THE INFLUENCE OF VASCULAR PATHOLOGY ON MACHINE LEARNING METHODS FOR ALZHEIMER'S DISEASE

Myrthe van Haaften

October 2022

Erasmus MC

Page intentionally left blank

THE INFLUENCE OF VASCULAR PATHOLOGY ON MACHINE LEARNING METHODS FOR ALZHEIMER'S DISEASE

MSc thesis

Myrthe van Haaften

Student number : 4547470

11-10-2022

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University ; Delft University of Technology ; Erasmus University Rotterdam

Master thesis project (TM30004 ; 35 ECTS)

Dept. of Radiology and Nuclear Medicine, Dept. of Neurology, ERASMUS MC

January 2022 – October 2022

Supervisors:

Dr. Esther Bron

Prof. dr. Meike Vernooij

Dr. Harro Seelaar

Thesis committee members:

Prof. dr. Meike Vernooij, Erasmus MC (chair)

Dr. Esther Bron, Erasmus MC

Dr. Harro Seelaar, Erasmus MC

Dr. Berend Stoel, LUMC

Dr. Inge de Kok, Erasmus MC

Page intentionally left blank

Summary

Many machine learning models have been developed to aid in the diagnosis of dementia, to predict dementia risk and to determine cognitive performance. While it is well known that vascular pathology is a critical contributing factor to dementia, cerebral small vessel disease is often not addressed by these methods, possibly causing impeded generalizability of the models from research setting to clinical practice. The aim of this thesis is to evaluate whether vascular pathology, as represented by white matter hyperintensities (WMHs), influences the outcome of machine learning prediction models for dementia and as such hampers generalizability of these models from the research setting to clinical practice.

To answer this research question, we developed a convolutional neural network (CNN) for prediction of cognitive performance measured by MMSE and ADAS13. We used ADNI data and included all subject timepoints that adhered to our inclusion criteria (N = 4846). Next, we derived gray matter density maps (GMDMs) and WMH density maps (WMHDMs) from T1 and FLAIR MRI scans using voxel-based morphometry. Six CNN models were implemented, differing in input: model 1 (GMDM), model 2 (GMDM, age, sex), model 3 (WMHDM), model 4 (WMHDM, age, sex), model 5 (GMDM, WMHDM, age, sex) and model 6 (GMDM, log(WMH ratio), age, sex). WMH ratio was defined as WMH volume corrected for intracranial volume. We split the data into a low, middle and high WMH load group based on WMH ratio tertiles and performed two sets of experiments. First, we trained all models on the low WMH load group with the intention of testing on the high WMH load group as a measure of generalizability. Second, we trained model 2 and 5 on data from all WMH load groups. For these sets of experiments, we split both the low WMH load group and all WMH load groups into a train, validation and test set.

The models trained on the low WMH load group obtained a similar performance to prediction of a constant value and prediction by a model developed on randomized cognitive scores. These models likely suffered from the underrepresentation of lower cognitive performance data (i.e. high ADAS13 scores and low MMSE scores). This is supported by the fact that model 2 and 5 outperformed constant value prediction when trained on data from all WMH load groups. The current models did in general not benefit from adding WMHDM input to the model, while they did benefit slightly from log(WMH ratio) information. However, we were unable to answer our research question due to the suboptimal performance of the models. Our future research will focus on revisiting our research question through brain age prediction in order to improve generalizability of ML methods for AD to clinical populations with prominent vascular pathology.

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
AD	Alzheimer’s disease
ADAS-Cog	Alzheimer’s disease assessment scale-cognitive subscale
ADNI	Alzheimer’s disease neuroimaging initiative
AUC	Area under the curve
CDR	Clinical dementia rating
CN	Cognitively normal
CNN	Convolutional neural network
CSF	Cerebrospinal fluid
FC	Fully connected
FLAIR	Fluid-attenuated inversion recovery
GM	Gray matter
GMDM	Gray matter density map
ICV	Intracranial volume
MAE	Mean absolute error
MCI	Mild cognitive impairment
ML	Machine learning
MMSE	Mini-mental state examination
MNI	Montreal neurological institute
MoCA	Montreal cognitive assessment
MRI	Magnetic resonance imaging
MSLE	Mean squared logarithmic error
SVD	Small vessel disease
VaD	Vascular dementia
VBM	Voxel-based morphometry
WM	White matter
WMHs	White matter hyperintensites
WMHDM	White matter hyperintensity density map

1 Introduction

Dementia is generally defined as cognitive decline that is severe enough to impair occupational, domestic, or social functioning [1]. The most common type of dementia is Alzheimer's disease (AD), which has a prevalence of around 50 million patients worldwide [2]. AD is a slowly progressive neurodegenerative disease characterized by neuritic plaques and neurofibrillary tangles. Current therapies manage to relieve AD symptoms, but no curative treatment has been developed yet. On radiological imaging, the typical characteristic of AD is cortical atrophy, which is caused by neuronal degeneration. Often, the hippocampus is most severely affected [3]. Following AD, vascular dementia (VaD) is considered to be the second most common type of dementia [4]. VaD has ischemic and hemorrhagic etiologies that damage the cerebral blood vessels. VaD subtypes include multi-infarct dementia, strategic infarct dementia and subcortical ischemic VaD. Other common types of dementia include Lewy body dementia and frontotemporal dementia.

Early diagnosis of dementia and its underlying diseases is highly important for delivering the right care to the right patients. Furthermore, early diagnosis is crucial in the development of treatments, as it is expected that late treatment plays a major role in the failure of clinical trials [5]. Yet, diagnosing dementia is challenging and often takes multiple years [6, 7]. Currently, the definite diagnosis of AD and other dementia types can only be made by post mortem pathological evaluation of brain tissue [8]. In clinical practice, a multimodal approach is used to come to a diagnosis, possibly including neuropsychological examination, imaging, cerebrospinal fluid (CSF) analysis and/or blood tests.

Another factor complicating correct diagnosis in clinical practice is mixed pathology. Boyle et al. [9] analysed brain autopsy data of two longitudinal cohorts (N = 1079) consisting of residents of retirement communities and catholic clergy. Of all samples, 9% had pure AD, while 40% showed both AD and vascular pathology, and 44% exhibited AD in combination with both vascular and additional neurodegenerative pathology [9, 10]. In clinical practice, about 30 - 40% of the late onset dementia patients present with a combination of Alzheimer pathology and vascular pathology [11]. This specific combination of AD and vascular pathology is often referred to as mixed dementia. Cerebral small vessel disease (SVD) is the most common vascular cause of dementia and a major contributor to mixed dementia [12, 13]. MRI characteristics of SVD include white matter hyperintensities (WMHs) [13], which are white matter lesions that are hyperintense on T2 weighted MRI. Their presence and volume increase with age [14, 15] and they are associated with a number of disorders and conditions such as stroke, depression and dementia [3]. There is increasing evidence indicating that cerebrovascular pathology, including SVD, is a critical contributing factor to dementia, and that dementia is a result of synergistic interactions between vascular and neurodegenerative pathology [10]. Even though the importance of addressing mixed dementia has become clear, there is no consensus on diagnostic criteria and there is a lack of studies addressing this dementia type [16].

Many machine learning (ML) prediction models have been developed to aid in the diagnosis of dementia. These models extract features from high dimensional data to determine the diagnosis of an individual patient or even predict conversion from mild cognitive impairment (MCI), a prodromal dementia stage, to dementia. Most ML prediction models are imaging based and developed solely for AD. They show a good performance in distinguishing AD patients from cognitively normal (CN) individuals in research setting, with area under the curve (AUC) values of 85% - 98% [6]. Lower performances are obtained in predicting conversion from MCI to AD, with AUC values of 62% - 82% [6]. In addition to predicting diagnosis and MCI conversion, models have been developed for the determination of cognitive scores, e.g. the mini-mental state examination (MMSE) and Alzheimer's disease assessment scale-cognitive subscale (ADAS-Cog) [17, 18, 19]. A convolutional neural network (CNN) is a popular model often used for these classification and prediction tasks, which belongs to a ML subdomain called deep learning.

It is expected that the performance of the previously described ML prediction models is substantially lower in clinical groups compared to the highly selected research populations in which they were developed. This is related to the generalizability of the model, which measures how well the model performs (generalizes to) new, unseen data. Currently, most features are gray matter (GM) atrophy related and extracted from T1 weighted brain MRI scans. Vascular pathology, including SVD, is often not taken into account by these models. By ignoring the role that vascular pathology plays in the development of dementia and its contribution to cognitive impairment at the time of diagnosis, and by therefore also ignoring mixed dementia, these models are expected to have an impeded generalizability from research setting to clinical practice.

Therefore, the aim of this thesis is to evaluate whether vascular pathology, as represented by WMHs, influences the outcome of ML prediction models for dementia and as such hampers generalizability of these models from the research setting to clinical practice. More specifically, in this thesis, the influence of WMHs on the prediction of cognitive functioning by a CNN is evaluated. We hypothesize that the performance and the generalizability of the model is better when the model is able to use both GM atrophy related input and WMH related input, compared to a model using only GM atrophy related input. Secondly, we hypothesize that the model is able to derive predictive information from the WMH related input when no GM atrophy related information is supplied. MMSE and ADAS-Cog-13 are used as measures of cognitive functioning.

The rest of this thesis is structured as follows. Chapter 2 outlines the methodology of the thesis. Chapter 3 in turn describes the results and Chapter 4 provides the discussion of the results. In Chapter 5, the conclusion of the thesis is provided. More elaborate background information on machine learning and CNNs is provided in Appendix A.

2 Methods

2.1. Dataset

Data used in the preparation of this thesis were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni-info.org). This database consists of four data cohorts: ADNI 1, ADNI GO, ADNI 2 and ADNI 3. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. ADNI includes the full AD dementia spectrum, represented by the following diagnostic groups:

1. Cognitively normal: subjects reporting no memory complaints;
2. Subjective memory complaints: subjects reporting cognitive symptoms or complaints, while objective psychometric testing shows no clear impairment [20];
3. Mild cognitive impairment: subjects with objectifiable cognitive impairment without impairment of daily activities [21];
4. Dementia, including AD.

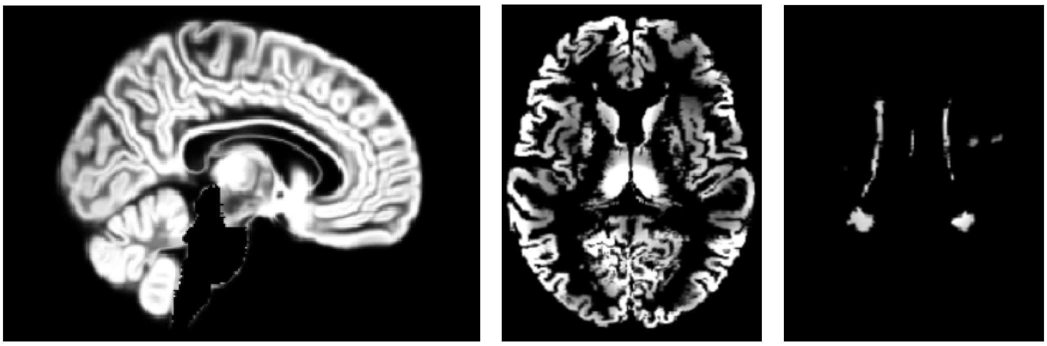
As ADNI only actively includes AD dementia subjects, other etiologies were present only in small numbers, i.e. through MCI converter subjects. Mixed dementia is not a separate diagnostic category in ADNI. In this study, we included subject timepoints if a T1 and FLAIR brain MRI as well as an MMSE and/or ADAS-Cog 13, further referred to as ADAS13, were available. FLAIR scans with a quality score of 4 (termed 'unusable' by ADNI) were not taken into account. Based on these criteria, ADNI 1 data was not used as it included FLAIR scans for only a handful of subjects. For each subject timepoint, we extracted the following data from the ADNI database: demographic information (e.g. age, sex), diagnostic information, image data (T1 and FLAIR brain MRI) and cognitive scores (MMSE and ADAS13).

2.2. Image processing

We processed the T1 and FLAIR scans using a voxel-based morphometry (VBM) pipeline to obtain GM and WMH density maps (GMDMs and WMHDMs). GMDMs were used instead of raw MRI scans, as the study of Bron et al. [6] showed higher model performance using GMDMs.

As a preprocessing step, we segmented the scans into GM, white matter (WM), CSF and WMHs. A brain tissue and WMH segmentation method implemented by Quantib was used. This method is based on the method described by de Boer et al. [22] and undertook the following steps using the T1 and FLAIR MRI as the input:

1. Segmentation of GM, WM and CSF on the T1 scan using an atlas-based k-nearest neighbor classifier.
2. Segmentation of WMHs by thresholding the FLAIR scan using automatic threshold calculation based on the histogram of all GM-labeled FLAIR voxels.



(a) Sagittal slice of gray matter template

(b) Axial slice of example GMDM

(c) Axial slice of example WMHDM, showing periventricular hyperintensities

Figure 2.1: Overview of voxel-based morphometry gray matter template and results. GMDM = Gray matter density map, WMHDM = White matter hyperintensity density map.

3. Removal of false positive WMH segmentations located outside the WM based on the tissue type of the neighboring voxels.

The output of this pipeline were the GM, WM, CSF and WMH segmentations and the WMH volume. The second step of the image processing extracted the GMDMs using a previously developed VBM pipeline consisting of the following steps [23]:

1. Nonlinear registration of the GM segmentations to the nonlinear symmetric standard Montreal Neurological Institute (MNI) GM template with a 1 × 1 × 1 mm voxel resolution. FSL version 5.0.2.2 was used. As the Quantib GM segmentation did not include the brainstem, we removed the brainstem from the GM template using the CerebrA atlas [24]. A remnant outline of the brainstem was present after applying the atlas, but the template was deemed to be of good enough quality for the intended use.
2. Correction of registration-related differences in the absolute GM volume using a spatial modulation procedure. For this, the voxel density values were multiplied by the Jacobian determinants estimated in the previous step.

We performed an automated quality check of the resulting GMDMs to identify density maps with outlier voxels based on their aberrant intensity. Lastly, the registration matrix and Jacobian determinants from the GM VBM results were applied to the WMH segmentations to generate WMHDMs in the same template space as the GMDMs. See Figure 2.1 for an overview of the GM template and VBM results.

2.3. CNN models

A CNN is a deep learning model that is able to automatically extract features from 2D and 3D data structures [25]. The convolution layers extract feature maps, which are then used by fully connected layers to compute the output of the network. For this thesis, we selected a CNN as the model type as CNNs are able to derive useful features from the data itself, meaning that the model is not restricted to pre-extracted features. We used the CNN architecture of the previously developed brain age model [23] as we expected this model to be suitable for cognitive score prediction. First, the model also uses GMDMs as input. Second, the model was developed for brain age prediction, which is closely linked with cognitive performance. The network architecture includes four convolutional blocks ending in a pooling layer, global average pooling and fully connected layers that propagate to a single regression output.

As measures of cognitive functioning, we used MMSE and ADAS13. The MMSE [26] is commonly used in the diagnosis of cerebral neurodegenerative diseases [27]. The maximum number of points is 30 and a higher score corresponds to better cognitive functioning. Contrarily to the MMSE, the ADAS13 is an AD-specific scale often used to measure the effects of antidementia treatments [28]. The scale runs from 0 to 85, with higher scores denoting more severe cognitive impairment. A more elaborate overview of the scores can be found in Appendix B. We considered MMSE as the main outcome, as this

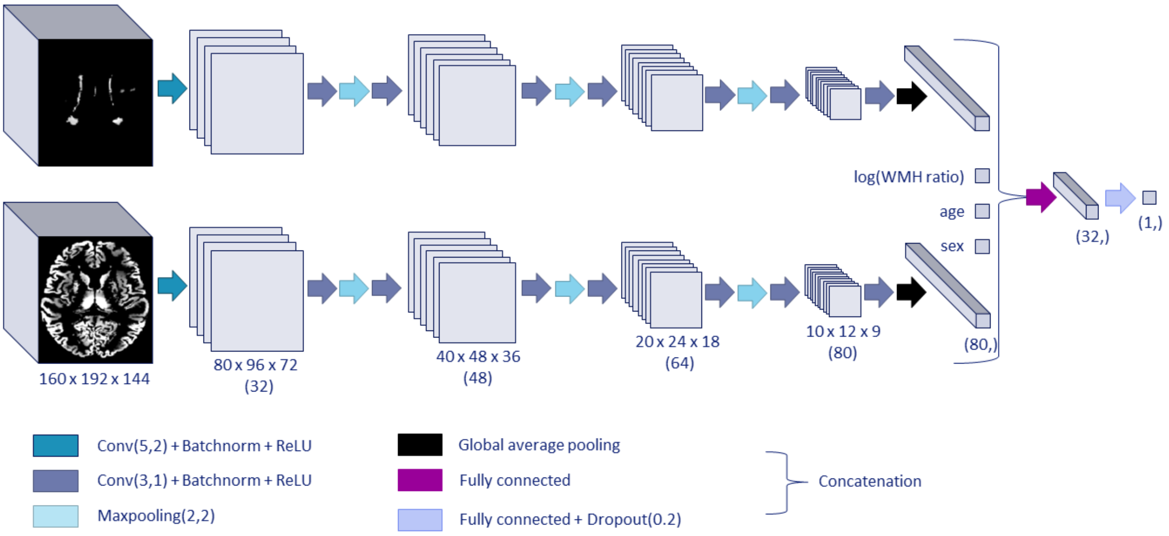


Figure 2.2: Total convolutional neural network architecture depicting all possible inputs, adapted from [23]. There are eight convolutional layers in total. Depending on the model, some inputs are absent. The imaging features and input from clinical variables (age, sex and/or log(WMH ratio)) are combined by concatenation. For the models using only imaging data (GMDM or WMHDM), the concatenation layer is absent, and the imaging features are directly passed on to the first fully connected layer. GMDM = Gray matter density map, WMH = White matter hyperintensity, WMHDM = White matter hyperintensity density map.

scale is widely used in clinical practice. ADAS13 served as an additional outcome measure to evaluate the influence of the exact cognitive functioning measure on the results. To evaluate the influence of WMHs on the prediction of cognitive functioning, CNN models were developed that differed in input:

1. GMDM;
2. GMDM, age and sex;
3. WMHDM;
4. WMHDM, age and sex;
5. GMDM, WMHDM, age and sex;
6. GMDM, log transformed WMH ratio, age and sex.

The WMH ratio is defined as $\frac{WMHvolume}{ICV}$, with ICV as the intracranial volume. The WMHDMs as well as WMH ratio were separately used as inputs to evaluate the difference between providing information solely on WMH load (i.e. WMH ratio) versus providing information on both WMH load and WMH location (i.e. WMHDM). For model 5, the imaging inputs (GMDM and WMHDM) were separately propagated throughout the CNN convolutional layers, before concatenation of the resulting flattened feature maps. Therefore, the parameters for the imaging inputs were not shared but learned separately. A total overview of the CNN architecture is depicted in Figure 2.2.

As a preparation to the rest of the thesis, a thesis feasibility study was conducted that consisted of selecting and implementing a machine learning model that could be used to answer the research question of this thesis, i.e. the brain age CNN model (Appendix C).

2.4. Experimental setup

We split the data into a low, middle and high WMH load group, consisting of timepoints in the lower tertile ($WMH\ ratio\ < t_1$), middle tertile ($WMH\ ratio\ \geq t_1$ and $WMH\ ratio\ \leq t_3$) and upper tertile ($WMH\ ratio\ > t_3$) respectively. Using these groups, we performed two sets of experiments. First, we developed the models on the low WMH load group and used the high WMH load group as a test

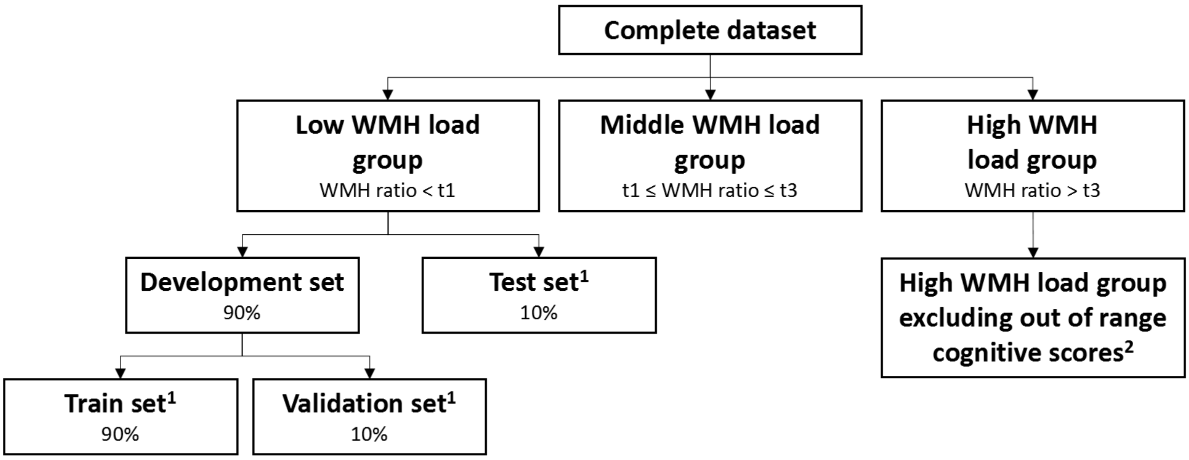


Figure 2.3: Overview of data splitting for models developed on low WMH load group. t1 = first tertile, t3 = third tertile, WMH = White matter hyperintensity.
[1] Subjects with WMH ratio < 5th percentile were excluded for the CNN models on WMHDMs only (models 3 and 4).
[2] Timepoints with MMSE $\leq$ 5th percentile and ADAS13 $\geq$ 95th percentile were excluded.

set. This setup evaluated the generalizability from a low to a high WMH load, which is clinically most relevant as the majority of ML models are developed on ADNI [29, 30], while ADNI subjects generally have a low WMH load [31, 32]. Second, we combined the WMH load groups and developed and tested models 2 (GMDM, age, sex) and 5 (GMDM, WMHDM, age, sex) on this full dataset. This was done to evaluate the effect of adding WMH information on the model performance when more data was available for training and validation, and when higher WMH loads (and consequently lower cognitive performance timepoints) were included in the train set.

Both MMSE and ADAS13 were used as outcome measures. Mean absolute error (MAE) of the real versus predicted cognitive scores was used to measure model performance. MAE was computed as follows, with $\widehat{y_i}$ as the predicted cognitive score of the i-th sample, y_i being the corresponding real cognitive score and $n_{samples}$ being the number of samples:

$$MAE(y, \widehat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i - \widehat{y_i}|$$

In addition to this, the Pearson correlation coefficient (r) of the predicted versus real cognitive score values is reported. A more elaborate description of the CNN method, the original brain age CNN architecture and a complete overview of the final CNN parameters are included in Appendix D.

2.4.1. Training on low WMH load group

In the first set of experiments, we developed the previously described models on the low WMH load group. For this, we split the low WMH load group into a train, validation and test set using stratification for the cognitive score outcome (Figure 2.3). The test set consisted of 10% of the complete low WMH load group, while the remaining data (development set) was split into a train (90%) and validation (10%) set. Subject timepoints were grouped during the split, i.e. all timepoints of one subject were assigned to the same set to maintain the independence of the sets. When a subject had timepoints belonging to both the low and high WMH load group, we only took the timepoints corresponding to the most frequent load group into account. The timepoints of the minority group were discarded to prevent subject overlap between the low and high WMH load group.

The log transformed WMH ratio was determined, for which WMH ratios of 0 were set to the lowest non-zero WMH ratio. For the CNN models on WMHDMs only (models 3 and 4), we excluded subject timepoints with WMH ratio < 5th percentile as we assumed that these maps contained too little information on their own.

We trained all previously described models on the low WMH load train set and tested them on the low WMH load test set. As the high WMH load group contained better cognitive score values that were not present in the train set, we tested the models twice on the high WMH load group:

1. Testing on the complete high WMH load group;
2. Testing on the high WMH load group excluding timepoints with out of range cognitive scores compared to the train set (MMSE score $\leq$ 5th percentile or ADAS13 score $\geq$ 95th percentile).

In order to compare the obtained performances to the performance of a model making random predictions, we developed a random model by training model 5 (GMDM, WMHDM, age, sex) while shuffling the cognitive score outcomes within the data batches.

2.4.2. Training on all WMH load groups

In the second set of experiments, we developed models 2 (GMDM, age, sex) and 5 (GMDM, WMHDM, age, sex) on the full dataset. For this, the data from all three WMH load groups was split into a train, validation and test set using stratification for the cognitive score outcome. The test set consisted of 10% of the full dataset, while the remaining data (development set) was split into a train (90%) and validation (10%) set.

3 Results

3.1. Data

In total, 4846 subject timepoints were included. Table 3.1 contains data characteristics for the low (N = 1615), middle (N = 1615) and high (N = 1616) WMH load group. All GMDMs passed the quality check. Table 3.2 contains the number of timepoints in the train and validation sets. When developing the models on only the low WMH load group, with the intention of testing on the high WMH load group, a number of WMH load converter timepoints were excluded for the low (N = 36) and high (N = 26) WMH load group. A flowchart of ADNI subject timepoint inclusion can be found in Appendix E. The number of timepoints in the intended test sets and distributions of different variables, including MMSE and ADAS13, can be found in and Appendix F.

Table 3.1: Characteristics of the low, middle and high WMH load group, computed over all timepoints, i.e. some subjects were included more than once in the statistics. CN = Cognitively normal, IQR = Interquartile range, MCI = Mild cognitive impairment, WMH = White matter hyperintensity.

	Low WMH load	Middle WMH load	High WMH load
# subject timepoints	1615	1615	1616
# subjects	732	753	631
# timepoints per subject (mean)	2.2	2.1	2.6
Male/female (%)	51.0/49.0	53.1/46.9	51.1/48.9
Age (years; mean ± std)	70.8 ± 7.3	74.0 ± 6.8	77.6 ± 6.9
CN/MCI/Dementia/Unknown (%)	48.7/42.2/8.2/0.9	37.2/47.4/14.7/0.7	28.3/48.9/22.4/0.4
MMSE (Median, IQR)[1]	29, 2	29, 4	28, 4
ADAS13 (median, IQR)[2]	9.7, 8.7	12.0, 11.3	15.0, 15.0
WMH ratio (median, IQR)	0.00053, 0.00067	0.0025, 0.0016	0.010, 0.010

[1] Excluding 4 timepoints in low, 3 timepoints in middle and 3 timepoints in high WMH load group with missing MMSE
[2] Excluding 16 timepoints in low, 20 timepoints in middle and 19 timepoints in high WMH load group with missing ADAS13

3.2. Experiments

The MAE and Pearson correlation coefficients of the CNN models on the train and validation sets can be found in Figure 3.1. The exact MAE values, correlation coefficients and scatter plots of model predictions are included in Appendix G, while learning curves of model training can be found in Appendix H.

Our first set of experiments trained the models on the low WMH load group. For MMSE, validation MAE differs little per model: it ranges from 1.55 - 1.79. For ADAS13, the spread is slightly larger with values ranging from 5.08 - 5.91. However, a simple prediction of a constant value for each subject timepoint, i.e. an MMSE of 29 or an ADAS of 10, already results in a good low WMH load validation MAE of 1.48 and 5.63 respectively, which is in the same range as our obtained performances. Moreover, when we retrained model 5 (GMDM, WMHDM, age, sex) while randomizing the cognitive scores in the data

Table 3.2: Number of subject timepoints in train and validation sets. For models 3 and 4, subjects with little to no WMHs are excluded, resulting in the lower number of timepoints. GMDM = Gray matter density map, NA = Not applicable, WMH = White matter hyperintensity, WMHDM = White matter hyperintensity density map.

Models	Cognitive score	Train set low WMH load	Validation set low WMH load	Train set all WMH loads	Validation set all WMH loads
1: GMDM 2: GMDM, age, sex 5: GMDM, WMHDM, age, sex 6: GMDM, log(WMH ratio), age, sex	MMSE	N = 1275	N = 142	N = 3916	N = 436
	ADAS13	N = 1265	N = 142	N = 3880	N = 431
3: WMHDM 4: WMHDM, age, sex	MMSE	N = 1204	N = 140	NA	NA
	ADAS13	N = 1197	N = 136	NA	NA

batches, we obtained a validation MAE of 1.63 for MMSE and 5.68 for ADAS13. As this indicates that our CNN models do not outperform simple predictions that are not learned on the data, we consider these models to be suboptimal. Therefore, we did not evaluate the performance of the models on the intended test sets. While all models have similar validation MAE, their correlation coefficients show differences. For MMSE, correlation coefficients range from 0.13 - 0.61 on the train set and -0.11 - 0.34 on the validation set. The correlation coefficients are in general higher for ADAS13, with ranges of 0.69 - 0.92 for the train set and 0.03 - 0.45 for the validation set. For both MMSE and ADAS13, the best validation coefficient is achieved by model 6 (GMDM, log(WMH ratio), age, sex).

For training on the full cohort (all WMH load groups combined), the validation MAE of model 2 (GMDM, age, sex) and 5 (GMDM, WMHDM, age, sex) was in general slightly higher than when these models were developed on the low WMH load data (MMSE: 1.99 for model 2 and 2.06 for model 5; ADAS13: 5.64 for model 2 and 6.31 for model 5). Model 2, developed on all data, outperformed the low WMH load models for MMSE with respect to the correlation coefficient on the train (0.81) and validation (0.50) set. For model 5 and MMSE, the train correlation coefficient was in the same range as coefficients obtained by the low WMH load group models (0.57), but the validation coefficient was slightly higher (0.43). When looking at the ADAS13 results, model 2 (GMDM, age, sex) obtained higher train and validation coefficients than model 5 (GMDM, WMHDM, age, sex) (0.82 vs 0.75 and 0.67 vs. 0.65). Figure 3.2 provides a visual overview of the prediction quality of these models. In general, the models show better predictive behavior for higher cognitive performances; for lower cognitive performances (i.e. higher ADAS13 and lower MMSE scores), the predictions differ largely from the correct cognitive scores.

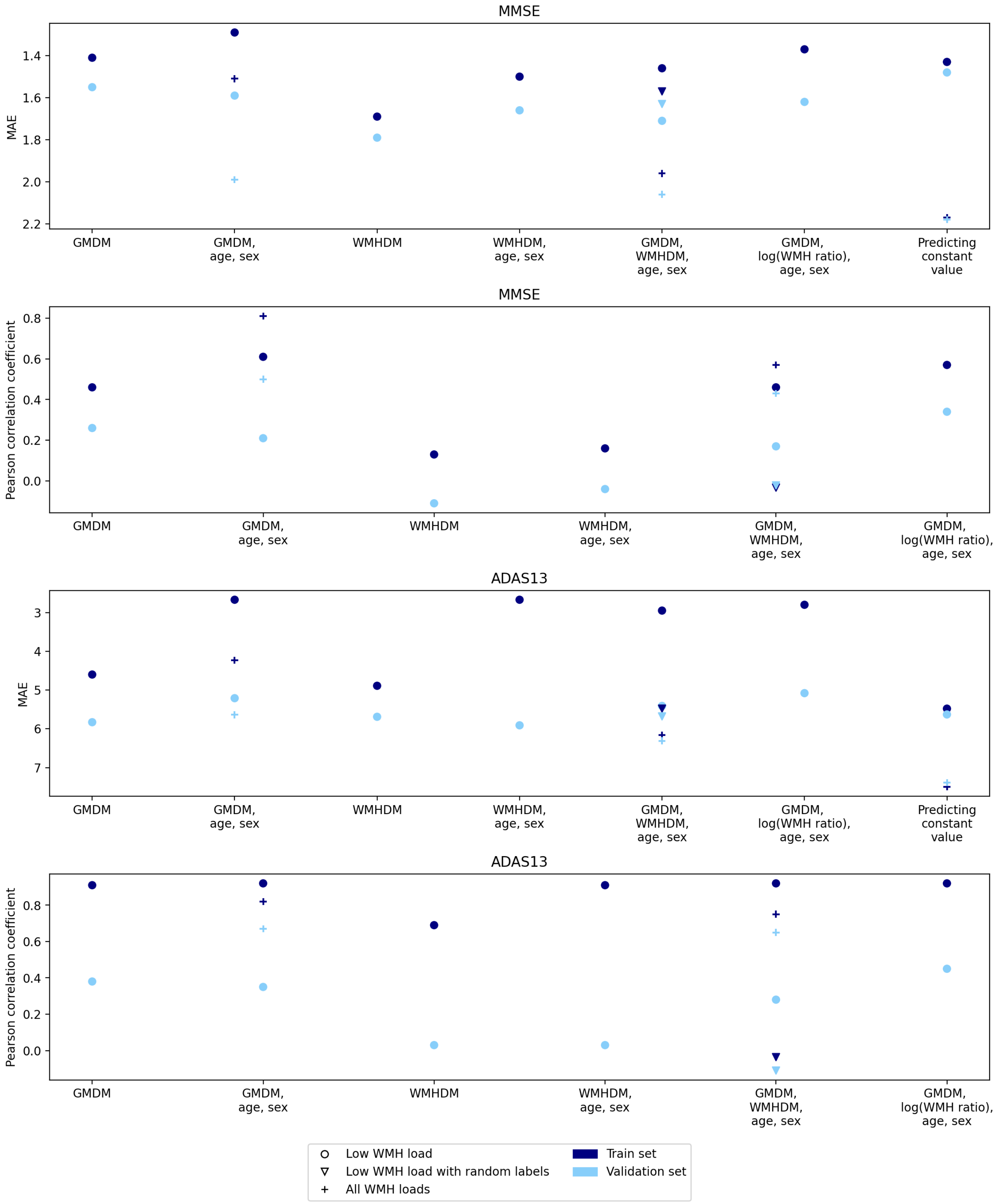


Figure 3.1: Mean absolute error (MAE) and Pearson correlation coefficients obtained for the CNN models. Color distinguishes the train and validation set, while marker shape indicates on which data the model was developed: low WMH load group (●), low WMH load group with randomized cognitive scores per batch (▽), and all WMH load groups (+). In addition to the experimental results, the MAE when predicting a constant value is included. Note that for MAE, the vertical axis is inverted for so that the plot can be interpreted more intuitively (smaller MAE values indicate better performance). GMDM = Gray matter density map, WMH = White matter hyperintensity, WMHDM = White matter hyperintensity density map.

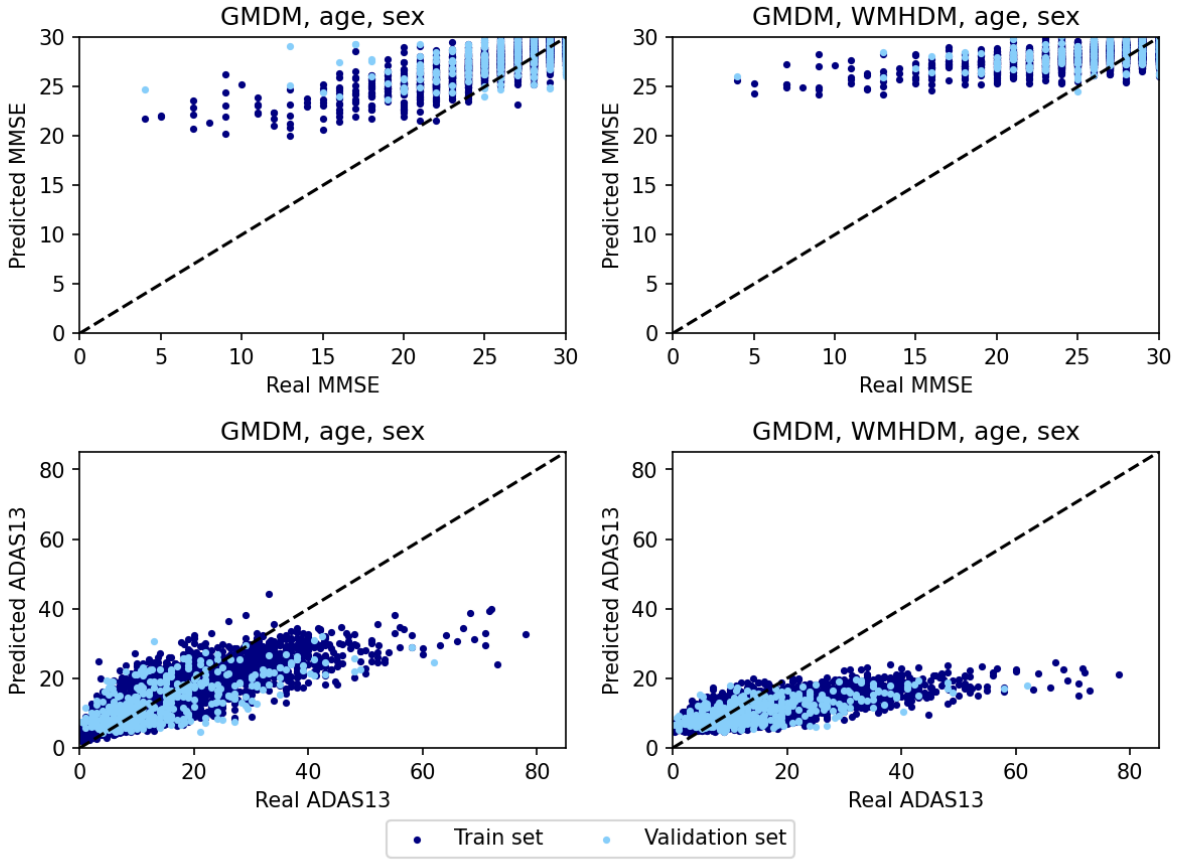


Figure 3.2: Scatter plots of model predictions for model 2 (GMDM, age, sex) and model 5 (GMDM, WMHDM, age, sex) when developed on all WMH load groups. Ideal predictions are indicated by the dashed line. GMDM = Gray matter density map, WMH = White matter hyperintensity, WMHDM = White matter hyperintensity density map.

4 Discussion

To our knowledge, this study is the first to evaluate the influence of vascular pathology (represented by WMHs on MRI) on ML prediction models in the context of AD. We developed multiple CNN models for prediction of cognitive performance based on several inputs. We aimed to evaluate the performance and generalizability of these models. We presented the performance on the train and validations sets of the low WMH load group and all WMH load groups. However, as our trained models obtained a similar performance to prediction of a constant value and prediction by a model developed on randomized cognitive scores, we were not able to answer our research question. Therefore, we did not evaluate the performance of the models on the intended test sets, as we wanted to keep these sets independent and unused for potential future research. The generalizability of the developed models to a higher WMH load was therefore not evaluated in this thesis.

Looking at the results, the validation MAE of the low WMH load group models is similar for the different inputs. This would indicate that adding age, sex or WMH input does not influence the model's performance. However, MAE is limited as a performance measure because most timepoints correspond to a good cognitive performance with closely grouped high MMSE and low ADAS13 scores. This aspect is better quantified by Pearson's correlation, which is lower for the MMSE and ADAS13 WMHDM models (model 3 and 4) than for the other models when looking at the validation set. This indicates that the WMHDMs contain less predictive information on their own than the GMDMs. This would be expected, especially for the low WMH load group. While adding WMHDM input to the GMDM, age, sex model results in a slightly lower validation coefficient, adding log(WMH ratio) results in a higher validation coefficient (MMSE: Δ_r = 0.13; ADAS13: Δ_r = 0.10). Adding age and sex information to the model generally results in more overfitting, as it mostly affects the training performance. However, age, sex and WMH information can be added to a CNN in multiple ways, and we only examined one possibility.

When using all data, our models were able to predict the cognitive measures succesfully. The ADAS13 model based on GMDM, age and sex (validation MAE: 5.64) outperforms constant value prediction (validation MAE: 7.39), and higher ADAS13 and lower MMSE values were predicted than when using only low WMH load data. While adding WMHDM input to this model results in a slightly lower performance (MAE: validation 6.31), it still outperforms constant value prediction. The Pearson correlation validation coefficients of these two models are similar (MMSE: 0.50 and 0.43; ADAS13: 0.67 and 0.65). Using all data for training instead of just the low WMH load group also reduces the difference between train and validation coefficients, indicating that information learned on the train set generalizes better to the validation set. It would be interesting to look into the weights of these models and to generate saliency maps, i.e. heatmaps of the imaging inputs indicating what areas the model focused on, to evaluate how the models used the WMH input.

Multiple explanations can be considered for the suboptimal performances of the low WMH load models. First, there are clear limitations in the cognitive score data. The ADNI MMSE distribution suffers from a pronounced ceiling effect because the majority of the subjects perform well on the MMSE. Lower MMSE scores are largely underrepresented, making it harder for a model to capture the relations between the input data and the outcome. Consequently, our models mostly predicted high MMSE values and failed to consider the lower MMSE scores. While the ADAS13 distribution suffers less from a floor effect, it is still skewed and higher ADAS13 scores (i.e. lower cognitive performance) occur much less than the lower (i.e. good) scores. For ADAS13, better model performance was obtained using all WMH load groups, indicating that the model benefits from more (lower cognitive performance) data. Additionally, the models could suffer from MMSE and ADAS13 measurement noise. Exact scores may fluctuate due to e.g. different raters, patient's education and cultural background, and the exact timing of the cognitive examination. Therefore, a change in MMSE or ADAS13 score does not necessarily reflect a change in cognitive performance. Lastly, a limitation of the input data is that MRI scans are a macroscopic measure of cerebral disease and do not always fully correspond to a patient's clinical presentation.

In addition to the cognitive score limitations, ADNI data may not be the most ideal study population for other reasons as well. We used ADNI data as it is publicly available and frequently used for the development of ML prediction models. Hence, our results are valid for most AD related ML prediction models. Also, ADNI used strict criteria and protocols resulting in a large harmonised multicenter dataset. As these criteria made for a well educated study population with few comorbidities and, as stated before, a relatively good cognitive performance, ADNI is not fully representative of the patient spectrum encountered in clinical practice. Most importantly for this study, it is known that ADNI has a relatively low WMH load [31, 32]. As most ML models have been developed on ADNI data [29, 30], they encounter lower WMH loads in their training population than those present in the clinical population. We simulated this phenomenon by splitting ADNI data into a low and high WMH load group, with the low WMH load group being used for training and the high WMH load group representing the clinical population as a test set. We aimed to evaluate the generalizability from a low to a high WMH load, providing an estimate of the generalizability to clinical practice. However, the WMH load in ADNI (especially in its low WMH load group) could be too low to capture the relation between WMHs and cognitive performance. Furthermore, the ADNI WMH load group had a low percentage of dementia timepoints (8.2%), which could make it harder for the model to learn the relationship between atrophy and cognitive performance. We expect that higher WMH loads are encountered in clinical practice than are included in ADNI, and that the high WMH load group is even older with more extensive cognitive complaints than encountered in ADNI.

Another explanation of our model's performance could lie in the model architecture (i.e. the number, types and order of layers) and hyperparameters, as these can greatly influence the performance. However, as this model has been used succesfully for brain age prediction, we believe the model architecture to be suitable for our research question. Also, we performed hyperparameter optimization, albeit with a limited amount of hyperparameters, and the learning curves show logical learning behavior. Therefore, we do not consider wrong hyperparameter settings to be a large underlying cause of the impeded performances.

Several other ML prediction models have been developed for MRI based cross-sectional determination of MMSE and ADAS on ADNI data. Bhagwat et al. [17] developed an artifical neural network for ADNI 1 and ADNI 2 data (N = 1359) based on hippocampal segmentations and cortical thickness measures (MMSE: mean r = 0.55; ADAS: mean r = 0.63). Yan et al. [18] implemented a sparse multitask learning model for ADNI 1 data (N = 718) using cortical thickness measures (MMSE: mean r = 0.56; ADAS: mean r = 0.64). Lastly, Stonnington et al. [19] applied relevance vector regression to ADNI subjects (N = 586) using GMDMs (MMSE: r = 0.48; ADAS: r = 0.57). As for our models, MMSE prediction was found to be harder than ADAS prediction. Although we cannot completely compare our performance to these values because we did not use our test sets, the Pearson correlation on the validation set of our GMDM, age, sex model is similar when using all data (MMSE: r = 0.50; ADAS: r = 0.67).

Even though our performances are not state of the art, this study takes an important methodological step towards inclusion of vascular pathology in AD related ML prediction models, thereby addressing mixed dementia. In relation to this, our study has several strengths. By developing multiple models, we assessed the added value of different inputs. Importantly, we evaluated adding only WMH volume information as well as adding both WMH volume and spatial information. To make the comparison between models as fair as possible, we optimized hyperparameters separately for each input. While this could result in different hyperparameters, this prevents the model optimization from being biased towards one particular input. Moreover, including multiple timepoins per subject resulted in a large dataset and an intrinsic form of data augmentation. Limitations of this study other than previously discussed include incorrect GM and WMH segmentation and incorrect registration to the GM template. Incorrect WMH segmentation could lead to subjects being included in the wrong WMH load group. For example, some subjects converted from the high to the low WMH load group over time, which is clinically not possible. However, if a subject had timepoints belonging to both WMH load groups, only timepoints from the most frequent group were used, thereby partly removing these unrealistic conversion timepoints.

As our research question has been left unanswered, many roads are open for future research. First, other cognitive scores could be used as outcomes of the CNN, e.g. the clinical dementia rating (CDR) or the Montreal cognitive assessment (MoCA). However, for ADNI data, these scores likely also suffer from skewed distributions with underrepresentation of subjects with lower cognitive performance. The influence of WMHs on dementia diagnosis (as opposed to cognitive performance) could also be evaluated. Yet, this setup would not permit taking VaD or mixed dementia into account, as the number of subjects with this diagnosis is probably too low for a model to recognize them as a separate diagnostic category. Therefore, the research population would mainly consist of AD patients, but the models already perform well on AD vs. CN classification. Most added value of WMHs would lie in populations with higher WMH loads and with vascular or mixed dementia. Other research populations might therefore be more suitable than ADNI for answering our research question, e.g. population-based cohorts like the Rotterdam Study or clinical datasets.

As a next step, we aim to use the methodology of this thesis for brain age prediction. The CNN model used in this thesis was specifically developed for brain age prediction, and brain age is related to cognitive functioning and dementia risk. Age as an outcome does not suffer from a largely skewed distribution, and we would be able to include even more data as age is known for each subject timepoint. Moreover, we plan to combine ADNI data with data from the Rotterdam Study, thereby generating a large dataset with a broader spread in WMH load, including vascular and mixed dementia cases. With this setup, we hope to revisit and answer our research question in order to improve generalizability of ML methods for AD to clinical populations with prominent vascular pathology.

5 Conclusion

In conclusion, we were not able to answer our research question due to the suboptimal performance of our models. The models developed on the low WMH load data likely suffered from an underrepresentation of lower cognitive performance data, which is supported by the fact that the ADAS13 models were able to perform better than constant value prediction when developed on all WMH load data. In general, the models did not benefit from adding WMHDM input to the model, while they did benefit slightly from log(WMH ratio) information. Adding age and sex to the current models only increased overfitting. In the end, other datasets may be more suitable for answering our research question as ADNI's study population has few comorbidities and a low WMH load in general. While our research question could not be answered, there is still a pressing need to address vascular pathology and thereby vascular and mixed dementia in machine learning research. Our future research will focus on revisiting our research question through brain age prediction and by combining ADNI data with Rotterdam Study data.

Bibliography

[1] Gale SA, Acar D, Daffner KR. Dementia. American Journal of Medicine. 2018 10;131:1161-9. doi:10.1016/j.amjmed.2018.01.022.

[2] Breijyeh Z, Karaman R. Comprehensive Review on Alzheimer's Disease: Causes and Treatment. Molecules (Basel, Switzerland). 2020 12;25. doi:10.3390/molecules25245789.

[3] Vernooij MW, van Buchem MA. Neuroimaging in Dementia. In: Hodler J, von Schulthess GK, Kubik-Huch RA, editors. Diseases of the Brain, Head and Neck, Spine 2020–2023: Diagnostic Imaging. Springer International Publishing; 2020. p. 131-42. doi:10.1007/978-3-030-38490-6_11.

[4] Khan A, Kalaria RN, Corbett A, Ballard C. Update on Vascular Dementia. Journal of Geriatric Psychiatry and Neurology. 2016 9;29:281-301. doi:10.1177/0891988716654987.

[5] Mehta D, Jackson R, Paul G, Shi J, Sabbagh M. Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010-2015. Expert Opinion on Investigational Drugs. 2017 6;26:735-9. doi:10.1080/13543784.2017.1323868.

[6] Bron EE, Klein S, Papma JM, Jiskoot LC, Venkatraghavan V, Linders J, et al. Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. NeuroImage: Clinical. 2021 1;31. doi:10.1016/j.nicl.2021.102712.

[7] Vliet DV, Vugt MED, Bakker C, Pijnenburg YAL, Vernooij-Dassen MJFJ, Koopmans RTCM, et al. Time to diagnosis in young-onset dementia as compared with late-onset dementia. Psychological Medicine. 2013 2;43:423-32. doi:10.1017/S0033291712001122.

[8] Elahi FM, Miller BL. A clinicopathological approach to the diagnosis of dementia. Nature Reviews Neurology. 2017 8;13:457-76. doi:10.1038/nrneurol.2017.96.

[9] Boyle PA, Yu L, Wilson RS, Leurgans SE, Schneider JA, Bennett DA. Person-specific contribution of neuropathologies to cognitive loss in old age. Annals of Neurology. 2018 1;83:74-83. doi:10.1002/ana.25123.

[10] Rundek T, Tolea M, Ariko T, Fagerli EA, Camargo CJ. Vascular Cognitive Impairment (VCI). Neurotherapeutics. 2021. doi:10.1007/s13311-021-01170-y.

[11] Nandigam RNK, Schneider JA, Arvanitakis Z, Bang W, Bennett DA. Mixed brain pathologies account for most dementia cases in community-dwelling older persons. Neurology. 2008 3;70:816-7. doi:10.1212/01.wnl.0000307675.38908.39.

[12] Wardlaw JM, Smith EE, Biessels GJ, Cordonnier C, Fazekas F, Frayne R, et al. Position Paper Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. The Lancet Neurology. 2013;12. doi:10.1016/S1474-4422(13)70124-8.

[13] Pantoni L. Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. The Lancet Neurology. 2010 7;9:689-701. doi:10.1016/S1474-4422(10)70104-6.

[14] Alber J, Alladi S, Bae HJ, Barton DA, Beckett LA, Bell JM, et al. White matter hyperintensities in vascular contributions to cognitive impairment and dementia (VCID): Knowledge gaps and opportunities. Alzheimer's and Dementia: Translational Research and Clinical Interventions. 2019 1;5:107-17. doi:10.1016/j.trci.2019.02.001.

[15] Jorgensen DR, Shaaban CE, Wiley CA, Peter X, Gianaros J, Mettenburg J, et al. A population neuroscience approach to the study of cerebral small vessel disease in midlife and late life: an invited review. American journal of physiology: heart and circulatory physiology. 2018;314:1117-36. doi:10.1152/ajpheart.00535.2017.

[16] Custodio N, Montesinos R, Lira D, Herrera-Pérez E, Bardales Y, Valeriano-Lorenzo L. Mixed dementia: A review of the evidence. Dementia and Neuropsychologia. 2017 10;11:364-70. doi:10.1590/1980-57642016dn11-040005.

[17] Bhagwat N, Pipitone J, Voineskos AN, Chakravarty MM. An artificial neural network model for clinical score prediction in alzheimer disease using structural neuroimaging measures. Journal of Psychiatry and Neuroscience. 2019 7;44:246-60. doi:10.1503/jpn.180016.

[18] Yan J, Li T, Wang H, Huang H, Wan J, Nho K, et al. Cortical surface biomarkers for predicting cognitive outcomes using group l2,1 norm. Neurobiology of Aging. 2015 1;36:S185-93. doi:10.1016/j.neurobiolaging.2014.07.045.

[19] Stonnington CM, Chu C, Klöppel S, Jack CR, Ashburner J, Frackowiak RSJ. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. NeuroImage. 2010 7;51:1405-13. doi:10.1016/j.neuroimage.2010.03.051.

[20] Steinberg SI, Negash S, Sammel MD, Bogner H, Harel BT, Livney MG, et al. Subjective memory complaints, cognitive performance, and psychological factors in healthy older adults. American Journal of Alzheimer's Disease and other Dementias. 2013;28:776-83. doi:10.1177/1533317513504817.

[21] Anderson ND. State of the science on mild cognitive impairment (MCI). CNS Spectrums. 2019 2;24:78-87. doi:10.1017/S1092852918001347.

[22] de Boer R, Vrooman HA, van der Lijn F, Vernooij MW, Ikram MA, van der Lugt A, et al. White matter lesion extension to automatic brain tissue segmentation on MRI. NeuroImage. 2009 5;45:1151-61. doi:10.1016/j.neuroimage.2009.01.011.

[23] Wang J, Knol MJ, Tiulpin A, Dubost F, Bruijne MD, Vernooij MW, et al. Gray matter age prediction as a biomarker for risk of dementia. Proceedings of the National Academy of Sciences of the United States of America. 2019 10;116:21213-8. doi:10.1073/pnas.1902376116.

[24] Manera A, Dadar M, Fonov V, Collins DL. CerebrA: Accurate registration and manual label correction of the Mindboggle-101 atlas for the MNI-ICBM152 template; 2020. Available from: `http://nist.mni.mcgill.ca/icbm-152-nonlinear-atlases-2009/`.

[25] Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights into Imaging. 2018 8;9:611-29. doi:10.1007/s13244-018-0639-9.

[26] Folstein MF, Folstein SE, Mchugh PR. ”Mini-mental state” A practical method for grading the cognitive state of patients for the clinician. Journal of psychiatric research. 1975;12:189-98. doi:10.1016/0022-3956(75)90026-6.

[27] Lewis TJ, Trempe CL. Diagnosis of Alzheimer's—Standard-of-Care. In: The End of Alzheimer's. Elsevier; 2017. p. 52-77. doi:10.1016/b978-0-12-812112-2.00003-3.

[28] Kueper JK, Speechley M, Montero-Odasso M. The Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog): Modifications and responsiveness in pre-dementia populations. A narrative review. Journal of Alzheimer's Disease. 2018;63:423-44. doi:10.3233/JAD-170991.

[29] Rathore S, Habes M, Iftikhar MA, Shacklett A, Davatzikos C. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. Academic Press Inc.; 2017. doi:10.1016/j.neuroimage.2017.03.057.

[30] Grueso S, Viejo-Sobera R. Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review. Alzheimer's Research and Therapy. 2021 12;13. doi:10.1186/s13195-021-00900-w.

[31] Ramirez J, McNeely AA, Scott CJM, Masellis M, Black SE. White matter hyperintensity burden in elderly cohort studies: The Sunnybrook Dementia Study, Alzheimer's Disease Neuroimaging Initiative, and Three-City Study. Alzheimer's and Dementia. 2016 2;12:203-10. doi:10.1016/j.jalz.2015.06.1886.

[32] Puzo C, Labriola C, Sugarman MA, Tripodis Y, Martin B, Palmisano JN, et al. Independent effects of white matter hyperintensities on cognitive, neuropsychiatric, and functional decline: A longitudinal investigation using the National Alzheimer's Coordinating Center Uniform Data Set. Alzheimer's Research and Therapy. 2019 7;11. doi:10.1186/s13195-019-0521-0.

[33] Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. The American journal of psychiatry. 1984 11;141:1356-64. doi:10.1176/ajp.141.11.1356.

[34] Mohs RC, Knopman D, Petersen RC, Ferris SH, Ernesto C, Grundman M, et al. Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's Disease Assessment Scale that broaden its scope. The Alzheimer's Disease Cooperative Study. Alzheimer disease and associated disorders. 1997;11 Suppl 2:S13-21. doi:10.1097/00002093-199700112-00003.

A
Technical background

Machine learning is used to learn patterns from high dimensional data to predict a specific outcome, e.g. a diagnosis or a cognitive score. Inputs to the model can be derived from clinical data, e.g. age or lab tests, but also from medical imaging data. In classical machine learning methods, the user decides which imaging data features are used as the input to the machine learning algorithm. In deep learning, however, the algorithm is able to extract useful features from the image itself.

The most commonly used deep learning model in medical image analysis is the CNN. In a CNN, the input image is propagated through different layers, including convolution layers, pooling layers and fully connected (FC) layers. The convolution and pooling layers carry out the feature extraction, whereas the FC layers determine the relation between the extracted features and the outcome [25]. The convolution layers perform a convolution of the image with a filter, which can be considered as filtering the image. The last layer of the network is an activation function, which maps the extracted feature values to the predicted outcome value. The term ’model architecture’ is used to refer to CNN characteristics such as the number, type, order and shape of the layers.

Once the model architecture is defined and the data is preprocessed, the model can start training (Figure A.1). In supervised machine learning, training refers to a stage in which the model learns the relationship between the input and the outcome. Deep learning models also learn which features to extract during this stage. During training, part of the dataset (the train set) is supplied to the model, including the outcome value. Throughout the training process, the CNN learns which filters should be used in the convolution layers and which weights in the FC layers [25]. The learning process is guided by the loss function, which compares the predicted outcome of the model to the actual outcome and is therefore a reflection of the model performance. A high loss indicates that the model makes wrong predictions, whereas a low loss indicates that the model predicts the outcome well. During the training phase, the goal is to minimize the loss function. Each training step, the CNN adjusts its learnable parameters in such a way that the loss function becomes lower and that in turn the predictions become better. This optimization of the learnable parameters is conducted through backpropagation and gradient descent. Examples of loss functions for regression are the mean absolute error and the mean squared error. During the training, data is supplied to the CNN in small parts that are called batches. Each batch contains data from a limited number of subjects. Standard batch sizes are 4, 8, 16, 32 or 64. An epoch is defined as the period during which the whole training set, i.e. all batches, has been passed to the CNN once. As CNNs have many parameters that are learned during training, these models are in general data hungry.

When developing a model, the dataset is generally split into a train, validation and test set (Figure A.2). The train set is used during the previously described training phase to learn the relationship between the input data and the outcome. The validation set is used during the training phase to evaluate the performance on and generalization to unseen data. Lastly, once the model is fully trained, the test set is used to evaluate the performance of the model on independent, unseen data that was not used during the development of the model. The comparison of the training performance with the validation

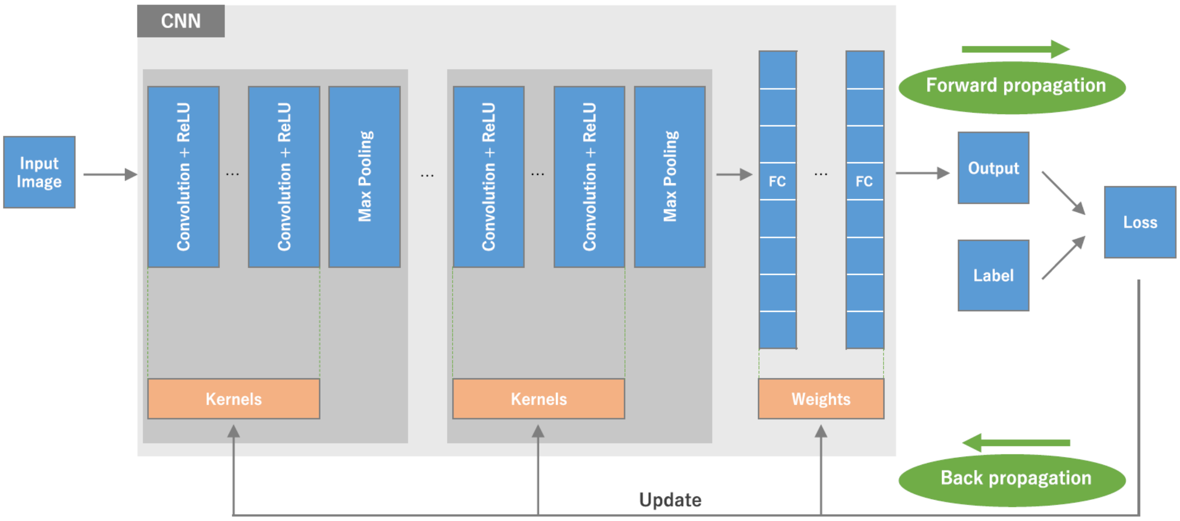


Figure A.1: Visual overview of a CNN and its training process [25]. An input image is propagated throughout different layers. The convolutional layers perform filtering of the image, with kernels being the applied filters. The last layers of the network are fully connected (FC). Forward propagation computes the loss of the model by comparing the output of the model with the real label. Next, backpropagation is conducted to adjust the learnable parameters, i.e. the kernels/filters of the convolution layers and the weights of the FC layers, in such way that the loss is minimized.

performance provides an indication of the model fit to the data. If the model performs well on the training data, but badly on the validation data, the model is overfit. This means that the model has learned to predict the training data too well by extracting patterns that do not generalize to other data, e.g. the validation data. If the model performs badly on both the training and validation data, the model is underfit. Underfitting occurs if the model was not able to learn relevant patterns from the data and was therefore not able to predict the outcome. Ideally, the performance on the training and validation data is similar.

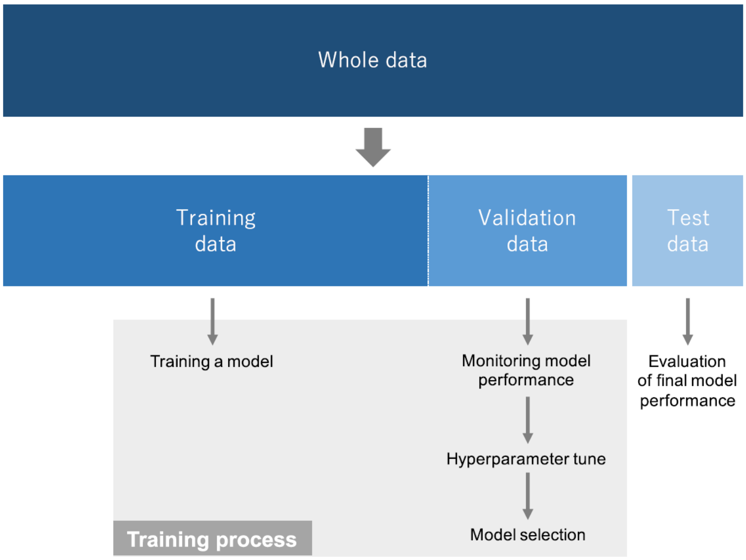


Figure A.2: Overview of the use of a train, validation and test set in machine learning [25]. The train set is used for training of the model, while the validation set is used to monitor the model performance, to tune model parameters or to select the most ideal model. Once the model is developed, the test set is used to evaluate the performance of the model on unseen data.

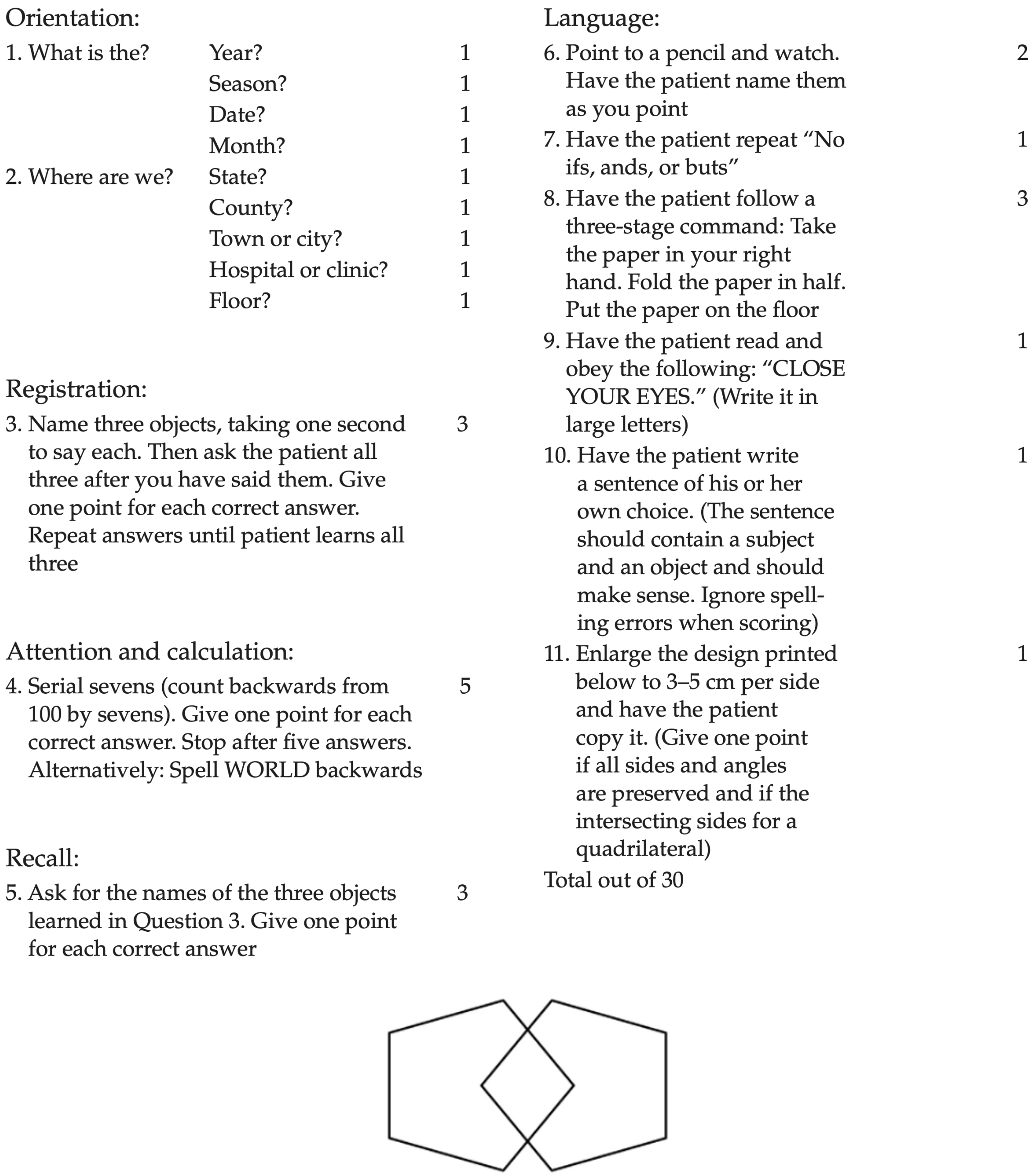
B

Scores of cognitive functioning

In this study, we used two scores of cognitive functioning: Mini-mental state examination (MMSE) and Alzheimer's disease assessment scale-cognitive subscale 13 (ADAS-Cog-13 or ADAS13).

The MMSE (Figure B.1) addresses a multitude of cognitive functions, including orientation to time and place, recall ability, short-term memory, and arithmetic ability [27]. Any score $\geq$ 25 can be considered normal. Abnormal scores can be divided into severe ($\leq$ 9), moderate (10 - 20) and mild (21 - 24) cognitive impairment.

The ADAS-Cog, in turn, consists of tasks assessing the memory, language and praxis domains [33]. The ADAS-Cog-11 evaluates eleven tasks depicted in Figure B.2. The ADAS-Cog-13 adds the two tasks of delayed word recall and a number cancellation or maze task to this score [28, 34].



Orientation:

1. What is the?	Year?	1
	Season?	1
	Date?	1
	Month?	1
2. Where are we?	State?	1
	County?	1
	Town or city?	1
	Hospital or clinic?	1
	Floor?	1

Registration:

3. Name three objects, taking one second to say each. Then ask the patient all three after you have said them. Give one point for each correct answer. Repeat answers until patient learns all three	3

Attention and calculation:

4. Serial sevens (count backwards from 100 by sevens). Give one point for each correct answer. Stop after five answers. Alternatively: Spell WORLD backwards	5

Recall:

5. Ask for the names of the three objects learned in Question 3. Give one point for each correct answer	3

Language:

6. Point to a pencil and watch. Have the patient name them as you point	2
7. Have the patient repeat "No ifs, ands, or buts"	1
8. Have the patient follow a three-stage command: Take the paper in your right hand. Fold the paper in half. Put the paper on the floor	3
9. Have the patient read and obey the following: "CLOSE YOUR EYES." (Write it in large letters)	1
10. Have the patient write a sentence of his or her own choice. (The sentence should contain a subject and an object and should make sense. Ignore spelling errors when scoring)	1
11. Enlarge the design printed below to 3–5 cm per side and have the patient copy it. (Give one point if all sides and angles are preserved and if the intersecting sides for a quadrilateral)	1
Total out of 30	

Figure B.1: Overview of the Mini-mental state examination [27].

Task	Description	Scoring
Word Recall	A list of 10 words is read by the subject, and then the subject is asked to verbally recall as many of the words as possible. Three trials of reading and recalling are performed.	Mean number of words not recalled across the three trials; scoring range is 0 to 10.
Naming Objects and Fingers	The subject is asked to name the fingers of their dominant hand as well as twelve objects: flower (plastic), bed (doll house furniture), whistle, pencil, rattle, mask, scissors, comb, wallet, harmonica, stethoscope, and tongs.	The number of fingers and objects correctly named; scoping range is 0 to 4.
Commands	The subject is asked to perform commands that involve one to five steps. For example, the two-step command is to "Point to the ceiling, then to the floor."	Scored from 0 to 5 based on the largest number of steps that are correctly performed (score is 0 if five step command is correctly performed).
Constructional Praxis	The subject is shown four geometric forms (circle, two overlapping rectangles, rhombus, cube) and asked to copy them on a piece of paper.	Scored from 0 to 5 based on the number of correctly drawn forms.
Ideational Praxis	The subject is asked to pretend to send a letter to themselves: fold letter, put letter in envelope, seal envelope, address envelope, and put a stamp on the envelope.	Scored from 0 to 5 based on difficulty of performing the five components.
Orientation	The subject is asked the date, month, year, day of the week, season, time of day, place, and person.	The number of correct responses; scoring range is 0 to 8.
Word Recognition	The subject reads twelve words aloud, and then these twelve words are randomly shuffled with twelve new words, and the subject is asked whether they have previously seen each of the twenty-four words. Three trials are performed.	Mean number of correct responses across the three trials; scoring range is 0 to 12.
Language	After the administration of the Word Recall task (Q1) ten minutes of open-ended conversation occur between the test administrator and subject, before the remainder of the tasks are presented. These ten minutes of conversation are used to assess language ability.	Quality of speech is given a global rating by the administrator that ranges from 0 to 5.
Comprehension of Spoken Language	This task also relies on the ten minutes of open-ended conversation. The administrator provides an assessment of how well the subject can understand speech.	The administrator provides a score from 0 to 5.
Word Finding Difficulty	During the aforementioned open-ended conversation, the administrator assesses how much difficulty the subject has in finding desired words.	The administer provides a score from 0 to 5.
Remembering Test Instructions	The administrator provides an assessment according to the number of times that the subject needed to be reminded of instructions for the Word Recognition task.	The administrator provides a score from 1 to 5.

Figure B.2: Overview of the ADAS-Cog-11 [28].

C

Thesis feasibility study

C.1. Introduction

The aim of this thesis was to evaluate whether vascular pathology influences the outcome of ML prediction models for dementia and as such hampers generalizability of these models from the research setting to clinical practice. Therefore, a machine learning model had to be selected and implemented that could be used to answer this research question, which was the goal of this thesis feasibility study. As our model input included imaging data, a CNN model was selected as the model type. Initially, we planned to use a CNN model developed by Bron et al. [6], which obtained a good performance for AD versus CN classification. However, consultation with clinical staff raised the question on the added value of vascular pathology to this model. First, the model distinguishes AD and CN subjects. If the AD subjects purely have AD, and little to no vascular component to their dementia, which is often the case in ADNI data, the vascular pathology would not be of added value to the diagnosis. Second, this model would not allow for (future) inclusion of vascular and mixed dementia subjects due to the small number of subjects with these diagnoses in datasets. Therefore, we selected cognitive performance as our outcome. It is well known that vascular pathology influences cognitive performance, and this outcome would also allow for (future) inclusion of non-AD dementia subjects in smaller numbers.

The next step was to select a CNN network, preferably already developed, that would be suitable for cognitive performance prediction. For this, we selected the brain age CNN model [23]. This model was previously developed within the department for determination of age from gray matter density maps (GMDMs). As brain age is related to cognitive performance, and as our study also used GMDM input, we deemed this model to be suitable for our cognitive performance regression task. However, when first implementing this model with GMDMs as the input and MMSE as the outcome, the model did not learn anything from the data. The model predicted a near constant value for each subject timepoint. This indicated that the model was unable to extract useful patterns from the data and could therefore not determine the relationship between the input and output. The rest of the feasibility study was focused on implementing the brain age model succesfully for our data and outcomes, i.e. making sure that the model was able to predict distinct cognitive score values.

C.2. Methods

In our first implementation of the brain age CNN model, the model was unable to extract predictive information for the data. In general, this can have multiple causes:

1. There is a bug in the code causing the model to be unable to learn the relationship between the input and the output.
2. The model architecture and/or hyperparameters are not suitable for the problem and should be adjusted.
3. The problem cannot be solved using deep learning as there is no coherent relationship between the input and the output.

To differentiate between these possible causes, a multitude of experiments were conducted:

1. Evaluate the data quality. If data quality is low, switch to data of higher quality. This provided an indication of whether the model itself was at fault (e.g. due to a bug) or whether the problem was too hard for the model to learn (e.g. due to noisy data).
2. Evaluate different sets of hyperparameters. This provided an indication of whether the model settings were not correct for the problem at hand.
3. Perform an experiment in which the outcome to predict also serves as an input to the model. This model should be able to predict the outcome with near perfect performance. If the model was indeed able to predict the outcome when the outcome value was also supplied as the input, it was more likely that the original problem was too hard to predict using the specified network architecture and parameters. If the model would still be unable to learn anything, a bug in the code is more likely.

We first addressed the data quality. In the initial experiment, MMSE was used as the outcome. However, this score suffered from a pronounced ceiling effect. Most subject timepoints were clustered around high MMSE scores of 26 - 30 and only a handful of subject timepoints corresponded to low MMSE scores. Therefore, the lack of distinct MMSE scores made for a complicated machine learning problem that was perhaps unsolvable due to intrinsic limits of the data itself. As a next step, we switched to the ADAS13 score as the outcome, as this score has a broader score range (maximum score of 85 for ADAS13 versus maximum score of 30 for the MMSE) and the subject timepoints were more evenly distributed amongst the different scores.

We performed experiments differing in the following hyperparameters: learning rate, batch size, early stopping patience and maximum number of epochs. Additionally, we evaluated both ADAS13 and log(ADAS13) as the outcome, as the ADAS13 score distribution still had a positive skew, thereby favoring lower ADAS13 scores. For the same reason, we evaluated multiple loss functions, including the mean squared logarithmic error, which penalizes underestimation more than overestimation. Lastly, we set up an experiment that supplied the outcome as an input to the model in addition to the GMDM input. If a persistent bug existed in the code, the model would still not be able to learn much from the input data. We repeated the previous experiments using this extra input.

C.3. Results and discussion

The experiments optimizing hyperparameters, outcome variables and loss functions all resulted in prediction of a near constant value. When repeating these experiments while adding the outcome as an input variable, the model was still not able to pick up much from the data. When the learning rate was increased, the model was able to predict a few distinct ADAS13 values, which could indicate that there was a problem of getting stuck in a local minimum. However, we were not able to optimize this model further, indicating that there was a bug present in the code. When going through the code, we found out that the outcome value of each subject (i.e. the cognitive score) was nested and supplied to the network with another shape than the output of the model. When we assured that the outcome values were of the same shape as the output of the model, the model started to learn and was able to predict the ADAS13 score successfully.

In Figure C.1, learning curves can be found of the models before and after the bug was uncovered. As can be seen, with the bug present, the model is unable to learn much from the data even if the outcome is supplied as the input. When the bug is removed and the outcome is again supplied as an input, the model is able to quickly predict the outcome with near perfect performance.

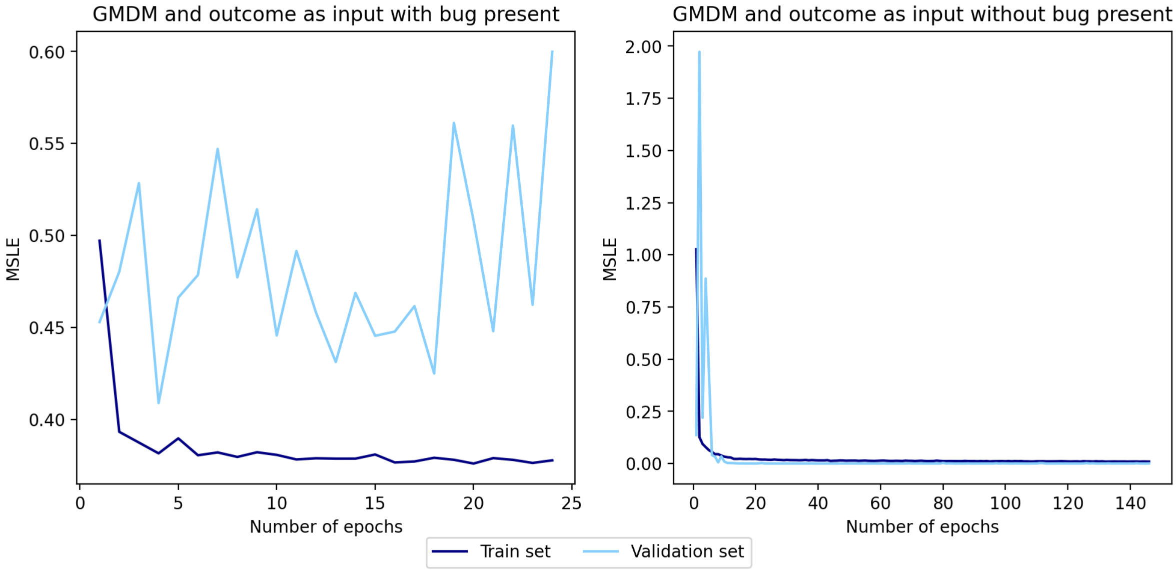


Figure C.1: Learning curves for brain age CNN model before and after the bug was uncovered. GMDM = Gray matter density map, MSLE = Mean squared logarithmic error.

D
CNN model

The architecture of the original brain age model is depicted in Figure D.1. For model 5 (GMDM, WMHDM, age, sex), two identical imaging branches were defined that performed feature extraction prior to the fully connected layers. Python 3.7.4, Tensorflow 2.2.0 and keras 2.3.0 were used to implement the CNN model. The models were trained on the high memory GPU nodes (gpu005, gpu006 and gpu-hm-001) of the GPU cluster of the Biomedical Imaging Group Rotterdam. The image input size to the network was 160 x 192 x 144. A mask excluding the cerebellum was applied to the images, followed by cropping of the images to the input size. Unnecessary background was removed, but some zero padding was preserved around the density maps. As a preprocessing step, age and log(WMH ratio) were normalized using Z-score normalization, for which the mean and standard deviation were determined on the train set. The model obtaining the lowest validation loss during training was saved as the final model. Adam optimization and early stopping were used.

The following CNN parameters were shared between all models: Adam optimizer β_1 (0.9), β_2 (0.999), ϵ (10^{-8}) and decay (10^{-4}); early stopping monitor (validation loss), Δ_{min} (0) and patience (20); maximum number of epochs (200). The learning rate $(10^{-3}$, 10^{-4}, 10^{-5}, $10^{-6})$ and batch size (8, 16) were optimized based on the validation set performance for the models. However, the ideal batch size was 16 for all final models. The learning rate was set to 10^{-4} for the MMSE models, except for model 5 trained on all data (10^{-5}), and 10^{-5} for the ADAS13 models. Loss functions were the mean absolute error for MMSE and mean squared logarithmic error for ADAS13. Data batches were stratified on cognitive score to make sure that each batch was representative of the cognitive score distribution.

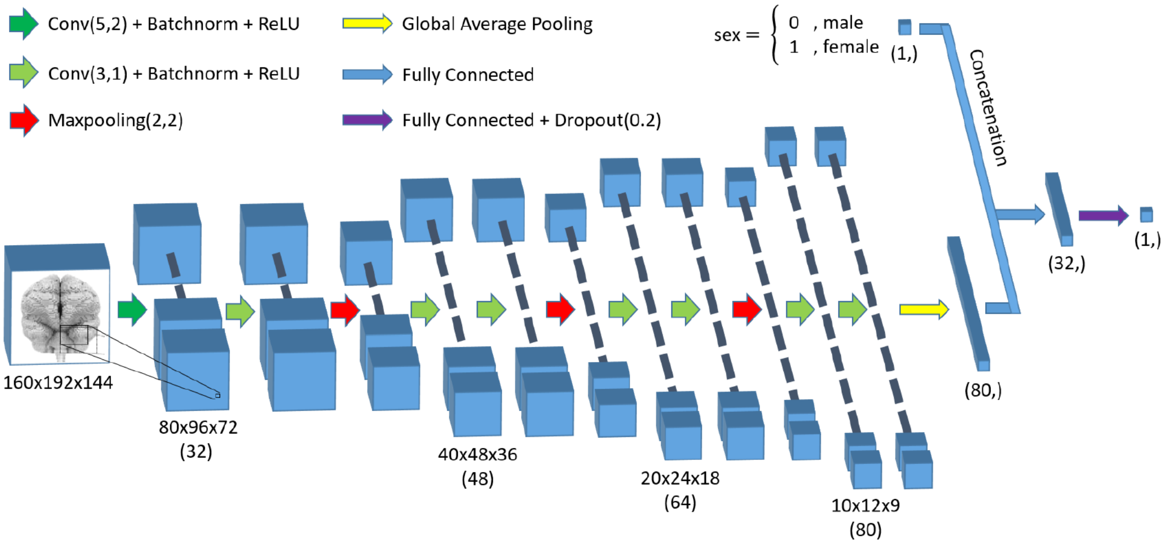


Figure D.1: Original brain age CNN model architecture [23].

E

Flowchart of ADNI data processing and inclusion

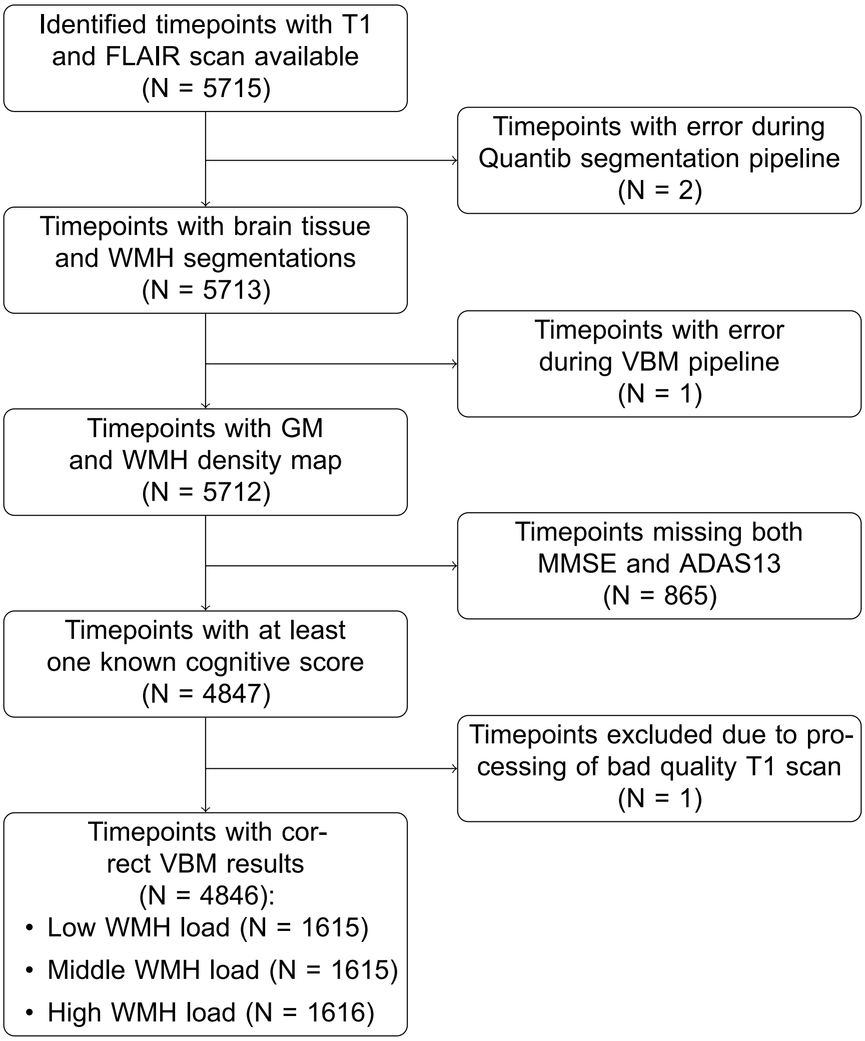

Identified timepoints with T1 and FLAIR scan available (N = 5715)
Timepoints with error during Quantib segmentation pipeline (N = 2)
Timepoints with brain tissue and WMH segmentations (N = 5713)
Timepoints with error during VBM pipeline (N = 1)
Timepoints with GM and WMH density map (N = 5712)
Timepoints missing both MMSE and ADAS13 (N = 865)
Timepoints with at least one known cognitive score (N = 4847)
Timepoints excluded due to processing of bad quality T1 scan (N = 1)
Timepoints with correct VBM results (N = 4846):
• Low WMH load (N = 1615)
• Middle WMH load (N = 1615)
• High WMH load (N = 1616)

Figure E.1: Flowchart of subject timepoint inclusion for ADNI data. GM = Gray matter, VBM = Voxel-based morphometry, WMH = White matter hyperintensity.

F

ADNI data distributions

Table F.1 shows the number of timepoints in each of the test sets (that were in the end not used). In Figure F.1 and Figure F.2, data distributions of respectively general clinical variables and scan variables can be found. Figure F.3 contains scatter plots depicting the relations between age, WMH ratio and the cognitive scores. Figure F.4 and Figure F.5 show the MMSE and ADAS13 distributions over the train, validation and test sets of respectively the low WMH load group and all three WMH load groups. Figure F.6 shows the scores that were considered to be (not) out of range when determining the subset of the high WMH load group excluding out of range scores.

Table F.1: Number of subject timepoints in the defined test sets. For models 3 and 4, subjects with little to no WMHs are excluded, resulting in a lower number of timepoints. GMDM = Gray matter density map, NA = Not applicable, WMH = White matter hyperintensity, WMHDM = White matter hyperintensity density map.

Models	Cognitive score	Test set low WMH	Test set high WMH 1[1]	Test set high WMH 2[2]	Test set all WMH
1: GMDM 2: GMDM, age, sex 5: GMDM, WMHDM, age, sex 6: GMDM, log(WMH ratio), age, sex	MMSE	N = 158	N = 1587	N = 1498	N = 484
	ADAS13	N = 157	N = 1571	N = 1492	N = 480
3: WMHDM 4: WMHDM, age, sex	MMSE	N = 152	N = 1587	N = 1498	NA
	ADAS13	N = 152	N = 1571	N = 1492	NA

[1] Complete high WMH load group

[2] High WMH load group excluding timepoints with out of range cognitive scores

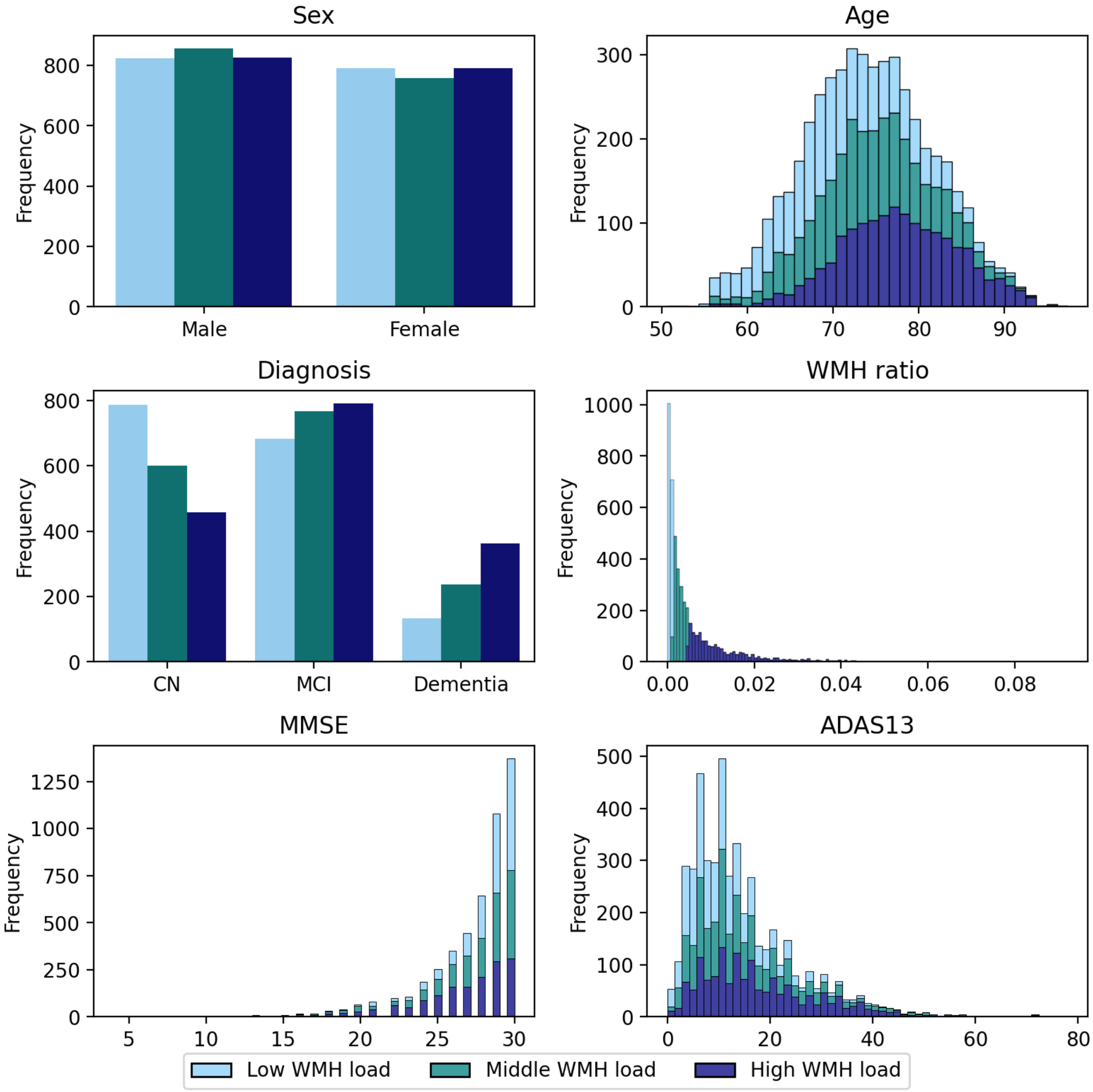


Figure F.1: Distribution of general variables for all timepoints in the low, middle and high WMH load group. Please note that the histograms are stacked. CN = Cognitively normal, MCI = Mild cognitive impairment, WMH = White matter hyperintensity.

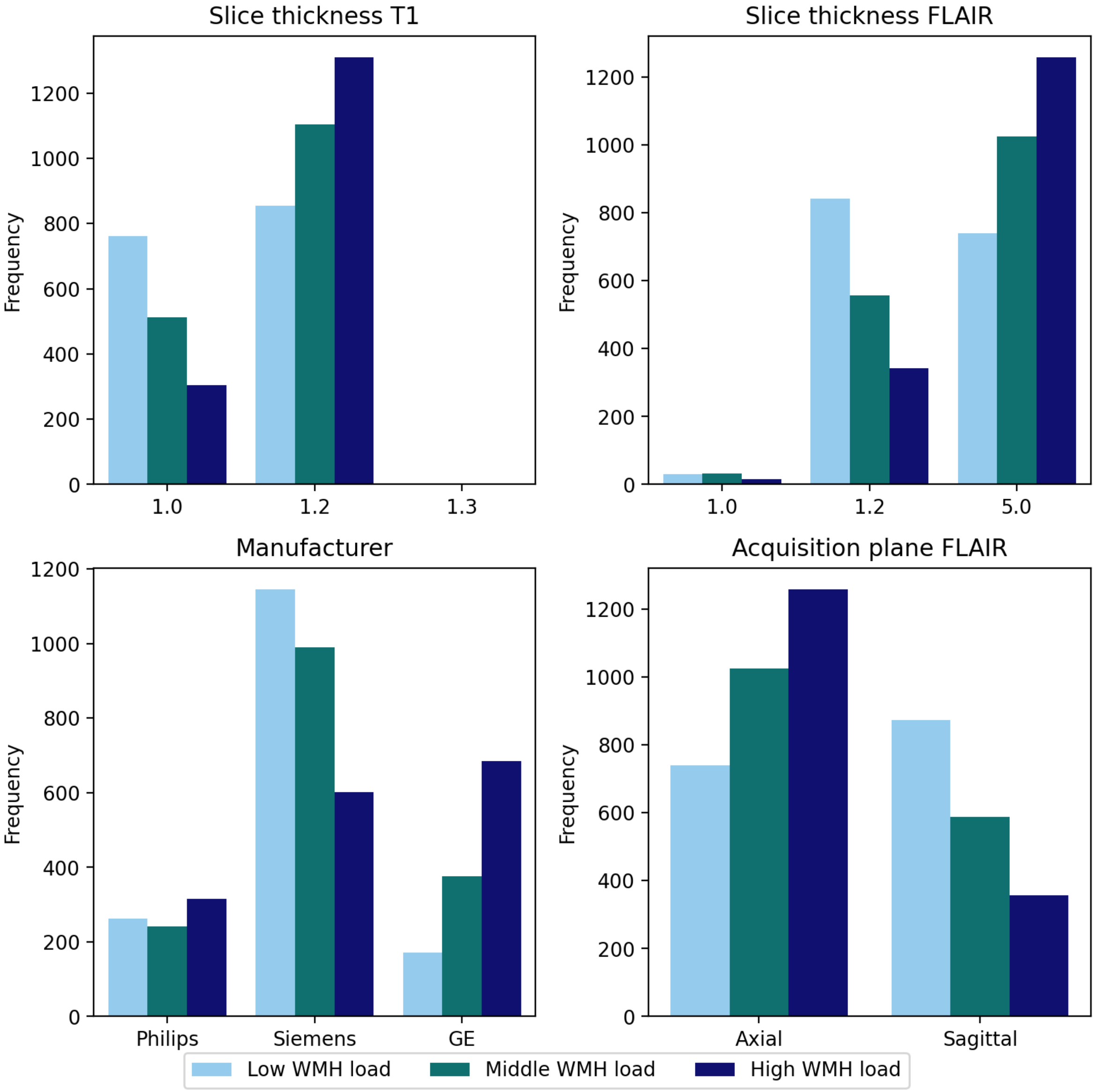


Figure F.2: Distribution of scan parameters for all timepoints in the low, middle and high WMH load group. WMH = White matter hyperintensity.

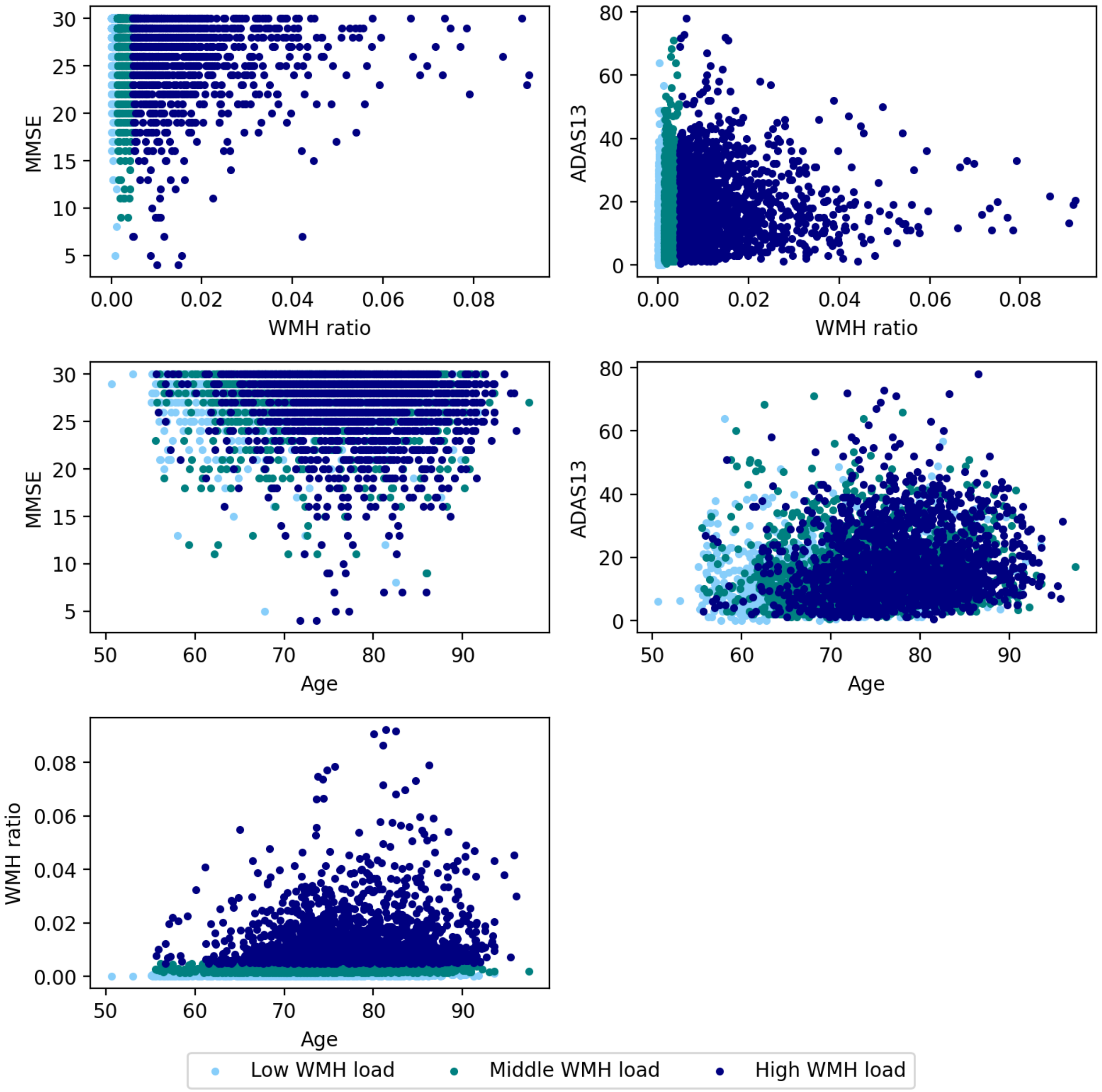


Figure F.3: Scatter plots of age, WMH ratio and cognitive scores. WMH = White matter hyperintensity.

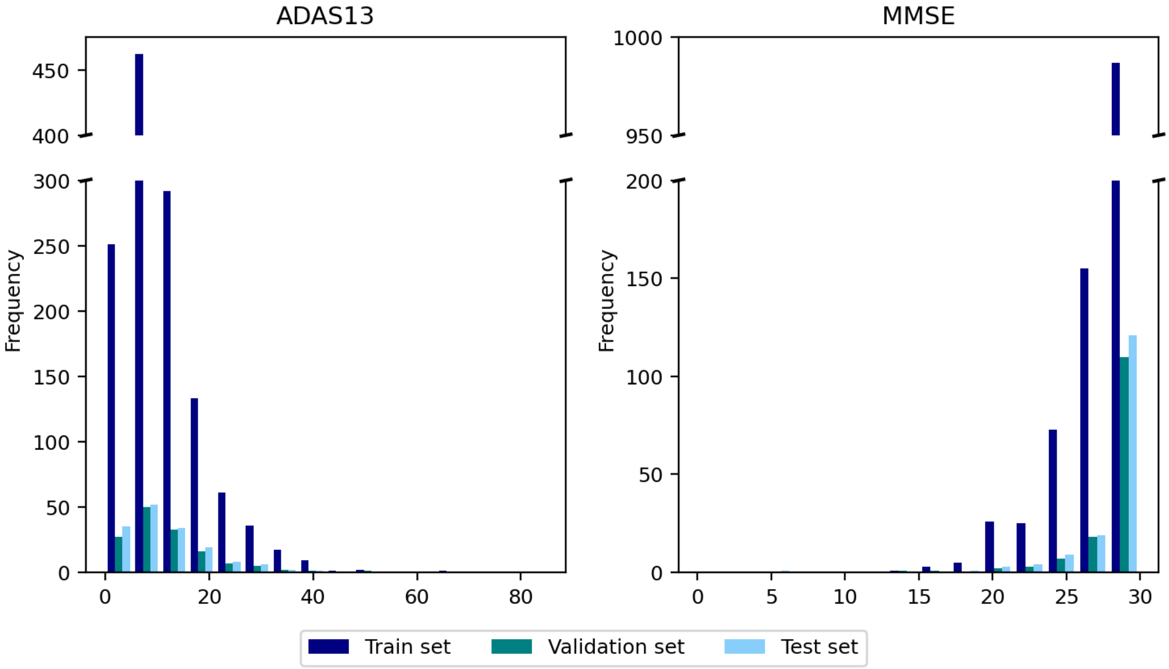


Figure F.4: Distribution of cognitive scores over the train, validation and test sets of the low white matter hyperintensity load group.

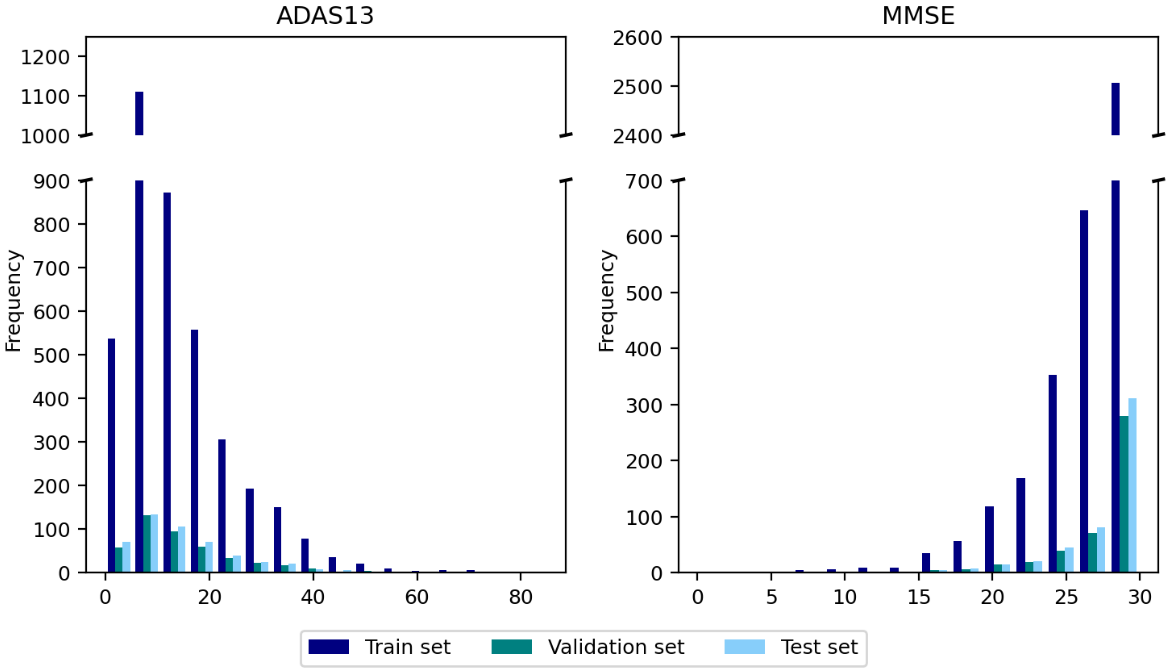


Figure F.5: Distribution of cognitive scores over the train, validation and test sets of all three white matter hyperintensity load groups.

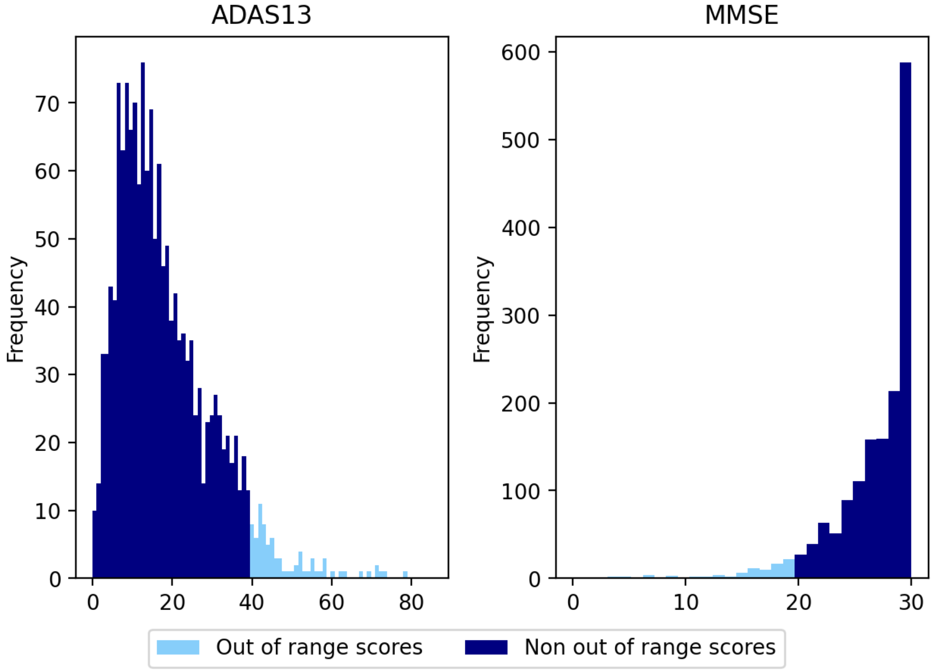


Figure F.6: Overview of (not) out of range cognitive scores when defining the high WMH load group subset excluding out of range cognitive scores.

G

CNN performance

The MAE of the CNN models on the train and validation sets can be found in Table G.1. The Pearson correlation coefficients of the predicted versus real cognitive score values are included in Table G.2. Table G.3 lists performances obtained when predicting a constant value, while Table G.4 contains the performance of model 5 (GMDM, WMHDM, age, sex) when the cognitive scores are randomly shuffled per batch. Scatter plots of the real versus predicted cognitive scores are included for the MMSE models developed on the low WMH load group (Figure G.1), the ADAS13 models developed on the low WMH load group (Figure G.2) and the models trained on random cognitive scores within the batches (Figure G.3).

Table G.1: Mean absolute error of the CNN models over the train and validation sets. GMDM = Gray matter density map, NA = Not applicable, WMHDM = White matter hyperintensity density map.

MMSE				
Model	Train set low WMH load	Validation set low WMH load	Train set all WMH loads	Validation set all WMH loads
1: GMDM	1.41	1.55	NA	NA
2: GMDM, age, sex	1.29	1.59	1.51	1.99
3: WMHDM	1.69	1.79	NA	NA
4: WMHDM, age, sex	1.50	1.66	NA	NA
5: GMDM, WMHDM, age, sex	1.46	1.71	1.96	2.06
6: GMDM, log(WMH ratio), age, sex	1.37	1.62	NA	NA
ADAS13				
Model	Train set low WMH load	Validation set low WMH load	Train set all WMH loads	Validation set all WMH loads
1: GMDM	4.60	5.83	NA	NA
2: GMDM, age, sex	2.67	5.21	4.23	5.64
3: WMHDM	4.89	5.69	NA	NA
4: WMHDM, age, sex	2.67	5.91	NA	NA
5: GMDM, WMHDM, age, sex	2.95	5.41	6.16	6.31
6: GMDM, log(WMH ratio), age, sex	2.80	5.08	NA	NA

Table G.2: Pearson correlation coefficients of the CNN predictions over the train and validation sets. GMDM = Gray matter density map, NA = Not applicable, WMHDM = White matter hyperintensity density map.

MMSE				
Model	Train set low WMH load	Validation set low WMH load	Train set all WMH loads	Validation set all WMH loads
1: GMDM	0.46	0.26	NA	NA
2: GMDM, age, sex	0.61	0.21	0.81	0.50
3: WMHDM	0.13	-0.11	NA	NA
4: WMHDM, age, sex	0.16	-0.04	NA	NA
5: GMDM, WMHDM, age, sex	0.46	0.17	0.57	0.43
6: GMDM, log(WMH ratio), age, sex	0.57	0.34	NA	NA
ADAS13				
Model	Train set low WMH load	Validation set low WMH load	Train set all WMH loads	Validation set all WMH loads
1: GMDM	0.91	0.38	NA	NA
2: GMDM, age, sex	0.92	0.35	0.82	0.67
3: WMHDM	0.69	0.03	NA	NA
4: WMHDM, age, sex	0.91	0.03	NA	NA
5: GMDM, WMHDM, age, sex	0.92	0.28	0.75	0.65
6: GMDM, log(WMH ratio), age, sex	0.92	0.45	NA	NA

Table G.3: Lowest mean absolute error to be obtained when predicting a constant value. MAE = Mean absolute error, WMH = White matter hyperintensity.

Data	Predicted value MMSE	MAE MMSE	Predicted value ADAS13	MAE ADAS13
Low WMH load train set	29	1.43	10	5.48
Low WMH load validation set	29	1.48	10	5.63
Low/middle/high WMH load train set	29	2.17	12	7.50
Low/middle/high WMH load validation set	29	2.18	12	7.39

Table G.4: Mean absolute error and Pearson correlation coefficients of model 5 (GMDM, WMHDM, age, sex) over the train and validation sets when the cognitive scores are randomly shuffled for each batch. GMDM = Gray matter density map, MAE = Mean absolute error, WMHDM = White matter hyperintensity density map.

Outcome	MAE train set	MAE validation set	Pearson correlation train set	Pearson correlation validation set
MMSE	1.57	1.63	-0.033	-0.022
ADAS13	5.48	5.68	-0.036	-0.11

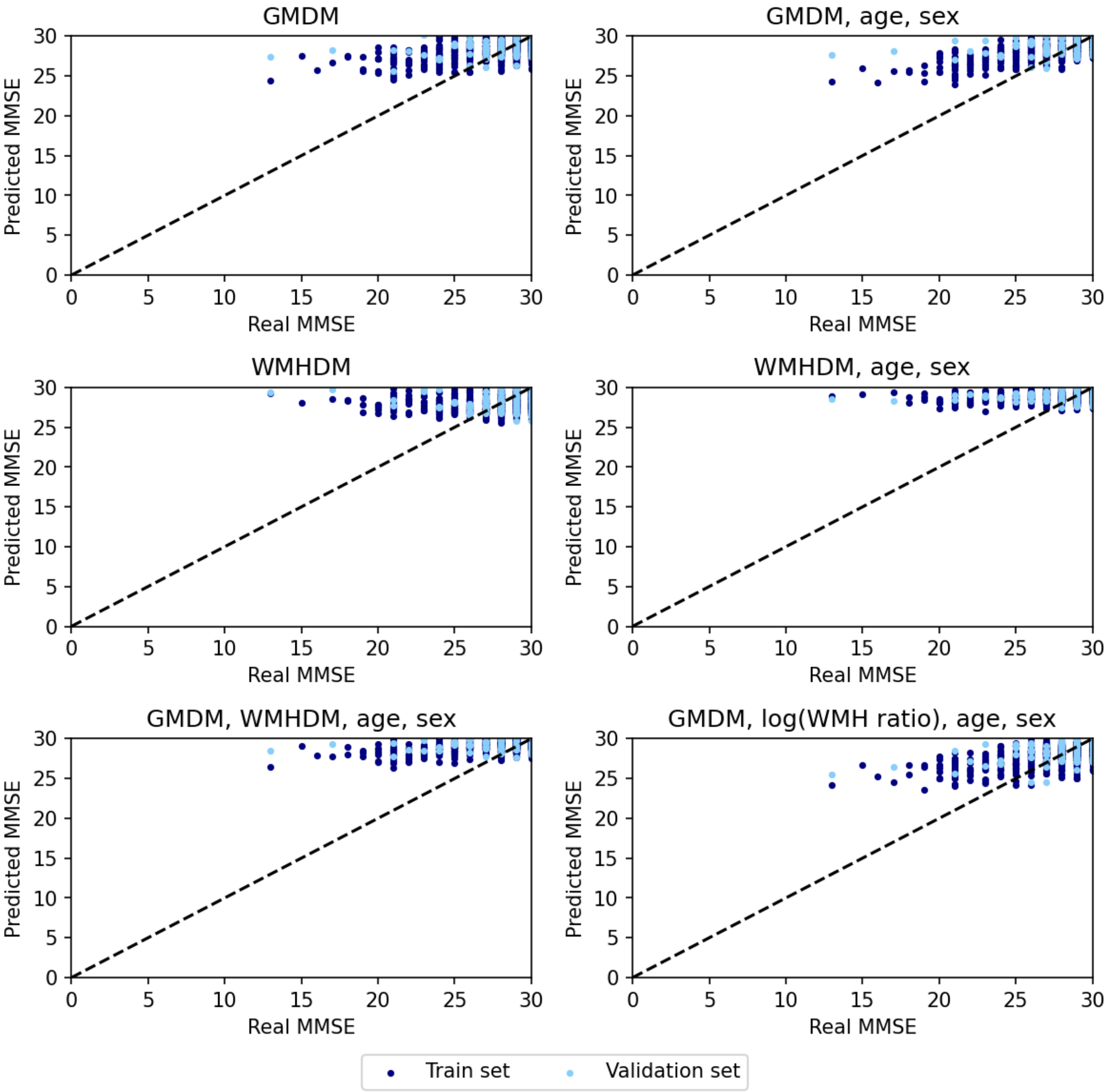


Figure G.1: Scatter plots of model predictions for MMSE models when developed on the low WMH load group. Ideal predictions are indicated by the dashed line. GMDM = Gray matter density map, WMH = White matter hyperintensity, WMHDM = White matter hyperintensity density map.

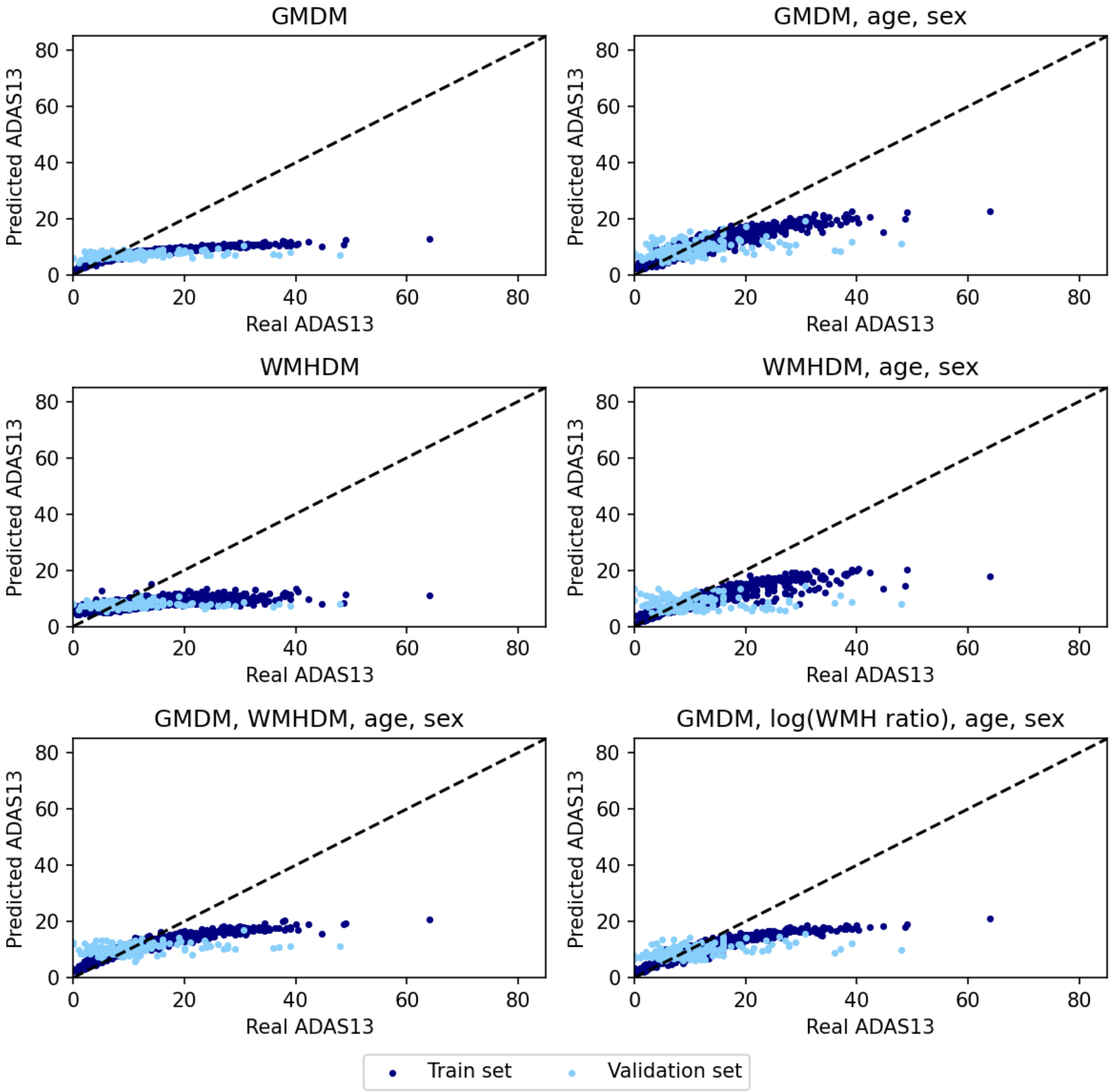


Figure G.2: Scatter plots of model predictions for ADAS13 models when developed on the low WMH load group. Ideal predictions are indicated by the dashed line. GMDM = Gray matter density map, WMH = White matter hyperintensity, WMHDM = White matter hyperintensity density map.

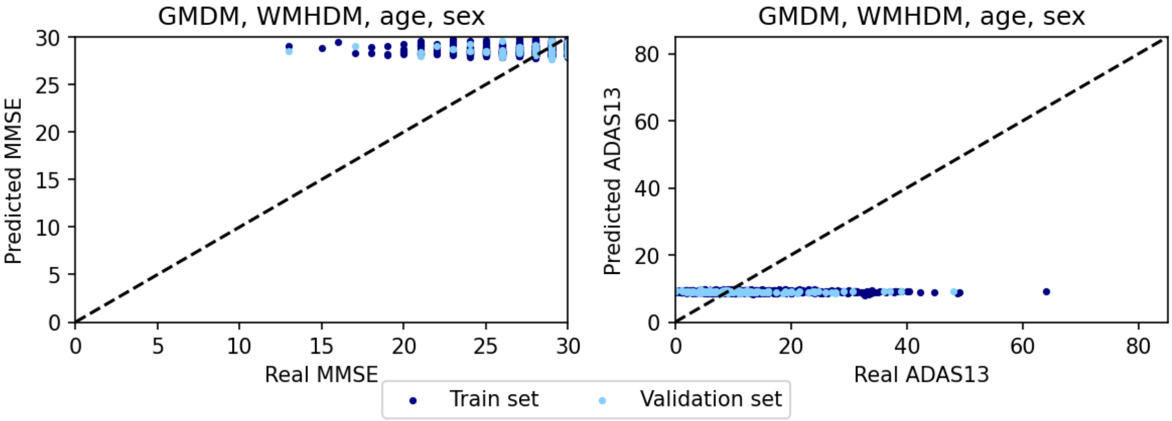


Figure G.3: Scatter plots of model predictions for model 5 (GMDM, WMHDM, age, sex) when cognitive scores are randomized within the batches. Ideal predictions are indicated by the dashed line. GMDM = Gray matter density map, WMHDM = White matter hyperintensity density map.

H

Learning curves

Learning curves are a visualization of the training process of a CNN. The horizontal axis is a representation of training progress, often the number of epochs. The vertical axis shows a performance measure, e.g. the loss or another performance metric, for both the train and validation set. During training, the loss should decrease for the train and validation set. Ideally, the train and validation losses converge towards the same plateau. Often, early stopping is used to stop model training once the validation loss starts increasing, which indicates that the model starts overfitting. See Figure H.1 for a depiction of an expected learning curve. Figure H.2 and Figure H.3 contain the learning curves for respectively the MMSE and ADAS13 models developed on the low WMH load group. Learning curves of model 2 (GMDM, age, sex) and model 5 (GMDM, WMHDM, age, sex) when developed on all WMH load groups are included in Figure H.4. Lastly, see Figure H.5 for learning curves of model 5 (GMDM, WMHDM, age, sex) when cognitive scores are shuffled within each batch.

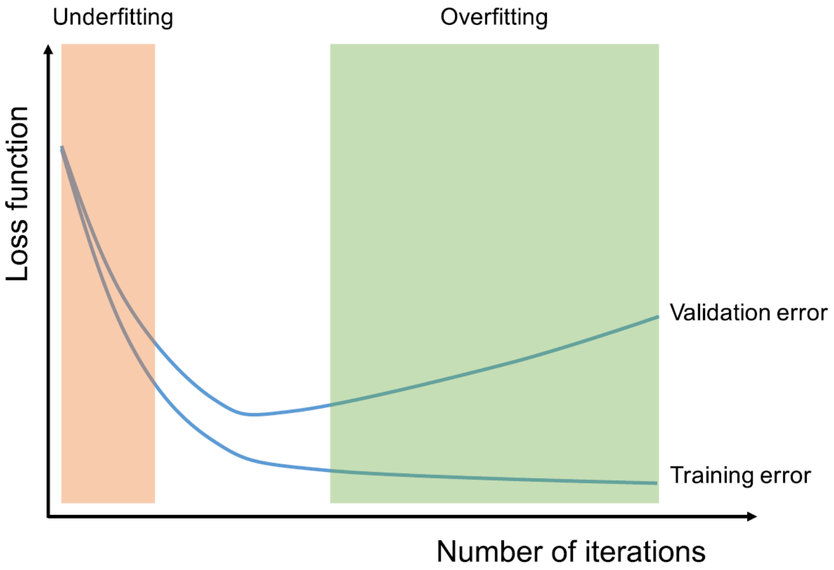


Figure H.1: Expected learning curves of a CNN [25]. The white interval in between underfitting and overfitting is the most ideal configuration of the model that can be reached by early stopping of the training process.

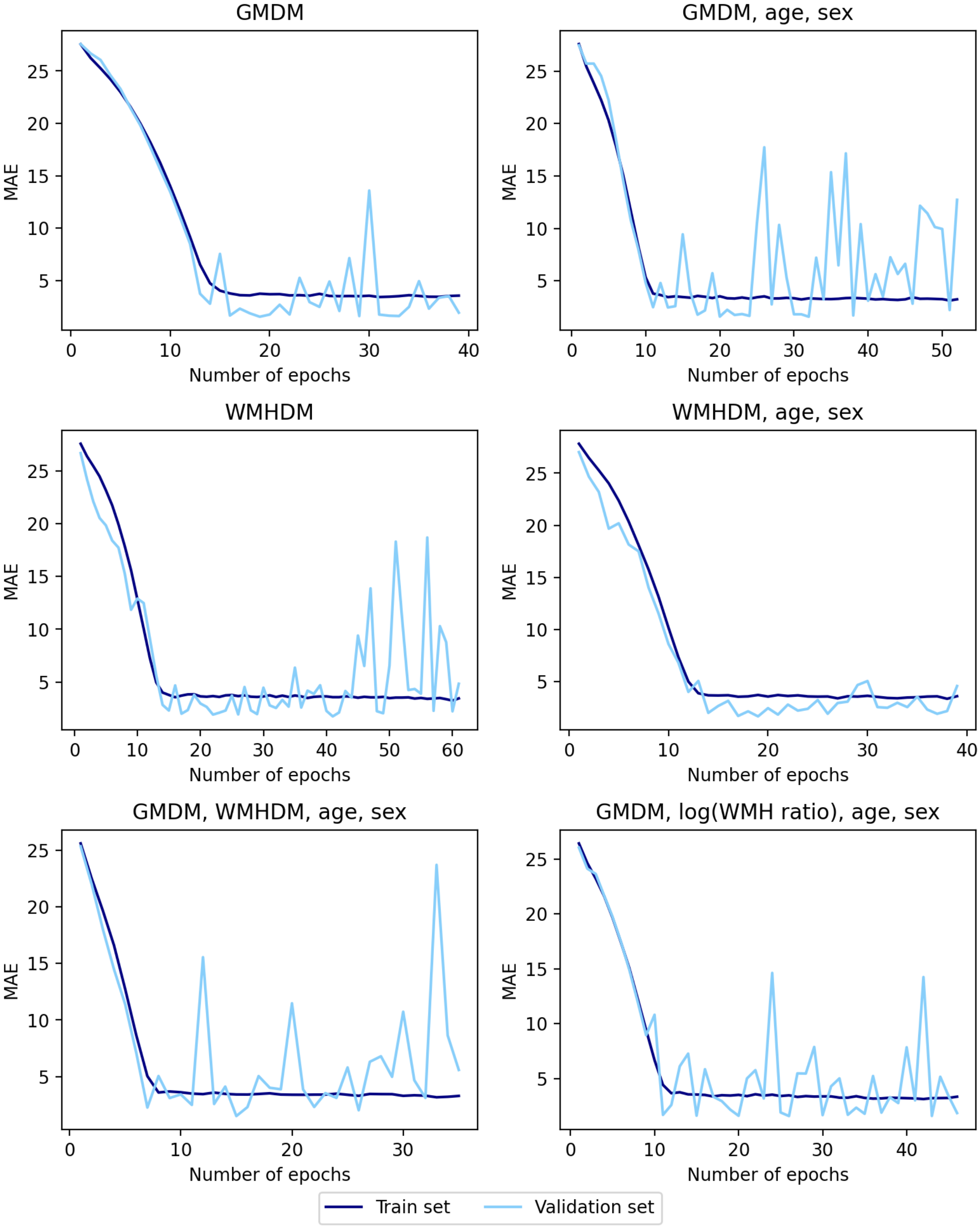


Figure H.2: Learning curves for MMSE models developed on low WMH load group. GMDM = Gray matter density map, MAE = Mean absolute error, WMH = White matter hyperintensity, WMHDM = White matter hyperintensity density map.

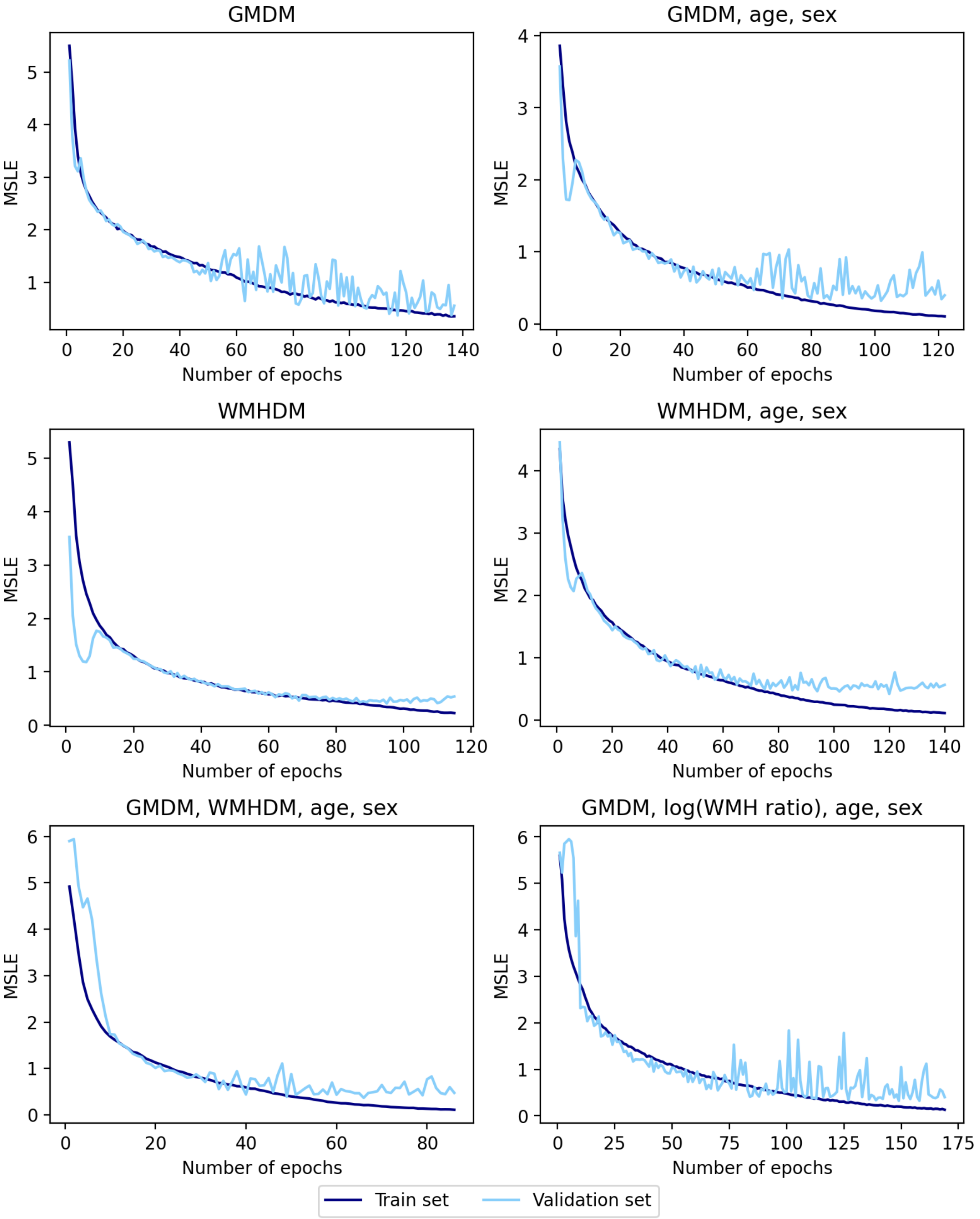


Figure H.3: Learning curves for ADAS13 models developed on low WMH load group. GMDM = Gray matter density map, MSLE = Mean squared logarithmic error, WMH = White matter hyperintensity, WMHDM = White matter hyperintensity density map.

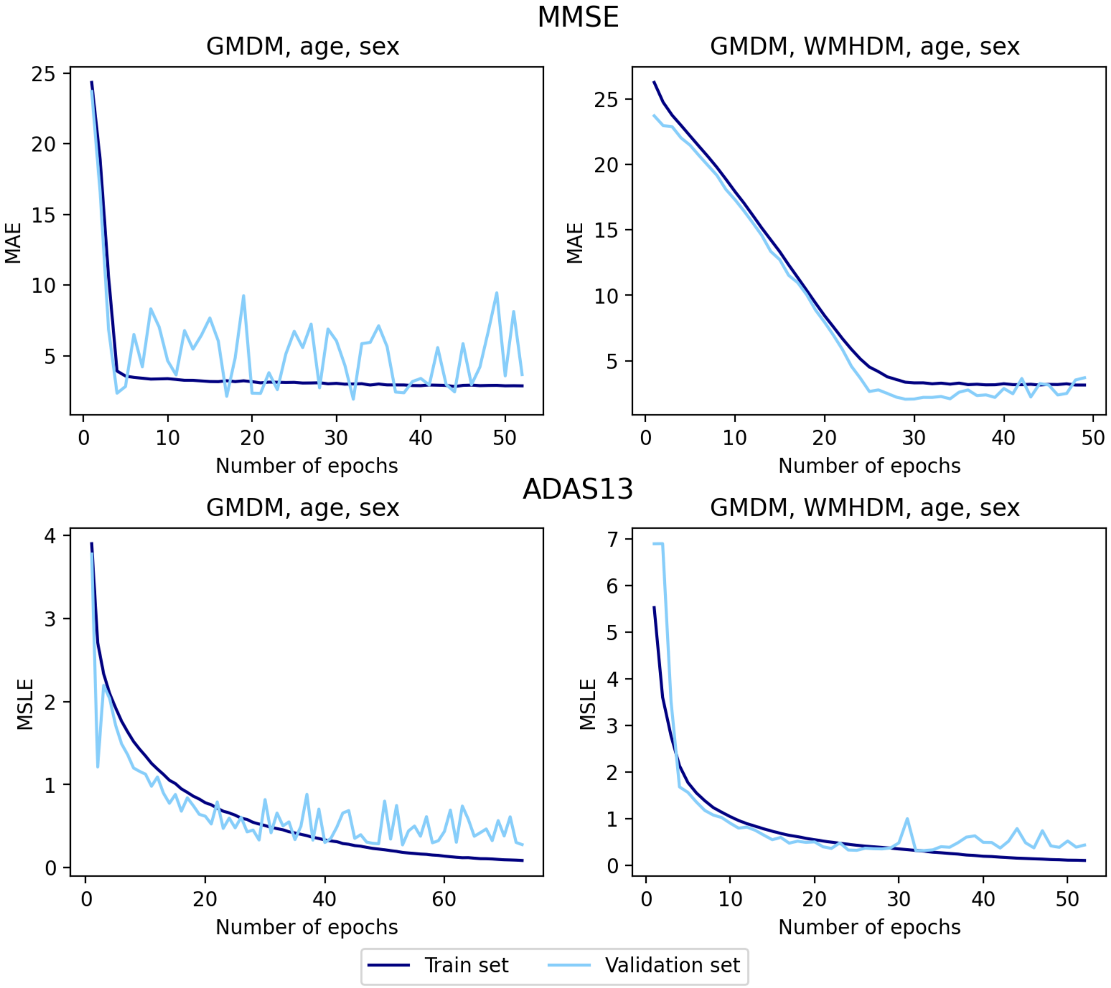


Figure H.4: Learning curves for models developed on all WMH load groups. GMDM = Gray matter density map, MAE = Mean absolute error, MSLE = Mean squared logarithmic error, WMHDM = White matter hyperintensity density map.

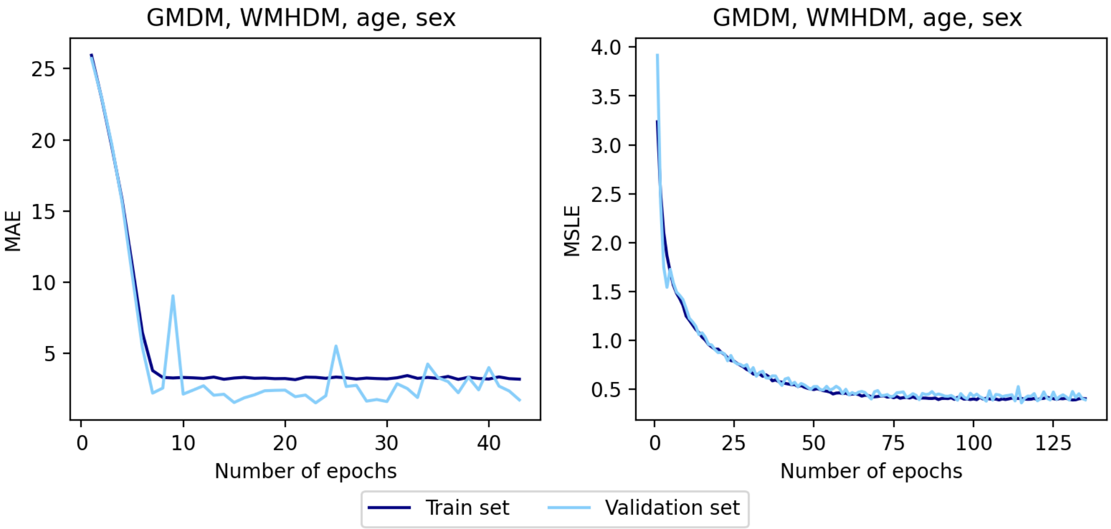


Figure H.5: Learning curves for model 5 when trained using shuffled batch cognitive scores. The left plot is the learning curve for MMSE, while the right plot shows the learning curve for ADAS13. GMDM = Gray matter density map, MAE = Mean absolute error, MSLE = Mean squared logarithmic error, WMHDM = White matter hyperintensity density map.