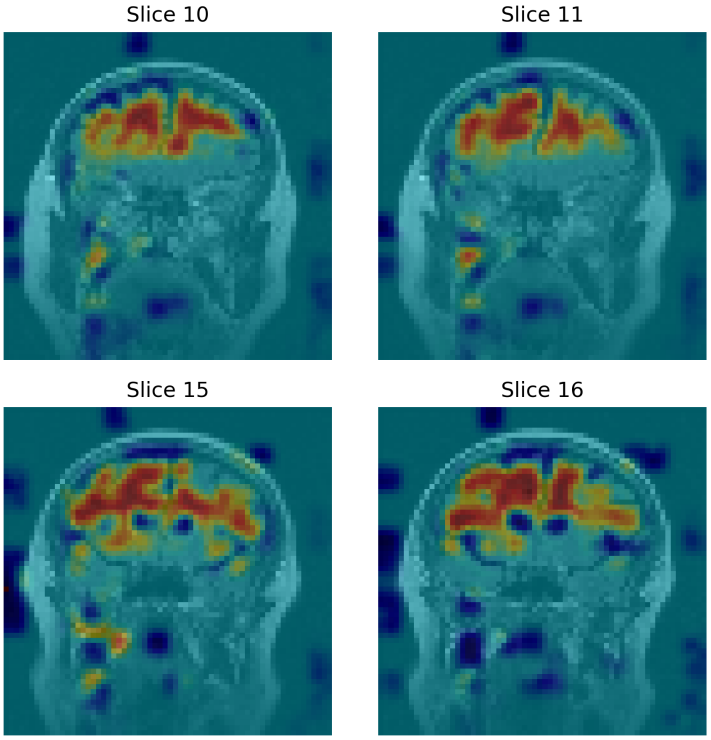


Machine learning-based detection of vascular pathologies in MRI images

A multi-task learning approach for identifying white matter hyperintensities and vascular cognitive impairment in FLAIR MRI images

Master's thesis in Biomedical Engineering

AMANDA HOVEKLINT
KAJSA HOMANN

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

Master's thesis 2025

Machine learning-based detection of vascular pathologies in MRI images

A multi-task learning approach for identifying white matter hyperintensities and vascular cognitive impairment in FLAIR MRI images

AMANDA HOVEKLINT
KAJSA HOMANN

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical engineering
Division of Signal Processing and Biomedical Engineering
Chalmers University of Technology
Gothenburg, Sweden 2025

Machine learning-based detection of vascular pathologies in MRI images
A multi-task learning approach for identifying white matter hyperintensities and vascular cognitive impairment in FLAIR MRI images
Amanda Hoveklint
Kajsa Homann

Supervisor: Petronella Kettunen, Institute of Neuroscience and Physiology, University of Gothenburg
Examiner: Jennifer Alvén, Department of Electrical Engineering, Chalmers University of Technology

Master's Thesis 2025
Department of Electrical Engineering
Division of Signal Processing and Biomedical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: The resulting Grad-CAM heatmap for a patient using three-class classification. The figure shows that the implemented model focuses more on regions within the brain and less on the skull. This patient had a fairly small amount of WMH in the brain and was diagnosed with SCI.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2025

Abstract

This thesis investigates the potential of using FLAIR imaging data and machine learning (ML) to identify patients with vascular cognitive disease (VCD). VCD is currently underdiagnosed, and diagnosis usually happens at the late stages of the disease. Currently, there are no treatments available for VCD, but if caught early, it is possible to stop the progression of the disease by, for example, lifestyle changes. Therefore, early detection of VCD is essential.

A total of 750 magnetic resonance imaging (MRI) scans from a cohort of 506 women and men between 50 and 79 years of age from the Gothenburg mild cognitive impairment (MCI) study were analyzed. The cohort included controls, preclinical participants (subjective cognitive impairment and MCI), and patients with Alzheimer's disease (AD), subcortical small-vessel disease (SSVD), and mixed AD/SSVD. FLAIR sequences from a 1.5 Tesla MRI scanner were used, together with previously analyzed white matter hyperintensity (WMH) volumes acquired from the FreeSurfer 5.3 software.

The input to the model consisted of FLAIR MRI volumes, while the ground truth data included clinical diagnoses and WMH volume measurements. A multi-task learning (MTL) model based on ResNet18 was implemented, combining regression of WMH volumes and classification of patient diagnoses. The explainability method Guided Grad-CAM was implemented to visualize the regions of the MRI images that contributed most to the model's predictions or classifications. The model was analyzed in a binary and two multi-classification cases. The regression head produced WMH volume estimates comparable to those obtained from FreeSurfer. The classification head achieved an F1-score of 0.5806 for binary classification, but did not yield satisfactory results using three-class classification and five-class classification, with F1-scores of 0.5381 and 0.4465, respectively. This thesis demonstrates that it is possible to quantify WMH volumes and identify patients with vascular cognitive impairment based on FLAIR images using ML.

Keywords: machine learning, vascular cognitive disease, FLAIR, white matter hyperintensities, multi-task learning, guided Grad-CAM.

Acknowledgements

First, we would like to express our gratitude to the Gothenburg mild cognitive impairment study research team at the Memory Clinic, Sahlgrenska University Hospital. In particular, we extend our thanks to Associate Professor Petronella Kettunen and Emir Basic, MSc in Biomedical Engineering, who have served as our primary supervisors during this thesis. We are grateful for the opportunity to work on this project and for the support we have received in understanding and learning the clinical aspects of the thesis. We also appreciate the guidance and assistance provided by both of you throughout the project.

We would also like to thank our examiner, Assistant Professor Jennifer Alvén, for serving as both our examiner and technical supervisor. We are grateful for the support provided throughout the project, especially in addressing technical challenges. We are also thankful for all the meetings where you proposed solutions to the challenges we encountered during the work, as well as for the support provided for the overall project.

At last, we would like to thank Johannes Johansson and Anton Myhrberg, who have read through our report and provided feedback. We would also like to thank Kristoffer Gustafsson and Saga Frisell for their valuable constructive feedback on this work.

Amanda Hoveklint and Kajsa Homann, Gothenburg, June 2025

List of Acronyms

AD	Alzheimer's disease
Adam	Adaptive moment estimation
AI	Artificial intelligence
AUC	Area under the curve
CNN	Convolutional neural network
CSF	Cerebrospinal fluid
CT	Computed tomography
DICOM	Digital imaging and communications in medicine
FLAIR	Fluid-attenuated inversion recovery
GDS	Global deterioration scale
Grad-CAM	Gradient weighted class activation mapping
HC	Healthy controls
MAE	Mean absolute error
MCI	Mild cognitive impairment
ML	Machine learning
MMSE	Mini-mental state examination
MONAI	Medical open network for AI
MRI	Magnetic resonance imaging
MSE	Mean square error
MTL	Multi-task learning
NIfTI	Neuroimaging informatics technology initiative
PET	Positron emission tomography
ReLU	Rectified linear unit
RVM	Relevance vector machine
SCI	Subjective cognitive impairment
SPECT	Single-photon emission computed tomography
SSVD	Sub-cortical small vessel disease
T1	T1-weighted MRI
T2	T2-weighted MRI
VCD	Vascular cognitive disease
VCI	Vascular cognitive impairment
WMH	White matter hyperintensities

Contents

List of Acronyms **ix**

1 Introduction **1**
1.1 Background . . . 1
1.1.1 The Gothenburg mild cognitive impairment study . . . 2
1.2 Purpose and research questions . . . 2
1.3 Previous work . . . 3
1.4 Scope and limitations . . . 4

2 Preliminaries **5**
2.1 Stages and diseases . . . 5
2.1.1 Control subjects . . . 6
2.1.2 Subjective cognitive impairment . . . 6
2.1.3 Mild cognitive impairment . . . 6
2.1.4 Alzheimer's disease . . . 6
2.1.5 Vascular cognitive disease . . . 6
2.1.5.1 Mixed SSVD/AD . . . 7
2.2 Diagnostic procedure . . . 7
2.3 Machine learning . . . 8
2.3.1 ResNet18 . . . 8
2.3.2 Loss functions . . . 8
2.3.2.1 Mean squared error . . . 9
2.3.2.2 Mean absolute error . . . 9
2.3.2.3 Huber loss . . . 9
2.3.2.4 Cross-entropy loss . . . 9
2.3.2.5 Focal loss . . . 10
2.3.3 Overfitting . . . 10
2.3.3.1 Data augmentation . . . 10
2.3.3.2 Batch normalization . . . 10
2.3.3.3 Dropout . . . 10
2.3.3.4 Reduce learning rate on plateau . . . 11
2.3.3.5 Weight decay . . . 11
2.3.3.6 Label smoothing . . . 11
2.3.3.7 Early stopping . . . 11
2.3.4 Multi-task learning . . . 11

2.3.5 Explainability model 11
2.4 Evaluation metrics 12
2.4.1 Pearson correlation coefficient 13
2.4.2 Spearman's rank correlation coefficient 13
2.4.3 R^2 13
2.4.4 Accuracy 13
2.4.5 Precision 14
2.4.6 Recall 14
2.4.7 F1-score 14
2.4.8 K-fold cross-validation 14

3 Methods **15**
3.1 Data acquisition and preprocessing 15
3.1.1 Data acquisition 15
3.1.2 NIfTI conversion 17
3.1.3 Normalization 18
3.2 Machine learning model 18
3.2.1 Regression model 18
3.2.2 Classification model 18
3.2.3 Multi-task learning 19
3.2.4 Evaluation metrics 19
3.2.5 Using T1 and FLAIR vs FLAIR alone as input 20
3.2.6 Adding covariates or not 20
3.2.7 Choosing loss function 20
3.2.8 Overfitting 21
3.2.8.1 Data augmentation 21
3.2.8.2 Dropout 22
3.2.8.3 Weight decay 22
3.2.8.4 Batch normalization 22
3.2.8.5 Label smoothing 22
3.2.8.6 Reduce learning rate on plateau 23
3.2.8.7 Early stopping 23
3.2.9 K-fold cross-validation 23
3.3 Explainability model 24
3.4 Use of artificial intelligence 25

4 Results **27**
4.1 Using T1 and FLAIR vs FLAIR alone as input 27
4.2 Choosing loss function 28
4.2.1 Regression head 28
4.2.2 Binary classification 29
4.2.3 Three-class classification 31
4.2.4 Five-class classification 31
4.3 Overfitting 33
4.3.1 Data augmentation 33
4.3.2 Early stopping 34
4.4 K-fold cross-validation 34

4.5 Multitask learning . . . 35
4.5.1 Binary classification . . . 35
4.5.2 Three-class classification . . . 36
4.5.3 Five-class classification . . . 37
4.6 Explainability model . . . 39

5 Discussion **45**
5.1 Using T1 and FLAIR vs FLAIR alone as input . . . 45
5.2 Choosing loss function . . . 45
5.3 Overfitting . . . 46
5.4 Multi-task learning . . . 46
5.4.1 Performance analysis . . . 46
5.4.2 The formation of classification categories . . . 47
5.5 Explainabilty model . . . 48
5.6 Future work . . . 50

6 Conclusion **51**

A Appendix 1 **I**
A.1 Data augmentation ablation study for focal loss . . . I

1

Introduction

This chapter presents the background to this master's thesis, information on the Gothenburg mild cognitive impairment (MCI) study, previous work, and scope and limitations.

1.1 Background

Vascular cognitive disease (VCD), also known as vascular dementia, is the most common type of cognitive disease after Alzheimer's disease (AD) [1]. It is caused by vascular pathologies, the most common type being subcortical small vessel disease (SSVD), i.e., damage to the small vessels in the brain, with common symptoms being cognitive, behavioral, and motor problems. Cognitive impairment can be classified into seven stages of deterioration using the Global Deterioration Scale (GDS) [2]. The patients in the study are all in stages 1 to 4, which are defined in section 2.1.

One predictor of developing VCD is having large white matter hyperintensities (WMH) [3], [4]. WMH are frequent in VCD but are also found in 30% of AD cases [1]. Although they are a sign of vascular pathology, WMH can also be found in asymptomatic patients [5]. Among people over 60 years of age, 10–20% have WMH, while almost 100% of those 90 years of age or older exhibit WMH.

WMH can be seen as white patches on fluid-attenuated inversion recovery (FLAIR) magnetic resonance imaging (MRI) [5]. They also appear white on T2-weighted MRI and dark on T1-weighted images, but are most clearly visualized on FLAIR images [6]. The most common imaging method when clinically assessing cognitive disease is computed tomography (CT) [7]. However, CT fails to capture vascular pathologies such as WMH as accurately as FLAIR MRI imaging [6]. VCD is today an underdiagnosed disease [8], and difficulties in properly assessing vascular pathology in the brain might be one of the reasons why. WMH is often caused by incomplete infarction, where blood flow to deep areas of the brain is chronically reduced [1]. SSVD is the most common cause of these infarctions. Reduced blood flow in the brain leads to hypoxia and breaks down the blood-brain barrier, which in turn causes damage such as demyelination and axonal loss, causing cognitive issues [1], [5]. Studies have shown that it is possible for the WMH to both increase and decrease in size, as well as disappear altogether [5]. This suggests that the process may be

reversible before it causes demyelination and axonal loss. Therefore, identifying patients in the process of developing vascular pathologies in the brain might allow for the delay of the development of SSVD.

1.1.1 The Gothenburg mild cognitive impairment study

The Gothenburg MCI study is an ongoing longitudinal cohort study at the memory clinic at Sahlgrenska University Hospital in Mölndal, Gothenburg [9]. It started in 1999 and focuses on investigating the early phases of AD and VCD, when cognitive problems are emerging. The study includes patients with different stages of cognitive impairment, ranging from subjective and mild cognitive impairment (SCI and MCI) to mild cognitive disease (AD and VCD). Cognitively normal individuals are included as healthy controls (HC).

The cohort consists of approximately 1,000 patients between the ages of 50 and 79 who sought medical help for cognitive problems [9]. A baseline examination was conducted, followed by follow-up assessments in years two, four, six, and ten after the baseline. Data collected for each patient included neuropsychological tests, neuroimaging, and biomarker data. Follow-up visits allow researchers to track the progression to cognitive disease (AD or VCD) from the patient's initial state.

Guidelines for inclusion in the cohort were age between 50 and 79, mini-mental state examination (MMSE) score > 18 out of 30, and self- or informant-reported cognitive decline lasting at least six months [9]. Another criterion for participating in the study was that there were no obvious underlying causes of memory problems other than cognitive disease, such as brain tumours, subdural hematoma, or major stroke. A criterion for exclusion in the study was if the patient had a systemic or other somatic disease that may have caused or was the cause of cognitive impairment.

Patients included in the healthy control group had inclusion criteria similar to those of other participants [9]. However, they had to have an MMSE score greater than 26 and no current or past cognitive decline. All other inclusion and exclusion criteria remained the same as those for the other patients.

1.2 Purpose and research questions

The purpose of this master's thesis is to investigate the use of machine learning (ML) as a more accurate way to diagnose a patient with VCD compared to the clinical method used today. The project aims to implement regression and classification ML methods to predict the amount of WMH and distinguish between types of cognitive disease from FLAIR MRI images. This is necessary since VCD is difficult to diagnose, and the progression of the disease could be delayed if it is at GDS 4, and even prevented if caught at GDS 2 or 3.

Our specific research questions are:

- Is it possible to identify patients with VCD based on FLAIR images using ML

classification models?

- Is it possible to differentiate between different cognitive diseases based on FLAIR images using ML classification models?
- Can ML regression models be used to identify WMH volumes on brain FLAIR images?

1.3 Previous work

Many studies have been conducted on using ML in the medical field [10]. For this thesis, the review of previous work has been narrowed down to using ML methods to detect WMH and/or diagnose any type of cognitive disease.

In an article published by Ebrahim et al., ResNet-18 was used to detect AD using MRI images and applied transfer learning in both 2D and 3D model architectures [11]. Transfer learning was applied as a way of preventing overfitting when training the model, since the MRI dataset only had hundreds of images. The model was pre-trained on the ImageNet dataset. They achieved an accuracy of 96.88% and a sensitivity of 100% after applying an optimization method during model training.

Feng et al. presented a method that employed a logistic regression model for detecting cognitive impairment [12]. The study aimed to identify cognitive impairment in patients with WMH using MRI, and to analyze the relationship between cognitive decline and associated factors. They performed image registration to align the T1 with the FLAIR images. This was done to perform segmentation of WMH on the FLAIR images using their VB-net model. They also evaluated the ML models, Gaussian process, random forest, and quadratic discriminant analysis algorithm. Out of these four models, logistic regression based on WMH features performed the best, achieving an area under the curve (AUC) score of 0.819 on a test set.

Studies that use ML to detect WMH mainly use segmentation. Ghafoorian et al. implemented a convolutional neural network (CNN) to segment WMH in T1 and FLAIR images [13]. They used image registration to align the images and could then use location features to improve the accuracy of the model. Their model achieved a Dice score of 0.792, compared to 0.805 achieved by a human expert. Another study achieved a similar Dice score of 0.78 when segmenting WMH using a 2D UNet-based deep learning method [14]. Neither of the studies used the WMH segmentations to diagnose disease.

We have found few studies utilizing ML for diagnosing VCD. A PubMed search using the terms *machine learning*, *detection*, and *vascular dementia* yielded 10 results, of which two were identified as relevant to our work. In those studies, the ML model diagnosed patients with VCD using data from biomarkers [15], and neuropsychological and electrophysiological measurements [16].

Most studies on ML and cognitive disease are done on AD, as it is the most common

type of cognitive disease, and there are known fluid biomarkers for this disease [17]. The detection of WMH using ML is mostly done using segmentation methods. We have found studies that bear similarities to our thesis work. However, our work is distinguished from previous studies by combining WMH volume identification and VCD diagnosis using FLAIR MRI images and ML methods. We present a multi-task learning (MTL) model combining regression methods for WMH volume identification and classification for diagnosing VCD.

1.4 Scope and limitations

This project aims to implement an ML model that can, based on FLAIR MRI images, distinguish VCD from healthy individuals and patients with other cognitive diseases, such as AD. The focus on VCD rather than AD is what differentiates this work from previous studies.

This project was divided into two parts: (1) an MTL model to predict both WMH volumes and patient diagnosis based on FLAIR MRI images and ground truth WMH volumes and diagnosis; (2) an explainability model to show the regions the model focuses on when predicting WMH volume and classifying VCD.

One of the limitations of this project was the 20-week time frame of the master's thesis, which limited the overall scope and depth of the work. Another limitation was our knowledge within the medical field. With our background, the knowledge of ML was enough to apply different methods to the project and achieve an acceptable outcome. However, this project also required knowledge of the brain, various types of cognitive disease, and how they are reflected in MRI images. That led to time spent learning about these topics, and some knowledge had to be overlooked to complete the project on time. A deeper knowledge of the data and the medical conditions addressed in this project might have sped up processes such as data preprocessing, allowing a bigger focus on the ML models. An additional challenge was the amount of data. Reality will always be a limiting factor when working with medical data. It is hard to acquire new data since a patient with the specific disease is needed, and it is costly to perform an MRI examination. Strict patient confidentiality laws slow down the process of getting new data. Patients have to consent to their images being used in research, and ethical approval from the Ethical Review Authority is needed to collect and use the data. For this thesis work, the amount of data was enough to achieve an acceptable result, but more data would facilitate the work.

2

Preliminaries

The following chapter presents relevant background theory, including an overview of the different cognitive disease types, current diagnostic procedures, and the methods used in this work.

2.1 Stages and diseases

Cognitive impairment can be classified into seven stages of disease according to the GDS. In the Gothenburg MCI study, stages 1 to 4 are represented in the cohort [9]. They are defined as:

- GDS 1: No cognitive impairment
- GDS 2: SCI
- GDS 3: MCI
- GDS 4: Mild cognitive disease

GDS 1 is referred to as HC. GDS 4 is categorized into the type of cognitive disease the patient has, either AD or VCD. The different diseases, HC, SCI, MCI, AD, and VCD (including pure SSVD and mixed SSVD/AD), see Figure 2.1, are presented in more detail below.

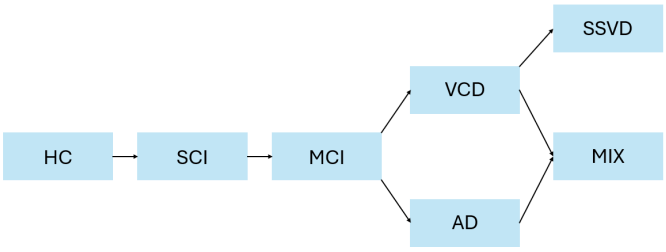


Figure 2.1: The different diseases and stages used in this thesis. HC = healthy control, SCI = subjective cognitive impairment, MCI = mild cognitive impairment, AD = Alzheimer's disease, VCD = vascular cognitive disease, SSVD = subcortical small vessels disease, MIX = mixed SSVD/AD.

2.1.1 Control subjects

The control subjects, HC, were recruited through senior citizen organizations and had to be cognitively healthy and score over 26 points on the MMSE test [9]. They were all in the same age range as the other patients.

2.1.2 Subjective cognitive impairment

SCI (GDS 2) is a less severe stage of cognitive disease than MCI, as the memory loss or cognitive symptoms experienced by the patient are not as severe as in MCI [9]. In follow-up visits, some SCI patients in the Gothenburg MCI study regressed to HC, or progressed to MCI or cognitive disease.

2.1.3 Mild cognitive impairment

MCI (GDS 3) is a pre-stage of cognitive disease. The patient experiences some measurable memory loss and cognitive symptoms, but it has not yet affected daily life [9]. In the Gothenburg MCI study, some patients regressed to HC after having previously been classified at GDS 3, while some progressed to GDS 4. Some patients stabilized and remained MCI. This showed that both progressive and regressive movements exist in the pre-stages of cognitive disease.

2.1.4 Alzheimer's disease

With 32 million patients worldwide [18], AD is the most common cause of cognitive disease, accounting for 50-75% cases [17]. The most common symptom of AD is memory issues, and it is also the first symptom. Other symptoms, such as difficulties multi-tasking and behavioral changes, appear as the disease progresses. AD is caused by β-amyloid plaques and neurofibrillary tangles in the brain. This, in turn, causes neurodegeneration and brain atrophy. β-amyloid in cerebrospinal fluid (CSF) is used as a fluid biomarker for AD. There exists a disease-modifying anti-body treatment available for AD that reduces the β-amyloid plaques in the brain and therefore reduces the disease progression [19]. Since November 2024, this treatment called Lecanemab/Leqembi is approved by the European Medicines Agency [20]. Patients will have regular MRI scans to monitor side effects such as oedema and haemorrhage of the brain. Unfortunately, AD patients who also have significant WMH or other vascular pathologies are not eligible for this treatment, since cerebrovascular disease is thought to increase the risk of previously mentioned side effects [21].

2.1.5 Vascular cognitive disease

VCD is the most common type of cognitive disease after AD, accounting for 15-20% of cases [22]. The prevalence of VCD is higher than diagnosis rates reported by memory clinics, indicating that it is underdiagnosed [8]. VCD is caused by vascular pathologies, the most common form being SSVD, which is damage to small vessels in the brain [1]. This causes locomotor, cognitive, and behavioral problems. Memory loss is less prevalent in VCD, where issues with executive function are more common

[22]. In the Gothenburg MCI study, patients with other causes of VCD (post-stroke dementia or multi-infarct dementia) were excluded, therefore, there are only SSVD and mixed SSVD/AD patients in the VCD data, see subsubsection 2.1.5.1 [9].

Currently, there is no curative treatment for VCD [23]. It is, however, possible to prevent it if the conditions leading to VCD are caught before the disease progresses. Changes can be made to lifestyle and diet, which have proven effects in preventing VCD. If it is too late for prevention, it is still possible to slow the progression if caught early. This makes early detection highly important. The progression can be slowed by treating the underlying causes [24]. VCD risk factors, such as hypertension and diabetes, can be treated with medication [23]. Further, untreated hypertension causes WMH, which is common in VCD. There are also risk factors that cannot be treated or changed [1]. Age and gender are known to have an impact on the risk of developing VCD [22], [25]. Patients of older age and male sex have a greater risk of VCD.

2.1.5.1 Mixed SSVD/AD

Mixed SSVD/AD is a subtype of VCD, where the vascular disease is combined with AD [1]. A patient with mixed SSVD/AD exhibits the overlapping characteristics of both conditions, i.e., they might have both memory problems and reduced cognitive speed [9]. They present with clinical AD symptoms according to specific criteria, along with a certain amount of WMH, combined with a marked frontal lobe syndrome. Mixed SSVD/AD contributes to about 20% of cognitive disease cases [9], [26].

2.2 Diagnostic procedure

According to the guidelines issued by the National Board of Social Affairs and Health, diagnosing cognitive disease should proceed as follows: a basal investigation is done as a first step, where interviews with relatives, cognitive screening tests such as the clock test and MMSE, CT, and blood tests are done to determine whether the patient has cognitive impairment, and what type [7]. Even though CT is the standard imaging procedure, MRI is sometimes used if the clinic has access to a machine. If the results of the initial investigation are not conclusive, a more extensive investigation is conducted in a specialist unit. This can include neuropsychological tests, MRI, single-photon emission computed tomography (SPECT), lumbar puncture, or fluorodeoxyglucose-positron emission tomography (PET), depending on what is deemed necessary for the patient or available in the region.

The basal investigation described above is the standard at primary healthcare centers in Sweden. The Memory Clinic at Sahlgrenska University Hospital conducts a more extensive initial investigation. They perform tests required in the basal investigation, but also MRI and lumbar puncture, to diagnose more accurately.

Clinicians accurately diagnose what type of cognitive disease a patient has in around 70% of cases [27]. Many patients suffering from cognitive disease have not been

diagnosed appropriately. A cognitive disease screening study in Germany found that out of all the patients that had cognitive disease, only 40% had previously been diagnosed [28]. When the patients are eventually diagnosed, it is typically at a late stage of the disease [29].

The method commonly used to assess WMH in MRI images in clinical settings is the Fazekas scale [30]. It is a visual rating scale, where periventricular and deep WMH are rated separately, with a score from 0 to 3, where 0 is no WMH and 3 is large WMH.

2.3 Machine learning

This section covers the theory related to the ML model, loss functions, and MTL. It also covers the theory behind the explainability method used in this work.

2.3.1 ResNet18

ResNet18 is an 18-layer deep Residual Neural Network, and was developed by a Microsoft Research team [31]. It consists of convolutional layers and residual skip/shortcut connections organized into residual blocks. The convolutional layers apply a set of filters, commonly of size 3x3. A shortcut connection is added to each pair of filters in the convolutional block. See Figure 2.2 for a representation of the ResNet18 architecture.

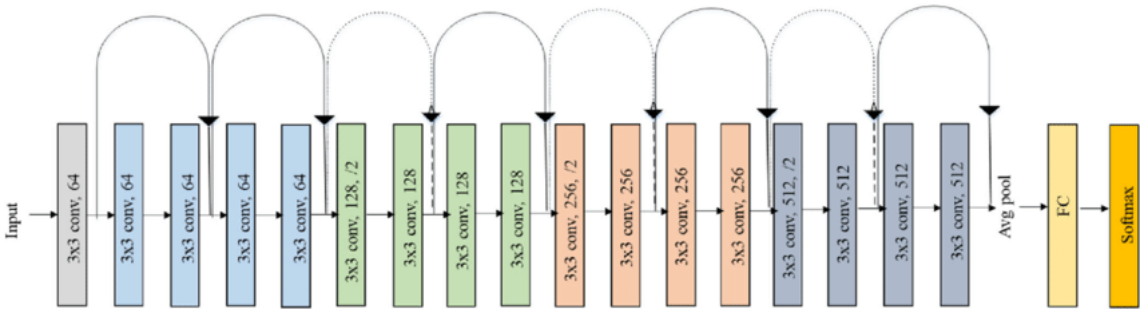


Figure 2.2: An illustration of the ResNet18 architecture with the added shortcut connections [32]. conv. = convolutional layer, FC = fully connected layer.

2.3.2 Loss functions

The loss function is a measure of the cost of the model's predictions [33]. The more mistakes the model makes, the higher the loss is. The goal of training is to minimize the loss. This section presents the theoretical background of the different loss functions evaluated during the project work. The first three loss functions presented are used for regression models, while the remaining ones are applied to classification models.

2.3.2.1 Mean squared error

Mean squared error (MSE) loss computes the error by taking the average squared difference between the model's predicted values and the ground truth values [34], see Equation 2.1. In the equation, y_i is the actual value, $\hat{\mathrm{y}}_i$ is the predicted value, and n is the total number of samples:

$$\mathrm{MSE} = \frac{1}{n}\sum_{\mathrm{i}=1}^{n}(y_i - \hat{y}_i)^2. \tag{2.1}$$

MSE loss assigns a higher penalty to larger errors.

2.3.2.2 Mean absolute error

Mean absolute error (MAE) loss calculates the average of the absolute differences between the model's predicted values and the actual values [34], see Equation 2.2. In the equation, y_i is the actual value, $\hat{\mathrm{y}}_i$ is the predicted value, and n is the total number of samples:

$$\mathrm{MAE} = \frac{1}{n}\sum_{\mathrm{i}=1}^{n}|y_i - \hat{y}_i|. \tag{2.2}$$

MAE loss treats all errors with equal weight [35].

2.3.2.3 Huber loss

Huber loss combines both MAE loss and MSE [34]. The Huber loss will be quadratic if the absolute value of the residuals is less than or equal to the threshold δ, and linear if the absolute value is greater than δ, see Equation 2.3. Huber loss uses the threshold to switch between MSE and MAE loss, as shown below:

$$L_\delta(a) = \begin{cases} \frac{1}{2}(y_i - \hat{y}_i)^2 & \text{if } |y_i - \hat{y}_i| \leq \delta, \\ \delta(|y_i - \hat{y}_i| - \frac{1}{2}\delta) & \text{if } |y_i - \hat{y}_i| > \delta. \end{cases} \tag{2.3}$$

2.3.2.4 Cross-entropy loss

The cross-entropy loss function is commonly used in multi-class classification [36]. In Equation 2.4, the loss function is defined for k classes [37]. Here, y_{mi} is the true probability of class m and data point i, 1 for the correct class, and 0 for the other classes, and x_i is the output of the network, as shown below:

$$\mathrm{L_{CE}} = -\sum_{\mathrm{i}=1}^{n}\sum_{\mathrm{m}=0}^{k} y_{m,i}\log(\mathrm{softmax}(x_i)). \tag{2.4}$$

Applying softmax to the classification model output produces k numbers that behave like probabilities, meaning they are all between 0 and 1, and sum to 1.

2.3.2.5 Focal loss

Focal loss is a modification to cross-entropy loss to combat class imbalances [38]. Adding the term $(1 - softmax(x_i))^\gamma$ to the cross-entropy loss, where γ is an adjustable focusing parameter ≥ 0, focal loss is defined as:

$$\mathrm{L_{FL}} = -\sum_{\mathrm{i=1}}^{n}\sum_{\mathrm{m=0}}^{k}(1 - softmax(x_i))^\gamma y_{mi}\log(softmax(x_i)). \tag{2.5}$$

2.3.3 Overfitting

Overfitting is a common issue in ML, where the model fits the training data too well and is therefore unable to generalize to unseen data [39]. This is caused by noise, a limited dataset, and a too complex model. This section presents methods that are known to alleviate this problem.

2.3.3.1 Data augmentation

In real-world ML problems, the amount of data available to train a model is often insufficient [40]. Data augmentation is an effective way to mitigate this problem, and still be able to train a generalized model. The training set can increase in size, quality, and diversity using data augmentation. This gives a more complete data set, which will help minimize overfitting [41].

Data augmentation on image data sets can be performed in two ways, synthetic data generation or data warping [41]. Synthetic data generation creates new synthetic images, while data warping transforms the existing images using transforms such as contrast changes, rotations, and translations.

2.3.3.2 Batch normalization

Batch normalization is used to normalize the activations of the hidden layers in the deep learning model during training [42]. Each mini-batch is normalized to have a mean of 0 and a variance of 1 [43]. The normalization during training ensures that the model is less sensitive to the initial training parameters, allowing larger learning rates. As a result, batch normalization can speed up training and improve accuracy and generalization.

2.3.3.3 Dropout

Dropout is used to prevent neural network units from co-adapting [44]. During training, units are randomly dropped from the neural network. The final model will then act like a combination of multiple network architectures, which will improve model performance and prevent overfitting. Research has found that performance can be improved by regularization using dropout [45]. However, dropout can slow down the training, since different units are dropped in every training step.

2.3.3.4 Reduce learning rate on plateau

Using a constant learning rate is not always ideal when learning a complex pattern [46]. To improve the learning of these complex patterns, the learning rate can be decayed during training. Learning rate decay is a method used to improve both optimization and generalization. One implementation is to automatically reduce the learning rate when the validation loss plateaus.

2.3.3.5 Weight decay

Weight decay is a popular method used for regularization [47]. The weights of the model are kept small by penalizing large weights [48]. This reduces the model's complexity, which improves the model's generalization.

2.3.3.6 Label smoothing

Label smoothing is a regularization method that creates soft targets from the original, hard targets [49]. The soft targets are drawn from a weighted average of the hard targets and a uniform distribution. This improves accuracy and generalization by preventing the model from becoming overconfident.

2.3.3.7 Early stopping

Early stopping is a regularization technique that uses the average error on the validation set to determine when to stop training the model [50]. It is based on the assumption that the validation error approximates the test error, and is used to improve generalization by stopping training once the validation error stops decreasing. Early stopping is widely used since it is easy to implement and has been found to be superior to other methods in mitigating overfitting. However, it is possible that early stopping can occur before the model has converged, and the validation error can temporarily increase before decreasing further.

2.3.4 Multi-task learning

MTL trains multiple ML tasks simultaneously [51]. This can improve generalization by sharing representations across related tasks [52]. The assumption is that related tasks share a common feature representation of the input features, and training them together could produce a more robust representation [53]. Apart from minimizing overfitting, MTL can improve the performance of the original task. However, the tasks must be related; training unrelated tasks can instead greatly reduce performance. The most common implementation of MTL is by using hard parameter sharing. Hidden layers are shared between all tasks, while each task has its task-specific output layer (head) [52].

2.3.5 Explainability model

The explainability method selected for this work was Guided gradient weighted class activation mapping (Grad-CAM), which combines Grad-CAM and guided backprop-

agation.

Grad-CAM is a white-box ML model. A white-box model refers to a transparent ML model, which allows users to understand the internal workings and what features influence the outcome and decision [54]. It identifies and analyzes the gradient information that affects the final convolutional layer in the network, assigning importance values to individual neurons [55]. The class-discriminative localization map, Grad-CAM, can be defined according to Equation 2.6. Here, c represents the class, and α_k^c is defined in Equation 2.7, and is the neuron importance weights. The weights are obtained by first evaluating how the class score y^c changes in response to the feature map activations A^k from a convolutional layer. These weights are applied to the forward activation maps, and a rectified linear unit (ReLU) function is then used to generate the final class-specific localization map (heatmap), as shown in:

$$\mathrm{L}^c_{\text{Grad-CAM}} = \mathrm{ReLU}(\sum_k \alpha_k^c A^k). \tag{2.6}$$

The neuron importance weights α_k^c are computed as follows:

$$\alpha_k^c = \frac{1}{\mathrm{Z}} \sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k}. \tag{2.7}$$

The generated heatmap then shows where the model fetches information from in the image.

Guided backpropagation identifies which parts of an image contribute to activating specific neurons in the model. This is achieved by filtering out negative gradients using backpropagation through a ReLU layer [55]. It merges the principles of standard backpropagation and deconvolutional networks [56]. While deconvolutional networks allow only positive gradients to pass backward, backpropagation focuses on features with positive activations in earlier layers.

Guided Grad-CAM combines the visualizations of Grad-CAM and Guided backpropagation using element-wise multiplication [55].

2.4 Evaluation metrics

Seven different evaluation metrics were used to evaluate how well the models perform. Pearson correlation coefficient, Spearman's rank correlation coefficient, and R-squared (R^2) were used to evaluate the regression head. Accuracy, precision, recall, and F1-score were used to assess the classification head.

2.4.1 Pearson correlation coefficient

One of the most common performance metrics for evaluating regression models is the Pearson correlation coefficient (R) [57]. It measures the linear correlation between two variables, the target (t) and the predicted value (p). R is a normalized measurement of correlation calculated in the following equation,

$$R = \frac{\text{Cov}(t, p)}{\text{Std}(t) \cdot \text{Std}(p)}, \tag{2.8}$$

and is a value between -1 and 1. A value of 0 indicates no correlation, while -1 and 1 represent strong negative and positive relationships, respectively.

2.4.2 Spearman's rank correlation coefficient

When the target (t) and predicted value (p) do not follow a normal probability distribution, the assumptions required for calculating the Pearson correlation coefficient are violated [58]. Instead, the non-parametric measure Spearman's rank correlation coefficient (ρ) can be used to measure the correlation. When calculating ρ, t and p need to be ordered and assigned ranks, so that the largest value has rank N and the smallest 1. The calculation is then the same as for Pearson, except on the ranked t and p, as in the following equation:

$$\rho = \frac{\text{Cov}(\text{rank}(t), \text{rank}(p))}{\text{Std}(\text{rank}(t)) \cdot \text{Std}(\text{rank}(p))}. \tag{2.9}$$

2.4.3 R^2

The third evaluation metric used for the regression head is R^2. The R^2 value indicates how well the data fits the regression model [59]. It shows the proportion of variance in the dependent variable that is accounted for by the independent variable in the model. R^2 is a value between 0 and 1, and a higher R^2 value indicates a better fit of the model to the data. The formula for calculating R^2 can be seen in the equation below:

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}. \tag{2.10}$$

2.4.4 Accuracy

Accuracy measures the proportion of correctly predicted data points by the model [34]. It indicates how many percent of the model's predictions are correct, and is defined below:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{2.11}$$

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

2.4.5 Precision

Precision is the ratio of true positives to the total number of predicted positives [34]. It measures how often the model correctly identifies positive data points. Precision also accounts for false positives, where the model incorrectly labels negative data points as positive, and is defined below:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}. \tag{2.12}$$

2.4.6 Recall

Recall measures the model's ability to identify all of the actual true positive cases [34]. Recall accounts for false negatives, where the model falsely identifies data points as negative, and is defined below:

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}. \tag{2.13}$$

2.4.7 F1-score

F1-score is a combination of precision, see subsection 2.4.5, and recall [34], and is defined below:

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.14}$$

F1 score is useful in scenarios where accuracy can be misleading, such as when dealing with imbalanced datasets where positive cases are rare.

2.4.8 K-fold cross-validation

K-fold cross-validation is a method to evaluate how well an ML model generalizes [60]. The training data is split into K equally sized sets, where one is used as validation data, and the rest as training data [61]. Different model training runs are then performed, where each of the K sets is used as validation once. The results of the model are then averaged over the runs, which gives a more robust estimate of how well the model performs. K-fold cross-validation can also be used to find the optimal hyperparameters to train the model [60].

3

Methods

This section describes the work carried out during the project and outlines the implementation process. It covers everything from data retrieval and preprocessing to efforts aimed at minimizing overfitting and developing the ML models.

3.1 Data acquisition and preprocessing

This section describes the data available in the project and how it was collected. It also presents the class divisions and the preprocessing steps applied to the data. This includes conversion and normalization.

3.1.1 Data acquisition

The Gothenburg MCI study's data contains unprocessed brain MRI images in T1, T2, and FLAIR modality using a 1.5 Tesla MRI scanner, clinical data, each patient's diagnosis, and brain region masks and WMH measurements made from the T1 volume using the software FreeSurfer 5.3. The data was reviewed, and data irrelevant to this project was discarded. Only FLAIR MRI images, WMH measurements, and diagnosis data were used.

Instead of using personal information, each patient was given a random pseudo-number, which was connected to the patient's MRI images at a specific time point during the study. The imaging dataset contains three different MRI sequencing techniques: T1, T2, and FLAIR, and can differ in scan angle (sagittal, axial, and coronal angles), but only the FLAIR MRI was used for this work in the end. Every slice in the FLAIR MRI had a resolution of 416 x 512 (height x width) and was made up of approximately 28 slices.

The dataset contained more scans than patients, because some patients had done the MRI scan more than once. If a patient had more than one MRI examination, all images were connected to the same pseudo-number. When splitting into training, validation, and test sets, we made sure that every FLAIR MRI connected to the same pseudo-number was in the same set. This gave us a cohort of 506 patients and 750 FLAIR MRI images.

No additional curation of the MRI images was required, as the dataset had already

been used in a previous master's thesis project conducted in 2022 [62]. The dataset was well-sorted, and duplicates had been removed. In our project, an assessment was conducted to determine which of the available FLAIR modalities were best suited. The FLAIR images used in this project were acquired from a coronal angle, meaning the patient's brain was scanned from top to bottom. Two types of FLAIR images were available: slice thickness of 4 mm or 5 mm.

Figure 3.1 shows the number of images with their diagnosis category, which are: HC (146), SCI (212), MCI (207), AD (103), and VCD (82), where VCD includes pure SSVD and mixed SSVD/AD. This gave us a total of 750 MRI images to use.

Originally, we started with a five-class classification setup, where all the diagnostic categories presented above were treated as separate classes. However, due to issues with overfitting (see subsection 3.2.8) and the limited performance observed in this configuration, we decided to explore a three-class and a binary classification setup to investigate whether these approaches would yield improved results. The three-class classification consisted of "healthy" patients, AD, and VCD. While the binary classification case consisted of one class with vascular pathology and one without. Figure 3.2 shows the distribution of the images for the three-class classification and the binary classification. The three-class classification is divided such that HC and SCI belong to one class, AD and MCI patients with WMH volumes less than 4820 mm^3 belong to the second class, and VCD and MCI patients with 4820 mm^3 or more belong to class three. The binary classification case was similar, but the AD and MCI with less WMH were placed in the same class as HC and SCI. The WMH volume threshold of 4820 mm^3 for classifying vascular pathology in MCI patients was derived from a separate, ongoing study conducted at the memory clinic and was given to us verbally.

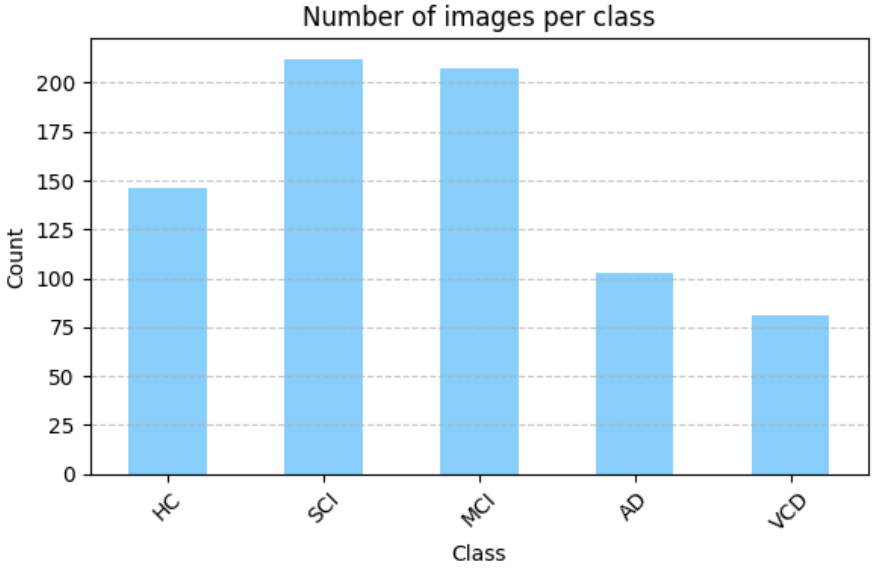
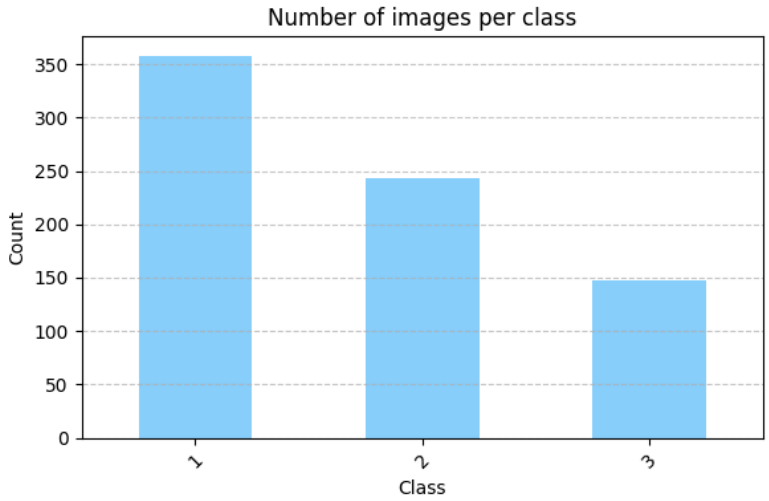
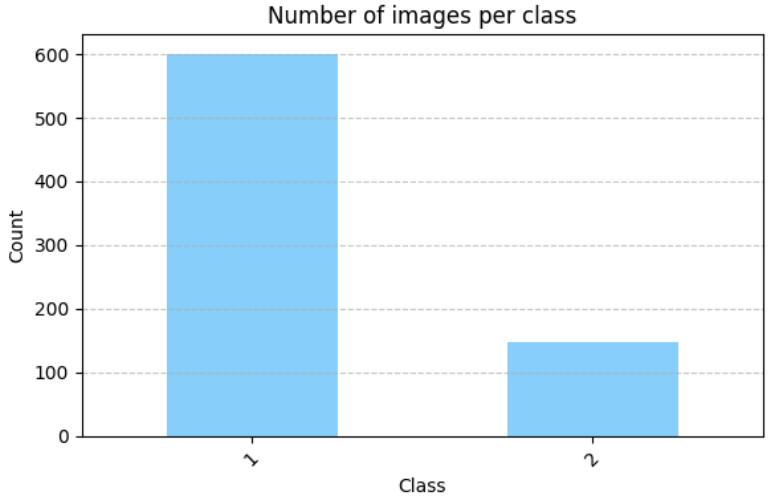


Figure 3.1: The dataset diagnosis distribution for five-class classification.



(a) Distribution for three-class classification. 1 = HC/SCI, 2 = AD/MCI with WMH < 4820 mm^3, 3 = VCD/MCI with WMH ≥ 4820 mm^3.



(b) Distribution for binary classification. 1 = HC/SCI/AD/MCI with WMH < 4820 mm^3, 2 = VCD/MCI with WMH ≥ 4820 mm^3.

Figure 3.2: The dataset diagnosis distribution for the cases with fewer categories.

3.1.2 NIfTI conversion

The original MRI images were Digital Imaging and Communications in Medicine (DICOM) files, but they were converted to Neuroimaging Informatics Technology Initiative (NIfTI) format. When doing this conversion, only the necessary metadata is stored in the header along with the image [63]. We created a Python script to convert the DICOM volumes to NIfTI format using the Python library `nibabel`[64].

3.1.3 Normalization

We normalized both the FLAIR MRI images and the ground truth volumes. The MRI images were individually normalized using an ML library called Medical Open Network for AI (MONAI)'s `ScaleIntensity` function [65]. The WMH volumes were normalized due to their high values, which caused the loss function to return extremely large values without normalization. The formula for the Z-normalization can be seen in Equation 3.1, where μ is the average WMH volume and σ is the standard deviation.

$$\text{Normalized WMH} = \frac{WMH - \mu}{\sigma} \tag{3.1}$$

3.2 Machine learning model

To address the two main problems our work focused on, we initially implemented a regression model to predict WMH volumes using FLAIR MRI images. The plan was to use these predicted volumes to classify VCD based on WMH burden. However, this approach was not successful, as we were only able to implement a linear classification model. As a result, we implemented a more advanced MTL model, which allowed for simultaneous regression and classification, see Figure 3.3, aiming to improve the classification performance.

For all models described in this section, measured WMH volumes were used as ground truth for the regression methods, while diagnosis labels served as ground truth for the classification methods.

3.2.1 Regression model

A regression task involves training a model on labeled data, where each input is associated with a real-valued target (ground truth), and the goal is to minimize the difference between predicted and actual values. As an initial step, several model architectures were compared to identify the most suitable approach for our regression task. The models evaluated included relevance vector machine (RVM) [66], a simple CNN [67], UNet (MONAI) [68], and ResNet18 (MONAI) [31]. In the initial experiments, the RVM performed poorly and was therefore excluded from further evaluation. Both the simple CNN and UNet models yielded reasonably good results, but the ResNet18 model outperformed both models when looking at the prediction plots and metrics. As a result, ResNet18 was selected as the base model for the regression task and later used as the foundation for the MTL model.

3.2.2 Classification model

A classification model is used to classify the correct label of the input data. In this project, the aim is to use classification to classify what cognitive disease diagnosis a patient has. Initially, the predicted WMH volumes from the regression model

were the only input to the classification model. This meant it was not possible to implement a more advanced model than a linear model. A logistic regression model was implemented using the scikit-learn LogisticRegression and applied to both multi-classification and binary classification.

3.2.3 Multi-task learning

To improve the results of the classification model, an MTL model was implemented. It was implemented with one regression head and one for classification, see Figure 3.3. The rationale was that by training shared weights for both regression and classification tasks, the classification head could benefit from knowing which image regions are important for estimating WMH volumes, based on the hypothesis that VCD can be diagnosed from WMH volume measurements.

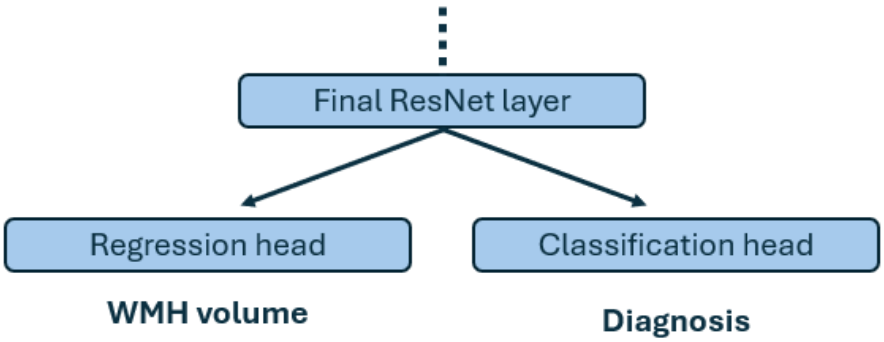


Figure 3.3: A schematic of the MTL model architecture.

To implement the MTL model, a classification head was added to the MONAI ResNet18 model used for regression. Two different loss functions were used to penalize both classification and regression error. They were combined as follows:

$$\text{Total loss} = 3 \cdot \text{Regression loss} + \text{Classification loss}. \tag{3.2}$$

The learning rate was 10^{-4}, and the optimizer used was adaptive moment estimation (Adam). The base MTL model was then implemented using binary classification, three-class classification, and five-class classification, where some layers and parameters were changed based on the results of experiments described in subsection 3.2.8 and subsection 3.2.9.

3.2.4 Evaluation metrics

We used the evaluation metrics presented in section 2.4 to evaluate the results from the MTL model. For Pearson correlation coefficient [69], we used `numpy.corrcoef`, and for Spearman's rank coefficient, we used a function from SciPy [70]. For R^2, accuracy, F1-score, precision, and recall, we used scikit-learn's built-in functions [71]. The weighted versions of F1-score, precision, and recall were selected to account for

class imbalance. The difference between the original and the weighted versions of the evaluation metrics is that the weighted metrics compute an average that is weighted by the number of true instances in each class.

We used the weighted metrics for the cases with three and five classes. Additionally, we applied them to the binary classification task to evaluate an ideal scenario. This could demonstrate the model's potential performance if the dataset had a more balanced distribution between the classes. However, since this does not reflect the actual distribution in our dataset (around 600 images for the non-vascular patients, and 150 for the vascular), we also used the original, unweighted equations presented in section 2.4 to illustrate how the model performs on our dataset.

3.2.5 Using T1 and FLAIR vs FLAIR alone as input

Since the dataset included both T1 and FLAIR MRI images, an experiment was conducted to evaluate whether combining both modalities as input would improve model performance compared to using FLAIR alone. The results of this comparison are presented in more detail in section 4.1. Although the results show improvements when including T1, we ultimately chose to proceed with FLAIR-only input due to its faster training time and clinical relevance in diagnosing VCD.

3.2.6 Adding covariates or not

To explore whether performance could be improved for both regression and classification, we included the covariates age and gender as additional inputs. They were added both separately and together using feature concatenation before the final output layers of the model. However, this did not lead to any improvement in results, and in one case, performance even deteriorated. Therefore, it was ultimately decided not to include any covariates. Due to limited time, we were only able to test a subset of the covariates suggested as risk factors for WMH. As a result, we cannot rule out the possibility that other covariates might have had a different impact on the outcome, but there was not sufficient time to investigate these.

3.2.7 Choosing loss function

We compared different loss functions to evaluate which were most suitable for our problem, three for the regression head and three for the classification head. For the regression head, we compared Huber loss, MSE, and MAE loss. For the classification head, we evaluated different loss functions for each classification case separately.

Each loss function tested on the regression head was trained for 10 epochs with a learning rate of 0.0001. The models were then evaluated using metrics commonly used for regression tasks: R, ρ, and R^2.

For the classification head, cross-entropy loss and focal loss were compared in both the five-class and three-class classification setups. The performance of each loss function was evaluated by analyzing the loss curves over time and assessing the

standard classification metrics: accuracy, F1-score, precision, and recall. Each model was trained for 15 epochs with a learning rate of 0.0001. In addition, for the binary classification case, binary cross-entropy with logits loss [72] was also included in the comparison and trained under the same conditions.

The results of this comparison are presented in more detail in section 4.2, but we ultimately chose to proceed with MSE for the regression head, cross-entropy for binary classification and the three-class classification, and focal loss for the five-class classification case.

3.2.8 Overfitting

This section presents the methods used to minimize overfitting. Some of the methods described below were applied in combination with each other. Due to the limited time frame of the project, we aimed to evaluate as many strategies as possible to minimize overfitting. Therefore, we chose to combine certain methods when it was feasible and efficient, while still testing others individually to isolate their effects. See Figure 3.4 for an example of how the overfitting could look.

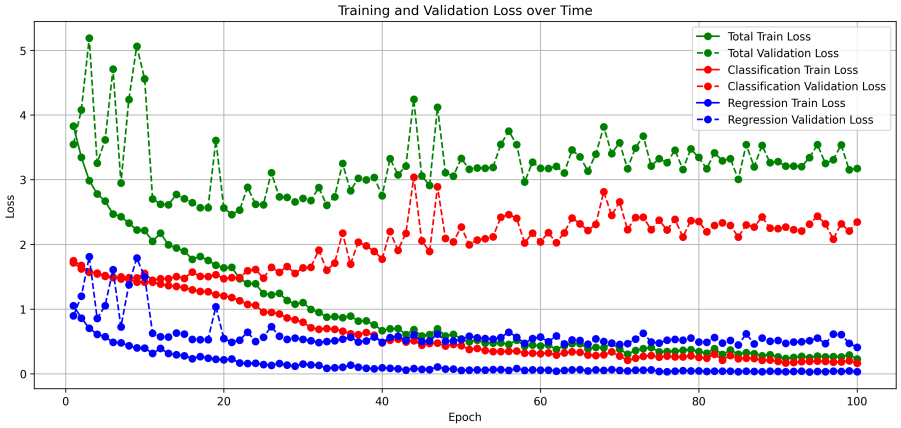


Figure 3.4: An example of the loss curves when the overfitting was at its worst. The loss function used in the figure was cross-entropy.

3.2.8.1 Data augmentation

To improve the generalization of the model, data augmentations were performed on the dataset. Augmentations that would be representative of the dataset were included and visually inspected to ensure that the images were not changed beyond recognition. To further ensure that this would help the model, an ablation study was done.

Starting with six different augmentations thought to be relevant in medical imaging tasks, each one was removed once to see what the effect of each one was. This was tested using the five-class classification case with both focal loss and cross-entropy loss functions. When applying focal loss, none of the different augmentations tested improved the results, nor did a combination of them. In the cross-entropy case,

when using Gaussian smoothing and Gaussian noise the results declined, so the final data augmentations were image translations up to 5 pixels, rotations up to 5°, scaling up to 1.1x, and contrast adjustments using gamma transformations with values sampled between 0.8 and 1.2, all chosen at random with a probability of 0.3.

3.2.8.2 Dropout

To minimize overfitting, we added dropout layers at two different locations in the ResNet18 model and tested at two separate stages. First, dropout was applied in the backbone of the model, before it branched into the classification and regression heads. In the second case, the dropout layer was placed within the classification head, as it was primarily the classification loss that overfitted, see Figure 3.4. In both cases, dropout was set to 0.5 for a first try and 0.3 for a second. Adding a dropout layer did not affect the overfitting and was therefore not used when running the final tests. Dropout was not combined with any other method when attempting to reduce overfitting.

3.2.8.3 Weight decay

Weight decay was incorporated into the optimizer as a regularization technique to prevent the model weights from becoming excessively large and improve generalization. The weight decay was set to 10^{-5} since it penalizes larger weights, which can help prevent overfitting, but it did not lead to any significant improvement in reducing overfitting. Due to a limited time frame, we were unable to test additional values. As a result, we cannot rule out the possibility that this method could help minimize overfitting; however, the value we had time to test did not improve the results. Weight decay was combined with batch normalization and label smoothing, and once with batch normalization only.

3.2.8.4 Batch normalization

We implemented batch normalization in an attempt to minimize overfitting. However, the results did not improve, and batch normalization was not included in the final tests. Batch normalization was combined with label smoothing and weight decay, and once with weight decay only.

3.2.8.5 Label smoothing

When using cross-entropy as the loss function, we experimented with label smoothing to prevent the model from becoming overly confident in its predictions and minimize the overfitting. We tried label smoothing values of 0.1 and 0.3, but neither resulted in significant improvements, so they were not included in the final tests. Label smoothing was used in combination with weight decay and batch normalization. It was also used alone once.

3.2.8.6 Reduce learning rate on plateau

Another measure to minimize overfitting was to reduce learning rate on plateau. This method was implemented to monitor the classification validation loss and to reduce the learning rate after 10 epochs, if no improvement had occurred. The initial learning rate was 10^{-4}. This method was evaluated using cross-validation k-fold, see section 4.4. These results showed that this method helped reduce overfitting to some extent, but the improvement was not as substantial as with other methods that we tried, and for one of the classification cases, the result worsened. Therefore, it was not included in the final tests.

3.2.8.7 Early stopping

We implemented early stopping together with checkpointing, where the model's weights were saved at the best validation loss, so testing was done on the best state of the model. This was done by exploring two different approaches. The first approach monitored the total average validation loss. The model was set to stop training if the total validation loss did not improve after a certain number of consecutive epochs.

The second approach monitored the total average validation loss for the classification, since this was where overfitting occurred. Instead of stopping based on the total validation loss, training was stopped if the classification validation loss did not improve after a certain number of epochs. We experimented with early stopping thresholds of 10 and 15 epochs, depending on the total number of planned training epochs.

3.2.9 K-fold cross-validation

K-fold cross-validation was applied to all three classification cases. To ensure that data from the same patient did not appear in multiple folds, the `GroupKFold` from scikit-learn was employed. For each task, 10% of the data was held out as a test set, and the remaining 90% was partitioned into eight folds. The model was trained eight times, using a different fold for validation in each iteration. Validation metrics were computed after each training run and subsequently averaged across all folds. For the five-class and three-class classification cases, weighted F1-score, recall, and precision were used, but for the binary case, only unweighted metrics were evaluated. Early stopping was applied during cross-validation to determine the median number of training epochs across folds, which was then used for final model training.

Cross-validation was also utilized to evaluate different regression loss functions. The mean validation metrics, particularly the F1-score, were compared across loss functions, and the one yielding the best performance was selected for the final training phase. A similar approach was applied to assess the impact of reducing the learning rate on plateau. However, as shown in section 4.4, this technique did not lead to a substantial improvement and was therefore not adopted in the final model configuration.

After the k-fold cross validation was performed, final training of the three models (binary, three-class, and five-class) was carried out using the full training dataset, excluding the reserved test set. The full training dataset included both the prior training and validation sets, resulting in only a training and a test set. The final training was not based on the best-performing fold, but on a combination of all folds. The choice of loss function and number of training epochs was informed by the cross-validation results. Final model performance was evaluated on the held-out test set, with results presented in section 4.5.

3.3 Explainability model

To verify that the model focused on and extracted information from areas commonly affected by WMH, an explainability method was employed. Guided Grad-CAM is a visual explainability technique that highlights activations from different layers of the model, which can then be overlaid on the original image to identify the specific brain regions involved in the model's decision-making process. The Guided Grad-CAM method was imported from Captum, a model interpretability library developed for PyTorch, and the implementation script was developed based on this framework [73]. A threshold value of 0.05 and a gamma value of 0.8 were chosen.

A preprocessing stage is conducted to enhance the clarity of the Grad-CAM visualizations. This includes data cleaning and normalization, removing background and minor noise, and enhancing important patterns. In this step, the threshold value is applied to remove background noise, while the gamma value is used to enhance weaker signals for improved visualization. We then normalized the heatmap to improve visualization quality and enhance the clarity of the resulting images. See Listing 3.1 for the code used for the preprocessing and normalization of the heatmap.

We added the possibility to select which patient's FLAIR MRI to visualize. The model generated and displayed the resulting Guided Grad-CAM heatmap over the selected FLAIR image by entering the random number corresponding to the patient. It was possible to display all the slices from one FLAIR image, but due to it resulting in 64 images, we limited it to slices 10 to 35. For some cases, slice 15 to 40 was used instead. A slice corresponded to each 2D FLAIR MRI image, with a 4 or 5 mm space between each slice.

Listing 3.1: The preprocessing code and normalization of the heatmap.

```
def preprocess_gradcam(attributions, threshold=0.2, gamma=0.5):
    attributions = attributions.squeeze()

    p5, p95 = np.percentile(attributions, [5, 95])
    attributions = np.clip((attributions - p5) / (p95 - p5), 0, 1)

    attributions[attributions < threshold] = 0
    attributions = np.power(attributions, gamma)

    return attributions
def normalize_heatmap(hm, min_val, max_val):
```

```
    """Normalizing heatmap"""
    return ((hm - min_val) / (max_val - min_val) * 255).clip(0,
        255).astype(np.uint8)
```

3.4 Use of artificial intelligence

During this thesis work, we utilized artificial intelligence (AI) tools to support both code development and report writing. ChatGPT was used to assist in generating code, troubleshooting, and improving existing code. For refining the written report, ChatGPT was primarily used to correct grammar, improve formatting, suggest better word choices, eliminate repetition, and provide synonyms. Additionally, Grammarly was used to further enhance the quality of the text, and LaTeX's built-in AI assistance was also employed.

4

Results

This section presents the relevant results of the project work. It includes outcomes related to the choice of input, selection of loss function, efforts to reduce overfitting, and model performance.

4.1 Using T1 and FLAIR vs FLAIR alone as input

The results showed that a difference in performance was notable. The model performed slightly better when using both T1 and FLAIR images. As shown in the Figure 4.2, the loss becomes more stable toward the end of training, compared to the curve in Figure 4.1. The ρ, R, and R^2 coefficients were also slightly higher when both image types were used, see Table 4.1. Ultimately, FLAIR was selected as the sole input modality. The reasoning behind this decision is discussed in section 5.1.

	R	ρ	R^2
FLAIR	0.7254	0.6073	0.5087
T1 and FLAIR	0.7690	0.6705	0.5642

Table 4.1: The resulting metrics on the validation dataset for FLAIR and T1/FLAIR as input.

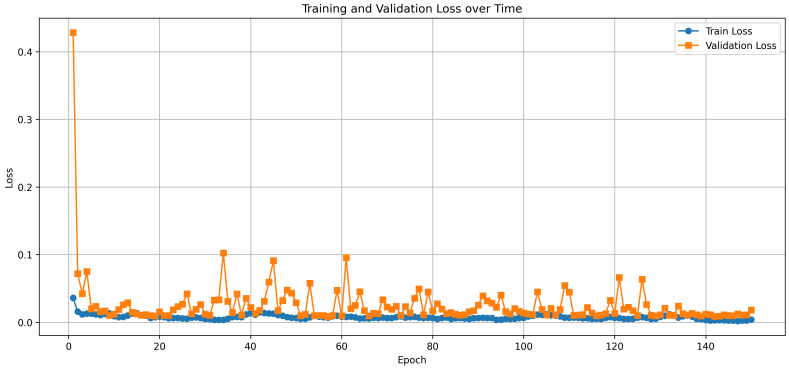


Figure 4.1: The loss curve using FLAIR as input over 150 epochs.

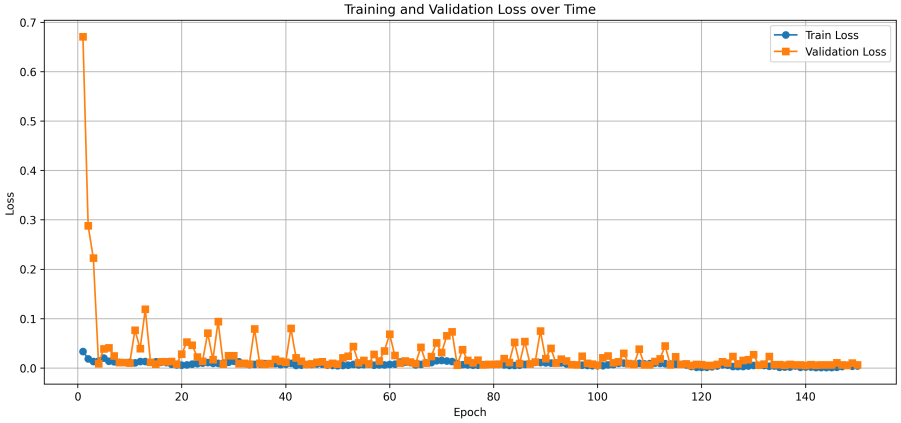


Figure 4.2: The loss curve using both T1 and FLAIR as input over 150 epochs.

4.2 Choosing loss function

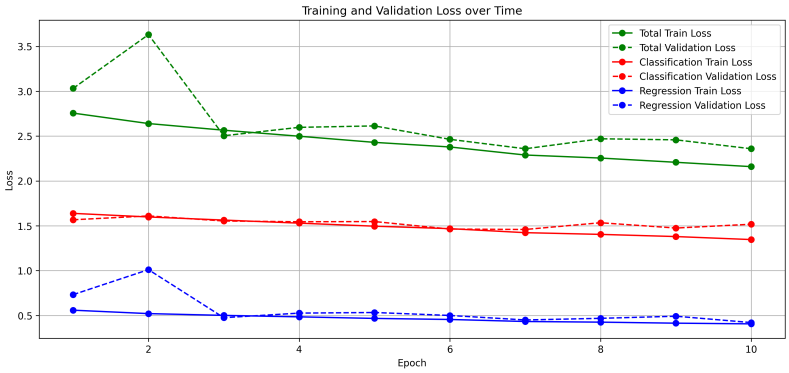
This section presents the results of the loss function comparison for both the regression head and classification head, and their three different classification cases.

4.2.1 Regression head

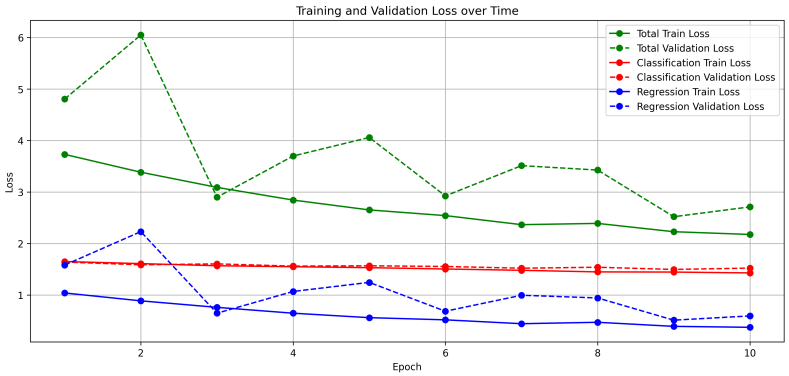
Table 4.2 shows the resulting metrics for the loss functions Huber loss, MSE, and MAE loss for the regression head on the validation dataset. As shown in Table 4.2, MSE was the loss function that yielded the best metric results in two out of three metrics, while Huber loss achieved the highest result in the third. Although MSE and Huber loss were relatively close in performance, MAE loss performed notably worse in two out of the three metrics compared to the other two loss functions. Figure 4.3 shows the loss curves for each loss function. While they all appear similar, the figure illustrates that the Huber loss may be better suited, as it results in both a lower total loss and a lower regression loss compared to the other two loss functions evaluated. Since MSE yielded the best metric values and the loss curves appeared similar across the tested functions, it was selected for continued use.

	R	ρ	R^2
Huber Loss	0.7039	0.7673	0.4533
Mean squared error	0.7161	0.7307	0.4642
Mean absolute error	0.5992	0.7073	0.2962

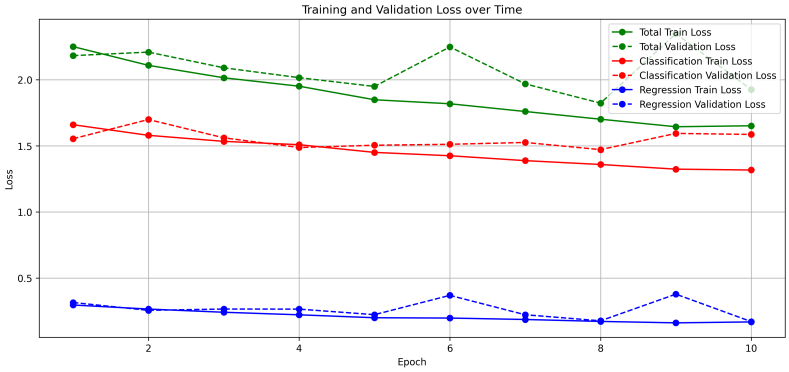
Table 4.2: The resulting metrics for the different loss functions for the regression head on the validation set.



(a) The loss curves using MAE loss.



(b) The loss curves using MSE loss.



(c) The loss curves using Huber loss.

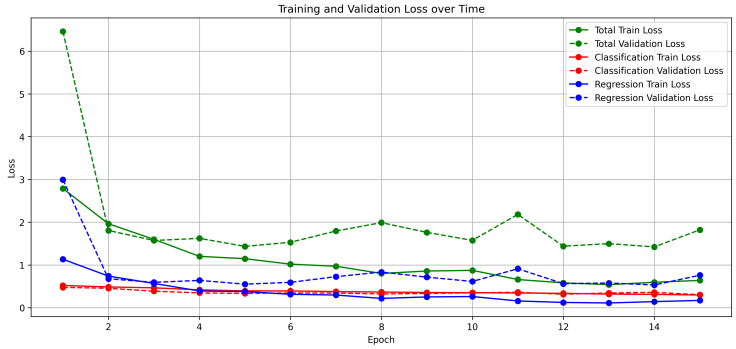
Figure 4.3: The loss curves for MAE loss, MSE, and Huber loss.

4.2.2 Binary classification

Table 4.3 shows the resulting metrics for the loss functions: binary cross-entropy with logits loss, cross-entropy, and focal loss for the classification head on binary classification. Figure 4.4 shows the resulting loss curves for each loss function. It can be seen in Table 4.3 that cross-entropy was the loss function yielding the best results for binary classification and was therefore selected for further use.

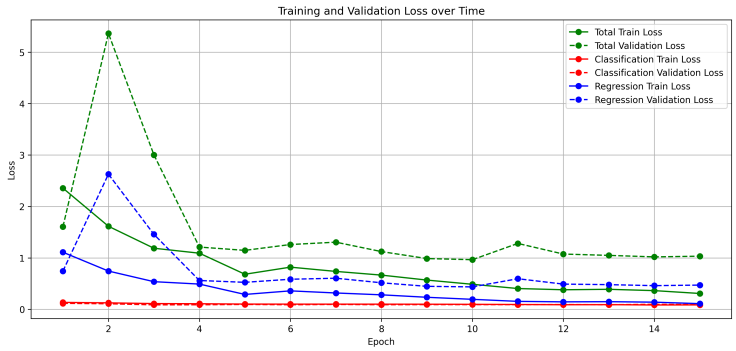
	Accuracy	F1-score	Precision	Recall
Binary Cross-Entropy with logits loss	0.8091	0.7237	0.6546	0.8091
Cross-Entropy	0.8455	0.8408	0.8378	0.8455
Focal loss	0.8273	0.7860	0.8041	0.8273

Table 4.3: The resulting metrics for the different loss functions using binary classification on the validation set.



(a) The loss curve using binary cross-entropy with logistic loss.

(b) The loss curve using cross-entropy loss.



(c) The loss curve using focal loss.

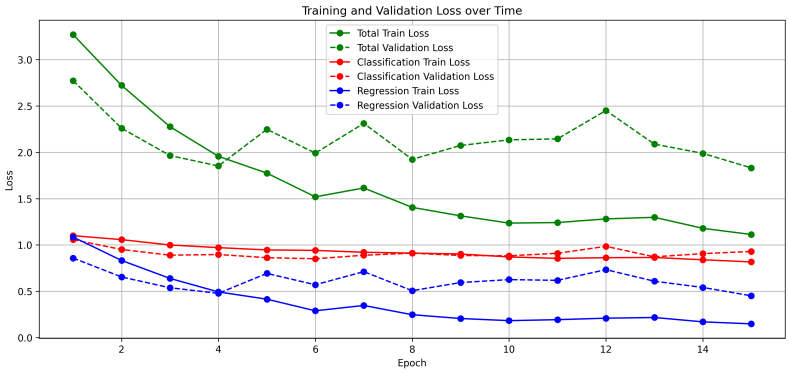
Figure 4.4: The loss curves for binary cross-entropy with logistic loss, cross-entropy, and focal loss.

4.2.3 Three-class classification

Table 4.4 presents the resulting metrics for the loss functions cross-entropy and focal loss, applied to the three-class classification task. Figure 4.5 displays the corresponding loss curves. Based on the metrics in Table 4.4, cross-entropy provided the most favorable results and was therefore chosen for further use.

	Accuracy	F1-score	Precision	Recall
Cross-Entropy	0.5909	0.4941	0.4322	0.5909
Focal loss	0.5727	0.4680	0.4193	0.5727

Table 4.4: The resulting metrics for the different loss functions for the three-class classification case on the validation set.



(a) The loss curve using cross-entropy loss.

(b) The loss curve using focal loss.

Figure 4.5: The loss curves for cross-entropy and focal loss applied on three-class classification.

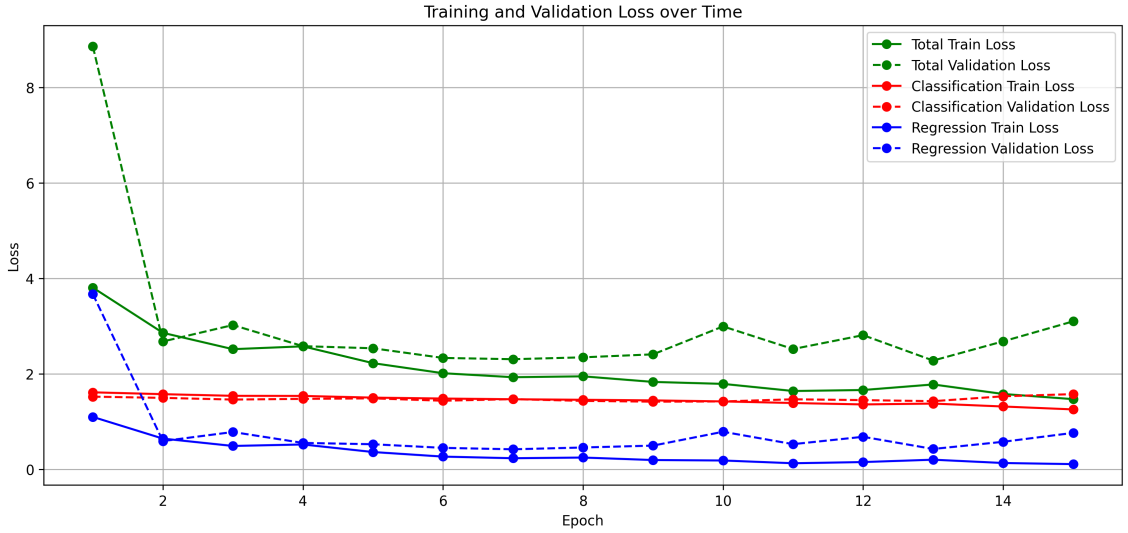
4.2.4 Five-class classification

Table 4.5 shows the resulting metrics for the loss functions cross-entropy and focal loss, for the classification head using five classes. Figure 4.6 shows the resulting loss curves for each loss function. It can be seen in Table 4.5 that focal loss yielded the

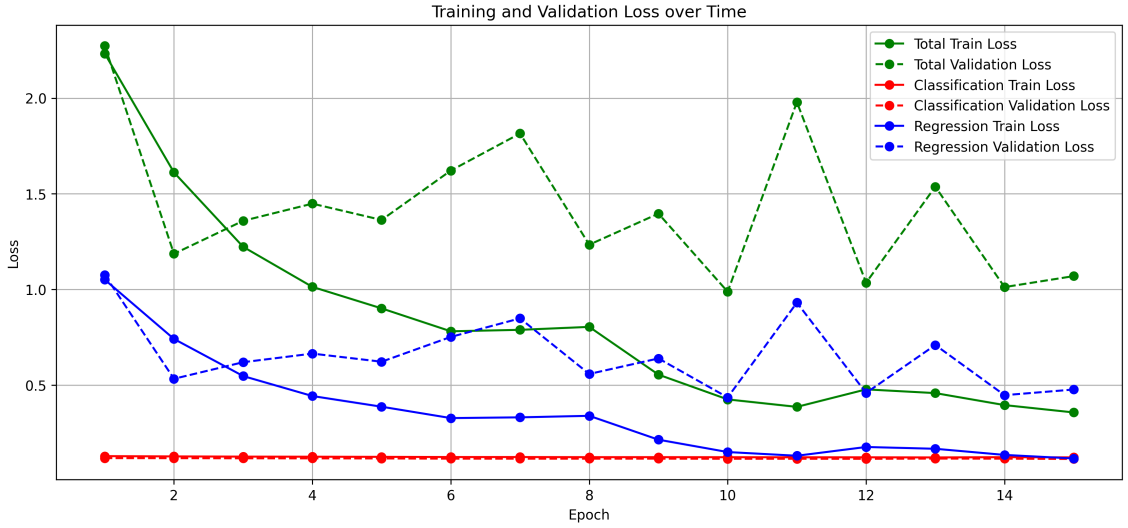
best metrics results and loss curves. Therefore, focal loss was selected for further use.

	Accuracy	F1-score	Precision	Recall
Cross-Entropy	0.3455	0.3209	0.3368	0.3455
Focal loss	0.4455	0.3300	0.4108	0.4455

Table 4.5: The resulting metrics for the different loss functions for the five-class classification on the validation set.



(a) The loss curve using cross-entropy loss.



(b) The loss curve using focal loss.

Figure 4.6: The loss curves for cross-entropy and focal loss for five-class classification.

4.3 Overfitting

Among the investigated methods to combat overfitting, only data augmentation and early stopping led to substantial improvement. The other approaches discussed in subsection 3.2.8 showed no notable results and are therefore not reported. Presented below are the results of using data augmentation and early stopping.

4.3.1 Data augmentation

The data augmentation ablation study was applied to the five-class classification case. It showed that removing the transforms Gaussian noise and Gaussian smoothing improved the F1-score when using cross-entropy loss, as seen in Table 4.6. When using focal loss, the results did not improve when using data augmentation, as seen in Table A.1 in the Appendix.

	Accuracy	F1-score	Precision	Recall
All augmentations	0.3277	0.2573	0.3151	0.3277
No augmentations	0.1750	0.1426	0.1828	0.1750
No Gaussian noise	0.3950	0.3521	0.4436	0.3950
No scaling	0.3445	0.2894	0.3479	0.3445
No rotation	0.3193	0.2467	0.2098	0.3193
No contrast adjustment	0.3025	0.2480	0.2426	0.3025
No Gaussian smoothing	0.3361	0.3227	0.3375	0.3361
No translation	0.2689	0.2532	0.3079	0.2689

Table 4.6: The validation set results of the data augmentation ablation study using cross-entropy loss using five-class classification.

The example in Figure 4.7 shows the loss graph when removing the Gaussian smoothing transform, using cross-entropy loss.

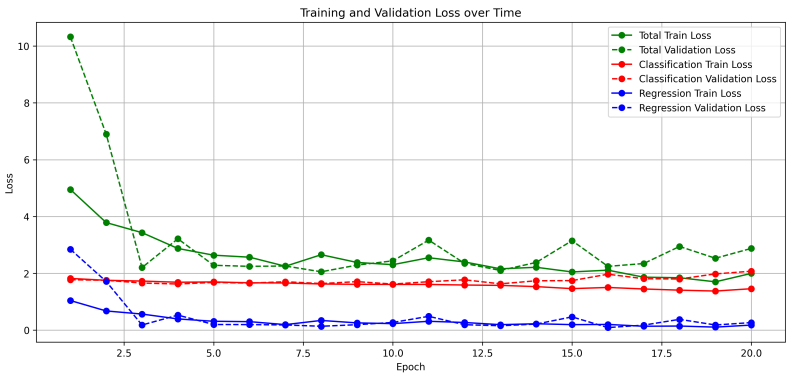


Figure 4.7: The loss graph when not including Gaussian smoothing in the data augmentation transforms.

4.3.2 Early stopping

Early stopping and checkpointing generated better results for all evaluation metrics, both regression and classification. See Table 4.7 for comparison.

	R	ρ	R_2	Accuracy	F1-score	Precision	Recall
Without	0.7873	0.6827	0.5241	0.2182	0.2246	0.4881	0.2182
With	0.8061	0.8132	0.5540	0.3091	0.3040	0.3734	0.3091

Table 4.7: The validation set metrics with and without using early stopping on the five-class classification case.

4.4 K-fold cross-validation

The average validation metrics for the three cases with their best hyperparameters chosen are presented in Table 4.8, Table 4.9, Table 4.10.

	Mean ± Std
Accuracy	0.8326 ± 0.0667
F1-score	0.6533 ± 0.1040
Precision	0.5708 ± 0.1354
Recall	0.7893 ± 0.0845

Table 4.8: The average validation set results for binary classification using K-fold cross-validation.

	Mean ± Std
Accuracy	0.5394 ± 0.0488
F1-score	0.5182 ± 0.0671
Precision	0.5408 ± 0.1073
Recall	0.5394 ± 0.0488

Table 4.9: The average validation set results for the three-class classification case using K-fold cross-validation.

	Mean ± Std
Accuracy	0.2991 ± 0.0760
F1-score	0.2315 ± 0.0805
Precision	0.2315 ± 0.0798
Recall	0.2991 ± 0.0760

Table 4.10: The average validation set results for five-class classification using K-fold cross-validation.

Reducing the learning rate on plateau did not improve the results notably, and for five-class classification, it worsened the results, which can be seen in Table 4.11. The result shown in Table 4.11 uses focal loss and early stopping, with or without reducing the learning rate on plateau.

	Accuracy	F1-score	Precision	Recall
Without	0.2991 ± 0.0760	0.2315 ± 0.0805	0.2315 ± 0.0799	0.2991 ± 0.0760
With	0.2954 ± 0.0505	0.2167 ± 0.0557	0.2143 ± 0.0804	0.2954 ± 0.0505

Table 4.11: The results of reducing the learning rate on plateau in the five-class classification case, evaluated using K-fold cross-validation. At the top is without using learning rate reduction, and below is with. The results are presented as mean ± standard deviation.

4.5 Multitask learning

This section presents the final results from the three cases. The final outcomes were obtained by training on the entire training and validation set combined, and testing on an unseen test set. Early stopping was used during K-fold cross-validation to find out the best number of epochs for training the three classification cases.

4.5.1 Binary classification

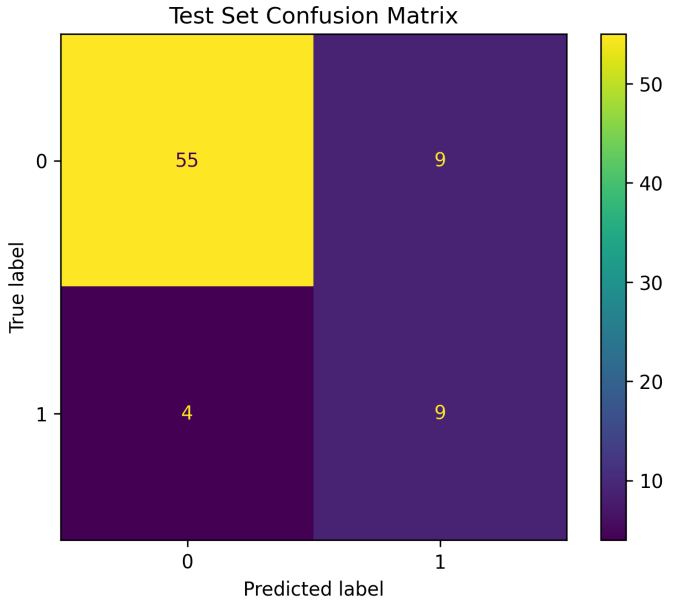
Table 4.12 presents the evaluation metrics for the test set in the binary classification case. The weighted evaluation metrics are shown in Table 4.13, illustrating the potential improvement in performance if a balanced dataset were used. Figure 4.8 shows the final confusion matrix along with a plot over the predicted volumes. In the prediction plot, pink dots represent the test set predictions, while blue dots correspond to the predictions from the training and validation sets.

R	ρ	R_2	Accuracy	F1-score	Precision	Recall
0.8387	0.8679	0.5969	0.8312	0.5806	0.5000	0.6923

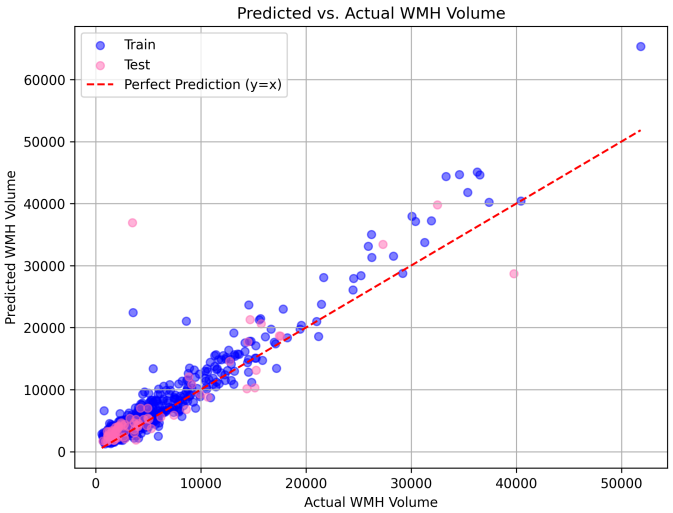
Table 4.12: The resulting test set metrics for binary classification using the original equations for F1-score, precision, and recall. This result represents the reality of the image distribution for our work.

R	ρ	R_2	Accuracy	F1-score	Precision	Recall
0.8387	0.8679	0.5969	0.8312	0.8414	0.8592	0.8312

Table 4.13: The resulting test set metrics for binary classification using weighted F1-score, precision, and recall, creating an ideal case where the distribution of images is even between the classes.



(a) The confusion matrix for the test set. 0 = HC/SCI/AD/MCI with WMH < 4820 mm^3, 1 = VCD/MCI with WMH ≥ 4820 mm^3.



(b) The model's final volume predictions.

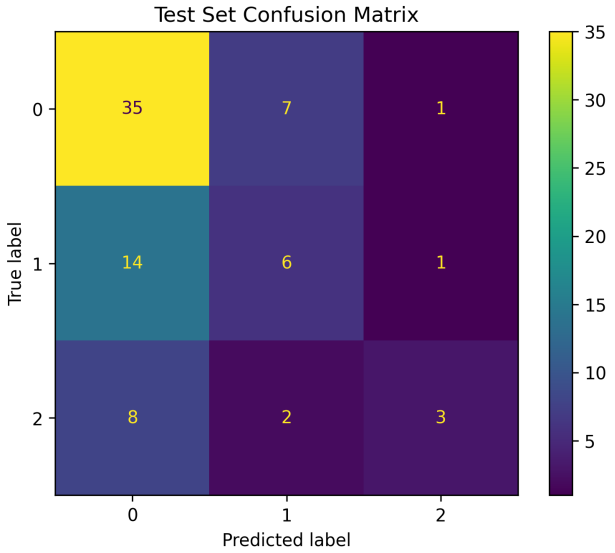
Figure 4.8: Confusion matrix and prediction plot for the test set in the binary classification case.

4.5.2 Three-class classification

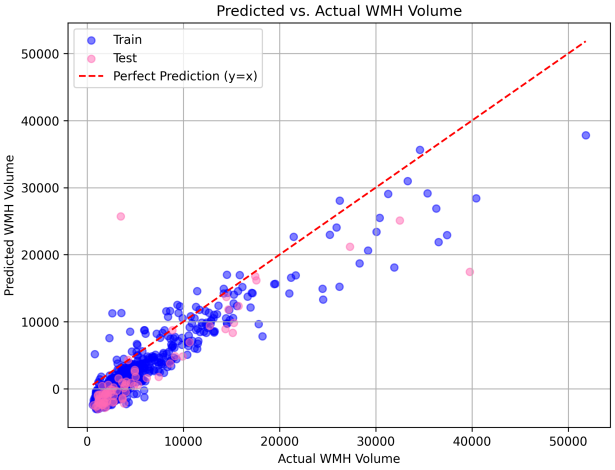
Table 4.14 shows the resulting evaluation metrics for the test set on the three-class classification case. Figure 4.9 presents the confusion matrix and a plot over the predicted values. In the volume prediction plot, pink dots represent predictions from the test set, and blue dots represent predictions from the training/validation set.

R	ρ	R_2	Accuracy	F1-score	Precision	Recall
0.8419	0.8584	0.4818	0.5714	0.5381	0.5533	0.5714

Table 4.14: The resulting test set metrics for three-class classification.



(a) The confusion matrix for the test set. 0 = HC/SCI, 1 = AD/MCI with WMH < 4820 mm^3, 2 = VCD/MCI with WMH ≥ 4820 mm^3.



(b) The model's final volume predictions.

Figure 4.9: Confusion matrix and prediction plot for the test set in the three-class classification case.

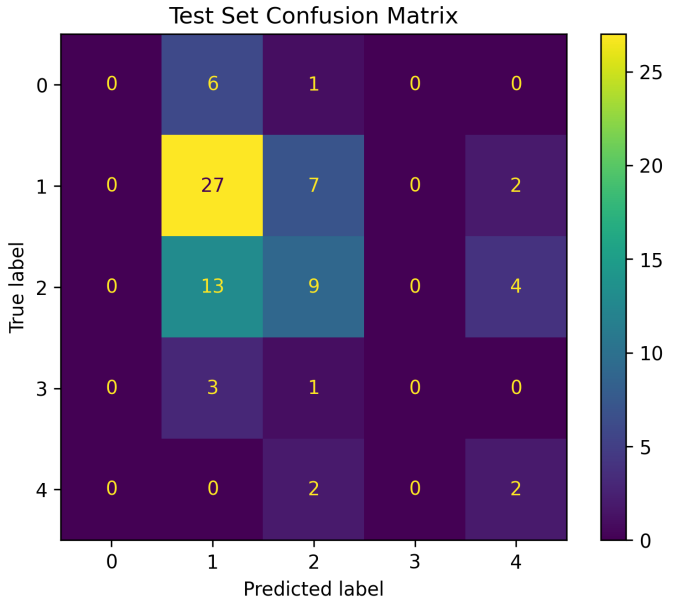
4.5.3 Five-class classification

Table 4.15 contains the final evaluation metrics for the five-class classification case. Figure 4.10 displays the confusion matrix and plot over the predicted values. Like

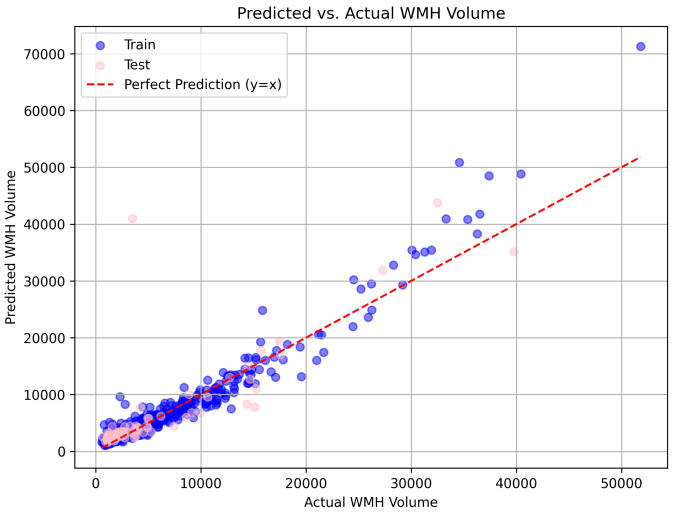
the above prediction plots, pink represents the test set prediction, and blue dots the training/validation set predictions.

R	ρ	R_2	Accuracy	F1-score	Precision	Recall
0.8204	0.8612	0.5409	0.4935	0.4465	0.4226	0.4935

Table 4.15: The resulting test set metrics for five-class classification.



(a) The confusion matrix for the test set. 0 = HC, 1 = SCI, 2 = MCI, 3 = AD, 4 = VCD.



(b) The model's final volume predictions.

Figure 4.10: Confusion matrix and prediction plot for the test set in the five-class classification case.

4.6 Explainability model

Figure 4.11 shows a Guided Grad-CAM representation for the five-class classification case. This patient exhibited a high amount of WMH and was diagnosed with VCD. In the heatmap, red indicates regions that the model relies on when making its predictions, while blue represents areas from which the model does not retrieve information. More Grad-CAM results can be found after Figure 4.11, where visualizations are presented for both the binary and three-class classification cases using different patient random numbers.

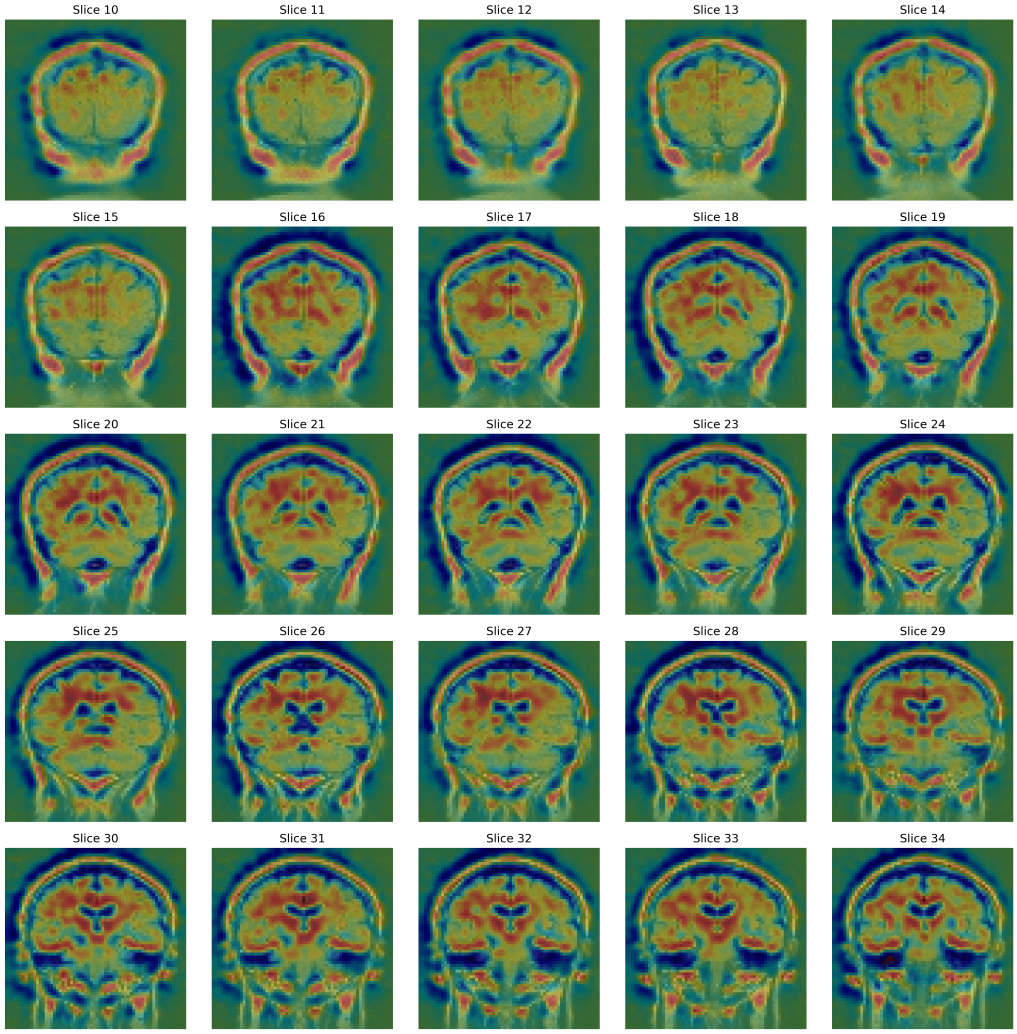


Figure 4.11: The resulting heatmap for a patient when using five-class classification shows that the model in this case captured regions within the brain, but also included some areas from the skull. This patient had a very large amount of WMH and was diagnosed with VCD.

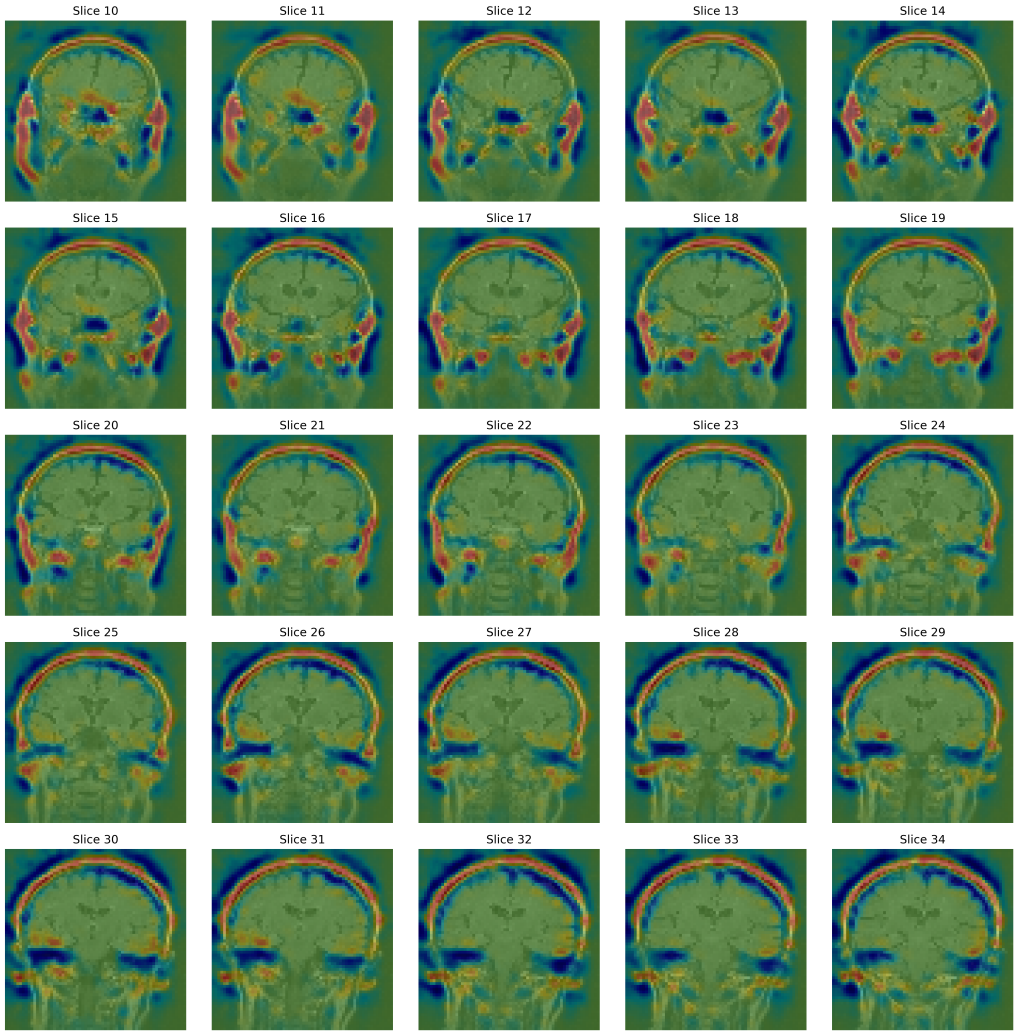


Figure 4.12: The resulting heatmap for a patient using binary classification shows that more activation was captured in the skull than in regions within the brain. This patient had a low amount of WMH and was diagnosed with SCI.

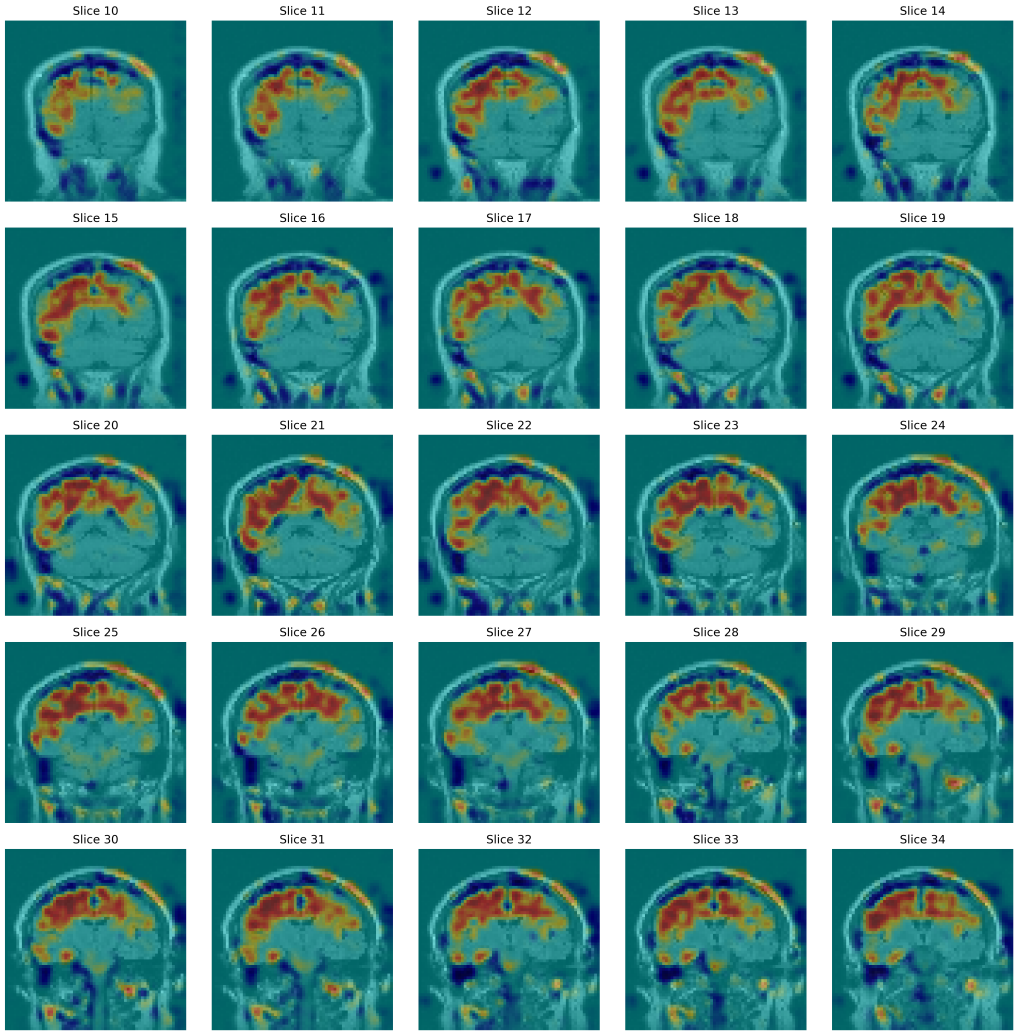


Figure 4.13: The resulting heatmap for a patient using three-class classification. It shows that, in this case, the model captured more regions within the brain and less of the skull. This patient had a fairly small amount of WMH in the brain and was diagnosed with MCI.

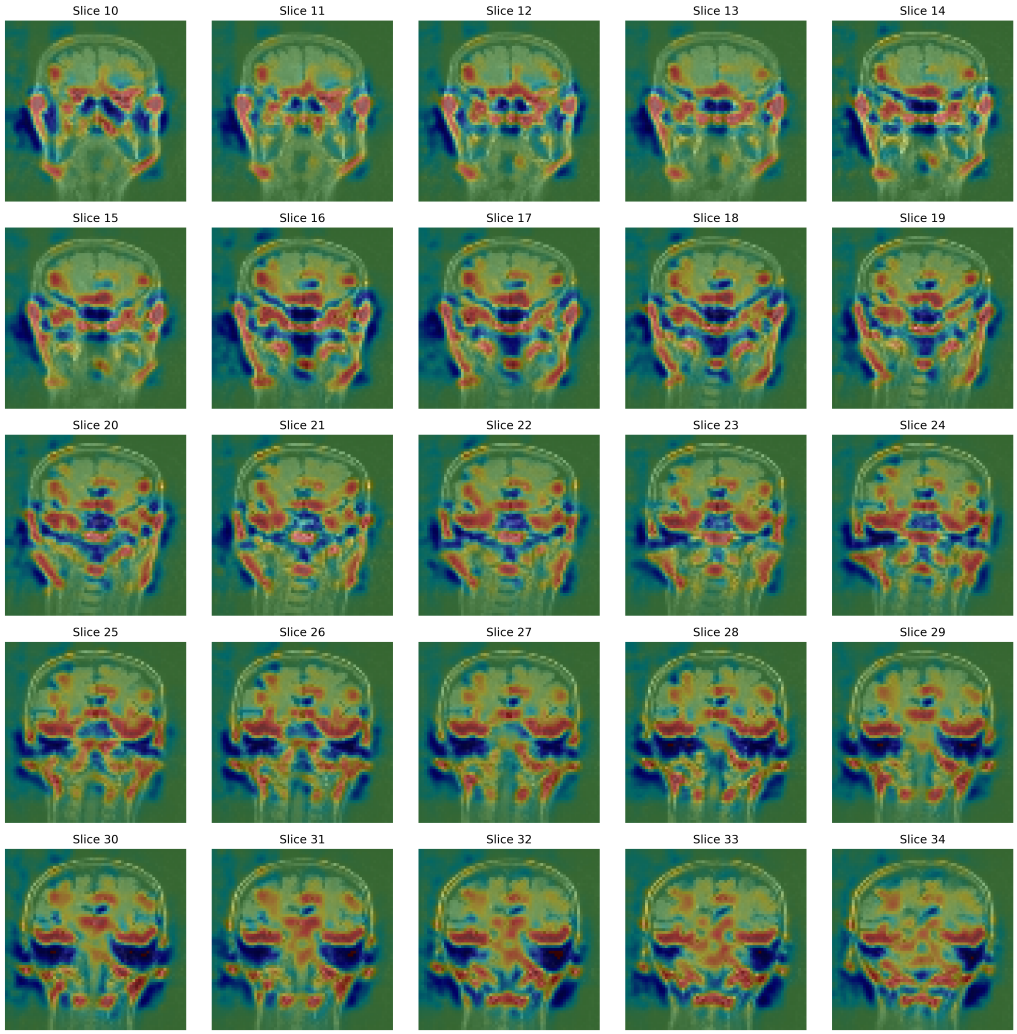


Figure 4.14: The resulting heatmap for a patient using binary classification shows that the model captured more regions within the brain, although some areas of the skull and nearby regions were still included. This patient had a very high amount of WMH in the brain, higher than the patient in Figure 4.11, and was diagnosed with MCI.

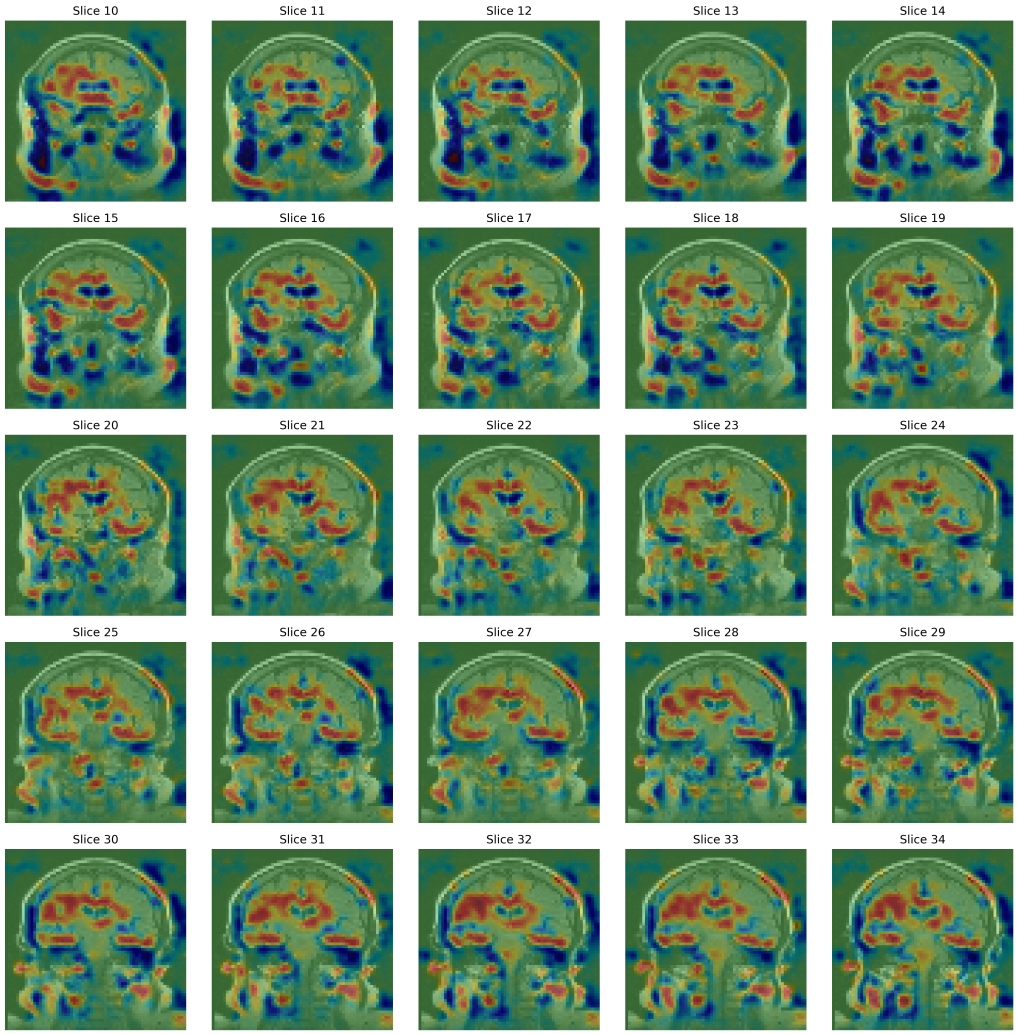


Figure 4.15: The resulting heatmap for a patient using five-class classification. The model captured regions within the brain, although some areas of the skull and nearby regions were still included. This patient had a very low amount of WMH in the brain and was an HC.

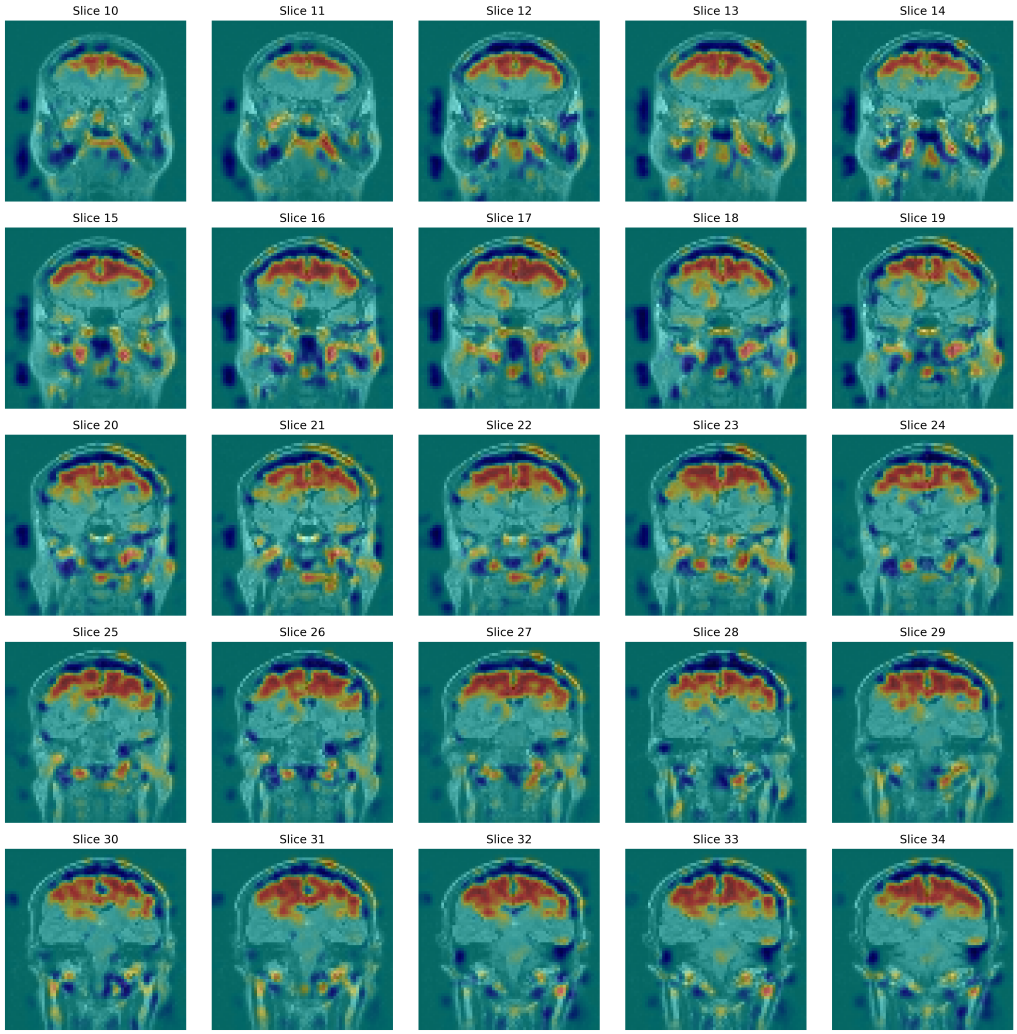


Figure 4.16: The resulting heatmap for a patient using three-class classification. The model captured regions within the brain, with minimal inclusion of the skull and surrounding areas. This patient had a relatively low amount of WMH, but more than the patient in Figure 4.13, and was diagnosed with SCI.

5

Discussion

5.1 Using T1 and FLAIR vs FLAIR alone as input

Even though using FLAIR and T1 images in combination produced better results than using FLAIR alone, we decided to use only FLAIR images. One of the two main reasons for this was the clinical relevance of FLAIR. The original project proposal was to investigate whether it was possible to diagnose VCD using FLAIR MRI images, and therefore, we wanted to continue with that plan. Another reason for selecting FLAIR as the input modality was that including both FLAIR and T1 images doubled the training time. Looking at the results using both T1 and FLAIR, the improvement was not notable enough to justify the increase in training time.

5.2 Choosing loss function

When selecting which loss functions to evaluate, we chose those most commonly used for both regression and classification tasks. The final choice of loss function for each task was primarily based on the resulting evaluation metrics, with some consideration given to the appearance of the loss curves. In certain cases, the differences in evaluation metrics between the loss functions were minimal, while in others, a clear distinction was observed. For the binary classification case, the focal loss curve appeared smoother and more stable; however, cross-entropy yielded better evaluation metrics and was therefore selected. In contrast, for the five-class classification case, the focal loss curve appeared less stable than cross-entropy, yet it produced better metric scores.

The final decision to use different loss functions for the various classification cases was driven by the evaluation metrics obtained. While it would have been possible to apply either cross-entropy or focal loss consistently across all cases, the performance differences for some metrics were too significant to overlook. In the binary and three-class classification cases, the F1-score was notably higher when using cross-entropy compared to focal loss. For the five-class classification, accuracy, precision, and recall were all better using focal loss compared to using cross-entropy. Therefore, to achieve the best performance, we used both loss functions.

5.3 Overfitting

Many different methods to combat overfitting were experimented with, both individually and in combination. Still, it seems that the model overfits to some extent in all three classification cases. Looking at Figure 4.7, even though the data augmentation improves the overfitting, the validation loss still increases more than the training loss before 20 epochs. The figure also shows that the classification validation loss starts increasing before the regression validation loss. It seems that the classification head is trained faster than the regression head, which is one reason for the model overfitting. If the tasks used in MTL are not related enough, the model can overfit to the unrelated task, which results in reduced performance on the original task. However, we believe the two tasks are related, and therefore consider it unlikely that this is the primary reason for the model overfitting.

Weight decay, label smoothing, dropout, and batch normalization did not combat the overfitting. While these are methods for regularization, such as penalizing large weights and preventing overconfident predictions, this may not address the cause of overfitting in our case. The most probable reason the model continued to overfit, and why the only methods that proved effective were data augmentation and early stopping, is that the dataset is too small to properly train this regression and classification model. Future work will therefore focus on improving augmentation and dataset expansion strategies, or ML methods that do not rely as heavily on a large dataset.

5.4 Multi-task learning

This section contains the discussion around the MTL model's performance for each classification case and the regression. It also includes the division of diagnosis classes.

5.4.1 Performance analysis

The binary model developed in this project demonstrated the best performance among the three classification approaches. However, an F1-score of 0.5806 indicates that the model is not yet suitable for clinical application. With access to a larger and more balanced dataset, the model would likely perform better and could potentially become a valuable tool for evaluating vascular contributions to cognitive impairment. The ideal scenario, in which weighted metrics were used, illustrated that a more balanced dataset could lead to improved results. The primary research question posed was whether it is possible to identify patients with VCD based on FLAIR images using ML. Our results suggest that this is possible, as the model shows signs of learning from the training data rather than merely guessing the disease type. However, further work is needed before any definitive conclusions can be drawn.

This project was not able to achieve an acceptable outcome in the five-class classification case and has therefore not proven that it is possible to differentiate between different cognitive diseases based on FLAIR images using ML. There are a few

different reasons for this, which are discussed in subsection 5.4.2. Being able to differentiate between five different cognitive diseases may be of limited clinical value, since there are already fluid biomarkers, such as β-amyloid, available to detect AD. By combining biomarker information with our model, it is possible to distinguish between AD, SSVD, and mixed SSVD/AD, which are the three most common types of cognitive disease.

Regarding the three-class classification case, this project was unable to achieve an acceptable outcome. However, it performed better than the five-class case and achieved higher evaluation metric scores. It reached an accuracy score of 0.5714 and an F1-score of 0.5381, which is not as good as the performance observed in the binary classification case, but still better than for the five-class classification. As mentioned above, this indicates that the model struggles to differentiate between different cognitive diseases but performs relatively well in distinguishing between VCD and non-vascular conditions.

Although the ability to differentiate between various cognitive diseases was also desirable, the main focus of this project was to be able to detect vascular influence in cognitive impairment based on FLAIR MRI images, by focusing on areas commonly affected by WMH. In the binary classification case, the test set metrics for the regression task were: $R = 0.8387$, $\rho = 0.8679$, $R^2 = 0.5969$. Using regression for quantification of WMH volumes was used to make the model focus on areas in the brain commonly affected by WMH.

The regression results for all three models are generally good and relatively similar. Each model achieves comparable scores across the evaluated regression metrics. For both the R and ρ coefficients, the models score between approximately 0.80 and 0.87, depending on the specific metric. Regarding the R^2 score, all three models achieve values in the range of 0.50 to 0.60. These results indicate that the regression part of the MTL performs consistently across all classification cases, regardless of whether there are two, three, or five classes. This outcome is expected, as the disease labels do not influence the regression model but are solely dependent on the ground truth WMH volume measurements.

5.4.2 The formation of classification categories

The initial plan was to conduct experiments only on the five-class classification case. However, due to issues with overfitting, we also tested a binary and a three-class classification case. The decision to include a binary and a three-class classification case was made after discussions with the research team at the Memory Clinic, as a strategy to help reduce overfitting.

The five-class classification case was proved to be quite difficult to work with from the start, and not just because of overfitting. The number of classes was too high with the amount of imaging data available, resulting in each class having few data points. The HC, SCI, and MCI classes had many more images than the disease categories AD and VCD, see 3.1, so the dataset was also highly imbalanced. This

affected the three-class and binary classification cases as well.

The five different diagnosis categories are heterogeneous groups, which means that there are differences in the type of disease within the groups. MCI is, for example, patients who in the future will develop either AD or SSVD, or even regress to SCI or HC. Diagnosing based on MRI images by focusing on WMH-affected areas will be challenging, as WMH volume varies significantly within diagnostic groups.

When determining how to group the disease categories for the binary classification case, we chose to combine HC and SCI subjects, who were either healthy or showed only subjective cognitive symptoms, with AD and MCI patients whose WMH volume was below 4820 mm^3. This formed the non-vascular class. The vascular class consisted of VCD and MCI patients with WMH volumes equal to or greater than 4820 mm^3. For the three-class classification case, AD and MCI patients were grouped into a separate class. In these groupings, we assumed that HC, SCI, and AD patients would typically have WMH volumes below the threshold, while VCD patients would have volumes above it. However, we acknowledge that there are some cases where the diagnoses do not follow this assumption. Some HC, SCI, and AD patients have a WMH volume greater than 4820 mm^3, and some VCD patients have a volume less than 4820 mm^3. For HC, there are 33 data instances with a WMH volume greater than 4820 mm^3, SCI has 42, and AD has 32. VCD (not including SSVD/AD) has 2 data instances lower than 4820 mm^3, and SSVD/AD has 15.

The division into five groups used in this case represents just one of many possible ways to categorize the different stages. We chose to divide the patients based on the clinical diagnoses established by physicians at the Memory Clinic. This approach was selected because the necessary diagnostic information was readily available, and it aligned best with our research focus on MRI imaging and vascular impairment. Other possible ways to categorize the disease groups could include stratifying based on the extent of WMH in the brain or incorporating relevant biomarkers that may influence the group distribution. Including biomarkers, for example, β-amyloid, could help clarify the classification of patients with the diagnosis SSVD/AD, making it easier to assign them to the appropriate category. However, stratifying based on WMH volume and then training the model to identify WMH introduces bias.

5.5 Explainabilty model

The decision to implement an explainability model was made to visualize the regions within the FLAIR MRI images from which the model retrieves information when making predictions or classifications. We wanted to assess whether the model retrieved information from regions where WMH was present, and to verify this, we implemented Guided Grad-CAM. We chose to use Guided Grad-CAM because it is one of the most intuitive visualization methods to interpret, and it was straightforward to implement using the Captum library. It does not require any ground truth segmentation masks of the target area to highlight relevant regions in the image. This was one of the most important reasons for choosing to use an explainability

model, as we did not have any segmentation masks available for the FLAIR MRI images. The only available segmentation masks were for T1 images, and using them would have required image registration.

The choice of the gamma value (0.8) and the threshold (0.05) was made after testing. The threshold value was effective in removing irrelevant areas such as the black background and noise, thereby allowing the visualization to focus on more important regions. A gamma value of 0.8 was chosen to enhance the visibility of meaningful patterns without overexposing the entire image. We tried various values for both gamma and the threshold. Ultimately, the values for gamma and the threshold were chosen because they provided the best balance and clarity in the resulting image.

We wanted the ability to select specific patients' FLAIR MRI images for visualization using Grad-CAM. This allowed us to compare how well the model highlighted the WMH areas in individuals with a high amount of WMH versus those with a low amount. It also enabled us to observe any differences in the resulting images between individuals with the clinical diagnoses of VCD, AD, or HC.

The results from the Guided Grad-CAM implementation varied. We obtained both satisfactory outcomes and cases where the results were not usable. In the unusable cases, the entire image was red, and therefore, we could not extract any information about where the model believed WMH was present. If a more precise method had been desired, a segmentation model would have been required.

The visualization results provided by Grad-CAM indicated that the MTL model lacked precision. In some cases, it showed that the model not only focused on regions within the brain but also included parts of the skull when generating predictions or classification. Although it typically focused on areas within the brain, these regions were often larger than the actual WMH areas observed in the FLAIR MRI images. Using three-class classification produced the most reliable Grad-CAM visualizations, with the highlighted regions predominantly located within the brain and minimal inclusion of the skull or areas outside the brain. The five-class classification case followed closely, with the MTL model still retrieving a substantial amount of information from within the brain. However, it also showed more of the skull and other non-relevant parts of the FLAIR image. The binary classification case yielded the poorest results; although the MTL model showed to retrieve some information from within the brain, most of the highlighted regions were located near the skull or adjacent areas. In some instances, regions around the eye sockets were also included.

The Grad-CAM algorithm does not provide a complete explanation of the model's decision-making process, as it highlights activations in specific layers based on the gradients, which may not fully capture all aspects of the model's reasoning. As a result, it is not possible to draw complete conclusions from the visualizations alone. While certain insights can be gained from the Grad-CAM heatmaps, they do not provide sufficient evidence to confirm or exclude the importance of specific features or regions. Even if a heatmap appears visually credible, this does not guarantee that

it accurately reflects the areas the model truly relies on when making predictions. This allows us to draw some conclusions, but we cannot definitively state that the model based its decisions solely on the specific regions highlighted in the heatmap.

5.6 Future work

To advance the identification of patients with VCD, or even differentiate between cognitive diseases, based on FLAIR images using ML, the next step is to explore a larger and more balanced dataset. Publicly available datasets may offer improvements over the limited and imbalanced dataset used in this project, as shown in Figure 3.1 and Figure 3.2. If continuing with the same dataset, efforts should focus on addressing the imbalance, potentially through improved data augmentation strategies or transfer learning. Although some augmentation has been applied, further exploration of both augmentation and other data expansion strategies is needed. Transfer learning offers an alternative by pre-training on a larger, publicly available dataset. This enables the model to learn general features and structures. The pre-trained model can then be fine-tuned on a smaller, task-specific dataset to capture the details relevant to the specific application. However, it is not certain that it is possible to expand the dataset or access a larger dataset, so we provide a future direction that could improve results without needing a larger dataset.

One such promising direction is the use of segmentation-based methods, on which several previous studies using ML to identify WMH are based. This is a promising lead for this project since the current dataset includes brain region masks that could serve as ground truth for segmentation tasks. Segmentation not only has the potential to provide more accurate WMH volume estimates but also enables spatial analysis of WMH distribution, which could provide valuable information for distinguishing between cognitive diseases. The working hypothesis is that WMH volume, location, and appearance vary between types of cognitive disease, and segmentation could support testing this hypothesis.

Although segmentation was one of the methods discussed at the beginning of this project, the need to align the T1-derived brain masks and the FLAIR images prevented its implementation. The fact that the brain masks were based on T1 images only became clear later in the project, and there was not enough time to implement the necessary image registration. If planned from the beginning, image registration could have been completed within 20 weeks. Therefore, segmentation is a promising objective for a future master's thesis project building on this work.

6

Conclusion

The main goal of this thesis was to investigate the possibility of using ML to identify patients with VCD, focusing on WMH-affected areas of the brain in FLAIR images. We showed that this is possible by implementing an MTL model, where the regression head of the binary classification model, predicting WMH volumes, achieved a ρ of 0.8612 and the classification head got an F1-score of 0.5806. However, the three-class classification and five-class classification did not perform as well, and it was not possible to differentiate between cognitive diseases to the same extent. Further investigations are needed on the cases using multi-classification, and one major aspect is solving the overfitting problem. Future work, therefore, includes expanding the dataset to both increase the size and improve class balance. Another aspect of future work is implementing image registration methods to be able to use segmentation to allow the model to better focus on WMH-affected areas of the brain, to hopefully improve classification.

Bibliography

[1] R. N. Kalaria, "The pathology and pathophysiology of vascular dementia," en, *Neuropharmacology*, vol. 134, pp. 226–239, May 2018, ISSN: 00283908. DOI: 10.1016/j.neuropharm.2017.12.030. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0028390817306275 (visited on 01/24/2025).

[2] E. Prost, *Global Deterioration Scale*, en, Sep. 2005. [Online]. Available: https://geriatrictoolkit.missouri.edu/cog/Global-Deterioration-Scale.pdf (visited on 01/27/2025).

[3] S. Bombois, S. Debette, A. Bruandet, *et al.*, "Vascular Subcortical Hyperintensities Predict Conversion to Vascular and Mixed Dementia in MCI Patients," en, *Stroke*, vol. 39, no. 7, pp. 2046–2051, Jul. 2008, ISSN: 0039-2499, 1524-4628. DOI: 10.1161/STROKEAHA.107.505206. [Online]. Available: https://www.ahajournals.org/doi/10.1161/STROKEAHA.107.505206 (visited on 05/06/2025).

[4] K. Meguro, H. Ishii, M. Kasuya, *et al.*, "Incidence of dementia and associated risk factors in Japan: The Osaki-Tajiri Project," en, *Journal of the Neurological Sciences*, vol. 260, no. 1-2, pp. 175–182, Sep. 2007, ISSN: 0022510X. DOI: 10.1016/j.jns.2007.04.051. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0022510X0700319X (visited on 05/06/2025).

[5] J. G. Merino, "White Matter Hyperintensities on Magnetic Resonance Imaging: What Is a Clinician to Do?" en, *Mayo Clinic Proceedings*, vol. 94, no. 3, pp. 380–382, Mar. 2019, ISSN: 00256196. DOI: 10.1016/j.mayocp.2019.01.016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0025619619300758 (visited on 01/24/2025).

[6] J. M. Wardlaw, M. C. Valdés Hernández, and S. Muñoz-Maniega, "What are White Matter Hyperintensities Made of?: Relevance to Vascular Cognitive Impairment," en, *Journal of the American Heart Association*, vol. 4, no. 6, e001140, Jun. 2015, ISSN: 2047-9980. DOI: 10.1161/JAHA.114.001140. [Online]. Available: https://www.ahajournals.org/doi/10.1161/JAHA.114.001140 (visited on 05/15/2025).

[7] *Nationella riktlinjer för vård och omsorg vid demenssjukdom: stöd för styrning och ledning*, sv. Stockholm: Socialstyrelsen, 2016, ISBN: 978-91-7555-433-4.

[8] M. Barbay, H. Taillia, C. Nedelec-Ciceri, *et al.*, "Vascular cognitive impairment: Advances and trends," en, *Revue Neurologique*, vol. 173, no. 7-8, pp. 473–480, Jul. 2017, ISSN: 00353787. DOI: 10.1016/j.neurol.2017.06.009.

[Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S003537871730468X (visited on 05/14/2025).

[9] A. Wallin, A. Nordlund, M. Jonsson, *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," en, *Journal of Cerebral Blood Flow & Metabolism*, vol. 36, no. 1, pp. 114–131, Jan. 2016, ISSN: 0271-678X, 1559-7016. DOI: 10.1038/jcbfm.2015.147. [Online]. Available: https://journals.sagepub.com/doi/10.1038/jcbfm.2015.147 (visited on 03/05/2025).

[10] Y. Mintz and R. Brodie, "Introduction to artificial intelligence in medicine," en, *Minimally Invasive Therapy & Allied Technologies*, vol. 28, no. 2, pp. 73–81, Mar. 2019, ISSN: 1364-5706, 1365-2931. DOI: 10.1080/13645706.2019.1575882. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/13645706.2019.1575882 (visited on 05/09/2025).

[11] A. Ebrahimi, S. Luo, and R. 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)*, Wellington, New Zealand: IEEE, Nov. 2020, pp. 1–6, ISBN: 978-1-7281-8579-8. DOI: 10.1109/IVCNZ51579.2020.9290616. [Online]. Available: https://ieeexplore.ieee.org/document/9290616/ (visited on 05/09/2025).

[12] J. Feng, D. Hui, Q. Zheng, *et al.*, "Automatic detection of cognitive impairment in patients with white matter hyperintensity and causal analysis of related factors using artificial intelligence of MRI," *Computers in Biology and Medicine*, vol. 178, p. 108 684, Aug. 2024, ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2024.108684. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482524007698 (visited on 01/28/2025).

[13] M. Ghafoorian, N. Karssemeijer, T. Heskes, *et al.*, "Location Sensitive Deep Convolutional Neural Networks for Segmentation of White Matter Hyperintensities," en, *Scientific Reports*, vol. 7, no. 1, p. 5110, Jul. 2017, ISSN: 2045-2322. DOI: 10.1038/s41598-017-05300-5. [Online]. Available: https://www.nature.com/articles/s41598-017-05300-5 (visited on 05/09/2025).

[14] Y. Zhang, Y. Duan, X. Wang, *et al.*, "A deep learning algorithm for white matter hyperintensity lesion detection and segmentation," en, *Neuroradiology*, vol. 64, no. 4, pp. 727–734, Apr. 2022, ISSN: 1432-1920. DOI: 10.1007/s00234-021-02820-w. [Online]. Available: https://doi.org/10.1007/s00234-021-02820-w (visited on 01/28/2025).

[15] G. Murdaca, S. Banchero, M. Casciaro, *et al.*, "Potential Predictors for Cognitive Decline in Vascular Dementia: A Machine Learning Analysis," en, *Processes*, vol. 10, no. 10, p. 2088, Oct. 2022, ISSN: 2227-9717. DOI: 10.3390/pr10102088. [Online]. Available: https://www.mdpi.com/2227-9717/10/10/2088 (visited on 01/28/2025).

[16] C. Carrarini, C. Nardulli, L. Titti, *et al.*, "Neuropsychological and electrophysiological measurements for diagnosis and prediction of dementia: A review on Machine Learning approach," *Ageing Research Reviews*, vol. 100, p. 102 417, Sep. 2024, ISSN: 1568-1637. DOI: 10.1016/j.arr.2024.102417. [Online].

Available: https://www.sciencedirect.com/science/article/pii/S1568163724002356 (visited on 01/28/2025).

[17] C. A. Lane, J. Hardy, and J. M. Schott, "Alzheimer's disease," en, *European Journal of Neurology*, vol. 25, no. 1, pp. 59–70, Jan. 2018, ISSN: 1351-5101, 1468-1331. DOI: 10.1111/ene.13439. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/ene.13439 (visited on 03/28/2025).

[18] A. Gustavsson, N. Norton, T. Fast, *et al.*, "Global estimates on the number of persons across the Alzheimer's disease continuum," en, *Alzheimer's & Dementia*, vol. 19, no. 2, pp. 658–670, Feb. 2023, ISSN: 1552-5260, 1552-5279. DOI: 10.1002/alz.12694. [Online]. Available: https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.12694 (visited on 04/03/2025).

[19] C. H. Van Dyck, C. J. Swanson, P. Aisen, *et al.*, "Lecanemab in Early Alzheimer's Disease," en, *New England Journal of Medicine*, vol. 388, no. 1, pp. 9–21, Jan. 2023, ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMoa2212948. [Online]. Available: http://www.nejm.org/doi/10.1056/NEJMoa2212948 (visited on 04/29/2025).

[20] European Medicines Agency, *Leqembi recommended for treatment of early Alzheimer's disease*, en, Nov. 2024. [Online]. Available: https://www.ema.europa.eu/en/news/leqembi-recommended-treatment-early-alzheimers-disease (visited on 05/05/2025).

[21] J. Cummings, L. Apostolova, G. D. Rabinovici, *et al.*, "Lecanemab: Appropriate Use Recommendations," eng, *The Journal of Prevention of Alzheimer's Disease*, vol. 10, no. 3, pp. 362–377, 2023, ISSN: 2426-0266. DOI: 10.14283/jpad.2023.30.

[22] F. J. Wolters and M. A. Ikram, "Epidemiology of Vascular Dementia: Nosology in a Time of Epiomics," en, *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 39, no. 8, pp. 1542–1549, Aug. 2019, ISSN: 1079-5642, 1524-4636. DOI: 10.1161/ATVBAHA.119.311908. [Online]. Available: https://www.ahajournals.org/doi/10.1161/ATVBAHA.119.311908 (visited on 04/03/2025).

[23] P. Prajjwal, M. D. M. Marsool, P. Inban, *et al.*, "Vascular dementia subtypes, pathophysiology, genetics, neuroimaging, biomarkers, and treatment updates along with its association with Alzheimer's dementia and diabetes mellitus," en, *Disease-a-Month*, vol. 69, no. 5, p. 101 557, May 2023, ISSN: 00115029. DOI: 10.1016/j.disamonth.2023.101557. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0011502923000378 (visited on 04/03/2025).

[24] T. Erkinjuntti, "Vascular Dementia: Challenge of Clinical Diagnosis," en, *International Psychogeriatrics*, vol. 9, pp. 51–58, Dec. 1997, ISSN: 10416102. DOI: 10.1017/S1041610297004699. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1041610224056400 (visited on 01/27/2025).

[25] Q. Cao, C.-C. Tan, W. Xu, *et al.*, "The Prevalence of Dementia: A Systematic Review and Meta-Analysis," en, *Journal of Alzheimer's Disease*, vol. 73, no. 3, L.-Q. Zhu, Ed., pp. 1157–1166, Feb. 2020, ISSN: 1387-2877, 1875-8908. DOI: 10.3233/JAD-191092. [Online]. Available: https://journals.sagepub.com/doi/10.3233/JAD-191092 (visited on 04/09/2025).

[26] N. Custodio, R. Montesinos, D. Lira, E. Herrera-Pérez, Y. Bardales, and L. Valeriano-Lorenzo, "Mixed dementia: A review of the evidence," *Dementia & Neuropsychologia*, vol. 11, no. 4, pp. 364–370, Dec. 2017, ISSN: 1980-5764, 1980-5764. DOI: 10.1590/1980-57642016dn11-040005. [Online]. Available: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1980-57642017000400364&lng=en&tlng=en (visited on 04/03/2025).

[27] E. Basic, F. Locatelli, M. Jonsson, *et al.*, *Reclassification of manifest disease diagnoses using Alzheimer's disease biomarkers in the Gothenburg mild cognitive impairment study*, International Conference on Alzheimer's and Parkinson's Diseases and related neurological disorders, Gothenburg, 2023.

[28] T. Eichler, J. R. Thyrian, J. Hertel, *et al.*, "Rates of Formal Diagnosis in People Screened Positive for Dementia in Primary Care: Results of the DelpHi-Trial," *Journal of Alzheimer's Disease*, vol. 42, no. 2, pp. 451–458, Aug. 2014, ISSN: 18758908, 13872877. DOI: 10.3233/JAD-140354. [Online]. Available: https://journals.sagepub.com/doi/full/10.3233/JAD-140354 (visited on 01/27/2025).

[29] M. Prince, D. R. Bryce, and D. C. Ferri, "World Alzheimer Report 2011: The benefits of early diagnosis and intervention," en, *Alzheimer's Disease International*, p. 12, Sep. 2011. [Online]. Available: https://www.alzint.org/u/WorldAlzheimerReport2011.pdf (visited on 01/27/2025).

[30] F. Fazekas, J. Chawluk, A. Alavi, H. Hurtig, and R. Zimmerman, "MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging," en, *American Journal of Roentgenology*, vol. 149, no. 2, pp. 351–356, Aug. 1987, Publisher: American Roentgen Ray Society, ISSN: 0361-803X, 1546-3141. DOI: 10.2214/ajr.149.2.351. [Online]. Available: https://www.ajronline.org/doi/10.2214/ajr.149.2.351 (visited on 05/23/2025).

[31] K. He, X. Zhang, S. Ren, and J. Sun, *Deep Residual Learning for Image Recognition*, Version Number: 1, 2015. DOI: 10.48550/ARXIV.1512.03385. [Online]. Available: https://arxiv.org/abs/1512.03385 (visited on 05/09/2025).

[32] F. Ramzan, M. U. G. Khan, A. Rehmat, *et al.*, "A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer's Disease Stages Using Resting-State fMRI and Residual Neural Networks," en, *Journal of Medical Systems*, vol. 44, no. 2, p. 37, Feb. 2020, ISSN: 0148-5598, 1573-689X. DOI: 10.1007/s10916-019-1475-2. [Online]. Available: http://link.springer.com/10.1007/s10916-019-1475-2 (visited on 05/09/2025).

[33] C. M. Bishop, *Pattern recognition and machine learning* (Information science and statistics), eng. New York: Springer, 2006, ISBN: 978-0-387-31073-2.

[34] J. Terven, D.-M. Cordova-Esparza, J.-A. Romero-González, A. Ramírez-Pedraza, and E. A. Chávez-Urbiola, "A comprehensive survey of loss functions and metrics in deep learning," en, *Artificial Intelligence Review*, vol. 58, no. 7, p. 195, Apr. 2025, ISSN: 1573-7462. DOI: 10.1007/s10462-025-11198-7. [Online]. Available: https://link.springer.com/10.1007/s10462-025-11198-7 (visited on 05/09/2025).

[35] J. Terven, D. M. Cordova-Esparza, A. Ramirez-Pedraza, E. A. Chavez-Urbiola, and J. A. Romero-Gonzalez, "Loss Functions and Metrics in Deep Learning," 2023, Publisher: arXiv Version Number: 5. DOI: 10.48550/ARXIV.2307.

02694. [Online]. Available: https://arxiv.org/abs/2307.02694 (visited on 05/09/2025).

[36] A. Mao, M. Mohri, and Y. Zhong, *Cross-Entropy Loss Functions: Theoretical Analysis and Applications*, Version Number: 2, 2023. DOI: 10.48550/ARXIV.2304.07288. [Online]. Available: https://arxiv.org/abs/2304.07288 (visited on 05/06/2025).

[37] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An Introduction to Statistical Learning: with Applications in Python* (Springer Texts in Statistics), en. Cham: Springer International Publishing, 2023, ISBN: 978-3-031-38747-0. DOI: 10.1007/978-3-031-38747-0. [Online]. Available: https://link.springer.com/10.1007/978-3-031-38747-0 (visited on 05/06/2025).

[38] W. Liu, L. Chen, and Y. Chen, "Age Classification Using Convolutional Neural Networks with the Multi-class Focal Loss," *IOP Conference Series: Materials Science and Engineering*, vol. 428, p. 012 043, Oct. 2018, ISSN: 1757-899X. DOI: 10.1088/1757-899X/428/1/012043. [Online]. Available: https://iopscience.iop.org/article/10.1088/1757-899X/428/1/012043 (visited on 05/06/2025).

[39] X. Ying, "An Overview of Overfitting and its Solutions," *Journal of Physics: Conference Series*, vol. 1168, p. 022 022, Feb. 2019, ISSN: 1742-6588, 1742-6596. DOI: 10.1088/1742-6596/1168/2/022022. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1168/2/022022 (visited on 05/07/2025).

[40] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," en, *Array*, vol. 16, p. 100 258, Dec. 2022, ISSN: 25900056. DOI: 10.1016/j.array.2022.100258. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2590005622000911 (visited on 05/06/2025).

[41] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," en, *Journal of Big Data*, vol. 6, no. 1, p. 60, Dec. 2019, ISSN: 2196-1115. DOI: 10.1186/s40537-019-0197-0. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0 (visited on 05/06/2025).

[42] J. Bjorck, C. Gomes, B. Selman, and K. Q. Weinberger, *Understanding Batch Normalization*, Version Number: 4, 2018. DOI: 10.48550/ARXIV.1806.02375. [Online]. Available: https://arxiv.org/abs/1806.02375 (visited on 05/27/2025).

[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pmlr, 2015, pp. 448–456.

[44] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014, ISBN: 1532-4435 Publisher: JMLR. org.

[45] K. Sanjar, A. Rehman, A. Paul, and K. JeongHong, "Weight Dropout for Preventing Neural Networks from Overfitting," in *2020 8th International Conference on Orange Technology (ICOT)*, Daegu, Korea (South): IEEE, Dec. 2020,

pp. 1–4, ISBN: 978-1-6654-1852-2. DOI: 10.1109/ICOT51877.2020.9468799. [Online]. Available: https://ieeexplore.ieee.org/document/9468799/ (visited on 05/07/2025).

[46] K. You, M. Long, J. Wang, and M. I. Jordan, *How Does Learning Rate Decay Help Modern Neural Networks?* Version Number: 2, 2019. DOI: 10.48550/ARXIV.1908.01878. [Online]. Available: https://arxiv.org/abs/1908.01878 (visited on 05/07/2025).

[47] G. Zhang, C. Wang, B. Xu, and R. Grosse, *Three Mechanisms of Weight Decay Regularization*, Version Number: 1, 2018. DOI: 10.48550/ARXIV.1810.12281. [Online]. Available: https://arxiv.org/abs/1810.12281 (visited on 05/08/2025).

[48] A. Krogh and J. Hertz, "A simple weight decay can improve generalization," *Advances in neural information processing systems*, vol. 4, 1991.

[49] C.-B. Zhang, P.-T. Jiang, Q. Hou, *et al.*, "Delving Deep Into Label Smoothing," *IEEE Transactions on Image Processing*, vol. 30, pp. 5984–5996, 2021, ISSN: 1057-7149, 1941-0042. DOI: 10.1109/TIP.2021.3089942. [Online]. Available: https://ieeexplore.ieee.org/document/9464693/ (visited on 05/08/2025).

[50] L. Prechelt, "Early Stopping - But When?" In *Neural Networks: Tricks of the Trade*, G. Goos, J. Hartmanis, J. Van Leeuwen, G. B. Orr, and K.-R. Müller, Eds., vol. 1524, Series Title: Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 55–69, ISBN: 978-3-540-49430-0. DOI: 10.1007/3-540-49430-8_3. [Online]. Available: http://link.springer.com/10.1007/3-540-49430-8_3 (visited on 05/07/2025).

[51] R. Caruana, "Multitask Learning*," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997, ISSN: 08856125. DOI: 10.1023/A:1007379606734. [Online]. Available: http://link.springer.com/10.1023/A:1007379606734 (visited on 05/06/2025).

[52] S. Ruder, *An Overview of Multi-Task Learning in Deep Neural Networks*, Version Number: 1, 2017. DOI: 10.48550/ARXIV.1706.05098. [Online]. Available: https://arxiv.org/abs/1706.05098 (visited on 05/06/2025).

[53] Y. Zhang and Q. Yang, "A Survey on Multi-Task Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5586–5609, Dec. 2022, ISSN: 1041-4347, 1558-2191, 2326-3865. DOI: 10.1109/TKDE.2021.3070203. [Online]. Available: https://ieeexplore.ieee.org/document/9392366/ (visited on 05/13/2025).

[54] P. L. Fung, M. A. Zaidan, H. Timonen, *et al.*, "Evaluation of white-box versus black-box machine learning models in estimating ambient black carbon concentration," en, *Journal of Aerosol Science*, vol. 152, p. 105 694, Feb. 2021, ISSN: 00218502. DOI: 10.1016/j.jaerosci.2020.105694. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0021850220301798 (visited on 05/09/2025).

[55] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," en, *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Feb. 2020, ISSN: 0920-5691, 1573-1405. DOI: 10.1007/s11263-

019-01228-7. [Online]. Available: http://link.springer.com/10.1007/s11263-019-01228-7 (visited on 05/09/2025).

[56] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, *Striving for Simplicity: The All Convolutional Net*, arXiv:1412.6806 [cs], Apr. 2015. DOI: 10.48550/arXiv.1412.6806. [Online]. Available: http://arxiv.org/abs/1412.6806 (visited on 05/09/2025).

[57] V. Plevris, G. Solorzano, N. Bakas, and M. Ben Seghier, "Investigation of performance metrics in regression analysis and machine learning-based prediction models," in *8th European Congress on Computational Methods in Applied Sciences and Engineering*, CIMNE, 2022. DOI: 10.23967/eccomas.2022.155. [Online]. Available: https://www.scipedia.com/public/Plevris_et_al_2022a (visited on 05/08/2025).

[58] N. Dendukuri and C. Reinhold, "Correlation and Regression," en, *American Journal of Roentgenology*, vol. 185, no. 1, pp. 3–18, Jul. 2005, ISSN: 0361-803X, 1546-3141. DOI: 10.2214/ajr.185.1.01850003. [Online]. Available: https://www.ajronline.org/doi/10.2214/ajr.185.1.01850003 (visited on 05/08/2025).

[59] J. Fernando, *R-Squared: Definition, Calculation, and Interpretation*, en, Nov. 2024. [Online]. Available: https://www.investopedia.com/terms/r/r-squared.asp (visited on 04/11/2025).

[60] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto, and S. Ridella, "The'K'in K-fold Cross Validation.," in *ESANN*, vol. 102, 2012, pp. 441–446.

[61] I. Guyon, A. Saffari, G. Dror, and G. Cawley, "Model Selection: Beyond the Bayesian/Frequentist Divide.," *Journal of Machine Learning Research*, vol. 11, no. 1, 2010, ISBN: 1532-4435.

[62] K. S. Tang and L. Fallqvist, "Machine learning techniques to understand cognitive decline," 2022. [Online]. Available: https://hdl.handle.net/20.500.12380/304706.

[63] F. Krones, U. Marikkar, G. Parsons, A. Szmul, and A. Mahdi, "Review of multimodal machine learning approaches in healthcare," en, *Information Fusion*, vol. 114, p. 102 690, Feb. 2025, ISSN: 15662535. DOI: 10.1016/j.inffus.2024.102690. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1566253524004688 (visited on 05/15/2025).

[64] nipy.org, *NiBabel*. [Online]. Available: https://nipy.org/nibabel/ (visited on 05/15/2025).

[65] MONAI, *Medical Open Network for Artificial Intelligence*. [Online]. Available: https://monai.io/ (visited on 05/15/2025).

[66] M. Tipping, "The Relevance Vector Machine," in *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K. Müller, Eds., vol. 12, MIT Press, 1999. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1999/file/f3144cefe89a60d6a1afaf7859c5076b-Paper.pdf.

[67] Y. LeCun, B. Boser, J. Denker, *et al.*, "Handwritten Digit Recognition with a Back-Propagation Network," in *Advances in Neural Information Processing Systems*, D. Touretzky, Ed., vol. 2, Morgan-Kaufmann, 1989. [Online]. Avail-

able: https://proceedings.neurips.cc/paper_files/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf.

[68] O. Ronneberger, P. Fischer, and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, arXiv:1505.04597 [cs], May 2015. DOI: 10.48550/arXiv.1505.04597. [Online]. Available: http://arxiv.org/abs/1505.04597 (visited on 05/15/2025).

[69] NumPy, *Numpy.corrcoef*. [Online]. Available: https://numpy.org/doc/2.2/reference/generated/numpy.corrcoef.html (visited on 06/04/2025).

[70] SciPy, *Scipy.stats. spearmanr*. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html (visited on 06/04/2025).

[71] scikit-learn, *Sklearn.metrics*. [Online]. Available: https://scikit-learn.org/stable/api/sklearn.metrics.html (visited on 06/04/2025).

[72] PyTorch, *BCEWithLogitsLoss*. [Online]. Available: https://docs.pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html (visited on 05/28/2025).

[73] Captum, *Guided GradCAM*. [Online]. Available: https://captum.ai/api/guided_grad_cam.html (visited on 05/15/2025).

A

Appendix 1

A.1 Data augmentation ablation study for focal loss

	Accuracy	F1-score	Precision	Recall
All augmentations	0.2605	0.1470	0.1724	0.2605
No augmentations	0.3109	0.2806	0.3392	0.3109
No Gaussian noise	0.2941	0.2115	0.2005	0.2941
No scaling	0.2941	0.1982	0.2745	0.2941
No rotation	0.2857	0.2033	0.2778	0.2857
No contrast adjustment	0.3277	0.2419	0.2664	0.3277
No Gaussian smoothing	0.2521	0.1022	0.0641	0.2521
No translation	0.2773	0.2285	0.2135	0.2773

Table A.1: The validation set results of the data augmentation ablation study using focal loss using five-class classification.

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY