Self-supervised Feature Extractor Training for Alzheimer's Disease Classification

A thesis

submitted in partial fulfilment

of the requirements for the Degree

of

Doctor of Philosophy in Computer Science

at

The University of Waikato

by

Chen Zheng

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2024

Abstract

Deep learning achieves encouraging performance in natural image classification. It has huge potential for detecting neural degenerative anomalies, but this is often limited by the availability of well-segmented neuroimages and computational resources. One way to address this problem is to apply self-supervised learning methods that utilise unsegmented neuroimages and artificial labels. Another solution is developing surrogate tasks to learn the representations of neuroimages for classification. This thesis reports the investigations of different variants of self-supervised learning and pretext tasks to train feature extractors for downstream Alzheimer's Disease classification.

Firstly, this thesis reviews the literature regarding Alzheimer's Disease classification and possible data leakage issues. Then, a lightweight 3D CNN-based ensemble is trained to predict brain age using the 3D MRI data of cognitively normal subjects from the OASIS-3 dataset. The extracted features are evaluated in the binary classification of CN vs. AD patients from their brain MRI scans. This approach achieved competitive performance compared with state-of-the-art methods in the literature.

The next part of this thesis developed four different self-supervised learning pretext tasks for feature extractor training: brain age prediction, brain sMRI reconstruction, brain sMRI rotation classification, as well as a combination of all three approaches into one single multi-task predictor. To further explore the feasibility of employing synthetic neuroimaging data in the self-supervised learning setting, the proposed approaches are trained on the LDM100K dataset followed by evaluation using real-world OASIS and ADNI datasets.

The real-world data training and testing leads to the best classification performance. The random cropping data augmentation technique can improve feature extractor training on 3D MRI data. Due to high computational expense and time limitations, the results of the training using synthetic data are not as satisfactory as those using real-world data. Future research is needed to develop more advanced feature extractor architectures and more complex pretext tasks that can learn more discriminative features. Another area of research to improve training efficiency would involve developing specialised software and hardware for processing 3D neuroimaging data.

Acknowledgements

Words cannot express my gratitude to Prof. Bernhard Pfahringer and Dr. Michael Mayo for their invaluable guidance and patience. They have been incredibly helpful in providing feedback on my work. They have helped me to improve my writing and my presentation skills. They have also been patient with me when I have made mistakes. They have taught me so much about machine learning and about how to be a better researcher.

I am also grateful to my lab mates for their feedback, suggestions, and moral support. I have had many interesting discussions with them, particularly HongYu Wang, Tim Leathart, and Attaullah Sahito. I want to express my heartfelt gratitude to Henry Gouk for the stimulating conversations. Their insights have been invaluable to me.

I want to thank everyone in the computer science department for making my time enjoyable. Thanks should also go to the librarians from the university.

Lastly, I would thank my family, especially my parents, for their love and support throughout my life. They have always been there for me, no matter what. Their support for me has kept my spirits and motivation high during this process. They have made it possible for me to achieve my goals.

Contents

Chapter 1

Introduction

1.1 Alzheimer's Disease

Age is one of the primary risk factors for neurodegenerative diseases, especially dementia and its most common form, Alzheimer's Disease (AD) [92]. AD causes an irreversible progressive impairment of memory and cognitive functions that considerably disrupts a patient's daily life. Because of the ageing population, the cost of care for AD is expected to increase from the current 47 million to 152 million by 2050 [6] just in the US. In view of the lack of progress in developing effective treatment for AD [92] and the rapidly increasing costs of medical care and associated socioeconomic impact, defeating AD is a priority for science and society.

In the context of neurodegenerative diseases, particularly AD, brain changes are consistently found in the form of atrophy (e.g. tissue loss or shrinkage) compared to a population of healthy brains, as shown in Fig. 1.1. Well-segmented data is often unavailable in the medical domain, particularly in the field of AD neuroimaging [64, 42]. The main reason is that the AD disease procession is not limited to specific brain regions or sites [11]. Therefore AD neuroimaging data are often only categorised into possible diagnostic groups in clinical practice. There are four groups: CN, cognitive normal or healthy controls; EMCI, early mild cognitive impairment; LMCI, late mild cognitive impairment; AD,

Alzheimer's disease.

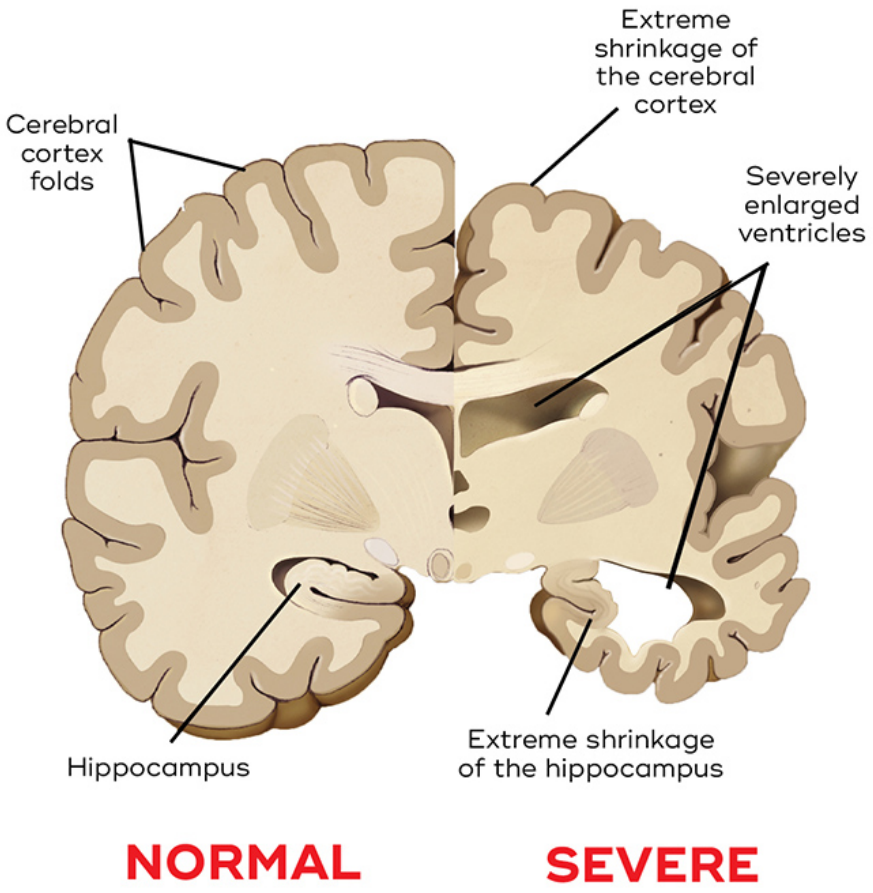


Figure 1.1: Comparison of brain tissue between a normal brain (left) and severe brain atrophy caused by Alzheimer's Disease (right) from a coronal (frontal) point of view. This figure is copied from website [77].

Due to human brain inaccessibility, structural magnetic resonance imaging (MRI) is a widely used neuroimaging technique to assess an individual's brain. MRI is a tomographic technology that produces two-dimensional images that consist of individual slices of the brain. The MRI scanner can only measure the signals in one plane, thus a coordinate system consists of three planes to describe the standard anatomical position of a human. Fig. 1.2 shows the three imaging planes of a 3D MRI scan.

The task of differentiating subjects with AD from cognitively normal (CN) subjects is popular in the literature. As shown in Fig. 1.3, before the development of AD, subjects go through an early stage called mild cognitive impair-

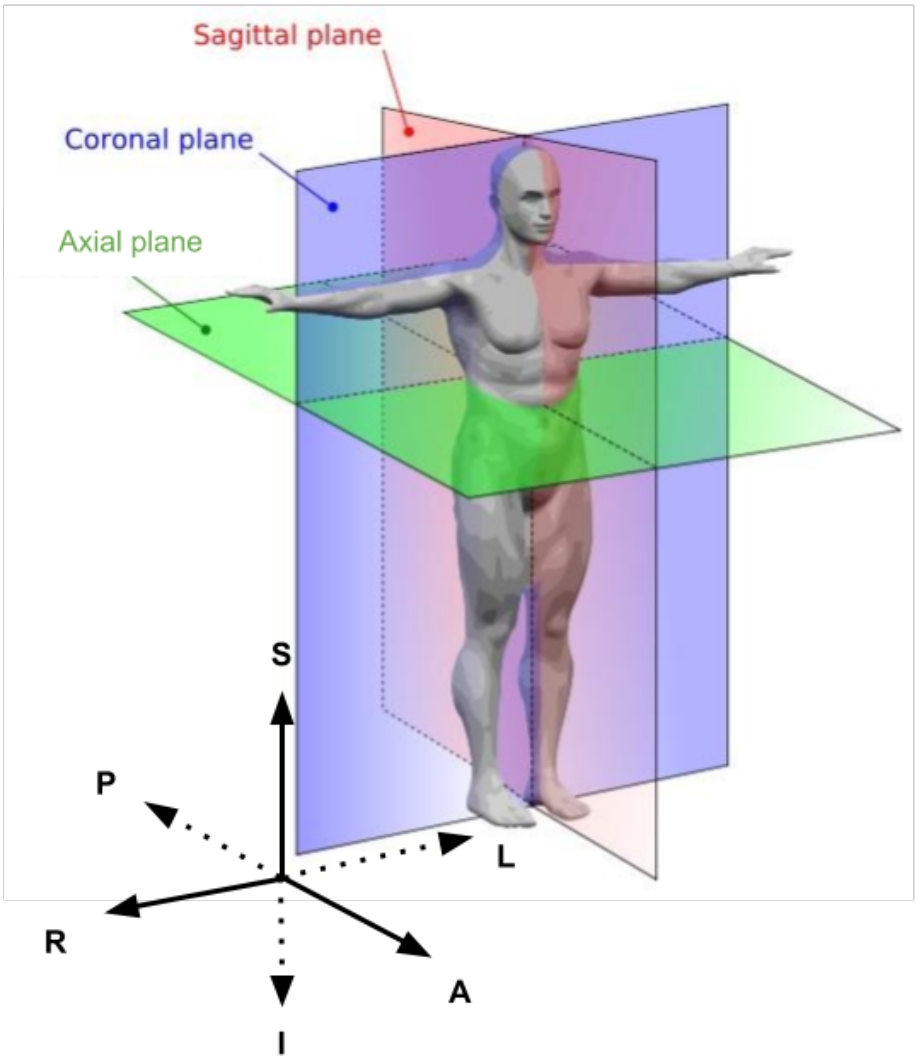


Figure 1.2: An MRI is defined by the plane (direction) of the image that is taken. Typically, three planes are used to describe the standard anatomical position of a human body [65]. The basic orientation terms for an MRI with respect to the human body: from the inferior (I) to superior (S) is the axial plane; from the left (L) to right (R) is the sagittal plane; and from the anterior (A) to posterior (P) is the coronal plane. Figure source [4].

ment (MCI) during which their cognitive functions deteriorate. Subjects with MCI may remain cognitively stable or subsequently progress to AD. Therefore, identifying MCI subjects from CN subjects is another task of interest.

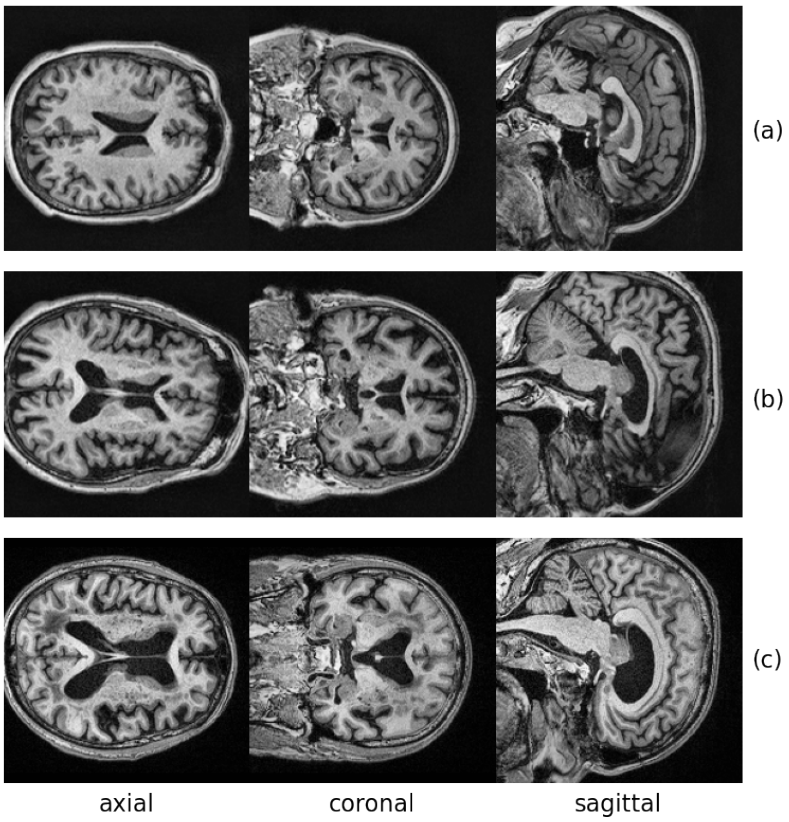


Figure 1.3: Three examples of structural MRI scans slices of subjects from the OASIS-3 [42] dataset in three planes: axial, coronal and sagittal. Row (a) shows images from a healthy subject, whereas row (b) comes from a MCI subject with obvious brain changes in all planes. Lastly, row (c) reveals the severe brain tissue loss of an Alzheimer's subject.

Inspired by the success of natural image classification, many publications proposed end-to-end convolutional neural networks (CNN) to classify brain MRI scans into AD or CN. To utilise 3D MRI data, slice-level, patch-level and scan-level approaches are widely used. Slice-level and patch-level approaches are computationally efficient, but only scan-level methods are able to take

advantage of the full 3D spatial information. The state-of-the-art results are very promising, however, a significant proportion of the studies performed a biased evaluation. Their data leakage issues will be discussed in the next chapter.

1.2 Neuroimaging and Computer Science

Automated medical image analysis systems have been built since it has been possible to scan and load medical images into a computer. From the 1970s to the 1990s, sequential application of low-level pixel processing (edge and line detector filters, region growing) and mathematical modelling (fitting lines, circles and ellipses) were used for medical image analysis [48]. In the same period, artificial intelligence concepts were implemented with many if-then-else statements to form rule-based image processing systems [79].

At the end of the 1990s, supervised machine learning techniques using training data to build medical image analysis systems began increasing in popularity. Examples include the use of active shape models for segmentation [21], the employment of atlas methods to differentiate new data from training data [43], and the concept of pattern recognition [52] and use of statistical classifiers [5, 61] for computer-aided detection and diagnosis. Many commercially available medical image analysis systems [15, 53, 73] are based on such pattern recognition or machine learning approaches.

Although there is a shift from systems that are completely designed by humans to systems that are trained by computers using example data from which features are extracted, the extraction of discriminant features from the images is still done by human researchers [48]. Conventionally, constructing such machine learning systems requires meticulous engineering and considerable domain knowledge to transform the raw data (i.e. the pixel values of an image) into discriminative features, such as the shape, colour, and/or texture roughness as well as their combinations. The learning system then determines

the optimal decision boundary in the feature space so that it can detect patterns (e.g. tumor tissue) in the raw data. However, this transformation relies on handcrafted features, thus the conventional machine-learning techniques are limited in their ability to process raw data. Logically, one possible next step is to let the computer learn such features for discrimination directly from raw data.

1.3 Deep Learning

Recently, the Machine Learning (ML) field has received enormous attention due to the exciting breakthrough of the Deep Learning technique[55]. Deep learning is a sub-field of machine learning, but it is different from traditional machine learning in that features are learned from the raw data. Deep learning aims to discover high-level data representations by utilising hierarchical architectures [44]. For example, a cat's different combinations of shape, colour, and texture can be represented at a higher abstract level. Deep learning computational models consist of multiple processing layers to learn such representations of data with multiple levels of abstraction. With the composition of enough such layers, appropriate discriminative features can be learned as high-level representations in a latent space. The reason is that higher levels of representation amplify aspects of the input that are important for discrimination and suppress irrelevant variations [44].

Deep learning-based algorithms and techniques gained significant attention when they started outperforming other approaches on the ImageNet classification benchmark in 2012 [41] due to using progressively deeper networks. Deep learning methods even exceed human performance in natural image classification to a level that the error rate is close to the Bayes rate [41]. Henceforth, deep learning is the state-of-the-art approach for a wide variety of computer vision problems.

There are generally two categories of neural networks: feedforward and re-

current networks. A feedforward network (e.g. convolutional neural networks) is a network that contains inputs, outputs, and hidden layers. The inputs only travel in the forward direction. Each neuron (node) computes the output based on the weighted sum of its inputs. Then the output is fed into the next processing layer as inputs. This feedforward computation continues until the final layer and determines the output of the network. Feedforward networks are often used in classification tasks.

In contrast, a recurrent network has feedback paths. This allows data to travel in both directions between the layers as well as each connection between neurons. Due to the looping nature, recurrent networks aim to reach a state of equilibrium by continuously changing themselves based on the inputs. When the input changes, the network tries to achieve a new state of equilibrium. This type of network is widely used in optimisation problems.

1.3.1 Supervised Learning

The majority of modern deep learning architectures are based on feedforward neural networks. They are built upon many layers of non-linear processing units for feature extraction. Data is fed into the input layer and propagates through the network to reach a final output. An error value is obtained by comparing the network output against the expected value. The training procedure propagates the error backwards and updates the weights of the network to minimise the error [45]. This supervised learning/training process is repeated a given number of epochs to optimise the weights.

Supervised learning is the most common training technique when the data are well-labelled. The idea is to learn the relevance of different features by fitting the known outcome. Supervised learning can be grouped into two main types: classification and regression. A classification task uses algorithms to classify data into particular categories (e.g. email vs. spam), whereas regression tasks aim to predict a numerical output (e.g. temperature forecast).

1.3.2 Self-supervised Learning

CNN-based classifiers in the literature show promising performance in some tasks. However, achieving such performance requires a vast amount of well-labelled data. The collection of labelled data can be costly and extremely difficult or even impossible to obtain. Therefore, self-supervised learning techniques can be very beneficial.

Self-supervised learning belongs to the unsupervised learning method family which can be applied to suitable data formats (e.g. images) where labels are not available. Generally, computer vision pipelines that employ self-supervised learning involve performing two tasks, a pretext task and a target task. The target task can be anything like a classification or detection task that requires a vast amount of well-labelled data. The pretext task is the self-supervised learning task solved to learn visual representations, with the aim of using the learned representations or model weights obtained in the process, for the target task. This is based on the assumption that the unlabelled data can be represented in a semantic and structurally meaningful way by learning the pretext task.

A pretext task is designed by fabricating artificial labels from unlabelled data without human annotation. Predicting rotations is one of the most popular pretext tasks which has a simple and straightforward architecture and requires minimal sampling. For example, the rotations of $0, 90, 180, 270$ degrees can be applied to the image and a network is trained to predict such rotations. Equivalently, the network performs a 4-way classification to identify the rotation. Then this network or its weights can be passed to the downstream (target) task.

1.4 Contributions

The main thesis hypothesis is that the appropriate application of self-supervised learning can obtain superior deep learning-based feature extractors that im-

prove the classification accuracy of Alzheimer's Disease neuroimages. Based on the hypothesis, this thesis investigates feature extractor training approaches based on (a) brain neuroimage age prediction, (b) brain neuroimage rotation classification, (c) brain neuroimage reconstruction, and (d) multi-head training.

The main contribution of the thesis is as follows:

- Proposed a 3D CNN-based brain neuroimage age prediction approach as a pretext task to train feature extractors.
- Proposed a 3D CNN-based brain neuroimage rotation classification approach as a pretext task to train feature extractors.
- Proposed a 3D CNN-based AutoEncoder for brain neuroimage reconstruction approach as a pretext task for feature extractor training.
- Proposed self-supervised single-head and multi-head training approach for brain neuroimage feature extractors including reconstruction and rotation classification.
- Application of 3D Latent Diffusion Model-generated high-resolution brain MRI data comprising 100k subjects to train 3D CNN-based feature extractors.
- The extracted features are then evaluated in the binary classification of CN vs AD patients from their brain MRI scans. The evaluation of the above feature extractors uses a training approach based solely on subject-wise train-test-splits. The unbiased results show promising performance of the self-supervised approaches.

The proposed 3D CNN-based brain age prediction approach [97] has been published in the 2022 International Joint Conference on Neural Network, Italy. We are working on publishing the other approaches.

1.5 Thesis Outline

This thesis is structured as follows:

Chapter 2 introduces the fundamental blocks to build neural networks and the loss functions and the optimisation methods for training. This section also describes the neuroimage datasets used in this study.

Chapter 3 reviews the literature related to training feature extractors for neuroimage classification using deep learning.

Chapter 4 presents an approach that uses CNN-based brain age prediction as a pretext task for feature extractor training.

Chapter 5 explores the idea of using synthetic neuroimaging data for feature extractor training.

Chapter 6 investigates an AutoEncoder-based feature extractor training approach as a pretext for AD classification.

Chapter 7 investigates the possibility of using brain sMRI rotation classification as a pretext task for feature extractor training.

Chapter 8 explores the possibility of feature extractor training by utilising the multi-head configuration of brain age prediction, brain rotation classification, and brain image reconstruction.

Chapter 9 firstly concludes the contribution of this thesis, followed by a summary of the experimental results from the previous chapters. Then this section provides recommendations for further research that could build on this work.

Chapter 2

Background

This chapter starts by revisiting the basic concepts of deep learning, followed by a brief overview of the building blocks of artificial neural networks (ANN). The next section reviews the loss functions and optimisation methods for ANN training. The last section describes the datasets used in this study.

2.1 Artificial Neural Network Architectures

Artificial neural networks are the backbone of deep learning algorithms and techniques. The most common ANN is the feedforward architecture. In the context of supervised learning or self-supervised learning, ANNs are trained to approximate some function f. For example, an ANN can map an input x to a target y, and is parameterised by θ. This ANN defines a mapping $y = f(x; \theta)$ and learns the optimal values of the parameters θ to minimise the loss. *theta* is often randomly initialised [28] in the literature. The training process adjusts θ by backpropagating a gradient value with respect to the calculated loss between the target and the output.

In practice, a network is normally composed of many layers. For example, functions $f^{(1)}$, $f^{(2)}$ and $f^{(3)}$ can be chained together as $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$. In this case, function $f^{(1)}$ is called the input layer, whereas functions $f^{(3)}$ is referred as the output layer, and $f^{(2)}$ is the hidden layer.

2.1.1 Convolutional Layers

The convolutional layers are the keystone of neural networks to operate on image data. Their purpose is to extract elementary visual features such as edges, endpoints, and corners. The most common type of convolution that is used is the 2D convolution layer which is usually abbreviated as conv2D. A kernel in a conv2D layer "slides" over the 2D input data, in 2-direction (x, y), performing an element-wise multiplication. Finally, the products are summed up into a single output pixel. The kernel will perform the same operation for every location it slides over, transforming a 2D matrix of features into a different 2D matrix of features. An example of 2D convolution is shown in Fig. 2.1.

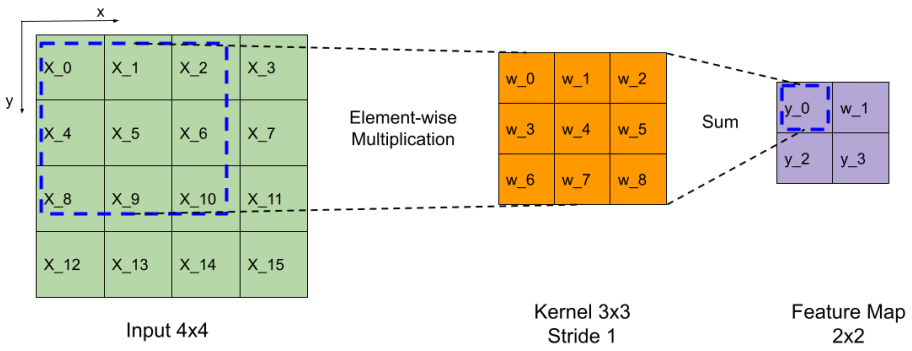


Figure 2.1: Examples of 2D convolution operation. A kernel size of 3×3 is applied to an 4×4 image, producing a 2×2 feature map.

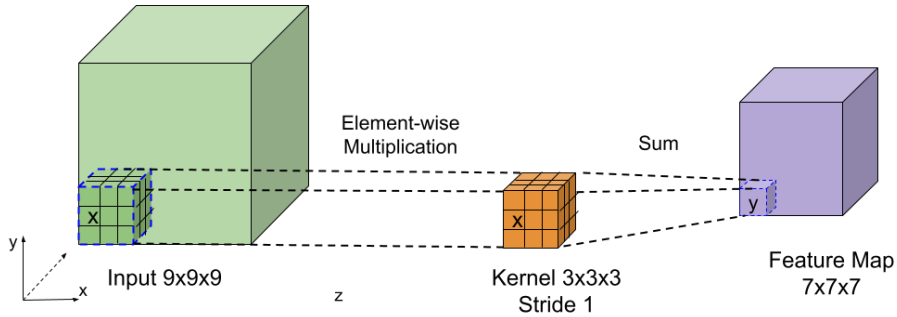


Figure 2.2: Examples of 3D convolution operation. A kernel size of $3 \times 3 \times 3$ is applied to an $4 \times 4 \times 4$ image, producing a $2 \times 2 \times 2$ feature map.

The convolution operation can be used with three-dimensional data. 3D convolution kernels move in 3-direction (x, y, z) to calculate the convolutional output which produces a 3D volume as a result of the operation. The idea is to preserve spatial information between pixels in the MRI data, or temporal information between frame stacks of video data. An example of 3D convolution is shown in Fig. 2.2.

2.1.2 DeConvolutional Layers

DeConvolution layers [94] (also called transposed-convolution layers) are similar to convolutional layers, but fairly different in operations. Unlike convolution layers which concentrate the information of various pixels into a single pixel, deconvolution layers spread the information from one pixel to various pixels. As shown in Fig. 2.3, zero padding is added to the original input followed by a convolution operation. This thesis employs 3D deconvolution, with an example provided in Fig. 2.4.

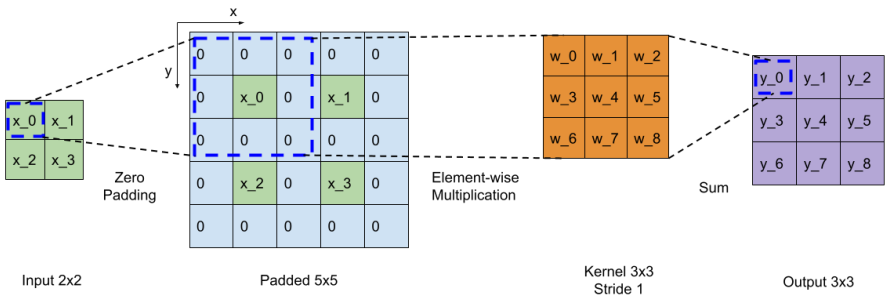


Figure 2.3: Examples of 2D deconvolution operation. A kernel size of 3×3 is applied to an 4×4 image, producing a 2×2 feature map.

2.1.3 Batch Normalisation

Training Deep Neural Networks is a complicated task. The problem is that the distribution of each layer's input changes because the parameters of the previous layers change [32]. This requires lowering the learning rate and metic-

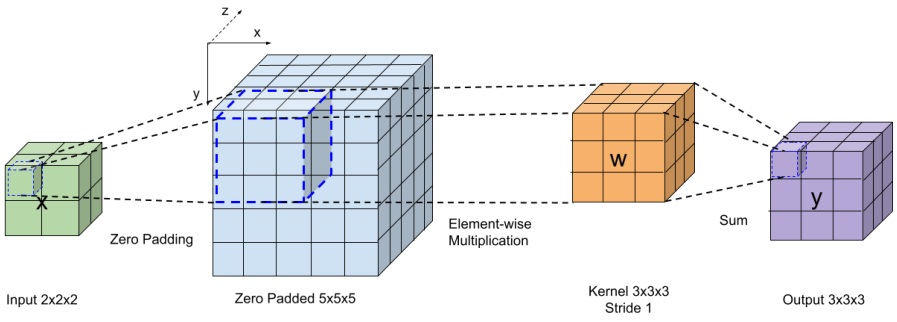


Figure 2.4: Examples of 3D deconvolution operation. A kernel size of $3 \times 3 \times 3$ is applied to an $4 \times 4 \times 4$ image, producing a $2 \times 2 \times 2$ feature map.

ulous parameter initialisation which makes training deep networks notoriously time-consuming.

To accelerate deep neural network training, [32] proposed a batch normalisation technique. It is implemented in the form of a batch normalisation layer that is part of the deep architecture. The idea is to apply normalisation to the activation x over a mini-batch followed by scaling and shifting, which can be expressed as:

$$y = \frac{x - \mu[x]}{\sqrt{\sigma[x] + \epsilon}} * \gamma + \beta \tag{2.1}$$

where μ is the mini-batch mean, σ is the mini-batch variance, ϵ is a constant added to the mini-batch variance for numerical stability, and γ and β are the learnable scaling and shifting parameters, respectively. There are also alternatives like LayerNorm Ba et al. [7] that are well suited for sequence models such as transformers. They are not considered in this thesis.

2.1.4 Pooling Layers

Pooling layers are normally added after convolution layers for downsampling. The idea is to reduce the dimensions of hidden layers by combining the outputs of the previous layer into a single neuron in the next layer. The most common ones are max pooling and average pooling are shown in Fig. 2.5. This thesis

uses the 3D version pooling layers that are shown in Fig. 2.6.

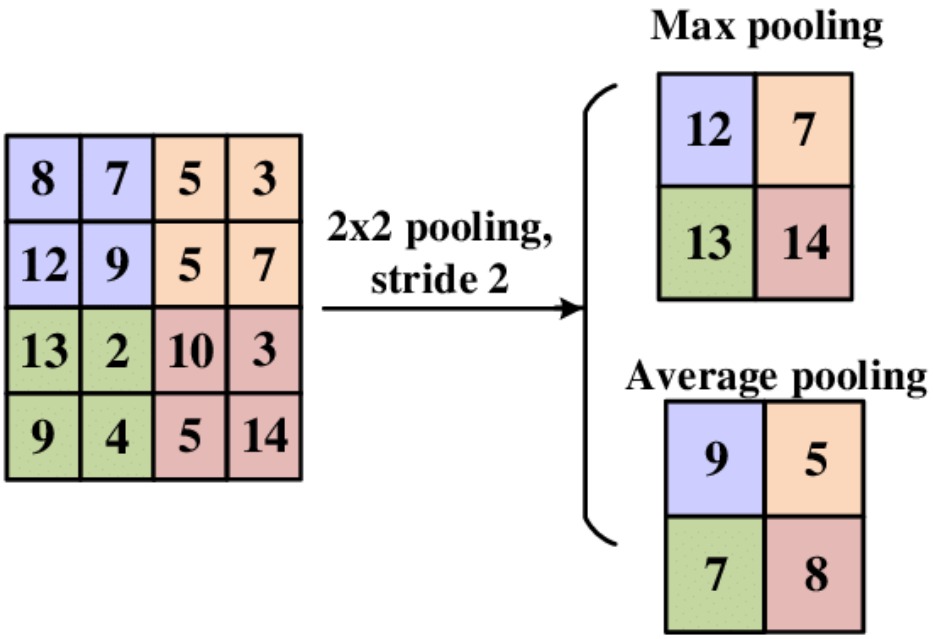


Figure 2.5: Examples of max pooling and average pooling with a stride of 2. The max pooling operation passes the maximum value within a 2×2 kernel, whereas the average pooling produces the mean within a 2×2 kernel.

2.1.5 Fully-Connected Layers

Feedforward networks often use a fully-connected layer as the output layer that can be written as:

$$f(x; W, b) = x^T W + b \tag{2.2}$$

where W is the weight matrix and b is the bias vector. The size of W and b is defined according to the input and output dimensions. For example, input dimension 4 and output dimension 5 of a fully-connected layer require a weight matrix with dimensions 5 by 4.

2.1.6 Activation Function

To introduce nonlinearities, non-linear activation functions are used in the networks. The most common one is the rectified linear unit (ReLU). The idea

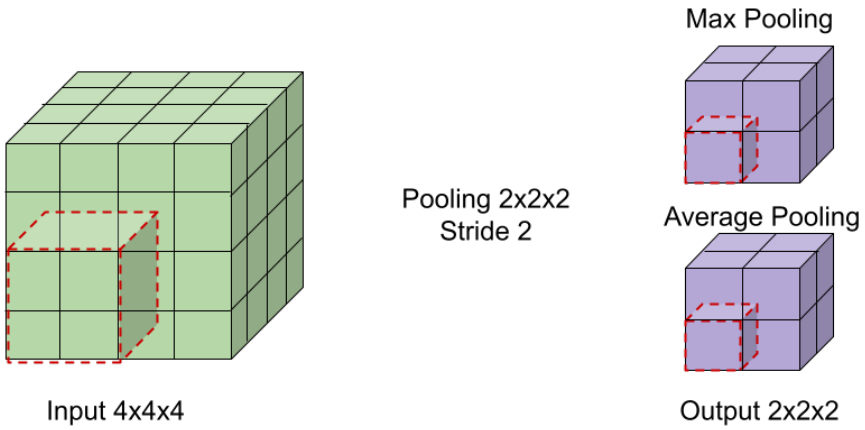


Figure 2.6: Examples of max pooling and average pooling with a stride of 2. The max pooling operation passes the maximum value within a 2×2 kernel, whereas the average pooling produces the mean within a 2×2 kernel.

is to convert the negative inputs to zero therefore not activating such neurons. It is defined as:

$$\text{ReLU}(x) = max(0, x) \tag{2.3}$$

where the input x represents element-wise input. One major benefit is the reduced likelihood of the gradient vanishing [13]. This arises when the values of gradients are smaller than one. The gradients are multiplied in backpropagation to get the gradients of lower layers. The effect of multiplying the gradients makes the value of gradients to be even smaller for lower layers, leading to a very small change or even no change in the weights of lower layers. Therefore, the deeper the network, the more the effect of vanishing gradients. This makes learning per iteration slower when activation functions that suffer from vanishing gradients such as sigmoid and tanh activation functions. On the other hand, the gradient of a ReLu function is either 0 for $x <= 0$ or 1 for $x > 0$. Therefore, there is no need to worry about the depth of layers as the gradients will neither vanish nor explode when multiplied.

Depending on the task, the choice of activation function greatly impacts the training efficiency and predictive performance. For the classification tasks, the softmax function is widely used as the final layer to transform the raw

outputs of the network into a vector of probabilities, essentially a probability distribution over the input classes. It is defined as:

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}} \tag{2.4}$$

where C is the number of classes and the label is generally given to the class with the highest probability.

2.2 Loss Functions

This section discusses the most widely used loss functions for regression, reconstruction, and classification that this thesis used.

2.2.1 Mean Absolute Error Loss

This is the most widely used function to measure the absolute error of each network output and target value. It can be expressed as:

$$\text{MAE}(y, \hat{y}) = \frac{\sum_{i=0}^{N-1} |y_i - \hat{y}_i|}{N} \tag{2.5}$$

where N is the number of data samples, $\hat{y}$ is the network output and y is the target value.

2.2.2 Categorical Cross-Entropy Loss

Softmax cross-entropy is the most frequently used classification loss function which measures the difference between the expected probability distribution and the predicted probability distribution. It can be written as:

$$\text{CE}(y, \hat{y}) = -\sum_{i=1}^{C} y_i \cdot log(\hat{y}_i) \tag{2.6}$$

where C is the number of classes, $\hat{y}_i$ is the output of the softmax activation value (probability) for the i^{th} class and y_i is the corresponding ground truth probability. The value is usually 1 for the correct class and 0 for all other classes.

2.3 Training & Optimisation

This section briefly describes the backbone algorithms for training neural networks and optimisation methods to improve training efficiency.

2.3.1 Backpropagation

Training a neural network means adjusting the weights of the model in such a way that the loss of the overall dataset is minimised. Backpropagation [81], short for "backward propagation of errors", is an algorithm for supervised learning of artificial neural networks using gradient descent. The "backward" part of the name stands for the gradient calculation that proceeds backwards through the network, with the gradient of the final layer of weights being calculated first and the gradient of the first layer of weights being calculated last. Partial computations of the gradient from one layer are reused in the computation of the gradient for the previous layer. This backward flow of the error information allows for efficient computation of the gradient at each layer versus the naive approach of calculating the gradient of each layer separately.

Given an artificial neural network f and a loss function $\mathcal{L}$, the method calculates the gradient of the error function with respect to the neural network's weights θ, which can be written as:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} [\mathcal{L}(f(x_i; \theta), y_i)] \tag{2.7}$$

where N is the number of of training samples, x_i is the feature vector of the i^{th} sample and y_i is the corresponding target label.

Using gradient descent, the updating rule for network parameters θ_{t+1} at iteration t can be described as:

$$\theta_{t+1} = \theta_t - \alpha \nabla \mathcal{L}(f(x; \theta_t), y) \tag{2.8}$$

where α is the user-defined learning rate to control the step size of parameter updating, whereas $\alpha \nabla \mathcal{L}(f(x; \theta), y)$ is the gradient that combines the partial derivatives of each input-output pair over the loss function with respect to

the network parameters. The network parameters θ then can be updated by gradient descent. This process is repeated until a local minimum is found or some convergence criterion is met.

2.3.2 Optimisation Methods

The vanilla gradient descent method requires the gradient from all training samples. This is computationally infeasible when the training set is very large. Therefore, stochastic optimisation methods are employed to update network parameters. Practically, a large dataset is often broken down into smaller mini-batches and then fed into the network. The gradient is calculated per mini-batch and then the parameters are updated accordingly. This is known as Stochastic Gradient Descent (SGD) and has proven to converge at a sublinear rate [87].

To accelerate the speed of convergence speed, [67] proposed a momentum mechanism for the updating rule. The idea is to accumulate the previous gradient as momentum and update the network parameters by adding the momentum, which can be expressed as:

$$M_t = \gamma M_{t-1} + \alpha \nabla \mathcal{L}(f(x; \theta_t), y) \tag{2.9}$$

$$\theta_{t+1} = \theta_t + M_t \tag{2.10}$$

where M_{t-1} is the accumulated momentum and γ is the momentum scale which is often set to a large value such as 0.9.

Fine-tuning the global learning rate for the optimisation algorithms can be very difficult in certain tasks, thus, adaptive learning rate optimisation methods are preferred. Different from keeping track of the sum of gradients (momentum), the Adaptive Gradient algorithm [19], or "AdaGrad" for short, keeps track of the sum of squared gradients and uses that to adapt the gradient for different features. The idea is that the more a feature has been updated already, the less chance it needs an update in the future, thus giving other features a higher chance to be updated. The update rule for AdaGrad can be

expressed as:

$$V_t = \sqrt{\sum_{i=1}^{t} \nabla\mathcal{L}(f(x;\theta_t), y)^2} \tag{2.11}$$

$$\theta_{t+1} = \theta_t - \alpha \frac{\nabla\mathcal{L}(f(x;\theta_t), y)}{V_t + \epsilon} \tag{2.12}$$

where V_t is the accumulated sum of gradients of parameters θ at iteration t and ϵ is a smoothing term (often set to $1e^{-5}$) to avoid division by zero.

However, AdaGrad is incredibly slow because the sum of the squared gradients only grows and never shrinks. The Root Mean Square Propagation (RMSProp) algorithm fixes this issue by adding a decay factor to V_t, which can be written as:

$$V_t = \sqrt{\beta V_{t-1} + (1-\beta)\sum_{i=1}^{t} \nabla\mathcal{L}(f(x;\theta_t), y)^2} \tag{2.13}$$

$$\theta_{t+1} = \theta_t - \alpha \frac{\nabla\mathcal{L}(f(x;\theta_t), y)}{V_t + \epsilon} \tag{2.14}$$

where β is the decay rate. The idea is that only recent $gradient^2$ values matter, and the ones from long ago are basically forgotten. The decay rate also helps with the scaling of gradients so that the gradients will have a smaller chance to overshoot near minima.

Adam [38] (short for Adaptive Moment Estimation) optimisation algorithm takes the best of both worlds of momentum and root mean square mechanism. It can be written as:

$$M_t = \beta_1 M_{t-1} + (1-\beta_1)\nabla\mathcal{L}(f(x;\theta_t), y) \tag{2.15}$$

$$V_t = \sqrt{\beta_2 V_{t-1} + (1-\beta_2)\sum_{i=1}^{t} \nabla\mathcal{L}(f(x;\theta_t), y)^2} \tag{2.16}$$

$$\theta_{t+1} = \theta_t - \alpha \frac{M_t}{V_t + \epsilon} \tag{2.17}$$

where β_1 is the decay rate for the sum of gradients (momentum), commonly set at 0.9; β_2 is the decay rate for the sum of squared gradients, commonly set at 0.999; α is the learning rate, and it is commonly set between 0.001 and 0.0001. Adam empirically works well [38], and thus in recent years, it is commonly the go-to choice for deep learning problems.

Adam has an inherent problem with the $L2$ regularisation term. In deep learning libraries such as PyTorch [71], Tensorflow [1] or Keras [16], instead of modifying the loss function, the $L2$ regularisation is implemented by adding the sum of all parameters squared ($\lambda \sum_{i=1}^{W} w_i^2$) to the loss function. The user-defined hyperparameter λ is a positive real number. The idea is to reduce each network parameter by an amount proportional to its current value during the optimisation step. $L2$ regularisation is also referred to as weight decay.

However, if one adds the weight decay term at this point, then moving averages of the gradient M and its square V keep track not only the values of the gradients of the loss function but also additionally the values of the regularisation term. To address the ineffective handling of weight decaying, AdamW [54] was proposed, where weight decay is only performed when updating the parameters:

$$\theta_{t+1} = \theta_t - \alpha \frac{M_t}{V_t + \epsilon} + \alpha\lambda \sum_{i=1}^{W} w_i^2 \tag{2.18}$$

Where W represents the total number of parameters in the network. The idea is that the weight decay term does not end up in the exponential moving averages and is thus only proportional to the weight itself. The authors experimentally demonstrated that AdamW yields better training loss and that the models generalise much better than models trained with Adam. Therefore, this thesis chose AdamW as the optimiser for all the training tasks.

2.4 Data Augmentation Techniques

Data augmentation is a technique of artificially increasing the training set by creating modified copies of a dataset using existing data. This can be done by applying minor changes to the data, such as flipping, rotating, or cropping. Data augmentation can be used to improve the performance of machine learning models by making them more robust to noise and variations in the data. It is especially useful for tasks where the amount of available data is limited. Fig. 2.7 depicts some of the most common data augmentation

techniques.

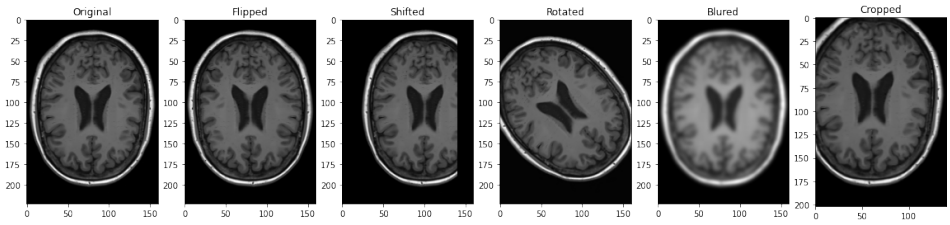


Figure 2.7: An example of data augmentation techniques including flipping, shifting, rotation and blurring. The original image is the No. 80 slice in the axial plane of subject sub-000000 from the LDM100K dataset.

2.5 Magnetic Resonance Imaging

As explained by the Medical School of Case Western Reserve University [76]:

> MRI is based on the magnetisation properties of atomic nuclei. A powerful, uniform, external magnetic field is employed to align the protons that are normally randomly oriented within the water nuclei of the tissue being examined. This alignment (or magnetisation) is next perturbed or disrupted by the introduction of an external Radio Frequency (RF) energy. The nuclei return to their resting alignment through various relaxation processes and in so doing emit RF energy. After a certain period following the initial RF, the emitted signals are measured. Fourier transformation is used to convert the frequency information contained in the signal from each location in the imaged plane to corresponding intensity levels, which are then displayed as shades of grey in a matrix arrangement of pixels. By varying the sequence of RF pulses applied & collected, different types of images are created. Repetition Time (TR) is the amount of time between successive pulse sequences applied to the same slice. Time to Echo (TE) is the time between the delivery of the RF pulse and the receipt of the echo signal.
>
> The most common MRI sequences are T1-weighted and T2-weighted scans. T1-weighted images are produced by using short TE and TR times. The contrast and brightness of the image are predominately determined by T1 properties of tissue. Conversely, T2-weighted images are produced by using longer TE and TR times. In these images, the contrast and brightness are predominately determined by the T2 properties of tissue.

> As shown in Fig. 2.8, T1- and T2-weighted images can be easily differentiated by looking at the cerebrospinal fluid (CSF). CSF is dark on T1-weighted imaging and bright on T2-weighted imaging. A third commonly used sequence is the Fluid Attenuated Inversion Recovery (Flair). The Flair sequence is similar to a T2-weighted image except that the TE and TR times are very long. By doing so, abnormalities remain bright but normal CSF fluid is attenuated and made dark. This sequence is very sensitive to pathology and makes the differentiation between CSF and an abnormality much easier.

This thesis used T1-weighted (T1w) structural MRI scans as it is the most commonly available neuroimaging modality in the OASIS-3 dataset. Due to availability issues, there are other types of MRI imaging are not considered in this work.

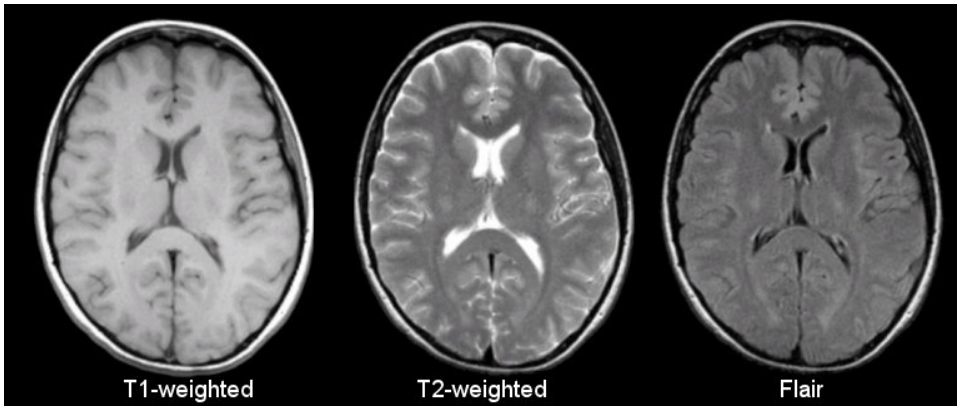


Figure 2.8: A comparison of T1w vs. T2w vs. Flair neuroimaging of the same subject. This figure is copied from [76].

2.6 Datasets

This section describes the publicly available brain neuroimaging datasets for Alzheimer's disease (AD) that were used in this thesis, including two real-world datasets: OASIS and ADNI and a synthetic dataset LDM100K. Both the OASIS and ADNI focus on Alzheimer's disease, but none of them has detailed segmentation for the disease lesion, which would indicate the regions

in an organ or tissue that have suffered damage caused by the disease. A brief summary of the datasets is shown in Table.2.1.

Dataset	**Classes**	**#Subjects**	**#Scans**	$MMSE^{*}$	CDR^{**}	**Sex**	**Age**
ADNI	AD	192	530	23.3 +/- 2.1	2	Male 1,426 Female 1,316	55 - 94
	MCI	398	1126	27.0 +/- 1.8	1, 0.5		
	CN	229	877	29.1 +/- 1.0	0		
OASIS-3	AD	104	237	15.0 +/- 5.8	2	Male 487 Female 611	42 - 95
	MCI	389	1131	24.9 +/- 3.3	1, 0.5		
	CN	605	2021	29.1 +/- 1.2	0		
LDM-100k	CN	100,000	100,000	N/A	N/A	N/A	44 - 82

Table 2.1: A brief summary of the datasets. *The Mini-Mental State Examination (MMSE) measures general cognitive status, with scores ranging from 0 (severe impairment to 30 (no impairment). **The Clinical Dementia Rating (CDR) is a numeric scale used to quantify the severity (stages) of dementia in clinical practice. Clinical Dementia Rating Assignment Qualitative equivalences are as follows: 0 is none; 0.5 is very mild; 1 is mild; 2 is moderate and 3 is severe.

2.6.1 ADNI-3

In 2004, the Alzheimer's Disease Neuroimaging Initiative (ADNI) [64] began as a longitudinal study to develop clinical, imaging, genetic, and biochemical biomarkers for AD. The long-term goal is to improve diagnostic methods for the early detection of AD. This cooperative study made a global impact by developing a set of standardised protocols that allows for the comparison of results across different medical centres and scanner manufacturers. Furthermore, ADNI's data-sharing policy significantly simplifies the data access procedure for qualified researchers worldwide.

ADNI includes participants between the ages of 55 and 90 across 57 sites in the United States and Canada. The participants undergo a series of initial tests that are repeated at intervals over subsequent years, including a clinical evaluation, neuropsychological tests, genetic testing, lumbar puncture, and MRI and PET scans.

After more than a decade of dedicated effort, ADNI released the phase 3 data in 2016 which consolidated the existing data from ADNI-1, ADNI-GO and ADNI-2. In total, there are 483 cognitively normal, 300 early mild cognitive impairments, 551 mild cognitive impairments, 150 late mild cognitive impairments and 437 AD subjects. An example from the ADNI dataset is shown in Fig. 2.9.

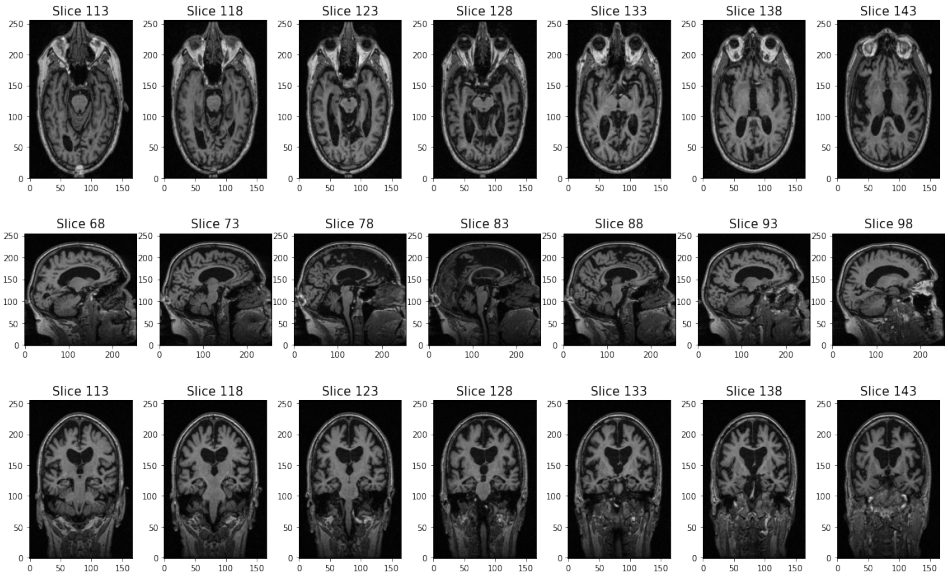


Figure 2.9: Slice samples of subject ID S15147 from the ADNI database. The overall shape of this MRI image is $166 \times 256 \times 256$.

2.6.2 OASIS-3

Open Access Series of Imaging Studies (OASIS) [42] is a project that aimed at making neuroimaging datasets freely available to the scientific community. The OASIS-3 dataset became available in early 2020. This thesis uses the OASIS-3 dataset as it is the newest at the time of writing.

OASIS-3 is a longitudinal multimodality neuroimaging dataset for normal ageing and Alzheimer's Disease. It is a retrospectively compiled dataset of 1378 participants that was collected across many projects over 30 years. Ranging from 42 to 95 years old, there are 755 cognitively normal adults and another 622 adults at different stages of cognitive decline. To protect privacy, each

participant was randomly given a subject identifier. Then all dates were normalised to reflect the number of days from the beginning of the study. Each subject recorded at least one magnetic resonance imaging (MRI) session. An example from the OASIS dataset is shown in Fig. 2.10.

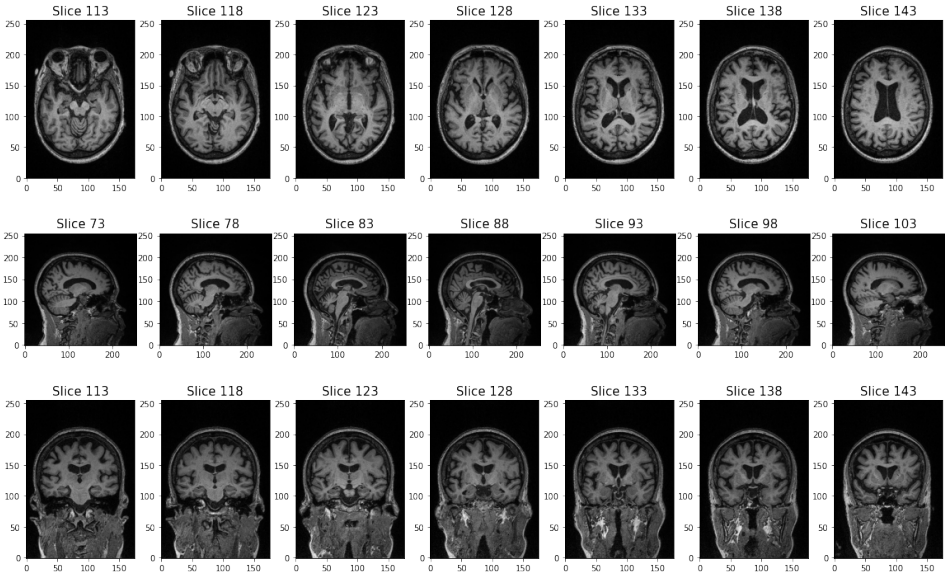


Figure 2.10: Slice samples of subject ID OAS30001 from the OASIS database. The overall shape of this MRI image is $176 \times 256 \times 256$.

OASIS-3 also includes the Clinical Dementia Rating (CDR) Score [63] for each subject upon each clinical assessment along with the participation. The real dates of clinical visits for CDR scoring are also removed and reformatted to show the days from the first visit. In most cases, the clinical visits and the MR sessions are not carried out on the same day.

2.6.3 LDM-100k

Given the data-hungry nature of deep learning-based approaches, generating synthetic data provides a promising alternative for training superior models. More medical data allows researchers to conduct experiments on a larger scale. Inspired by the photorealistic synthetic images produced by a latent diffusion model [80], the authors of [74] used T1w MRI images from the UK Biobank dataset [86] to generate brain MRI images.

Their proposed method selected 31,740 healthy individuals aged between 44 and 82 years from the UKBioBank database. There are 14,942 male subjects and 16,798 female subjects, respectively. They also utilised the volume of ventricular cerebrospinal fluid (min-max: 6,995.68 - 171,375.0 mm^3) and the brain volume normalised for head size (min-max: 1,144,240 - 1,793,910 mm^3).

The T1w MRI scans were used to train the models to learn about the probabilistic distribution of brain images, while conditioning on covariables including age, sex, and brain structure volumes. Then the authors evaluated the generated brain scans by computing the correlation between the obtained values and the inputted conditioning values for brain volume and brain age, resulting in a high correlation coefficient of $r = 0.972$ and $r = 0.692$, respectively.

The authors used the conditioning variables to control the data generation for a synthetic dataset with 100,000 brain scans. They made the dataset publicly available along with the conditioning information. An example from the LDM100K dataset is shown in Fig. 2.11.

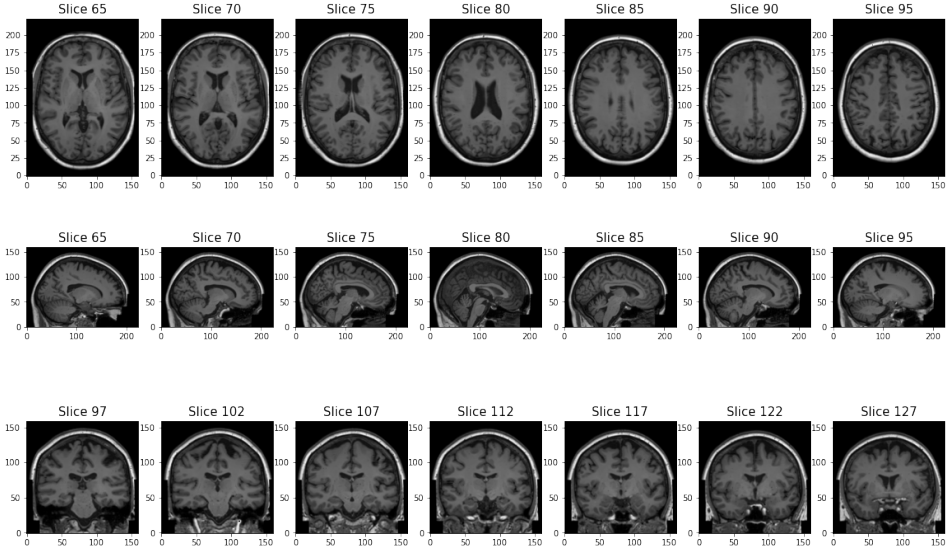


Figure 2.11: Slice samples of subject ID sub-000000 from the LDM100K database. The overall shape of this MRI image is $160 \times 224 \times 160$.

2.7 Evaluation Metrics

To measure the performance, we computed the true positive (TP), false positive (FP), true negative (TN), false negative (FN) and area under the receiver operating characteristic curve (AUC). The Classification accuracy (ACC) represents the proportion of correctly classified scans among the whole population. The sensitivity (SEN) shows the rate of being predicted positively when the disease is present, whereas the specificity (SPE) shows the rate of being predicted negatively when the disease is absent. The AUC represents the probability that a randomly selected positive scan is ranked more highly than a randomly selected negative scan. Table. 2.2 shows the metrics for evaluation.

In addition to the metrics used previously, Youden's J statistic (J_stat) is added to the list of metrics. As shown in Table. 2.2, J_stat summarises the performance of a diagnostic test. It combines sensitivity and specificity into a single measure (Sensitivity + Specificity - 1) and has a value between -1 and 1. An example of J_stat is given in Fig. 2.12. In a perfect test in which there are no false positives or false negatives, J_stat equals 1. Values close to 0 mean that the diagnostic test gives the same likelihood of positive results for groups with and without the disease, which renders the test not very informative. The J_stat value of -1 shows that the predictions are exactly opposite to the ground truth. In this case, a flip of the labels could resolve the issue.

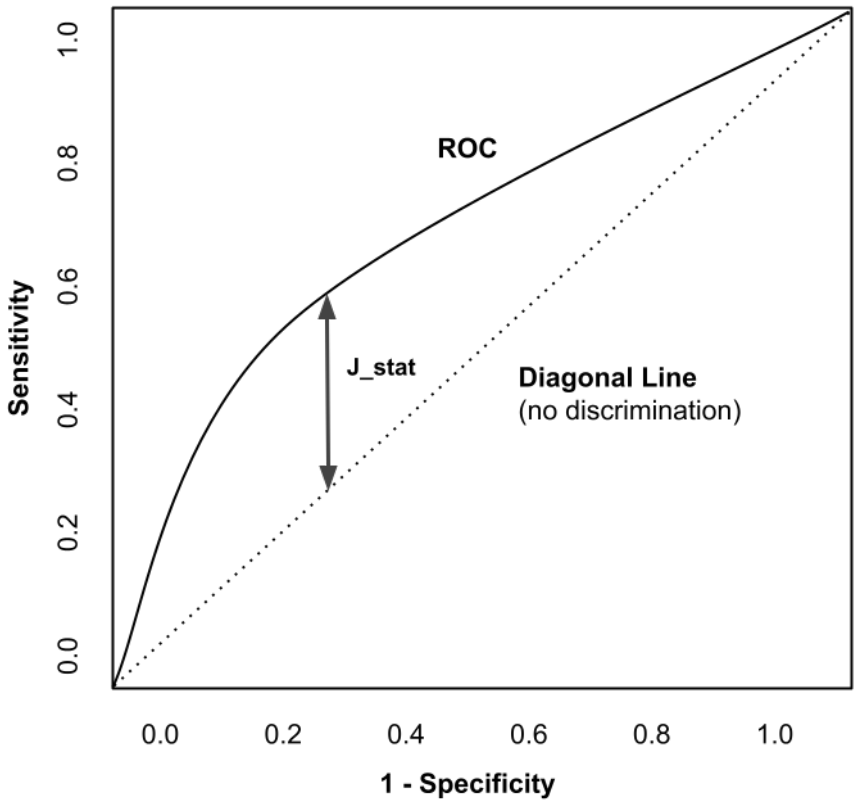


Figure 2.12: An example of J_stat. The vertical line represents the maximum value of Youden's index for the ROC curve. The diagonal line corresponds to the ROC curve of a classifier that predicts the class at random. The more area under the ROC curve indicates better discriminative performance of the classifier.

Table 2.2: Evaluation Metrics

Metric	***Formula***
$Accuracy(ACC)$	$\frac{TP+TN}{TP+TN+FP+FN}$
$Sensitivity(SEN)$	$\frac{TP}{TP+FN}$
$Specificity(SPE)$	$\frac{TN}{FP+TN}$
$Youden's Jstatistic(J\_stat)$	$\frac{TP}{TP+FN} + \frac{TN}{FP+TN} - 1$

Chapter 3

Related Work

The first section reviews the literature regarding Alzheimer's Disease classification. This section also shows the data leakage problem commonly found in related works. The next three sections review the development of brain age prediction, autoencoder-based brain image reconstruction and image rotation classification methods, respectively.

3.1 Alzheimer's Disease Classification

As shown in Fig.1.1 and Fig.1.3, the brain tissue loss is most obvious between the CN and AD classes, therefore, the class difference is most distinguishable on structural MRI scans. As a result, the classification between Alzheimer's Disease (AD) and cognitively normal (CN) subjects is the most studied task in the literature. Before the development of dementia, patients go through a phase called mild cognitive impairment (MCI) in which they experience the deterioration of certain cognitive functions, such as having difficulty reasoning. MCI patients can be subdivided into stable, progressive and converting stages or early and old stages, denoted as sMCI, pMCI or cMCI, EMCI and LMCI respectively. Identifying this cognitive stage of a patient is another task of interest.

As Structural MRI (sMRI) is the most available form of neuroimaging modality, this thesis mainly focuses on publications that use CNN-based meth-

ods on T1w data. It is worth noting that other modalities such as functional MRI and positron emission tomography (PET) are slowly gaining popularity in the literature due to their increasing availability in the datasets [42].

Inspired by the success of natural image classification using 2D CNNs (e.g. ResNet [30] and VGGNet [85]), extracting 2D slices from 3D MRI volumes is a popular choice of input in many studies. Many studies report state-of-the-art classification accuracy on various datasets. However, many of the proposed approaches are suspected to have data leakage or biased evaluation metrics. It is speculated that the authors are primarily from medical domains and are not experienced in machine learning [90]. The rapid adoption of machine learning methods in a certain domain could lead to errors. Moreover, the classification results are often unclear about slice-level or subject-level accuracy.

A very recent publication [36] highlights this issue which is apparently present in all application areas of machine learning. This thesis critically reviewed some recent studies using machine learning techniques in the field of Alzheimer's Disease classification. The common data leakage found in the literature are:

- Leakage A: Wrong Data Split. The data split is not performed on the subject-level when defining the train, test and validation subset, resulting in data from the same subject appearing in multiple subsets. This problem can occur when 2D patches or 2D slices are extracted from a 3D image, or when 3D images of the same subject are available at multiple time points.

- Leakage B: Late split. Procedures such as data augmentation, feature selection or pre-training tasks are performed before subject-level training, validation and testing splits. For example, the generated image of the same subject could appear in several subsets if the augmentation is performed before the subject-level split.

- Leakage C: Biased transfer learning. Transfer learning can cause data

leakage if the source and target domains overlap. For example, the CN subjects used in a pre-training task of an AutoEncoder are also used in the downstream task of CN vs. MCI or CN vs. AD classification.

3.1.1 Prior Work without Data Leakage

In the literature, there are publications that evaluated their approach without clear data leakage. For example, some 2D slice-level and 3D patch-level approaches performed train-test-split on the subject-level before extracting 2D slices, whereas some 3D volume-level studies only selected one scan per subject. A brief summary of each paper in this subsection is listed in Table.3.1.

Using predefined brain regions of interest (ROIs) is a popular information extraction choice to reduce the number of dimensions of the input. Aderghal et al. [2] proposed a 2D slice-based approach (Fig.3.1). From the ADNI database, they selected the screening T1w Structural MRI images of 188 AD, 228 CN and 399 MCI subjects. The 3D hippocampus ROI is firstly extracted from the MRI volume by a shape of $(28 \times 28 \times 28)$. Then 2D slices are extracted from the 3D ROI from the sagittal, coronal and axial planes. After that, the median slice of each plane was chosen and its two closest neighbours were also included as a shape of $(28 \times 28 \times 3)$. Each image undertakes a preprocessing of alignment, intensity normalisation and augmentation of flips, volume translations and blurring. Then their approach trained three independent 2D CNN for each plane followed by a majority vote of probability to predict the class. Their paper reported the binary classification results for AD vs. NC with 91.02% accuracy, 92.72% sensitivity, and 89.94% specificity. Although the authors did not explain the train-test-split procedure, the subject-level 3D ROI extraction is unlikely to have data leakage issues.

In their follow-up paper [3], a transfer learning approach (Fig.3.2) is proposed to pre-train the models on sMRI data to improve the performance on a different neuroimaging modality. This paper used the same set of subjects [2] for the pre-training task. The authors included an additional small number of

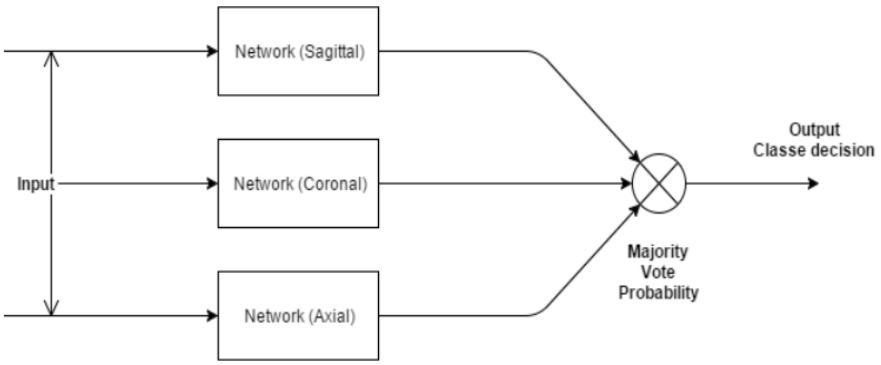


Figure 3.1: Three independent 2D CNNs for each plane followed by a majority vote of probability to predict the class. This figure is copied from [2].

subjects with both sMRI and Diffusion Tensor Imaging (DTI) for testing. The MRI volumes are processed in a similar pipeline with the addition of skull-stripping and brain tissue segmentation, followed by the same augmentation techniques of flips, volume translations and blurring. After that, the median slice of each plane was chosen with its two closest neighbours as a three-channel image ($28 \times 28 \times 3$). They trained the sMRI models using the same scheme as shown in Fig.3.1, and then the best-trained parameters are reused to initialise the DTI model and fine-tune it on the DTI dataset. The DTI classification achieved 92.5% accuracy, 94.7% sensitivity, and 90.4% specificity. The authors also reported poor performance while using random initialisation for the DTI model. Their train-test-split method is clearly explained as a subject-level split thus there is no data leakage issue.

The authors in [8] proposed a 3D CNN classifier achieving 90% accuracy on AD and CN classification. They selected 199 AD and 141 CN subjects from the ADNI dataset where each subject has multiple MRI scans. The pre-processing includes conforming, motion correction, non-uniform intensity normalisation, affine transformation, max-min intensity normalisation, and skull removal. They also reported that training their network without preprocessing yielded a drop in accuracy of 38%. Their proposed architecture is described in Fig.3.3. They clearly reported a subject-level data partition strategy to avoid

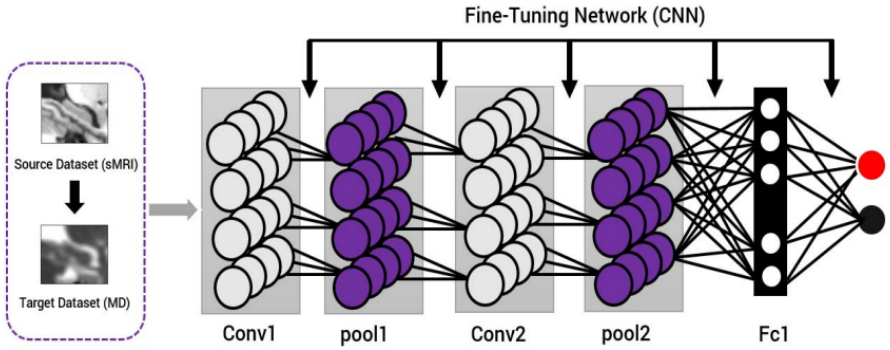


Figure 3.2: An overview of the transfer learning approach from sMRI to MD imaging. This figure is copied from [3].

data leakage.

Layer	Filter size	# Neurons	Output size
Conv1 + Stride-2 + ReLU	7x7x7x64		55x55x55
Conv2 + ReLU	3x3x3x64		55x55x55
Conv3 + ReLU + MaxPool	3x3x3x128		27x27x27
Conv4 + ReLU + MaxPool	3x3x3x128		13x13x13
Conv5 + ReLU +MaxPool	3x3x3x128		6x6x6
FC1 + ReLU		256	1x256
FC2 + ReLU		256	1x256
FC3 + SoftMax		2	1x2

Figure 3.3: An overview of the 3D Convolutional Network architecture in [8]. This figure is copied from [8].

[40] proposed two 3D CNN-based approaches, as shown in Fig.3.4. They proposed a VoxCNN (similar to VGG) and ResNet network with modified components for 3D inputs. They selected 50 AD, 43 LMCI, 77 EMCI and 61 CN subjects from the ADNI subject. Only the first sMRI image of each subject is selected followed by alignment, normalisation, and skull-stripping. The evaluation reported 79% and 80% AD vs CN accuracy for 3D VoxCNN and 3D ResNet, respectively. The authors are aware of the data leakage issues and performed subject-level splits. The size of the chosen data is relatively small compared to other studies.

Li et al. [46] proposed a classification method based on the combination of multi-model convolutional networks to learn the various features from brain

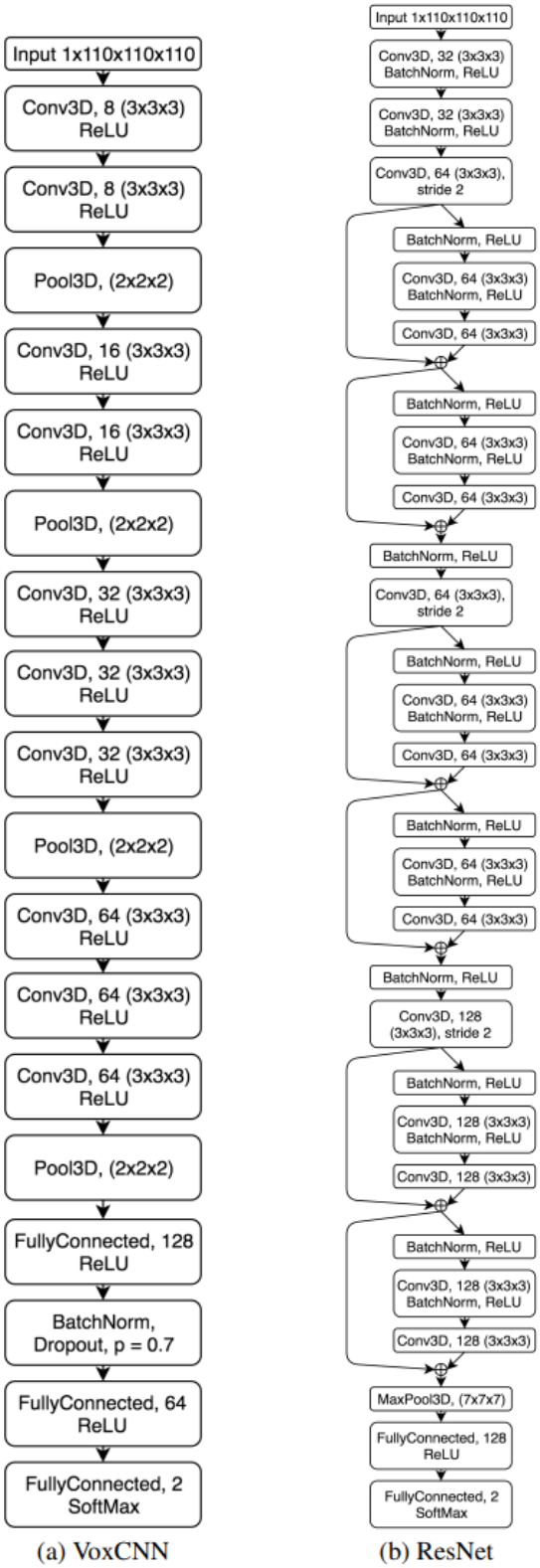


Figure 3.4: An overview the 3D VoxCNN and 3D ResNet architectures in [40]. This figure is copied from [40].

sMRI and classify AD and NC subjects. As shown in Fig.3.5, their approach consists of two feature extractors: part one is a deep 3D CNN, whereas part two has multiple 3D CNN AutoEncoders (AEs). The AutoEncoders are pre-trained for sMRI reconstruction followed by a fine-tuning for AD vs. CN classification. The input sizes are varied for the AEs to obtain multiple feature extractors. Then the features from both CNN and AEs are concatenated for fully connected layers and softmax activation. To evaluate their approach, they selected the T1w sMRI data of 199 AD and 229 CN subjects from the ADNI baseline visits. Each MRI image is processed for linear registration, intensity normalisation, skull-stripping and cerebellum-removal. The testing result showed 88.31% accuracy for AD vs. CN classification task. As the authors only used the baseline visit per subject, there is no data leakage detected.

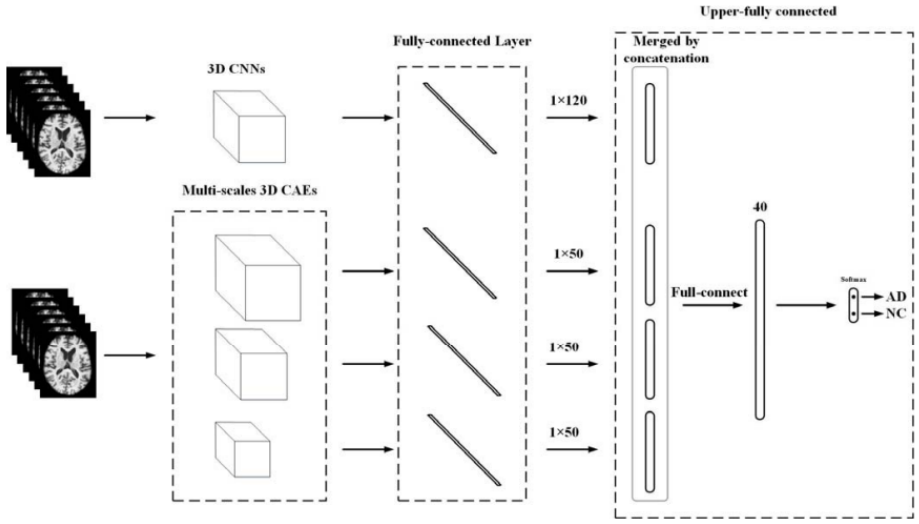


Figure 3.5: An overview of the multi-model convolutional networks in [46]. This figure is copied from [46].

Li et al. [47] proposed a classification framework by fusing multiple DenseNets to classify AD and NC subjects on T1w sMRI data, as shown in Fig.3.6. They selected 199 AD, 403 MCI and 229 NC subjects for evaluation. All MRI images are nonuniform intensity normalised, skull-stripped and cerebellum-removed, followed by a linear registration to align all the images. They partitioned each brain MRI image into a number of local regions and extract 3D patches

$32 \times 32 \times 32$ from each region for clustering. Then each cluster of 3D patches is fed into a 3D DenseNet followed by a concatenation for final prediction. The average classification performance of k-clustering is optimal with $K = 10$, resulting in 89.5% accuracy. Their method can effectively reduce the dimension of 3D MRI. Similar to their previous study [46], only the baseline scan of each subject is selected, therefore no data leakage is detected.

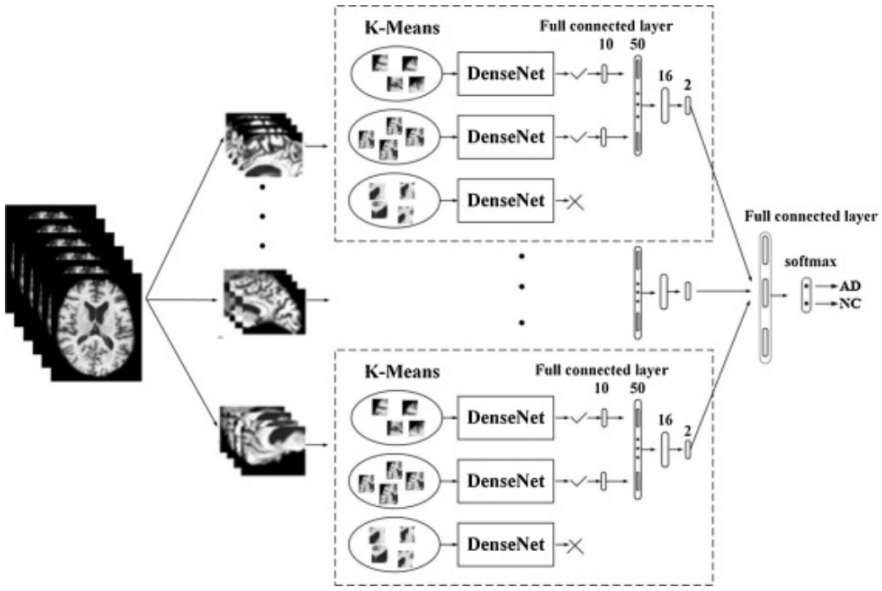


Figure 3.6: An overview of 3D patch k-clustering and 2D DenseNet ensemble for 3D sMRI classification in [47]. This figure is copied from [47].

A 3D multi-modality fusion approach is proposed in [50] for AD classification. They selected 93 AD, 76 MCI converters (pMCI) and 128 MCI non-converters (sMCI), 100 NC with both sMRI and PET data from the ADNI dataset to evaluate the proposed approach. All baseline visit T1w sMRI images were preprocessed by applying the typical bi-commissural line correction, skull-stripping, cerebellum removal, intensity normalisation and affine registration, whereas each PET image is registered to its corresponding sMRI orientation. A number of 27 3D local patches are uniformly extracted from each sMRI and PET image. As shown in Fig.3.7, each group of local patches is utilised to pre-train an independent 3D CNN, followed by a joint fine-tuning of 2D CNNs before feature fusion and final prediction. The fusion approach resulted in a

higher 93.26% accuracy compared to either single modality sMRI 84.97% and PET 88.08% in the AD vs CN task. This study only selected one image for each modality per subject, thus there is no risk of data leakage.

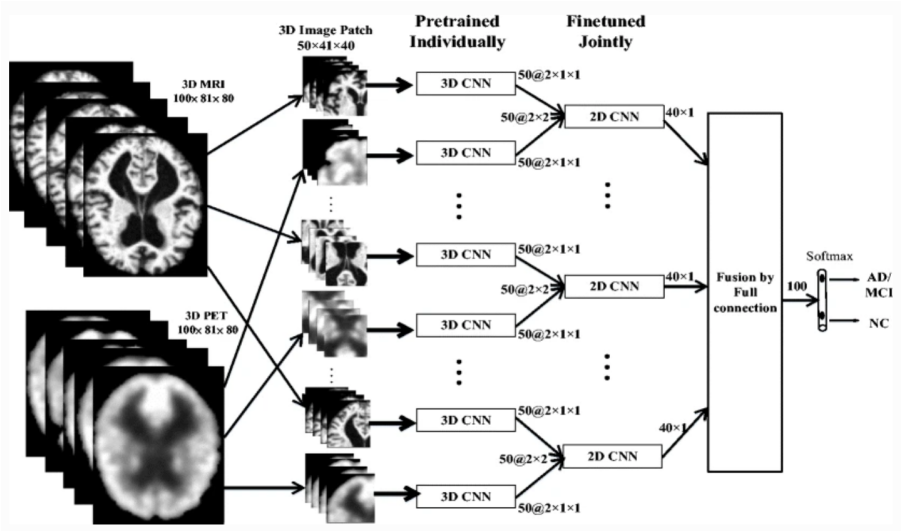


Figure 3.7: An overview of the multi-modality fusion approach for AD classification in [50]. This figure is copied from [50].

Senanayake et al. [83] proposed a multimodality fusion approach for AD classification. They utilised the 3D sMRI volumes and neuropsychological measurements (35 features) from the ADNI database. The resulting dataset has 515 MR volumes that belong to three classes: 161 AD volumes, 193 MCI volumes and 161 CN volumes. There is no proprocessing step explained in their paper. Their proposed approach used dilated convolutions, residual connections and dense connections to reduce the dimension of 3D MRI to a comparable dimension so the two feature vectors can be meaningfully merged together. They reported about 80% accuracy for AD vs CN classification. Although they clearly stated the subject-level train-test-split, they did not employ the cross-validation technique to demonstrate the robustness of their proposed approach.

Valliani and Soni [88] proposed an approach that uses pre-trained residual network models to classify AD. They selected 188 AD, 243 MCI and 229 CN subjects from the ADNI dataset. The authors clearly stated that only the

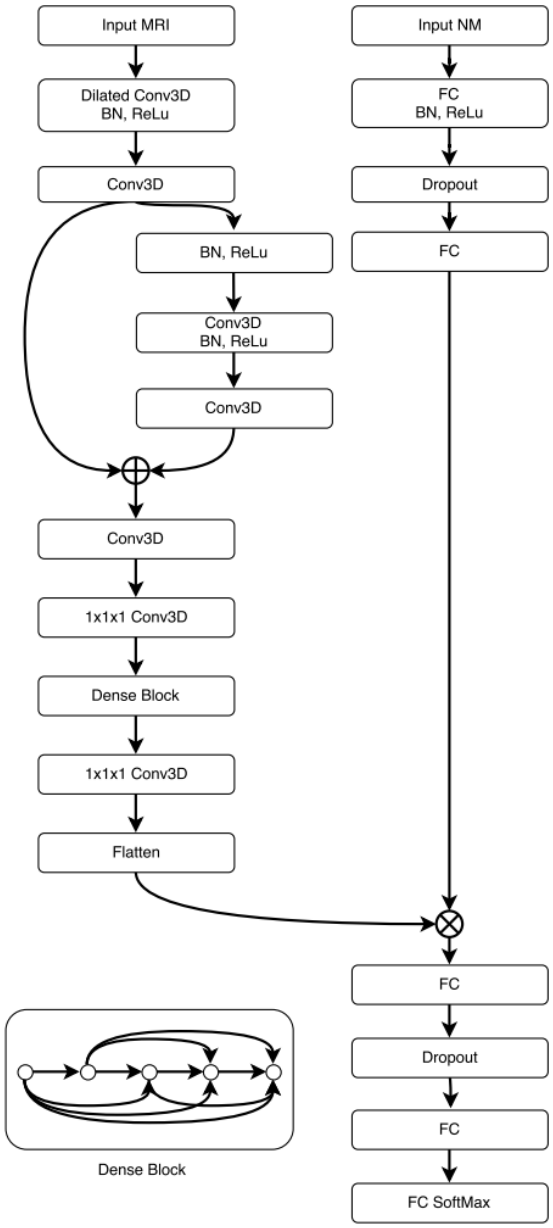


Figure 3.8: An overview of the fusion approach for AD classification in [83]. $\oplus$ represents addition operation and $\otimes$ represents concatenation operation. This figure is copied from [83].

median axial slice of the first sMRI from each subject is used to avoid data leakage, but they did not explain the preprocessing pipeline in the paper. Their proposed architecture consists of a ResNet-18 network, which is initialised using ImageNet weights. Then the network is fine-tuned on MRI data with on-the-fly affine transformation. Their evaluation showed 81.3% accuracy in the AD vs. CN scenario. Their approach only utilised one plane of the 3D MRI volume and a relatively shallow architecture for transfer learning, which underused the spatial information in the 3D volume.

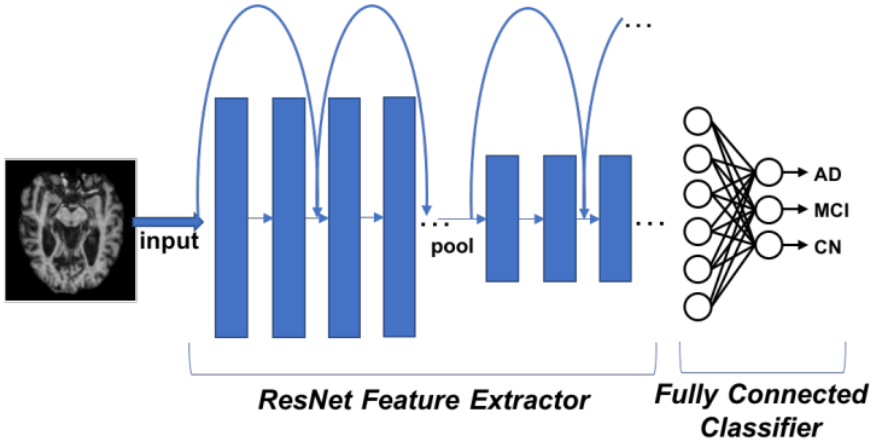


Figure 3.9: An overview of the transfer learning approach for AD classification using pre-trained ResNet-18 in [88]. This figure is copied from [88].

Oh et al. [68] proposed an unsupervised approach to extract discriminative features for AD classification on T1w sMRI data. Their approach first trains a 3D AutoEncoder to reconstruct AD and CN MRI images with minimal loss, then the encoder is fine-tuned for AD and CN classification. They also conducted transfer learning that uses weights of the AD/CN classifier to initialise the sMCI vs. pMCI classifier for later fine-tuning. To evaluate the approach, they selected the baseline scans of 198 AD, 166 pMCI, 101 sMCI and 230 CN subjects from the ADNI database. All MRI images are preprocessed including realignment, normalisation, and smoothing. The testing result showed 86.6%, 77.37% and 63.04% accuracy for AD vs. CN, pMCI vs. CN, and sMCI vs CN classification tasks, respectively. As they only selected one sMRI image from

each subject, there no risk of data leakage.

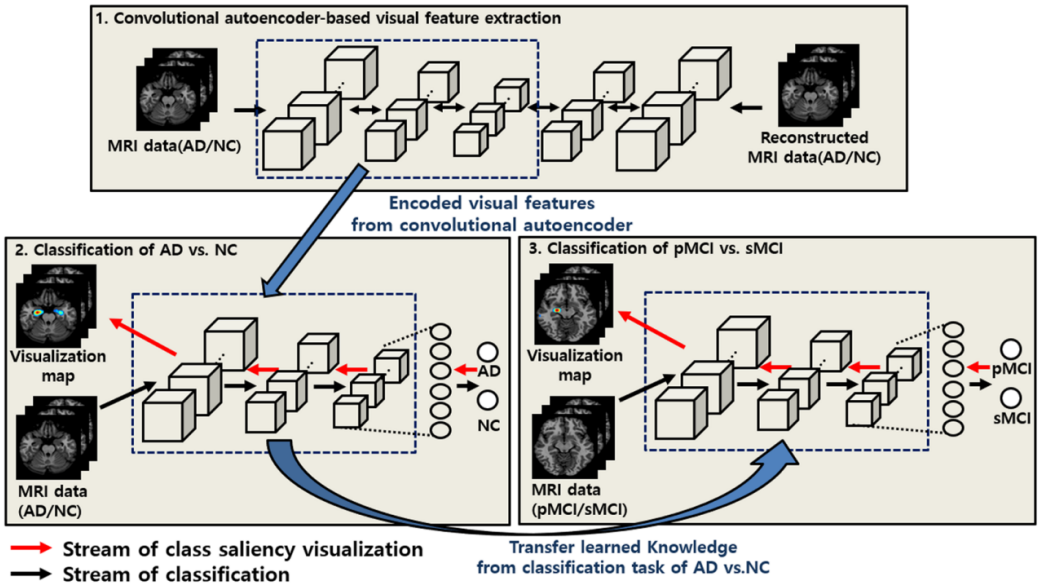


Figure 3.10: An overview of the AutoEncoder-based feature extraction approach for AD classification in [68]. This figure is copied from [68].

Liu et al. [51] proposed a method that combines two models to classify AD on T1w sMRI images. As shown in Fig.3.11, the first multi-task model consists of multiple residual and upsampling layers that generate a binary segmentation mask of the hippocampus region along with bottleneck features. The second model utilises a 3D patch from the mask to learn features and then concatenates with the bottleneck features for a final prediction. To evaluate their method, they randomly selected 449 participants, including 97 AD, 233 MCI, and 119 CN subjects, from whom the ADNI baseline collection. All MR images were resized and intensity normalised then skull-stripped and cerebellum-removed before extracting a $64 \times 64 \times 64$ 3D patch around the hippocampus region. The obtain ground truth segmentation of each MRI image, they used FIRST from the FMRIB Software Library [62] to gain a coarse segmentation then three radiologists manually corrected it. For AD vs. CN classification task, they reported 80.1% accuracy. There is no data leakage detected as they only used one 3D patch per MRI image from each subject.

Basher et al. [10] proposed a compound approach to classify AD on T1w

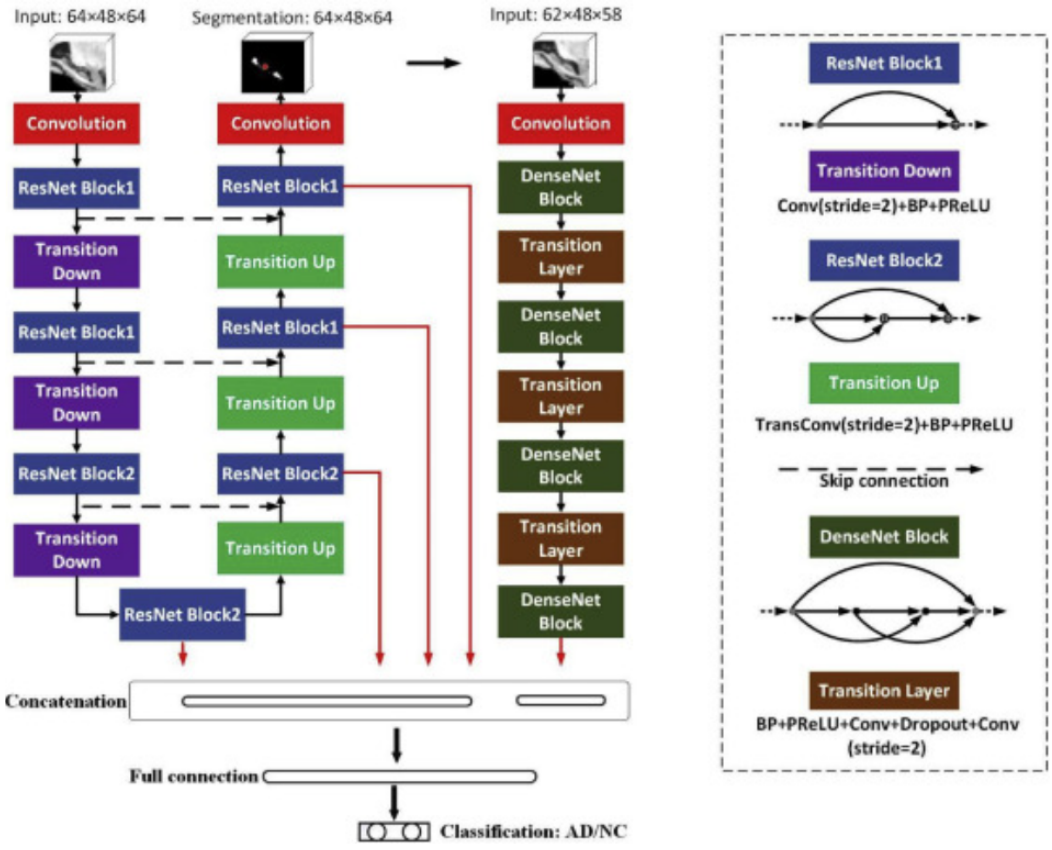


Figure 3.11: An overview of the AutoEncoder-based feature extraction approach for AD classification in [51]. This figure is copied from [51].

sMRI images. As shown in Fig.3.12, the first localisation CNN model learns to estimate the 3D hippocampal region in a 3D MRI image. Secondly, plane-wise discrete volume 2D CNN models are trained to measure the number of voxels of each slice in the hippocampus region as a feature matrix. Then the final CNN is trained on the features matrix to predict a binary label of the input 3D image. The authors selected 171 CN and 80 AD subjects from a private dataset, the Gwangju Alzheimer's and Related Dementia (GARD), to evaluate their approach. They only used z-score normalisation to process each MRI image. The testing results show near 95% accuracy for the binary classification task. Although their chosen dataset is private and the number of subjects is comparatively small, their approach is using the MRI images in a 3D fashion thus no data leakage is detected.

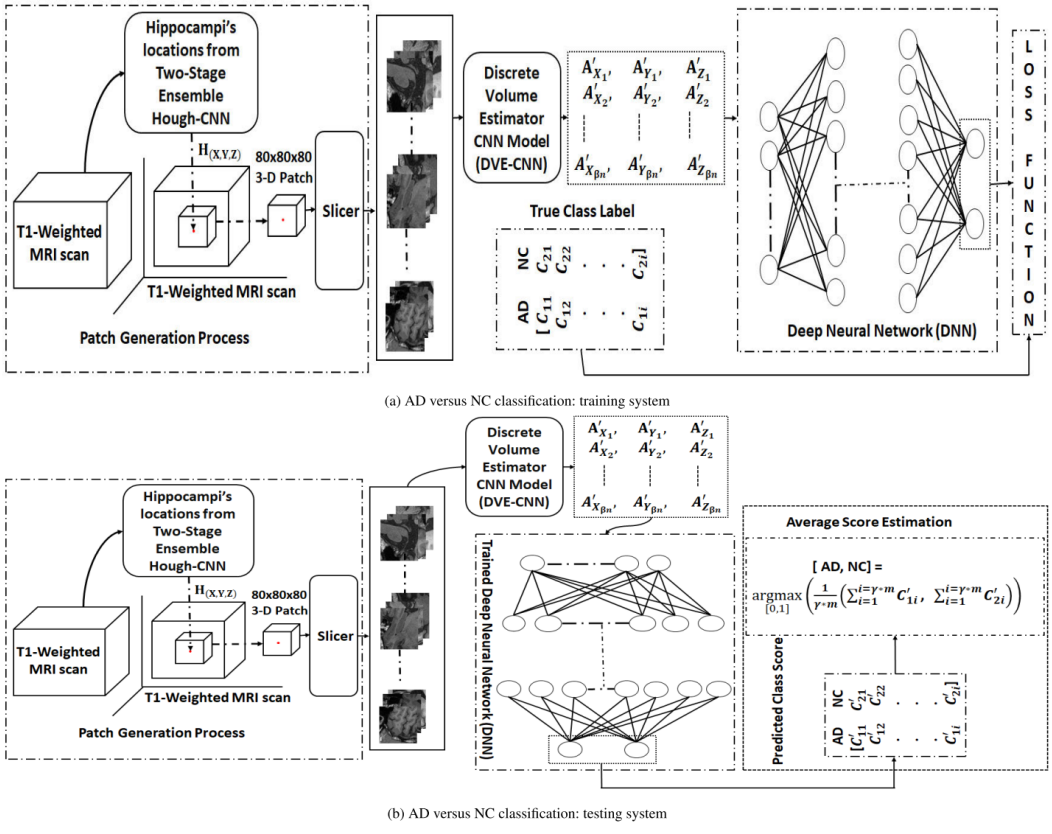


Figure 3.12: An overview of the compound feature extraction approach for AD classification in [10]. This figure is copied from [10].

Venugopalan et al. [89] proposed a fusion approach to classify AD on T1w sMRI images, clinical and genetic data. As shown in Fig.3.13, they used

3D CNN for imaging data and stacked Denoising AutoEncoders for clinical and genetic data to extract features for fusion followed by classification. The authors selected 266 AD, 104 MCI and 132 CN subjects for their last visit from the ADNI database. There are 220 patients that have all three data modalities (testing dataset), 588 patients have genetic and clinical, 283 patients have imaging and clinical, and the remaining patients have only clinical data. The preprocessing pipelines for MRI images are registration, segmentation and normalisation, whereas the clinical and genetic features with missing values $> 70\%$ are discarded in clinical and genetic data. Their evaluation showed 86% accuracy performance. As they are only using one MRI image per subject, there is no data leakage.

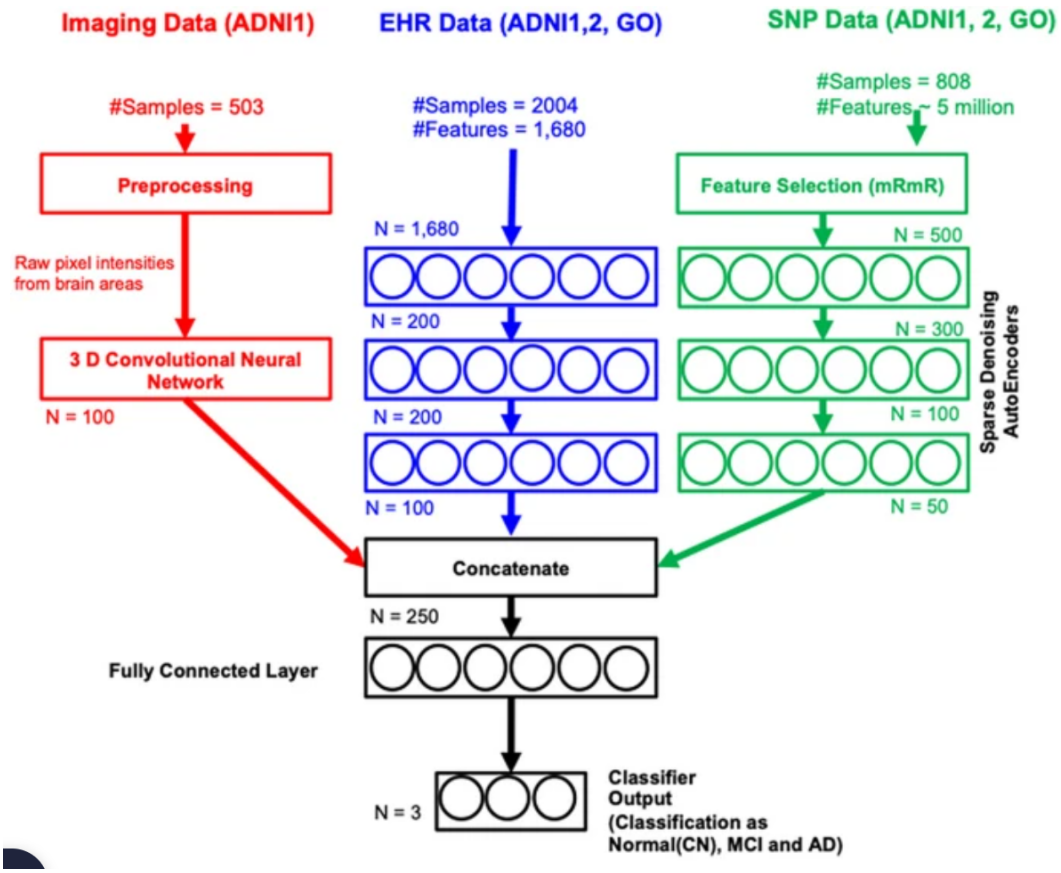


Figure 3.13: An overview of the multi-modality feature fusion approach for AD classification in [89]. This figure is copied from [89].

Wu et al. [93] proposed a transfer learning and self-supervised method to classify AD on the T1w sMRI modality. As shown in Fig.3.14, 32 axial slices

closest to the centre of each MRI image are fed into a pre-trained network to generate the bottleneck features for dimension reduction. Then an AutoEncoder is trained on these features and the bottleneck features are utilised again to further reduce dimension. Finally, the slice-wise features are merged for the final classification. The authors selected 100 AD subjects and 316 CN subjects for training from the OASIS-1 database. An additional 100 subjects of each class are chosen for testing purposes. Each MRI image is processed through facial features, smoothing, correction, normalisation and registration. Using the MobileNet as the pre-trained bottleneck feature extractor showed the best classification accuracy of 80.5%. The authors are aware of the data leakage problem and clearly performed subject-level train-test-split.

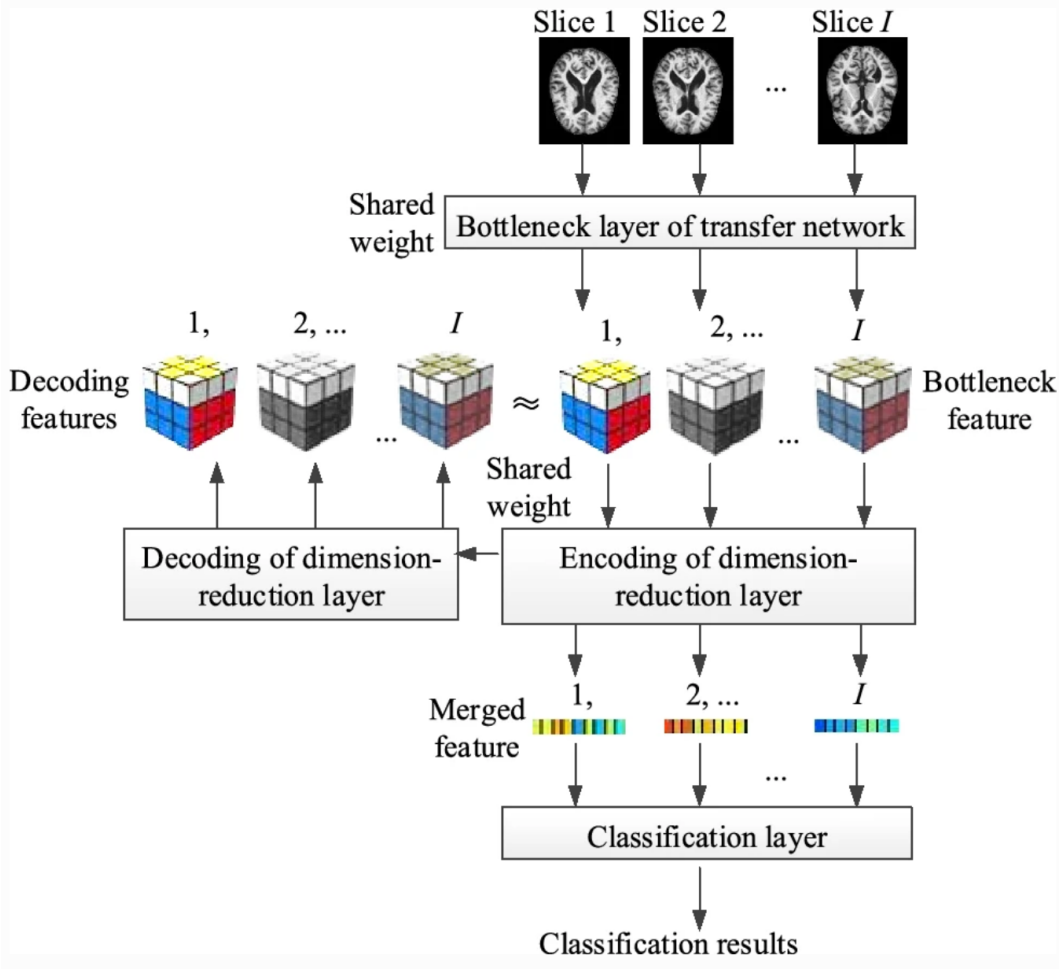


Figure 3.14: An overview of the multi-modality feature fusion approach for AD classification in [93]. This figure is copied from [93].

3.1.2 Prior Work with Potential Data Leakage

Many other scan-level studies either performed non-subject-level train-test-split or did not provide a clear enough explanation of their method to ensure reproducibility, therefore their reported results are highly likely to data leakage. A brief summary of each paper in this subsection is listed in table Table.3.2.

Islam and Zhang [33] proposed an approach for multiclass classification of disease stages using 2D Dense CNN. They selected 416 subjects of four stages from the OASIS database. They clearly reported a subject-level data partition strategy to avoid data leakage. Multiple 2D patches (112×112) are extracted from each plane of a 3M MRI volume. The 2D patches are normalised to zero-mean and unit variance. Random cropping is applied before the inputs are fed into a 3-in-1 ensemble model, as shown in Fig.3.15. The authors reported an average 94% accuracy for a four-class classification task. However, the number of samples in the test for each class is imbalanced and a baseline evaluation on the AD vs CN task is not reported. Although their proposed approach utilises 3D sMRI volume as input, the train-test-split is suspected to have *Leakage A*.

Basaia et al. [9] proposed a 3D CNN classification approach to identify AD on T1w sMRI data. They have selected 352 AD and 510 CN from the ADNI dataset, they also have access to 124 AD and 55 CN subjects in a private dataset, namely the Milan dataset [58]. They proposed a plain 3D CNN architecture and reported 99.2% accuracy on AD vs. CN task. However, they did not state the number of sMRI volumes selected from each subject thus it is suspected to have a data leakage A issue.

Raju et al. [78] also proposed a plain 3D CNN based approach to classify AD on T1w sMRI images. From the ADNI database, they selected 152 NC, 181 MCI, and 132 AD subjects for evaluation. Each image goes through a preprocessing pipeline including centre-cropped resizing, motion correction, non-uniform intensity normalisation, affine transformation, intensity normalisation and skull stripping. Then 27 half-overlapping patches ($50 \times 41 \times 40$)

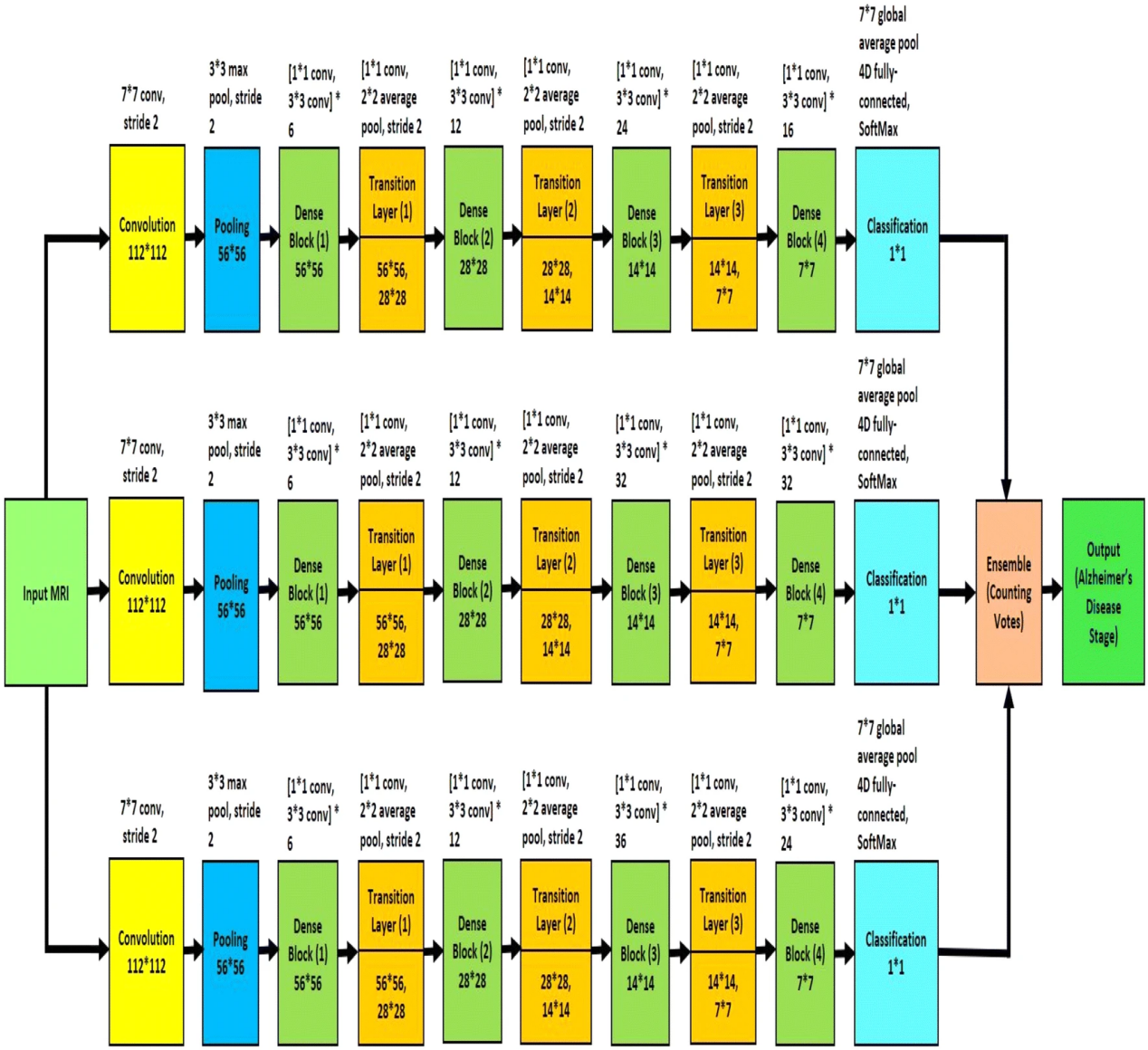


Figure 3.15: An overview of the 2D Dense CNN ensemble for AD classification. This figure is copied from [33].

are extracted. As shown in Fig.3.16, each 3D patch is given to an individual 3D CNN to generate a feature vector. A fusion of the 27 feature vectors and a support vector machine is used for the final prediction. The author reported 97.77% accuracy for a three-class classification task. They did not explain the sMRI selecting and splitting process for each subject thus it is suspected to have a data leakage type B.

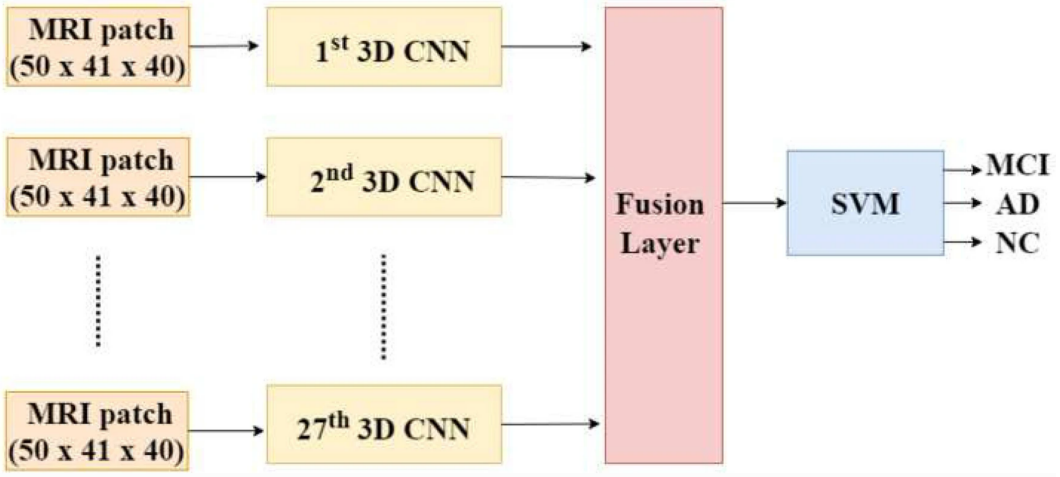


Figure 3.16: An overview of the 3D CNN based feature extraction approach for AD classification in [78]. This figure is copied from [78].

3.1.3 Prior Work with Data Leakage

There are a few papers that have clear data leakage issues. The non-subject-level train-test-split is the most common issue, whereas the data leakage of representation or features is less common but more difficult to avoid. Publications in this section are also listed in Table. 3.2.

Jain et al. [34] proposed a 2D CNN (VGG-like) approach for AD, MCI and CN classification. Using a relatively simple VGG-like 2D CNN, they reported over 99% accuracy on a collection consisting of 50 subjects for each class from the ADNI database. The extracted 2D slices from each MRI image are then split into training and testing subsets for evaluation, risking that the slices of the same subject could end up in both training and testing subsets (i.e. Data Leakage A).

Paper	Dataset #Subjects (#Scans)	Data Selection	Preprocessing & Augmentation	Classifier & Approach	Modality	Accuracy (%) AD vs CN
Aderghal et al. [2]	ADNI AD 188 MCI 399 CN 228	Hippocampus ROI (28x28x28), 3 slices from each plane, Subject-level	alignment, intensity normalisation, flips, blurring, volume translations,	2D CNNs (one for each plane) and majority voting	T1w sMRI	91.02
Aderghal et al. [3]	ADNI AD 188 MCI 399 CN 228	Hippocampus ROI (28x28x28), 3 slices from each plane, Subject-level	alignment, intensity normalisation	2D CNNs (one for each plane), Transfer and Fine-Tuning for DTI data, then feature fusion	of T1w sMRI and Diffusion Tensor Imaging	92.5
Bäckström et al. [8]	ADNI AD 199 CN 141	3D MRI (110x110x110), Subject-level	Cortical Reconstruction, Edge Trim, Resize, Intensity Min-Max Normalisation	3D CNN	T1w sMRI	98.37
Korolev et al. [40]	ADNI AD 50 LMCI 43 EMCI 77 CN 61	3D MRI (110x110x110), Subject-level	alignment, normalisation, skull-stripping	3D VoxCNN & 3D ResNet	T1w sMRI	79 & 80
Li et al. [46]	ADNI AD 199 CN 229	Multiscale 3D MRI, Subject-level	linear registration, intensity normalisation, skull-stripping, cerebellum-removal	Ensemble of 3D CNN and 3D ConvAE	T1w sMRI	88.31
Li et al. [47]	ADNI AD 199 MCI 403 CN 229	Multiple 3D patches (32x32x32) from each ROI, Subject-level	linear registration, intensity normalisation, skull-stripping, cerebellum-removal	K-means clustering of ROI patches followed by the Ensemble of Multi-cluster DenseCNN	T1w sMRI	89.5
Liu et al. [50]	ADNI AD 93 MCI 204 NC 100	Multiple 3D ROIs (50x41x40), Subject-level	sMRI: skull-stripping, intensity normalisation, cerebellum removal, affine registration PET: registration to sMRI	Fusion of 3D CNNs (one for each ROI) for sMRI and PET	T1w sMRI and PET	93.26
Senanayake et al. [83]	ADNI AD (161) MCI (193) CN (161)	3D MRI and 1D neuropsychological features, Subject-level	Not Explained	Dilated 3D CNN, dense connections and residual connections fusing with neuropsychological features	T1w sMRI and 1D Feature Vector	76
Valliani and Soni [88]	ADNI AD 188 MCI 243 CN 229	median axial 2D slice of 3D MRI, Subject-level	Not Explained	ImageNet ResNet (pretrained) plus FullyConnected Layers (training)	T1w sMRI	81
Oh et al. [68]	ADNI AD 198 pMCI 166 sMCI 101 CN 230	3D MRI, Subject-level	realignment, normalisation, smoothing	3D ConvAE on AD and CN reconstruction, then transfer learning for classification	T1w sMRI	86.6
Liu et al. [51]	ADNI AD 97 MCI 233 CN 119	3D patch raw and segmentation of hippocampus, Subject-level	resize, affine registration, intensity normalisation, skull stripping, cerebellum removal,	3D ResNet and 3D DenseNet then concatenation for classification	T1w sMRI	92.5
Basher et al. [10]	GARD AD 80 CN 171	One 3D patch from hippocampus, Subject-level.	z-score normalisation	volumetric features transformation, CNN classification	T1w sMRI	94.82
Venugopalan et al. [89]	ADNI AD 266 MCI 104 CN 132 Clinical 2004 genetic 808	21 ROIs from each sMRI, EHR 1680 features, SNPS 5 million features.	For sMRI registration, segmentation, normalisation For EHR and SNP: discard features missing more than 70 percent	3D CNN for sMRI (100 features), Denoising AE for EHR (100 features), Denoising AE for SNP (50 features), Fusion of all features for classification	T1w sMRI, Electronic Health Records, Single Nucleotide Polymorphisms	86
Ebrahimi et al. [23]	ADNI AD 132 CN 132I	3D MRI volume	intensity normalisation, registration, tissue segmentation, resize, scaling, pixel translation	3D CNN ResNet-18 from pre-trained 2D ImageNet transfer learning by duplicating 2D weights for the third dimension	T1w sMRI	96.88
Wu et al. [93]	OASIS Training AD 100 CN 316 OASIS Testing AD 100 CN 100	2D Slices of a 3D MRI volume	facial features, smoothing, correction, normalisation, registration	Pre-trained CNN (MobileNet) to extract features, then train autoencoder to reduce dimension for classification	T1w sMRI	80.50

Table 3.1: Table of papers without data leakage. EMCI: early MCI, LMCI: late MCI, pMCI: progressive MCI, sMCI: stable MCI.

Unlike many other papers, Maqsood et al. [56] extracted subjects from the OASIS-2 database based on CDR scores. They also divided the 3D MRI into 2D slices that might have a data leakage issue (type A). There is no explicit explanation for the slice extraction process, so it is very likely to have a data leakage issue (type A). They used a pre-trained ImageNet-based transfer learning approach resulting in 92.85% accuracy in a four-class classification setting.

Mehmood et al. [59] also used the OASIS database, but they extracted full-size 2D slices from each plane. Their preprocessing pipeline includes resizing, linear contrast stretching, and intensity normalisation. They also used augmentation techniques such as rotation, zoom, linear transformation and channel shift. The authors chose a plain 2D CNN for the binary classification task. Their evaluation shows 99.05% accuracy in the binary classification task. They used two parallel VGG-like 2D CNNs followed by concatenation to obtain better features of the 2D slices. However, their train-test-split is not performed on the subject-level, thus it is almost certain to have a type A leakage issue.

Later on, Mehmood et al. [60] proposed a transfer learning approach for the AD classification task. From the ADNI database, they selected sMRI for grey matter tissue segmentation. Then they extracted 2D slices of each plane from the segmented voxels for learning. The authors utilised pre-trained VGG-like 2D CNN along with the layer freezing technique. Their evaluation showed 98.73% accuracy in the AD vs. CN task. However, their train-test-split is performed on the slice level after the grey matter segmentation. It is highly likely to have a data leakage that processed or augmented images of the same subject are in both subsets (i.e. leakage type B).

Huang et al. [31] proposed a 3D CNN approach but they extracted 3D hippocampus ROI from both sMRI and PET data. The ROI is selected after MRI realignment, and resizing. The 3D ROI for PET modality is chosen as the same location as sMRI images. Each modality data is used to train a

model for feature concatenation followed by AD vs. CN classification. They reported 90.10% accuracy using a weight-sharing strategy for the two models. They also reported single modality-model results of 81.19% and 89.11% for sMRI and PET, respectively. As a single subject might have multiple sMRI images, there is a risk that the MRI image of the same subject could be in both training and testing subsets (type A leakage).

Zhang et al. [95] also introduced a 3D CNN based approach for AD classification that is similar to [31]. They selected subjects from the ADNI database then the sMRI volumes are prepared for Skull-stripping, spatial normalisation, background removal and resizing. From each 3D sMRI volume per subject, 64 overlapping 3D patches are extracted. As the authors did not provide an explicit explanation of the train-test-slip process, it is possible that their evaluation result (97.35% accuracy) is flawed by a data leakage (type A). Nevertheless, they proposed a novel network architecture that incorporates an attention mechanism along with dense connectivity between layers, as shown in Fig.3.17.

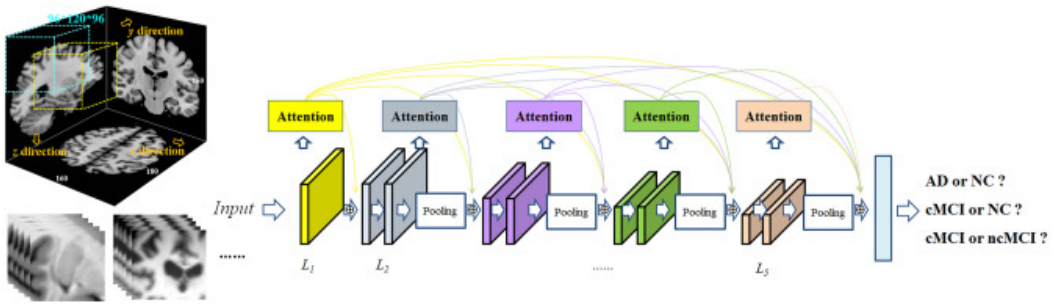


Figure 3.17: An overview of the 3D CNN attention mechanism with dense connectivity for AD classification. This figure is copied from [95].

Liu et al. [49] also used the OASIS database to evaluate their proposed depth-wise separable convolution (DSC) network. They used 1×1 kernels for the convolution layers instead of 3×3 or larger kernels. The smaller kernels greatly reduced the computational costs. They first trained a 2D CNN to obtain 78.02% accuracy as a baseline, then they replaced the CNN layers with DSC layers using significantly fewer parameters that showed similar perfor-

mance. However, they performed the train-test-split on the slice-level after the MRI selection and augmentation. As no explanation was given for their choice of only using the axial plane, it is almost certain that there is data leakage of type B.

Kang et al. [35] utilised a generative adversarial network (GAN) as part of their approach for the AD classification task. The authors selected subjects from the ADNI database under the preprocessing pipeline of affine transformation, non-linear registration, bias field correction, and smoothing. Then a 2D CNN is trained using coronal plane slices to select the top 11 slices per subject with the highest classification. As shown in Fig.3.18, their approach uses the selected slices to train the GAN part, then the discriminator parameters are transferred into multiple independent 2D CNNs. Each CNN is fine-tuned using a random subset of the 2D slices. The final prediction is obtained by majority voting over an ensemble of the CNN predictions. The authors reported an accuracy value of 90.36% when classifying AD vs. CN. The utilisation of GAN networks is the highlight of their approach, but the lack of clarification of train-test-split raises the concern of type B data leakage in their evaluation.

Shanmugam et al. [84] explored the transfer learning approach for AD classification using pre-trained networks. They selected subjects from the ADNI database and then applied linear contrast stretching and intensity normalisation. Their evaluation showed that the overall accuracy of the GoogleNet, AlexNet and ResNet-18 in detecting AD is 96.39%, 94.08% and 97.51% respectively. As the explanation of 2D slice selection is not clearly stated, it is very likely to have a data leakage of type A.

Goenka and Tiwari [29] proposed a workflow for AD classification using sMRI data on volume-level inputs, patch-level inputs and slice-level inputs. As shown in Fig.3.19, the authors established a workflow to examine the classification efficiency of the three inputs. For network inputs, the whole 3D volume can be fed into the network or smaller 3D patches are selected from a volume as inputs, whereas the slice-level method takes 2D slices from each plane

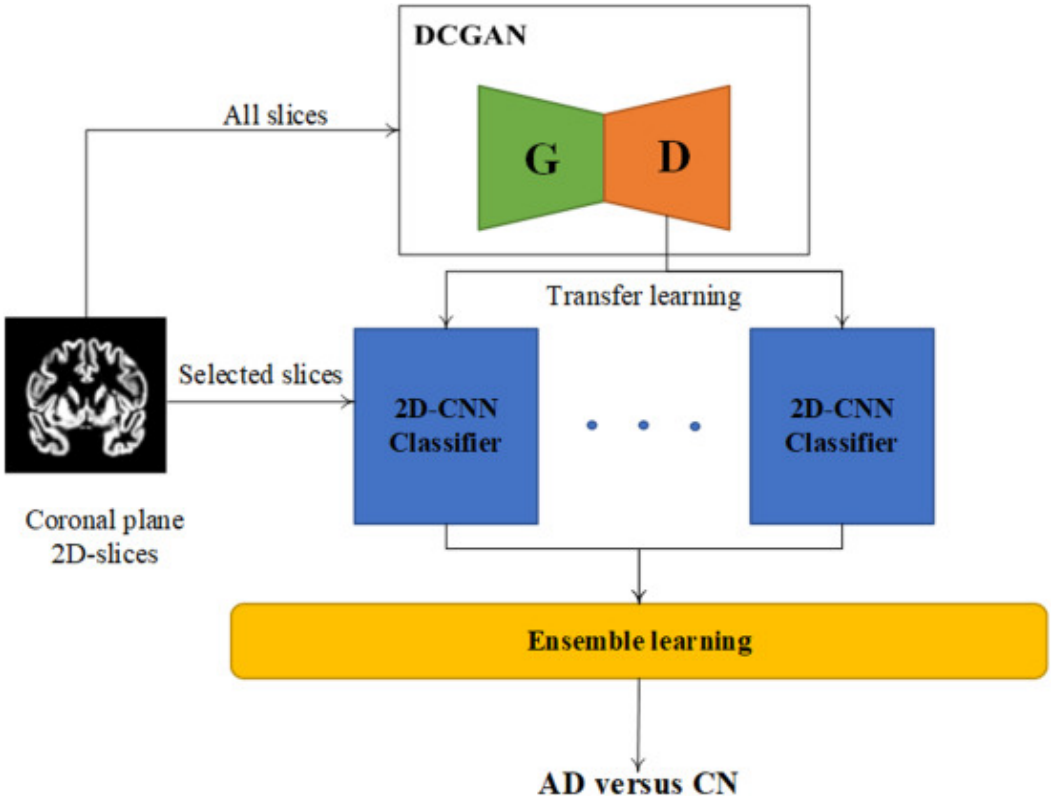


Figure 3.18: An overview of the approach using GAN for AD classification in [35]. This figure is copied from [35].

of a volume. Their evaluation shows that the 3D volume-level approach was the most efficient with a classification accuracy of 98.26%, followed by the 3D patch-level approach (97.48%) and then the 2D slice-level approach(95.40%). The authors selected subjects from the ADNI database, and they utilised multiple scans per subject. This leads to the potential data leakage that the scans of one subject might be in both training and testing subsets for either of the volume, patch or slice approaches.

Savaş [82] also attempted the transfer learning approach for AD classification. From the ADNI database, the authors extract 2D slices from the sagittal plane of T1w sMRI volumes. They established a comparison framework for a variety of pre-trained models including ResNet, EfficientNet, AlexNet, ImageNet, VGGNet, MobileNet, XceptionNet. Among them, the EfficientNetB0 reported the highest binary classification accuracy of 92.98%. While splitting the slices into training and testing subsets, each slice is organised as an

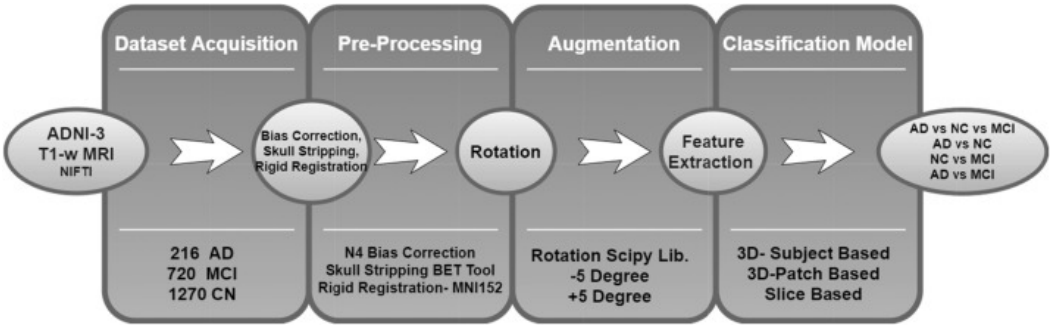


Figure 3.19: An overview of the workflow for AD classification. This figure is copied from [29].

independent data point, thus there is a risk of type A data leakage.

Orouskhani et al. [69] proposed an approach for AD classification using Triplet Networks. Their idea is to employ a triplet loss function to calculate the difference between various AD stage images. By minimising the loss, the triplet networks can be trained as feature extractors for downstream classification. As shown in Fig.3.20, three input images from different classes of AD are given to the network with identical architecture and weights. Then, each network generates the corresponding features for each input, followed by the calculation of the distance between features. Using subjects from the OASIS database, their approach achieved 99.41% accuracy in a multiclass classification task. Despite the novelty of their approach, the authors did not clarify the slice selection process and train-test-split process. Therefore, it is very like to have a data leakage that the slices of the same subject could be in both subsets.

3.2 Brain Age Prediction

Along with the advancement of neuroimaging technologies, neuroimage-driven brain age prediction has received increasing research attention due to its potential contribution to clinical practice. Study [25] shows that non-demented individuals with type II diabetes have a brain age 4.6 years greater than healthy individuals. [66] found that individuals with schizophrenia show increased brain age by 2.6 years.

Paper	Dataset #Subjects (#Scans)	Data Selection & Leakage	Preprocessing & Augmentation	Classifier & Approach	Modality	Accuracy AD vs CN
Islam and Zhang [33]	OASIS Total 416 Subjects of CDR 0,0.5,1,2	2D Patch (112x112) from each plane, (Suspected Leakage A)	Z-score normalisation, 3 crops of each plane	2D Dense Connectivity CNNs (one CNN for each plane) then majority voting	T1w sMRI	93 (multiclass)
Basaia et al. [9]	ADNI / Milan AD 352 / AD 124 sMCI 294 / MCI 50 cMCI 253 / CN 55 CN 510	3D MRI, Subject-level (Suspected Leakage A)	tissue segmentation, registration, affine transformation	3D CNN classification	T1w sMRI	99.2
Jain et al. [34]	ADNI AD 50 MCI 50 CN 50	32 most informative MRI slices (256x256x256) based on entropy (Leakage A)	motion correction, conform, affine transformation, intensity normalisation, skull stripping	2D CNN (VGG-like)	T1w sMRI	99
Maqsood et al. [56]	OASIS CDR-0 167 CDR-0.5 87 CDR-1 105 CDR-2 23	2D Slices (Leakage A)	intensity normalisation	ImageNet-based transfer learning	T1w sMRI	92.85 (multiclass)
Huang et al. [31]	ADNI AD(1355) CN(1506)	3D slices of hippocampus ROI (Leakage A)	realignment, resize, hippocampus ROI selection	3D CNN x 2 for both modalities then feature concatenation	T1w sMRI, PET	90.10
Mehmood et al. [59]	OASIS CDR-0 167 CDR-0.5 87 CDR-1 105 CDR-2 23	2D slices (Leakage B)	resize,linear contrast stretching, intensity normalisation, rotation, zoom, linear transformation, channel shift	Two parallel VGG-like 2D CNNs followed by feature concatenation	T1w sMRI	99.05
Raju et al. [78]	ADNI AD 132 MCI 181 NC 152	27 3D patches (50x41x40) per subject, Subject-level (Suspected Leakage A)	resize, motion correction, non-uniform intensity normalisation, affine transformation, intensity normalisation, skull stripping.	One 3D CNN for each 3D patch followed by feature concatenation and SVM classification	T1w sMRI	97.77 (multiclass)
Mehmood et al. [60]	ADNI AD 75 EMCI 70 LMCI 70 CN 85	2D Slices (Leakage B)	tissue segmentation, affine regularization, spatial normalisation, smoothing, resize	CNN with frozen layer training then transfer learning	T1w sMRI	98.73
Zhang et al. [95]	ADNI AD 280 cMCI 162 sMCI 251 CN 275	64 3D patches (96x120x96) per subject (Leakage B)	skull-strip, spatial normalisation, background removal, resize	3D connection-wise attention and densely connected convolution neural network	T1w sMRI	97.35
Liu et al. [49]	OASIS AD (90) MCI (136) CN (266)	2D slices (Leakage B)	resize, intensity normalisation; clipping, flipping, increase contrast, rotate	2D CNN with 1 by 1 kernel	T1w sMRI	91.40
Kang et al. [35]	ADNI AD 187 MCI 382 CN 229	2D Slices (Leakage A)	affine transformation, non-linear registration, bias field correction, smoothing	GAN Discriminator transfer learning to multiple classifiers for a majority voting ensemble	T1w sMRI	90.36
Shanmugam et al. [84]	ADNI AD 277 LMCI 234 MCI 228 EMCI 264 CN 162	2D Slices (Leakage A)	linear contrast stretching, intensity normalisation	Transfer learning from pre-trained networks GoogLeNet, ResNet-18, and AlexNet	T1w sMRI	98.63 94.08 97.51
Goenka and Tiwari [29]	ADNI AD 70 MCI 224 CN 475	3D MRI 3D patches 2D slices (Leakage B)	Bias Field Correction, Skull Stripping, Rigid Registration, Image Rotation	3D CNN for 3D volume, 3D CNN for 3D patches, 2D CNN for 2D slices	T1w sMRI	98.26 97.48 95.40
Savaş [82]	ADNI AD 99 MCI 148 NC 135	2 middle slices (Leakage A)	resize	2D CNN with transfer learning from ResNet, EfficientNet, AlexNet, ImageNet, VGGNet, MobileNet, XceptionNet.	T1w sMRI	92.98 (EfficientNetB0)
Orouskhani et al. [69]	OASIS CDR-0 167 CDR-0.5 87 CDR-1 105 CDR-2 23	2D slices (Leakage A)	Not Explained	VGG16-based Triplet Loss Network	T1w sMRI	99.41 (multiclass)

Table 3.2: Table of papers that might have data leakage. pMCI: progressive MCI, cMCI: convert MCI, sMCI: stable MCI, ncMCI: non-convert MCI.

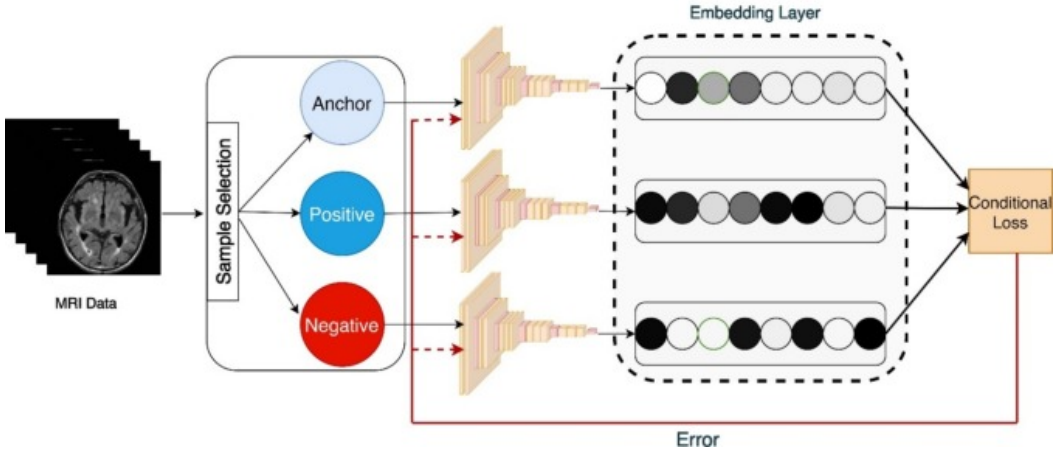


Figure 3.20: An overview of the Triplet Network for AD classification. This figure is copied from [69].

Researchers utilise this approach to introduce a reliable biomarker for neurodegeneration diagnosis and monitoring in individuals with AD. The idea is that an "older-appearing" brain MRI indicates an accelerated brain ageing process. [26] presented a brain age estimation framework with 81% accuracy in predicting the conversion of MCI to AD within 3 years of follow-up, which outperformed methods based on two state-of-the-art biomarkers: hippocampus volume and cerebrospinal fluid. [24] reported a five years mean gap between the predicted age and chronological age of cognitively normal individuals, whereas the mean estimated age difference of AD individuals is ten years. [72] reported mean absolute error (MAE) of 2.14 years in brain age prediction and 99.5% accuracy in sex classification using a 3D CNN model trained on the UKBiobank T1w MRI data ($N = 14,503$). Despite using 3D MRI images, their CNN model only has 3 million parameters, therefore the computational complexity and memory cost are significantly reduced. [12] first generated grey and white matter images using a nonlinear deformation template based on 839 healthy individuals, then they trained support vector regression models for 129 AD subjects to estimate the brain age. Although the AD images show significantly higher estimated brain age than actual age, the number of preprocessing steps required is significant, making their approach inflexible and difficult to reproduce. These studies show that predicted brain

age is correlated with neurodegenerative disease progression, therefore it can be used as a valid feature for classification.

3.3 Self-supervised Approach for AD Classification

The self-supervised approach achieved promising results in the literature for natural image processing. Among the few attempts dedicated to the AD classification task, Dufumier et al. [20] proposed a cross-entropy-based loss function for a self-supervised approach that maps subjects' T1w sMRIs with a similar age close to each other in a latent space. They applied the new loss function to pre-train 3D DenseNet and 3D UNet as feature extractors on $10,000$ 3D sMRI volumes from 13 publicly available datasets. Their evaluation using a subset from the ADNI dataset reported 96.58% and 96.84% AUC, respectively. All the MRI data for pre-training and testing are preprocessed with non-linear registration followed by grey matter segmentation. To the best knowledge of this thesis, Dufumier et al. [20] involved the most extensive amount of real-world sMRI data in the literature. Although the train-test-split is performed on the subject-level without the concern of data leakage, this approach heavily depends on the quality and accuracy of the grey matter segmentation.

Zhao et al. [96] proposed a self-supervised learning approach for AD classification using AutoEncoders. As shown in Fig.3.21 they embedded a dummy unit vector τ in the network by applying a fully connected layer to it. This made τ a global variable and independent of the input volumes. The autoencoder is pre-trained on a synthetic dataset with image pairs to learn the development of the AD over time, in which the vector z is colinear with τ. The preprocessing consisted of denoising, bias field correction, skull stripping, affine registration to a template, re-scaling to a $64 \times 64 \times 64$ volume, and transforming image intensities within the brain mask to z-scores. Using the longitudinal sMRI image pairs of subjects from the ADNI and the Alcohol

dataset, a fine-tuning process improved the classification accuracy from 72.0% 62.9% to 84.1% 71.7%, respectively. The authors ensured that the image pairs from the same subject belong to the same subset during evaluation. However, the authors did not provide much information regarding the synthetic dataset, therefore the overlapping of the subjects between training and testing is unknown.

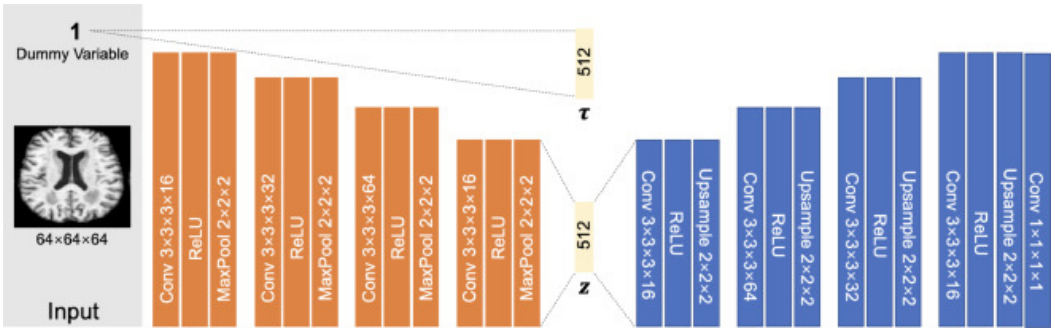


Figure 3.21: An overview of the AutoEncoder-based self-supervised learning approach for AD classification in [96]. The Orange blocks are the encoder and the blue blocks are the decoder networks. This figure is copied from [96].

Ouyang et al. [70] proposed an autoencoder-based self-supervised approach for AD classification. The novelty of their approach lies in utilising each subject's longitudinal sMRI scans to train an autoencoder for feature extraction. The autoencoder learns a latent space in which the neighbourhoods are age-consistent (i.e. MRI volumes of different subjects but with similar brain ages are close to each other) and progression-consistent (i.e. the latent space vector of the MRI volumes pairs from the same subject captures the age development). The authors linked the MRI volumes of each subject into "pairs" according to their chronological order. To augment the inputs, random shift (within 4 pixels), rotation (within 2 degrees), and random flipping of brain hemispheres are applied to each pair of MRIs. Their approach is very unlikely to have a data leakage as their train-test-split and volume usage is on the subject-level. The evaluation showed an 83.5% accuracy for AD vs. CN classification.

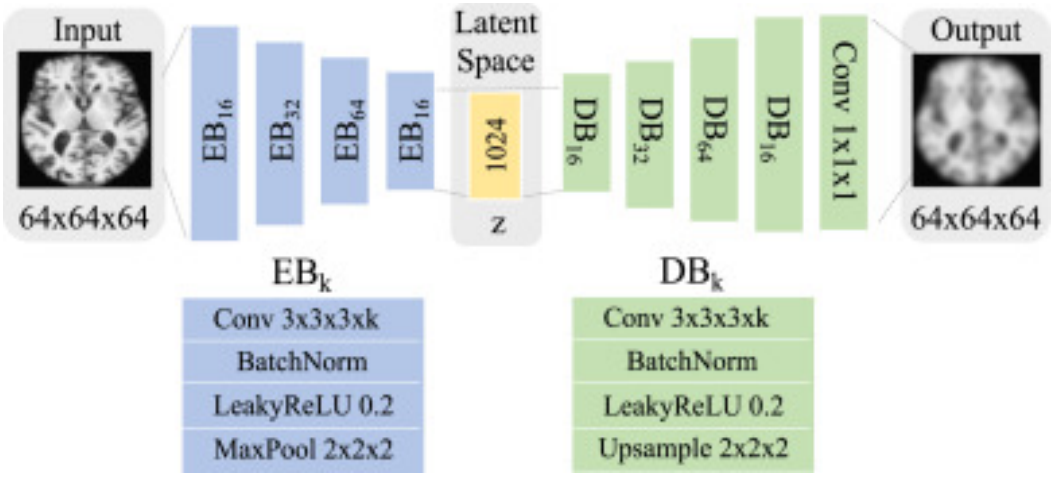


Figure 3.22: An overview of the AutoEncoder Network used in [70] for AD classification. This figure is copied from [70].

3.4 Discussion

This section briefly discusses the choice of input shapes, the variety of pre-processing methods and deep networks, and the data leakage issues in recent publications on AD classification.

3.4.1 Input Shape

Many papers extracted 2D slices from 3D MRI volumes as inputs. It is an instinctive choice as the convolutional networks showed promising performance in natural images. However, slice-level inputs are more prone to data leakage if the train-test-split is not performed on the subject level. Another issue with slice-level inputs is that the spatial information in 3D MRI volumes is not comprehensively utilised. Many studies treated the slices from the same volume as independent images. Some studies extracted patch-level data (from the volume) as inputs as a way of compensation. One advantage of this method is increased training samples as the number of patches per volume can be quite high.

To reduce the computational cost of large 3D CNN, some authors divided the whole 3D volume into smaller 3D patches. The hippocampus region is the most common location for single patch selection. While extracting multiple

patches, the number of patches and the option of overlapping are different across papers. As shown in the previous section, selecting multiple patches from a single volume has the same risk of data leakage as choosing multiple slices. Nevertheless, patch-level methods can partially utilise the spatial information within the volumetric data with reduced computational cost.

The improved computational power in more recent times has allowed direct model training on 3D MRI volumes as opposed to 2D slices. As a result, the number of papers using 3D volume as input is on the rise. However, one issue with volume-level methods is that the total number of training samples is often limited. Given the increased number of parameters to learn in 3D CNN, the risk of model overfitting is significantly increased.

3.4.2 Preprocessing and Augmentation

In the literature, there is a wide range of preprocessing methods has been employed. For example, resizing, registration/realignment, and intensity normalisation are the most popular ones. Some papers used skull-stripping to extract the brain tissue as a way to reduce noise. Some papers used third-party software to generate brain grey matter tissue segmentation for learning. The evaluation and performance of the downstream task heavily depend on the choice of preprocessing methods.

In AD classification, the authors used augmentation techniques that are similar to what has been used in natural image classification problems. For example, geometric transformations such as random flip, crop, rotate, stretch, and zoom images can be directly applied to 2D slices. For 3D volumes, random crop and linear shift are more popular in the literature due to their simplicity for the third dimension.

The choice of preprocessing pipeline and data augmentation techniques is not consistent across the publications. The reason for the choice of preprocessing and augmentation is often not clearly explained by the authors. This makes the comparison of different approaches very difficult to reach a system-

atic conclusion.

3.4.3 Architecture and Learning

It is good to see the wide variety of deep network architectures in the literature. From VGG-inspired networks to DenseNet, ResNet and GANs, the promising architectures in the natural image domain have been adapted into AD classification. These models are often the go-to choice for slice-level approaches due to their readiness. The availability of pre-trained weights also encouraged the transfer learning approaches to utilise those well-studied models. Some authors replaced the 2D convolutional layers with 3D ones for 3D volume-level inputs. One can observe that the tremendous number of trainable parameters of 3D convolution is limiting the usage of whole 3D volume as inputs.

In terms of learning schemes, most papers used a supervised learning approach for AD classification. However, given the number of trainable parameters in their networks, the amount of data is far from ideal. Only a few papers attempted the unsupervised approach, but the potential risk of data leakage in their work increased the difficulty of a systematic comparison.

3.4.4 Summary

As mentioned above, a significant proportion of the studies in the literature are suspected to have a data leakage issue. The approaches using slice-level inputs are the most sensitive to data leakage as the slices of each subject are mixed together before the train-test split. The patch-level approaches are can be liable to leakage as well due to the wrong split. The volume-level approach can be flawed as well if the multiple 3D MRI volumes extracted from the same subject are mixed before splitting into train and test subsets.

Furthermore, it is difficult to compare studies using different datasets, the different choices of subjects, and the lack of explanation about the preprocessing pipeline and augmentation techniques. Due to the potential data leakage issues, conducting a systematic comparison of various approaches is faced with

unique challenges. In addition, the choice of neural architecture and the choice of hyperparameters often lacks explanation. Moreover, the working code of implementation is often not available, thus it is very difficult if not impossible to reproduce the results.

Chapter 4

Feature Extractor Training using Brain Age Prediction

Inspired by the success of natural image classification, many publications proposed end-to-end convolutional neural networks (CNN) to classify the brain MRI scans of AD subjects from CN subjects. To utilise 3D MRI data, slice-level, patch-level and scan-level approaches are widely used. Slice-level and patch-level approaches are computationally efficient, but scan-level methods are able to take advantage of the 3D spatial information. The state-of-the-art results are very promising, however, a significant proportion of the studies performed a biased evaluation. Their data leakage issues will be shown in the next section.

In the literature, the neuroimage-driven approach uses brain structural variations and alternations to predict brain age [12]. A comparison of predicted age with true chronological age can be used to determine the brain condition of subjects. Studies [17, 39, 25, 66, 26, 24] show that there is a link between neurodegenerative disease progression and ageing. Utilising this link, we present a deep learning-based brain age prediction approach as a pretext task for obtaining a robust feature extractor for AD vs CN classification on T1-weighted (T1w) MRI scans. The following assumptions are made and assessed:

1. AD individuals suffer from accelerated brain ageing.

2. The appearance of the brain in MRI scans of AD subjects is significantly older than CN subjects of the same chronological age.

3. CNN-based models trained to predict the age of CN-only MRI scans will predict the age of AD MRI scans higher (older) than CN scans.

4. Trained models for CN age prediction can generate discriminable latent representations from MRI for CN vs AD classification.

The motivation behind a pretext task is to leverage the available relative information of the vast amount of unsegmented data to train discriminating models. Each MRI scan has a timestamp in the OASIS database. The timestamp may be used as an accurate description of the chronological age of the subject. Cost-efficiency is another consideration as this thesis has no resources to annotate neuroimaging data with segmentation. This brain age prediction pretext task is scalable to large datasets such as synthetic datasets.

Using the 3D T1-weighted MRI data of cognitive normal (CN) subjects from the OASIS-3 dataset, lightweight 3D CNN-based models are trained to predict brain age. The extracted features are then used in the binary classification of CN vs. AD patients from their brain MRI scans. To establish a baseline, we used support vector machines and random forest classifiers as base classifiers. Our results suggest that the 3D MRI-driven CNN brain age prediction pretext task approach can learn AD-relevant features with high discriminative power without a complicated pipeline of preprocessing or data augmentation. Highlighting the novelty of the approach: the train-test-split is carefully performed on the subject-level to avoid data leakage; the spatial information of 3D volumetric data is fully utilised; the robustness is proven by not using complicated data preprocessing and augmentation techniques.

4.1 Subject Selection

The proposed method was evaluated on T1w MRI scans from the public dataset OASIS-3 [42]. OASIS-3 is a collection of 2,168 T1w MRI sessions for 609 cognitively normal adults and 489 individuals at various stages of cognitive decline ranging in age from 42-95 years. To eliminate the bias of younger CN subjects, we selected only CN subjects older than 59 years to match the youngest subject in the AD group. In the OASIS-3 dataset, most MRI sessions are in T1w format, thus we chose to download T1w scans. Each subject has multiple MRI sessions and each session might have multiple MRI scans. We used the entry timestamp of each MRI scan as the brain's chronological age.

To evaluate the cognitive change of subjects, the Clinical Dementia Rating (CDR) [18] was proposed in 1993, and then globally used due to its diagnostic accuracy and reliability. The CDR system quantifies dementia severity from very mild (CDR 0.5) to mild (CDR 1), moderate (CDR 2), and severe (CDR 3). Research [37] found that CDR scores can provide very useful information to evaluate mild cognitive impairment (MCI) to AD progression in individuals in clinical practice. We have organised the subjects according to their CDR scores of the first and last clinical assessments. The final number of AD and CN subjects selected is shown in Table. 4.1

By convention, Nibabel [14] axes are always in RAS+ orientation (from left towards right, from posterior towards anterior and from inferior towards superior) as shown in Fig. 1.2. However, the 3D scans in the OASIS-3 dataset have various orientations due to different scanners and acquisition protocols. To establish a uniform baseline, this study only included scans that are RAS+ oriented. Scans with less than 128 slices on any plane are excluded.

4.2 Preprocessing

The whole brain scan size of $178 \times 256 \times 256$ voxels is not GPU memory-efficient due to a large number of black voxels on the scan border. Therefore, resizing all

Table 4.1: Characteristics of the Subjects

	CDR Score thresholds	***# of Subjects***
CN	$first = 0$, $last = 0$	513
AD	$first < 1$, $last > 1$	97

3D MRI scans to $192 \times 192 \times 192$ voxels is a logical choice since it still maintains all brain tissue voxels without consuming extensive GPU memory. Preliminary results show that smaller scan sizes (e.g. $156 \times 156 \times 156$ and $128 \times 128 \times 128$) significantly accelerate the training, however, they cannot yield good results. The possible reason could be that the subtle yet informative structural brain changes caused by AD are lost during drastic downsampling and the resizing operation might produce artefacts in the image.

As previously shown [90], preprocessing has a significant impact on the performance of brain age estimation and AD classification. Therefore, for establishing a baseline, this study only employed minimum steps of preprocessing. Sequentially, each MRI scan undertakes a pipeline (shown in Fig. 4.1) of:

1. Resizing: To utilise the 3D scans that have different shapes, and also to remove the excessive background image volume outside the brain tissue, all of them are resampled to a pre-selected shape $192 \times 192 \times 192$.

2. Max-Min Intensity Normalisation: To eliminate the inconsistency of 3D scan voxel intensities, e.g. under or overexposure, the range of intensities is normalised by $I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}}$.

3. Contrast Limited Adaptive Histogram Equalisation (CLAHE) [75]: To enhance the contrast between different brain tissue types, adaptive histogram equalisation is performed on each 3D scan. The contrast-limited variation is chosen to reduce the overamplification of noise.

Although many studies reported that data augmentation techniques signif-

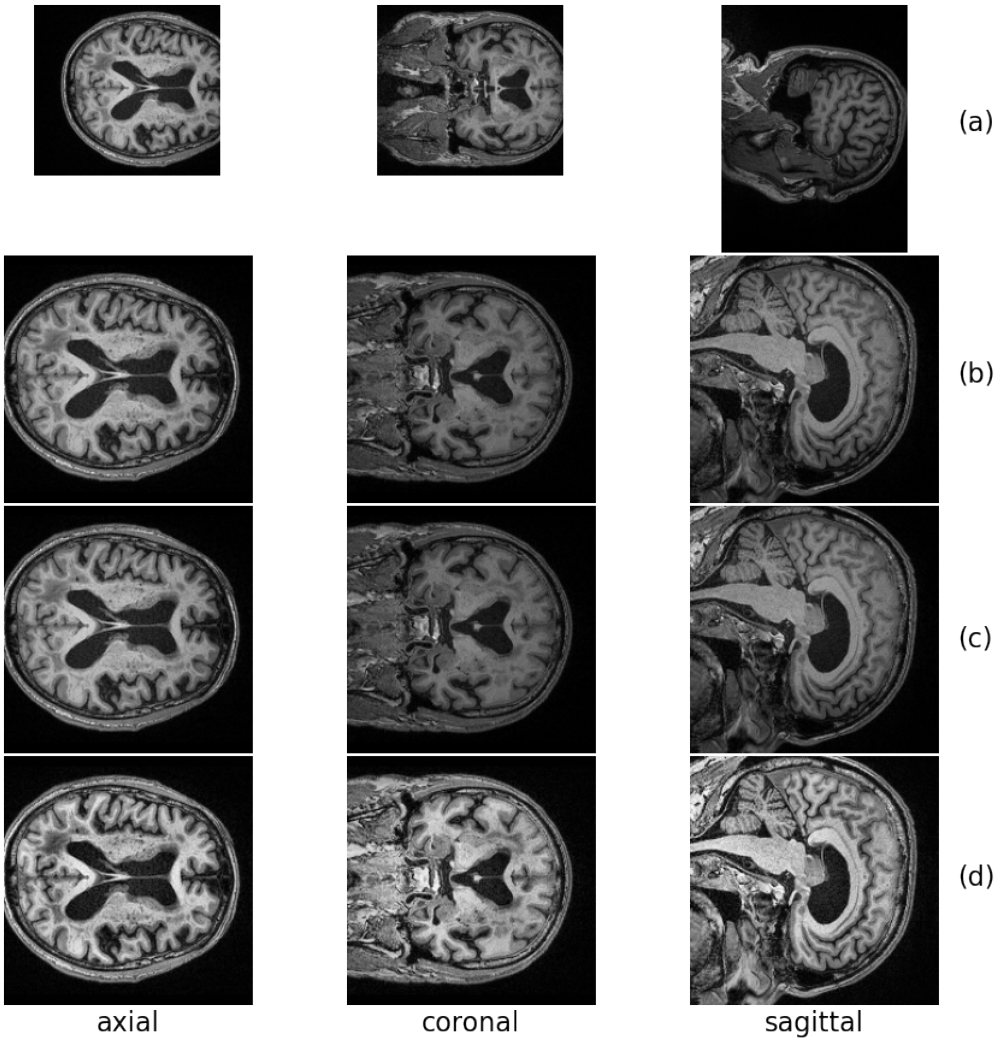


Figure 4.1: Visualisation of each preprocessing step. Row (a) shows the raw image ($176 \times 256 \times 256$ voxels) of a 3D MRI scan, whereas Row (b) presents the resized images ($192 \times 192 \times 192$ voxels). Row (c) shows the max-min intensity normalised image followed by row (d) of images that are processed by contrast limited adaptive histogram equalisation [75].

icantly impact classification performance, they often did not state a thorough explanation to support their selection of techniques. This study did not apply any image augmentation techniques after preprocessing to establish a baseline for our proposed approach.

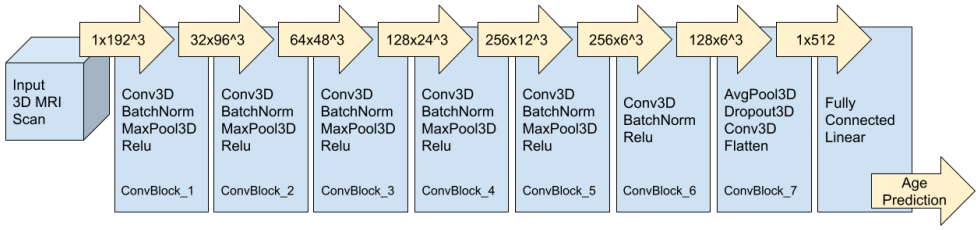
4.3 Method

Our proposed approach is independent of neuroimaging modalities and since T1w sMRI is widely used for AD diagnosis, we thus use 3D T1w MRI scans to estimate brain age. Due to the computational cost of 3D CNN, we developed a lightweight 3D CNN model based on [72]. Our model consists of 3 million learnable parameters, which is more compatible with small dataset sizes and 3D volume data. For comparison, the 3D ResNet-18 has 33.2 million parameters in [22].

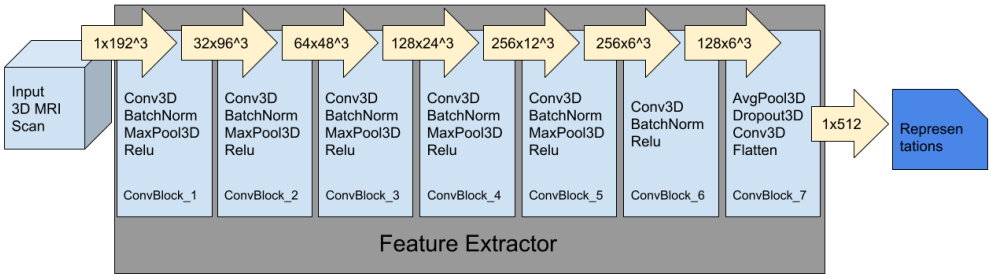
Our proposed model consists of seven 3D convolution blocks. Each one of the first 5 blocks has one Convolutional layer, one BatchNorm layer, one MaxPool and one ReLU layer. Followed by a sixth block without the MaxPooling layer. Then the seventh block of single AvgPool layer, Dropout layer, Convolutional layer, Flatten layer and Linear layer. The final linear layer regress these feature maps to an age prediction. The detail of the proposed of brain age estimation scheme is shown in Fig. 4.2a. After the regression training, the fully connected layer is removed, and then the feature extractor portion of the 3D CNN model is used to generate representations for the scans. The detail of the proposed feature extractor is shown in Fig. 4.2b.

4.4 Experiment Settings

As shown in Fig. 4.3, firstly we train the pretext regression task to minimise the error between the chronological age and estimated brain age on the CN subjects only. We performed 10 runs of regression model training using different random seeds for subject-level train-test-split only on the CN subjects. In



(a) Brain Age Estimation



(b) Brain MRI scan to Representation

Figure 4.2: Overview of the proposed pretext brain age prediction task using 3D convolutional neural networks for Alzheimer's Disease detection. (a) Transforms input 3D MRI scan into a numerical prediction, (b) Generate representations from input scans. Note that the term $ch \times dim\text{^}3$ in the "arrows" denotes the shape of intermediate data between convolutional layers. The "ch" denotes the number of channels, whereas "$dim\text{^}3$" indicates the size of data. All the ConvBlocks are configured to perform $3 \times 3 \times 3$ convolution except the 6^{th} block which has a $1 \times 1 \times 1$ kernel for downsampling purposes.

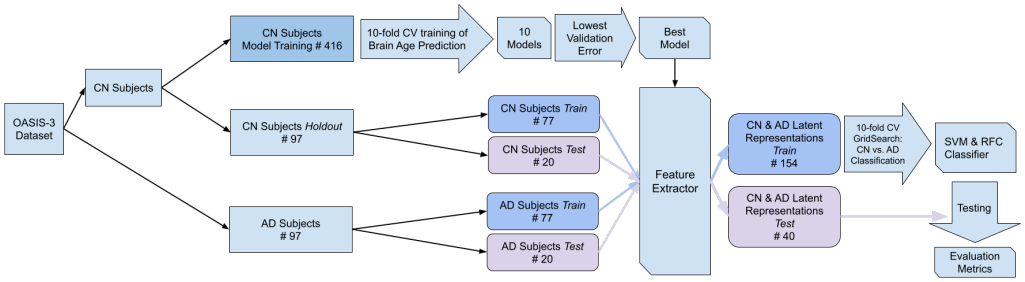


Figure 4.3: Overview of the experimental settings for the brain age-based approach. This figure represents one out of ten runs where each run uses a different random seed for train-test-split. The "#" shows the number of subjects, e.g. CN Subjects Model Training set has 416 subjects for the pretext regression task of feature extractor training. Note that only 97 AD subjects meet the CDR-based selection criteria, so we decided to split the same number of CN subjects into the test set for balance purposes.

each run, we execute 10-fold cross-validation on the CN-only training set. The mean absolute error (MAE) was used as the loss function and the CNN was optimised using Adamw [54] with learning rate = 0.0001 and decay = 0.01, which are chosen based on preliminary experiments. A checkpointing strategy was employed to alleviate overfitting. Specifically, if the validation error does not improve in the current epoch, the checkpoint of saved weights with the lowest validation error will not be updated. The brain age estimation is trained for 100 epochs since we found that our checkpointing strategy normally stops between $40 - 80$ epochs. In each run, we selected the best model with the lowest validation error among the 10 folds.

Subsequently, we conducted experiments to evaluate the proposed feature extractor on scan-level classification performance of AD vs CN subjects using both the best model per run and an ensemble of all models per run. We split the unseen CN holdout set and AD set on subject-level to avoid data leakage. Then the scans of each subject are transformed into latent representations. Followed by 10-fold cross-validation and empirical grid search for hyperparameters while using support vector machines (SVM) and random forester classifier (RFC). To avoid potential type A data leakage, a subject ID grouped split is performed during the cross-validation.

4.5 Results

4.5.1 Brain Age Prediction

The train and validation MAE loss of the best-selected model of each run is shown in Fig. 4.4. Averaging 10 runs, the selected best models have a training MAE loss of 1.6 ± 0.2 years and a validation MAE loss of 3.3 ± 0.4 years. We can observe that the training loss steadily improves, whereas the validation loss has more variation. A possible reason for this behaviour is that the size of the validation set is relatively small compared to the training set.

Then we use the selected best models to predict the brain age of unseen CN

and AD MRI scans. At this point, we split the unseen CN subjects and AD subjects on subject-level, resulting in the train set and test set, as depicted in Fig. 4.3. We can observe that the predicted ages of AD scans are higher than their true ages, whereas the CN scans have a relatively smaller age difference, as shown in Fig. 4.4.

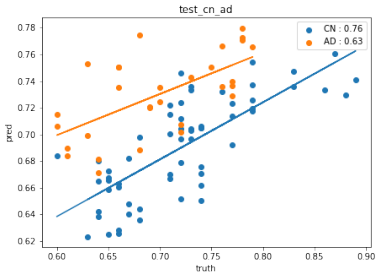
4.5.2 AD vs CN Classification

The classification performance of the ten-run extracted features is shown in Table. 4.2. Each method is repeated for all of the ten runs using SVM and RFC and the final results are averaged. The "Best Model" denotes the classification results using representations generated by the feature extractor with the lowest MAE loss per run, Whereas "All Models" is a majority voting ensemble using all models per run.

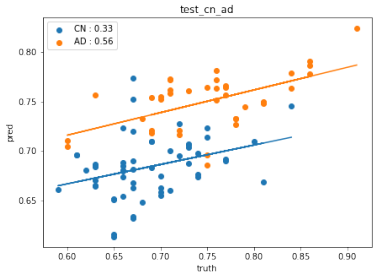
As we can see from Table. 4.2, the overall performance achieved by majority voting of all models is significantly better than using the best model per run. For example, the majority voting RFC ensemble using all models of ten runs achieved 0.847 ± 0.08 ACC and 0.822 ± 0.08 AUC, meanwhile, the majority voting SVM ensemble performed almost the same. The ensembles outperformed their single-model counterpart by roughly 10%.

Our approach achieved the same level of performance compared to three state-of-the-art scan-level data leakage-free approaches [40, 46, 8] for AD/CN classification using the same neuroimaging modality without any data augmentation or complicated pipeline of preprocessing.

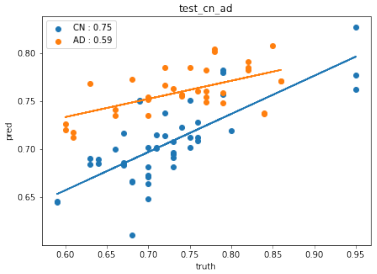
The ADNI dataset is also employed to test the proposed pretext task. As shown in 4.3, the classification performance is not as good as for the OASIS counterpart. One reason could be that the number of subjects (CN: 107, AD: 93) selected from the ADNI datasets is significantly less than in the OASIS experiments, thus the age prediction task cannot train representative feature extractors. The RFC using all models achieved the highest classification ACC, SPE and AUC, a similar trend to the OASIS dataset's results. However, the



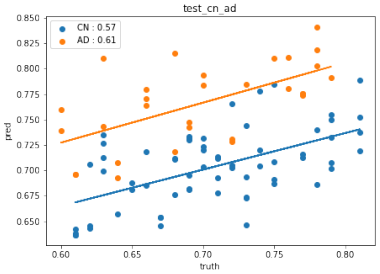
Run 1



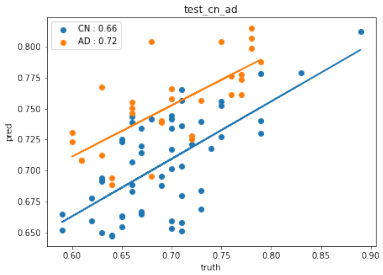
Run 2



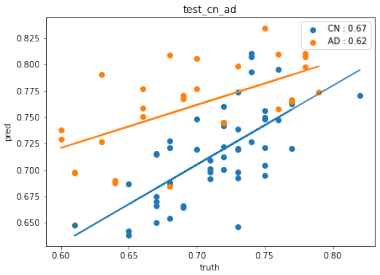
Run 3



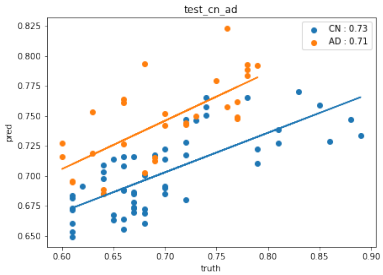
Run 4



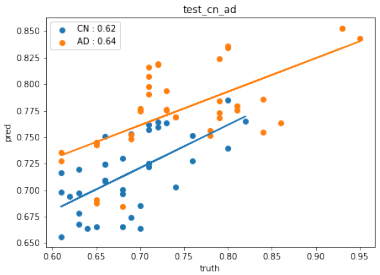
Run 3



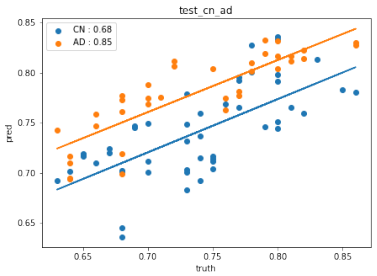
Run 4



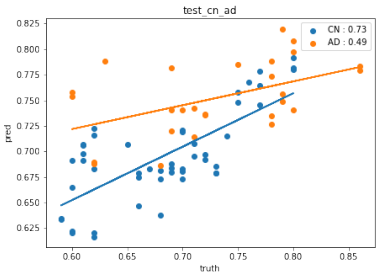
Run 5



Run 6



Run 9



Run 10

Figure 4.4: Scatter plots of predicted age vs true age of test set CN and AD MRI scans using the selected best model per run. The x-axis is the true age, whereas the y-axis shows the predicted age (both been divided by 100).

OASIS Classification	**AD vs. CN**			
	ACC	**SEN**	**SPE**	**AUC**
SVM (Best Model)	0.752 ± 0.06	0.622 ± 0.07	0.834 ± 0.10	0.728 ± 0.06
SVM (All Models)	0.842 ± 0.08	$\mathbf{0.733 \pm 0.12}$	0.908 ± 0.06	0.821 ± 0.08
RFC (Best Model)	0.745 ± 0.07	0.584 ± 0.14	0.844 ± 0.05	0.714 ± 0.07
RFC (All Models)	$\mathbf{0.847 \pm 0.08}$	0.711 ± 0.12	$\mathbf{0.934 \pm 0.05}$	$\mathbf{0.822 \pm 0.08}$

Table 4.2: OASIS classification performance of using age prediction as a pretext to train feature extractors.

classification performance values of the ADNI dataset are considerably lower than the OASIS dataset.

ADNI Classification	**AD vs. CN**			
	ACC	**SEN**	**SPE**	**AUC**
SVM (Best Model)	0.541 ± 0.078	$\mathbf{0.283 \pm 0.246}$	0.668 ± 0.219	0.475 ± 0.065
SVM (All Models)	0.564 ± 0.077	0.168 ± 0.197	0.767 ± 0.185	0.468 ± 0.047
RFC (Best Model)	0.559 ± 0.084	0.195 ± 0.217	0.745 ± 0.177	0.470 ± 0.090
RFC (All Models)	$\mathbf{0.572 \pm 0.124}$	0.145 ± 0.209	$\mathbf{0.785 \pm 0.218}$	$\mathbf{0.489 \pm 0.060}$

Table 4.3: ADNI classification performance of using age prediction as a pretext to train feature extractors.

The baseline experiments using OASIS and ADNI data and the base model are shown in 4.4. The idea is to briefly explore the performance of an end-to-end classification approach without pretext tasks. The base model is composed of the convolution layers in the age prediction model with an added classification output final layer. The train-test-split is on the subject level thus the imbalance of subject numbers. The binary cross entropy loss function is chosen to train the model for 100 epochs. The results are the average and standard deviation of the 10-fold cross-validation on both OASIS and ADNI data.

As shown in the baseline table, both results indicate that the trained model is only predicting the majority class 0 (CN). The accuracy is as high as the proportion of class 0, however, the precision and recall are also 0 for class 1 (AD). The extreme class imbalance might be the reason for the OASIS dataset.

In contrast, classes 0 and 1 are closely represented in the ADNI dataset, the classification results are still biased towards class 0. This poor performance indicates that the baseline approach is not efficient for the AD classification task.

Baseline AD vs. CN	**Classes**	**precision**	**recall**	**f1-score**	**accuracy**
OASIS3	0 (300+) 1 (30+)	0.933 ± 0.012 0.0	1.0 0.0	0.964 ± 0.006 0.0	0.933 ± 0.012
ADNI	0 (50+) 1 (30+)	0.643 ± 0.082 0.0	1.0 0.0	0.767 ± 0.072 0.0	0.644 ± 0.082

Table 4.4: The baseline classification performance using the base model of the brain age prediction model.

4.6 Discussion

4.6.1 Comparison with Other Methods

Instead of classifying directly using 3D MRI scans, we proposed a novel 3D CNN-based pretext brain age prediction task to train the feature extractor for AD detection. The experimental results clearly and convincingly show that the trained feature extractors can generate higher-level discriminative features to achieve comparable classification performance with methods in the literature. The proposed approach has the following advantages:

- The subject-level train-test-split fundamentally eliminates the risk of data leakage.
- Our 3D CNN architecture can fully utilise the spatial information of 3D MRI scans, which completely eliminates the need for hyperparameter tuning for the number of slices and the number or the location of patches.

- Our approach explicitly incorporates the subject's meta-information age into the model learning process, which brings more prior knowledge into the feature extractor.
- Our approach can achieve comparable performance in the absence of complicated preprocessing techniques such as scan registration and scan skull-stripping, therefore greatly reducing the overall complexity.
- Unlike other studies that rely heavily on data augmentation techniques for performance, the classification results demonstrated the robustness of our approach.

4.6.2 Computational Cost

We implemented our proposed approach in Python 3.7 using PyTorch 1.10 version on CUDA 11.3 runtime environment. The mini-batch size is set to 18 so that the mini-batch can be distributed on 3 NVIDIA A100 GPUs. Due to the nature of 3D MRI scans, the computation is significantly higher than 2D images. The average runtime for a single 3D MRI scan to undertake the significant procedures are as follows:

1. Resizing: 2.73 *seconds*.
2. Max-Min Intensity Normalisation: 0.36 *seconds*.
3. CLAHE: 1.37 *seconds*.
4. 3D CNN Brain Age Estimation (batch size 1 for 1 epoch): 0.174 *seconds*.
5. 3D MRI scan to representation: 0.046 *seconds*.

4.6.3 Limitations and Future Work

There are still some limitations to the proposed approach. Firstly, the architecture of the employed 3D CNN model, such as the number and types of convolutional layers, may not be the optimal choice. Secondly, it is not easy

to visualise the extracted features for interpretation of relevant brain changes in clinical practice, and how to visualise these features is an open research question. Thirdly, this study only utilised the T1w MRI scans in the OASIS-3 dataset, which might have limited the learning ability of our proposed approach.

Formulating brain age as a prediction target is inspired by the related work in the medical domain. The main benefit is to utilise the vast amount of unsegmented data to train feature extractors for AD classification. Compared to the direct disease label prediction task, this pretext task is relatively easy to train. This pretext task may open the possibility of adapting other regression models as feature extractors. The features learned during pretext learning may also be transferred and fine-tuned for other brain disease classification tasks, leading to future research potential.

In future work, it is interesting to conduct experiments to assess the effectiveness of 3D MRI augmentation techniques for this approach. Also, the more difficult task of MCI vs CN classification needs to be investigated in the future. Given the available scan modalities (e.g. T2w, PET) and demographic information (e.g. gender, education) in the OASIS-3 dataset, it is interesting to study how to integrate them into our approach. To explore the potential of synthetic neuroimaging data, the next chapter re-uses part of the brain age-based self-supervised pretext task to train feature extractors.

Chapter 5

Representation Learning using Brain Age Prediction

The previous chapter investigated the performance of using brain age prediction as a pretext task on real-world data to train feature extractors for AD classification in a self-supervised setting. In this chapter, the primary motivation is to explore the idea of using synthetic neuroimaging data in the pretext task for the same feature extractor training. This chapter first trains the feature extractor for brain age prediction using the LDM100K synthetic dataset. Then the discriminative performance of the learnt features is evaluated on the real-world OASIS-3 and ADNI datasets, respectively.

5.1 Subject Selection

Contrary to the previous chapter, here we utilise synthetic data from the LDM-100k dataset to train the AutoEncoder. For testing, we selected AD and CN subjects from the OASIS-3 dataset. The overview of the experimental settings for this chapter approach is depicted in Fig.5.1.

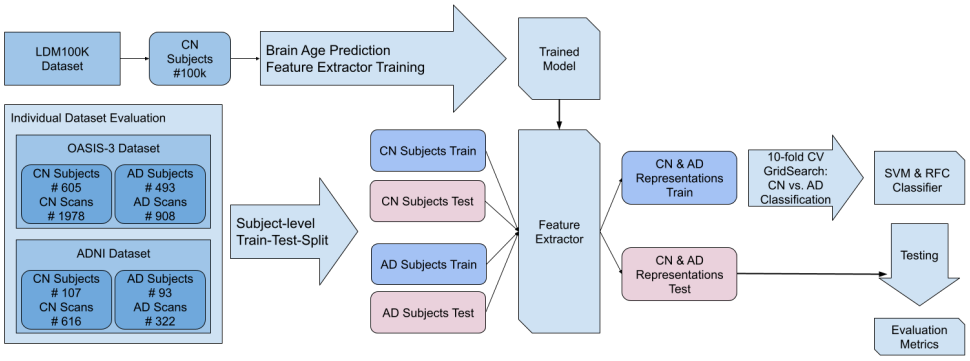


Figure 5.1: Overview of the experimental settings for the brain age-based approach. The "#" shows the number of subjects and scans, e.g. there are 605 CN Subjects with 1,978 scans from the OASIS-3 dataset.

5.1.1 Training Data

As shown in Fig.5.1, the training data is selected from the LMD100K dataset. It is a large dataset recently created by [74]. The authors utilised the latent diffusion models to generate synthetic images from high-resolution 3D brain images. The models are trained to learn about the probabilistic distribution of brain images while conditioned on covariables, such as age, sex, and brain structure volumes.

Given the quantity of data points in LDM100k, it can be used to train large and complex models. However, the high-resolution 3D images are computationally expensive. The previous chapter used a 10-fold cross-validation setup to obtain multiple models. Due to hardware limitations, the training process was reduced to 1-fold for this data.

5.1.2 Testing Data

As shown in Fig.5.1, the testing data are chosen from the OASIS-3 dataset and ADNI dataset. The subject IDs are inconsistent in these two datasets, and there is no documentation regarding the generation of the subject IDs. Therefore, evaluation is performed independently on both datasets to eliminate the change of data leakage. The number of subjects used for testing is signifi-

cantly increased compared to the previous chapter due to the involvement of the LDM100k dataset for training.

For the OASIS dataset, the CDR-based subject selection is refined for this chapter and the following ones. More specifically, each sMRI scan has a timestamp, as well as each CDR assessment. The problem is that they are performed on different dates. We associated each sMRI scan with a CDR score that is closest to the date of the scan. A subject is selected based on the highest CDR score. Then the lower-scored sMRI scans are discarded for a better representation of the real-world situation. For example, a subject is added to the AD group if their CDR score is 2 or 3. The subjects with CDR 0.5 and 1 are also included in the AD group to counter the data imbalance issue. Therefore, the CN group only consist of subjects with CDR 0. In total, there are 605 CN subjects and 493 AD subjects, yielding 1978 and 908 sMRI images, respectively.

From the ADNI database, CN and AD subjects are selected based on the labels given by the ADNI database host. For comparison purposes, the sMRI data are chosen from their screening visit. This is a popular selection criterion in the literature as the screening visit includes more amount of sMRI data per subject than any of the follow-up visits. In total, there are 107 CN subjects and 93 AD subjects, yielding 616 and 322 sMRI images, respectively.

A subject-level train-test-split is carried out during the evaluation for both OASIS-3 and ADNI datasets to address the data leakage issue. However, as there might be more than one sMRI scan per subject, a data leakage type A might occur during the 10-fold cross-validation. For example, the sMRI images of one subject could be split into both the training set and the validation set. To avoid that, a subject ID grouped split is configured for the cross-validation process.

5.2 Preprocessing

As in the previous chapter, the input sMRI images are resized, normalised and contrast equalised for the consistency of experiments. More specifically, the images are processed through the same pipeline of:

1. Resizing: To utilise the 3D scans that have different shapes, and also to remove the excessive background image volume outside the brain tissue, all of them are resampled to a pre-selected shape $200 \times 200 \times 200$.

2. Max-Min Intensity Normalisation: To eliminate the inconsistency of 3D scan voxel intensities, e.g. under or overexposure, the range of intensities is normalised by $I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}}$.

3. Contrast Limited Adaptive Histogram Equalisation (CLAHE) [75]: To enhance the contrast between different brain tissue types, adaptive histogram equalisation is performed on each 3D scan. The contrast-limited variation is chosen to reduce the overamplification of noise.

5.3 Method

5.3.1 Brain Age Prediction

For comparison purposes, we reused the same model (see Fig.4.2) architecture as in the previous chapter. The main difference is the utilisation of the synthetic LDM100K dataset. Due to the computational hardware limitation, the significantly increased cost led to the change of training strategy and hyperparameters. The model is only trained in a 1-fold setting for 50 epochs. The learning rate decays by a factor of 0.5 after every 20 epochs. The size of each mini-batch is set to 15 to fit in the GPU memory. The hyperparameter for training is shown in Table. 5.1.

Based on our preliminary experiments on various data augmentation techniques, random cropping (RandCrop) showed the best improvement in the

classifications. Therefore, we employed random cropping on the fly during training. To maintain comparability to the previous chapter, the 200×200×200 3D volumes are cropped to be the shape of 192 × 192 × 192.

Hyperparameter	**Value**
BatchSize	15
Epochs	50
LearningRate	0.001
LearningRate Step	20
LearningRate Gamma	0.5

Table 5.1: The hyperparameter setting for the brain age prediction training process. The learning rate step and gamma are configured to reduce the learning rate every 20 epochs at a rate of 0.5.

5.3.2 Brain Representation Generation

Once the brain age prediction model has been trained for 50 epochs, the feature extractor generates latent space representations of the test data. The latent space representation is a compressed version of the original data, they also contain the most important features of the data, so it can be used for classification. As shown in Fig.5.1, the real-world train and test subjects are selected from the OASIS-3 and ADNI dataset for evaluation.

5.4 Results

As shown in Table. 5.2, the overall classification performance on the OASIS-3 dataset is worse than in the previous chapter. The best ACC (73.9%), SPE (93.3%), AUC (0.627) and J_stat (0.253) can be found by using the Support Vector Classifier (SVC) on features with RandCrop augmentation. The Random Forester Classifier(RFC) classifier results in a slightly higher SEN on features without RandCrop. One can see that the RFC results are relatively

similar with or without the RandCrop augmentation, while the SVC on the age without augmentation shows the lowest across the metrics.

The classification results on the ADNI dataset show lower performance than the OASIS-3 dataset. As shown in Table. 5.3, the ACC results are similar to each other across the rows. The best SEN, SPE, AUC and J_stat can be found in the row of Age with RandCrop. Three rows are showing 0.0 of SEN with 1.0 of SPE. The AUC values of each row are very close to 0.5. Both SVC and RFC resulted in 0.0 J_stat on feature from Age without RandCrop training. The augmentation technique improved the AUC and J_stat by an insignificant margin.

LDM100K Training OASIS Testing					
SVC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
Age	0.707 +/-0.050	0.278 +/-0.086	0.912 +/-0.044	0.595 +/-0.036	0.190 +/-0.071
Age RandCrop	**0.739 +/-0.037**	0.320 +/-0.052	**0.933 +/-0.025**	**0.627 +/-0.032**	**0.253 +/-0.064**
RFC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
Age	0.724 +/-0.044	**0.321 +/-0.056**	0.919 +/-0.044	0.620 +/-0.026	0.239 +/-0.051
Age Randcrop	0.728 +/-0.052	0.304 +/-0.066	0.925 +/-0.021	0.614 +/-0.035	0.228 +/-0.071

Table 5.2: The classification performance using the feature extractor trained on the LDM100K dataset via the brain age prediction approach tested on the OASIS-3 dataset.

LDM100K Training ADNI Testing					
SVC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
Age	0.653 +/-0.063	0.000 +/-0.000	1.000 +/-0.000	0.500 +/-0.000	0.000 +/-0.000
Age RandCrop	0.643 +/-0.091	0.175 +/-0.285	0.865 +/-0.290	0.520 +/-0.033	0.040 +/-0.067
RFC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
Age	0.653 +/-0.063	0.000 +/-0.000	1.000 +/-0.000	0.500 +/-0.000	0.000 +/-0.000
Age RandCrop	0.635 +/-0.077	0.009 +/-0.020	0.966 +/-0.037	0.488 +/-0.021	-0.025 +/-0.041

Table 5.3: The classification performance using the feature extractor trained on the LDM100K dataset via the brain age prediction approach tested on the ADNI dataset.

5.5 Discussion

5.5.1 Comparison with Other Methods

This chapter trains the feature extractors using a large amount of synthetic neuroimaging data. Using the same base model along with brain chronological age made the results meaningfully comparable with the previous approach. The experimental results demonstrate that the trained feature extractors were able to generate discriminative features. As shown in the previous section, the obtained features are not as good as in the previous chapter, but not far behind. The low sensitivity demonstrates a high number of false positives, whereas the high specificity points to the number of false negatives is low. The J_stat is not much greater than 0 which indicates that the discriminative features need some improvement. One reason could be that the previous chapter trained the model on real-world data and tested the model within the same dataset. Overall the classification performance on the OASIS-3 dataset is more promising than the

ADNI dataset. This chapter contributes the following points:

- The subject-level train-test-split fundamentally eliminates the risk of data leakage.
- The 3D CNN architecture can fully utilise the spatial information of 3D MRI scans, which entirely removes the hyperparameter tuning for the number of slices and the number or the location of patches. This simplification significantly saves the computational cost when the training data is at a large scale.
- It shows the potential of utilising the vast amount of synthetic neuroimaging data in the medical domain.
- The experiment results indicate that random cropping is a promising augmentation technique for the brain age prediction pretext task.

5.5.2 Computational Cost

The computational cost in this chapter has been significantly increased due to the use of the LDM100K dataset. Same as in the previous chapter, the approach is implemented in Python 3.7 using PyTorch 1.10 version on CUDA 11.3 runtime environment. The mini-batch size is set to 15 so that the mini-batch can be distributed on 4 NVIDIA A100 GPUs. The overall training time for 50 epochs on the LMD100K dataset is approximately 39.0 hours. Due to the time limitation, this chapter did not conduct a 10-fold cross-validation for ensemble classification.

5.5.3 Limitations and Future Work

This chapter builds on previous work by exploring the use of synthetic neuroimaging data to train feature extractors for real-world AD classification tasks. Inherently, the architecture of the 3D model is limited by the available computational hardware. The interpretability of the learnt features has

been further complicated as the feature extractors are trained on synthetic data. The large amount of synthetic data significantly increases the training cost. Also, the significant increase in computational cost made training multiple feature extractors as part of an ensemble infeasible. To further explore the potential of synthetic neuroimaging data, the next chapter investigates an AutoEncoder-based self-supervised pretext task for feature extractor training.

Chapter 6

Representation Learning using Brain Image Reconstruction

The previous chapter investigated the performance of using brain age prediction as a pretext task to train feature extractors for AD classification in a self-supervised setting. In this chapter, the primary motivation is to explore the idea of using an AutoEncoder as a self-supervised pretext task for feature extractor training.

AutoEncoders are widely used as a self-supervised method to extract information from unlabelled data. The idea is to transform the input data into a latent space representation followed by reconstruction as output. A CNN-based AutoEncoder typically utilises convolution layers in the encoder as CNN classifiers, whereas the decoder consists of deconvolution layers. The generality of the representation is learned by minimising the error between the original input and the reconstructed input. Then encoder component is utilised as a feature extractor for the downstream task. This chapter contributes towards the exploration of using AutoEncoders to learn representation for Alzheimer's Disease classification in the context of self-supervised training.

6.1 Subject Selection

As shown in Fig.6.1, the training subjects selection is identical to the previous chapter for ease of comparison between different approaches. The training data is the full LMD100K synthetic dataset while the testing data are chosen from the OASIS-3 and ADNI dataset, respectively. Again, the selection of testing subjects follows the same protocol as the previous chapter to eliminate any chance of data leakage. The train-test-split during the feature generation process is on the subject level as well as the classifier cross-validation during testing.

As shown in Fig.6.1, the training and testing subjects selection is identical to the previous chapter for ease of comparison between different approaches. For testing, we evaluated the feature extractors using subjects from the OASIS-3 dataset and ADNI dataset independently.

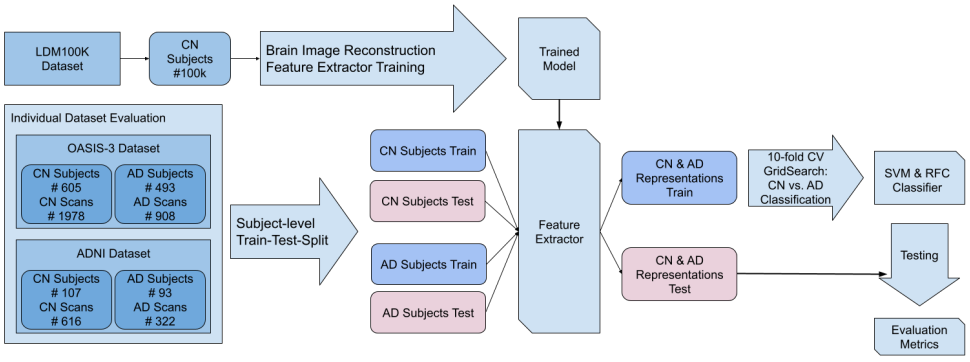


Figure 6.1: Overview of the experimental settings for the AutoEncoder-based approach. The "#" shows the number of subjects and scans, e.g. there are 605 CN Subjects with 1,978 scans from the OASIS-3 dataset.

6.2 Preprocessing

As in the previous chapter, the input sMRI images are resized, normalised and contrast equalised for the consistency of experiments. More specifically, the images are processed through the same pipeline as in the previous chapter:

1. Resizing: To utilise the 3D scans that have different shapes, and also to remove the excessive background image volume outside the brain tissue, all of them are resampled to a pre-selected shape $200 \times 200 \times 200$.

2. Max-Min Intensity Normalisation: To eliminate the inconsistency of 3D scan voxel intensities, e.g. under or overexposure, the range of intensities is normalised by $I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}}$.

3. Contrast Limited Adaptive Histogram Equalisation (CLAHE) [75]: To enhance the contrast between different brain tissue types, adaptive histogram equalisation is performed on each 3D scan. The contrast-limited variation is chosen to reduce the overamplification of noise.

6.3 Method

6.3.1 Brain sMRI Reconstruction

As this reconstruction approach is independent of neuroimaging modalities and since T1w sMRI is widely used for AD diagnosis, we thus use 3D T1w MRI scans for AutoEncoder reconstruction. For comparison purposes, we used the same lightweight model with a 3 million parameter from the age prediction model as the encoder. Then we develop its deconvolution counterpart as the decoder, which consists of 3.2 million learnable parameters.

The encoder network consists of seven 3D convolution blocks. Each one of the first 5 blocks has one Convolutional layer, one BatchNorm layer, one MaxPool and one ReLU layer. The sixth block has no MaxPooling layer while the seventh block consists of a single AvgPool layer, a Dropout layer, a Convolutional layer, and a Flatten layer. The flattened representations are then fed into the decoder. Th detail of the proposed brain sMRI reconstruction sche is shown in Fig. 6.2a. After the reconstruction training, the decoder is removed, and then the encoder is used as a feature extractor to generate representations for testing. The detail of the proposed feature extractor is

shown in Fig. 6.2b.

The hyperparameter for training is shown in Table.6.2b. Due to the increased complexity of the AutoEncoder model, the size of the mini-batch is decreased to 12 to fit in the GPU memory. For the same reason, the training process was reduced to 1-fold and 50 epochs.

Hyperparameter	**Value**
BatchSize	12
Epochs	50
LearningRate	0.001
LearningRate Step	20
LearningRate Gamma	0.5

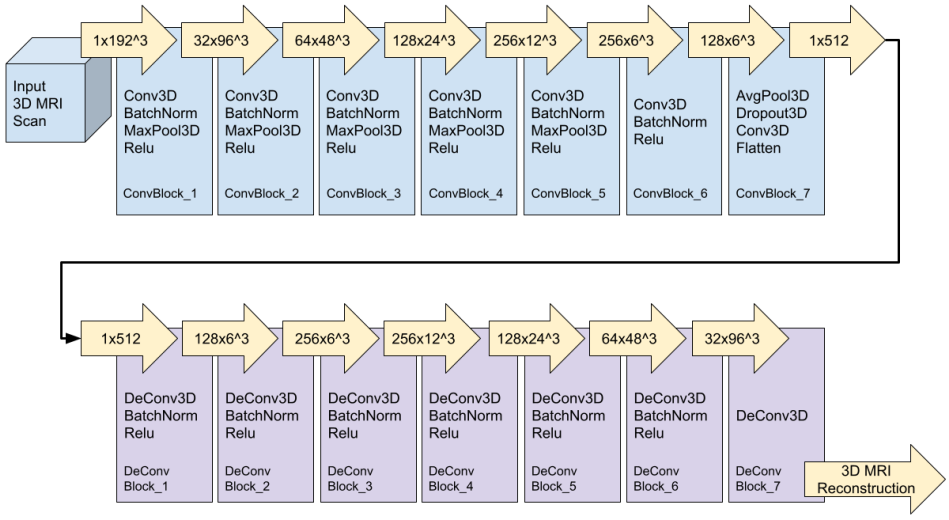
Table 6.1: The hyperparameter setting for the AutoEncoder training process. The learning rate step and gamma are configured to reduce the learning rate every 20 epochs at a rate of 0.5.

6.3.2 Brain Representation Generation

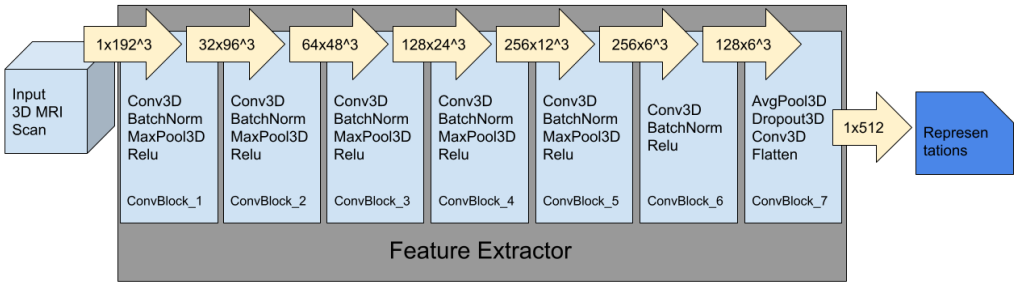
Once the AutoEncoder training is finished after 50 epochs, the encoder network is used to generate latent space representations of the test data. The latent space representation is a compressed version of the original data, they also contain the most important features of the data, so it can be used for classification. As shown in Fig.6.1, the real-world train and test subjects are selected from the OASIS-3 and ADNI dataset for evaluation.

6.4 Results

Using the metrics described in Table. 2.2, the evaluation results of the AutoEncoder-based feature extractor training method are shown in Table. 6.2. On the OASIS-3 dataset, the SVC outperformed the RFC on almost all metrics. The RandCrop helped SVC achieve the highest ACC of 73.8% with less standard



(a) Brain MRI Reconstruction



(b) Brain MRI to Representation

Figure 6.2: Overview of the proposed pretext brain image reconstruction task using 3D convolutional neural networks for Alzheimer's Disease detection. (a) compress the input 3D MRI scan into a latent representation, followed by a reconstruction, (b) Generate representations from input scans. Note that the term $ch \times dim\text{^}3$ in the "arrows" denotes the shape of intermediate data between convolutional layers. The "ch" denotes the number of channels, whereas "$dim\text{^}3$" indicates the size of data. All the ConvBlocks are configured to perform $3 \times 3 \times 3$ convolution except the 6^{th} block which has a $1 \times 1 \times 1$ kernel for downsampling purposes.

deviation than other combinations of techniques. The same trend can be observed in SEN, AUC and J_stat while the SPE is slightly behind the RFC with RandCrop.

Unfortunately, both SVC and RFC performed less than ideal on the ADNI dataset. As shown in Table. 6.3, all the SEN values are 0 and most of SPE values are near 1. Most of AUC values are around 0.5 while the values of J_stat are zero. The RandCrop augmentation improved the results of RFC by a tiny amount.

LDM100K Training OASIS Testing					
SVC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
AutoEncoder	0.738 +/-0.046	0.402 +/-0.056	0.897 +/-0.049	0.649 +/-0.037	0.299 +/-0.073
AutoEncoder RandCrop	**0.738 +/-0.037**	**0.417 +/-0.088**	0.892 +/-0.042	**0.654 +/-0.042**	**0.309 +/-0.084**
RFC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
AutoEncoder	0.721 +/-0.045	0.314 +/-0.093	0.909 +/-0.051	0.612 +/-0.040	0.223 +/-0.079
AutoEncoder RandCrop	0.712 +/-0.050	0.266 +/-0.096	**0.927 +/-0.052**	0.597 +/-0.037	0.194 +/-0.075

Table 6.2: The classification performance using the feature extractor trained on the LDM100K dataset via the brain image reconstruction approach tested on the OASIS-3 dataset.

6.5 Discussion

6.5.1 Comparison with Other Methods

Instead of classifying directly using real-world 3D MRI scans, we proposed a 3D CNN-based AutoEncoder reconstruction pretext to train the feature extractor using synthetic data for AD classification. The experimental results

LDM100K Training ADNI Testing					
SVC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
AutoEncoder	0.653 +/-0.063	0.000 +/-0.000	1.000 +/-0.000	0.500 +/-0.000	0.000 +/-0.000
AutoEncoder RandCrop	0.653 +/-0.063	0.000 +/-0.000	1.000 +/-0.000	0.500 +/-0.000	0.000 +/-0.000
RFC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
AutoEncoder	0.653 +/-0.063	0.000 +/-0.000	1.000 +/-0.000	0.500 +/-0.000	0.000 +/-0.000
AutoEncoder RandCrop	0.647 +/-0.073	0.000 +/-0.000	0.989 +/-0.032	0.495 +/-0.016	-0.011 +/-0.032

Table 6.3: The classification performance using the feature extractor trained on the LDM100K dataset via the brain image reconstruction approach tested on the ADNI dataset.

demonstrate that the trained feature extractors were able to generate informative features, which resulted in comparable classification performance. The proposed approach has the following advantages:

- The subject-level train-test-split fundamentally eliminates the risk of data leakage.
- Using each 3D MRI as a whole fully incorporates the spatial information into the learning scheme.
- Using AutoEncoder for image reconstruction needs no clinical information or human inputs thus it is well-suited for self-supervised learning.
- AutoEncoder is an effective choice to reduce the dimensionality of input data while preserving the most important information.
- The random cropping augmentation techniques can improve classification performance in the brain image reconstruction pretext task.

6.5.2 Computational Cost

Similar to the previous chapter, the LDM100K dataset significantly increased the training time. On top of that, the decoder component in the AutoEncoder network has doubled the learnable parameters. The mini-batch size has to be reduced to 15 so that the mini-batch can be distributed on 4 NVIDIA A100 GPUs. The overall training time for 50 epochs on the LMD100K dataset is approximately 43 hours. Because of time constraints, this chapter did not perform a 10-fold cross-validation for ensemble classification.

6.5.3 Limitations and Future Work

As shown in the previous subsection, the increased computation cost of the AutoEncoder-based approach results in expensive training time. Also, the sheer amount of training data from LDM100k 3D significantly boosts the training time. Last but not least, the learned representations in AutoEncoders can be difficult to interpret, which can make it difficult to understand how the model is making its predictions. This can be a problem for tasks where interpretability is important, such as medical diagnosis.

Limited by the availability of computational hardware, the architecture of the AutoEncoder has to be reduced. It is interesting to explore the optimal architecture with more learning capability when more computational resource is at the disposal. Another promising area of future work is model compression. This involves reducing the size of a model without significantly impacting its accuracy. The performance of the AutoEncoder-based features encourages the possibility of using synthetic data to improve data-scarce training tasks. Improving the quality of synthetic data is another exciting area to be explored in the future.

Chapter 7

Representation Learning using Brain sMRI Rotation Classification

The previous chapter investigated the performance of using AutoEncoder-based brain MRI reconstruction as a pretext task to train feature extractors for AD classification. In this chapter, the primary motivation was to explore the idea of using 3D MRI rotation classification as a pretext task for feature extractor training. This chapter further extends the utilisation of brain sMRI data to learn discriminative features for real-world AD classification.

Another popular self-supervised training approach is the classification of rotation. The idea is to engineer a pretext task by rotating an input image to certain degrees and then train the feature extractor to identify the degrees of rotation. The underlying assumption is that if an image is rotated, the semantic content of the image should remain the same, but the appearance of the image should change. Commonly the rotation of a 2D image is set to 90, 180 and 270 degrees to formulate a classification task. For 3D images, the possibility of 90 degrees of rotation can be around 3 different and independent axis, thus the total number of classes is 64 classes. This part is further discusses in the next section.

7.1 Subject Selection

As shown in Fig.7.1, the training data is the entire LMD100K synthetic dataset while the testing is conducted on subjects selected from the OASIS-3 and ADNI datasets, separately. The purpose is to make the experimental results comparable across various approaches. Similar to the previous chapter, the split into train and test subjects is performed before training and testing to avoid data leakage.

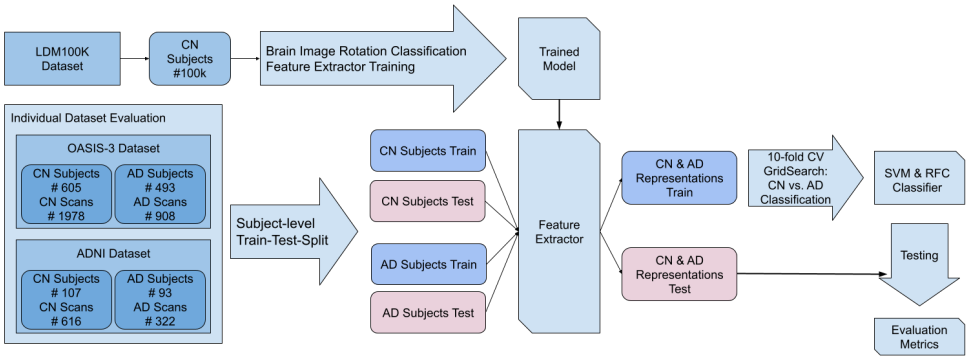


Figure 7.1: Overview of the experimental settings for the AutoEncoder-based approach. The "#" shows the number of subjects and scans, e.g. there are 493 AD Subjects with 908 scans from the OASIS-3 dataset.

7.1.1 Rotation Generation

Rotating 2D images as a pretext task is commonly used in the literature. This category of methods utilises the structural information of data itself and defines pretext tasks that relate to the final application. Rotation computation is fully automatic, not needing any human input, thus readily available image datasets can be utilised for training. Features learned in this method reportedly generalise well in the downstream tasks and achieve state-of-the-art performance [27, 91, 57].

A 2D image can be rotated around the centre of the image for a given degree. For the human eye, the rotation can be developed in clockwise or counter-clockwise directions. The 3D MRI images can be rotated in a similar

way: each dimension can be seen as a stack of 2D images/slices then all dimensions are rotated around a shared centre point. Therefore, the rotation of a 3D brain image can be performed by 3 separate 2D rotations in each plane.

As shown in Fig.7.2, the first row is a 3D MRI in its original orientation, which we conveniently assign as class 0. The number in the brackets indicates the number of 90-degree clockwise rotations performed in each plane. For example, the second row shows $(0, 0, 1)$ where the third dimension is rotated 90 degrees clockwise once, as class 1. Not all 64 rotations lead to unique images, therefore only the first 32 out of 64 possible combinations of rotations are considered as classes in this thesis. Then the remaining 32 classes are developed accordingly.

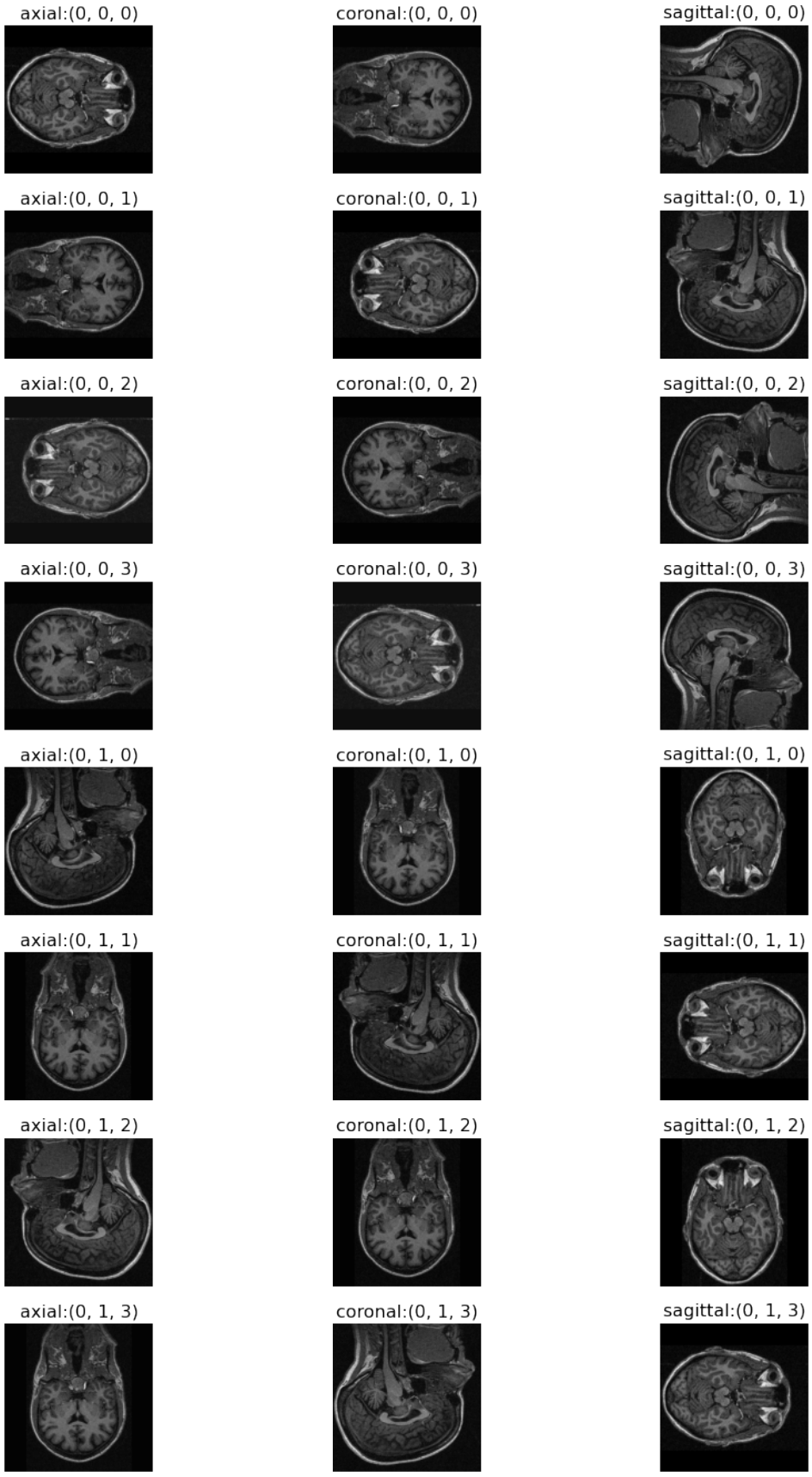
7.2 Preprocessing

Same as in the previous chapter, the input sMRI images are resized, normalised and contrast equalised for the consistency of experiments. More specifically, the images are processed through the same pipeline of:

1. Resizing: To utilise the 3D scans that have different shapes, and also to remove the excessive background image volume outside the brain tissue, all of them are resampled to a pre-selected shape $200 \times 200 \times 200$.

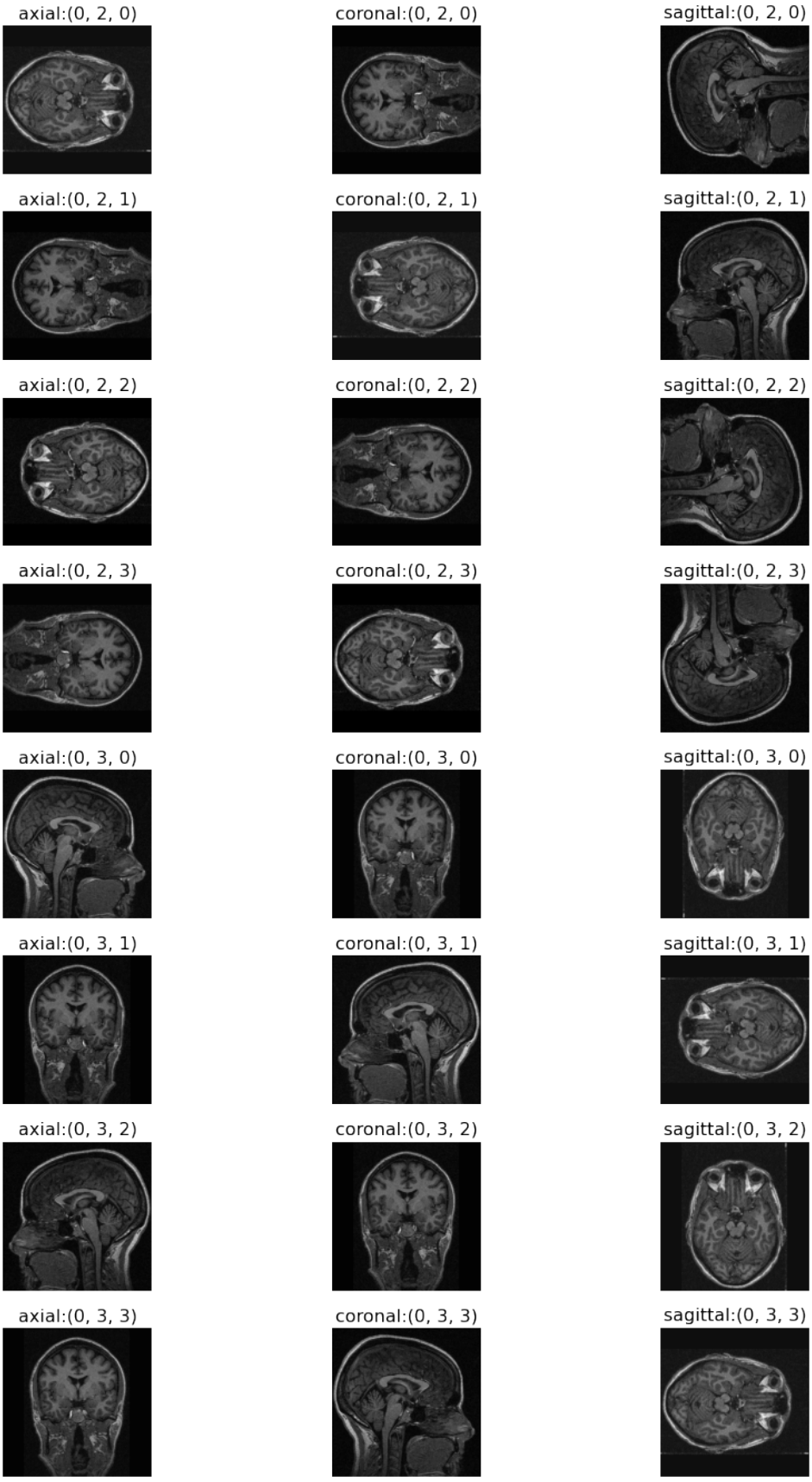
2. Max-Min Intensity Normalisation: To eliminate the inconsistency of 3D scan voxel intensities, e.g. under or overexposure, the range of intensities is normalised by $I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}}$.

3. Contrast Limited Adaptive Histogram Equalisation [75]: To enhance the contrast between different brain tissue types, adaptive histogram equalisation is performed on each 3D scan. The contrast-limited variation is chosen to reduce the overamplification of noise.



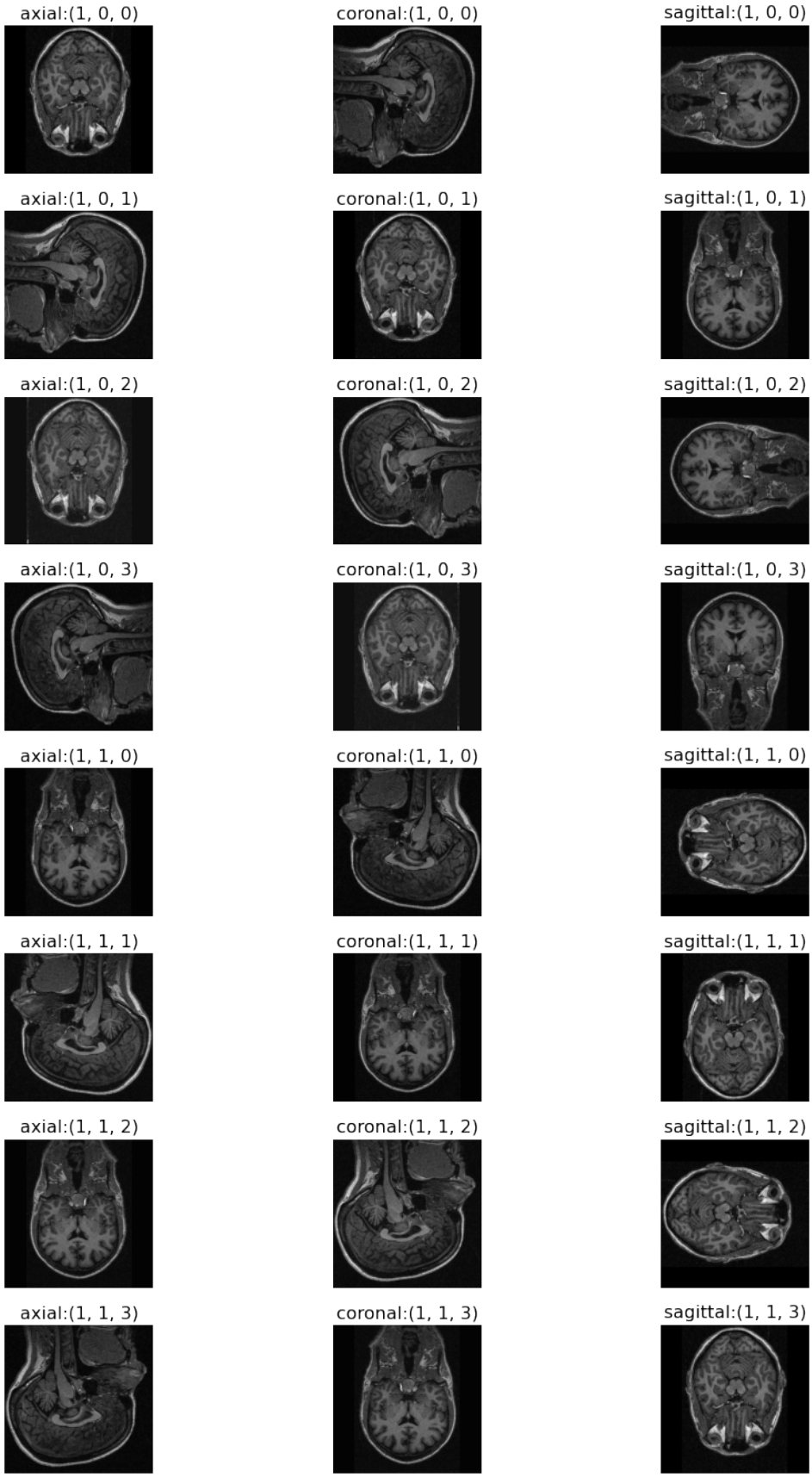
(a)

Figure 7.2: Examples of Brain sMRI Rotation. Part (a) shows classes 1 to 8.



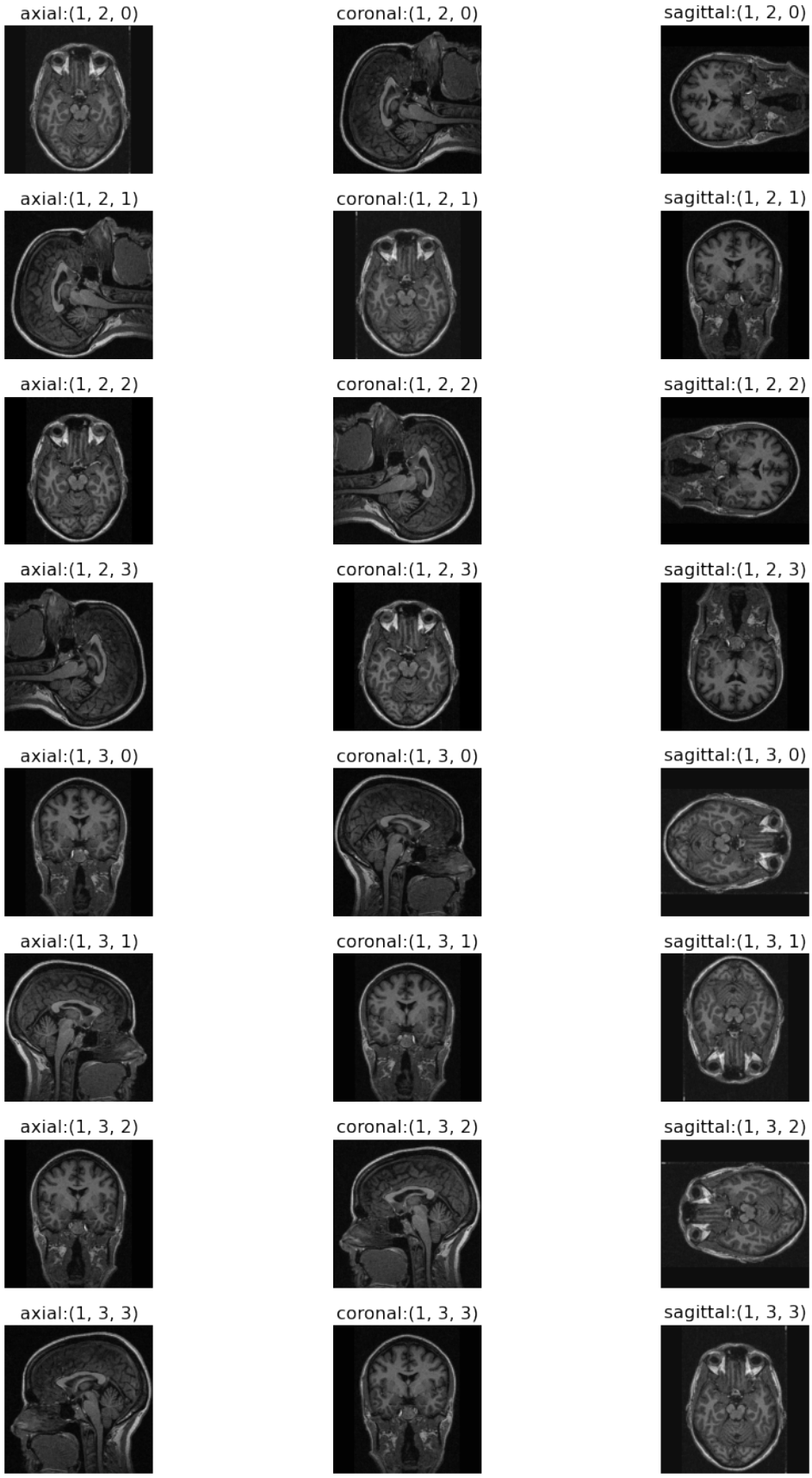
(b)

Figure 7.2: Examples of Brain sMRI Rotation. Part (b) shows classes 9 to 16.



(c)

Figure 7.2: Examples of Brain sMRI Rotation. Part (c) shows classes 17 to 24.



(d)

Figure 7.2: Examples of Brain sMRI Rotation. Part (d) shows classes 25 to 32.

7.3 Method

7.3.1 Brain Rotation Classification

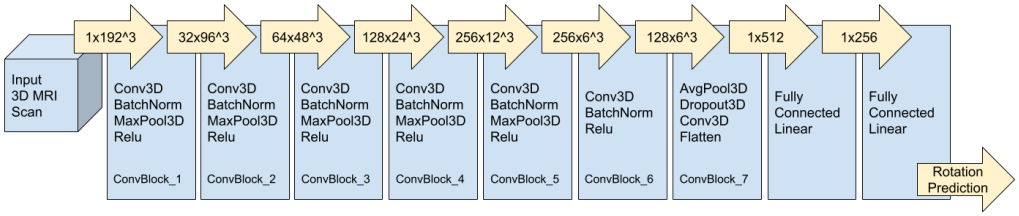
As this reconstruction approach is independent of neuroimaging modalities and since T1w sMRI is widely used for AD diagnosis, we thus use 3D T1w MRI scans for brain image rotation classification. For comparison purposes, we used the same 3 million parameter lightweight model as the feature extractor, followed by a fully connected layer and a softmax activation. The details of the architecture are depicted in Fig.7.3. The hyperparameters for training are listed in Table.7.1.

Hyperparameter	**Value**
BatchSize	20
Epochs	50
LearningRate	0.001
LearningRate Step	20
LearningRate Gamma	0.5

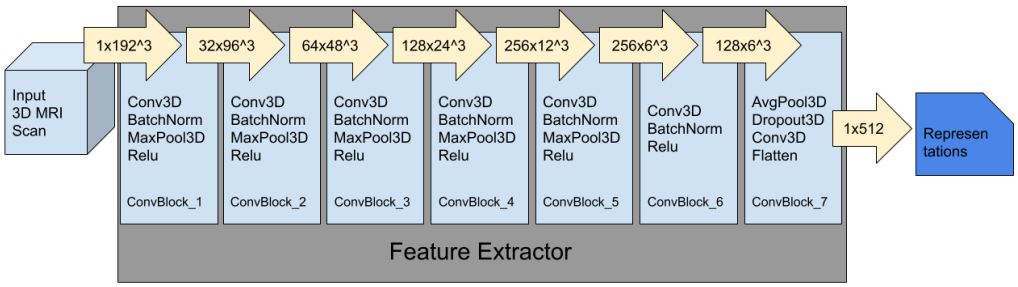
Table 7.1: The hyperparameter setting for the brain image rotation classification training process. The learning rate step and gamma are configured to reduce the learning rate every 20 epochs at a rate of 0.5.

7.3.2 Brain Representation Generation

After 50 epochs of training, the classifier network (without softmax layer) is used to generate latent space representations of the test data. The representations are then used for the downstream classification of CN vs AD subjects. As shown in Fig.7.1, the real-world train and test subjects are selected from the OASIS-3 and ADNI dataset for evaluation.



(a) Brain MRI Reconstruction



(b) Brain MRI to Representation

Figure 7.3: Overview of the proposed pretext brain rotation classification task using 3D convolutional neural networks for Alzheimer's Disease detection. (a) compress the input 3D MRI scan into a latent representation, followed by classification, (b) Generate representations from input scans. Note that the term $ch \times dim\text{^}3$ in the "arrows" denotes the shape of intermediate data between convolutional layers. The "ch" denotes the number of channels, whereas "$dim\text{^}3$" indicates the size of data. All the ConvBlocks are configured to perform $3 \times 3 \times 3$ convolution except the 6^{th} block which has a $1 \times 1 \times 1$ kernel for downsampling purposes.

LDM100K Training OASIS Testing					
SVC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
Rotation	0.738 +/-0.046	0.402 +/-0.056	0.897 +/-0.049	0.649 +/-0.037	0.299 +/-0.073
Rotation RandCrop	**0.738 +/-0.037**	**0.417 +/-0.088**	0.892 +/-0.042	**0.654 +/-0.042**	**0.309 +/-0.084**
RFC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
Rotation	0.721 +/-0.045	0.314 +/-0.093	0.909 +/-0.051	0.612 +/-0.040	0.223 +/-0.079
Rotation RandCrop	0.712 +/-0.050	0.266 +/-0.096	**0.927 +/-0.052**	0.597 +/-0.037	0.194 +/-0.075

Table 7.2: The classification performance using the feature extractor trained on the LDM100K dataset via the brain image rotation classification approach tested on the OASIS-3 dataset.

7.4 Results

Using the same metrics as in the previous chapter and as described in Table. 2.2, we evaluate the effect of the Rotation-based feature extraction on the OASIS-3 dataset first. As shown in Table. 7.2, the highest ACC of 73.8% is achieved by SVC in conjunction with RandCrop augmentation. The highest SEN (41.7%), AUC (0.654) and J_stat (0.309) are also obtained by this combination. The RFC with RandCrop resulted in the highest SPE of 92.7%.

The evaluation results on the ADNI dataset are displayed in Table. 7.3. Similar to the previous chapter, both SVC and RFC performed poorly on the ADNI dataset. The ACC is just a tiny bit higher than random guesses. ALL the SEN values are close to 0 and most of SPE values are near 1. Most of the AUC values are around 0.5 while the values of J_stat are zero. The RandCrop augmentation improved the results of RFC by a small margin.

LDM100K Training ADNI Testing					
SVC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
Rotation	0.649 +/-0.062	0.000 +/-0.000	0.994 +/-0.017	0.497 +/-0.008	-0.006 +/-0.017
Rotation RandCrop	0.651 +/-0.064	0.007 +/-0.015	0.993 +/-0.015	0.500 +/-0.005	0.001 +/-0.010
RFC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
Rotation	0.643 +/-0.059	0.003 +/-0.010	0.983 +/-0.023	0.493 +/-0.014	-0.014 +/-0.028
Rotation RandCrop	0.651 +/-0.060	0.013 +/-0.021	0.990 +/-0.019	0.502 +/-0.014	0.004 +/-0.029

Table 7.3: The classification performance using the feature extractor trained on the LDM100K dataset via the brain image rotation classification approach tested on the ADNI dataset.

7.5 Discussion

7.5.1 Comparison with Other Methods

Unlike traditional classification tasks predicting class labels, self-supervision requires the fabrication of artificial labels as prediction targets. This chapter presents an approach that fabricates rotation-based labels as a self-supervised pretext task to train feature extractors for AD classification. The evaluation results support that the trained feature extractors were able to learn discriminative information, in turn showing reasonable classification performance. The proposed approach has the following advantages:

- The subject-level train-test-split fundamentally eliminates the risk of data leakage.
- This self-supervised learning approach generates 90 degrees of rotations as classification labels. There is no need for explicit class labels or segmentation of disease lesions.

- Using each 3D MRI as a whole fully incorporates the spatial information into the learning scheme.
- The random cropping augmentation techniques can improve classification performance in the rotation classification pretext task.

7.5.2 Computational Cost

As the architecture is similar to the brain age prediction approach, the amount of learnable parameters is also relatively the same. Similarly, the network is implemented in Python 3.7 using PyTorch 1.10 version on CUDA 11.3 runtime environment. The mini-batch size is set to 20 so that the mini-batch can be distributed on 4 NVIDIA A100 GPUs. As the 3D MRI rotation operation is performed on the fly during training, the overall runtime is increased to approximately 57.3 hours for 50 epochs on the LMD100K dataset. Due to time limitations, this chapter was unable to conduct a 10-fold cross-validation for ensemble classification.

7.5.3 Limitations and Future Work

There are two possible main limitations of the rotation-based training methods for medical data. The first one might be the sensitivity to image scale and intensity. This means that if an image is scaled up or down, or if the illumination of an image is changed, the model may not be able to learn accurate features. One feasible solution could be to engineer additional pretext tasks for such changes. For example, instead of simply predicting the rotation angle of an image, the model could be trained to predict the scale and illumination of an image as well. This would require the model to learn features that are invariant to different types of transformations, which is an intriguing topic to research.

Another limitation might be the sensitivity to noise which has made it difficult to learn invariant features on datasets with a lot of noise. The common

inherent noises with MRI images are fat and blood tissues or slight body movements during the image acquisition process resulting in misalignment. Noise reduction techniques such as brain tissue extraction or skull removal are interesting research directions for the future.

Chapter 8

Representation Learning using Multi-Head Tasks

This chapter explores the possibility of feature extractor training by utilising the multi-head configuration of brain age prediction, brain rotation classification, and brain image reconstruction. As the experiments of the previous chapters suggested, there is a potential improvement to be had by combining all the tasks into one multi-head task to train better feature extractors.

Engineering multi-head tasks is a widely used method to utilise different modalities of inputs such as images, text, and numeric and nominal values. For example, functional MRI, blood tests and other clinical assessments are often collected along with structural MRI acquisition. A multi-head task can incorporate this information for feature extractor training. However, the additional model complexity can result in significant computational costs. Therefore, the design of multi-head tasks normally takes incremental steps during the design process. This thesis designed the multi-head task as a combination of all the methods from the four previous chapters.

8.1 Subject Selection

Same as in the previous chapter, we utilised the synthetic data from the LDM-100k dataset to train the rotation classifier. We also selected the same AD and

CN subjects from the OASIS and ADNI datasets for testing. The overview of the experimental settings for the rotation classification approach is shown in Fig.8.1.

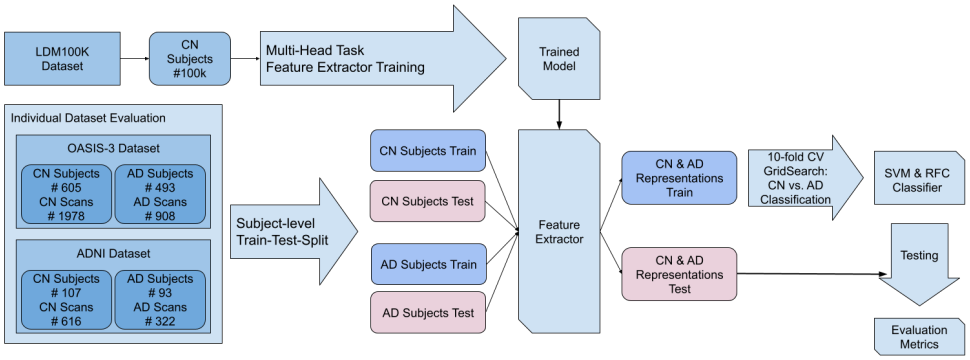


Figure 8.1: Overview of the experimental settings for the AutoEncoder-based approach. The "#" shows the number of subjects and scans, e.g. there are 107 CN Subjects with 616 scans and 93 AD Subjects with 322 scans from the ADNI dataset.

8.2 Preprocessing

Same as in the previous chapters, the input sMRI images are resized, normalised and contrast equalised for the consistency of experiments. More specifically, the images are processed through the same pipeline of:

1. Resizing: To utilise the 3D scans that have different shapes, and also to remove the excessive background image volume outside the brain tissue, all of them are resampled to a pre-selected shape $200 \times 200 \times 200$.

2. Max-Min Intensity Normalisation: To eliminate the inconsistency of 3D scan voxel intensities, e.g. under or overexposure, the range of intensities is normalised by $I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}}$.

3. Contrast Limited Adaptive Histogram Equalisation [75]: To enhance the contrast between different brain tissue types, adaptive histogram equalisation is performed on each 3D scan. The contrast-limited variation is chosen to reduce the overamplification of noise.

8.3 Method

8.3.1 Multi-head Tasks

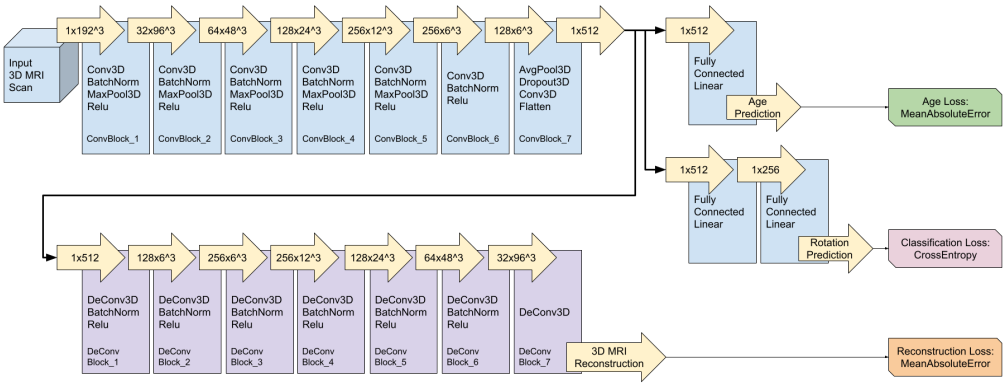
The details of the architecture are depicted in Fig.8.2. The idea is to combine different types of output heads into one base feature extractor. In our case, we combined the brain age prediction task, the brain image reconstruction task and the brain image rotation classification task as three output heads. In the case of the reconstruction task, the output head is the decoder part. For comparison purposes, we used the same 3 million parameter lightweight model as the base feature extractor. While training the Multi-head task, the base feature extractor is shared between three different output heads. Each mini-batch is fed into each of the heads where the order of the head is randomly chosen during training. The parameters of the base model are updated according to each loss of the head task. The hyperparameters for training are listed in Table.8.1. Due to the added decoder network of the reconstruction task, the mini-batch size is reduced to 12 due to GPU memory limitations.

Hyperparameter	**Value**
BatchSize	12
Epochs	50
LearningRate	0.001
LearningRate Step	20
LearningRate Gamma	0.5

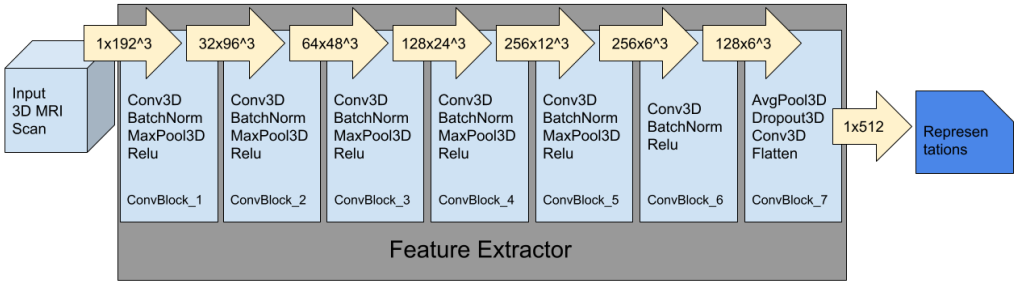
Table 8.1: The hyperparameter setting for the Multi-Head training process. The learning rate step and gamma are configured to reduce the learning rate every 20 epochs at a rate of 0.5.

8.3.2 Brain Representation Generation

After 50 epochs of training, the feature extractor (without the final output heads) is used to generate latent space representations of the test data. The



(a) Brain MRI Reconstruction



(b) Brain MRI to Representation

Figure 8.2: Overview of the proposed pretext multi-head task using 3D convolutional neural networks for Alzheimer's Disease detection. (a) compress the input 3D MRI scan into a latent representation, followed by the brain age prediction, brain image reconstruction and brain rotation classification, (b) Generate representations from input scans. Note that the term $ch \times dim\text{^}3$ in the "arrows" denotes the shape of intermediate data between convolutional layers. The "ch" denotes the number of channels, whereas "$dim\text{^}3$" indicates the size of data. All the ConvBlocks are configured to perform $3 \times 3 \times 3$ convolution except the 6^{th} block which has a $1 \times 1 \times 1$ kernel for downsampling purposes.

representations are then used for the downstream classification of CN vs AD subjects. As shown in Fig.8.1, the real-world train and test subjects are selected from the OASIS-3 and ADNI datasets for evaluation.

8.4 Results

Using the same metrics as in the previous chapter and as described in Table. 2.2, the quality of learned features are evaluated on the OASIS-3 and ADNI datasets. As shown in Table. 8.2, the highest ACC of 73.3% is achieved by SVC in conjunction with RandCrop augmentation. This classifier with RandCrop also resulted in the highest SEN (37.9%), AUC (0.636) and J_stat (0.272) while the highest SPE of 94.4% is obtained by RFC without data augmentation.

The evaluation results on the ADNI dataset are shown in Table. 8.3. Same as in the previous chapter, the discriminative features learned by the Multi-Head task are far from ideal. The ACC is not much better than random guesses. Most of the SEN values are close to 0 and all of SPE values are near 1. The AUC values are fluctuating around 0.5 while the J_stat values swing near 0. Similar to the trend on the OASIS-3 dataset, a small improvement can be seen in the SVC with RandCrop augmentation.

8.5 Discussion

8.5.1 Comparison with Other Methods

This chapter combines the approaches of the previous chapters into a Multi-Head task to train feature extractors for AD classification. As each individual task is self-supervised, the combined task is also a self-learning task. The evaluation performance suggests that the Multi-Head task could learn discriminative features. Although the AD classification results are comparable to the previous chapters, there is no significant improvement despite the increased

LDM100K Training OASIS Testing					
SVC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
Multi-Head	0.719 +/-0.045	0.342 +/-0.082	0.903 +/-0.053	0.623 +/-0.041	0.245 +/-0.083
Multi-Head RandCrop	**0.733 +/-0.031**	**0.379 +/-0.080**	0.894 +/-0.028	**0.636 +/-0.037**	**0.272 +/-0.073**
RFC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
Multi-Head	0.684 +/-0.052	0.133 +/-0.066	**0.944 +/-0.039**	0.539 +/-0.021	0.077 +/-0.041
Multi-Head RandCrop	0.702 +/-0.053	0.243 +/-0.067	0.910 +/-0.039	0.577 +/-0.025	0.154 +/-0.049

Table 8.2: The classification performance using the feature extractor trained on the LDM100K dataset via the Multi-Head approach tested on the OASIS-3 dataset.

LDM100K Training ADNI Testing					
SVC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
Multi-Head	0.618 +/-0.068	0.057 +/-0.117	0.927 +/-0.131	0.492 +/-0.040	-0.016 +/-0.079
Multi-Head RandCrop	0.661 +/-0.052	0.144 +/-0.104	0.930 +/-0.047	0.537 +/-0.038	0.075 +/-0.075
RFC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
Multi-Head	0.616 +/-0.051	0.021 +/-0.027	0.938 +/-0.080	0.479 +/-0.035	-0.041 +/-0.070
Multi-Head RandCrop	0.638 +/-0.063	0.016 +/-0.017	0.968 +/-0.032	0.492 +/-0.019	-0.016 +/-0.039

Table 8.3: The classification performance using the feature extractor trained on the LDM100K dataset via the Multi-Head approach tested on the ADNI dataset.

computational cost. The proposed combined approach has the following advantages:

- The subject-level train-test-split fundamentally eliminates the risk of data leakage.
- Using each 3D MRI as a whole fully incorporates spatial information into the learning process.
- There is no need for explicit class labelling or lesion segmentation as all sub-tasks are in self-supervised learning configuration.
- The evaluation performance shows improved robustness, thus demonstrating the feasibility of combining multiple self-supervised learning tasks into a multi-head task using one base model.

8.5.2 Computational Cost

Same as in the previous chapters, the model is implemented in Python 3.7 using PyTorch 1.10 version on CUDA 11.3 runtime environment and trained on 4 NVIDIA A100 GPUs. The mini-batch size is reduced to 12 to enable parallel distribution for faster training. As explained in the previous subsections, each head computes its own loss followed by one pass of the base feature extractor parameter updates. Therefore, the computational cost is an accumulation of each head, resulting in 177.7 hours of runtime for 50 epochs of training on the LDM100K dataset. Unfortunately, a 10-fold cross-validation for ensemble classification could not be executed in this chapter due to the significantly increased computational expense.

8.5.3 Limitations and Future Work

The first major limitation of the multi-head task approach is the dramatic increase in computational cost. Even though the base model is rather lightweight,

the amount of training data in the 3D format made this approach very expensive to increment. Also, the potential of using more complex base models such as ResidualNetwork or VisionTransformer is limited by the availability of powerful computational hardware. It is speculated that performance improvement could be gained by acquiring more computational resources.

Another possible future work direction is expanding the multi-head task using more neuroimaging modalities to improve feature extractors. There are many neuroimaging data associated with the subjects such as functional MRI, positron emission tomography and computerised tomography that can be engineered into the Multi-head task scheme. Along with the development of neuroimaging technologies, the availability of multi-modality neuroimage data is also on the rise. Utilising more neuroimaging modalities might improve the feature extractors for downstream tasks. However, the integration of multi-modality neuroimaging data requires sophisticated preprocessing methods which need to be further investigated. Moreover, the increase in training data also escalates the computational cost as a consequence.

Chapter 9

Conclusion

This thesis started with a description of Alzheimer's Disease and the basics of neuroimaging, followed by an introduction to convolutional neural networks and their components. Then we presented a literature review of CNN-based AD classification. We compared various published approaches and closely examined their data usage in both 2D and 3D forms. As a result, we identified three main types of data leakage issues. Then, we conducted experiments to evaluate the following research hypothesis:

1. Training 3D CNN-based models for healthy brain age prediction as a pretext task can generate discriminable latent representations from MRI for CN vs. AD classification.

2. Synthetic neuroimaging data can be used to train 3D CNN-based models on the pretext task of brain age prediction.

3. 3D AutoEncoder-based reconstruction can be used as a pretext task in conjunction with synthetic neuroimaging data to train feature extractors.

4. Rotating 3D neuroimages can generate artificial labels that can be used in a pretext classification task for feature extractor training.

5. Combining multiple pretext tasks into one model can be used as a multi-head task to train feature extractors.

6. Using random cropping as a data augmentation technique can improve feature extractor training on 3D MRI data.

For the first question, Chapter. 4 presented an approach by using subjects' brain chronological age as a pretext target to train feature extractors for downstream AD classification. The results suggest that 3D CNN-based models trained to predict the age of CN-only MRI scans will predict the age of AD MRI scans higher (older) than CN scans. The proposed approach utilised a fundamental pipeline of preprocessing methods and a lightweight 3D CNN model, making it very robust. Not employing any data augmentation technique also demonstrated the robustness of the proposed approach. The subject-level train-test-split fundamentally eliminates the risk of data leakage.

Regarding the second question, Chapter. 5 presented the results using synthetic neuroimaging data using the same brain age prediction approach. For ease of comparison between different approaches, the preprocessing pipeline is reused. Although the classification results are not as good as training and testing on real-world data, it demonstrate the feasibility of utilising large-scale synthetic data for pretext task training. The results also suggest that random cropping is a promising data augmentation technique for 3D sMRI data.

Chapter. 6 shows the experimental results to support the third question. A 3D AutoEncoder consisting of Convolutional and DeConvolutional layers has been introduced as the backbone of the reconstruction pretext task. This self-supervised learning approach needs no explicit class labels or segmentation of disease lesions. The results suggest that utilising a large amount of unlabelled or segmented data in the medical field is possible. Again, the random cropping shows an improvement in the training of feature extractors. This approach also addresses the data leakage issues by using subject-level train-test-split.

Chapter. 7 focuses on the fourth research question. The proposed approach fabricated artificial labels for the 3D sMRI data by rotating a combination of 90 degrees in each plane. Then a CNN classifier is trained to identify the combination of rotation. The evaluation performance suggests that identifying

the rotation of 3D images without labelling or segmentation can be used to train feature extractors for downstream AD classification. Same as before, the random cropping data augmentation technique demonstrates its positive impact on training. This chapter also carefully performs train-test-split on the subject-level to eliminate the chance of data leakage.

Chapter.8 proposed a Multi-Head pretext task to examine this research question. The idea is to design different output layers according to the pretext tasks using the same base model. The loss of each head task is backpropagated to update the weights of the base model. The evaluation results suggest that this Multi-Head approach can be used as a pretext task for feature extractor training. The results also support that random cropping is a promising data augmentation technique for 3D sMRI data. Same as in the previous experiments, data leakage is not a concern in this part.

9.1 Summary of Results

The best classification performance is achieved by Chapter 4. The training and testing are performed on real-world data chosen from the OASIS-3 dataset. As shown in Table. 9.1, the overall best performance was achieved by the majority voting ensemble of the models from a cross-validation process. The random forest classifier ensemble obtains 0.847 ± 0.08 ACC, 0.934 ± 0.05 SPE and 0.822 ± 0.08 AUC, while the support vector machine ensemble shows a slightly higher SEN of 0.733 ± 0.12. The ensemble performances are approximately 10% better than the result of a single model.

The results from Chapters 5 to 8 using LDM100K synthetic dataset for feature extractor training are shown in Tables 9.2 to 9.5. Each table presents the classification performance of 4 proposed approaches trained on the LDM100k dataset with and without random cropping using a support vector classifier or random forest classifier on either the OASIS or the ADNI dataset.

Among these 4 tables trained on the LDM100K dataset, the highest clas-

Task	**AD vs. CN**			
Classifier	**ACC**	**SEN**	**SPE**	**AUC**
SVM (Best Model)	0.752 ± 0.06	0.622 ± 0.07	0.834 ± 0.10	0.728 ± 0.06
SVM (All Models)	0.842 ± 0.08	$\mathbf{0.733 \pm 0.12}$	0.908 ± 0.06	0.821 ± 0.08
RFC (Best Model)	0.745 ± 0.07	0.584 ± 0.14	0.844 ± 0.05	0.714 ± 0.07
RFC (All Models)	$\mathbf{0.847 \pm 0.08}$	0.711 ± 0.12	$\mathbf{0.934 \pm 0.05}$	$\mathbf{0.822 \pm 0.08}$

Table 9.1: Classification performance of using brain age prediction as a pretext to train feature extractors.

sification ACC of 73.9% is achieved by the SVC classifier on features obtained by brain age prediction in conjunction with RandCrop data augmentation on the OASIS-3 dataset. The AutoEncoder with RandCrop shows a fraction less ACC but the highest SEN of 41.7%, AUC of 0.654 and J_stat of 0.309 while the rotation-based approach using RandCrop resulted in the highest SPE of 0.947. The RFC shows a slightly lower number of ACC (72.8%) and AUC (0.614). A lower SEN of 31.4% is obtained by AutoEncoder without data augmentation while a similar SPE of 96.5% is come by the same rotation and RandCrop combination. The RFC earned a slightly poor J_stat of 0.239 compared to SVC. Both SVC and RFC performed poorly on the ADNI testing data, which is not ideal compared with the OASIS-3 testing. One reason could be that the OASIS-3 has a higher resolution of CDR recordings that allows fine-grain subject selection.

The main results of this thesis can be summarised as follows:

- Brain age prediction trained on the OASIS dataset achieved competitive performance compared with state-of-the-art methods in the literature.
- Using real-world data, the brain age prediction approach achieved the overall best classification performance.
- Applying cross-validation as a method to obtain multiple feature extractors as an ensemble can improve the classification performance.
- Feature extractors trained on the LDM100K synthetic dataset achieved

LDM Training & OASIS Testing					
SVC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
Age	0.707 +/-0.050	0.278 +/-0.086	0.912 +/-0.044	0.595 +/-0.036	0.190 +/-0.071
Age RandCrop	**0.739 +/-0.037**	0.320 +/-0.052	0.933 +/-0.025	0.627 +/-0.032	0.253 +/-0.064
AutoEncoder	0.738 +/-0.046	0.402 +/-0.056	0.897 +/-0.049	0.649 +/-0.037	0.299 +/-0.073
AutoEncoder RandCrop	0.738 +/-0.037	**0.417 +/-0.088**	0.892 +/-0.042	**0.654 +/-0.042**	**0.309 +/-0.084**
Rotation	0.696 +/-0.054	0.187 +/-0.055	0.933 +/-0.030	0.560 +/-0.029	0.120 +/-0.058
Rotation RandCrop	0.674 +/-0.061	0.092 +/-0.068	**0.947 +/-0.044**	0.520 +/-0.018	0.040 +/-0.035
Multi-Head	0.719 +/-0.045	0.342 +/-0.082	0.903 +/-0.053	0.623 +/-0.041	0.245 +/-0.083
Multi-Head RandCrop	0.733 +/-0.031	0.379 +/-0.080	0.894 +/-0.028	0.636 +/-0.037	0.272 +/-0.073

Table 9.2: The classification performance using the feature extractor trained on the LDM100K dataset via all four approaches tested on the OASIS-3 dataset using the support vector classifier.

similar performance compared to the same model using real-world data. This supports the feasibility of utilising large-scale synthetic data for pretext task training.

- All the training and testing splits are performed on the subject-level to prevent data leakage issues.
- Random cropping data augmentation technique shows consistent improvement across different experiments.

LDM Training & OASIS Testing					
RFC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
Age	0.724 +/-0.044	0.321 +/-0.056	0.919 +/-0.044	0.620 +/-0.026	**0.239 +/-0.051**
Age Randdcrop	**0.728 +/-0.052**	0.304 +/-0.066	0.925 +/-0.021	**0.614 +/-0.035**	0.228 +/-0.071
AutoEncoder	0.721 +/-0.045	**0.314 +/-0.093**	0.909 +/-0.051	0.612 +/-0.040	0.223 +/-0.079
AutoEncoder Randdcrop	0.712 +/-0.050	0.266 +/-0.096	0.927 +/-0.052	0.597 +/-0.037	0.194 +/-0.075
Rotation	0.701 +/-0.055	0.163 +/-0.047	0.954 +/-0.027	0.559 +/-0.026	0.117 +/-0.051
Rotation Randdcrop	0.692 +/-0.058	0.102 +/-0.041	**0.965 +/-0.024**	0.534 +/-0.013	0.067 +/-0.026
Multi-Head	0.684 +/-0.052	0.133 +/-0.066	0.944 +/-0.039	0.539 +/-0.021	0.077 +/-0.041
Multi-Head RandCrop	0.702 +/-0.053	0.243 +/-0.067	0.910 +/-0.039	0.577 +/-0.025	0.154 +/-0.049

Table 9.3: The classification performance using the feature extractor trained on the LDM100K dataset via all four approaches tested on the OASIS-3 dataset using the random forest classifier.

LDM Training & ADNI Testing					
SVC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
Age	0.653 +/-0.063	0.000 +/-0.000	**1.000 +/-0.000**	0.500 +/-0.000	0.000 +/-0.000
Age RandCrop	0.643 +/-0.091	**0.175 +/-0.285**	0.865 +/-0.290	0.520 +/-0.033	0.040 +/-0.067
Rotation	0.649 +/-0.062	0.000 +/-0.000	0.994 +/-0.017	0.497 +/-0.008	-0.006 +/-0.017
Rotation RandCrop	0.651 +/-0.064	0.007 +/-0.015	0.993 +/-0.015	0.500 +/-0.005	0.001 +/-0.010
AutoEncoder	0.653 +/-0.063	0.000 +/-0.000	**1.000 +/-0.000**	0.500 +/-0.000	0.000 +/-0.000
AutoEncoder RandCrop	0.653 +/-0.063	0.000 +/-0.000	**1.000 +/-0.000**	0.500 +/-0.000	0.000 +/-0.000
Multi-Head	0.618 +/-0.068	0.057 +/-0.117	0.927 +/-0.131	0.492 +/-0.040	-0.016 +/-0.079
Multi-Head RandCrop	**0.661 +/-0.052**	0.144 +/-0.104	0.930 +/-0.047	**0.537 +/-0.038**	**0.075 +/-0.075**

Table 9.4: The classification performance using the feature extractor trained on the LDM100K dataset via all four approaches tested on the ADNI dataset using the support vector classifier.

LDM Training & ADNI Testing					
RFC	**ACC**	**SEN**	**SPE**	**AUC**	**J_stat**
Age	**0.653 +/-0.063**	0.000 +/-0.000	**1.000 +/-0.000**	0.500 +/-0.000	0.000 +/-0.000
Age RandCrop	0.635 +/-0.077	0.009 +/-0.020	0.966 +/-0.037	0.488 +/-0.021	-0.025 +/-0.041
Rotation	0.643 +/-0.059	0.003 +/-0.010	0.983 +/-0.023	0.493 +/-0.014	-0.014 +/-0.028
Rotation RandCrop	0.651 +/-0.060	0.013 +/-0.021	0.990 +/-0.019	**0.502 +/-0.014**	**0.004 +/-0.029**
AutoEncoder	**0.653 +/-0.063**	0.000 +/-0.000	**1.000 +/-0.000**	0.500 +/-0.000	0.000 +/-0.000
AutoEncoder RandCrop	0.647 +/-0.073	0.000 +/-0.000	0.989 +/-0.032	0.495 +/-0.016	-0.011 +/-0.032
Multi-Head	0.616 +/-0.051	**0.021 +/-0.027**	0.938 +/-0.080	0.479 +/-0.035	-0.041 +/-0.070
Multi-Head RandCrop	0.638 +/-0.063	0.016 +/-0.017	0.968 +/-0.032	0.492 +/-0.019	-0.016 +/-0.039

Table 9.5: The classification performance using the feature extractor trained on the LDM100K dataset via all four approaches tested on the ADNI dataset using the random forest classifier.

9.2 Discussion

As shown in tables 9.2 to 9.5, the brain age prediction pretext task shows the best AD classification accuracy across the board. It is not as excellent as some approaches in the literature, but it is has a lot of potential to be explored in the future in terms of advanced regression models and training techniques.

The results of utilising the synthetic LDM-100k dataset are not as good as using real-world data, especially since the sensitivity is low. This might be due to the extreme class imbalance in the dataset. The CN class is the majority up to 90% in the train set. Even though the test set is close to the 1:1 class ratio, the model is biased toward the majority class during prediction.

Future work is needed to address this severe imbalance in the train set. For example, upsampling of the minority while downsampling the majority class is a widely used technique during model training, it might help mitigate the impact of imbalance and improve sensitivity. Adjusting class weights is another popular method to address the class imbalance issue. Also, data augmentation could be another choice to leverage the minority class. Last but not least, synthetic data generation for the AD class has not yet been explored either.

9.3 Limitation & Future Work

Although the evaluation results support the feasibility of the proposed approaches, there are some limitations:

1. The architecture of the 3D CNN base model, such as the number and types of convolutional layers, may not be the optimal choice. The usage of more complex models is limited by the availability of computational hardware.

2. The proposed approaches are only using T1w structural MRI data for feature extractor training, which leaves other available neuroimaging modalities underutilised.

3. Using 3D CNN and synthetic data for training feature extractors dramatically increases the computational cost.

4. It is not easy to visualise the extracted features for the interpretation of relevant brain changes in clinical practice.

This thesis obtained experimental results that open possibilities for future research, which can be summarised as follows:

1. The potential of using more complex base models such ResidualNetwork or VisionTransformer can be further investigated.

2. There are data available in other neuroimaging modalities (e.g. T2w, PET) and demographic information (e.g. gender, education) that can be further explored.

3. More efficient computational hardware and software can be further developed for efficient processing of 3D inputs.

4. The brain sMRI preprocessing methods and data augmentation techniques used in this thesis can be further extended to improve feature extractor training.

5. Machine learning model behaviour interpretability is extremely important in the medical domain. How to visualise the extracted features from MRI images is an open research question.

In conclusion, all four research directions in this thesis are based on the simple idea that using self-supervised learning methods can improve the feature extractor for Alzheimer's Disease classification performance. This might be the best choice of approach when neuroimaging data labelling or segmentation is unavailable. This study solely scratched the surface of the self-supervised learning field in regard to AD classification. For future work, the age-based approach can be further expanded by including other clinical information (e.g. cognitive assessment, gender and education) in the learning target. Other

neuroimaging modalities (e.g. T2 MRI and FLAIR) might be a great asset when enough amount of them have been collected and labelled.

The rotation-based approach might be improved by further utilising the spatial information of the 3D MRI data. One popular pretext task for 2D images is to predict the degree of rotation, but the same task is not well-studied for 3D images. Also, the pretext tasks proposed in the existing literature are relatively simple. It would be interesting to explore more complex rotation-based pretext tasks that can learn more discriminative features.

The subjects' clinical visits and MRI scans are temporally organised in the datasets. The temporal relationship between the visits and scans is under-utilised in this thesis. The similarities between two adjacent MRI scans might be used to enforce an AutoEncoder to minimise as part of the loss function. It is also intriguing to further explore the temporal relationship, which might lead to a more comprehensive representation of AD development and progression.

The current Multi-Head task only includes brain age prediction, brain image reconstruction and brain image rotation classification. It would be interesting to develop other pretext tasks and combine them into the multi-head framework. Also, it might be beneficial to incorporate additional clinical information and neuroimaging modalities into the multi-head framework.

Currently, many computational software are developed for 2D inputs. Dedicated software libraries for 3D neuroimaging data are still in its infancy. Also, specialised hardware for 3D neuroimaging data is either extremely expensive to buy, very difficult to access or even does not exist. Therefore, advancing computational software and hardware for 3D neuroimaging data would be another interesting direction for future research.

Although Explainable AI (XAI) is not the focus of this thesis, medical AI tools are demanded to be transparent and accountable. Regulatory bodies such as the European Union and the U.S. Federal Trade Commission's initiatives are calling for transparency and accountability in AI systems. To meet such demands and requirements, XAI aims to develop methods and techniques

to enhance the interpretability of AI models. Among the proposed approaches, methods based on attention mechanisms show promising progress in healthcare to interpret medical AI-made diagnoses. Despite significant progress, challenges remain in the research and development of XAI. The most pressing one is the balancing between interpretability and model performance while ensuring the reliability of explanations.

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Karim Aderghal, Jenny Benois-Pineau, Karim Afdel, and Catheline Gwenaëlle. Fuseme: Classification of smri images by fusion of deep cnns in 2d+ ε projections. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, pages 1–7, 2017.

[3] Karim Aderghal, Alexander Khvostikov, Andrei Krylov, Jenny Benois-Pineau, Karim Afdel, and Gwenaelle Catheline. Classification of alzheimer disease on imaging modalities with deep cnns using cross-modal transfer learning. In *2018 IEEE 31st international symposium on computer-based medical systems (CBMS)*, pages 345–350. IEEE, 2018.

[4] Kevin Adistambha, Stephen Davis, Christian Ritz, Ian S Burnett, and David Stirling. Enhancing multimedia search using human motion. *Multimedia-A Multidisciplinary Approach to Complex Issues, InTech*, pages 161–174, 2012.

[5] Maria-Luiza Antonie, Osmar R Zaiane, and Alexandru Coman. Application of data mining techniques for medical image classification. In *Proceedings of the Second International Conference on Multimedia Data Mining*, pages 94–101. Springer-Verlag, 2001.

[6] Alzheimer's Association. 2019 alzheimer's disease facts and figures. *Alzheimer's & dementia*, 15(3):321–387, 2019. URL `https://www.alz.org/alzheimers-dementia/facts-figures`.

[7] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

[8] Karl Bäckström, Mahmood Nazari, Irene Yu-Hua Gu, and Asgeir Store Jakola. An efficient 3d deep convolutional network for alzheimer's disease diagnosis using mr images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 149–153. IEEE, 2018.

[9] Silvia Basaia, Federica Agosta, Luca Wagner, Elisa Canu, Giuseppe Magnani, Roberto Santangelo, Massimo Filippi, Alzheimer's Disease Neuroimaging Initiative, et al. Automated classification of alzheimer's disease and mild cognitive impairment using a single mri and deep neural networks. *NeuroImage: Clinical*, 21:101645, 2019.

[10] Abol Basher, Byeong C Kim, Kun Ho Lee, and Ho Yub Jung. Volumetric feature-based alzheimer's disease diagnosis from smri data using a convolutional neural network and a deep neural network. *IEEE Access*, 9: 29870–29882, 2021.

[11] Randall J Bateman, Chengjie Xiong, Tammie LS Benzinger, Anne M Fagan, Alison Goate, Nick C Fox, Daniel S Marcus, Nigel J Cairns, Xianyun Xie, Tyler M Blazey, et al. Clinical and biomarker changes in dominantly inherited alzheimer's disease. *N Engl J Med*, 367:795–804, 2012.

[12] Iman Beheshti, Shiwangi Mishra, Daichi Sone, Pritee Khanna, and Hiroshi Matsuda. T1-weighted mri-driven brain age estimation in alzheimer's disease and parkinson's disease. *Aging and disease*, 11(3):618, 2020.

[13] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[14] Matthew Brett, Christopher J. Markiewicz, Michael Hanke, Marc-Alexandre Côté, Ben Cipollini, Paul McCarthy, Dorota Jarecka, Christopher P. Cheng, Yaroslav O. Halchenko, Michiel Cottaar, Eric Larson, Satrajit Ghosh, Demian Wassermann, Stephan Gerhard, Gregory R. Lee, Hao-Ting Wang, Erik Kastman, Jakub Kaczmarzyk, Roberto Guidotti, Or Duek, Jonathan Daniel, Ariel Rokem, Cindee Madison, Brendan Moloney, Félix C. Morency, Mathias Goncalves, Ross Markello, Cameron Riddell, Christopher Burns, Jarrod Millman, Alexandre Gramfort, Jaakko Leppäkangas, Anibal Sólon, Jasper J.F. van den Bosch, Robert D. Vincent, Henry Braun, Krish Subramaniam, Krzysztof J. Gorgolewski, Pradeep Reddy Raamana, Julian Klug, B. Nolan Nichols, Eric M. Baker, Soichi Hayashi, Basile Pinsard, Christian Haselgrove, Mark Hymers, Oscar Esteban, Serge Koudoro, Fernando Pérez-García, Nikolaas N. Oosterhof, Bago Amirbekian, Ian Nimmo-Smith, Ly Nguyen, Samir Reddigari, Samuel St-Jean, Egor Panfilov, Eleftherios Garyfallidis, Gael Varoquaux, Jon Haitz Legarreta, Kevin S. Hahn, Oliver P. Hinds, Bennet Fauber, Jean-Baptiste Poline, Jon Stutters, Kesshi Jordan, Matthew Cieslak, Miguel Estevan Moreno, Valentin Haenel, Yannick Schwartz, Zvi Baratz, Benjamin C Darwin, Bertrand Thirion, Carl Gauthier, Dimitri Papadopoulos Orfanos, Igor Solovey, Ivan Gonzalez, Jath Palasubramaniam, Justin Lecher, Katrin Leinweber, Konstantinos Raktivan, Markéta Calábková, Peter Fischer, Philippe Gervais, Syam Gadde, Thomas Ballinger, Thomas Roos, Venkateswara Reddy Reddam, and freec84. nipy/nibabel: 3.2.1, November 2020. URL `https://doi.org/10.5281/zenodo.4295521`.

[15] Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A Guertin, Joo Han Chang, Robert A Lindquist, Jason Moffat, et al. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7(10):R100, 2006.

[16] François Chollet et al. Keras, 2015. URL `https://github.com/fchollet/keras`.

[17] Roy G Cutler, Jeremiah Kelly, Kristin Storie, Ward A Pedersen, Anita Tammara, Kimmo Hatanpaa, Juan C Troncoso, and Mark P Mattson. Involvement of oxidative stress-induced abnormalities in ceramide and cholesterol metabolism in brain aging and alzheimer's disease. *Proceedings of the National Academy of Sciences*, 101(7):2070–2075, 2004.

[18] Correne A DeCarlo, Holly A Tuokko, Dorothy Williams, Roger A Dixon, and Stuart WS MacDonald. Bioage: Toward a multi-determined, mechanistic account of cognitive aging. *Ageing research reviews*, 18:95–105, 2014.

[19] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[20] Benoit Dufumier, Pietro Gori, Julie Victor, Antoine Grigis, Michele Wessa, Paolo Brambilla, Pauline Favre, Mircea Polosan, Colm Mcdonald, Camille Marie Piguet, et al. Contrastive learning with continuous proxy meta-data for 3d mri classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 58–68. Springer, 2021.

[21] Nicolae Duta and Milan Sonka. Segmentation and interpretation of mr brain images. an improved active shape model. *IEEE Transactions on Medical Imaging*, 17(6):1049–1062, 1998.

[22] Amir Ebrahimi, Suhuai Luo, and Raymond Chiong. Introducing transfer learning to 3d resnet-18 for alzheimer's disease detection on mri images. In *2020 35th international conference on image and vision computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2020.

[23] Amir Ebrahimi, Suhuai Luo, and for the Alzheimer's Disease Neuroimaging Initiative. Convolutional neural networks for alzheimer's disease detection on mri images. *Journal of Medical Imaging*, 8(2):024503–024503, 2021.

[24] Katja Franke, Gabriel Ziegler, Stefan Klöppel, Christian Gaser, Alzheimer's Disease Neuroimaging Initiative, et al. Estimating the age of healthy subjects from t1-weighted mri scans using kernel methods: exploring the influence of various parameters. *Neuroimage*, 50(3):883–892, 2010.

[25] Katja Franke, Christian Gaser, Brad Manor, and Vera Novak. Advanced brainage in older adults with type 2 diabetes mellitus. *Frontiers in aging neuroscience*, 5:90, 2013.

[26] Christian Gaser, Katja Franke, Stefan Klöppel, Nikolaos Koutsouleris, Heinrich Sauer, and Alzheimer's Disease Neuroimaging Initiative. Brainage in mild cognitive impaired patients: predicting the conversion to alzheimer's disease. *PloS one*, 8(6):e67346, 2013.

[27] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[28] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[29] Nitika Goenka and Shamik Tiwari. Alzvnet: A volumetric convolutional neural network for multiclass classification of alzheimer's disease through multiple neuroimaging computational approaches. *Biomedical Signal Processing and Control*, 74:103500, 2022.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[31] Yechong Huang, Jiahang Xu, Yuncheng Zhou, Tong Tong, Xiahai Zhuang, and Alzheimer's Disease Neuroimaging Initiative (ADNI). Diagnosis of alzheimer's disease via multi-modality 3d convolutional neural network. *Frontiers in neuroscience*, 13:509, 2019.

[32] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. URL `http://arxiv.org/abs/1502.03167`.

[33] Jyoti Islam and Yanqing Zhang. Brain mri analysis for alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks. *Brain informatics*, 5:1–14, 2018.

[34] Rachna Jain, Nikita Jain, Akshay Aggarwal, and D Jude Hemanth. Convolutional neural network based alzheimer's disease classification from magnetic resonance brain images. *Cognitive Systems Research*, 57:147–159, 2019.

[35] Wenjie Kang, Lan Lin, Baiwen Zhang, Xiaoqi Shen, Shuicai Wu, Alzheimer's Disease Neuroimaging Initiative, et al. Multi-model and multi-slice ensemble learning architecture based on 2d convolutional neural networks for alzheimer's disease diagnosis. *Computers in Biology and Medicine*, 136:104678, 2021.

[36] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 2023.

[37] Jee Wook Kim, Min Soo Byun, Bo Kyung Sohn, Dahyun Yi, Eun Hyun Seo, Young Min Choe, Shin Gyeom Kim, Hyo Jung Choi, Jun Ho Lee, Ik Seung Chee, et al. Clinical dementia rating orientation score as an ex-

cellent predictor of the progression to alzheimer's disease in mild cognitive impairment. *Psychiatry investigation*, 14(4):420, 2017.

[38] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[39] William E Klunk, Henry Engler, Agneta Nordberg, Yanming Wang, Gunnar Blomqvist, Daniel P Holt, Mats Bergström, Irina Savitcheva, Guo-Feng Huang, Sergio Estrada, et al. Imaging brain amyloid in alzheimer's disease with pittsburgh compound-b. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 55(3):306–319, 2004.

[40] Sergey Korolev, Amir Safiullin, Mikhail Belyaev, and Yulia Dodonova. Residual and plain convolutional neural networks for 3d brain mri classification. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, pages 835–838. IEEE, 2017.

[41] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[42] Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *MedRxiv*, 2019.

[43] JL Lancaster, LH Rainey, JL Summerlin, CS Freitas, PT Fox, AC Evans, AW Toga, and JC Mazziotta. Automated labeling of the human brain: A preliminary report on the development and evaluation of a forward-transform method. *Human brain mapping*, 5(4):238–242, 1997.

[44] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[45] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[46] Fan Li, Danni Cheng, and Manhua Liu. Alzheimer's disease classification based on combination of multi-model convolutional networks. In *2017 IEEE international conference on imaging systems and techniques (IST)*, pages 1–5. IEEE, 2017.

[47] Fan Li, Manhua Liu, Alzheimer's Disease Neuroimaging Initiative, et al. Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks. *Computerized Medical Imaging and Graphics*, 70:101–110, 2018.

[48] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *arXiv preprint arXiv:1702.05747*, 2017.

[49] Junxiu Liu, Mingxing Li, Yuling Luo, Su Yang, Wei Li, and Yifei Bi. Alzheimer's disease detection using depthwise separable convolutional neural networks. *Computer Methods and Programs in Biomedicine*, 203: 106032, 2021.

[50] Manhua Liu, Danni Cheng, Kundong Wang, Yaping Wang, and Alzheimer's Disease Neuroimaging Initiative. Multi-modality cascaded convolutional neural networks for alzheimer's disease diagnosis. *Neuroinformatics*, 16:295–308, 2018.

[51] Manhua Liu, Fan Li, Hao Yan, Kundong Wang, Yixin Ma, Li Shen, Mingqing Xu, Alzheimer's Disease Neuroimaging Initiative, et al. A multi-model deep convolutional neural network for automatic hippocam-

pus segmentation and classification in alzheimer's disease. *Neuroimage*, 208:116459, 2020.

[52] Shih-Chung B Lo, Heang-Ping Chan, Jyh-Shyan Lin, Huai Li, Matthew T Freedman, and Seong K Mun. Artificial convolution neural network for medical image pattern recognition. *Neural networks*, 8(7-8):1201–1214, 1995.

[53] Andreas Markus Loening and Sanjiv Sam Gambhir. Amide: a free software tool for multimodality medical image analysis. *Molecular imaging*, 2(3):15353500200303133, 2003.

[54] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[55] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

[56] Muazzam Maqsood, Faria Nazir, Umair Khan, Farhan Aadil, Habibullah Jamal, Irfan Mehmood, and Oh-young Song. Transfer learning assisted classification and detection of alzheimer's disease stages using 3d mri scans. *Sensors*, 19(11):2645, 2019.

[57] Pratik Mazumder, Pravendra Singh, and Vinay P Namboodiri. Few-shot image classification with composite rotation based self-supervised auxiliary task. *Neurocomputing*, 489:179–195, 2022.

[58] Guy M McKhann, David S Knopman, Howard Chertkow, Bradley T Hyman, Clifford R Jack Jr, Claudia H Kawas, William E Klunk, Walter J Koroshetz, Jennifer J Manly, Richard Mayeux, et al. The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia*, 7(3):263–269, 2011.

[59] Atif Mehmood, Muazzam Maqsood, Muzaffar Bashir, and Yang Shuyuan. A deep siamese convolution neural network for multi-class classification of alzheimer disease. *Brain sciences*, 10(2):84, 2020.

[60] Atif Mehmood, Shuyuan Yang, Zhixi Feng, Min Wang, Al Smadi Ahmad, Rizwan Khan, Muazzam Maqsood, and Muhammad Yaqub. A transfer learning approach for early diagnosis of alzheimer's disease on mri images. *Neuroscience*, 460:43–52, 2021.

[61] David J Miller and Hasan S Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. In *Advances in neural information processing systems*, pages 571–577, 1997.

[62] Rajendra A Morey, Christopher M Petty, Yuan Xu, Jasmeet Pannu Hayes, H Ryan Wagner II, Darrell V Lewis, Kevin S LaBar, Martin Styner, and Gregory McCarthy. A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. *Neuroimage*, 45(3):855–866, 2009.

[63] John C Morris. The clinical dementia rating (cdr): current version and scoring rules. *Neurology*, 1993.

[64] Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics*, 15(4):869–877, 2005.

[65] My-MS.org. Basic plane mathematics of mri, 2008. Accessed August 1, 2023. https://my-ms.org/mri_planes.html.

[66] Igor Nenadić, Maren Dietzek, Kerstin Langbein, Heinrich Sauer, and Christian Gaser. Brainage score indicates accelerated brain aging in schizophrenia, but not bipolar disorder. *Psychiatry Research: Neuroimaging*, 266:86–89, 2017.

[67] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady an ussr*, volume 269, pages 543–547, 1983.

[68] Kanghan Oh, Young-Chul Chung, Ko Woon Kim, Woo-Sung Kim, and Il-Seok Oh. Classification and visualization of alzheimer's disease using volumetric convolutional neural network and transfer learning. *Scientific Reports*, 9(1):1–16, 2019.

[69] Maysam Orouskhani, Sahar Rostamian, Firoozeh Shomal Zadeh, Mehrzad Shafiei, and Yasin Orouskhani. Alzheimer's disease detection from structural mri using conditional deep triplet network. *Neuroscience Informatics*, page 100066, 2022.

[70] Jiahong Ouyang, Qingyu Zhao, Ehsan Adeli, Greg Zaharchuk, and Kilian M Pohl. Self-supervised learning of neighborhood embedding for longitudinal mri. *Medical Image Analysis*, 82:102571, 2022.

[71] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-perfor`

[72] Han Peng, Weikang Gong, Christian F Beckmann, Andrea Vedaldi, and Stephen M Smith. Accurate brain age prediction with lightweight deep neural networks. *Medical image analysis*, 68:101871, 2021.

[73] William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel,

and Thomas E Nichols. *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

[74] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *Deep Generative Models: Second MICCAI Workshop, DGM4MICCAI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pages 117–126. Springer, 2022.

[75] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.

[76] David C Preston. Magnetic resonance imaging (mri) of the brain and spine: Basics, 2016. Accessed August 21, 2023. https://case.edu/med/neurology/NR/MRI

[77] The University Of Queensland Queensland Brain Institute. What causes dementia? https://qbi.uq.edu.au/dementia/dementia-causes-and-treatment, 2018. Accessed: 2018-9-30.

[78] Manu Raju, Varun P Gopi, VS Anitha, and Khan A Wahid. Multi-class diagnosis of alzheimer's disease using cascaded three dimensional-convolutional neural network. *Physical and Engineering Sciences in Medicine*, 43:1219–1228, 2020.

[79] Sai Prasad Raya. Low-level segmentation of 3-d magnetic resonance brain images-a rule-based system. *IEEE Transactions on Medical Imaging*, 9 (3):327–337, 1990.

[80] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion

models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[81] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[82] Serkan Savaş. Detecting the stages of alzheimer's disease with pre-trained deep learning architectures. *Arabian Journal for Science and Engineering*, 47(2):2201–2218, 2022.

[83] Upul Senanayake, Arcot Sowmya, and Laughlin Dawes. Deep fusion pipeline for mild cognitive impairment diagnosis. In *2018 IEEE 15th international symposium on biomedical imaging (isbi 2018)*, pages 1394–1997. IEEE, 2018.

[84] Jayanthi Venkatraman Shanmugam, Baskar Duraisamy, Blessy Chittattukarakkaran Simon, and Preethi Bhaskaran. Alzheimer's disease classification using pre-trained deep networks. *Biomedical Signal Processing and Control*, 71:103217, 2022.

[85] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[86] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[87] Shiliang Sun, Zehui Cao, Han Zhu, and Jing Zhao. A survey of optimization methods from a machine learning perspective. *IEEE transactions on cybernetics*, 50(8):3668–3681, 2019.

[88] Aly Valliani and Ameet Soni. Deep residual nets for improved alzheimer's diagnosis. In *Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics*, pages 615–615, 2017.

[89] Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D Wang. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific reports*, 11(1):3254, 2021.

[90] Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, Olivier Colliot, et al. Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. *Medical image analysis*, 63:101694, 2020.

[91] Zaidao Wen, Zhunga Liu, Shuai Zhang, and Quan Pan. Rotation awareness based self-supervised learning for sar target recognition with limited training samples. *IEEE Transactions on Image Processing*, 30:7266–7279, 2021.

[92] Bengt Winblad, Philippe Amouyel, Sandrine Andrieu, Clive Ballard, Carol Brayne, Henry Brodaty, Angel Cedazo-Minguez, Bruno Dubois, David Edvardsson, Howard Feldman, et al. Defeating alzheimer's disease and other dementias: a priority for european science and society. *The Lancet Neurology*, 15(5):455–532, 2016.

[93] Haifeng Wu, Jinling Luo, Xiaoling Lu, and Yu Zeng. 3d transfer learning network for classification of alzheimer's disease with mri. *International Journal of Machine Learning and Cybernetics*, 13(7):1997–2011, 2022.

[94] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. In *2010 IEEE Computer Society Conference on computer vision and pattern recognition*, pages 2528–2535. IEEE, 2010.

[95] Jie Zhang, Bowen Zheng, Ang Gao, Xin Feng, Dong Liang, and Xiaojing Long. A 3d densely connected convolution neural network with connection-wise attention mechanism for alzheimer's disease classification. *Magnetic Resonance Imaging*, 78:119–126, 2021.

[96] Qingyu Zhao, Zixuan Liu, Ehsan Adeli, and Kilian M Pohl. Longitudinal self-supervised learning. *Medical image analysis*, 71:102051, 2021.

[97] Chen Zheng, Bernhard Pfahringer, and Michael Mayo. Alzheimer's disease detection via a surrogate brain age prediction task using 3d convolutional neural networks. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.