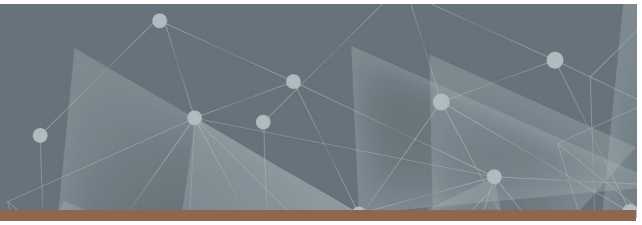
CHALMERS
UNIVERSITY OF TECHNOLOGY



# Machine Learning Techniques to Understand Cognitive Decline

## Using multiple MRI imaging techniques for multiclass classification of cognitive decline

Master's thesis in Data Science & AI

King Sang Tang
Lucas Fallqvist

DEPARTMENT OF ELECTRICAL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2022
www.chalmers.se

Master's thesis 2022

# Machine Learning Techniques to Understand Cognitive Decline

Using multiple MRI Imaging Techniques for Multiclass Classification of Cognitive Decline

King Sang Tang
Lucas Fallqvist


CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
*Computer vision and medical imaging analysis*
Institute of Neuroscience and Physiology, Gothenburg MCI study
Chalmers University of Technology
Gothenburg, Sweden 2022

Machine Learning Techniques to Understand Cognitive Decline
Using Multiple MRI Imaging Techniques for Multiclass Classification of Cognitive Decline
King Sang Tang
Lucas Fallqvist



Supervisor: Roman Naeem, Department of Electrical Engineering
Examiner: Fredrik Kahl, Electrical Engineering

Master's Thesis 2022
Department of Electrical Engineering
Computer vision and medical imaging analysis
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2022

# Abstract

This thesis project aims at understanding the progression of cognitive decline using machine learning models. This work acts as a support to the research done in the Gothenburg Mild Cognitive Impairment study. This clinical cohort focuses on five disease categories of patients and includes a control group. Specifically, the use of state-of-the-art CNN models such as ResNet[1] and GoogleNet[2] applied to the imaging techniques T1, T2 and FLAIR [3] has been investigated. The data posed some challenges in this work, for instance some patient categories have overlapping disease features, which could mean that in some patients could share MR features as well. Additionally, for the patient group Vascular Cognitive Disease, only minuscule changes in the brain blood vessels distinguish it from other classes. The overlap of MRI features in addition to the inherent complexity of the patient categories poses a challenge of being able to separate all six classes from each other. In order to handle the problem, an extensive search for separability between classes under each image capturing technique was done where binary classifiers were trained on each pair of the six classes, a total of 15 binary models trained on two classes at a time. However, tests concluded that a multi-class model predicting all six classes does not perform significantly above baseline. It can be seen from the binary models that many class pairs are not separable under certain image capturing techniques, most likely, due to feature sharing across classes. However, with other classes such as Control and AD it was possible to accurately predict and separate around 80% of the test data, showing a promise that Machine Learning can serve as a useful tool to better understand the MCI study cohort using available data and hardware.

# Acknowledgements

First and foremost, we want to extend our gratitude to our supervisor Roman Naeem who has given us continuous feedback and help throughout the project. With invaluable expertise and guidance you have helped us to push forward during times of doubt. Without your help this project would not have been possible.

We would also like to thank the Gothenburg MCI group and Sahlgrenska Academy, particularly Petronella Kettunen who has acted as our main supervisor from the research group. We are grateful for your trust in us during this whole project and for the many interesting and insightful meetings where we could utilize your expertise within this research field. We would also like to thank Grégoria Kalpouzos and Farshad Falahati for their support during the project. The advice given for medical tools selection has greatly helped during the preprocessing procedures, which was crucial for preparing the data.

Finally we want to thank our examiner Prof. Fredrik Kahl for acting as our examiner and giving your support for the project.

# List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

| | |
|---|---|
| AD | Alzheimer's disease |
| AI | Artificial Intelligence |
| FLAIR | Fluid-Attenuated Inversion Recovery |
| CNN | Convolutional Neural Network |
| DICOM | Digital Imaging and Communications in Medicine standard |
| MCI | Mild Cognitive Impairment |
| ML | Machine Learning |
| MRI | Magnetic Resonance Imaging |
| Mix | Mixture of AD and VaD |
| SCI | Subjective Cognitive Impairment |
| VaD | Vascular Cognitive Disease |
| CTRL | Control subjects |
| Pre-AD | Patients that express strong signs of developing into AD |
| Pre-VaD | Patients that express strong signs of developing into VaD |
| T1 | T1-Weighted MRI |
| T2 | T2-Weighted MRI |

# Contents

# List of Figures

# 1
# Introduction

In the following sections more detailed information about the Gothenburg MCI study, previous work in the field and the context of this report will be presented.

## 1.1 Gothenburg MCI study

The Gothenburg Mild Cognitive Impairment (MCI) study is a well-defined longitudinal cohort at the Memory Clinic in Mölndal that started in 1999[4]. Patients were actively recruited until 2015 and already recruited patients are given follow-up investigations. The cohort contains around 1000 patients seeking help for memory problems as well as control subjects, between the ages 50 to 79 years. Each patient file contains neuropsychological, neuroimaging and biomarker/lab data collected repeatedly from the same patient over the years (2, 4, 6 and 10 years after baseline). This makes it possible to follow the progressive development to Alzheimer's Disease(AD), from patients reporting subjective cognitive impairment(SCI) or MCI. As SCI and MCI are regarded as risk factors or early stages of dementia, there is a high chance that the molecular mechanisms underlying AD or subcortical small vessel disease (SSVD) onset can be found. Apart from this, data points from 136 controls are available. The cohort also contains patients of related diagnoses, such as mixed AD/SSVD and pure SSVD cases. This brings a total of six separate classes that the project actively focuses on.

In the dataset, patients with other forms of dementia (cortical vascular dementia, primary progressive aphasia, Lewy body dementia, frontotemporal dementia, or unspecified dementia) were excluded. The participants were recruited from the Gothenburg MCI study, a mono-center study of patients seeking help for cognitive complaints at the memory clinic at Sahlgrenska University Hospital. The inclusion and exclusion criteria were designed to exclude somatic and psychiatric conditions associated with increased risk of cognitive impairment. Thus, the inclusion criteria comprised age $> 40$ and $< 79$ years, Mini Mental State Examination (MMSE) score $> 19$, and self- or informant-reported cognitive decline with a duration $\leq 6$ months. The exclusion criteria included severe somatic disease (e.g., subdural hemorrhage, brain tumor, untreated hypothyroid state, encephalitis, and unstable heart disease), psychiatric disorder (e.g., major affective disorder or schizophrenia), substance abuse, and confusion. The healthy controls were primarily recruited through senior citizen organizations, e.g., information meetings on cognitive disorders and some were relatives of the patients. Present, or history of, cognitive decline was

an exclusion criterion in the controls, otherwise the exclusion criteria as well as the study procedures were similar as those applied for the patients.

For a large portion of the samples from patients and control subjects, the study has collected T1, T2 and Fluid-Attenuated Inversion Recovery(FLAIR) MRI sequences. For this project, the assumption is that data collected at different time-points can be regarded as two separate records. In other words, because the structure of the brain changes with time, one patient with more than one visit can be regarded as several independent data samples. Furthermore, not all subjects have the necessary data collected for all time-points, which is why the number of samples used in this project is smaller than the total number of records in the cohort. This will be further explained in section 3.1.1.

## 1.2 Purpose and research questions

Since the Gothenburg MCI study have collected a lot of data dating back to 1999 and never before used Machine Learning to analyze it, this project will be the first time the data is processed with ML tools. In order to support the group to research the ML field a few research questions have been formulated.

1. Can the dataset collected in the MCI study be used for classification with Machine Learning techniques?
2. Is it possible to use available ML tools in order to classify the diagnosis of the patients? (and correctly identify control subjects)

## 1.3 Previous Work

In this section previous related work that has been considered in this project is presented. The focus will be on binary- and multi-class classification as well as the differences of working with an open or closed dataset.

### 1.3.1 Binary classification

Binary classification for medical imaging data is a well researched field, where most of the work has come in recent years due to the progress of ML and AI techniques[5]. For instance, Tufail et al. [6] achieved a great cross-validation accuracy of 99% when using an Inception version 3 model. This model was trained on the dataset proposed by Hon et al. [7]. However, no testing was reported by Tufail et. al. there is no way to know how well this model would generalize to unseen samples. Therefore, the result should at its height be regarded as a proof that Inception Version 3 can at least converge on the training data features.

Sarraf et al.[8] used the two different CNN architectures, LeNet and GoogleNet to perform binary classification on AD patients and control subjects. They achieved

a great performance of 98.84% accuracy in the best case scenario when training on structural T1 images from the ADNI dataset[9].

In another binary classification work done by Wang et al. [10] an proposed eight-layer CNN model. Moreover, they showed that using six layers of convolutional layers with two fully connected layers and using Leaky ReLU activations was the best option for their task.

In a paper written by Khagi et al. [11] the authors tried to compare how well AlexNet, GoogleNet and ResNet could adapt towards medical image classification tasks, in this case binary classification of AD and control subjects. The goal here was to utilize transfer learning on the networks to be able to take advantage of the pre-training the networks can come with from natural image classification tasks. Even though this task is very different, as natural images often contain very different features and also come in full color, all three RGB-channels, the authors concluded that transfer learning could be utilized, successfully transferring weights pretrained on non-medical images for a medical image classification task. However, as also stated in the paper, the best accuracy was achieved when the authors did not utilize a pretrained network, but instead trained a network from scratch. The best accuracy achieved was 98% for a network trained from scratch and 94% accuracy for a network pretrained on natural images.

### 1.3.2 Multi-class classification

Korolev et al. [12] used a custom made model they called VoxCNN that was inspired by the VGG model as well as the ResNet model [1]. They trained the model on the ADNI dataset using T1 MRI images and with AD, control, early MCI and late MCI as their target classes. What they found was that with their proposed method both networks was able to separate AD and Control subjects but struggled with both early and late MCI.

Wang et al. [13] proposed a novel multimodal CNN method in order to classify AD, amnestic MCI and Control subjects. With this proposed method they were able to achieve a classification accuracy of 92.06% as the highest. This was achieved by combining diffusion tensor imaging (DTI) and functional magnetic resonance imaging (fMRI) as input to a 2D CNN model simultaneously. When only using DTI or fMRI the highest achieved accuracy was 87.30% and 82.54% respectively.

Farooq et al. [14] proposed a method utilizing established state-of-the-art network architectures. In their paper they suggest using Resnet or GoogleNet and training the from scratch over 100 epochs using ADNI data [9]. Their work targeted 4 different classes, namely AD, MCI, late MCI and healthy subjects. Using GoogleNet they managed to achieve a 4-way classification accuracy of 98.8% and with ResNet they managed to achieve an accuracy of 98.14% as the highest. This shows that there is great promise to utilize state-of-the-art architectures even in the field of medical imaging classification.

### 1.3.3 Open vs closed dataset

Working with open datasets such as Open Acess Series of Imaging Studies (OASIS) [15] or the Alzheimer's Disease Neuroimaging Initative (ADNI) dataset [9] comes with many benefits. The benefits are that the data is highly standardized and controlled such that all images have been captured with the same method and in the same format. In addition to this these datasets are quality checked several times and bad datapoints are discarded from the final data. Furthermore, you can get high quantities of data as these datasets are continuously added to from multiple sources. Altogether this means that the data is both of high quality and high quantity.

However, both the ADNI and the OASIS dataset only contain control and AD subjects [15], as these are the most commonly studies within medical data pertaining to mental cognition. The restriction with working with open datasets is thus that one is limited to what diseases ( classes ) to study. Furthermore, there is a restriction in which type of data imaging technique you can attain, for instance in the OASIS dataset there is only T1 MRI images, whilst in the ADNI one could also attain T2 and FLAIR MRI.

With closed datasets there is instead high possibility of customization, at the very least each closed dataset comes with their own customized target classes and image capturing techniques. There is also a higher possibility to extend the data and restructure it to fit ones need. This high customization comes at a cost however, which is that both the quality and quantity will generally be lower then compared to an open dataset. As closed datasets often are captured by local institutions or research groups practices can change over time and thus the data will not always follow a strict standard. For instance, in long going studies, the doctor/radiologists appointed to capture datapoints may change and with that comes some variance in the data gathering procedure. As closed datasets also often follow very strict rules by internal practices and by law, such as GDPR, it is also hard to outsource or let external expertise consult on standards. Therefore it is very important that one keeps this in mind when working with closed dataset and thoroughly look at the data and go through the necessary pre-processing steps in order to prepare the data.

## 1.4 Context and scope of this thesis

For this master thesis the aim is to utilize state-of-the-art network architectures, such as previous work has done [14][11][12]. In this thesis work the number of classes under interest is however extended from alot of previous work. Normally AD, MCI or some variant of these two diseases are the focus, apart from also having control subjects to compare with. For the work of this thesis however, the additional classes SCI, VaD and Mix are also within the scope of focus. More about each individual class can be read under chapter 2.1.

The reason for this work is that it is intended to help the Gothenburg MCI study [4], to explore the use of ML tools to understand their data and cohort. In order to succeed with this, the work should show that available software tools, that can be handled with available hardware should be able to interpret and handle the data the group has gathered in a meaningful way. For instance, one should be able to utilize a ML tool to distinguish between the different classes, as mentioned above, from one another.

This thesis will utilize previous work and build upon that and at the same time make use of the data gathered by and that is available from and within the Gothenburg MCI study. However, it is not intended that this thesis work will make use of external, open datasets such as ADNI [9] or OASIS [15].

# 2

# Theoretical Background & Framework

In the following chapter we discuss about the relevant theory behind the target classes and the utilized methods. In particular the different classes that are to be distinguished as well as what a CNN is and how transfer learning can be utilized.

## 2.1 Classes

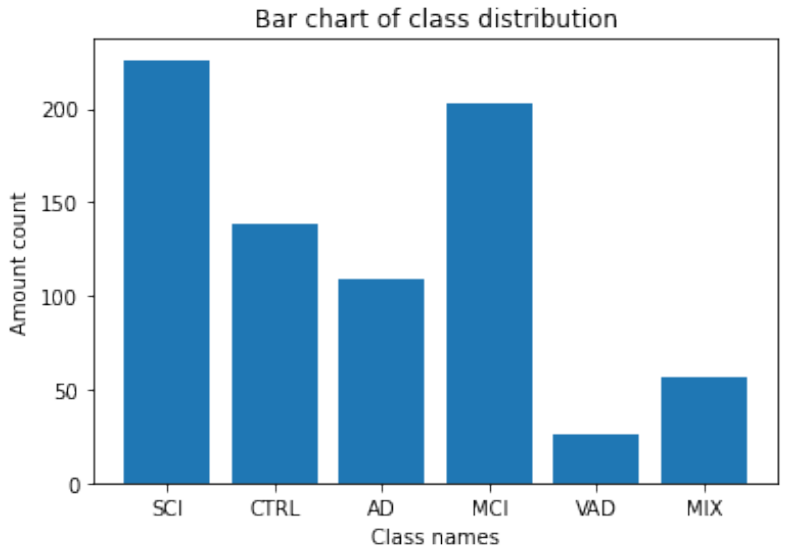



**Figure 2.1:** Class distribution of all the six classes in the dataset showing the number of complete samples with T1, T2 and FLAIR images. Total count of each class being SCI(226), CTRL(138), AD(109), MCI(201), VaD(26), MIX(56) and Total(756)

Figure 2.1 above depicts the distribution of the six target classes. For each class it is presented how many complete samples available with T1, T2 and FLAIR images captured. As it can be seen, the distribution is heavily unbalanced with the smallest class being VaD, which only has 26 complete samples and the largest class being SCI with 226 samples.

This unbalanced problem needs to be addressed properly as it otherwise creates issues when training a ML model. For instance, if one would run a binary model with SCI and VaD as the targets one could get an accuracy of roughly 90% by only predicting SCI all the time. This result could give the appearance of a very good performance when in reality it is actually only the baseline performance.

There are a number of ways to handle unbalanced class distributions[16]. For instance one can give different weights to the classes when calculating the loss cost for a wrong prediction. This way one can give the unbalanced data "as-is" and then specify higher loss cost for the underrepresented classes and lower loss cost for the large classes. Another method is to utilize under- and over-sampling[17]. By under-sampling one can exclude samples from the larger classes and thus create a more even distribution. Under-sampling comes with the risk of excluding useful information from the samples which are removed from the majority classes. When over-sampling one instead creates new samples from the underrepresented classes. The most straight-forward approach to do so is to simply make copies of the minority classes. Copying runs the risk of overfitting the model to the minority classes however and is therefore not a robust solution. To counter overfitting when over-sampling one can instead artificially create new samples by augmenting the available data [18].

### 2.1.1 Alzheimer's disease

AD is a cognitive disorder which is related to loss of memory, loss of logical thinking and other cognitive functions [19]. Even though AD is well studied it is not yet fully understood, in particular the origins of the disease as well as a fully covering diagnosis of it has not been established. With a combination of the most advanced diagnostic techniques one can only achieve around 90% diagnostic accuracy [20]. This accuracy is achieved after a combination of the standard Mini-Mental State Examination (MMSE) [21] which in itself have an diagnostic accuracy of about 85%. After the MMSE test additional information of family history, laboratory test as well as analysis of brain MRI scans can then give a combined accuracy which approaches 90% diagnostic accuracy.

The subjects in the cohort which has been classified as AD patients have been examined by doctors within the Gothenburg MCI study. The classification of AD is made by a combination of diagnostic measures. The subject goes through tests to determine it's cognitive fitness and memory capacity and also an analysis of the brain MR images. They also collect data regarding cerebral spinal fluid marker values for P-tau, T-tau and Beta-Amyloid 1-42 (A$\beta$1-42). The AD subjects amount to around 14% of the final data used in the project.

### 2.1.2 Control subjects

The control subjects included in the study are subjects which has no identified disease related to cognitive disability or functional decline. These control subjects are used in order to verify and compare the gathered data from the patients with the target diseases.

### 2.1.3 Subjective Cognitive Impairment

These patients have a self-proclaimed problem with their cognitive abilities and can in some cases show starter signs of developing a cognitive disease. However, it is important to note that these individuals do not have an expressed disease, at least not enough for them to be distinctly diagnosed. However, that they are experiencing troubles in their daily lives and feel like they have a cognitive impairment qualifies them to be diagnosed with Subjective Cognitive Impairment. Some causes of SCI can be due to high stress or depression disorder.

This disease naturally has a lot of connections to the other diseases, even more so then MCI, since SCI also have connections to the control subjects. It is not entirely certain that SCI subjects will develop into a distinct cognitive impairment in the future and thus it is possible for a MCI patient to develop back into a healthy state. However, the opposite is also true and they might just be a pre-AD or pre-VaD. This makes the SCI as an ambiguous class and according to recent findings from the Gothenburg MCI group SCI patients and Control subjects are highly similar in both MRI and biomarker data, suggesting that a lot of the SCI patients will devolve in the future and become healthy.

### 2.1.4 Mild Cognitive Impairment

Mild cognitive impairment is referred to as a circumscribed cognitive syndrome focused on memory loss or a comprehensive cognitive syndrome irrespective of the cognitive domains involved [22]. While the cognitive impairment in MCI is objectively measurable it should not constrain daily life. The MCI entity has been used mainly in the context of AD due to an increased risk to develop the disease but it has not yet been clarified which of the circumscribed or the comprehensive forms of MCI that is characteristic for early AD. Cerebrovascular disease, systemic, and other neurodegenerative disorders may also cause MCI at early stages although there are hitherto comparatively few studies in vascular and other non-AD forms of MCI.

### 2.1.5 Mixture of diseases

The mixture of diseases are a class where the patient has a combination of both AD and VaD, such that the cognitive decline cannnot fully be attributed to one or the other. These patients do not however have to have a worse overall cognitive ability than that of a patient only suffering from one of the diseases. The mental cognitive ability might have deteriorated to a point where it is troublesome, but not yet be an impediment on every aspect of the patients life. A patient only sustaining on of

the diseases might have worse cognitive decline, as their disease is further gone. The important aspect here is therefore not that the patient has a severe case of cognitive decline, rather that the symptoms and diagnose cannot be attributed exclusively to neither AD or VaD.

### 2.1.6 Vascular Disease

Vascular cognitive disorder (VaD) and AD belong to the most common cognitive disorders in the elderly population. Several forms of VaD exist but in this paper we use the singular denomination for all variants of VaD. VaD is similar to "vascular cognitive impairment" but refer more clearly to phenotypically characteristic subgroups and is broader than "vascular dementia" as milder forms of cognitive impairment also are included. The subcortical small-vessel type of disease (SSVD) has been estimated to be the most common form of VaD [23]. The disease affects the small vessels deep in the brain, including perforating arterioles, capillaries and venules. In these patients, magnetic resonance imaging (MRI) reveals increased occurrence of cerebral microbleeds (CMBs), infarcts and lacunes, as well as white matter hyperintensities (WMHs) [24] that correspond to lesions of the brain white matter, see figures 2.2 and 2.3. Moreover, SSVD patients exhibit reduced executive function, decreased processing speed and only mild memory loss, whereas AD patients are characterized by disturbances in interpreting sensory information and pronounced loss of memory. However, the clinical phenotype may resemble that of AD. Especially, the continuously progressive disease course is characteristic of both SSVD and AD. There are so far no established disease-specific biochemical markers for SSVD, but the blood-brain barrier (BBB) has been suggested to be involved in the pathogenesis of SSVD [23].

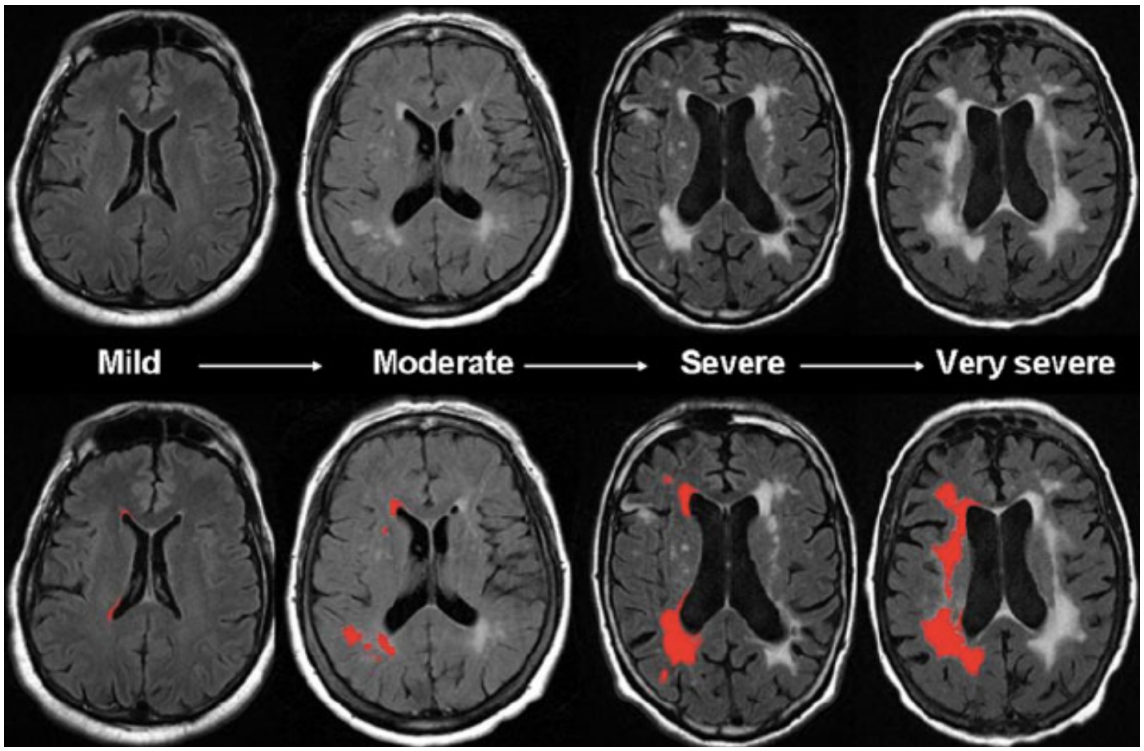



**Figure 2.2:** The Severity of MRI-detected white matter hyper-intensity in a FLAIR MRI as given by Chutinet et. al. [25]

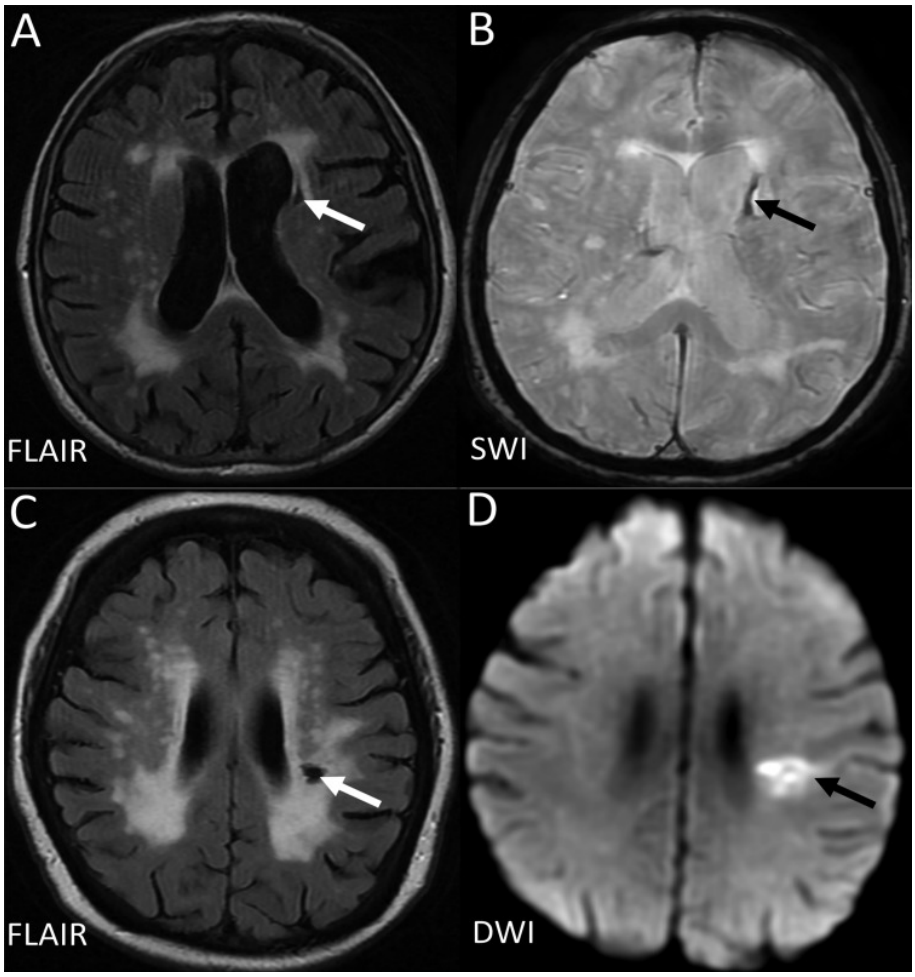



**Figure 2.3:** An example of a lacune in FLAIR MRI, susceptibility-weighted imaging(SWI) and diffusion-weighted imaging (DWI) MRIs as given by Shi et. al. [26]

## 2.2 Diagnostic procedures

The subjects that are included in the study and in the Gothenburg MCI studies cohort was classified using the Global Deterioration Scale (GDS)[27]. The GDS ranges between a rating of 1-7, where 1 is a healthy control subject and 7 is very severe cognitive decline. The classification into the different GDS levels were done by 4 main steps. (1) Using Stepwise Comparative Status Analysis (STEP), where focus was on disorientation, memory disturbance, poverty of language, reduced abstract thinking, sensory aphasia, visuospatial disturbance, visual agnosia and apraxia. (2) Through I-FLEX, which is a shorter version of Executive Interview (EXIT)[28]. EXIT focuses on anamolies in sentence repetitions, counting tasks, number-letter tasks and word fluency among others. (3) The Mini-Mental State Examination (MMSE) [21]. Finally, (4) Clinical Dementia Rating(CDR) which is an assessment based on both the subject being classified and an outside observer and informant. For example, the guidelines to grade a subject with GDS 4, were that STEP $> 1$, I-FLEX $> 3$, CDR $> 1$ and MMSE $\leq 25$. Once these requirements were met a consensus decision of each subject was made among the physicians at the Gothenburg MCI Study [23].

The practitioners at the Gothenburg MCI study who later classified patients into the specific types of dementia diagnoses used their research protocol and criteria focused on the clinical symptoms displayed by the subjects and MRI data.

Apart from the MRI images, fluid biomarker data has often been used for AD diagnosis as it guarantees a high accuracy value. However, biomarker data was not considered in the original diagnoses. For instance, when classifying an AD subject, the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria was used [29], which means that for a patient to be classified with AD, predominant parietotemporal lobe symptoms was necessary with non existent or very mild White Matter Hyperintensities visible with MRI [4].

Although originally the classification of patient diagnosis was not set using the CSF biomarkers, the classification was later revisited and updated using the criteria developed for diagnostics with biomarker data. It means that when applicable for diagnoses such as AD, the diagnosis would have to be in line with the International Working Group-2 (IWG-2) criteria [30]. This approach gives the possibility to classify patients into the Mix category. Thus, patients with biomarker data suggesting a strong AD case that was previously classified with VaD would now be qualified for a mixed AD/VaD diagnosis[23].
The diagnosis of VaD disease was set according to the Erkinjuntti criteria[31]. The patients therefore had to have cerebral WMHs detectable by MRI-scans and prevalent frontal lobe symptoms. The WHMs had to be classified as mild, moderate or severe in line with the Fazenka criterion[32]. When it comes to a mild case, the VaD diagnosis was only determined if lobe syndromes were not marked out. In addition for an VaD diagnosis to be valid, the biomarker data had to be out-ruled from an AD diagnosis, otherwise a mixed diagnosis would be set. The classifications set by the Gothenburg MCI study is aligned with the Vascular Impairment of Cognition Classification Consensus Study (VICCCS) [33].

## 2.3 Neuropsychological testing

The Gothenburg MCI study also utilized neuropsychological tests in addition to the GDS classification test that was used[34]. The neuropsychological tests used was the Rey Auditory Verbal Learning Test (RAVLT) as well as the Trail Making Test A (TMT-A) and B (TMT-B). The RAVLT tests was used to assess the episodic memory of the patients, the TMT-A for visual scanning capabilities and TMT-B for complex attention span.

## 2.4 Convolutional Neural Networks

Convolutional neural network has been widely used in medical image diagnosis field [35] and [36]. It is found that the paper "A deep CNN based multi-class classification of Alzheimer's disease using MRI" [14], mentioned the outstanding classification result for AD using GoogleNet, GoogleNet has been selected for the project. After the attempts of using GoogleNet, ResNet34 with Medical Open Network for Artificial Intelligence (MONAI) library support has also been selected for the project. This section briefly highlights the theory behind the model architectures of GoogleNet[2] and ResNet[1].

Both networks utilize the residual network design, which is to have skipped connection linkage between layers. By having this design, the problem of the vanishing gradient problem can be reduced to have a more robust model.

### 2.4.1 GoogLeNet

GoogLeNet is a 22-layer deep CNN developed by Szegedy et. al. [2]. The model architecture was able to perform state-of-the-art in the ImageNet Large-Scale Visual Recognition Challenge 2014 [37]. GoogLeNet has also been able to perfrom well in other domains such as medical image classification for cognitive decline [14]. For an overview of the model architecture see figure 2.4 below.

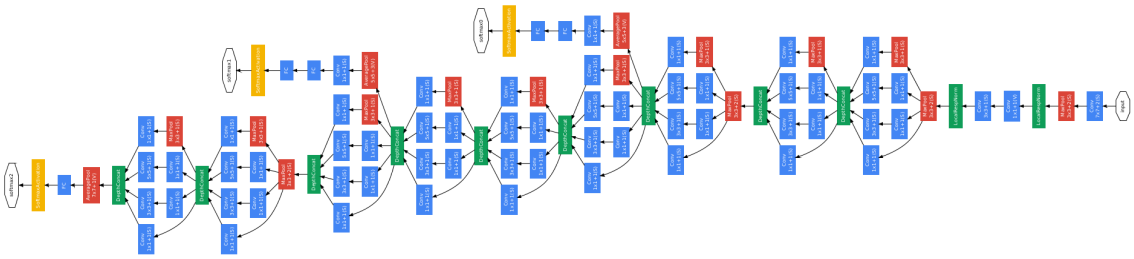


**Figure 2.4:** The GoogLeNet architecture layout as given by Szegedy1 et. al.[2] depicting the whole layout of the GoogLeNet architecture.

### 2.4.2 ResNet34

ResNet34 is a 34 layers residental deep CNN developed by Microsoft research team [38]. See figure 2.5 below. It won the ImageNet competition in 2015 by using the residual network design. The model has been used in 3D medical image prediction and it has been used in 3D medical image predictions [39]. There have been previous papers using ResNet in MRI brain for detecting AD [40] [41].

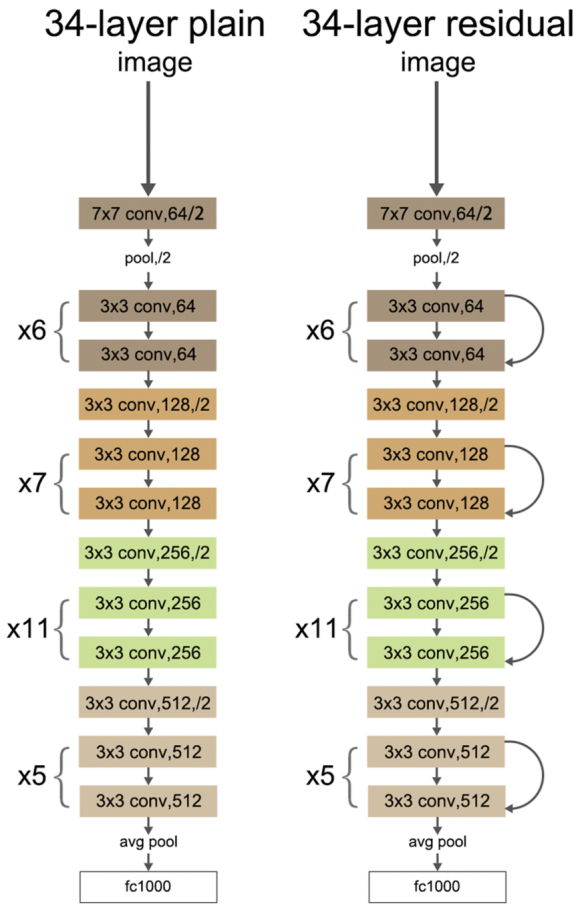



**Figure 2.5:** The resnet34 architecture layout as given by Zhang et. al.[42] depicting a regular CNN network on the left-hand side and a network with residual connections on the right-hand side.

## 2.5 Transfer Learning

In this section we will go through the different transfer learning approaches that was researched and implemented in this project. It is considered that transfer learning is recommended if possible, as it gives your model a better starting point, utilizes previous work and often leads to better results. In particular it has been shown that transfer learning from imagenet can give good results in many domains as it give a good general pretraining on image data [43]. However, when utilizing transfer learning for medical image task there are a few main challenges that could be to be addressed, these include converting 2D weights to 3D and utilizing RGB color images to gray-scale.

Transfer learning is a widely used and researched concept in the field of ML and AI and can often result in faster training times and better end results if positive transfer can be achieved [44]. It is however possible that negative transfer can be realized, this is a much less studies subject but means that trying to utilize pretrained networks has a negative impact on the target application. The most common case where transfer learning can have a positive impact is when the pretrained model is trained on similar data in both domain and format. However it has been shown that transfer learning can be used in cross-domain applications with positive transfer effects [7][11][45].

When utilizing transfer learning in a cross-domain application some challenges might need to be addressed. One of these challenges is that the data format might be different. For instance if you have pretrained weights on a 2D image set and want to utilize it for 3D image classification the weights needs to be changed to handle a 3D context instead. In a paper by Yang et. al. [46] a proposed technique called ACS Convolutions was put forward. This method, promises that it is theoretically possible to convert 2D weights into the 3D domain for any model architecture with a decrease in model size and computational costs. Yang et. al. also states that through extensive testing it can be shown that pretrained models converted to 3D via ACS Convolution outperform 3D CNN networks that are initalized from start.

# 3
# Methodology

## 3.1 Processing data

In this section we will give a detailed explanation as to which steps were taken and how the processing of the data applied in this project was done. In order to get an abstract overview of the steps, see figure 3.1 below.

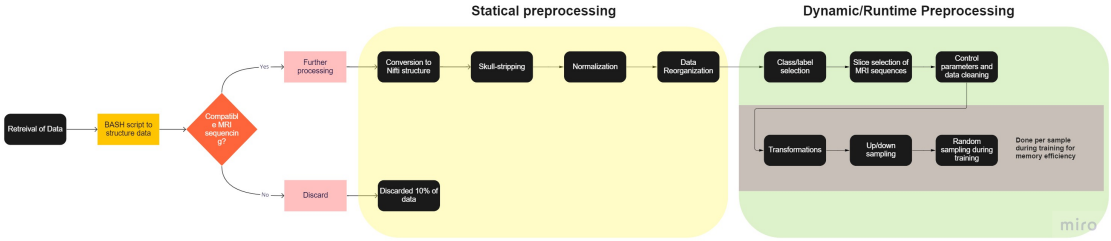



**Figure 3.1:** Flowchart giving a high-level overview of the static and dynamic preprocessing steps performed on the cohort dataset

### 3.1.1 Retrieving data

The dataset given by the Gothenburg MCI study was unprocessed 3D gray-scale brain MRI data in DICOM format with their respective biomarker data and diagnosis. In this project, in order to maximize the number of records, biomarker data is not used in the model as it contains many missing data. Only the diagnosis data and the MRI records have been included in the model.

For the MRI scans, each MRI datapoint was assigned a random number used to map the images to a specific patient and a specific time point. For instance, the scan record with random number: 42 could be assigned to patient ID 10 at year 0 visit. This dataset has around 30 scan types which differs in scan angle (sagittal, axial and coronal angles), MRI sequencing technique (T1, T2 and FLAIR) and different parameters used in the scanning machine to perform the capture such as slice depth. Every slice from the given scan type has a resolution of 416 x 512 (height x width). There are 1155 scan records given by 576 patients. The number of patients is less than the number of scan record because some patients have conducted the MRI scan test for more than 1 times as mentioned. Note that each patient can only participate in the test at least 2 years after their last scan. Additionally, scan data from each visit has been treated as completely different records regardless of the

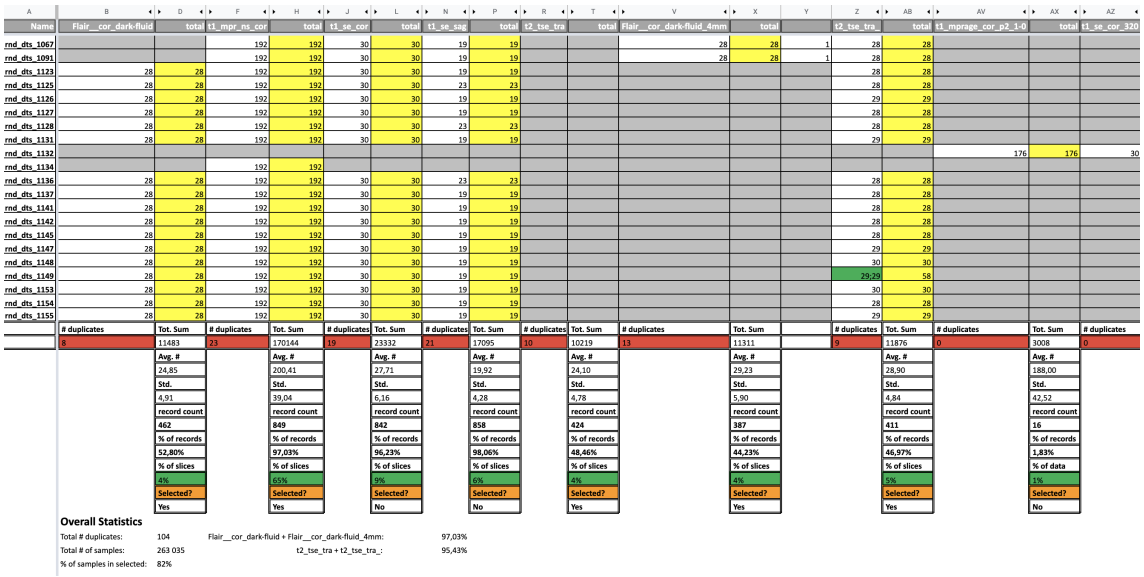
time points. This assumption is made so that it is possible to have more records of each individual disease type.

Apart from this, it is found that for each visit, the types of MRI scan each patient has conducted can be different because there could be some patients only agreed to participated in certain brain scan operations but not every type. Thus, to handle the high data inconsistency, a bash program has been written to analyse the dataset and decide the best scan types for the machine learning model. In total, 30 different MRI sequencing techniques was used, most of these were incompatible with each other and the sequencing techniques with the most amount of data coverage was chosen. It was possible to retain over 90% of the original data with the most commonly used T1, T2 and FLAIR sequences.

#### 3.1.1.1 Structuring data & Selecting data


**Figure 3.2:** An example of DICOM file structure

As the DICOM images save each slice image into a separate .dicom file as shown above in figure 3.2, it will be possible to use a terminal bash script program to count the number of slices from each MRI sequencing technique group for each patient and calculate the numbers of patients in each group. The overview result was been printed and exported into a csv file for further analysis in Excel spreadsheets, see figure 3.3. By doing so it was possible to get an overview of the data coverage as shown below for each MRI sequencing technique and thus make a selection of which part of the data to be used.



| Name | Flair__cor_dark-fluid | total | t1_mpr_ns_cor | total | t1_se_cor | total | t1_se_sag | total | t2_tse_tra | total | Flair__cor_dark-fluid_4mm | total | | t2_tse_tra_ | total | t1_mprage_cor_p2_1-0 | total | t1_se_cor_320 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| rnd_dts_1067 | | | 192 | 192 | 30 | 30 | 19 | 19 | | | 28 | 28 | 1 | 28 | 28 | | | |
| rnd_dts_1091 | | | 192 | 192 | 30 | 30 | 19 | 19 | | | 28 | 28 | 1 | 28 | 28 | | | |
| rnd_dts_1123 | 28 | 28 | 192 | 192 | 30 | 30 | 19 | 19 | | | | | | 28 | 28 | | | |
| rnd_dts_1125 | 28 | 28 | 192 | 192 | 30 | 30 | 23 | 23 | | | | | | 28 | 28 | | | |
| rnd_dts_1126 | 28 | 28 | 192 | 192 | 30 | 30 | 19 | 19 | | | | | | 29 | 29 | | | |
| rnd_dts_1127 | 28 | 28 | 192 | 192 | 30 | 30 | 19 | 19 | | | | | | 28 | 28 | | | |
| rnd_dts_1128 | 28 | 28 | 192 | 192 | 30 | 30 | 23 | 23 | | | | | | 28 | 28 | | | |
| rnd_dts_1131 | 28 | 28 | 192 | 192 | 30 | 30 | 19 | 19 | | | | | | 29 | 29 | | | |
| rnd_dts_1132 | | | | | | | | | | | | | | | | 176 | 176 | 30 |
| rnd_dts_1134 | | | 192 | 192 | | | | | | | | | | | | | | |
| rnd_dts_1136 | 28 | 28 | 192 | 192 | 30 | 30 | 23 | 23 | | | | | | 28 | 28 | | | |
| rnd_dts_1137 | 28 | 28 | 192 | 192 | 30 | 30 | 19 | 19 | | | | | | 28 | 28 | | | |
| rnd_dts_1141 | 28 | 28 | 192 | 192 | 30 | 30 | 19 | 19 | | | | | | 28 | 28 | | | |
| rnd_dts_1142 | 28 | 28 | 192 | 192 | 30 | 30 | 19 | 19 | | | | | | 28 | 28 | | | |
| rnd_dts_1145 | 28 | 28 | 192 | 192 | 30 | 30 | 19 | 19 | | | | | | 28 | 28 | | | |
| rnd_dts_1147 | 28 | 28 | 192 | 192 | 30 | 30 | 19 | 19 | | | | | | 29 | 29 | | | |
| rnd_dts_1148 | 28 | 28 | 192 | 192 | 30 | 30 | 19 | 19 | | | | | | 30 | 30 | | | |
| rnd_dts_1149 | 28 | 28 | 192 | 192 | 30 | 30 | 19 | 19 | | | | | | 29/29 | 58 | | | |
| rnd_dts_1153 | 28 | 28 | 192 | 192 | 30 | 30 | 19 | 19 | | | | | | 30 | 30 | | | |
| rnd_dts_1154 | 28 | 28 | 192 | 192 | 30 | 30 | 19 | 19 | | | | | | 28 | 28 | | | |
| rnd_dts_1155 | 28 | 28 | 192 | 192 | 30 | 30 | 19 | 19 | | | | | | 29 | 29 | | | |
| | # duplicates | Tot. Sum | # duplicates | Tot. Sum | # duplicates | Tot. Sum | # duplicates | Tot. Sum | # duplicates | Tot. Sum | # duplicates | Tot. Sum | | # duplicates | Tot. Sum | # duplicates | Tot. Sum | # duplicates |
| | 8 | 11483 | 23 | 170144 | 19 | 23332 | 21 | 17095 | 10 | 10219 | 13 | 11311 | | 9 | 11876 | 0 | 3008 | 0 |
| | | Avg. # | | Avg. # | | Avg. # | | Avg. # | | Avg. # | | Avg. # | | | Avg. # | | Avg. # | |
| | | 24,85 | | 200,41 | | 27,71 | | 19,92 | | 24,10 | | 29,23 | | | 28,90 | | 188,00 | |
| | | Std. | | Std. | | Std. | | Std. | | Std. | | Std. | | | Std. | | Std. | |
| | | 4,91 | | 39,04 | | 6,16 | | 4,28 | | 4,78 | | 5,90 | | | 4,84 | | 42,52 | |
| | | record count | | record count | | record count | | record count | | record count | | record count | | | record count | | record count | |
| | | 462 | | 849 | | 842 | | 858 | | 424 | | 387 | | | 411 | | 16 | |
| | | % of records | | % of records | | % of records | | % of records | | % of records | | % of records | | | % of records | | % of records | |
| | | 52,80% | | 97,03% | | 96,23% | | 98,06% | | 48,46% | | 44,23% | | | 46,97% | | 1,83% | |
| | | % of slices | | % of slices | | % of slices | | % of slices | | % of slices | | % of slices | | | % of slices | | % of data | |
| | | 4% | | 65% | | 5% | | 6% | | 4% | | 4% | | | 5% | | 1% | |
| | | Selected? | | Selected? | | Selected? | | Selected? | | Selected? | | Selected? | | | Selected? | | Selected? | |
| | | Yes | | Yes | | No | | No | | Yes | | Yes | | | Yes | | No | |

**Overall Statistics**

| | | | |
|---|---|---|---|
| Total # duplicates: | 104 | Flair__cor_dark-fluid + Flair__cor_dark-fluid_4mm: | 97,03% |
| Total # of samples: | 263 035 | t2_tse_tra + t2_tse_tra_: | 95,43% |
| % of samples in selected: | 82% | | |

**Figure 3.3:** The statistic data of the given MRI dataset featuring the selected scan types.

When analysing the data it was also discovered that for roughly 15% of the data samples duplicate records was captured in the final T1, T2 and FLAIR sequences. The exact reason why this duplicating was done was never found as it may have been done several years ago and no documentation of it is available. Together with the group and with manual overview of the samples the correct duplicate was extracted from the data such that no duplicates was included in the final versions.

T1__mpr__se__cor are MRI scans captured using T1 technology, coronal angle and with 192 slices taken for each patient. FLAIR__cor__dark-fluid and FLAIR__cor__dark-fluid__4mm are FLAIR MRI scans taken from coronal angle. Both scan type contains around 28 slices. The difference between these 2 scan types is the parameter to decide the distance between each discrete brain scan. As images taken using coronal angle scan the patient's brain from top to bottom, the parameter, slice depth, will be used to control the vertical resolution of the 3D scan data. Samples taken inside FLAIR__cor__dark-fluid__4mm used 4mm and scans in FLAIR___cor__dark-fluid used 5mm as the parameter. However, as it is observed that patients either has data inside FLAIR___cor__dark-fluid__4mm or FLAIR___cor__dark-fluid, these 2 data has been merged together for further usage. A function to align the number of slices for each scan type will be introduced in later preprocessing section. T2__tse__tra and T2__tse__tra__ are scans taken from T2 scanning method, axial angle and with 23 slices taken for each brain.

### 3.1.2 Static Pre-processing

There are 2 types of data preprocessing, static and dynamic. Static data preprocessing are one-off procedures, which will be implemented directly on the dataset stored inside disk space. While dynamic data preprocessing procedures changes with the hyper-parameter selected for the run. Static preprocessing procedures con-

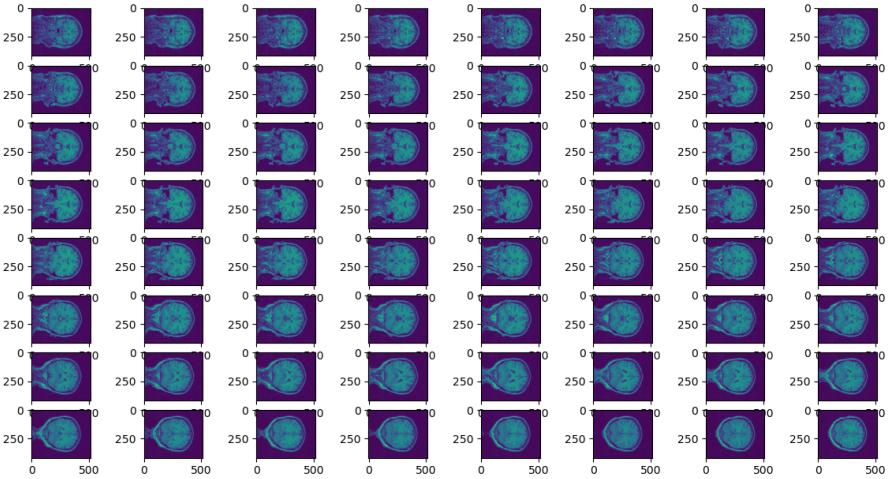
sists of 4 parts, which are Nifti conversion, Skull-stripping, Normalization and Data Reorganization.

#### 3.1.2.1 Nifti conversion

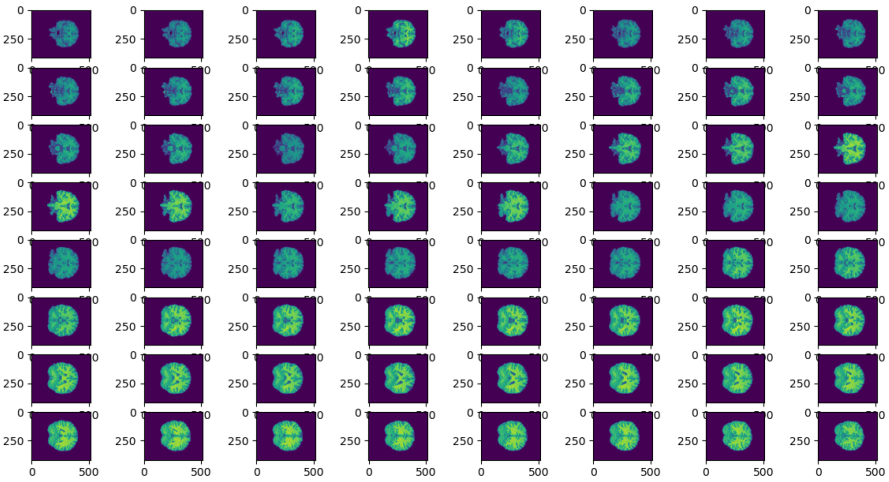
First, instead of directly using multiple DICOM slices as inputs, the files have been converted into Nifti format (.nii) using a C++ based library named dcm2niix [47]. The new input format can increase the overall I/O speed during data access as the program can now directly read the scans as a whole. Another reason for the conversion will be the extensive amount of support offered by existing frameworks such as FSL[48] for brain scan data in Nifti format instead of DICOM.

#### 3.1.2.2 Skull-stripping

As an brain scan contains an extensive amount of unrelated information such as the eyes and the skull, which will make the model hard to focus on the brain pixels. The C++ based program BET2 [49] provides with a tool to perform skull-stripping for the purpose. Apart from this, because the skull-stripping process replaces a huge amount of non-brain pixels by zero pixels. And therefore the resulting images would contain more zero pixels compared to the original slices. See figure 3.4 below for before and after skullstripping.



(a) Before Skull-stripping



(b) After Skull-stripping

**Figure 3.4:** Showing a raw MR Image with 64 slices in (A) and the skull-stripped with 64 slices In (B)

Hence, compression can therefore effectively reduce the overall size of the dataset for further purposes. It can especially further improve the I/O speed when the model loads the 3D images into the RAM during the training process. The BET program gives a skull-stripped and compressed Nifti 3D images as outputs (.nii.gz). The zipped Nifti will be unzipped during training, testing and validation phase when the data needs to be loaded.

#### 3.1.2.3 Normalization

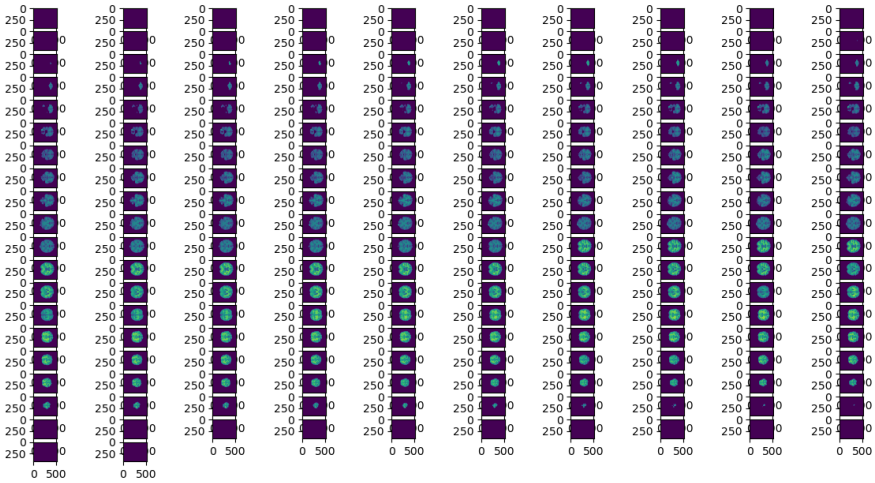
As it is observed the dataset does not have a common pixel intensity range across records, a python based normalization process of min-max scaling on pixel intensity has been implemented. Individual normalization for each sample has been used in this project, which is to normalize sample using the highest and lowest values from the 3D slices of the same patient. It is believed that per sample normalization will be better than per dataset normalization because the process will not be greatly affected by outlier samples. An intensity range of [0,255] and [0,1] have been used as different parameters during the training phase to search for the model with the best performance. The slices are normalized to [0,255] in this step, whether to use 0 and 255 for further min-max scaling to convert the range to be [0,1] has been placed as one of steps in dynamic preprocessing.

#### 3.1.2.4 Data Reorganization

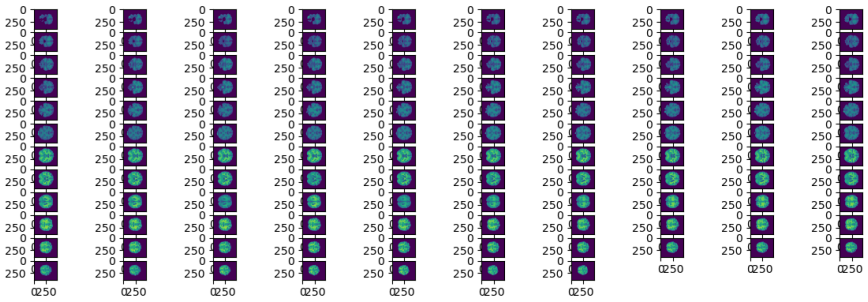
In a brain MRI there is a varying amount of images/slices of the brain depending on the thickness of each slice going through the brain. Since the human head is not completely symmetric from all sides and the brain is in the top of the head, the slices that you capture with MR will contain varying amounts of useful images. That is, in the start and the end of the slices there will always be less brain and more skull or organs such as the eyes or neck. So naturally, in the beginning and end of each brain MRI you have either a lot of black pixels or a very small sample of the end of the brain. This means that there will be a lot of redundant slices which were to be removed. Removing the redundant slices comes with the multiple benefits of speeding up training time, reducing memory usage and at the same time making it easier for the ML model to converge and not interpret nonsensical black images as valuable input.

Because we are dealing with gray scale images which have a possible value range of [0, 255] ranging from black(0) and white(255) it will be possible to set up a threshold using the sum of pixels to decide whether a slice will be discarded from the data. The possible range for the sum of pixels in an image is [0, 416x512x255=54,312,960] so if the sum of the pixels in an image is 0 then it should obviously be discarded. However, as mentioned, the edge images will also contain very little information as they will have a very small percentage of the image containing actual brain matter, so a threshold can be used to determine whether or not a slice should be discarded. In order to this empirical searches were done to find that a threshold value of 2,000,000 representing about 3.5% of the max value was a good threshold. It is observed that slices with enough information are greater than this threshold, and thus images with a sum less than 2,000,000 will be discarded. See Figure 3.5(a) and (b).

In addition to completely removing slices, the brains can also be observed to be positioned in different parts of the image. In order to remedy this and also remove even more excess information an optimal frame size was calculated and each and every brain is then placed in the center of this frame. Both the removal of slices with threshold value and the center & cropping was done by writing python scripts



(a) Before removing slices 192 slices in total



(b) After removing slices, 107 slices remaining to be used

**Figure 3.5:** Showing an original MR Image with all the slices in (A) and the processed MR Image with the black slices and any slice with a pixel sum of less then 3.5% of the maximum value. In (B) all the brain scans have also been centered in the middle.

and iterating over all the images to perform the operations.

In addition to this, the labels for each and every sample had to be gathered through Excel sheets as they were not readily available for use in combination with the MRI images. In order to do so a combination of BASH and Excel scripts was used as the MRI images was labeled with a random number which was connected to a certain patient ID in combination with the visit that the patient was on. For example, random number 25 and 512 could both be patient ID 2 but at the first and second visit. In addition to this, some of the diagnosis was missing for certain patients after the first visit that they made. However, this is expected since if there is no diagnosis filled in at the second visit, this then means that the doctor responsible determined that no change in diagnosis was done and the diagnosis from the previous visit could be considered as the correct one. However, if a diagnosis changed between two visit then a new diagnosis was filled in at the most recent visit. So in order to create the label data for this project the random number assigned to the MRI was extracted using BASH scripts and later combined with Excel scripts in order to find which random number was connected to which patient and visit together with the correct diagnosis at that time point.

### 3.1.3 Dynamic Preprocessing

Dynamic preprocessing procedures are steps that can be different from trials to trials. It is controlled by hyperparameters that can be dynamically changed between runs.

#### 3.1.3.1 Preprocessing steps

First, as instructed by the Gothenburg MCI group, duplicated records inside the dataset has been disregarded during the run. And then, Because the model will use both 3 types of scan type images, only records with diagnosis data and all 3 scan types ready will be performed in training and testing process.

#### 3.1.3.2 Monai preprocessing

Then, because the scan data is too large to be fitted inside RAM, the actual images will only be loaded from the paths in this step. After the images are loaded, a Windows method will be used to sample brain scans from the loaded 3D volumes, this step can also handle the problems occurred when merging both FLAIR___cor_dark-fluid_4mm and FLAIR___cor_dark-fluid scan type together.

The step is to ensure all the data shares the same numbers of slice by sampling. Therefore, individual image's resolution has been aligned from the step of Data Reorganization and the numbers of slice has been aligned in this step. Given 3 parameters: window_size_t1, window_size_t2 and window_size_FLAIR to control the numbers of slice used for each scan type volume, the program will sample a certain amount of continuous brain scans from the raw data.

The following python function was used in order to cut away slices from the start and the end of a image sequence. This way more control over how many slices that should be used for each image sequencing technique can be gained. The parameter window_size controls how many slices of an image that should be included and the parameter skew takes an input between [0,1] and controls what proportion of these slices should be cut from the top or bottom. Setting skew to zero would start the selected slices from the bottom and setting it to one would end the sequence at the last slice.

```
class Window(Transform):

  def __init__(self, window_size, skew):
    self.window_size = window_size
    self.skew = skew

  def __call__(self, inputs):
    self.n_slices = inputs.shape[-1]
    # Determining the cutting and window size
    cut_size = self.n_slices - self.window_size
    start_cut = cut_size*self.skew
    end_cut = self.n_slices - cut_size*(1-self.skew)

    # Handling uneven numbers
    start_cut = int(np.floor(start_cut))
    end_cut = int(np.floor(end_cut))


    # Return the cut file
    return inputs[:,:,start_cut:end_cut]
```

As it is found that the brain scans contain the most amount of information in the middle of the scans. Thus, the program will always extract middle slices from the raw data, leaving out some unused slices from the start and from the end of the 3D brain slices. Normally, these 2 amount will have a ratio closed to 0.5, but this will be controlled by the parameter skew. For example, by setting skew parameter to 0.9, 90% of the unused slices will be assigned to be start of the raw data while only 10% will be from the end of the record, which means by setting a high skew value, slices from the end will have a greater weight than that from the start.

As mentioned in static preprocessing, whether to use the intensity range of [0,1] or [0,255] will be controlled as one of the parameters. The parameter will control whether to apply min-max scaling using the range [0,255] to [0,1].
Then, the preprocessing procedure focuses on loading the input images with an appropriate input size for the model. First, as the input images are all in gray-scale images, the images will be added with only one color channel. Second, the program will convert the data to be tensor for the PyTorch's library setting.

## 3.2 Model choice and architecture

When choosing the model architecture, the motivation was to not start from scratch. The field of CNN networks is not new and a lot of work has been done by many talented researchers before. In addition to this, the aim of the project is to support the Gothenburg MCI study in their approach of using Machine learning tools. Therefore a decision was made where the implementations would be done with models that have showed a proof of concept in similar task previously. By that motivation the two model architectures chosen was GoogLeNet[2] and ResNet[1].

### 3.2.1 Binary model

Binary classification has been implemented to demonstrate the separability between classes under each scan type. To obtain the binary results, the ResNet34 model was chosen with 100 epoches and learning rate being 1e-6. This model was initialized by the orginial ResNet34 architecture described in section 2.4.2. As stated before this model was and is a state-of-the-art model for image tasks, including but not limited to classification tasks. ResNet34 was initalized from the Monai[50] library with 3D components, 1 in channel and 2 outchannels for the binary classification task. New copies of the model was created for T1, T2 and FLAIR sequences and saved back into local files in order to load the models at a later stage.

### 3.2.2 Multiclass ResNet Ensamble

In this section we go through the architecture and training method of the multiclass ResNet ensamble (MRE) model. This model was made out out three separate ResNet34 models, one for each image sequencing technique (T1, T2 and FLAIR) and one aggregating model which was a custom made model consisting of fully connected layers as well as ReLU activation functions. The model was built such that the output from all three sub-models for the imaging techniques as well as the final output from the aggregating model was outputted from the forward function. This way separate loss value could be fed back into different parts of the model depending on the individual outputs.

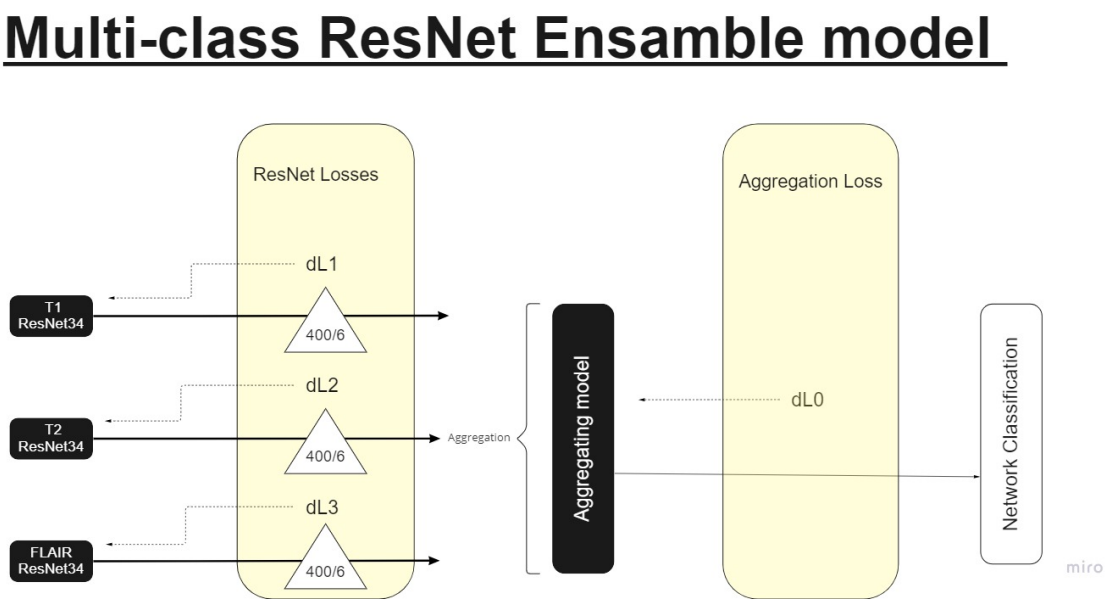



**Figure 3.6:** Model architecture of the multiclass ResNet ensamble, utilizing different loss values for different parts of the network.

Figure 3.6 above depicts the high-level design of the MRE model. As can be seen from the Image the ResNet34 models and the aggregating model received different loss values depending on their respective output. The way this was built was that the submodels for T1, T2 and FLAIR could output different size vectors depending on if they were to be fed to the aggregation model or to be outputted for classification. In the last layer of each ResNet34 the fully connected layer took an input of 400 logits and outputted 6, one for each class. This output was modified depending on if it was to be outputted to the aggregating model, in which case the full 400 logits vector was outputted or if it was to be outputted to for classification and loss value calculation, in which case a 6 logits vector was outputted. The reason why a SoftMax layer was not used in the final output was because the loss function utilized was the pytorch CrossEntropyLoss function [51] which performs a SoftMax operation on the input that it get.

# 4
# Experimental setups

This section goes through some of the experimental setups that was tried along the way. These vary from different types of pre-processing steps, architecture structures and hyper-parameter configurations.

## 4.1 Weight initialization

As mentioned earlier in the report, the data is highly unbalanced with the Mix and VaD classes being particularly underrepresented. The first attempt at fixing this issue was to utilize weight initialization when creating the loss function. The weights was initialized proportionally to the amount of samples each class contained. The weights was then inputted in two versions, the "raw" proportional weights and a normalized version of the weights where the sum of the weights was one.

$$\overrightarrow{W_i^0} = \frac{\sum c_i}{c_i} \tag{4.1}$$

Equation 4.1 depicts the calculation used for the raw proportional weights, which will give a higher value for the underrepresented classes. However, as it was noticed that inputting these weights, which have a value range of between

$$\overrightarrow{W_i^1} = \frac{\overrightarrow{W_i^0}}{\sum \overrightarrow{W_i^0}} \tag{4.2}$$

Equation 4.2 shows how we can then use the raw weights and normalize them such that the sum of $\overrightarrow{W^1}$ becomes one.
These weights are then inputted to the loss function in order to control how much loss that should be attributed to each class label. In our case we used the Pytorch library [52] to determine our loss function which was set to Softmax loss.

## 4.2 Loss function tweaking

Changing the way that the loss was fed back to the model was experimented with quite a bit. There was attempts to modify, change, enhance, interpolate between states and many other attempts. No tweaking of the loss function however yielded significant results.

## 4.3 Window size

After the preprocessing procedures, T1 MRI has reduced the number of slices from 192 to around 100 slices. The window size parameter for T1 has been chosen to be 70. A windows size test from 20 to 100 has been made using AD vs CTRL in T1.

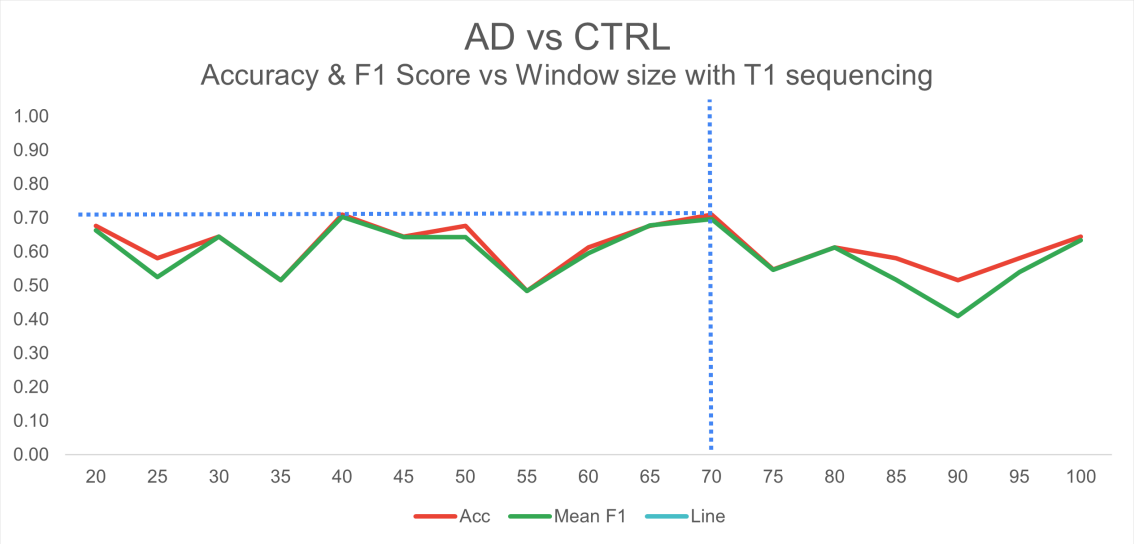



**Figure 4.1:** Linechart showing the accuracy and weighted mean F1 score as a function of the window size, i.e. the amount of slices used from each sample. The best metrics was obtained with a window size of 40 and 70.

It is clear that the accuracy reaches the best in 40 and 70. And it is believed that 70 slices allows the model to have more information for the learning process, 70 has been selected for T1. As unlike T1, these T2 and FLAIR scan types only have 23 to 30 slices for each record. Thus, For T2 and FLAIR, windows size 20 has been chosen for both image type as it ensures the model to obtain enough information from the dataset. However, as it is observed that from the the starting slices and the ending slices from T2 suffers from skull-stripping preprocessing artifacts, binary test with windows size being 10 and 20 for T2 have both been executed.

## 4.4 Model architecture choice

Seeing that AD vs CTRL has been implemented from previous research mentioned and a results with high level accuracy has been obtained, AD vs CTRL in this project has been used to validate whether it is possible to use the dataset for the classification tasks. Using ResNet34, a result with around 80% testing accuracy has been obtained using AD vs CTRL in T1 from the dataset. Thus, it demonstrates the level of correctness and the data integrity of the dataset. Then, attempts to use GoogleNet for the prediction has been made.

### 4.4.1 Googlenet

Since GoogLeNet is a well know high performing model that was originally created to be used on natural image classification [2], there exists pre-trained version of the model. These pre-trained models come with varying amounts of training on natural images, meaning real-world objects in the from of three channel RGB images from the Imagenet dataset [37]. Since it has been shown in previous work that using these pre-trained version can be utilized on medical imaging data via transfer learning, this was the approach chosen in this project as well.

Since the Imagenet dataset only contains 2D images and the MRI scans in the Gothenburg MCI cohort are 3D. The first step needed in order to use GoogLeNet for this project was to convert the pre-trainined weights from the 2D to 3D domain. This was performed using the method proposed by Yang et. al. [46] discussed in section [44]

In addition to converting the 2D weights in the GoogLenet model one also have to consider that the original model was pretrained on natural images in RGB. One therefore have to make a conversion of the three channeled input into one-channel gray scale images. There is no commonly accepted way of making this conversion that guarantees that the three channel input layer will be able to make sense for one channel gray-scale images. There is a number of proposed ways of doing these that have been explored in this project. Those are (1) inputting three copies of the image into each channel and giving the correct value for each channel corresponding to the RGB value of the gray-scale image. This way one can artificially utilize all three RGB channels and create the correct corresponding gray-scale image. For this project however we also have three different image capturing sequences T1, T2 and FLAIR. Therefore the second approach (2) was to input all three image sequences T1, T2 and FLAIR directly into the three channels of the model.

After applying the ACS conversion and trying both (1) inputting three copies and (2) to input T1, T2 and FLAIR into the GoogLeNet the training showed that the model did not converge/learn on the training data. The model struggled to even perform at baseline performance suggesting that it confused different classes with each other, see 5.3 and 5.2 for more details. Because the model was pretrained on natural 2D images and since it struggled with learning the experiments with GoogLeNet was quickly discontinued. This is because the error source becomes multi-fold as it could either be the 2D to 3D or three to one channel conversion, it could also be that the cross-domain transfer has a negative transfer effect, or it could be a problem with the data. To eliminate the source of error for easier analysis ResNet34 was chosen instead.

### 4.4.2 ResNet

The ResNet architecture was initalized using the MONAI library. The model there is adapted from the paper by Kaiming et. al. [38] and pretrained ResNet for 3D medical imaging context can be downloaded manually from the work done by Chen

et. al. [53]. As we wanted to utilize transfer learning to the extent possible, ResNet was firstly initialized from the pretrained weights available from the work done by Chen et. al. these weights, albeit from medical imaging data, was originally trained to perform lung segmentation. However, since it is still MRI gray-scale images within the medical field and pretrained on 3D CNN this seemed like a better starting point then that of pretrained models on imagenet [37].

The pretrained ResNet34 did not however have any positive transfer effects for this thesis. The initial starting point was as good as the random initialization that one gets when just starting the training from scratch. Therefore the attempts of trying to use a pretrained version was discontinued as loading the weights requires time and memory usage, which was not put to any good use.

# 5
# Results and Discussion

## 5.1 Multi-class results

The first attempt for the multi-class model was to utilize the GoogleNet architecture pretrained on the Imagenet dataset [2]. As described in section 4.4.1 these required to transform the 2D weights in the model to 3D. Additionally since it was pre-trained on three channel RGB images it also required conversion to gray-scale images.

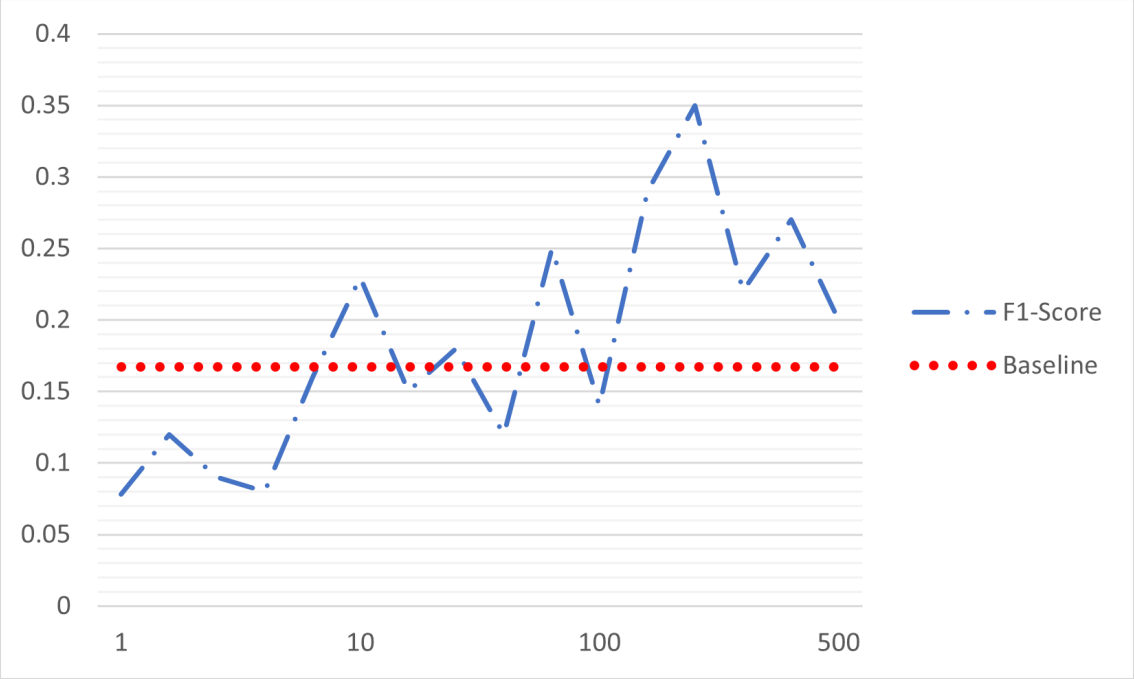



**Figure 5.1:** Linechart showing the performance of the multi-class model, y-axis depicting the F1-score over epoch 1-500 compared to baseline of 0.17. As can be observed, the performance is very unstable and does not seem to converge over time. The best result obtained was an F1-score of 0.36. This result was obtained using a multi-class model consisting of an ensemble of ResNets with an aggregation layer.

The above result in figure 5.1 was achieved using the architecture described in section 3.2.2.
However, the model was unfortunately not significantly better than baseline. Despite multiple trials with different model architectures, none of the trials was able to go

above 0.36 in F1-score. See figure 5.1 below for the performance of the best multi-class model over 500 epochs.

## 5.2 Binary results

As mentioned above, the project then focuses on investigating the separability between labels using binary models. The result obtained from the binary test uses both F1 score and the accuracy values as the measure of performance. Both validation and testing dataset has been included into the result. Noted that according to the MCI study group, it is expected that a real-world doctor can have a diagnosis with around 80% accuracy. It is considered that the result follows the underlined assumptions from previous research done within the field of cognitive impairment.

| Details | | | | Validation | | | | Test | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Classes | A | B | Scan Type | acc | F1 score A | F1 score B | mean F1 | acc | F1 score A | F1 score B | mean F1 |
| AD vs VaD | AD | VaD | FLAIR | 0.71 | 0.71 | 0.71 | 0.71 | 0.81 | 0.89 | 0.40 | 0.79 |
| CTRL vs VaD | CTRL | VaD | FLAIR | 0.82 | 0.79 | 0.85 | 0.82 | 0.65 | 0.75 | 0.46 | 0.70 |
| MCI vs VaD | MCI | VaD | FLAIR | 0.38 | 0.42 | 0.34 | 0.38 | 0.36 | 0.47 | 0.18 | 0.44 |
| SCI vs VaD | SCI | VaD | FLAIR | 0.64 | 0.72 | 0.50 | 0.61 | 0.87 | 0.93 | 0.00 | 0.84 |
| VaD vs MIX | VaD | MIX | FLAIR | 0.71 | 0.67 | 0.75 | 0.71 | 0.50 | 0.29 | 0.62 | 0.52 |
| CTRL vs VaD | CTRL | VaD | T1 | 0.76 | 0.79 | 0.73 | 0.76 | 0.75 | 0.85 | 0.29 | 0.76 |
| SCI vs VaD | SCI | VaD | T1 | 0.94 | 0.97 | 0.50 | 0.92 | 0.87 | 0.93 | 0.00 | 0.84 |
| VaD vs MIX | VaD | MIX | T1 | 0.79 | 0.77 | 0.80 | 0.78 | 0.50 | 0.00 | 0.67 | 0.47 |
| CTRL vs VaD | CTRL | VaD | T2 | 0.62 | 0.61 | 0.63 | 0.62 | 0.50 | 0.67 | 0.00 | 0.57 |
| MCI vs VaD | MCI | VaD | T2 | 0.52 | 0.43 | 0.59 | 0.51 | 0.39 | 0.51 | 0.19 | 0.48 |
| SCI vs VaD | SCI | VaD | T2 | 0.59 | 0.68 | 0.44 | 0.56 | 0.81 | 0.89 | 0.25 | 0.83 |
| VaD vs MIX | VaD | MIX | T2 | 0.50 | 0.36 | 0.59 | 0.48 | 0.40 | 0.00 | 0.57 | 0.40 |

**Figure 5.2:** Figure showing part of the performance in VaD or Mix prediction in terms of accuracy, individual F1 score and weighted mean F1 score in both testing and validation.

First, given underrepresented VaD and Mix labels, it is obvious that VaD and Mix prediction test pairs suffers from insufficient amount of training, validation and testing data in VaD and Mix labels. However, the model still seems to be able to pick up features from VaD to classify the disease from AD or CTRL in FLAIR with 81% accuracy and 65% accuracy respectively. This result is considered to follow the medical field's knowledge on FLAIR sequencing and the disease types. As mentioned before, according to [24], WMHs and lacunes has been recognized as the features that distinguishes VaD disease the most. And according to Norcliff et. al. [54] and Wardlaw et. al. [55], WHMs and lacunes are particularly visible from FLAIR.

Despite of the high accuracy in certain VaD pairs using FLAIR, the issue of insufficient data in VaD still plays an essential role. The F1 scores for VaD or Mix in those pairs are still not as high as the other labels. However, this value is already considered to be much bigger than that in T1 and T2. For example, in T1's CTRL vs VaD or T2's SCI vs VaD, although both pairs reveal good testing accuracy values, the F1 scores for VaD are still low (around 0.25). It is also considered to follow medical field's knowledge on T1 and T2 image that both image types are not used for VaD classification.

| Details | | | | Validation | | | | Test | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Classes | A | B | Scan Type | acc | F1 score A | F1 score B | mean F1 | acc | F1 score A | F1 score B | mean F1 |
| AD vs MCI | AD | MCI | T1 | 0.64 | 0.55 | 0.70 | 0.63 | 0.68 | 0.45 | 0.78 | 0.67 |
| AD vs MCI | AD | MCI | T2 | 0.69 | 0.40 | 0.79 | 0.65 | 0.71 | 0.48 | 0.80 | 0.69 |
| AD vs VaD | AD | VaD | FLAIR | 0.71 | 0.71 | 0.71 | 0.71 | 0.81 | 0.89 | 0.40 | 0.79 |
| CTRL vs AD | CTRL | AD | FLAIR | 0.71 | 0.74 | 0.67 | 0.70 | 0.80 | 0.83 | 0.75 | 0.80 |
| CTRL vs AD | CTRL | AD | T1 | 0.71 | 0.72 | 0.69 | 0.70 | 0.67 | 0.72 | 0.58 | 0.66 |
| CTRL vs AD | CTRL | AD | T2 | 0.68 | 0.75 | 0.55 | 0.66 | 0.63 | 0.62 | 0.65 | 0.63 |
| CTRL vs MCI | CTRL | MCI | T2 | 0.62 | 0.65 | 0.58 | 0.62 | 0.62 | 0.58 | 0.65 | 0.62 |
| CTRL vs VaD | CTRL | VaD | FLAIR | 0.82 | 0.79 | 0.85 | 0.82 | 0.65 | 0.75 | 0.46 | 0.70 |
| CTRL vs VaD | CTRL | VaD | T1 | 0.76 | 0.79 | 0.73 | 0.76 | 0.75 | 0.85 | 0.29 | 0.76 |
| MCI vs MIX | MCI | MIX | FLAIR | 0.68 | 0.70 | 0.65 | 0.68 | 0.69 | 0.79 | 0.38 | 0.70 |
| SCI vs AD | SCI | AD | FLAIR | 0.64 | 0.64 | 0.64 | 0.64 | 0.66 | 0.72 | 0.56 | 0.67 |
| SCI vs AD | SCI | AD | T2 | 0.74 | 0.81 | 0.59 | 0.74 | 0.73 | 0.80 | 0.59 | 0.73 |
| SCI vs MCI | SCI | MCI | FLAIR | 0.59 | 0.68 | 0.44 | 0.56 | 0.66 | 0.74 | 0.53 | 0.64 |
| SCI vs MIX | SCI | MIX | T2 | 0.82 | 0.84 | 0.80 | 0.82 | 0.66 | 0.78 | 0.25 | 0.67 |
| SCI vs VaD | SCI | VaD | T2 | 0.59 | 0.68 | 0.44 | 0.56 | 0.81 | 0.89 | 0.25 | 0.83 |

**Figure 5.3:** Figure showing test pairs with a good performance, i.e. >0.6 testing accuracy and >0 in both testing F1 scores.

It is defined that with an accuracy value >0.6, it reveals that the model is capable of extracting some features and learn from the training samples to perform a prediction. To be considered as high degree of separability to be used for multi-class classification, it is believed that it should be over 70% accuracy. Thus, from the result presented, only 6 pairs reveals an accuracy $> 0.7$. Hence, the experiment shows the reason why Multi-class label classification does not behave well with this dataset.

For this model, it shows from the result that the model can perform well in AD vs CTRL. It matches with the expectation as there have been research papers suggesting the possibility of the classification using AI models. As it is believed that AD diagnosis is mostly related to the overall mass of the whole brain, it should be possible for the model to capture this feature from these 3 scan types. In the results presented above, the model demonstrates the ability to classify AD patients from healthy subjects with all three scan types. FLAIR can obtain up to 80% testing accuracy with 0.8 F1 weighted score. T1 and T2 can both about 65% testing accuracy for the prediction.

Then, for SCI, as mentioned in the theory section 2.1.3, it is reported that the SCI label in this dataset is highly similar to CTRL. Thus, the fact the model does not perform well in SCI vs CTRL from all 3 scan types follows the expectation. Because of that, note that for the SCI vs other labels pairs, the model behaves similarly as CTRL vs other labels.

Lastly, for MCI, it is believed that the model in most cases does not separate between MCI, SCI and CTRL. As almost all the MCI vs SCI and MCI vs CTRL pair does not obtain an accuracy value >0.6 and also the model seems to perform well in MCI vs AD pairs, MCI has expressed a high similarly as CTRL, which matches with the speculation for the disease. As mentioned from 2.1.4, MCI means mild cognitive impairment and therefore it should share a certain amount of similarity with CTRL because the disease does not constrain daily life. Apart from this, MCI should still be considered as an early stage from Cerebrovascular disease, systemic,

and other neurodegenerative disorders including AD. Hence, it stands to reasons that the MCI class can be seen as a composite type of early diseases and thus, it is possible to have different features across samples. Therefore, it is believed that it can be possible to improve the accuracy by further dividing the MCI class into different groups such as pre-AD or pre-VaD by using some common features such as WHMs, lacune, cerebrovascular microbleed and perivascular space.

## 5.3 Separability

To assist in the separability analysis, graphs have been created showing the connection between each class pair. Note that it may not be completely accurate to determine whether VaD and Mix can be separated considering the amount of these 2 labels inside testing and Validation dataset. It is defined that for a pair to be separable, it should performed >0.6 in their mean F1 score while each individual F1 score cannot be lower than 0.4. And then, solid line (separable) is used to present whether a pair shows separability both in testing data and validation data. For the dashed line (often separable), the pair shows separability in only one of the testing or validation dataset. And lastly red dotted line is used to denoted non-separable pairs, see figures 5.4 5.5 and 5.6 below for the separability graph of T1, T2 and FLAIR respectively.

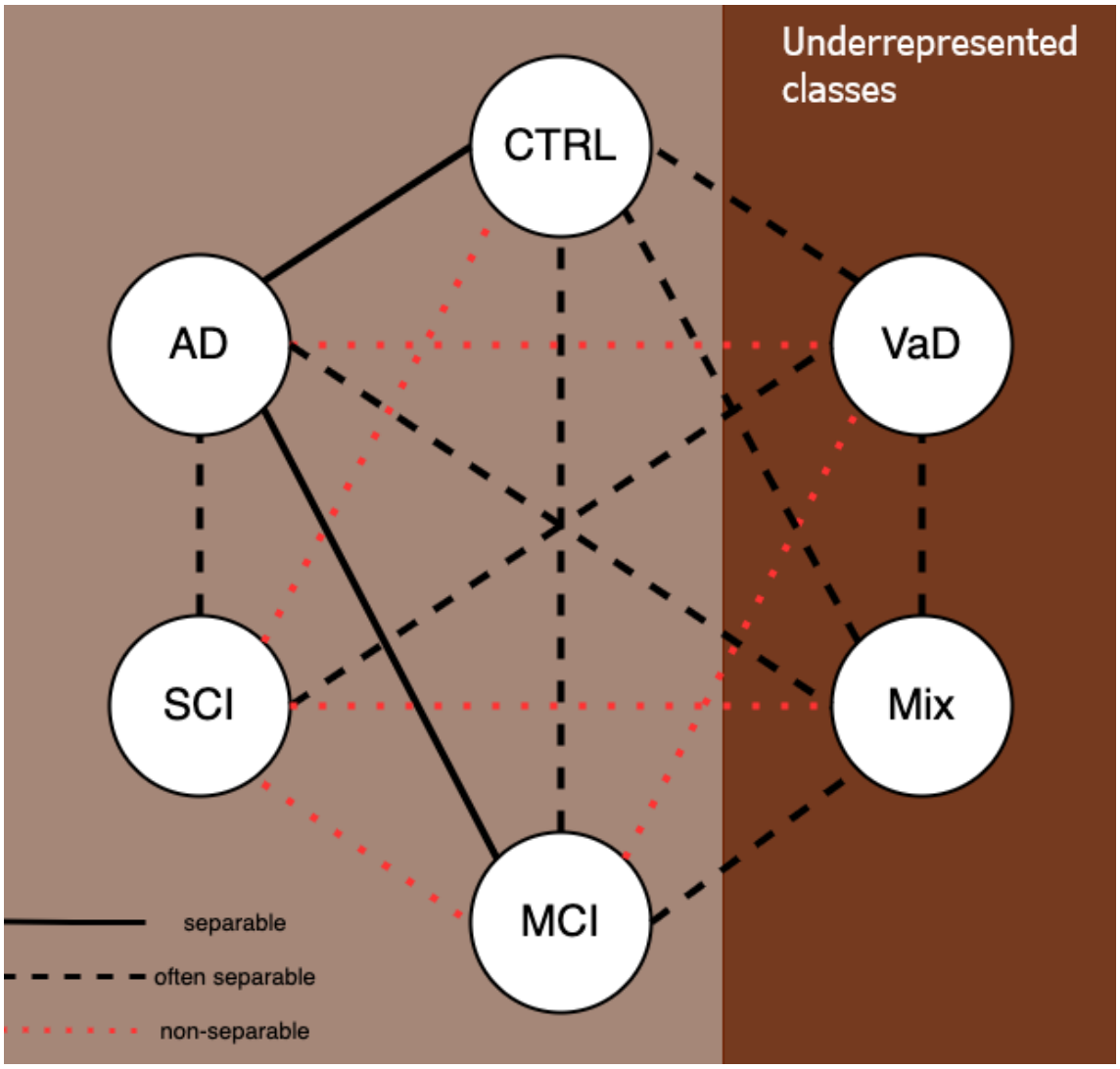



**Figure 5.4:** Figure showing separability of classes in T1 images

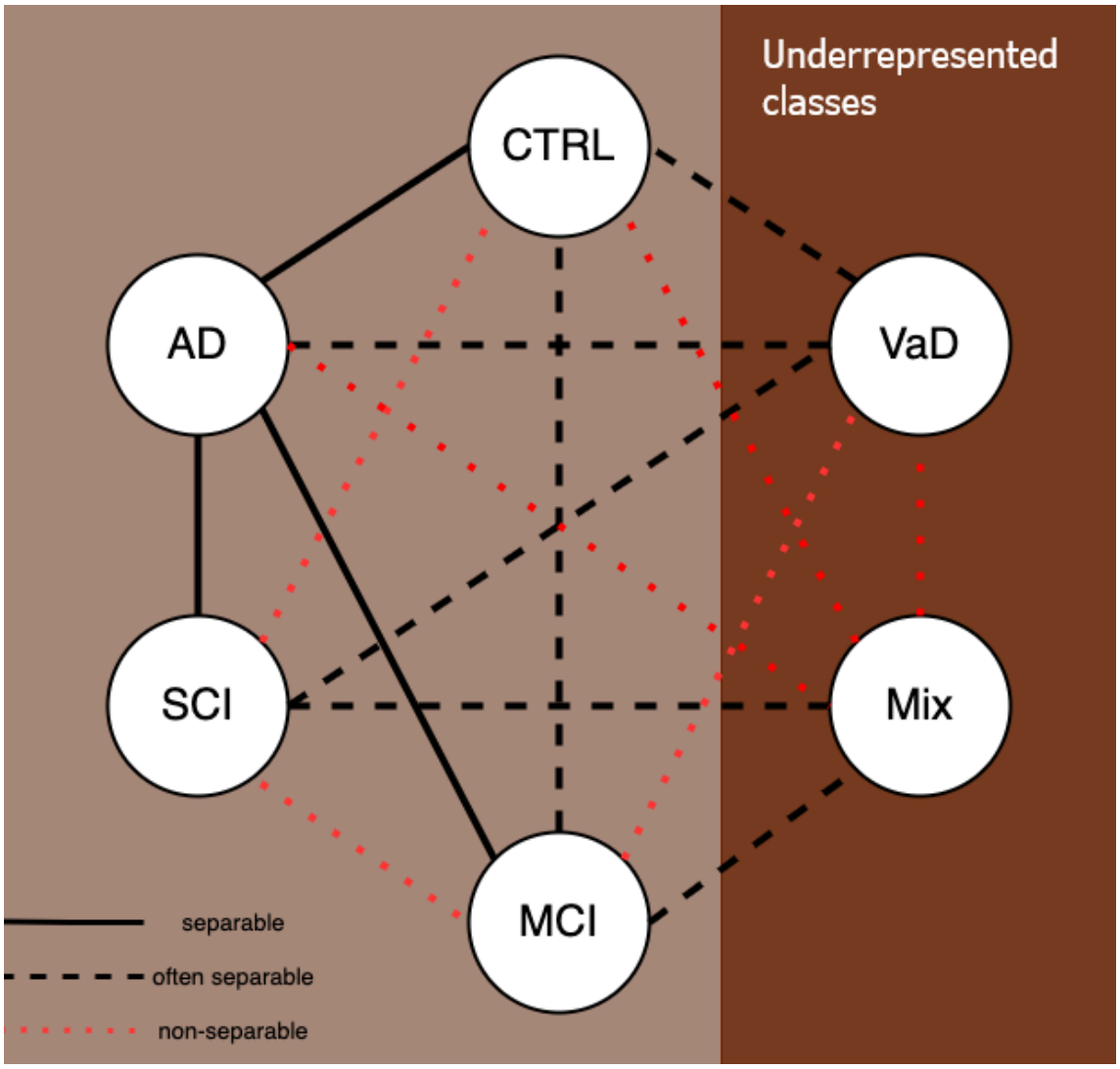



**Figure 5.5:** Figure showing separability of classes in T2 images

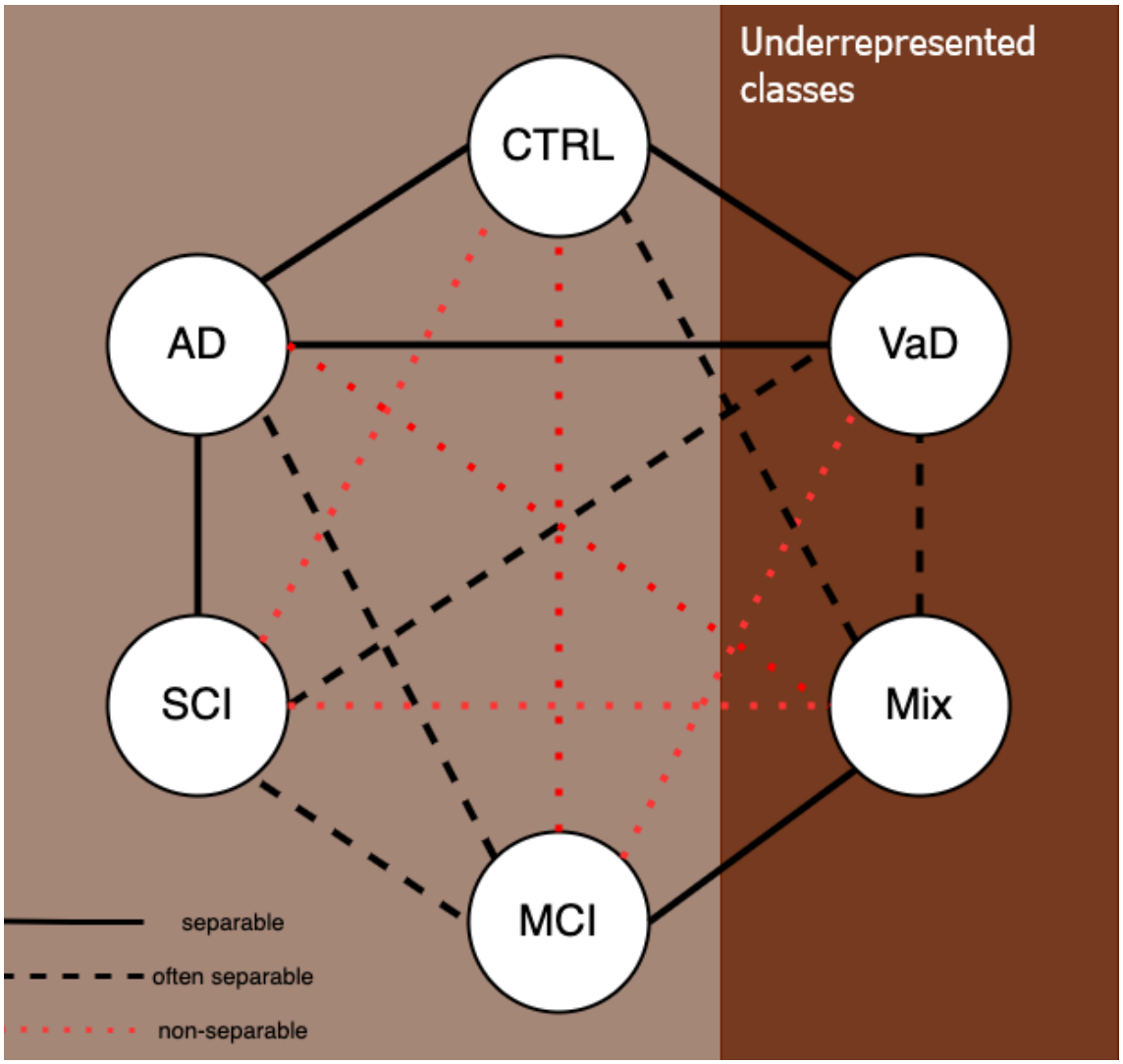



**Figure 5.6:** Figure showing separability of classes in FLAIR images

As mentioned in 5.2, the model is unable to tell the difference from SCI, MCI and CTRL classes but mostly able to separate them from AD. Then, although there is not enough VaD and MIX data, the model also shows the better performance from FLAIR when it tries to predict VaD or Mix from the other 4 classes.

## 5.4 Possible improvements

There are a number of ways that we would like to improve on our results and experiments that we would like to run but due to time and/or resource limitations was not able to during this project. The improvements mentioned below is mainly in order to improve the multi-class model, if not otherwise mentioned.

### 5.4.1 Further diagnosis

As mentioned in 5.2, if the MCI class can be further diagnosed and divided into sub-groups, it would be able to perform more tests on the label and possibly separate the MCI from CTRL or SCI. Second, instead of 3D images, if it is possible to have slices that demonstrates features from the 5 diseases, it would be possible to create a 2D classifier for the diagnosis. As it is observed that most sub-features for the target diseases including WHMs, lacune, cerebrovascular microbleed and perivascular space can be seen from 2D images, it should be able to train a machine learning model using key 2D slices instead of the whole 3D volumes. Then, by training the model with these 2D slices, it may be possible for a model to provide with a probability of the diseases by iterating the incoming 3D volumes. This vector can be used on it own or concatenated into the feature vector of the 3D prediction model for a better classification result.

### 5.4.2 Transfer learning

Transfer learning is as described in section 2.5 a very useful concept in order to not start over from scratch. In this project there was attempts at utilizing transfer learning, however, no specific pre-trained model on gray-scale 3D medical brain MRI data was found. The best approximation to the task was to utilize a ResNet pre-trained on medical imaging data, but for another task, namely medical imaging segmentation. When initializing with these weights however, it was found that the model did not perform better or worse than initializing from start. This seems to indicate that medical imaging segmentation does not carry over to diagnosis of brain MRI for cognitive decline.

There are however other ways one might utilize pre-training, especially since the dataset from the Gothenburg MCI cohort is relatively small. One could utilize an open source dataset such as ADNI [9], in order to pre-train a model on medical imaging data which is closely related to the task. This way the model might be able to pick up on important features at least of what constitutes a brain structure. Since the open-source datasets usually only have Control subjects, AD and in some cases MCI patients it is not presumed to give a large improvements to the task at hand but might nonetheless give some improvement or at least a better starting point when training.

### 5.4.3 Time and computational resources

A large bottleneck for this project has been that the time for running a training episode was long, a training of 100 epochs had a runtime of 18hrs. Additionally computational resources such as GPU memory was also restraining. In order to run the model on the data the batch size of each epoch had to be set to one, meaning that every data sample was processed one at a time. When only using one sample at a time the GPU memory was just enough to handle the computations but any larger batch size was not possible. Each processed sample consists of 70 T1 images, 20 T2 and 20 FLAIR, a total of 110 images per data sample. In our final ResNet

ensamble model we there was a total of around 190M parameters to be trained. So combining the amount of parameters needed to be trained and limiting the processing to one sample at a time naturally bring very long processing times. The GPU RAM available during this project was 12GB working memory on a NVIDIA Tesla K80. If the computational resources were higher then more experiments and faster progress could have been done.

### 5.4.4 Additional data collection

If it would be possible to collect more data then the obtained results would become more robust and the assumption is also that the accuracy and F1 score would increase. This is particularly true for disease type VaD and Mix as these have very few samples to begin with. With more data collected it is assumed that the model would be able to distinguish out even more what features are really relevant for each class. Thus the accuracy scores and weighted mean F1 scores would increase for each class.

For this project however, targeting Mix and VaD would yield the best marginal value since therse is only 56 and 26 samples of these classes respectively. This means that during training the model only has 34 and 16 samples to train on and during validation and testing only 22 (11 for validation and 11 for testing) and 10(5 for validation and 5 for testing) samples to evaluate on. The assumption is that this leads to a bad performance for two main reasons, the model does not have sufficient amount of samples to pick up features that will generalize to all cases of Mix and VaD. It is highly unlikely that within these small number of samples the model have enough information about the disease types in order to generalize to any instance of these disease. To begin with, the medical fields knowledge of the disease types are not fully mapped out such that all the relevant features of brain MRI scans are known. Secondly the small amount of validation and testing samples leads to big variations within the evaluation metrics. For instance, during training it could happen that the model classifies four VaD samples correctly and with only 5 samples available would then output a validation accuracy of 80%. This result seems very good but with such small amounts of validation samples it is impossible to say if the result is robust or not. It can also be seen from the test runs that for the Mix and VaD, the testing accuracies and weighted mean F1-score vary a lot from the validation, suggesting that the model indeed has not learned to correctly identify these classes, but rather due to the low amount of validation data shows good results.

The datapoints however are hard to get as it requires a patient to re-visit the Gothenburg MCI studies clinic, meet with a doctor and capture new brain MRI sequences. This is a time intensive and costly procedure that would require months in order to acquire new data with. Therefore it was not possible during this project to acquire new data related to these disease types but would be of interest in future studies on the cohort data.

# 6
# Conclusion

Previous work within the same field but with fewer target labels had shown great promise for a multi-class model to work. However the multi-class model was not able to perform significantly above baseline. Further research would be needed in order to see if a multi-class model would be able to distinguish between all six target classes. One main area of focus for future work would be to ensure that there are more samples across the underrepresented classes in the dataset. As of now there is just not a sufficient amount of data in order to train and perform robust tests on.

As mentioned in section 5.2, the F1 score offered from the pair tests, especially AD vs CTRL, validates the correctness of the dataset in certain extents as it matches the underlying expectation of the diseases classification. Thus, it is concluded that machine learning approaches can be used on the Gothenburg MCI study cohort to further their research. Then, it also reveals that even without sufficient amount of VaD and Mix labels, the model can still in certain level, demonstrates the separability between these 2 labels and the other 4 labels. It also indicates that given more VaD and Mix samples, it should be able to create a more robust prediction model for the diseases. Apart from that, it can be concluded that more research is needed to understand and possibly resolve the feature sharing, in particular between SCI, MCI and CTRL.

# A

# Appendix

| Details | | | Validation | | | | Test | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| A | B | Scan Type | Acc | F1 score A | F1 score B | Mean F1 | Acc | F1 score A | F1 score B | Mean F1 |
| AD | MCI | FLAIR | 0.74 | 0.77 | 0.70 | 0.73 | 0.42 | 0.31 | 0.50 | 0.44 |
| AD | MCI | T1 | 0.64 | 0.55 | 0.70 | 0.63 | 0.68 | 0.45 | 0.78 | 0.67 |
| AD | MCI | T2 | 0.69 | 0.40 | 0.79 | 0.65 | 0.71 | 0.48 | 0.80 | 0.69 |
| AD | MIX | FLAIR | 0.57 | 0.50 | 0.63 | 0.56 | 0.35 | 0.24 | 0.43 | 0.31 |
| AD | MIX | T1 | 0.68 | 0.61 | 0.73 | 0.67 | 0.55 | 0.64 | 0.40 | 0.56 |
| AD | MIX | T2 | 0.57 | 0.67 | 0.40 | 0.53 | 0.50 | 0.62 | 0.29 | 0.50 |
| AD | VaD | FLAIR | 0.71 | 0.71 | 0.71 | 0.71 | 0.81 | 0.89 | 0.40 | 0.79 |
| AD | VaD | T1 | 0.50 | 0.42 | 0.56 | 0.49 | 0.31 | 0.42 | 0.15 | 0.37 |
| AD | VaD | T2 | 0.64 | 0.44 | 0.74 | 0.59 | 0.25 | 0.25 | 0.25 | 0.25 |
| CTRL | AD | FLAIR | 0.71 | 0.74 | 0.67 | 0.70 | 0.80 | 0.83 | 0.75 | 0.80 |
| CTRL | AD | T1 | 0.71 | 0.72 | 0.69 | 0.70 | 0.67 | 0.72 | 0.58 | 0.66 |
| CTRL | AD | T2 | 0.68 | 0.75 | 0.55 | 0.66 | 0.63 | 0.62 | 0.65 | 0.63 |
| CTRL | MCI | FLAIR | 0.58 | 0.27 | 0.70 | 0.49 | 0.55 | 0.00 | 0.71 | 0.42 |
| CTRL | MCI | T1 | 0.64 | 0.55 | 0.71 | 0.64 | 0.57 | 0.00 | 0.73 | 0.43 |
| CTRL | MCI | T2 | 0.62 | 0.65 | 0.58 | 0.62 | 0.62 | 0.58 | 0.65 | 0.62 |
| CTRL | MIX | FLAIR | 0.74 | 0.76 | 0.71 | 0.73 | 0.46 | 0.58 | 0.24 | 0.48 |
| CTRL | MIX | T1 | 0.68 | 0.52 | 0.76 | 0.64 | 0.42 | 0.42 | 0.42 | 0.42 |
| CTRL | MIX | T2 | 0.47 | 0.53 | 0.40 | 0.46 | 0.50 | 0.63 | 0.25 | 0.52 |
| CTRL | VaD | FLAIR | 0.82 | 0.79 | 0.85 | 0.82 | 0.65 | 0.75 | 0.46 | 0.70 |
| CTRL | VaD | T1 | 0.76 | 0.79 | 0.73 | 0.76 | 0.75 | 0.85 | 0.29 | 0.76 |
| CTRL | VaD | T2 | 0.62 | 0.61 | 0.63 | 0.62 | 0.50 | 0.67 | 0.00 | 0.57 |
| MCI | MIX | FLAIR | 0.68 | 0.70 | 0.65 | 0.68 | 0.69 | 0.79 | 0.38 | 0.70 |
| MCI | MIX | T1 | 0.68 | 0.53 | 0.76 | 0.64 | 0.53 | 0.57 | 0.48 | 0.55 |
| MCI | MIX | T2 | 0.70 | 0.69 | 0.71 | 0.70 | 0.53 | 0.65 | 0.29 | 0.57 |
| MCI | VaD | FLAIR | 0.38 | 0.42 | 0.34 | 0.38 | 0.36 | 0.47 | 0.18 | 0.44 |
| MCI | VaD | T1 | 0.50 | 0.67 | 0.00 | 0.33 | 0.94 | 0.00 | 0.89 | 0.84 |
| MCI | VaD | T2 | 0.52 | 0.43 | 0.59 | 0.51 | 0.39 | 0.51 | 0.19 | 0.48 |
| SCI | AD | FLAIR | 0.64 | 0.64 | 0.64 | 0.64 | 0.66 | 0.72 | 0.56 | 0.67 |
| SCI | AD | T1 | 0.64 | 0.55 | 0.71 | 0.63 | 0.56 | 0.57 | 0.55 | 0.56 |
| SCI | AD | T2 | 0.74 | 0.81 | 0.59 | 0.74 | 0.73 | 0.80 | 0.59 | 0.73 |

| Details | | | Validation | | | | Test | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| A | B | Scan Type | Acc | F1 score A | F1 score B | mean F1 | Acc | F1 score A | F1 score B | mean F1 |
| SCI | CTRL | FLAIR | 0.57 | 0.59 | 0.56 | 0.57 | 0.40 | 0.47 | 0.31 | 0.41 |
| SCI | CTRL | T1 | 0.55 | 0.62 | 0.47 | 0.54 | 0.53 | 0.67 | 0.22 | 0.53 |
| SCI | CTRL | T2 | 0.57 | 0.59 | 0.56 | 0.57 | 0.40 | 0.47 | 0.31 | 0.41 |
| SCI | MCI | FLAIR | 0.59 | 0.68 | 0.44 | 0.56 | 0.66 | 0.74 | 0.53 | 0.64 |
| SCI | MCI | T1 | 0.57 | 0.57 | 0.57 | 0.57 | 0.42 | 0.49 | 0.31 | 0.41 |
| SCI | MCI | T2 | 0.55 | 0.50 | 0.59 | 0.54 | 0.55 | 0.56 | 0.54 | 0.55 |
| SCI | MIX | FLAIR | 0.57 | 0.70 | 0.25 | 0.48 | 0.80 | 0.89 | 0.00 | 0.71 |
| SCI | MIX | T1 | 0.52 | 0.49 | 0.54 | 0.52 | 0.51 | 0.60 | 0.37 | 0.51 |
| SCI | MIX | T2 | 0.82 | 0.84 | 0.80 | 0.82 | 0.66 | 0.78 | 0.25 | 0.67 |
| SCI | VaD | FLAIR | 0.64 | 0.72 | 0.50 | 0.61 | 0.87 | 0.93 | 0.00 | 0.84 |
| SCI | VaD | T1 | 0.94 | 0.97 | 0.50 | 0.92 | 0.87 | 0.93 | 0.00 | 0.84 |
| SCI | VaD | T2 | 0.59 | 0.68 | 0.44 | 0.56 | 0.81 | 0.89 | 0.25 | 0.83 |
| VaD | MIX | FLAIR | 0.71 | 0.67 | 0.75 | 0.71 | 0.50 | 0.29 | 0.62 | 0.52 |
| VaD | MIX | T1 | 0.79 | 0.77 | 0.80 | 0.78 | 0.50 | 0.00 | 0.67 | 0.47 |
| VaD | MIX | T2 | 0.50 | 0.36 | 0.59 | 0.48 | 0.40 | 0.00 | 0.57 | 0.40 |

**Table A.2:** Table showing the full binary test result with the validation dataset and the testing dataset. Each row presents a test pair using class from A against the class from B column. Acc denotes accuracy.

**DEPARTMENT OF SOME SUBJECT OR TECHNOLOGY**
**CHALMERS UNIVERSITY OF TECHNOLOGY**
Gothenburg, Sweden
www.chalmers.se


CHALMERS
UNIVERSITY OF TECHNOLOGY

# Bibliography


[1] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. “Wider or deeper: Revisiting the resnet model for visual recognition”. In: *Pattern Recognition* 90 (2019), pp. 119–133.

[2] Christian Szegedy et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[3] Tobias Lindig et al. “Evaluation of multimodal segmentation based on 3D T1-, T2-and FLAIR-weighted images–the difficulty of choosing”. In: *Neuroimage* 170 (2018), pp. 210–221.

[4] Anders Wallin et al. “The Gothenburg MCI study: design and distribution of Alzheimer’s disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up”. In: *Journal of Cerebral Blood Flow & Metabolism* 36.1 (2016), pp. 114–131.

[5] Mingyu Kim et al. “Deep learning in medical imaging”. In: *Neurospine* 16.4 (2019), p. 657.

[6] Ahsan Bin Tufail, Yong-Kui Ma, and Qiu-Na Zhang. “Binary classification of Alzheimer’s disease using sMRI imaging modality and deep learning”. In: *Journal of digital imaging* 33.5 (2020), pp. 1073–1090.

[7] Marcia Hon and Naimul Mefraz Khan. “Towards Alzheimer’s disease classification through transfer learning”. In: *2017 IEEE International conference on bioinformatics and biomedicine (BIBM)*. IEEE. 2017, pp. 1166–1169.

[8] Saman Sarraf and Ghassem Tofighi. “Classification of alzheimer’s disease structural MRI data by deep learning convolutional neural networks”. In: *arXiv preprint arXiv:1607.06583* (2016).

[9] http://adni.loni.usc.edu/. *Alzheimer’s disease neuroimaging initiative*.

[10] Shui-Hua Wang et al. “Classification of Alzheimer’s disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling”. In: *Journal of medical systems* 42.5 (2018), pp. 1–11.

[11] Bijen Khagi et al. “CNN Models Performance Analysis on MRI images of OASIS dataset for distinction between Healthy and Alzheimer’s patient”. In: *2019 International Conference on Electronics, Information, and Communication (ICEIC)*. IEEE. 2019, pp. 1–4.

[12] Sergey Korolev et al. “Residual and plain convolutional neural networks for 3D brain MRI classification”. In: *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*. IEEE. 2017, pp. 835–838.

[13] Yan Wang et al. “A novel multimodal MRI analysis for Alzheimer’s disease based on convolutional neural network”. In: *2018 40th Annual Interna-*

*tional Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2018, pp. 754–757.

[14] Ammarah Farooq et al. “A deep CNN based multi-class classification of Alzheimer’s disease using MRI”. In: *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*. 2017, pp. 1–6. DOI: 10.1109/IST.2017.8261460.

[15] https://www.oasis-brains.org/. *Open Access Series of Imaging Studies*.

[16] Gary M Weiss, Kate McCarthy, and Bibi Zabar. “Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?” In: *Dmin* 7.35-41 (2007), p. 24.

[17] Ricardo Barandela et al. “The imbalanced training sample problem: Under or over sampling?” In: *Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR)*. Springer. 2004, pp. 806–814.

[18] Sebastien C Wong et al. “Understanding data augmentation for classification: when to warp?” In: *2016 international conference on digital image computing: techniques and applications (DICTA)*. IEEE. 2016, pp. 1–6.

[19] Rudy J Castellani, Raj K Rolston, and Mark A Smith. “Alzheimer disease”. In: *Disease-a-month: DM* 56.9 (2010), p. 484.

[20] R Nick Bryan. *Machine learning applied to Alzheimer disease*. 2016.

[21] Joseph R Cockrell and Marshal F Folstein. “Mini-mental state examination”. In: *Principles and practice of geriatric psychiatry* (2002), pp. 140–141.

[22] Florence Portet et al. “Mild cognitive impairment (MCI) in medical practice: a critical review of the concept and new diagnostic procedure. Report of the MCI Working Group of the European Consortium on Alzheimer’s Disease”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 77.6 (2006), pp. 714–718.

[23] Petronella Kettunen et al. “Blood-brain barrier dysfunction and reduced cerebrospinal fluid levels of soluble amyloid precursor protein-$\beta$ in patients with subcortical small-vessel disease”. In: *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring* 14.1 (2022), e12296.

[24] Silvia Koton et al. “Microvascular brain disease progression and risk of stroke: the ARIC study”. In: *Stroke* 51.11 (2020), pp. 3264–3270.

[25] Aurauma Chutinet and Natalia S Rost. “White matter disease as a biomarker for long-term cerebrovascular disease and dementia”. In: *Current treatment options in cardiovascular medicine* 16.3 (2014), pp. 1–12.

[26] Yulu Shi and Joanna M Wardlaw. “Update on cerebral small vessel disease: a dynamic whole-brain disease”. In: *Stroke and vascular neurology* 1.3 (2016).

[27] Barry Reisberg et al. “The Global Deterioration Scale for assessment of primary degenerative dementia.” In: *The American journal of psychiatry* (1982).

[28] Donald R Royall, Roderick K Mahurin, and Kevin F Gray. “Bedside assessment of executive cognitive impairment: the executive interview”. In: *Journal of the American Geriatrics Society* 40.12 (1992), pp. 1221–1226.

[29] Guy McKhann et al. “Clinical diagnosis of Alzheimer’s disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer’s Disease”. In: *Neurology* 34.7 (1984), pp. 939–939.

[30] Bruno Dubois et al. “Advancing research diagnostic criteria for Alzheimer’s disease: the IWG-2 criteria”. In: *The Lancet Neurology* 13.6 (2014), pp. 614–629.

[31] Timo Erkinjuntti et al. “Research criteria for subcortical vascular dementia in clinical trials”. In: *Advances in dementia research* (2000), pp. 23–30.

[32] Lars-Olof Wahlund et al. “A new rating scale for age-related white matter changes applicable to MRI and CT”. In: *Stroke* 32.6 (2001), pp. 1318–1322.

[33] Olivia Anna Skrobot et al. “A validation study of vascular cognitive impairment genetics meta-analysis findings in an independent collaborative cohort”. In: *Journal of Alzheimer’s Disease* 53.3 (2016), pp. 981–989.

[34] Anders Wallin et al. “Alzheimer’s disease—Subcortical vascular disease spectrum in a hospital-based setting: Overview of results from the Gothenburg MCI and dementia studies”. In: *Journal of Cerebral Blood Flow & Metabolism* 36.1 (2016), pp. 95–113.

[35] Samir S Yadav and Shivajirao M Jadhav. “Deep convolutional neural network based medical image classification for disease diagnosis”. In: *Journal of Big Data* 6.1 (2019), pp. 1–18.

[36] Waseem Rawat and Zenghui Wang. “Deep convolutional neural networks for image classification: A comprehensive review”. In: *Neural computation* 29.9 (2017), pp. 2352–2449.

[37] Olga Russakovsky et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115 (2015), pp. 211–252.

[38] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[39] Satya P Singh et al. “3D deep learning on medical images: a review”. In: *Sensors* 20.18 (2020), p. 5097.

[40] 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)*. 2020, pp. 1–6. DOI: 10.1109/IVCNZ51579.2020.9290616.

[41] Xin Xing et al. “Dynamic image for 3d mri image alzheimer’s disease classification”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 355–364.

[42] Hua Zhang et al. “Deep Learning Model for the Automated Detection and Histopathological Prediction of Meningioma”. In: *Neuroinformatics* 19 (July 2021). DOI: 10.1007/s12021-020-09492-6.

[43] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. “What makes ImageNet good for transfer learning?” In: *arXiv preprint arXiv:1608.08614* (2016).

[44] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. “A survey of transfer learning”. In: *Journal of Big data* 3.1 (2016), pp. 1–40.

[45] Ahmad Waleed Salehi et al. “A CNN Model: Earlier Diagnosis and Classification of Alzheimer Disease using MRI”. In: *2020 International Conference on Smart Electronics and Communication (ICOSEC)*. 2020, pp. 156–161. DOI: 10.1109/ICOSEC49089.2020.9215402.

[46] Jiancheng Yang et al. “Reinventing 2d convolutions for 3d images”. In: *IEEE Journal of Biomedical and Health Informatics* 25.8 (2021), pp. 3009–3018.

[47] Xiangrui Li et al. “The first step for neuroimaging data analysis: DICOM to NIfTI conversion”. In: *Journal of neuroscience methods* 264 (2016), pp. 47–56.

[48] Mark W Woolrich et al. “Bayesian analysis of neuroimaging data in FSL”. In: *Neuroimage* 45.1 (2009), S173–S186.

[49] Mark Jenkinson, Mickael Pechaud, Stephen Smith, et al. “BET2: MR-based estimation of brain, skull and scalp surfaces”. In: *Eleventh annual meeting of the organization for human brain mapping*. Vol. 17. 3. Toronto. 2005, p. 167.

[50] *Docs*. URL: https://monai.io/docs.html.

[51] *Crossentropyloss¶*. URL: https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html.

[52] Nikhil Ketkar and Jojo Moolayil. “Introduction to pytorch”. In: *Deep learning with python*. Springer, 2021, pp. 27–91.

[53] Sihong Chen, Kai Ma, and Yefeng Zheng. “Med3D: Transfer Learning for 3D Medical Image Analysis”. In: *arXiv preprint arXiv:1904.00625* (2019).

[54] Lucy Norcliffe-Kaufmann, Felicia B Axelrod, and Joel Gutierrez. “ENCYCLOPEDIA OF NEUROLOGICAL SCIENCES 2ND EDITION”. In: ().

[55] Joanna M Wardlaw et al. “Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration”. In: *The Lancet Neurology* 12.8 (2013), pp. 822–838.