mediator regulation of TRP channels | 15 | 3.7E-2 |
| Dopaminergic synapse | 18 | 4.3E-2 |
| Tuberculosis | 23 | 4.5E-2 |
| Sphingolipid signaling pathway | 17 | 4.7E-2 |

Table 5.10: Differentiate expression probe KEGG-enriched pathway

| Term | Correlation |
|---|---|
| Alzheimer's disease | 1.00 |
| Parkinson's disease | 0.62 |
| Non-alcoholic fatty liver disease (NAFLD) | 0.59 |
| Huntington's disease | 0.58 |
| Oxidative phosphorylation | 0.57 |
| Mitochondrion inner membrane | 0.42 |
| Electron transport | 0.36 |
| Respiratory chain | 0.35 |
| Cardiac muscle contraction | 0.32 |
| Hydrogen ion transmembrane transport | 0.31 |

Table 5.11: AD-associated pathways in high weight probes

## 5.4 Summary

In this study, we proposed an interpretable deep learning approach for the diagnosis of AD using gene expression data. By integrating a sparse AE for dimensionality reduction and XGBoost for classification, our method effectively addresses the challenges posed by high-dimensional gene expression data. The experimental results on the ADNI dataset, as well as the validation on ANM1 and ANM2 datasets, demonstrate the robustness and generalisability of our approach. The enrichment analysis further confirms the biological relevance of the extracted features, providing valuable insights into the underlying mechanisms of AD.

Our method not only achieves high diagnostic accuracy but also ensures interpretability, making it a promising tool for early diagnosis and intervention in AD. The ability to interpret the model's decisions is crucial in medical applications, as it allows clinicians to understand the underlying reasons for a diagnosis, thereby increasing trust in the model's predictions. This interpretability is achieved through the use of enrichment analysis, which helps in identifying the biological pathways and processes associated with the features extracted by the sparse AE. These insights can potentially lead to the discovery of new biomarkers and therapeutic targets for AD.

Furthermore, the integration of deep learning and traditional ML techniques in our approach highlights the importance of combining different methodologies to leverage their respective strengths. The sparse AE effectively reduces the dimensionality of the gene expression data, capturing the most relevant features, while XGBoost provides robust and accurate classification. This combination not only improves the performance of the model but also enhances its interpretability, making it more suitable for clinical applications.

Future work will focus on enhancing the model's performance and exploring its applicability to other neurodegenerative diseases. One potential direction is to incorporate more advanced techniques such as attention mechanisms, which can further improve the interpretability of the model by highlighting the most important features for each prediction. Additionally, integrating multi-omics data, including

proteomics, metabolomics, and imaging data, could provide a more comprehensive view of the disease pathology, leading to more accurate and reliable diagnoses.

The potential for clinical application of our model could revolutionise the approach to AD diagnosis, offering a non-invasive, cost-effective, and accurate method that could be widely adopted in clinical settings. The use of gene expression data for diagnosis is particularly advantageous as it can be obtained through less invasive procedures compared to traditional methods such as brain imaging or CSF analysis. This makes the diagnostic process more accessible and less burdensome for patients.

Moreover, the insights gained from the biological interpretation of the model's features could pave the way for new therapeutic strategies. By understanding the specific biological pathways and processes involved in AD, researchers can develop targeted therapies that address the root causes of the disease. This could lead to more effective treatments and improved patient outcomes, ultimately contributing to a better quality of life for individuals affected by AD.

In conclusion, our study presents a novel and interpretable deep learning approach for the diagnosis of AD using gene expression data. The combination of sparse AE and XGBoost not only improves diagnostic accuracy but also supports interpretability, making it a valuable tool for early diagnosis research. The potential clinical applications and the insights gained from the biological interpretation of the model's features highlight the significance of this approach in advancing AD research. Future work will continue to build on these findings, exploring new techniques and data sources to further enhance model performance and applicability to other neurodegenerative diseases.

The selected probes and their biological interpretation provide the molecular foundation for the multimodal work in the next chapter. Chapter 6 extends the thesis from gene expression alone to heterogeneous fusion, asking whether the molecular information identified here can be integrated with neuroimaging features to improve early AD diagnosis.

# Chapter 6

# Heterogeneous Multimodal Data Fusion for AD Diagnosis

In this chapter, we propose an MML method, which can effectively fuse heterogeneous neuroimaging data and genetic data, and mine their direct relationship, thus achieving effective early diagnosis of AD.

## 6.1 Introduction

In this chapter, we developed an MML method for fusing neuroimaging and gene expression data for early diagnosis of AD. We detail the experimental settings used to evaluate the proposed algorithm for early AD diagnosis. The primary goal of the experiments is to assess the ability of the MML algorithm to classify pMCI from sMCI based on multimodal data, which include both gene expression data and neuroimaging data. In this thesis, pMCI means MCI that progresses to AD during follow-up, whereas sMCI means MCI that remains stable. Correct classification of these stages is critical for early intervention in AD, making this task important for advancing clinical diagnostics.

The experimental design incorporates several traditional ML algorithms – SVM, RR, and RF – which serve as baseline methods for comparison. These methods are widely used in the ML community for classification tasks and offer a solid benchmark for evaluating the proposed MML algorithm. Additionally, we introduce GCNs as

a deep learning approach that can handle non-Euclidean data, which is typical of neuroimaging and genomic datasets. By comparing these methods, we aim to demonstrate the efficacy of MML in effectively fusing multimodal data for AD classification.

The chapter also provides a detailed discussion of the evaluation metrics used to assess the diagnostic performance of each method. These metrics – ACC, PRE, SEN, SPE, AUC, and F1 score – are essential for measuring both the overall classification performance and the ability of the model to distinguish between the two types of MCI, which are key to early AD detection.

Metric learning is selected over attention-based or generic joint-embedding approaches because the central challenge in this chapter is heterogeneity: gene expression and neuroimaging features differ in scale, dimension, and biological interpretation. Attention mechanisms can highlight influential inputs but usually require a shared representation before attention can be applied. Generic joint embeddings can align modalities, but they do not necessarily preserve modality-specific distances or complementary structure. MML is therefore used because it explicitly learns modality-specific projections and a shared distance metric, allowing the model to preserve within-modality structure while aligning information across modalities.

Finally, the experimental settings are structured around several specific analyses: Risk ROI analysis, Risk Gene analysis, ROI-Gene pair analysis, and ROI Connectivity analysis. These analyses aim to extract the most relevant features from both neuroimaging and gene expression data, which are then used to evaluate the performance of the MML algorithm and its ability to provide meaningful multimodal fusion. The results obtained from these experiments are crucial for understanding the potential of MML in improving the early diagnosis of AD through effective data fusion and classification.

| Diagnose | Age | Gender (Female/male) | MMSE | CDR-SB |
|---|---|---|---|---|
| pMCI | 73.01 ± 7.73 | 65 / 63 | 24.42 ± 2.92 | 3.70 ± 1.96 |
| sMCI | 71.91 ± 7.09 | 69 / 66 | 28.10 ± 1.72 | 1.36 ± 0.80 |

Table 6.1: Demographic and clinical summary of the multimodal ADNI subset used for pMCI and sMCI classification.

## 6.2 Methods

### 6.2.1 Dataset and Pre-Processing

The neuroimaging phenotypic data (PET and MRI) and gene expression data used in this work were obtained from the ADNI dataset, as summarised in Section 3.2. This study involved 263 samples, all of which had complete modality data, including MRI, PET, and genetic data. Detailed statistics on these samples are shown in Table 6.1.

The sMRI and fMRI data of the samples were pre-processed using the CAT12 toolbox, FreeSurfer image analysis suite and DPABI software package, respectively, and the VBM, SBM and functional time series of 210 cortical regions were obtained based on the Brainnetome brain template. At the same time, the sequences of SNPs corresponding to the genes were extracted using the PLINK software package.

The gene expression data are the high-weight probes selected in Chapter 5 by utilising the proposed shallow sparse AE. Their relevant genes (85 SNP loci within the APOE gene, including the AD-related risk locis APOEe4 SNP rs429358) were further been used for interpretability analysis.

### 6.2.2 Multimodal Metric Learning Overview

Metric learning algorithms aim to learn a distance metric that better captures the relationships between data points compared to traditional distance metrics, such as Euclidean distance. In the context of multimodal data, this involves learning a set of metrics, one for each modality, and finding a way to relate these metrics in a common space. The proposed MML method extends traditional metric learning to handle heterogeneous data sources, specifically focusing on MRI and gene expression data, both of which differ significantly in their nature and representation.

The general goal is to learn a shared metric that aligns the data from different modalities into a common space while maintaining the inherent structure and relationships within each modality. This is achieved by constructing a composite metric, where each modality's metric is represented by a modality-specific part and a common part shared across all modalities.

### 6.2.3 Mathematical Formulation of the MML Algorithm

Let $X^{MRI}$ denote the matrix of MRI data, where each row represents an MRI scan of a subject, and $X^{Gene}$ denote the matrix of gene expression data, where each row corresponds to the gene expression profile of a subject. The goal is to learn a metric space that can simultaneously handle both types of data and project them into a common space where the relationship between the subjects is well-preserved.

We define the distance metric for each modality as follows:

$$d^{MRI}(x_1, x_2) = (x_1 - x_2)^\top M^{MRI}(x_1 - x_2), \tag{6.1}$$

$$d^{Gene}(y_1, y_2) = (y_1 - y_2)^\top M^{Gene}(y_1 - y_2), \tag{6.2}$$

where $M^{MRI}$ and $M^{Gene}$ are the learned Mahalanobis distance matrices for the MRI and gene expression data, respectively. These matrices are symmetric and positive semi-definite (p.s.d), which ensures they define a valid distance metric.

The key challenge is to align these distances in a common space. We introduce a shared metric $M^{Common}$ that connects the modality-specific distances through a weighted fusion approach. The metric for each modality is defined as:

$$M^{Modality} = M^{Modality}_{specific} \times M^{Common}, \tag{6.3}$$

where $M^{Modality}_{specific}$ represents the modality-specific matrix that captures the relevant subspace for each modality, and $M^{Common}$ is a shared matrix that applies across all modalities. The multiplication of these matrices ensures that the distance metric captures both the specific structure of each modality and the common underlying relationship across modalities.

### 6.2.4 Modality-Specific Projections

Given the substantial differences in the representation of MRI and gene expression data, the first step in our approach is to apply modality-specific projections to each modality. This step transforms the data into a space where the features are more comparable across modalities. The projections are learned by optimising a projection matrix for each modality, denoted as $P^{MRI}$ for MRI and $P^{Gene}$ for gene expression data.

For MRI data, the projection is defined as:

$$X_{proj}^{MRI} = X^{MRI} P^{MRI}, \tag{6.4}$$

and for gene expression data, the projection is:

$$X_{proj}^{Gene} = X^{Gene} P^{Gene}. \tag{6.5}$$

These projection matrices are learned during the training process to map each modality into a common feature space, ensuring that the features from both modalities are comparable.

### 6.2.5 Fusion in Common Space

After applying the modality-specific projections, the features from both modalities are mapped into a common feature space using the shared metric $M^{Common}$. This ensures that the features from both MRI and gene expression data are aligned in a common space, where a meaningful distance measure can be computed.

The goal of this fusion step is to project the modality-specific features $X_{proj}^{MRI}$ and $X_{proj}^{Gene}$ into a shared space such that the resulting fused representation reflects the true similarity between samples across modalities.

Let $X_{proj}^{MRI} \in \mathbb{R}^{n \times d_{MRI}}$ and $X_{proj}^{Gene} \in \mathbb{R}^{n \times d_{Gene}}$ represent the projected features of the MRI and gene expression data, respectively, where $n$ is the number of samples, and $d_{MRI}$ and $d_{Gene}$ are the dimensionalities of the respective feature spaces after projection.

To perform the fusion, we define the distance between a pair of samples $i$ and $j$ in the common space as:

$$d^{Common}(X_{proj}^{MRI}, X_{proj}^{Gene}) = \|X_{proj}^{MRI} - X_{proj}^{Gene}\|_{M^{Common}}^2, \tag{6.6}$$

where $\|\cdot\|_{M^{Common}}$ denotes the weighted Mahalanobis distance in the common space, and $M^{Common}$ is the shared metric matrix. This distance is computed by applying the shared metric to the difference between the projected features of each modality.

To explicitly calculate this distance, let:

$$d^{Common}(X_i^{MRI}, X_j^{Gene}) = (X_i^{MRI} - X_j^{Gene})^\top M^{Common}(X_i^{MRI} - X_j^{Gene}), \tag{6.7}$$

where $X_i^{MRI}$ and $X_j^{Gene}$ represent the feature vectors of samples $i$ and $j$ in the MRI and gene expression modalities, respectively. This distance measures how similar the samples are in the common space defined by $M^{Common}$.

The fusion procedure involves ensuring that the learned distance metric $M^{Common}$ effectively aligns the data from both modalities. To do so, we need to consider how the shared metric $M^{Common}$ can be learned in such a way that it optimally represents the relationships between the features in the common space.

The projection of both the MRI and gene expression data into a common space is done by applying the shared metric $M^{Common}$ to the projected data points. The resulting fused features are:

$$X_{fused} = \left[X_{proj}^{MRI} \; X_{proj}^{Gene}\right] M^{Common}, \tag{6.8}$$

where $X_{fused}$ is the fused feature matrix representing the common space, combining both MRI and gene expression data. This fusion process ensures that both sets of features are integrated and aligned according to the learned common metric.

In practice, the fusion is performed by calculating the distances between all pairs of samples from both modalities in the common space. The distance matrix $D_{fused}$

for all sample pairs $i, j$ is computed as:

$$D_{fused} = \begin{bmatrix} d^{Common}(X_1^{MRI}, X_1^{Gene}) & d^{Common}(X_1^{MRI}, X_2^{Gene}) & \cdots & d^{Common}(X_1^{MRI}, X_n^{Gene}) \\ d^{Common}(X_2^{MRI}, X_1^{Gene}) & d^{Common}(X_2^{MRI}, X_2^{Gene}) & \cdots & d^{Common}(X_2^{MRI}, X_n^{Gene}) \\ \vdots & \vdots & \ddots & \vdots \\ d^{Common}(X_n^{MRI}, X_1^{Gene}) & d^{Common}(X_n^{MRI}, X_2^{Gene}) & \cdots & d^{Common}(X_n^{MRI}, X_n^{Gene}) \end{bmatrix}, \tag{6.9}$$

where $D_{fused} \in \mathbb{R}^{n \times n}$ is the pairwise distance matrix in the common space. This matrix captures the relationships between all samples across both modalities, and is used for further analysis or classification tasks.

The distance matrix $D_{fused}$ is then used to compute a similarity matrix $S_{fused}$, where each entry represents the similarity between two samples in the common space. The similarity matrix is derived from the distance matrix as:

$$S_{fused} = \exp(-D_{fused}), \tag{6.10}$$

where the exponential function is applied element-wise to the distance matrix to convert the distances into similarities, where smaller distances correspond to higher similarities. This similarity matrix $S_{fused}$ is then used as input for downstream tasks, such as classification or clustering.

To ensure the fusion process is effective, the shared metric $M^{Common}$ is learned to minimise the fusion loss function, which encourages similar samples from different modalities to be closer in the common space. The fusion loss function is defined as:

$$\mathcal{L}_{fusion} = \sum_{(i,j) \in S} \| d^{Common}(X_i^{MRI}, X_j^{Gene}) - y_{ij} \|^2, \tag{6.11}$$

where: - $(i, j) \in S$ represents the set of pairs of samples $(i, j)$ with known similarity (based on side information), - $y_{ij} \in \{0, 1\}$ is the similarity label for the pair $(i, j)$, where 1 indicates a similar pair and 0 indicates a dissimilar pair.

The goal of the fusion procedure is to minimise this loss function, ensuring that the features from both modalities are aligned in the common space in a way that

enhances their representational power for downstream tasks.

Thus, the fusion of multimodal data is achieved through the shared metric $M^{Common}$, which aligns the projected MRI and gene expression features into a common space, enabling more meaningful relationships to be captured and allowing for more accurate classification and diagnosis of AD.

### 6.2.6 Loss Function for Metric Learning

To learn the appropriate metrics for the MRI and gene expression data, we define a loss function that encourages similar samples to be closer in the common space while pushing dissimilar samples apart. The loss function is formulated as follows:

$$L = \sum_{(i,j)\in S} \left[ \|d^{Common}(X_{proj}^{MRI}, X_{proj}^{Gene}) - y_{ij}\|^2 + \lambda \|M^{MRI}\|_F^2 + \lambda \|M^{Gene}\|_F^2 \right], \quad (6.12)$$

where: - $S$ is the set of pairs of samples $(i, j)$ for which we have side information on their similarity (i.e., whether the samples belong to the same class or not), - $y_{ij} \in \{0, 1\}$ indicates the label of the pair $(i, j)$, where 1 denotes similar pairs and 0 denotes dissimilar pairs, - $\|\cdot\|_F$ denotes the Frobenius norm, - $\lambda$ is a regularisation parameter that controls the complexity of the learned metrics.

The first term encourages the model to minimise the distance between similar pairs and maximise the distance between dissimilar pairs. The second and third terms regularise the learned metrics by penalising the Frobenius norm of the metric matrices, ensuring that they are not too large and helping to prevent overfitting.

### 6.2.7 Optimisation Procedure

The goal of the proposed MML algorithm is to jointly optimise the projection matrices $P^{MRI}$, $P^{Gene}$, the modality-specific metric matrices $M^{MRI}$, $M^{Gene}$, and the shared metric matrix $M^{Common}$. This can be formulated as a joint optimisation problem:

$$\mathcal{L} = \mathcal{L}_{fusion} + \mathcal{L}_{reg}, \tag{6.13}$$

where: - $\mathcal{L}_{fusion}$ is the fusion loss that measures the discrepancy between the fused features in the common space (as described in Equation (5)), - $\mathcal{L}_{reg} = \lambda_1\|P^{MRI}\|_F^2 + \lambda_2\|P^{Gene}\|_F^2 + \lambda_3\|M^{MRI}\|_F^2 + \lambda_4\|M^{Gene}\|_F^2 + \lambda_5\|M^{Common}\|_F^2$ is the regularisation term that penalises large values of the projection matrices and metric matrices, preventing overfitting.

The gradient of the loss function with respect to the model parameters is computed via backpropagation, and the optimisation is performed using an iterative gradient-based method such as Stochastic Gradient Descent (SGD) or Adam.

The gradients with respect to the matrices $P^{MRI}$, $P^{Gene}$, $M^{MRI}$, $M^{Gene}$, and $M^{Common}$ are computed as follows:

$$\frac{\partial \mathcal{L}}{\partial P^{MRI}} = \frac{\partial \mathcal{L}_{fusion}}{\partial P^{MRI}} + \frac{\partial \mathcal{L}_{reg}}{\partial P^{MRI}}, \tag{6.14}$$

$$\frac{\partial \mathcal{L}}{\partial P^{Gene}} = \frac{\partial \mathcal{L}_{fusion}}{\partial P^{Gene}} + \frac{\partial \mathcal{L}_{reg}}{\partial P^{Gene}}, \tag{6.15}$$

$$\frac{\partial \mathcal{L}}{\partial M^{MRI}} = \frac{\partial \mathcal{L}_{fusion}}{\partial M^{MRI}} + \frac{\partial \mathcal{L}_{reg}}{\partial M^{MRI}}, \tag{6.16}$$

$$\frac{\partial \mathcal{L}}{\partial M^{Gene}} = \frac{\partial \mathcal{L}_{fusion}}{\partial M^{Gene}} + \frac{\partial \mathcal{L}_{reg}}{\partial M^{Gene}}, \tag{6.17}$$

$$\frac{\partial \mathcal{L}}{\partial M^{Common}} = \frac{\partial \mathcal{L}_{fusion}}{\partial M^{Common}} + \frac{\partial \mathcal{L}_{reg}}{\partial M^{Common}}. \tag{6.18}$$

Each of these gradients involves calculating the contribution of both the fusion loss and the regularisation term, ensuring that the optimisation process encourages both accurate fusion of multimodal features and well-regularised learned metrics.

After the gradients are computed, the model parameters are updated iteratively using a gradient descent-based optimisation algorithm. This process continues until the loss function converges to a minimum, indicating that the model has learned an optimal set of projection matrices and metric matrices for multimodal data fusion.

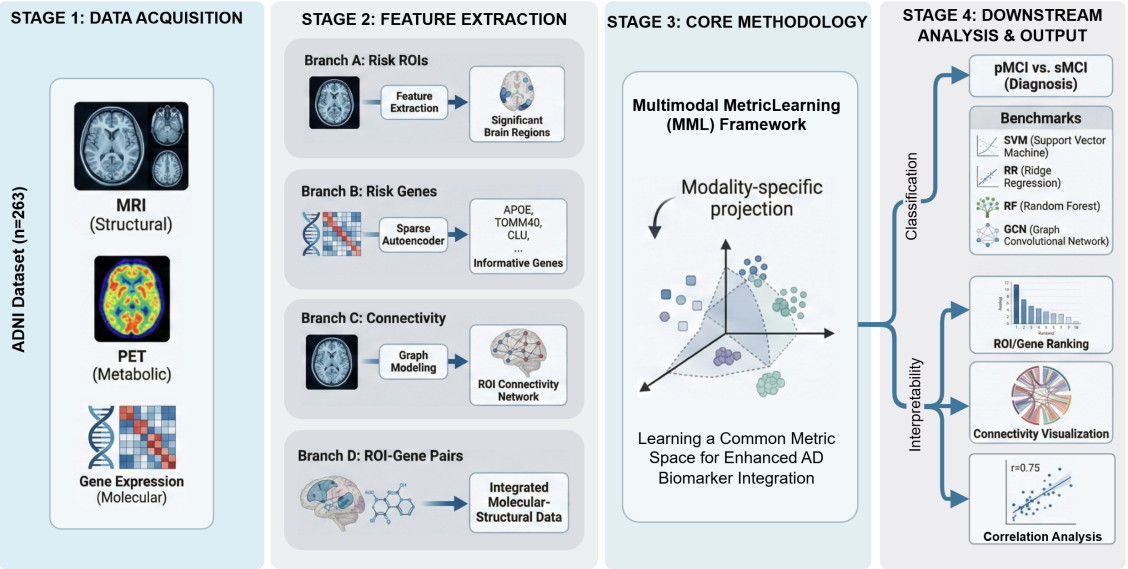



Figure 6.1: Overall analysis procedure of the proposed MML method, from modality-specific preprocessing through multimodal fusion, classification, and interpretation of genes, ROIs, SNP-ROI pairs, and functional connectivity.

## 6.3 Experimental Settings

In this section, we present the experimental settings used to evaluate the proposed MML algorithm for the early diagnosis of AD. The focus of the experiments is to classify pMCI and sMCI, which represent two key stages in the AD continuum. We compare the performance of the MML algorithm with traditional ML methods, such as SVM, RR, and RF, as well as deep learning models, specifically GCNs, which are better suited to handling non-Euclidean data.

### 6.3.1 AD Early Diagnostic Performance Analysis

The primary aim of the experiments is to assess the diagnostic performance of the proposed MML algorithm in classifying pMCI and sMCI from neuroimaging data (MRI) and gene expression data. These metrics are computed to evaluate how well the proposed model distinguishes between the two classes of MCI, which are critical for early AD diagnosis.

### Evaluation Metrics

To evaluate the performance of different models, we use the following standard evaluation metrics:

- **ACC**: The proportion of correctly classified samples out of all samples.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.19}$$

  where $TP$ is the number of true positives, $TN$ is the number of true negatives, $FP$ is the number of false positives, and $FN$ is the number of false negatives.

- **PRE**: The proportion of true positive predictions among all positive predictions.

$$PRE = \frac{TP}{TP + FP} \tag{6.20}$$

- **SEN**: The proportion of true positives correctly identified out of all actual positive instances.

$$SEN = \frac{TP}{TP + FN} \tag{6.21}$$

- **SPE**: The proportion of true negatives correctly identified out of all actual negative instances.

$$SPE = \frac{TN}{TN + FP} \tag{6.22}$$

- **AUC**: AUC measures the ability of the model to discriminate between classes. It is the area under the ROC curve, which plots the true positive rate (SEN) against the false positive rate (1 - SPE).

- **F1 score**: The harmonic mean of PRE and SEN, providing a balance between the two.

$$F1 = 2 \times \frac{PRE \times SEN}{PRE + SEN} \tag{6.23}$$

### Comparative Methods

The proposed MML algorithm is compared with several baseline models to evaluate its effectiveness in the multimodal fusion task:

- **SVM**: We use SVM with a radial basis function (RBF) kernel as a conventional non-linear classifier baseline.

- **RR**: RR is a linear regression method that applies L2 regularisation to prevent overfitting. In the context of classification, RR can be used with a logistic loss function to predict binary outcomes (progressive or stable MCI).

- **RF**: RF is an ensemble method that constructs multiple decision trees during training and outputs the class that is the mode of the classes predicted by individual trees. This method is known for its robustness and high generalisation performance.

- **GCNs**: GCNs are a deep learning method designed to operate on graph-structured data. GCNs extend the concept of convolutional layers to non-Euclidean spaces by processing data represented as graphs. In this study, we use GCNs to model the relationships between ROIs in MRI data and genes in gene expression data. GCNs naturally capture the structural dependencies between different nodes (e.g., brain regions or genes) in the data.

### 6.3.2 Experimental Design

As can be seen in Figure 6.1, the experimental project is designed around four key components that capture the various aspects of AD progression:

- **Risk ROI Analysis**: This step involves identifying and analysing the ROIs in the brain that are most strongly associated with the risk of AD. Using MRI data, we perform a feature extraction process to focus on ROIs that show significant differences between pMCI and sMCI groups. These ROIs are then used as input for the classification models.

- **Risk Gene Analysis**: This component analyses gene expression data to identify genes that are associated with the risk of AD progression. By leveraging feature selection techniques, such as sparse AEs, we focus on the most informative genes that help distinguish between pMCI and sMCI. These gene

expression features are used alongside MRI data in the multimodal classification models.

- **ROI-Gene Pair Analysis**: In this step, we investigate the relationships between brain regions (ROIs) and gene expression patterns. By pairing each ROI with corresponding gene expression data, we form multimodal feature vectors that represent both structural and molecular data. These paired features are used to evaluate the performance of the multimodal fusion methods, including the proposed MML algorithm.

- **ROI Connectivity Analysis**: This component focuses on the connectivity between brain regions and its relationship to AD. Using MRI data, we model the brain's network by considering the connectivity between different ROIs. This network representation is then used in the classification models to capture the interactions between brain regions, which could provide critical insights into the progression of AD.

### 6.3.3 Training and Testing Procedure

For all the methods, we perform a 5-fold cross-validation to evaluate the performance of the models. In each fold, the data is split into a training set (80% of the data) and a test set (20% of the data). The training set is used to train the models, and the test set is used to evaluate the classification performance. The performance metrics described in Section 3.4 are computed on the test set for each fold, and the final performance is obtained by averaging the results across all folds.

### 6.3.4 Implementation Details

For the traditional ML methods (SVM, RR, RF), we use standard implementations available in Python libraries such as scikit-learn. For SVM, we employ an RBF kernel with an optimised regularisation parameter, and for RF, we use the default settings with 100 trees.

The GCN is implemented using the PyTorch Geometric library, which provides

an efficient framework for graph-based deep learning. The GCN model is trained using the Adam optimiser with a learning rate of 0.001, and the number of layers is set to 2. We use a ReLU activation function between layers and a softmax activation at the output layer for classification.

The MML algorithm is implemented in TensorFlow, and the shared metric matrix $M^{Common}$ is learned via stochastic gradient descent (SGD). The regularisation parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ are tuned using cross-validation to balance the loss function's fusion and regularisation components.

## 6.4 Results and discussion

### 6.4.1 Results of comparative experiments

Traditional ML methods, SVM, RR, and RF, were used for comparison. Two types of deep learning-based methods, CNNs based on Euclidean space and GCN methods based on non-Euclidean space, were also considered. The classification performance of all comparison methods on varied modality data is documented in detail in Tables 6.2, 6.3, and 6.4. These tables present a comparative analysis of the proposed MML method across three datasets: gene expression data (Gene), ROIs extracted from sMRI and fMRI (ROI), and the fusion of gene expression data with ROI data (Dual). The metrics evaluated include ACC, SEN, SPE, AUC, and F1-score. The results demonstrate the superior performance of the proposed MML method across all modalities, particularly in the Dual dataset.

For the Gene dataset, MML achieves the highest ACC (0.6597), SPE (0.6301), and AUC (0.6577), significantly outperforming traditional methods. Notably, the F1-score of MML (0.6335) is slightly lower than CNN (0.6577), but the overall balance across metrics indicates robust performance. GCN achieves the highest SEN (0.7167) in this dataset but suffers from poor SPE (0.3808), highlighting its tendency to overfit positive samples.

For the ROI dataset, MML achieves the best performance across all metrics, including ACC (0.7900), SEN (0.8039), SPE (0.8045), AUC (0.7897), and F1-score

| Methods | Gene | | | | |
|---|---|---|---|---|---|
| | ACC | SEN | SPE | AUC | F1 |
| SVM | 0.4837 | 0.5077 | 0.4590 | 0.4833 | 0.4647 |
| RR | 0.5389 | 0.5141 | 0.5615 | 0.5378 | 0.5445 |
| RF | 0.5395 | 0.5327 | 0.5442 | 0.5385 | 0.5357 |
| CNN | 0.6329 | 0.6304 | 0.6026 | 0.6301 | **0.6577** |
| GCN | 0.5511 | **0.7167** | 0.3808 | 0.5487 | 0.3424 |
| MML | **0.6597** | 0.6853 | **0.6301** | **0.6577** | 0.6335 |

Table 6.2: Classification performance of the proposed MML method and baseline methods using gene expression data only.

(0.7750). These results underscore the method's capacity to handle complex structural and functional imaging data effectively. GCN performs well with high SEN (0.7744), but its SPE (0.6494) and overall ACC (0.7109) are lower than those of MML, indicating reduced reliability in distinguishing between classes.

The Dual dataset, which fuses gene expression data with ROI data, highlights the strength of MML in multimodal data integration. MML achieves the highest ACC (0.7968), SEN (0.7955), SPE (0.8564), AUC (0.7346), and F1-score (0.8066), outperforming all other methods by a substantial margin. The performance gains in this dataset validate the proposed method's capability to leverage complementary information from genomic and imaging data. While GCN also performs well in this dataset, its metrics are consistently lower than those of MML, particularly in SPE (0.6827), emphasising the advantage of MML's multimodal learning framework.

Overall, the proposed MML method consistently demonstrates superior performance across all datasets and metrics, particularly excelling in the integration of gene expression and ROI data. These results highlight the effectiveness of MML in handling both unimodal and multimodal data, making it a robust approach for early diagnosis of AD.

### 6.4.2 Interpretability Analysis

This section focuses on the interpretability of the proposed methodology. Firstly, we ranked the feature importance of ROIs and gene features extracted during MML preprocessing and visualised them for further analysis; then, we performed a corre-

| Methods | ROI | | | | |
|---|---|---|---|---|---|
| | ACC | SEN | SPE | AUC | F1 |
| SVM | 0.6445 | 0.5923 | 0.6955 | 0.6439 | 0.6575 |
| RR | 0.6289 | 0.6481 | 0.6096 | 0.6288 | 0.6158 |
| RF | 0.6883 | 0.6750 | 0.7026 | 0.6888 | 0.6892 |
| CNN | 0.6211 | 0.6186 | 0.7413 | 0.5673 | 0.6699 |
| GCN | 0.7109 | 0.7744 | 0.6494 | 0.7119 | 0.6824 |
| MML | **0.7900** | **0.8039** | **0.8045** | **0.7897** | **0.7750** |

Table 6.3: Classification performance of the proposed MML method and baseline methods using neuroimaging ROI features only.

| Methods | Multimodal Data | | | | |
|---|---|---|---|---|---|
| | ACC | SEN | SPE | AUC | F1 |
| SVM | 0.5629 | 0.5077 | 0.6179 | 0.5628 | 0.5766 |
| RR | 0.5508 | 0.5231 | 0.5782 | 0.5506 | 0.5584 |
| RF | 0.6952 | 0.6968 | 0.6936 | 0.6952 | 0.6918 |
| CNN | 0.6412 | 0.6413 | 0.7163 | 0.5462 | 0.7365 |
| GCN | 0.7183 | 0.7474 | 0.6827 | 0.7151 | 0.6911 |
| MML | **0.7968** | **0.7955** | **0.8564** | **0.7346** | **0.8066** |

Table 6.4: Classification performance of the proposed MML method and baseline methods using fused gene expression and neuroimaging ROI features.

lation analysis between the top 10 ROIs and the top 10 genes to confirm pairwise associations in each SNP-ROI pair. Finally, we compared brain functional connectivity patterns of sMCI and pMCI cases classified by the model to examine the medical significance of the classification results.

For neuroscientists and clinicians, fused features are most useful when they can be traced back to recognisable brain regions, genes, and gene-region relationships. The interpretation strategy in this chapter therefore decomposes the fused representation into modality-specific evidence: risk genes are examined through enrichment analysis, risk ROIs are mapped back to anatomical regions, SNP-ROI pairs are analysed as genotype-phenotype associations, and reconstructed functional connectivity is compared between sMCI and pMCI. This does not make the MML model fully transparent, but it provides a structured route from fused features to biological hypotheses that can be externally validated.

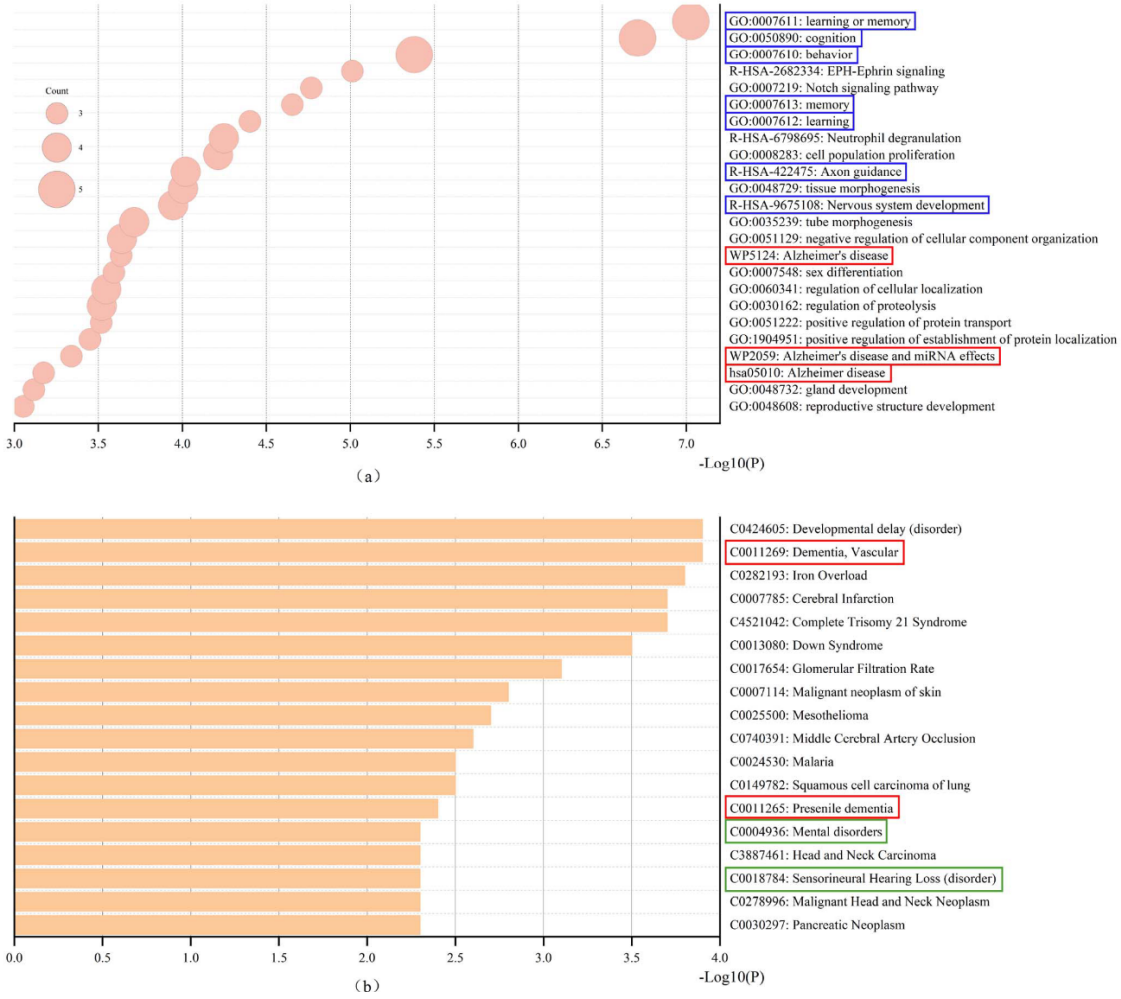



Figure 6.2: Gene enrichment visualisation for high-risk genes identified by MML, illustrating biological processes, pathways, and disease associations relevant to AD and neurological disorders.

### Significance Analysis of High Risk Genes

The enrichment visualisation in Figure 6.2.a illustrates how the genes identified by the MML interpretation analysis may be interpreted in relation to biological processes and pathways associated with AD and related neurological disorders. The enriched terms include processes such as learning, memory, cognition, and axon guidance, which are fundamental to proper neural function and cognitive health. Pathways like "Notch signalling" and "EPH-Ephrin signalling" further support this form of interpretation, as these pathways are essential for neural development, synaptic plasticity, and cellular organisation. The inclusion of AD-specific pathways, such as "miRNA effects in AD," demonstrates how the selected genes can be linked to mechanisms underlying neurodegeneration.

The disease-association panel in Figure 6.2.b also highlights associations with neurological diseases and comorbidities, including vascular dementia, presenile dementia, and sensorineural hearing loss. These conditions are known contributors to or manifestations of cognitive decline in AD, underscoring the genes' potential involvement in broader neurodegenerative and cerebrovascular processes. Additionally, links to "mental disorders" and other nervous system disorders point to a wide-ranging impact of these genes on brain health and disease.

### Significance Analysis of High Risk ROIs

The ROI visualisation in Figure 6.3 illustrates how high-contribution neuroimaging features can be mapped back to recognisable brain regions. The visualised regions include the hippocampus, amygdala, temporal lobe, posterior cingulate, precuneus, entorhinal cortex, insula, and fusiform gyrus. These regions are commonly implicated in memory, emotion regulation, and neural connectivity, and therefore provide a clinically interpretable route from the fused MML representation back to brain anatomy. The contribution ranking shows the type of ROI-level explanation that can be produced by the proposed interpretation pipeline.

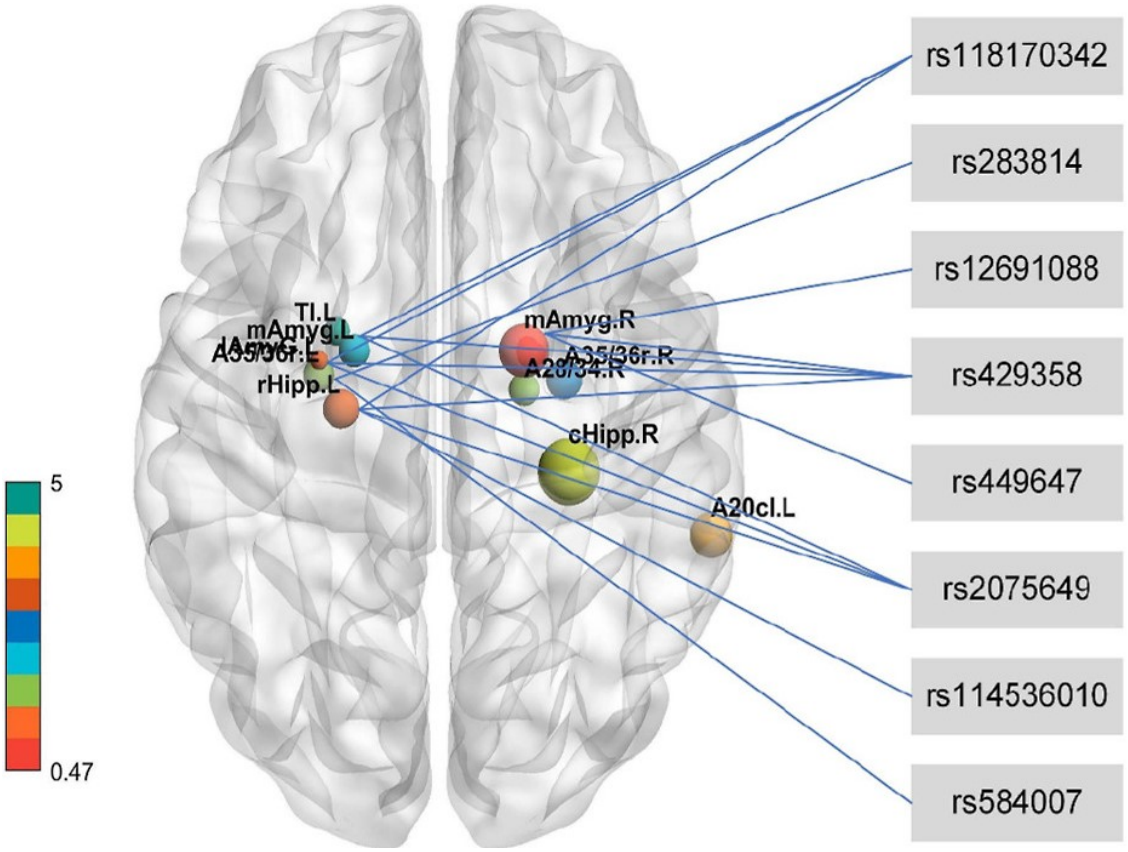



Figure 6.3: Visualisation of risk ROIs identified by MML, showing brain regions with high contribution to pMCI and sMCI classification.

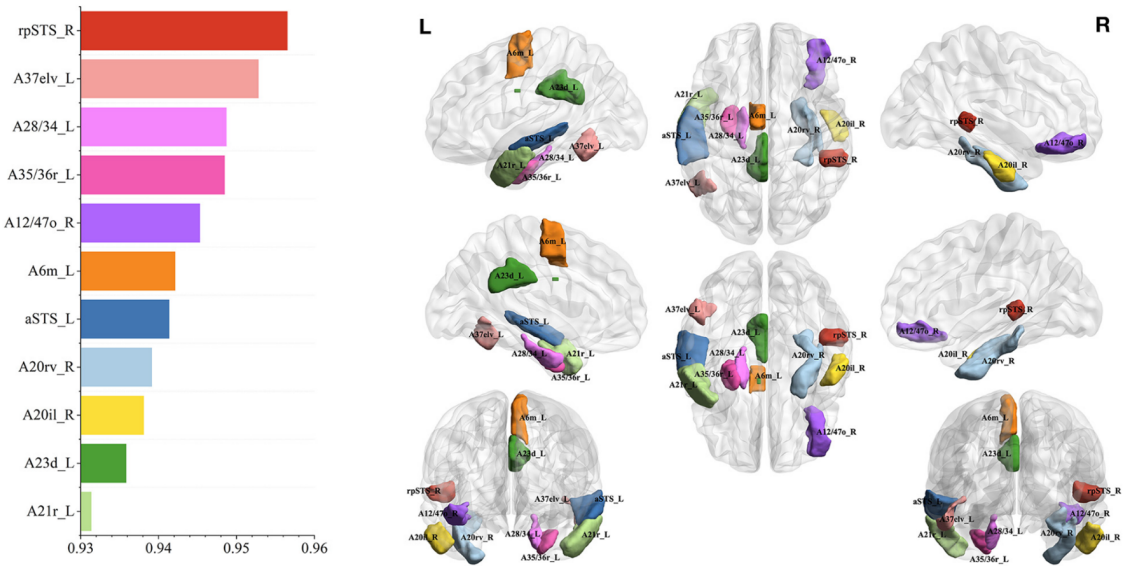



Figure 6.4: SNP-ROI pairs identified by the MML interpretation analysis, linking genetic variants with brain regions relevant to cognition and AD progression.

### High Risk ROI-Gene pairwise Correlation Analysis

The network in Figure 6.4 illustrates SNP-ROI pairwise associations, mapping genetic variants to brain regions associated with cognition and neurological function. Different edge colours represent the direction of the association, while line thickness indicates relative association strength. Notable ROIs include the hippocampus (rHipp, cHipp), amygdala (mAmyg), and temporal lobe, regions heavily implicated in memory, emotion regulation, and neural connectivity. These ROIs are critical for maintaining cognitive health, and disruptions in their function are strongly associated with neurodegenerative diseases such as AD.

Several key SNPs used in the visualisation (e.g., rs429358, rs449647) are already well-documented in AD research. For instance, rs429358 is part of the APOE gene, a well-known genetic risk factor for AD, linked to amyloid plaque deposition and neuroinflammation. The illustrated connections between these SNPs and the highlighted brain regions emphasise their potential influence on the structural and functional integrity of areas involved in learning and memory, making them strong candidates for further investigation in understanding the genetic underpinnings of neurodegeneration.

This mapping also provides insights into possible genotype-phenotype correlations. For example, the interaction of SNPs with specific ROIs such as the hippocampus, a region critical for memory consolidation, suggests a mechanistic link between genetic risk factors and the early symptoms of AD, such as memory loss. These findings pave the way for targeted research on how these genetic variations influence neural circuits, potentially informing biomarker development and individualised therapeutic strategies for AD and related neurological disorders.

### Interpretability analysis of classification results

The functional connectivity patterns for sMCI and pMCI illustrate potential differences in brain network integrity. In the sMCI visualisation in Figure 6.5, the functional connections appear more robust, with dense and widespread connectivity represented by coloured strings. This illustrates the expected interpretation

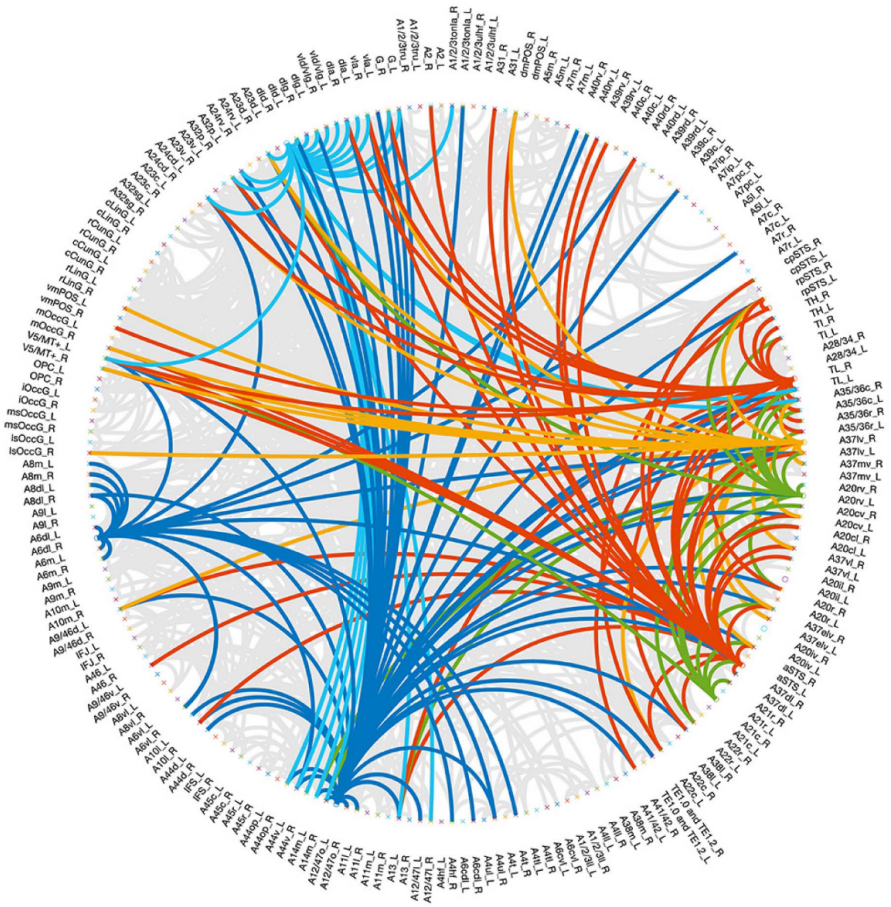


Figure 6.5: Functional connectivity pattern for sMCI cases, showing salience-associated ROI connections in the foreground and remaining ROI connections in the background.

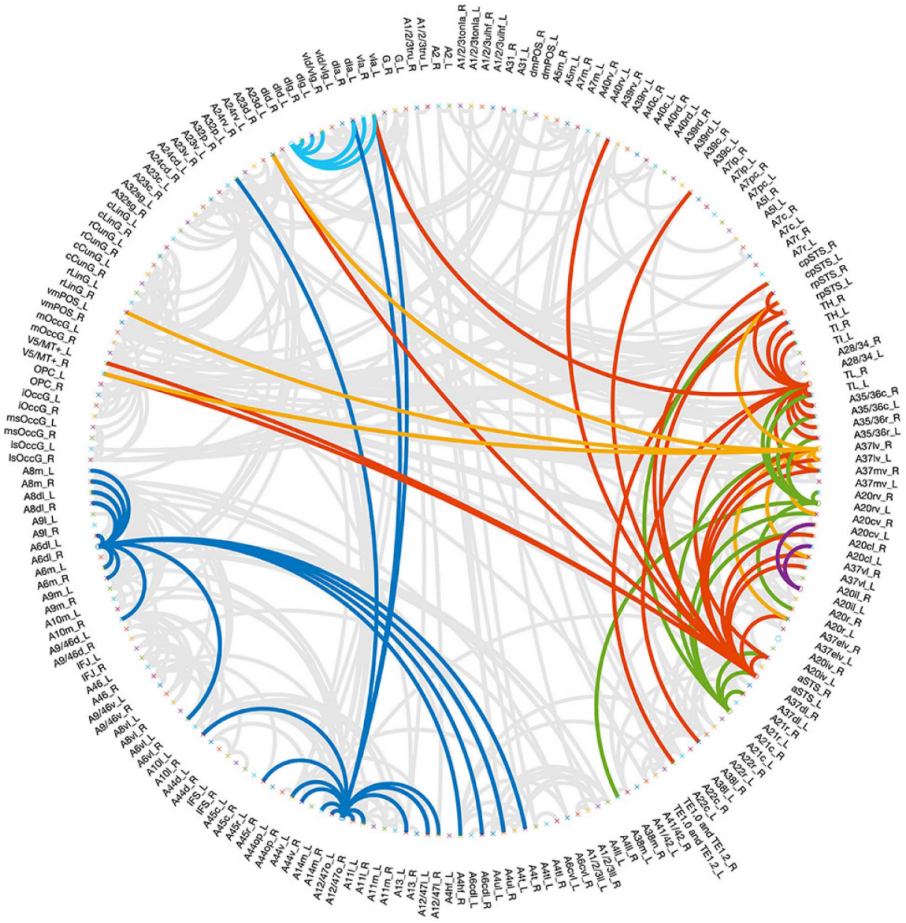


Figure 6.6: Functional connectivity pattern for pMCI cases, showing reduced salience-associated and background connectivity relative to sMCI.

that, while these individuals have cognitive impairment, their brain networks may maintain a degree of functional integration, possibly contributing to their stable condition.

In contrast, the pMCI visualisation in Figure 6.6 shows a marked reduction in functional connectivity, particularly in salience-associated connections. This reduction illustrates the interpretation that progressive breakdown of critical neural networks may affect higher-order cognitive functions. The decline in connectivity indicates that the brain's ability to compensate for cognitive decline may diminish over time, contributing to progression toward AD or more severe dementia.

By mapping brain functional connectivity, we found that the proposed MML not only improves the performance of sMCI and pMCI classification, but also provides biologically meaningful differences at the level of brain functional connectivity. The sMCI and pMCI cases classified by the proposed method show significant differences in brain connectivity.

## 6.5 Summary

In summary, the MML algorithm presented in this chapter demonstrates significant potential for advancing early AD diagnosis, particularly in the context of pMCI. Here, progressive refers to MCI cases that later convert to AD, while sMCI refers to cases that do not convert during follow-up. By combining data from PET and MRI scans with SNP information, the MML algorithm identifies high-risk ROIs and genetic markers associated with AD pathogenesis. The achieved accuracy of 84.4% and AUC of 0.843 in early pMCI diagnosis further underscore the efficacy of this approach.

Moreover, MML's ability to capture and synthesise complex interactions between multimodal data types offers valuable insights into the underlying mechanisms of pMCI progression. The identification of potential pathogenic factors highlights the algorithm's utility in unravelling the biological processes that may precede AD, fostering new avenues for early intervention.

Nevertheless, there are inherent limitations within this study, such as the un-

confirmed pathogenic factors and the need for additional collaborative efforts to validate these findings. Further research will be essential in refining the algorithm, expanding the dataset, and confirming the identified markers' relevance in larger and more diverse populations. Future work should also explore integrating additional modalities, especially EHRs and proteomics. EHRs could add longitudinal clinical context such as medication, comorbidity, and cognitive trajectories, while proteomics could provide a molecular layer closer to disease mechanisms than transcriptomic measurements alone. Ethical issues also require attention: algorithmic bias should be assessed across demographic groups, consent and data governance must be explicit for multimodal patient data, and clinically deployed models should provide auditable explanations for high-risk predictions.

Ultimately, the work presented here lays the foundation for more precise, data-driven tools in the early diagnosis and monitoring of AD, with the potential to improve patient outcomes through earlier interventions and personalised treatment strategies.

# Chapter 7

# Conclusion and Future Work

## 7.1 Conclusion

The early diagnosis of AD is critical for timely intervention, monitoring, and clinical trial recruitment, yet it remains challenging because AD is biologically complex and because available data are often high-dimensional, heterogeneous, and limited in sample size. This thesis addressed these challenges through three linked contributions: stacked AE-based dimensionality reduction for gene expression data, interpretable feature selection for biologically meaningful probes, and heterogeneous multimodal fusion of gene expression and neuroimaging data.

The first contribution was the development and evaluation of stacked sparse and denoising AEs for dimensionality reduction of high-dimensional gene expression data. Traditional dimensionality reduction methods, such as PCA and manifold-learning baselines, can struggle to capture non-linear gene expression patterns when sample sizes are limited. The results in Chapter 4 show that AE-based representations can improve downstream classification when the reduced dimension and network depth are selected carefully. This contribution addresses the high-dimensional, small-sample, non-linear data problem identified in the literature review.

The second contribution was an interpretable deep learning-based feature selection framework. Chapter 5 combined a shallow sparse AE with XGBoost-based feature ranking and Fast-RFE feature selection. The shallow architecture was deliberately chosen because the first hidden layer remains directly connected to input

probes, making it more suitable for biological interpretation than deeper latent representations. Enrichment analysis provided evidence that the selected probes were biologically meaningful and not merely predictive artefacts. This contribution addresses the interpretability gap in deep learning-based gene expression analysis.

The third contribution was a heterogeneous multimodal fusion method based on multimodal metric learning. Chapter 6 integrated neuroimaging and gene expression data while preserving modality-specific and complementary information. The proposed method outperformed unimodal baselines and several comparison methods in pMCI and sMCI classification, while the interpretation analysis linked fused features back to risk genes, ROIs, SNP-ROI pairs, and functional connectivity patterns. This contribution addresses the challenge of combining molecular and imaging biomarkers without collapsing their distinct biological meanings.

Collectively, these contributions show that diagnostic performance and interpretability should be developed together. For early AD diagnosis, improved classification accuracy is valuable only if the model can also support biological understanding, external validation, and eventual clinical trust.

## 7.2 Limitations

Several limitations should be acknowledged. First, the gene expression and multimodal datasets remain relatively small compared with the feature dimensionality, which increases the risk of overfitting and limits generalisability. Second, the experiments rely primarily on retrospective cohort data; prospective validation would be required before clinical use. Third, cross-cohort differences in gene expression platforms and preprocessing pipelines may affect reproducibility, even after probe harmonisation and normalisation. Fourth, the biological findings from enrichment and SNP-ROI analysis are hypothesis-generating and require independent experimental or clinical validation. Finally, although the models include interpretability analyses, they are not yet complete clinical explanation systems and would need further refinement for clinician-facing decision support.

## 7.3 Future Work

### Larger and More Diverse Validation

Future work should validate the proposed methods on larger multi-centre datasets with broader demographic coverage. This would allow more reliable assessment of generalisability across age, sex, ethnicity, scanner type, cohort recruitment criteria, and disease stage. External validation should include locked test cohorts and prospective studies to reduce the risk of optimistic performance estimates.

### Additional Modalities

The next modality to add should be EHR, because EHR data can provide longitudinal clinical context such as medication history, comorbidities, cognitive assessments, and disease progression timelines. Proteomics is also a strong candidate because protein-level changes are closer to many disease mechanisms than transcriptomic measurements alone. Combining EHR, proteomics, gene expression, and neuroimaging could improve diagnosis by linking molecular signals, brain structure and function, and real-world clinical trajectories.

### Clinical Deployment

Moving these methods toward clinical deployment would require external validation, calibration analysis, robustness testing, regulatory review, cost-effectiveness assessment, and integration with clinical workflow. The models would need to provide auditable outputs, clear uncertainty estimates, and explanations that clinicians can inspect. Deployment would also require standardised data acquisition and preprocessing pipelines so that predictions are stable across hospitals and laboratory platforms.

### Interpretability and Biological Validation

Further work should strengthen the biological validation of selected biomarkers. This could include systematic comparison with known AD-related genes, literature

mining, pathway databases, independent omics cohorts, and collaboration with laboratory researchers. For multimodal fusion, future work should quantify modality-specific contributions and develop visual explanations that connect model decisions to genes, ROIs, and clinical variables.

### Ethical and Governance Considerations

Ethical issues must be handled as part of future development rather than after deployment. Algorithmic bias should be evaluated across demographic and clinical subgroups, especially because AD datasets can under-represent some populations. Patient consent must clearly cover multimodal data linkage and secondary analysis. Data privacy, governance, and auditability are also essential, particularly when genetic and imaging data are combined. Clinically used systems should support human oversight rather than replacing clinical judgement.

## 7.4 Closing Remarks

This thesis has demonstrated the potential of interpretable deep learning and heterogeneous multimodal fusion for early AD diagnosis. By addressing dimensionality reduction, feature interpretability, and multimodal complementarity, the work provides a foundation for future diagnostic models that are not only accurate, but also biologically meaningful and more suitable for responsible clinical translation.

# Bibliography


Alzheimer's Association (2024), '2024 alzheimer's disease facts and figures', *Alzheimer's & dementia* **15**(3), 321–387.

Andrew, G., Arora, R., Lee, K. H. and Ng, A. Y. (2013), 'Deep canonical correlation analysis', *Proceedings of the 30th International Conference on Machine Learning (ICML)* **28**, 1247–1255.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32.
**URL:** *https://link.springer.com/article/10.1023/A:1010933404324*

Castellani, R. J., Rolston, R. K. and Smith, M. A. (2010), 'Alzheimer disease', *Disease-a-month: DM* **56**(9), 484.

Chen, J., Liu, Y. and Zhao, X. (2021), 'Autoencoder-based feature selection for gene expression data', *IEEE Transactions on Computational Biology and Bioinformatics* **18**(4), 1211–1219.

Chen, T. and Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining', pp. 785–794.

Chopra, S., Hadsell, R. and LeCun, Y. (2005), 'Learning a similarity metric discriminatively, with application to face verification', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 539–546.

Deng, X., Zhuang, H. and Zhang, X. (2016), 'Wrapper and filter-based feature selection combined with principal component analysis for dimension reduction in gene expression data', *Bioinformatics* **32**(5), 701–709.

Du, L., Liu, F., Liu, K., Yao, X., Risacher, S. L., Han, J., Saykin, A. J. and Shen, L. (2020), 'Associating multi-modal brain imaging phenotypes and genetic risk factors via a dirty multi-task learning method', *IEEE transactions on medical imaging* **39**(11), 3416–3428.

Du, L., Liu, K., Yao, X., Risacher, S. L., Han, J., Saykin, A. J., Guo, L. and Shen, L. (2020), 'Detecting genetic associations with brain imaging phenotypes in alzheimer's disease via a novel structured scca approach', *Medical image analysis* **61**, 101656.

Duda, R. O., Hart, P. E. and Stork, D. G. (2001), *Pattern Classification*, John Wiley & Sons, Inc.

Espezua, S., Villanueva, E., Maciel, C. D. and Carvalho, A. (2015), 'A projection pursuit framework for supervised dimension reduction of high dimensional small sample datasets', *Neurocomputing* **149**, 767–776.

Fathi, S., Ahmadi, M. and Dehnad, A. (2022), 'Early diagnosis of alzheimer's disease based on deep learning: A systematic review', *Computers in biology and medicine* **146**, 105634.

Fessel, J. (2019), 'Prevention of alzheimer's disease by treating mild cognitive impairment with combinations chosen from eight available drugs', *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **5**, 780–788.

Fisher, R. A. (1936), 'The use of multiple measurements in taxonomic problems', *Annals of Eugenics* **7**(2), 179–188.

Folstein, M. F., Folstein, S. E. and McHugh, P. R. (1975), 'Mini-mental state: a practical method for grading the cognitive state of patients for the clinician', *Journal of Psychiatric Research* **12**(3), 189–198.

Friedman, J. H. (2001), 'Greedy function approximation: A gradient boosting machine', *Annals of Statistics* **29**(5), 1189–1232.

Gatz, M., Reynolds, C. A., Fratiglioni, L., Johansson, B., Mortimer, J. A., Berg, S., Fiske, A. and Pedersen, N. L. (2006), 'Role of genes and environments for explaining alzheimer disease', *Arch. Gen. Psychiatry* **63**(2), 168–174.

Ghosal, S., Chen, Q., Pergola, G., Goldman, A. L., Ulrich, W., Berman, K. F., Blasi, G., Fazio, L., Rampino, A., Bertolino, A. et al. (2021), G-mind: an end-to-end multimodal imaging-genetics framework for biomarker identification and disease classification, *in* 'Medical Imaging 2021: Image Processing', Vol. 11596, SPIE, pp. 63–72.

Ghosal, S., Chen, Q., Pergola, G., Goldman, A. L., Ulrich, W., Weinberger, D. R. and Venkataraman, A. (2021), 'A biologically interpretable graph convolutional network to link genetic risk pathways and neuroimaging markers of disease', *bioRxiv* pp. 2021–05.

Goedert, M. and Spillantini, M. G. (2006), 'A century of alzheimer's disease', *science* **314**(5800), 777–781.

Gustavsson, A., Norton, N., Fast, T., Frölich, L., Georges, J., Holzapfel, D., Kirabali, T., Krolak-Salmon, P., Rossini, P. M., Ferretti, M. T. et al. (2023), 'Global estimates on the number of persons across the alzheimer's disease continuum', *Alzheimer's & Dementia* **19**(2), 658–670.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002*a*), 'Gene selection for cancer classification using support vector machines', *Machine Learning* **46**(1-3), 389–422.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002*b*), 'Gene selection for cancer classification using support vector machines', *Machine learning* **46**, 389–422.

Hall, M. A. (1999), 'Correlation-based feature selection for discrete and numeric class machine learning', *Proceedings of the Seventeenth International Conference on Machine Learning* pp. 359–366.

Hocking, R. R. (2013), 'Stepwise selection in regression: The need for caution', *Journal of the American Statistical Association* **78**(383), 265–275.

Hoerl, A. E. and Kennard, R. W. (1970), 'Ridge regression: biased estimation for non-orthogonal problems', *Technometrics* **12**(1), 55–67.

Horvath, S. and Raj, K. (2018), 'Dna methylation-based biomarkers and the epigenetic clock theory of ageing', *Nature reviews genetics* **19**(6), 371–384.

Horvath, S., Zhang, Y., Langfelder, P., Kahn, R. S., Boks, M. P., van Eijk, K., van den Berg, L. H. and Ophoff, R. A. (2012), 'Aging effects on dna methylation modules in human brain and blood tissue', *Genome biology* **13**, 1–18.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. (2003), 'Exploration, normalization, and summaries of high density oligonucleotide array probe level data', *Biostatistics* **4**(2), 249–264.

Jolliffe, I. T. and Cadima, J. (2016), 'Principal component analysis: a review and recent developments', *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* **374**(2065), 20150202.

Karlsson Linnér, R., Biroli, P., Kong, E., Meddens, S. F. W., Wedow, R., Fontana, M. A., Lebreton, M., Tino, S. P., Abdellaoui, A., Hammerschlag, A. R. et al. (2019), 'Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences', *Nature genetics* **51**(2), 245–257.

Kennedy, J. and Eberhart, R. (2015), 'Particle swarm optimization', *Proceedings of the IEEE International Conference on Neural Networks* **4**, 1942–1948.

Khagi, B. and Kwon, G.-R. (2020), '3d cnn design for the classification of alzheimer's disease using brain mri and pet', *IEEE Access* **8**, 217830–217847.

Khaire, U. M. and Dhanalakshmi, R. (2022), 'Stability of feature selection algorithm: A review', *Journal of King Saud University-Computer and Information Sciences* **34**(4), 1060–1073.

Kim, D.-G., Krenz, A., Toussaint, L. E., Maurer, K. J., Robinson, S.-A., Yan, A., Torres, L. and Bynoe, M. S. (2016), 'Non-alcoholic fatty liver disease induces signs of alzheimer's disease (ad) in wild-type mice and accelerates pathological signs of ad in an ad model', *Journal of neuroinflammation* **13**, 1–18.

Kira, K. and Rendell, L. A. (1992), 'A practical approach to feature selection', *Proceedings of the Ninth International Conference on Machine Learning* pp. 249–256.

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983), 'Optimization by simulated annealing', *Science* **220**(4598), 671–680.

Knopman, D. S., Amieva, H., Petersen, R. C., Chételat, G., Holtzman, D. M., Hyman, B. T., Nixon, R. A. and Jones, D. T. (2021), 'Alzheimer disease', *Nature reviews Disease primers* **7**(1), 33.

Kohavi, R. (1995), 'A study of cross-validation and bootstrap for accuracy estimation and model selection', pp. 1137–1143.

Kullback, S. and Leibler, R. A. (1959), *On Information and Sufficiency*, Vol. 22, The Annals of Mathematical Statistics.

Kumar, K., Kumar, A., Keegan, R. M. and Deshmukh, R. (2018), 'Recent advances in the neurobiology and neuropharmacology of alzheimer's disease', *Biomedicine & pharmacotherapy* **98**, 297–307.

Lee, S., Kwak, M., Tsui, K.-L. and Kim, S. B. (2019), 'Process monitoring using variational autoencoder for high-dimensional nonlinear processes', *Engineering Applications of Artificial Intelligence* **83**, 13–27.

Li, H., Li, J., Li, N., Li, B., Wang, P. and Zhou, T. (2011), 'Cognitive intervention for persons with mild cognitive impairment: A meta-analysis', *Ageing research reviews* **10**(2), 285–296.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. and Liu, H. (2017), 'Feature selection: A data perspective', *ACM computing surveys (CSUR)* **50**(6), 1–45.

Li, J., Wang, Y., Cao, Y. and Xu, C. (2016), 'Weighted doubly regularized support vector machine and its application to microarray classification with noise', *Neurocomputing* **173**, 595–605.

Li, L., Ma, Y. and Zhang, J. (2022), 'Integrating multi-omics data for cancer prognosis prediction using feature selection techniques', *Nature Communications* **13**(1), 1346.

Lissek, V. and Suchan, B. (2021), 'Preventing dementia? interventional approaches in mild cognitive impairment', *Neuroscience & Biobehavioral Reviews* **122**, 143–164.

Liu, C., Zhang, S. and Li, T. (2022), 'Lasso regression for identifying genetic markers associated with alzheimer's disease progression', *Frontiers in Neuroscience* **16**, 731482.

Liu, H., Yu, L. and Liu, H. H. (2005), 'Feature selection with competitive swarm optimizer', pp. 329–336.

Liu, J., Pan, Y., Li, M., Chen, Z., Tang, L., Lu, C. and Wang, J. (2018), 'Applications of deep learning to MRI images: A survey', *Big Data Min. Anal.* **1**(1), 1–18.

Liu, L., Zheng, T. and Zhang, X. (2020), 'Deep metric learning for alzheimer's disease diagnosis using fmri and clinical data', *Journal of Alzheimer's Disease* **75**(4), 1019–1032.

Liu, X., Zhang, Y., Li, Z. and Wang, J. (2018), 'Feature selection for alzheimer's disease diagnosis using wrapper methods', *Journal of Alzheimer's Disease* **62**(4), 1325–1337.

Liu, Y., Wu, X. and Zhang, H. (2019), 'Multimodal metric learning for alzheimer's disease classification', *IEEE Transactions on Medical Imaging* **38**(10), 2205–2214.

Lu, D., Popuri, K., Ding, G. W., Balachandar, R. and Beg, M. F. (2018), 'Multimodal and multiscale deep neural networks for the early diagnosis of alzheimer's disease using structural mr and fdg-pet images', *Scientific reports* **8**(1), 5697.

Lundberg, S. M. and Lee, S.-I. (2017), A unified approach to interpreting model predictions, *in* 'Advances in Neural Information Processing Systems', Vol. 30.

Lunnon, K., Sattlecker, M., Furney, S. J., Coppola, G., Simmons, A., Proitsi, P., Lupton, M. K., Lourdusamy, A., Johnston, C., Soininen, H. et al. (2013), 'A blood gene expression marker of early alzheimer's disease', *Journal of Alzheimer's Disease* **33**(3), 737–753.

Lyu, Y., Yu, X., Zhang, L. and Zhu, D. (2021), Classification of mild cognitive impairment by fusing neuroimaging and gene expression data: classification of mild cognitive impairment by fusing neuroimaging and gene expression data, *in* 'Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference', pp. 26–32.

Mandel, J., Ray, D. K. and Rao, P. A. S. N. (2015), 'Feature selection using backward elimination and stepwise methods', *Journal of Machine Learning* **12**(3), 445–457.

Mayblyum, D. V., Becker, J. A., Jacobs, H. I., Buckley, R. F., Schultz, A. P., Sepulcre, J., Sanchez, J. S., Rubinstein, Z. B., Katz, S. R., Moody, K. A. et al. (2021), 'Comparing pet and mri biomarkers predicting cognitive decline in preclinical alzheimer disease', *Neurology* **96**(24), e2933–e2943.

Michalewicz, Z. (1995), *Genetic Algorithms + Data Structures = Evolution Programs*, Springer.

Mila-Aloma, M., Brinkmalm, A., Ashton, N. J., Kvartsberg, H., Shekari, M., Operto, G., Salvadó, G., Falcon, C., Gispert, J. D., Vilor-Tejedor, N. et al. (2021), 'Csf synaptic biomarkers in the preclinical stage of alzheimer disease and their association with mri and pet: a cross-sectional study', *Neurology* **97**(21), e2065–e2078.

Mitchell, A. J. and Shiri-Feshki, M. (2009), 'Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies', *Acta psychiatrica scandinavica* **119**(4), 252–265.

Morris, J. C. (1993), 'The clinical dementia rating (cdr): current version and scoring rules', *Neurology* **43**(11), 2412–2414.

Nag, K. and Pal, N. R. (2015), 'A multiobjective genetic programming-based ensemble for simultaneous feature selection and classification', *IEEE transactions on cybernetics* **46**(2), 499–510.

Ng, A. et al. (2011), 'Sparse autoencoder', *CS294A Lecture notes* **72**(2011), 1–19.

Ngiam, J., Koh, P., Chen, Z. and Ng, D. R. K. J. (2011), 'Multimodal deep learning', *Proceedings of the 27th International Conference on Machine Learning (ICML)* pp. 689–696.

Prince, M., Knapp, M., Guerchet, M., McCrone, P., Prina, M., Comas-Herrera, A., Wittenberg, R. and Salimkumar, A. (2014), Dementia UK: update, PhD thesis, King's College London.

Qi, S., Sui, J., Pearlson, G., Bustillo, J., Perrone-Bizzozero, N. I., Kochunov, P., Turner, J. A., Fu, Z., Shao, W., Jiang, R. et al. (2022), 'Derivation and utility of schizophrenia polygenic risk associated multimodal mri frontotemporal network', *Nature communications* **13**(1), 4929.

Qin, Z., Liu, Z., Guo, Q. and Zhu, P. (2022), '3d convolutional neural networks with hybrid attention mechanism for early diagnosis of alzheimer's disease', *Biomedical Signal Processing and Control* **77**, 103828.

Rallabandi, V. S. and Seetharaman, K. (2023), 'Deep learning-based classification of healthy aging controls, mild cognitive impairment and alzheimer's disease using fusion of mri-pet imaging', *Biomedical Signal Processing and Control* **80**, 104312.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016), "why should i trust you?" explaining the predictions of any classifier, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 1135–1144.
**URL:** *https://dl.acm.org/doi/10.1145/2939672.2939778*

Romeu, P., Zamora-Martínez, F., Botella-Rocamora, P. and Pardo, J. (2015), Stacked denoising auto-encoders for short-term time series forecasting, *in* 'Artificial Neural Networks: Methods and Applications in Bio-/Neuroinformatics', Springer, pp. 463–486.

Rosen, W. G., Mohs, R. C. and Davis, K. L. (1984), 'A new rating scale for alzheimer's disease', *The American Journal of Psychiatry* **141**(11), 1356–1364.

Roweis, S. T. and Saul, L. K. (2000), 'Nonlinear dimensionality reduction by locally linear embedding', *Science* **290**(5500), 2323–2326.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986), 'Learning representations by back-propagating errors', *nature* **323**(6088), 533–536.

Saeys, Y., Inza, I. and Larranaga, P. (2007), 'A review of feature selection techniques in bioinformatics', *Bioinformatics* **23**(19), 2507–2517.

Schmidt, M. (1996), *Rey Auditory Verbal Learning Test: A Handbook*, Western Psychological Services.

Schölkopf, B., Smola, A. and Müller, K.-R. (1997), Kernel principal component analysis, *in* 'Lecture Notes in Computer Science', Lecture notes in computer science, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 583–588.

Schroff, F., Kalenichenko, D. and Philbin, J. (2015), 'Facenet: A unified embedding for face recognition and clustering', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 815–823.

Schultz, M. and Joachims, T. (2003), 'Learning a distance metric from relative comparisons', *Advances in Neural Information Processing Systems (NIPS)* **16**, 41–48.

Seo, S. W., Gottesman, R. F., Clark, J. M., Hernaez, R., Chang, Y., Kim, C., Ha, K. H., Guallar, E. and Lazo, M. (2016), 'Nonalcoholic fatty liver disease is associated with cognitive function in adults', *Neurology* **86**(12), 1136–1142.

Song, C., Shi, J., Zhang, P., Zhang, Y., Xu, J., Zhao, L., Zhang, R., Wang, H. and Chen, H. (2022), 'Immunotherapy for alzheimer's disease: Targeting $\beta$-amyloid and beyond', *Translational neurodegeneration* **11**(1), 18.

Song, J., Zheng, J., Li, P., Lu, X., Zhu, G. and Shen, P. (2021), 'An effective multimodal image fusion method using mri and pet for alzheimer's disease diagnosis', *Frontiers in digital health* **3**, 637386.

Sullivan, P. F., Fan, C. and Perou, C. M. (2006*a*), 'Evaluating the comparability of gene expression in blood and brain', *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **141**(3), 261–268.

Sullivan, P. F., Fan, C. and Perou, C. M. (2006*b*), 'Evaluating the comparability of gene expression in blood and brain', *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **141B**(3), 261–268.

Tenenbaum, J. B., de Silva, V. and Langford, J. C. (2000), 'A global geometric framework for nonlinear dimensionality reduction', *Science* **290**(5500), 2319–2323.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**(1), 267–288.

Tosic, M., Bakshi, A. and Jain, A. K. (2019), 'Wrapper-based feature selection for face recognition systems', *Pattern Recognition Letters* **123**, 97–104.

Vafaie, H. and DeJong, D. (1995), Genetic algorithms as a tool for feature selection in machine learning, *in* 'Proceedings of the Ninth IEEE International Conference on Tools with Artificial Intelligence', pp. 400–406.

Venugopalan, J., Tong, L., Hassanzadeh, H. R. and Wang, M. D. (2021), 'Multimodal deep learning models for early detection of alzheimer's disease stage', *Scientific reports* **11**(1), 3254.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A. and Bottou, L. (2010), 'Stacked denoising autoencoders: Learning useful representations in

a deep network with a local denoising criterion.', *Journal of machine learning research* **11**(12).

Vinyals, O., Blundell, C., Lillicrap, T. and Wierstra, D. (2015), 'Matching networks for one shot learning', *Advances in Neural Information Processing Systems (NIPS)* **29**, 3631–3639.

Wang, D. and Gu, J. (2018), 'Vasc: dimension reduction and visualization of single-cell rna-seq data by deep variational autoencoder', *Genomics, Proteomics & Bioinformatics* **16**(5), 320–331.

Wang, M., Shao, W., Hao, X. and Zhang, D. (2021), 'Identify complex imaging genetic patterns via fusion self-expressive network analysis', *IEEE Transactions on Medical Imaging* **40**(6), 1673–1686.

Wang, T., Li, C. and Sun, M. (2021), 'Biomarker discovery for colorectal cancer using feature selection and machine learning algorithms', *Bioinformatics* **37**(6), 883–890.

Wang, Y., Li, X. and Ruiz, R. (2018), 'Weighted general group lasso for gene selection in cancer classification', *IEEE transactions on cybernetics* **49**(8), 2860–2873.

Wattmo, C., Blennow, K. and Hansson, O. (2020), 'Cerebro-spinal fluid biomarker levels: phosphorylated tau (t) and total tau (n) as markers for rate of progression in alzheimer's disease', *BMC neurology* **20**, 1–12.

Whitley, D. (1991), 'The genitor algorithm and selection pressure: Why rank-based allocation of reproductive trials is best', pp. 116–121.

Wittenberg, R., Hu, B., Barraza-Araiza, L. and Rehill, A. (2019), 'Projections of older people with dementia and costs of dementia care in the united kingdom, 2019–2040', *London: London School of Economics* **2**.

World Health Organization (2022), *Ageing and health*, World Health Organization.

Xia, Y., Liu, H. and Zhang, Z. (2021), 'Metric learning for cancer classification based on gene expression data', *BMC Bioinformatics* **22**(1), 29.

Xia, Y., Liu, X. and Zhang, Y. (2019), 'Gene selection for cancer classification: A comprehensive review', *Computational Biology and Chemistry* **78**, 136–145.

Xie, R., Wen, J., Quitadamo, A., Cheng, J. and Shi, X. (2017), 'A deep auto-encoder model for gene expression prediction', *BMC genomics* **18**, 39–49.

Yang, H., Zhang, Z., Xu, J. and Wang, Y. (2020), 'Feature selection based on filter methods and its application to lung cancer classification', *Scientific Reports* **10**(1), 11207.

Yilmaz, Y. and Ozdogan, O. (2009), 'Liver disease as a risk factor for cognitive decline and dementia: an under-recognized issue', *Hepatology* **49**(2), 698–698.

Zanetti, O., Solerte, S. B. and Cantoni, F. (2009), 'Life expectancy in alzheimer's disease (ad)', *Archives of gerontology and geriatrics* **49**, 237–243.

Zhang, C. and Bom, S. (2021), Auto-encoder based model for high-dimensional imbalanced industrial data, *in* 'International Conference on Neural Information Processing', Springer, pp. 265–273.

Zhou, H., Zhang, Y., Zhang, Y. and Liu, H. (2019), 'Feature selection based on conditional mutual information: minimum conditional relevance and minimum conditional redundancy', *Applied Intelligence* **49**, 883–896.

Zhou, L., Wang, F. and Zhou, H. (2023), 'Random forest-based feature selection in cancer gene expression data analysis', *BMC Bioinformatics* **24**, 175.

Zhou, T., Liu, M., Thung, K.-H. and Shen, D. (2019), 'Latent representation learning for alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data', *IEEE transactions on medical imaging* **38**(10), 2411–2422.

Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320.