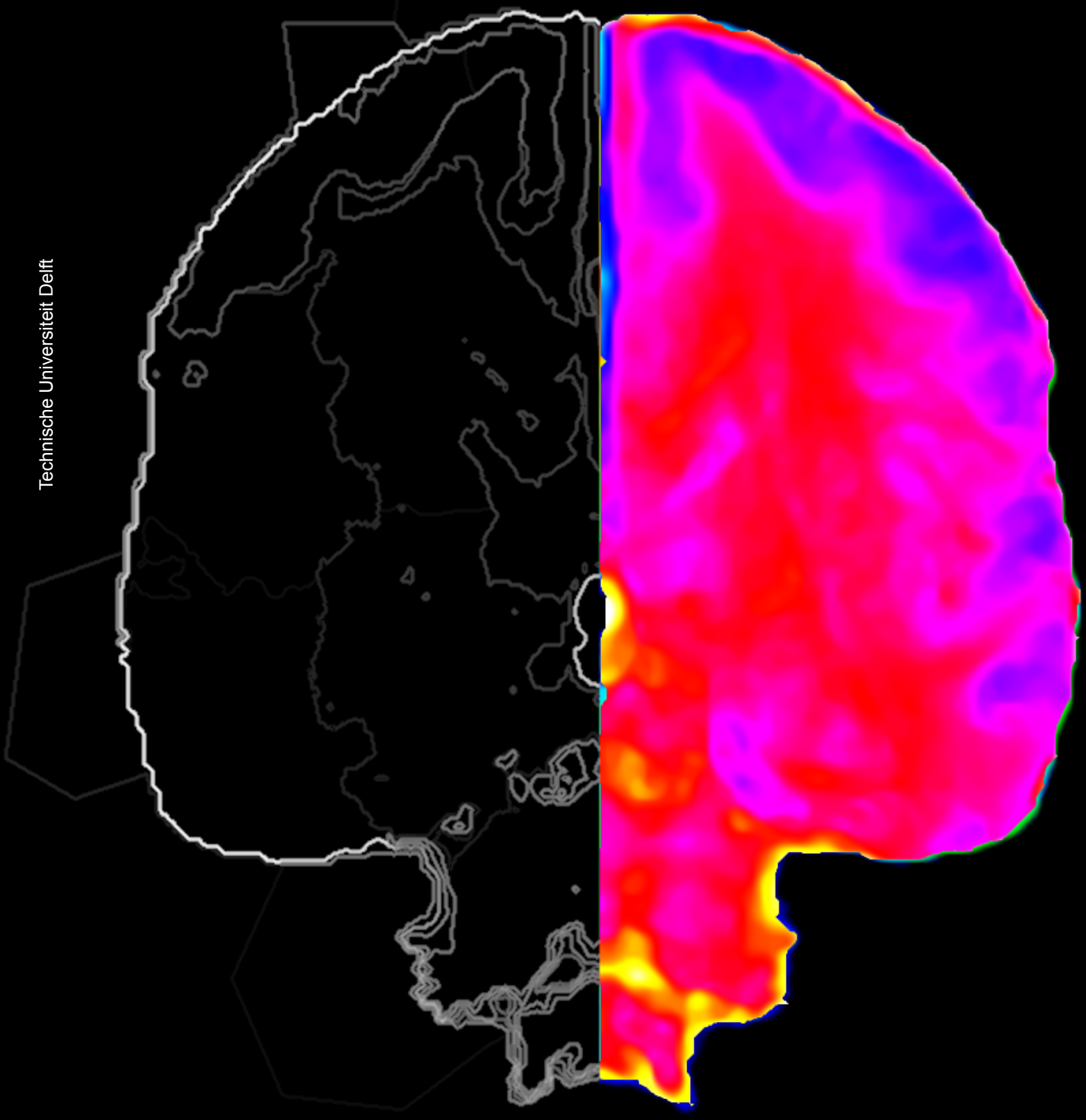
F.G. te Nijenhuis



Technische Universiteit Delft

Outcome prediction for endovascular therapy

Multimodal deep learning for acute ischemic events in the *arteria cerebri media*

Outcome prediction for endovascular therapy

Multimodal deep learning for acute ischemic events in the *arteria cerebri media*

by

F.G. te Nijenhuis

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday, the 30th of November at 10:00.

Student number: 5194342
Project duration: February 1, 2022 – October 1, 2022
Thesis committee: Dr. ir. J. C. van Gemert, TU Delft, supervisor
Dr. ir. X. Zhang, TU Delft
Dr. ir. T. Höllt, TU Delft

TU Delft

Preface

In these pages I describe my work on automated functional outcome prediction for stroke patients. The work was conducted at the Biomedical Imaging Group Rotterdam (BIGR), a research department within the Erasmus Medical Center, in conjunction with the Computer Vision Lab within the TU Delft.

I would like to thank dr. Theo van Walsum, dr. Xucong Zhang and Ruisheng Su for providing weekly supervision and guidance. Furthermore, I want to express my gratitude towards the additional members of the thesis committee, dr. Jan van Gemert and dr. Thomas Höllt. Additionally, I would like to thank the researchers and master's students at BIGR for the stimulating discussions, and for making the time spent at BIGR highly enjoyable.

Finally, I want to thank my family and friends for supporting me while I was studying to obtain a master's degree in Delft during the COVID-19 years. Overall, this has been a challenging period for me, and the work laid out before you in these pages is a testament to that, marking the culmination of multiple years of efforts. I would even go so far as to say that, symbolically, this thesis represents a major life event, and while I genuinely enjoyed working on it, I am excited and ready to move on to new things.

F.G. te Nijenhuis
Delft, September 2022

About the cover

The brain atlas which was used for image registration serves as the basis for the cover image. This atlas does not represent any single person's brain, instead, it is an 'average' brain of sorts. I performed a 3D projection of the CT images. I then rendered a 2D image at an interesting angle, showing the characteristic shape of a coronal section of the brain. By experimenting with image manipulation tools I was inspired to split the image into two separate hemispheres. I think this could very well represent the analytic and creative sides of the human brain, even though the myth that analytic and creative thought are located in opposite sides of the brain has been debunked. Nevertheless, significant amounts of technical as well as creative thought were required at different moments throughout the writing of this manuscript, so I think the image is still appropriate somehow. The reader is encouraged to come up with their own interpretation of the artwork, perhaps new ideas will emerge while reading this thesis!

Contents

1

Scientific article

Does multimodal deep learning improve functional outcome prediction of endovascular therapy for large vessel occlusion?

F.G. te Nijenhuis
Delft University of Technology

Abstract

The efficacy of endovascular therapy in large vessel occlusion (LVO) of the anterior circulation is dependent to a high degree on the selection of patients who are likely to benefit from this procedure. To this end, functional outcome prediction based on clinical parameters is an active area of research. In the preoperative screening of LVO patients, CT-Angiography (CTA) imaging is commonly acquired.

We compare the functional outcome prediction performance of multiple deep learning based classifiers with multiple conventional methods, including the clinically validated MR PREDICTS decision tool. Using a dataset composed of 1929 preprocedural CTA images combined with clinical data, we compare a clinical baseline model with an imaging based pipeline and a combined pipeline. For the imaging model backbone we train various state-of-the-art architectures (Med3D, Vision Transformer, Voxel Transformer). These models are used to predict dichotomized modified Rankin Scale score 90 days after mechanical thrombectomy. Binary classifier outcomes are quantified using Area-Under the receiver operating characteristic Curve (AUC). The activation maps of the best performing image based model are further investigated using the GradCAM++ post-hoc visualization method.

Combining clinical features with information extracted from CTA images does not significantly improve the performance of functional outcome prediction methods compared to the baseline model. The information extracted from the images does not seem to be complementary to the clinical features. This multimodal technique can however replace radiologically derived biomarkers, as its performance is non-inferior.

1. Introduction

Stroke is the second leading cause of death worldwide [20]. It leads to approximately 8,000 deaths in the Netherlands each year [51]. 366.000 Dutch people live with the sequelae of stroke, leading to a significant burden on the healthcare system [1].

In recent years, mechanical thrombectomy, also referred to as endovascular therapy (EVT), has emerged as an effective procedure for the treatment of acute ischemic stroke (AIS) in patients with a large vessel occlusion (LVO) [7–9,13,32]. Initial trials failed to convincingly show benefit of EVT over more conventional intravenous therapy (IVT); however it turned out that in selected subgroups of patients the beneficial effects of EVT compared to IVT were more substantial [26,40]. These results motivate the necessity of adequate selection of stroke patients, as better selection methods directly lead to improved outcomes after EVT.

Functional outcome after EVT is often quantified using the modified Rankin Scale (mRS), determined 90 days after occurrence of the stroke event. mRS ranges in multiple steps of increasing disability from a score of 0, indicating no sequelae, to 6, which indicates death. The mRS scale is shown in Table 1.

Multiple randomized trial based scoring methods have been developed to prognosticate functional outcome after EVT, using 90-day mRS (mRS90) as the outcome variable. Most of these methods are based on traditional statistical techniques. These methods are not equipped to directly extract information from more complicated forms of input data such as radiological images. Additionally, they often require radiological image biomarker information, which necessitates an arduous process of expert annotation. The process of biomarker extraction is further complicated by the oftentimes high degree of inter-observer variability [25].

AI based decision support systems might be of added benefit in this regard, either by automating the imaging biomarker extraction process, or by automating the entire functional outcome prediction process.

We hypothesize that information encoded in baseline imaging, which is performed for all stroke patients prior to EVT, can be extracted using deep learning based methods. A deep learning model that has been

mRS Score	Description	
0	No symptoms	Favorable outcome
1	No significant disability. Despite some symptoms, the patient is able to carry out their usual activities.	
2	Slight disability. Able to look out after own affairs without assistance, but unable to carry out all previous activities.	
3	Moderate disability. Requires some help, but able to walk unassisted.	Unfavorable outcome
4	Moderately severe disability. Unable to attend own bodily needs without assistance, and unable to walk unassisted	
5	Severe disability. Requires constant nursing care and attention, bedridden, incontinent.	
6	Death	

Table 1. Explanation of mRS scores. The scores can be dichotomized into a "Favorable" and "Unfavorable" outcome, as shown in the table [38].

trained to effectively predict functional outcome after EVT can potentially be utilized to inform clinical decision making.

1.1. Related work

Machine learning methods, as well as conventional statistical methods, have been used to predict functional outcome after thrombectomy.

1.1.1 Classical methods

Multiple prognostic scores of interventional outcome after EVT for LVO of the anterior circulation have been developed, such as Pittsburgh Response to Endovascular therapy (PRE), Totaled Health Risks in Vascular Events (THRIVE), THRIVE-c, Houston Intra-Arterial Therapy-2 (HIAT-2), Stroke Prognostication using Age and NIHSS (SPAN-100), NIHSS with age and volume (NAV) score, and MR PREDICTS [19, 21, 22, 30, 37, 39, 43, 48]. These scores use clinical information about the patient to predict functional outcome 90 days after the intervention, achieving moderate performance. Of the aforementioned models, MR PREDICTS attains the highest performance (Area-Under the ROC Curve (AUC) of 0.80) when all models are compared on a novel dataset [27]. Prognostic scores that incorporate neuroimaging based parameters, such as infarct size on CT or MRI or hemodynamic abnormalities on perfusion imaging, do not demonstrably improve performance, however, due to the heterogeneity of the different study populations it is difficult to directly compare results [23, 41, 44, 45].

1.1.2 Machine learning based methods

One potential approach to improve the performance of functional outcome prediction is by using machine learning based methods. Asadi et al. compared classical logistic regression methods with Multilayer Perceptron (MLP) and Support Vector Machine (SVM) based machine learning methods, showing that SVM works optimally [2]. SVM also performs better than classical methods when predicting the occurrence of symptomatic intracranial hemorrhage (SICH) following intravenous thrombolysis [5]. Nishi et al. compared pretreatment statistical methods to regularized logistic regression, random forest, and support vector machine, in the prediction of dichotomized mRS90, again demonstrating superiority of all machine learning methods [34]. Ramos et al. predicted mRS90 using multiple machine learning methods, with the best method, random forest, achieving an AUC of 0.81. Li et al. compared five machine learning methods with traditional methods and showed that, when predicting mRS90, machine learning yields superior results [29]. According to these results machine learning based methods can lead to superior functional outcome prediction performance compared to conventional methods when using only clinical input data.

1.1.3 Multimodal analysis

The potential for performance increase of machine learning methods over classical statistical methods becomes even greater when considering the incorporation of other types of data which are not readily amenable to meaningful statistical analysis, such as imaging data. Recent research has focused on predicting functional outcome using a combination of clinical features as

Paper	Image Modality	AUC
Zihni et al. [52]	TOF-MRA	0.76
Bacchi et al. [4]	NCCT	0.75
Samak et al. [42]	NCCT	0.75
Hilbert et al. [25]	CTA	0.71
De Graaf [15]	CTA	0.78

Table 2. Earlier work in mRS prediction after 90 days. TOF-MRA: Time Of Flight Magnetic Resonance Angiography, NCCT: Non-Contrast CT, CTA: CT Angiography.

well as imaging features extracted using deep learning based methods. Table 2 summarizes the previous work in this area. Zihni et al. used a combined pipeline with a Convolutional Neural Network (CNN) to extract imaging features as well as an MLP to process clinical data. The imaging data consisted of 3D volumes of TOF-MRA images. They show that an end-to-end multimodal pipeline integrating neuroimaging and clinical data leads to the best performance, with an AUC of 0.76 [52]. Bacchi et al. predict dichotomized mRS90 using several CNN and MLP based models, focusing on a combination of clinical and imaging data. For the imaging data, Non-Contrast CT (NCCT) scans are used. The best performing model is a combination of CNN and MLP, attaining an AUC of 0.75 [4]. Samak et al. also use NCCT imaging data, in combination with clinical information to achieve an AUC of 0.75 [42]. Hilbert et al. [25] and De Graaf [15] both successfully use deep learning models trained on CT Angiography (CTA) images, showing that these images also contain relevant information with regards to functional outcome prediction.

1.2. Contributions

Earlier work on automated functional outcome prediction has not convincingly shown improved performance of deep learning over conventional methods. We investigate a multimodal framework combining multilayer perceptron (MLP) processing of clinical features with the output of an image analysis backbone. We compare the performance of different state-of-the-art image processing backbone models with clinical baseline models, to investigate whether functional outcome prediction performance can be improved. The use of multiple different backbone models adds more weight to our conclusions regarding the difficulty of the problem, as we are able to draw model-agnostic conclusions. Training is performed on a dataset containing CTA images as well as clinical features of 1929 patients, which, to the best of our knowledge, is the largest dataset on which such an effort has been undertaken so far.

The remainder of this work is structured as follows. Section 2 describes the different learning strategies that were used. Section 3 outlines the datasets that were used, as well as the data preparation and preprocessing steps. Section 4 contains experimental results. Section 5 contains a discussion of the presented results and concludes the report.

2. Methods

We investigated three modeling strategies in predicting dichotomized mRS90. The simplest of these is the unimodal *clinical model*, which consists of a multilayer perceptron (MLP) containing four fully connected layers, which takes as its input patient data. The input features for the MLP correspond with those used by the MR PREDICTS model [48], and are shown in Table 3. For the nonlinearities we use Leaky ReLU [50]. The final layer of the *clinical model* contains a single unnormalized output, which is used for binary classification.

A schematic overview of the different processing pipelines is provided in Figure 1. As a baseline classifier we use a logistic regression with the coefficients derived from the updated MR PREDICTS decision tool [48].

The unimodal *imaging model* consists of a deep learning model, which we refer to as the backbone model, which takes CTA scans as input. We trained Med3D, VisionTransformer and VoxelTransformer architectures to serve as the backbone. The number of output features from the backbone model is variable, depending on the specific model used. A final linear layer again maps the output features from the backbone to a single unnormalized output value.

The third type of architecture we consider is the bimodal *combined model*, which concatenates the outputs of the *clinical* model with the outputs layer of the *imaging* model and feeds the combined output to a fully connected layer. The final output is again a single classification node. In this way, the *combined model* combines the imaging information with the clinical data by concatenating the features that were extracted by the imaging model with the patient features.

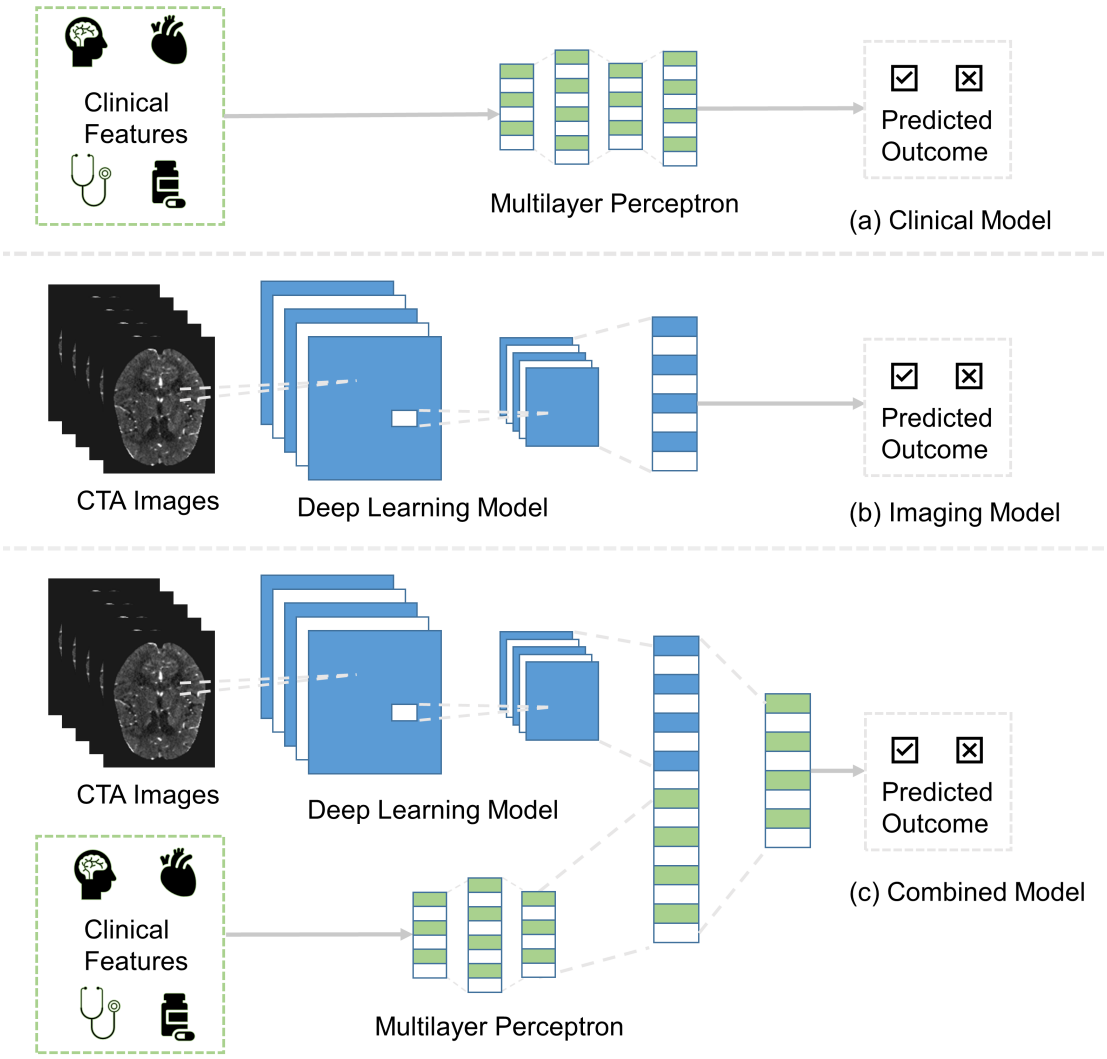


Figure 1. Schematic overview of the unimodal and multimodal architectures. (a) shows the unimodal clinical model, which processes clinical (and radiological) features, but not information extracted directly from CTA images, using an MLP model. (b) is a schematic representation of the unimodal imaging model, which uses a neural network as a backbone to directly infer functional outcome from the CTA images. Finally, (c) is a multimodal approach, combining the previous architectures by concatenating the outputs. DLM: Deep Learning Model. MLP: Multi-Layer Perceptron. FCL: Fully-Connected Layer.

2.1. Med3D

Med3D is a residual CNN specifically designed for medical image analysis. It is originally intended for segmentation purposes but can be reconfigured as a classification network. It consists of a modified ResNet backbone with an upsampling branch. The ResNet backbone was modified by changing the number of input channels from three to one, by expanding the 2D convolution operations to 3D convolutions, by setting the stride in layers three and four of the network to one, so there is no downsampling and by using dilated convolutions in the downstream convolutional layers. Med3D, like the ResNet on which it is based, comes in multiple sizes. In this work, we opt for the ResNet50 as our backbone model. The Med3D network was trained on a 3D segmentation dataset [12]. We initialize the Med3D architecture using the weights which were stored after training on the 3D segmentation dataset as described in [12]. We further modify the Med3D network by replacing the final segmentation layer with an average

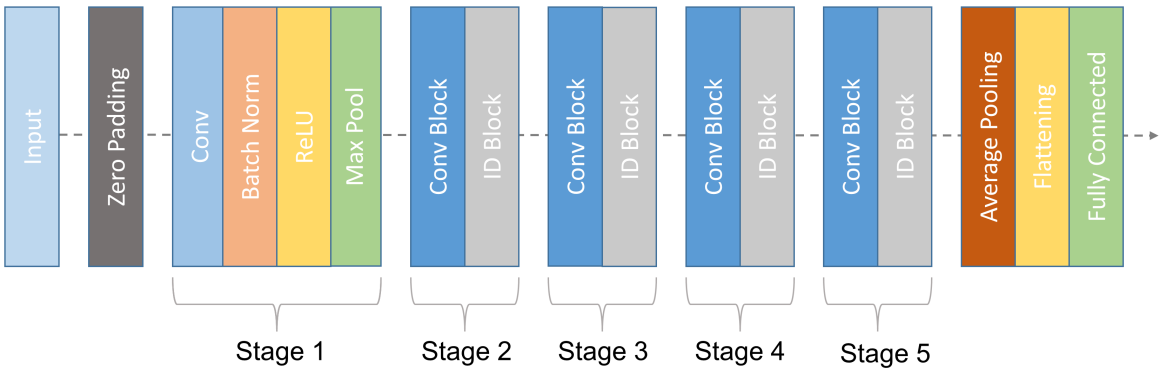


Figure 2. Schematic overview of the Med3D architecture, which is a modified version of the ResNet50 architecture. ResNet50 contains 50 layers in total, divided over multiple blocks. The identity (ID) block has the same input and output size, which the Conv block reduces the size of its input. For a detailed description of the original structure, we refer to [24]. Med3D modifies this structure mainly by extending the convolutions from 2D to 3D [12].

pooling layer followed by a linear layer mapping to 64 features.

2.2. Vision Transformer & Voxel Transformer

The Vision Transformer (ViT) is a modification of the Transformer natural language processing model, such that it can be used to handle visual tasks [17]. While the original ViT is described as a model which handles 2D input images, a three-dimensional extension of the ViT model is provided by the MONAI consortium, [14]. The Transformer relies on the self-attention mechanism as an alternative to the convolutional layer. An important benefit of the ViT model is that it requires less computational resources to train compared to conventional CNNs. Because the ViT employs the self-attention mechanism instead of convolutional layers it lacks the inductive biases commonly seen with CNNs, and as such it requires a relatively large training dataset to attain satisfactory performance. In the ViT model, we use the first token from the encoder output sequence as input for an MLP layer, using it as a classification token. A schematic representation of the ViT architecture is provided in Figure 3. We also use a modified version of the aforementioned architecture which performs global average pooling on the full output sequence, followed by a Fully-Connected Layer (FCL), which maps the output down to 2 nodes. We refer to this model as the Voxel Transformer (VoT).

2.3. Statistical analysis

We obtain AUC values and binary accuracy scores for each classifier on each validation dataset. For each classifier type, the fold with the best validation performance in terms of AUC is selected for subsequent statistical analysis. For each of the best validated classifiers, we obtain an Receiver Operating Characteristic (ROC) curve on the held out test data. We compare these curves using DeLong's test [16] to quantify the difference in performance. We also report calibration curves on the test set.

3. Data

We use the MR CLEAN registry, which is an ongoing prospective observational study involving 17 centers in the Netherlands. The registry contains data about patients who underwent EVT as a result of ischemic stroke. Inclusion of patients began in March 2014 [6].

For each patient the dataset contains demographic information, information about clinical as well as radiological parameters and patient outcome. Radiological features are defined as features which can only be determined by a radiologist looking at patient imaging. Clinical features can be determined using clinical information about the patient, which can be taken from patient records. The patient features are derived from the MR PREDICTS clinical decision tool and are displayed in Table 3. The accompanying preoperative CTA scan is also available for each included patient. For each patient, mRS90 was registered.

3.1. Image Data Selection

The dataset we investigate consists of two parts. Subset one contains 1000 patients, originating from previous work, where patients were selected from part one & two of the MR CLEAN registry [15]. Subset two contains 929 patients from part three of the MR CLEAN registry. Inclusion criteria differed between these cohorts.

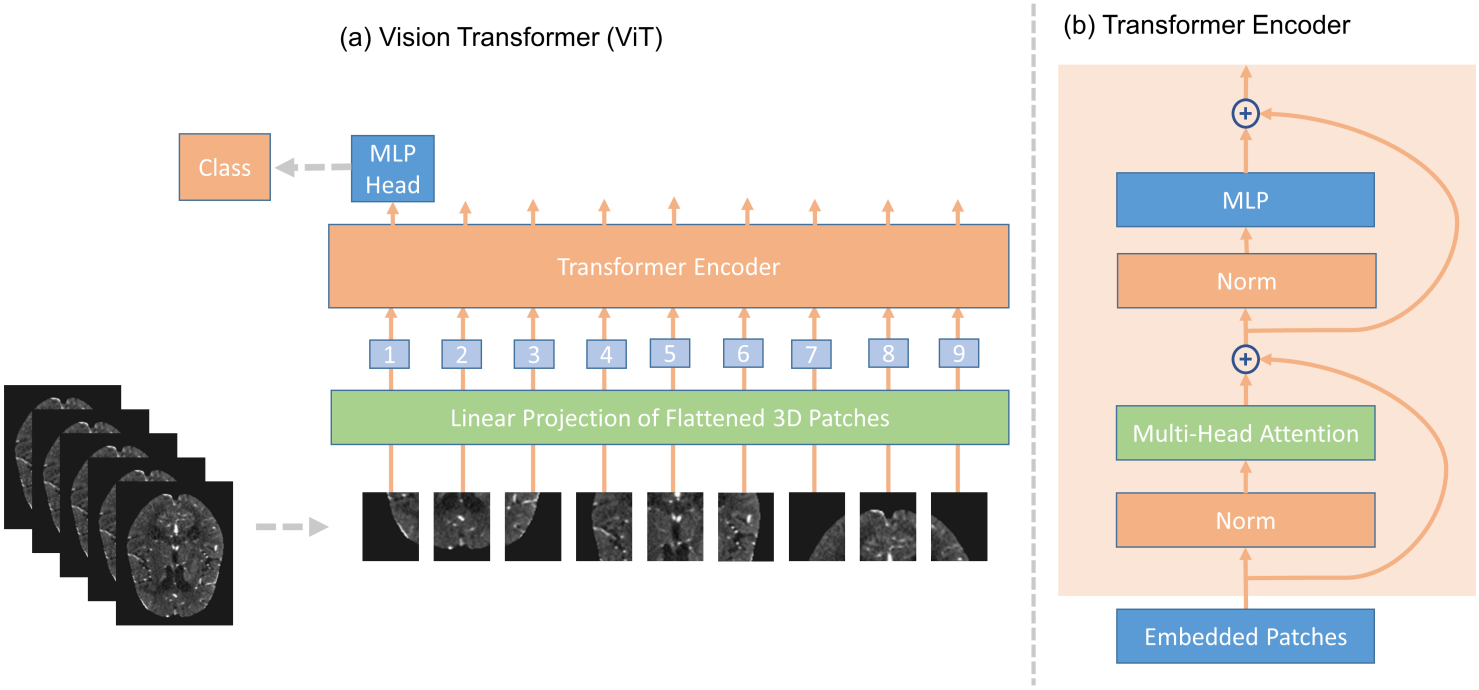


Figure 3. (a) shows a schematic representation of the Vision Transformer architecture. The CTA image is divided into patches. The patches are projected using a trainable linear projection. A positional encoding is added to each projected patch, and the patches are fed as a sequence to the transformer encoder. (b) shows the internal structure of a single layer in the encoder. The encoder contains multiple stacked encoding systems. For the Voxel Transformer variation of the architecture, see the supplement. Adapted from [17]

In both subsets, patients were included if the occlusion was located in the intracranial internal carotid arteries or in the M1, M2 or M3 segments of the middle cerebral arteries. Patients were only included if EVT was actually performed.

In subset one patients were excluded based on the images if the scan contained less than 50 slices, if slice thickness was greater than 1.5mm or if the slice spacing was greater than the thickness. Using manual inspection a single scan was selected for each patient. MR CLEAN registry parts one & two consisted of 3280 patients. Based on image selection criteria, 1711 patients were selected for preprocessing. Preprocessing was successful in 1480 cases. Of these patients, 480 were rejected due to an incorrect procedure, incorrect occlusion location or missing outcome variable. In the end, 1000 subjects were included in subset one.

In subset two, for each patient, the scan with minimal slice thickness was selected if there were multiple, excluding Maximum Intensity Projection (MIP) images. The slice thickness and spacing constraints were dropped for subset two. MR CLEAN registry part three contained 2003 subjects. Due to image selection criteria, 1603 were selected for registration. NifTI conversion and registration was successful in 1088 cases. These failure cases can be explained due to errors in the NIfTI conversion process. 159 cases were rejected due to the wrong procedure being performed, leaving 929 cases.

In total, 1929 patients were included. A flowchart depicting the inclusion process is shown in Figure 4.

3.2. Image data preprocessing

The raw CTA scans are heterogeneous in their image acquisition parameters, even after selection. To account for this, a number of processing steps are performed, which are also detailed in Figure 5. First, the images are converted from the DICOM format to NIFTI. NIFTI is a file format commonly used within the neuroimaging community, which is more suitable for image processing needs compared to the DICOM format [28].

After NiFTI conversion, affine registration to a brain atlas was performed using the ANTs software [3]. The construction of this atlas is described in Peter et al., [36]. Registration is performed in two steps. First, an affine transformation is performed to register the patient to the atlas. Second, a diffeomorphic trans-

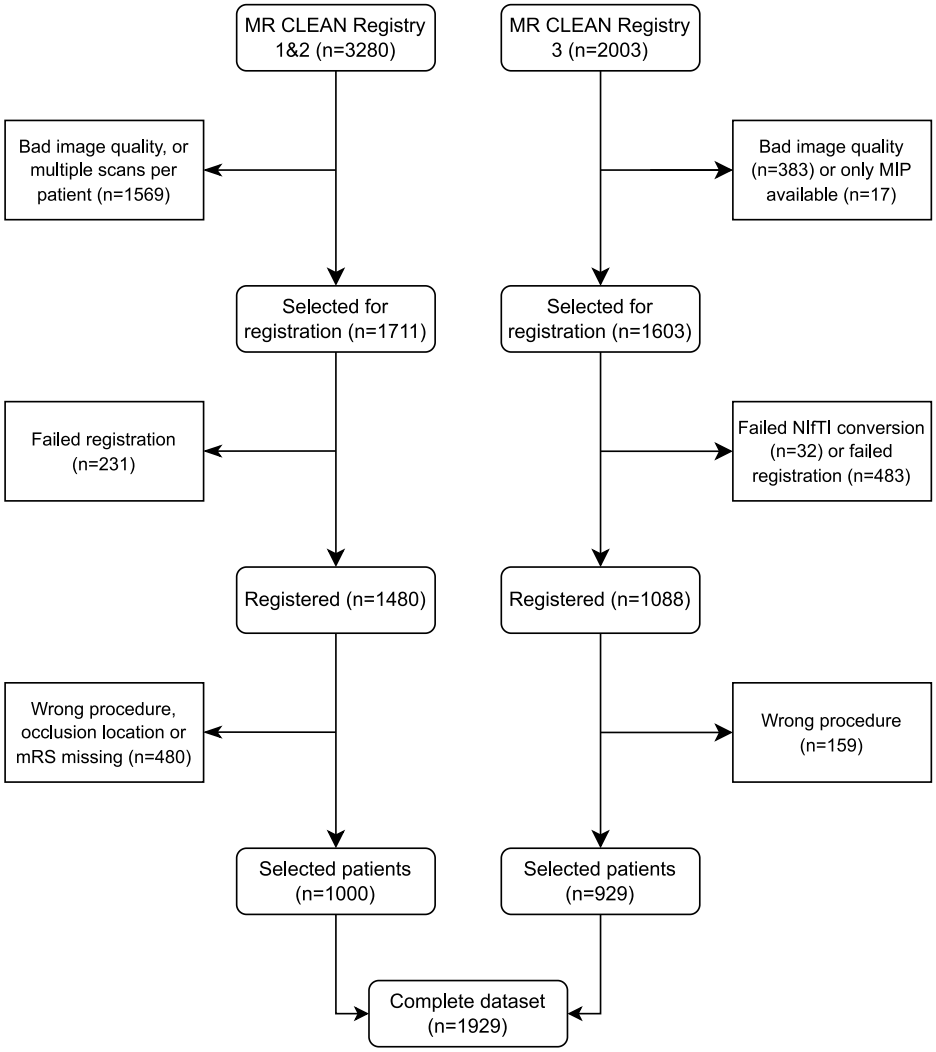


Figure 4. Flowchart depicting the inclusion process for scans from the MR CLEAN registry parts 1&2 (on the left) and for part 3 (on the right).

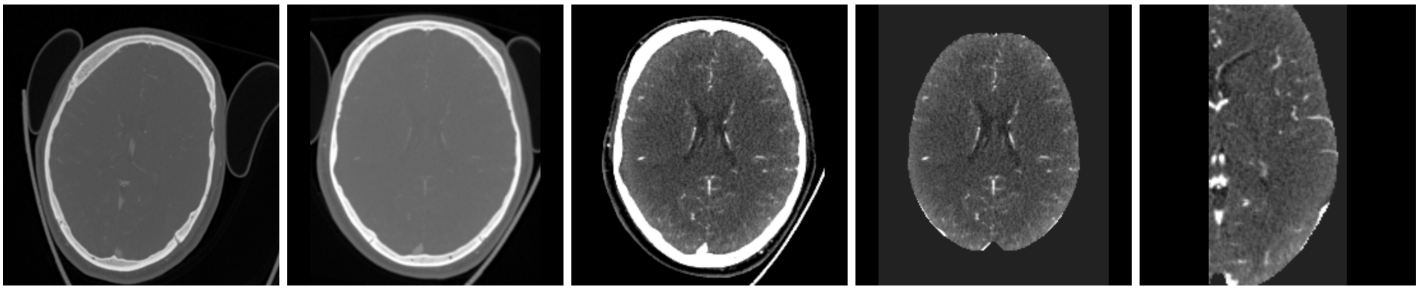


Figure 5. Illustration of the registration process, from left to right. A: original image, B: after registration to brain atlas, C: after normalization and clipping, D: after brain masking, E: after midline mirroring (if necessary) and hemisphere masking.

Feature	MR CLEAN Registry 1,2 & 3 (n=1929)	MR CLEAN Registry mRS90 0-2 (n=813)	MR CLEAN Registry mRS90 3-6 (n=1116)
Median Age (IQR)	72 (63-80)	69 (57-76)	76 (67-83)
Male % (n)	52% (1008)	57% (465)	49% (543)
Median ASPECTS (IQR)†	9 (8-10)	9 (8-10)	9 (8-10)
DM % (n)	17% (329)	11% (88)	22 % (241)
Mean Glucose (std)	7.40 (2.48)	7.01 (2.11)	7.70 (2.68)
Median baseline NIHSS (IQR)	16 (11-19)	13 (8-17)	17(13-21)
Median pre-stroke mRS (IQR)	0 (0-1)	0 (0-0)	0 (0-2)
Intravenous Alteplase % (n)	64% (1230)	71% (571)	59% (659)
Occlusion location % (n)†			
ICA	24.00 % (463)	19.80 % (161)	27.06% (302)
M1	57.02% (1100)	59.04 % (480)	55.56% (620)
M2	18.40 % (355)	20.66 % (168)	16.76% (187)
M3	0.10% (2)	0.25% (2)	0.00% (0)
Collateral score % (n)†			
Absent	4.56% (88)	1.85 % (15)	6.54 % (73)
<50%	36.13% (697)	28.17 % (229)	41.94 % (468)
>50<100%	38.00% (733)	41.94 % (341)	35.13% (392)
100%	20.11% (388)	26.69 % (217)	15.32% (171)
Median Systolic BP (IQR)	150 (132-167)	147 (130-163)	150 (135-170)
Median time-to-groin (IQR)	185 (136-265)	173 (130-250)	195 (145-275)

Table 3. Baseline characteristics for the patient, using the unimputed, unnormalized original data. Glucose in mmol/l, BP: Blood Pressure in mmHg.†: radiologically derived features.

formation was applied to register the atlas to the patient, so as to preserve the individual anatomical structure. Element-wise multiplication of the brain mask and the registered image was performed to remove the skull. Selected images were mirrored, such that the occluded vessel is always on the right side of the brain. The voxel values of the processed images were normalized to a range in $[0, 1]$. We only process the occluded hemisphere by cropping the image to a size of $(80 \times 112 \times 160)$ voxels, where each voxel represents a single cubic mm. Cropping is performed because early experiments showed no added benefit of using the complete brain versus only the affected hemisphere. Additionally, cropping the scan reduces the amount of computational resources required.

To standardize the depicted anatomy we clip the image by detecting the first slice which contains at least one voxel with an HU value greater than 1000, going in craniocaudal direction. We assume that detecting this voxel indicates the top of the skull. We define this slice as an anchor. We start 10mm in cranial direction from the anchor slice, and include 20cm worth of slices in caudal direction. This method ensures that we include the brain of the patient in a standardized way, excluding any additional anatomy, such as the neck or the aortic arch, that is sometimes included in the scans. Subsequently, we threshold the range of values of the image array between -40 and 260 HU.

3.3. Clinical and radiological data preprocessing

Clinical and radiological data contain both categorical as well as continuous variables. All values are normalized between zero and one to facilitate handling by the neural networks. In the complete dataset, the glucose value was missing in 8.7% of cases, and in 6.5% of cases there was no NIHSS follow-up information. For all other clinical variables values were present in at least 97.5% of cases. Missing data were imputed for all independent variables, using Multiple Imputation by Chained Equations (MICE, [46]) with a Gaussian Mixture estimator.

4. Experiments and Results

10% of the data ($n = 192$) is randomly sampled and held out as an independent test dataset. This test set is kept constant The remaining 90% of the data ($n = 1736$) is split using a stratified five-fold cross validation procedure, with a training set size of $n = 1388$ and a validation set containing $n = 348$ subjects.

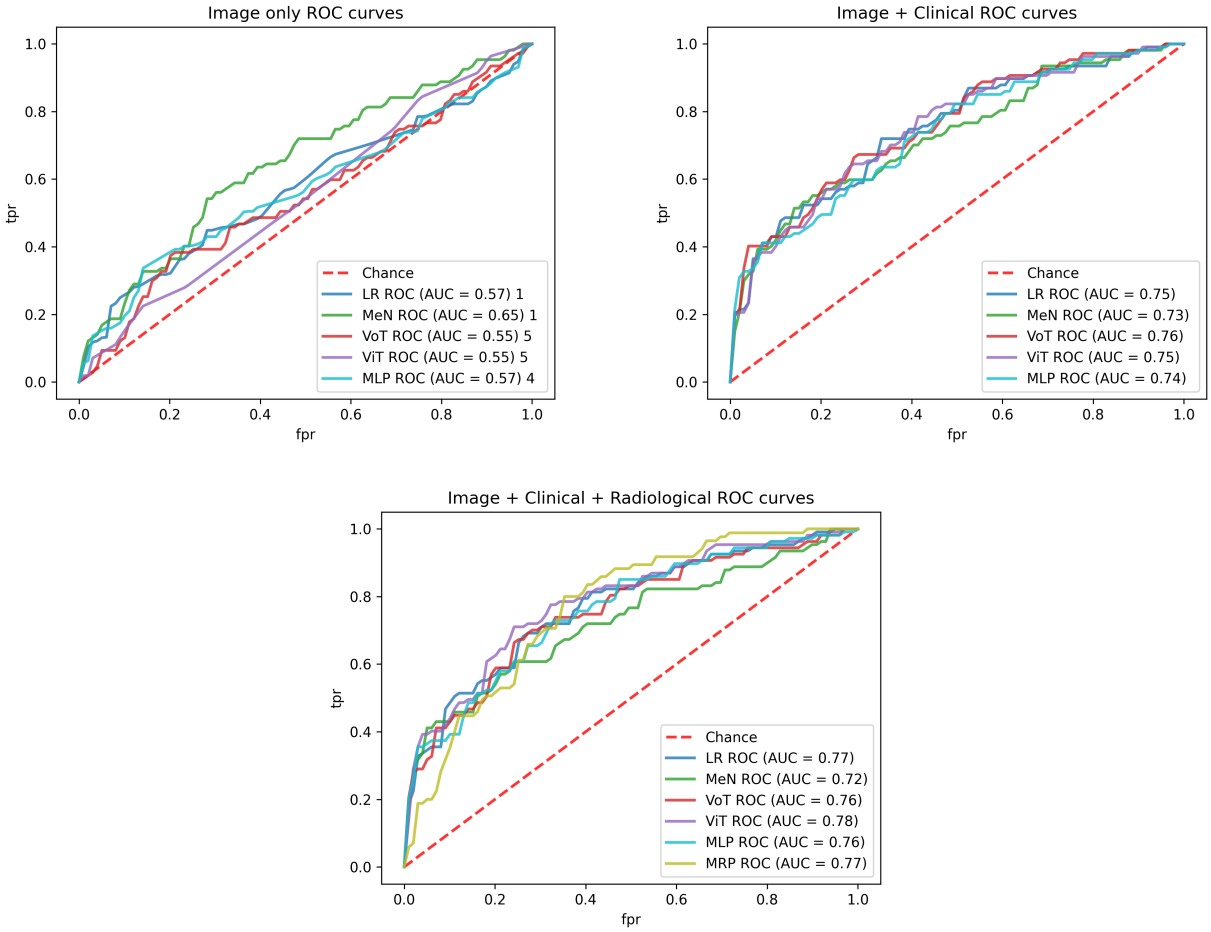


Figure 6. Receiver-Operating Characteristic (ROC) curves showing performance of the best fold (according to validation AUC) on the held out test set. LR: Logistic Regression, MeN: Med3D, VoT: Voxel Transformer, ViT: Vision Transformer, MLP: Multilayer Perceptron, MRP: MR PREDICTS logistic regression, tpr: true positive rate, fpr: false positive rate.

4.1. Implementation details

All experiments were performed on a single compute node on the Erasmus MC Radiology department GPU cluster. This node has an Intel Xeon Silver 4214 CPU, 192 GB of RAM and three P6000 24GB GPUs.

All software was written in Python, [47]. Machine learning code was written using the Pytorch framework, [35], augmented using the Pytorch Lightning framework, [18]. Registering the ML models was done using MLFlow, [11].

Each model was trained for 195 epochs. For each model, the initial learning rate was set at $1e^{-4}$. A plateau-based learning rate scheduler was used to multiply the learning rate by a factor of $1e^{-1}$ after 25 epochs without improvement of the validation loss, with a minimum achievable learning rate of $1e^{-6}$.

Early stopping was employed, such that training stopped when no noticeable increase in validation AUC is detected after 50 consecutive epochs. Additionally, a maximum of 50 epochs is set. Batch size was empirically selected to be 5 for the imaging based models due to limitations on the available GPU memory, and 40 for the clinical models.

Data augmentation was performed using random rotations and translations, with a probability of 0.5. Rotations were applied along each axis with a range $\in [-15°, 15°]$. Translations were applied at different settings per axis, such that the affected area would always remain visible. The translation settings were x-direction $\in [-5, 5]$ voxels, y-direction$\in [-15, 6]$ voxels and z-direction $\in [-5, 5]$ voxels.

Binary cross entropy is used as a loss function. This loss is calculated using the raw output logits. A sig-

		Image only	Image + Clinical	Image + Clinical + Radiological
LR†	AUC (± s.d)	0,58 (± 0,002)	0,75 (± 0,002)	0,70 (± 0,009)
	Accuracy (± s.d)	0,53 (± 0,005)	0,53 (± 0,005)	0,71 (± 0,009)
MLP†	AUC (± s.d)	0,62 (± 0,031)	0,79 (± 0,028)	0,74 (± 0,047)
	Accuracy (± s.d)	0,61 (± 0,016)	0,73 (± 0,015)	0,72 (± 0,056)
MeN	AUC (± s.d)	**0,68 (± 0,036)**	**0,82 (± 0,131)**	**0,88 (± 0,039)**
	Accuracy (± s.d)	**0,88 (± 0,082)**	**0,85 (± 0,120)**	**0,92 (± 0,030)**
VoT	AUC (± s.d)	0,62 (± 0,082)	0,69 (± 0,046)	0,72 (± 0,051)
	Accuracy (± s.d)	0,62 (± 0,071)	0,73 (± 0,024)	0,72 (± 0,019)
ViT	AUC (± s.d)	0,45 (± 0,015)	0,69 (± 0,046)	0,75 (± 0,051)
	Accuracy (± s.d)	0,57 (± 0,025)	0,72 (± 0,022)	0,72 (± 0,033)

Table 4. Average 5-fold validation performance of each classifier. Med3D shows the best average performance for each data category. †: These models cannot handle image inputs, so they only use clinical and radiological variables. In the case of the "Image only" dataset, they only received radiological features. See also Table 3 for the radiologically derived features.

moid operation is applied to the logits to obtain output probabilities. For optimization the AdamW optimizer is used [31]. Losses, AUC scores and binary accuracies were registered for the training and validation set during every epoch using the MLFlow framework [11].

4.2. Quantitative Evaluation

Average validation performance is displayed in Table 4. ROC curves produced by applying the best validated classifier folds on the held-out test set are displayed in Figure 6. Calibration curves are displayed in Figure 7. DeLong's test for comparison of ROC curves reveals that there is not a statistically significant difference between LR trained on the radiological features (AUC=0.57, n=192) and the Med3D model trained on CTA images only (AUC=0.65, z=1.19, p=0.235). Similarly, no statistically significant difference was found between LR and the Voxel Transformer (AUC=0.55, z=0.43, p=0.665), the Vision Transformer (AUC=0.55, z=0.46, p=0.643), or the Multilayer Perceptron (AUC=0.57, z=0.36, p=0.714).

Comparing logistic regression trained on the test set containing clinical features only (AUC=0.75, n=192) reveals no statistically significant difference with the Med3D model trained on CTA images and clinical features (AUC=0.73, z=0.72, p=0.471). Similar comparison between LR and the Voxel Transformer (AUC=0.76, z=0.62, p=0.535), the Vision Transformer (AUC=0.75, z=0.11, p=0.915) or the Multilayer Perceptron (AUC=0.74, z=0.89, p=0.373) again yield no statistically significant differences.

Finally, comparing the performance of the MR PREDICTS clinical decision tool ([48]) on the combined radiological and clinical features (AUC=0.77, n=192) with Med3D does not lead to a statistically significant result (AUC=0.73, z=1.12, p=0.263). The MR PREDICTS model performs comparably to the Voxel Transformer (AUC=0.76, z=0.46, p=0.645), the Vision Transformer (AUC=0.79, z=0.85, p=0.39) and the Multilayer Perceptron (AUC=0.76, z=0.35, p=0.721) on the test set, such that no statistically significant difference can be derived.

4.3. Post-Hoc Explainability

The best performing classifier on the "Image only" task, Med3D, was subjected to further analysis. Specifically, using the GradCam++ method [10] we visualize the activation maps of the final convolutional layer, overlaid on the input image for enhanced interpretability. Figure 8 shows the activations of the best model when presented with a single image in the testing dataset. The color intensity represents the contribution of the pixel with respect to the positive output class. From the image we can conclude that the network focuses on the vascular territory of the *arteria cerebri media* and the Sylvian fissure, as well as looking more diffusely at the frontal, temporal and parietal cortex. It seems like the ventricles, as well as the periventricular areas are also relevant to the classifier.

5. Discussion

We have demonstrated that multimodal deep learning can be used to predict functional outcome after mechanical thrombectomy in patients suffering from Large Vessel Occlusion of the anterior circulation, achieving similar performance to the MR PREDICTS clinical decision making tool. Inclusion of preprocessed imaging data using an end-to-end deep learning model does not, however, significantly improve performance compared to conventional statistical methods.

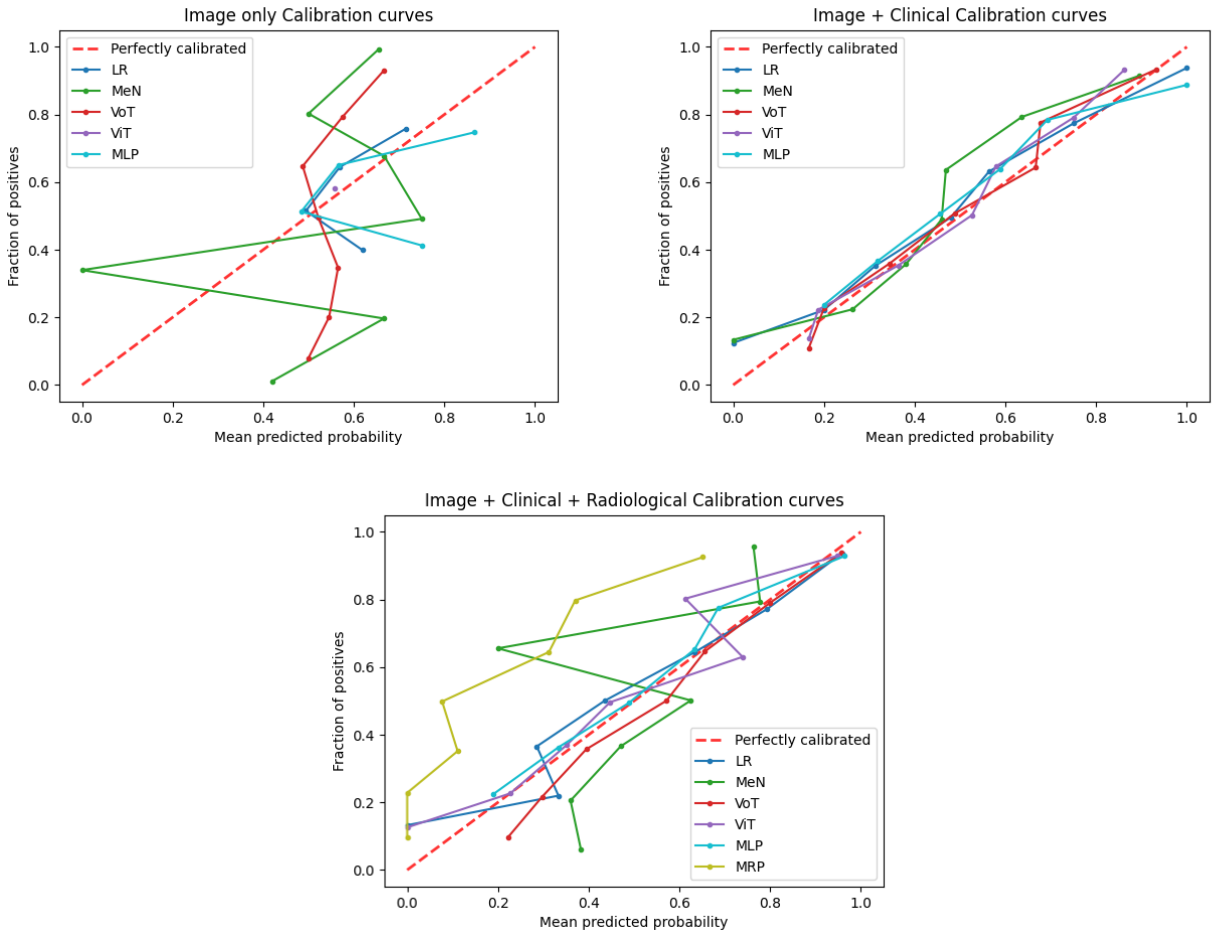


Figure 7. Calibration curves on the test set, for the best performing folds. LR: Logistic Regression, MeN: Med3D, VoT: Voxel Transformer, ViT: Vision Transformer, MLP: Multilayer Perceptron, MRP: MR PREDICTS logistic regression.

We have also shown that end-to-end deep learning can attain a similar performance when replacing the radiological features in the clinical model with a deep learning based backbone model. In particular, the Voxel Transformer model attains a similar performance on the test set using clinical features and CTA images compared to the MR PREDICTS model which uses radiological and clinical features.

When comparing the performance of the classifiers trained on the imaging data only with logistic regression trained on the radiological features, we see a moderate (but not statistically significant) increase in performance when using Med3D.

These results indicate that end-to-end deep learning models can extract latent information from imaging which is at least partially complementary to the information contained in clinical and radiological features. In particular, it seems like the imaging models look at cortical and periventricular atrophy to discern the age of the patient. Age is an important factor in the MR PREDICTS model. Other overlap of features is less readily explained. While deep learning does not improve functional outcome prediction performance, the results do suggest a potential role for deep learning in replacing the radiologically derived parameters. This is potentially clinically relevant as it means we can predict functional outcome without requiring a radiologist to manually extract radiological features from the scan, by relying on the Voxel Transformer.

The results are corroborated by post-hoc visualizations, which show that the model which performs best on the "imaging only" task, Med3D, mainly focuses on the vascular territory of the *arteria cerebri media*, as well as the surrounding cortical and ventricular regions. The surprising inclusion of the tissues surrounding the lesion can be interpreted as the model looking at cor-

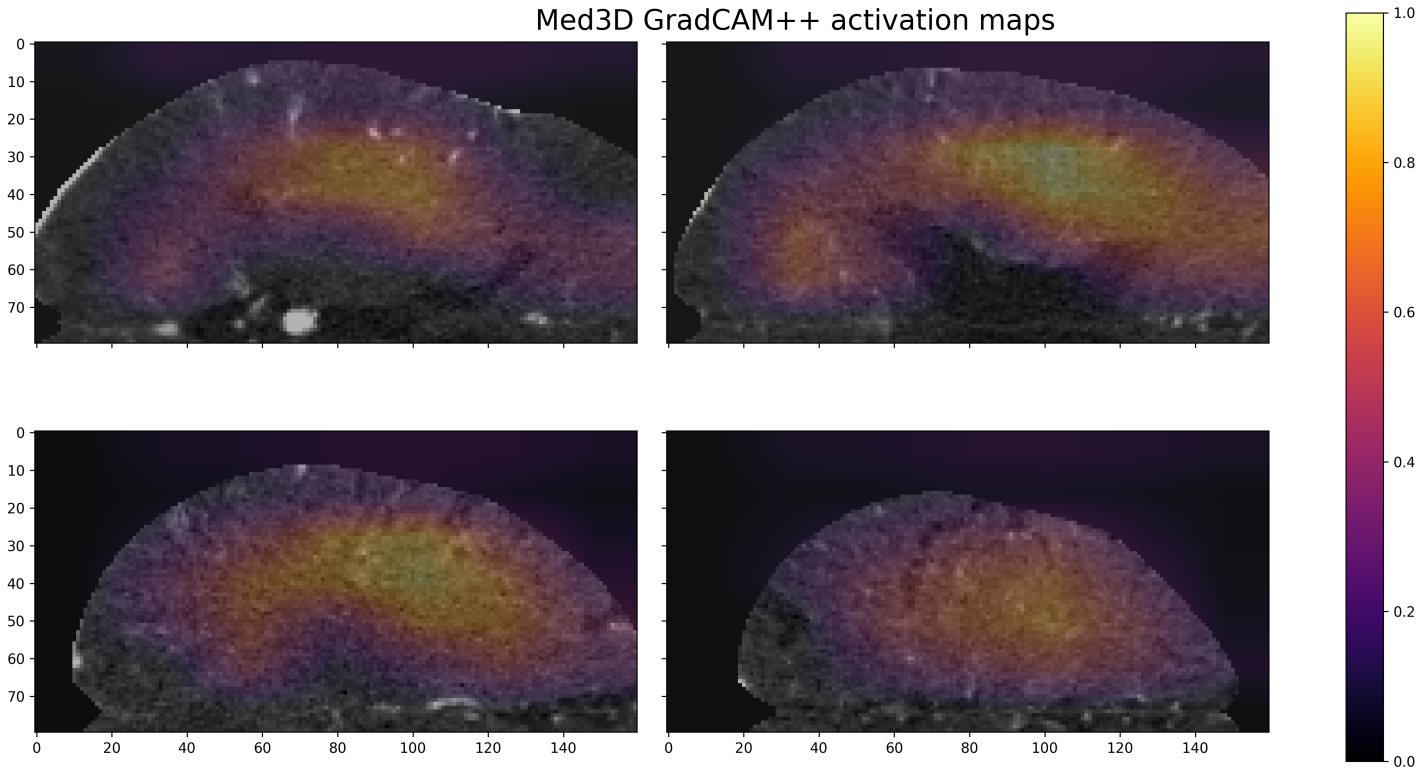


Figure 8. GradCAM++ output of the final activation layer of the Med3D model overlaid on four slices of a single input image. We see that the model mainly focuses on the lateral sulcus, where the M1 and M2 branches of the *arteria cerebri media* lie. The model also focuses on the periventricular space and on the cortex.

tical atrophy and subcortical leukoaraiosis, which are biomarkers for brain age [33]. Since age is a strong indicator of functional outcome in the clinical dataset, the information extracted from the image by the deep learning model might be partially redundant. Similarly, recent research has shown that gender related differences in brain structure exist at the macroscopic scale. Neural networks can learn to predict these differences [49]. It is possible that the classifier has learned to extract age as well as gender related features, explaining the moderately good performance on the "image only" task and lack of improvement once clinical and radiological features are added.

No significant difference is seen between the performance of the imaging based classifiers on the same datasets, suggesting that the problem itself is indeed hard. On the "Image + Clinical + Radiological" task, the performance of the Med3D architecture is relatively bad, while on the "Image only" task it is relatively good. This is most likely due to overfitting which rapidly occurs in the Med3D architecture. The Transformer based classifiers do not suffer from this phenomenon, possibly due to the lack of inductive biases, allowing these models to more readily extract information which is complementary to the clinical features.

We demonstrated non-inferiority of multimodal deep learning in combination with clinical features to the linear model with clinical and radiological features. While multimodal deep learning cannot be used for fully automated functional outcome prediction, there is a potential role for it in replacing the radiological features, potentially speeding up the outcome prediction process and alleviating some of the burden on the radiologist.

The MR PREDICTS clinical decision model was trained on a superset of the data on which it was tested, leading to potential bias through leakage of training data into the test set. Our reasoning for including the model was that it would allow for comparison between our classifiers and an actual decision tool used in clinical practice. As a simple logistic regression which did not suffer from this issue attained similar performance we accepted this potential bias.

The affine registration process which is performed

before inference in the imaging based classifiers takes approximately 5 minutes, and is prone to errors in some cases. This is potentially problematic when introducing such a classifier in clinical care.

We excluded subjects where the registration process was not successful. Registration often fails when artifacts were introduced, for instance if the patient moved. We hypothesize that there might be a correlation between the success chance of the registration and the outcome of the patient, as patient who are less capable of following instructions during the scanning process might be more seriously affected by the stroke, leading to a less favorable outcome in this patient group. This might result in a selection bias, where only relatively healthy patients are included due to their scans being usable.

One potential way to improve performance is to combine the information from multiple imaging modalities. In this study we only investigated the usability of neural networks when analyzing CTA images, but one advantage of deep learning is that it can combine information in unforeseen ways, leveraging imaging information at the sub-visual level. One potential new avenue of research would be to investigate whether addition of non-contrast CT, for instance as a different input channel to the classifier, would yield better results.

The preprocessing steps take time, which complicates clinical applicability of the classifiers. It would be interesting to see whether training the networks on the raw data would lead to similar performance.

It is not yet clear to what extent features extracted from the images are similar to the clinical and radiological features, future work should investigate which features are most relevant to the classifiers, and why. One potential way to do this is by regressing on the MR PREDICTS variables using the raw features extracted by the imaging classifiers, another potential approach is by calculating Shapley values for the different features.

6. Conclusion

We set out to investigate the potential usefulness of deep learning methods in improving functional outcome prediction after mechanical thrombectomy in stroke patients. Results did not indicate a significant performance improvement of multimodal, end-to-end deep learning methods, which combine imaging data with patient features, compared to conventional statistical methods based only on patient features. Future research must reveal whether such multimodal deep learning techniques can be of concrete clinical value.

References

[1] Cijfers beroerte. [Online, accessed 23 Jun 2022]. 1

[2] H. Asadi, R. Dowling, B. Yan, and P. Mitchell. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. *PLoS One*, 9(2):e88225, 2014. 2

[3] Brian Avants, Nick Tustison, and Gang Song. Advanced normalization tools (ants). *Insight J*, 1–35, 11 2008. 6

[4] Stephen Bacchi, Toby Zerner, Luke Oakden-Rayner, et al. Deep learning in the prediction of ischaemic stroke thrombolysis functional outcomes: A pilot study. *Academic Radiology*, 27(2):e19–e23, 2020. 3

[5] Paul Bentley, Jeban Ganesalingam, Anoma Lalani Carlton Jones, et al. Prediction of stroke thrombolysis outcome using ct brain machine learning. *NeuroImage: Clinical*, 2014. 2

[6] Olvert A. Berkhemer, Puck S.S. Fransen, Debbie Beumer, et al. A randomized trial of intraarterial treatment for acute ischemic stroke. *New England Journal of Medicine*, 372(1):11–20, 2015. PMID: 25517348. 5

[7] Debbie Beumer, Julie Staals, Jeannette Hofmeijer, et al. A randomized trial of intraarterial treatment for acute ischemic stroke. *The New England Journal of Medicine*, 2015. 1

[8] Joseph P. Broderick, Yuko Y. Palesch, Andrew M. Demchuk, et al. Endovascular therapy after intravenous t-pa versus t-pa alone for stroke. *The New England Journal of Medicine*, 2013. 1

[9] Bruce C.V. Campbell, Leonid Churilov, Nawaf Yassi, et al. Endovascular therapy for ischemic stroke with perfusion-imaging selection. *The New England Journal of Medicine*, 2015. 1

[10] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++ generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, mar 2018. 10

[11] Andrew Chen, Andy Chow, Aaron Davidson, et al. Developments in mlflow: A system to accelerate the machine learning lifecycle. In *Proceedings of the Fourth International Workshop on Data Management for End-to-End Machine Learning*, DEEM'20, New York, NY, USA, 2020. Association for Computing Machinery. 9, 10

[12] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *CoRR*, abs/1904.00625, 2019. 4, 5

[13] Alfonso Ciccone, Luca Valvassori, Michele Nichelatti, et al. Endovascular treatment for acute ischemic stroke. *The New England Journal of Medicine*, 2013. 1

[14] The MONAI Consortium. Project monai, Dec. 2020. 5

[15] Samantha de Graaf. Automated functional outcome prediction in stroke using combined imaging and clinical parameters. Master's thesis, TU Delft, the Netherlands, 2022. 3, 5

[16] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, Sep 1988. 5

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. 5, 6

[18] William Falcon et al. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3, 2019. 9

[19] Kyle M. Fargen, Imran Chaudry, Raymond D Turner, et al. A novel clinical and imaging based score for predicting outcome prior to endovascular treatment of acute ischemic stroke. *Journal of NeuroInterventional Surgery*, 2013. 2

[20] V. L. Feigin, B. A. Stark, C. O. Johnson, et al. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Neurol*, 20(10):795–820, 10 2021. 1

[21] Alexander C. Flint, Sean P. Cullen, Bonnie Faigeles, et al. Predicting long-term outcome after endovascular stroke treatment: the totaled health risks in vascular events score. *American Journal of Neuroradiology*, 2010. 2

[22] Alexander C. Flint, Vivek A. Rao, Sheila L. Chan, et al. Improved ischemic stroke outcome prediction using model estimation of outcome probability: The thrive-c calculation. *International Journal of Stroke*, 2015. 2

[23] Francesca Di Giuliano, Eliseo Picchi, Fabrizio Sallustio, et al. Accuracy of advanced ct imaging in prediction of functional outcome after endovascular treatment in patients with large-vessel occlusion. *The Neuroradiology Journal*, 32(1):62–70, 2019. PMID: 30303448. 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[25] A. Hilbert, L.A. Ramos, H.J.A. van Os, et al. Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke. *Computers in Biology and Medicine*, 115:103516, 2019. 1, 3

[26] David M. Kent, Peter M. Rothwell, John P. A. Ioannidis, et al. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials*, 2010. 1

[27] F. Kremers, E. Venema, M. Duvekot, et al. Outcome Prediction Models for Endovascular Treatment of Ischemic Stroke: Systematic Review and External Validation. *Stroke*, 53(3):825–836, 03 2022. 2

[28] Xiangrui Li, Paul S. Morgan, John Ashburner, et al. The first step for neuroimaging data analysis: Dicom to nifti conversion. *Journal of Neuroscience Methods*, 264:47–56, 2016. 6

[29] X. Li, X. Pan, C. Jiang, et al. Predicting 6-Month Unfavorable Outcome of Acute Ischemic Stroke Using Machine Learning. *Front Neurol*, 11:539509, 2020. 2

[30] John Liggins, Albert J Yoo, Albert J Yoo, et al. A score based on age and dwi volume predicts poor outcome following endovascular treatment for acute ischemic stroke. *International Journal of Stroke*, 2015. 2

[31] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. 10

[32] David J. McCarthy, Anthony Diaz, Dallas Sheinberg, et al. Long-term outcomes of mechanical thrombectomy for stroke: A meta-analysis. *The Scientific World Journal*, 2019. 1

[33] J. S. Meyer, S. Takashima, Y. Terayama, et al. CT changes associated with normal aging of the human brain. *J Neurol Sci*, 123(1-2):200–208, May 1994. 12

[34] Hidehisa Nishi, Naoya Oishi, Akira Ishii, et al. Predicting clinical outcomes of large vessel occlusion before mechanical thrombectomy using machine learning. *Stroke*, 2019. 2

[35] Adam Paszke, Sam Gross, Francisco Massa, et al. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019. 9

[36] Roman Peter, Bart J. Emmer, Adriaan C.G.M. van Es, and Theo van Walsum. Cortical and vascular probability maps for analysis of human brain in computed tomography images. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 1141–1145, 2017. 6

[37] Srikant Rangaraju, Amin Aghaebrahim, Christopher Streib, et al. Pittsburgh response to endovascular therapy (pre) score: optimizing patient selection for endovascular therapy for large vessel occlusion strokes. *Journal of NeuroInterventional Surgery*, 2015. 2

[38] J. RANKIN. Cerebral vascular accidents in patients over the age of 60. II. Prognosis. *Scott Med J*, 2(5):200–215, May 1957. 2

[39] Syed Ali Raza and Srikant Rangaraju. A review of pre-intervention prognostic scores for early prognostication and patient selection in endovascular management of large vessel occlusion stroke. *Interventional Neurology*, 2018. 2

[40] Peter M. Rothwell. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, 2005. 1

[41] F. Sallustio, N. Toschi, A. P. Mascolo, et al. Selection of anterior circulation acute stroke patients for mechanical thrombectomy. *J Neurol*, 266(11):2620–2628, Nov 2019. 2

[42] Zeynel Abidin Samak, Phil Clatworthy, and Majid Mirmehdi. Prediction of thrombectomy functional outcomes using multimodal data. 05 2020. 3

[43] Gustavo Saposnik, Amy K Guzik, Mathew J. Reeves, et al. Stroke prognostication using age and nih stroke scale: Span-100. *Neurology*, 2013. 2

[44] Fatima Soliman, Fatima Soliman, Ajay Gupta, et al. The role of imaging in clinical stroke scales that predict functional outcome: A systematic review. *The Neurohospitalist*, 2017. 2

[45] I.Y.L. Tan, Andrew M. Demchuk, Andrew M. Demchuk, et al. Ct angiography clot burden score and collateral score: correlation with clinical and radiologic outcomes in acute middle cerebral artery infarct. *American Journal of Neuroradiology*, 2009. 2

[46] Stef Van Buuren and Karin Oudshoorn. *Flexible multivariate imputation by MICE*. Leiden: TNO, 1999. 8

[47] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. 9

[48] Esmee Venema, Bob Roozenbeek, Maxim J.H.L. Mulder, et al. Prediction of outcome and endovascular treatment benefit: Validation and update of the mr predicts decision tool. *Stroke*, 52(9):2764–2772, 2021. 2, 3, 10

[49] Jiang Xin, Yaoxue Zhang, Yan Tang, and Yuan Yang. Brain differences between men and women: Evidence from deep learning. *Frontiers in Neuroscience*, 13, 2019. 12

[50] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network, 2015. 3

[51] R.H. Wimmers Y. Koop. Hart- en vaatziekten in nederland, 2021. [Online, accessed 15 Jun 2022]. 1

[52] Esra Zihni., Vince Madai., Ahmed Khalil., et al. Multimodal fusion strategies for outcome prediction in stroke. In *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF,*, pages 421–428. INSTICC, SciTePress, 2020. 3

A. Supplementary Material

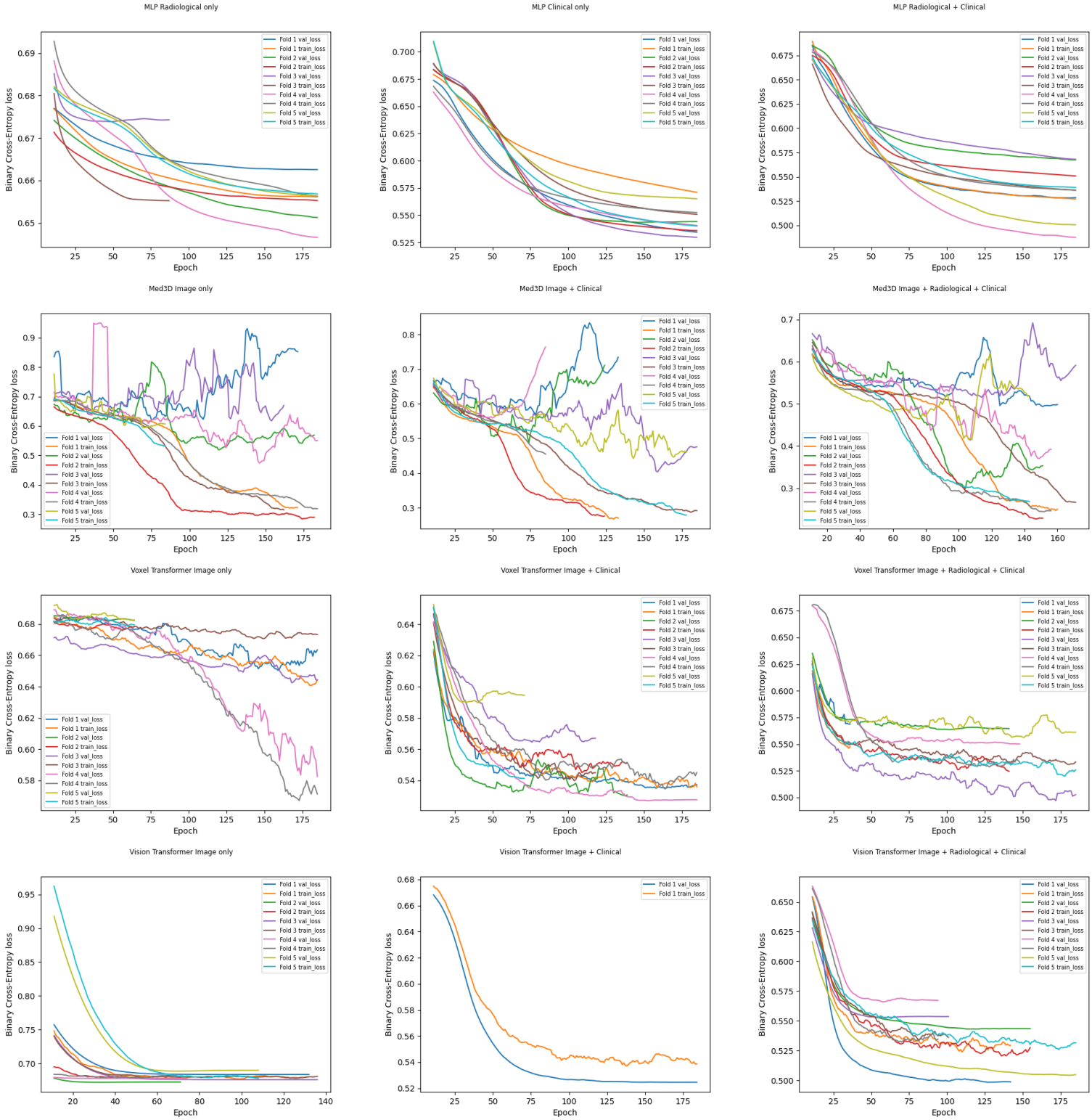


Figure 9. Learning curves of the different pipelines. For each model, validation and training loss curves over the five folds are displayed. Note that Vision Transformer Image + Clinical is missing some curves due to computational issues.

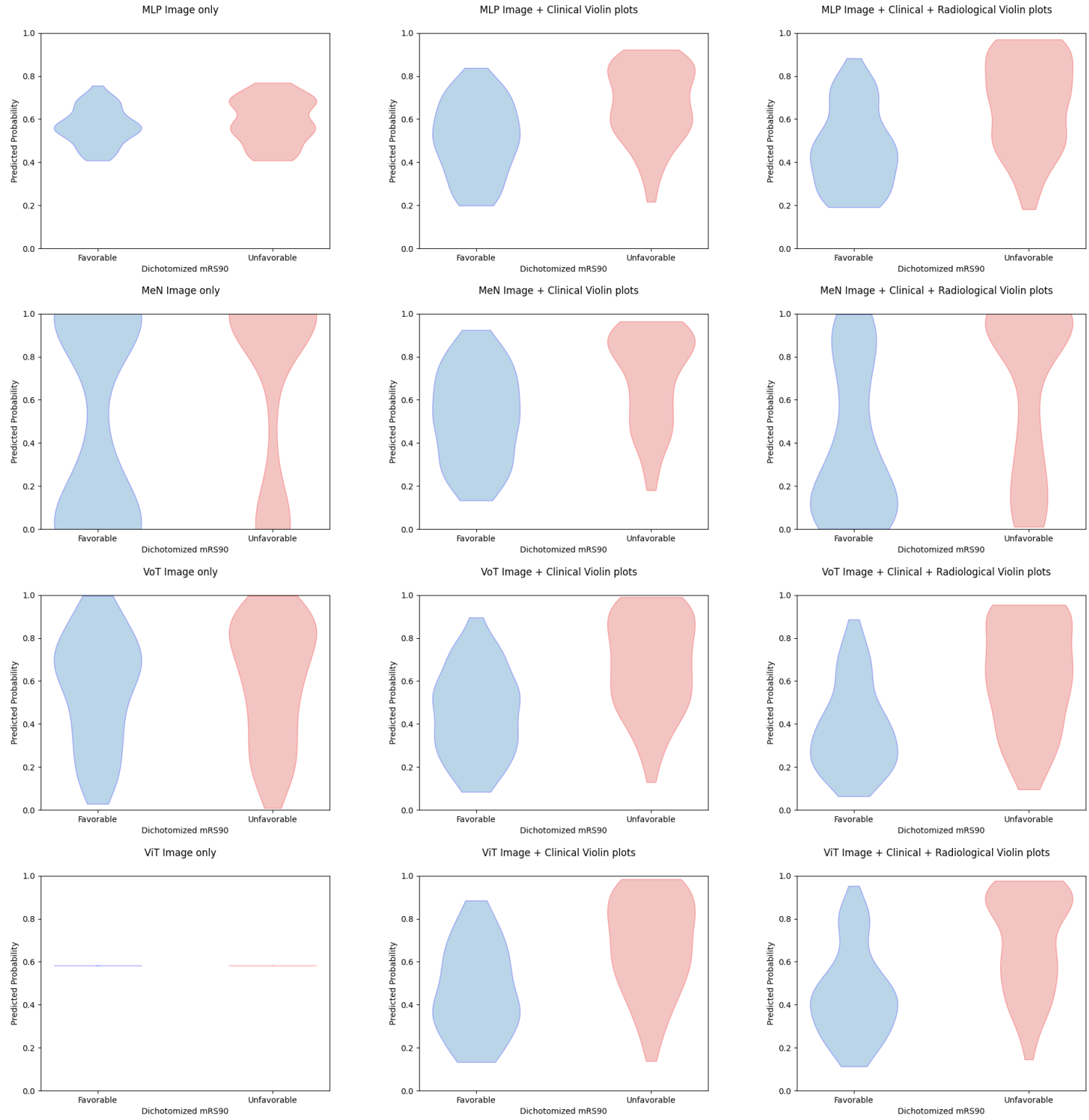


Figure 10. Violin curves. A favorable outcome corresponds to an output of 0, while an unfavorable outcome is predicted using a 1. The test set contains 192 subjects, and predicted probabilities are displayed for each outcome. An ideal plot would entail all probability mass for the favorable outcome concentrated around 0, while the unfavorable outcome would be concentrated around 1.

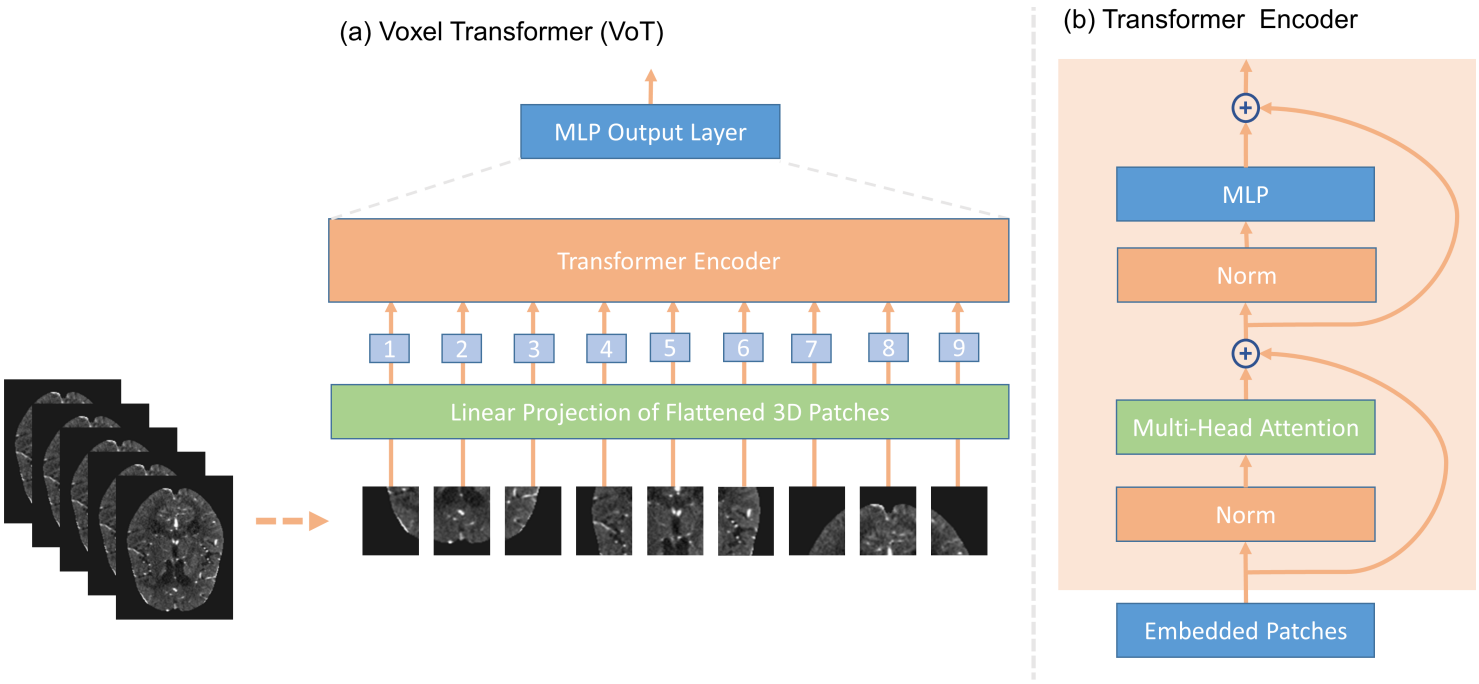


Figure 11. Voxel Transformer architecture. This architecture is similar to the one described in Figure 3, with the only difference being the output. Instead of using only the first token for classification as with the Vision Transformer, this model takes the entire output sequence and feeds it to a final MLP layer to arrive at an output.

2

Stroke & Functional Outcome Prediction

Stroke is a medical condition in which there is a reduced blood flow to a part of the brain. Approximately 80% of strokes are ischemic (caused by an obstruction in the blood vessel), the other 20% are hemorrhagic (caused by a bleeding). Each year, approximately 40.000 people suffer from a stroke in the Netherlands. Stroke is the cause of 8.000 deaths annually in the Netherlands [22]. The symptoms of stroke depend on the specific location of the obstruction, as the affected brain region dictates the exact complaints. Traditional symptoms of stroke include hemiparesis, aphasia, dysarthria, facial weakness, ataxia and vertigo [15], see Table 2.1a. As each cerebral hemisphere controls the contralateral (opposite) side of the body, an obstruction in the left hemisphere will result in hemiparesis of the right side of the body and vice-versa. Apart from these traditional symptoms, many different non-traditional symptoms may also occur if the the occlusion occurs in a less conventional location.

2.1. Pertinent Neuro-anatomy

The arterial supply of the brain can be divided into an anterior and a posterior circulation. The posterior part is mainly supplied by the vertebral arteries, which travel through openings (foramina) in the spinal column of the neck. The anterior part of the vasculature of the brain is supplied by the carotid arteries, which travel up through the neck. The anterior and posterior circulation connect in the circle of Willis, a circular arterial structure at the base of the brain. The purported anatomical purpose of the circle of Willis is to preserve some degree of blood flow to the brain in the event of an obstruction of one of the supplying blood vessels. There is great anatomical variation in the degree of completeness of this structure among different people. From the circle of Willis, three cerebral arteries arise to provide blood to almost the entire brain. The cerebral arteries are paired on the left and right, with anterior, middle and posterior arteries on each side leading to six arteries in total. Of these three blood vessels, the middle cerebral artery is the largest vessel. It is the vessel most commonly affected by stroke, accounting for over half of all cases. The middle cerebral artery consists of M1, M2, M3 and M4 segments. These vessels provide blood to a significant part of each hemisphere, supplying part of the frontal, temporal and parietal lobes, as well as other important structures such as the thalamus, the internal capsule and the caudate nucleus [17]. Figure 2.1 shows the course of the MCA through the brain, and Figure 2.2 shows a detail of the route of the vessel through the brain. A detailed explanation of the anatomy and neurophysiology of these structures is beyond the scope of this text.

2.2. Stroke imaging

According to most protocols in the Netherlands, initial non-contrast CT imaging should be performed on every patient with stroke symptoms. A CT (Computed Tomography) scan is an image of the anatomy of the patient. The CT scan is made by rotating an X-ray emitter tube, with detector arrays on the opposite side, around the patient. Different types of tissue attenuate the signal differently. Dense tissues such as bone lead to high attenuation of the X-ray beam, while soft tissues lead to less attenuation. By mathematically combining the attenuation at each rotation angle, an image can be recovered. Many different types of CT scans exist. The amount of attenuation at each voxel location in the scan can be quantified using Hounsfield Units (HU). The HU values for different tissues are provided in Table 2.2.

Symptom	Description
Hemiparesis	Weakness on one side of the body
Aphasia	Inability to formulate or comprehend language
Dysarthria	Difficulty speaking due to weakness of the speech muscles. Note that in contrast to aphasia, language comprehension is intact
Facial weakness	This affects only one side, leading to a drooping corner of the mouth with flattened wrinkles and an assymmetric face.
Ataxia	Muscle control problems leading to clumsy, disorganized movements
Vertigo	Dizziness

(a) Traditional stroke symptoms and their colloquial interpretation [15].

Table 2.1

(b) Illustration of facial symptoms in a stroke patient. Compared to the left side of the face, the facial weakness is noted on the right side, with a drooping mouth corner and assymmetric expression. Image available under Creative Commons license from [1].

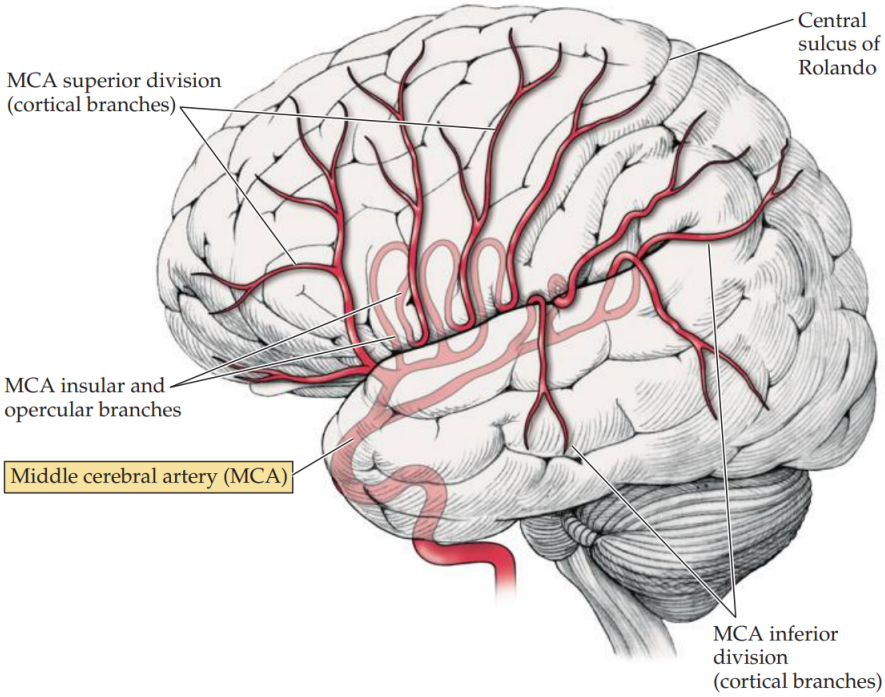


Figure 2.1: Schematic drawing which shows the course of the middle cerebral artery through the brain. Image from [4].

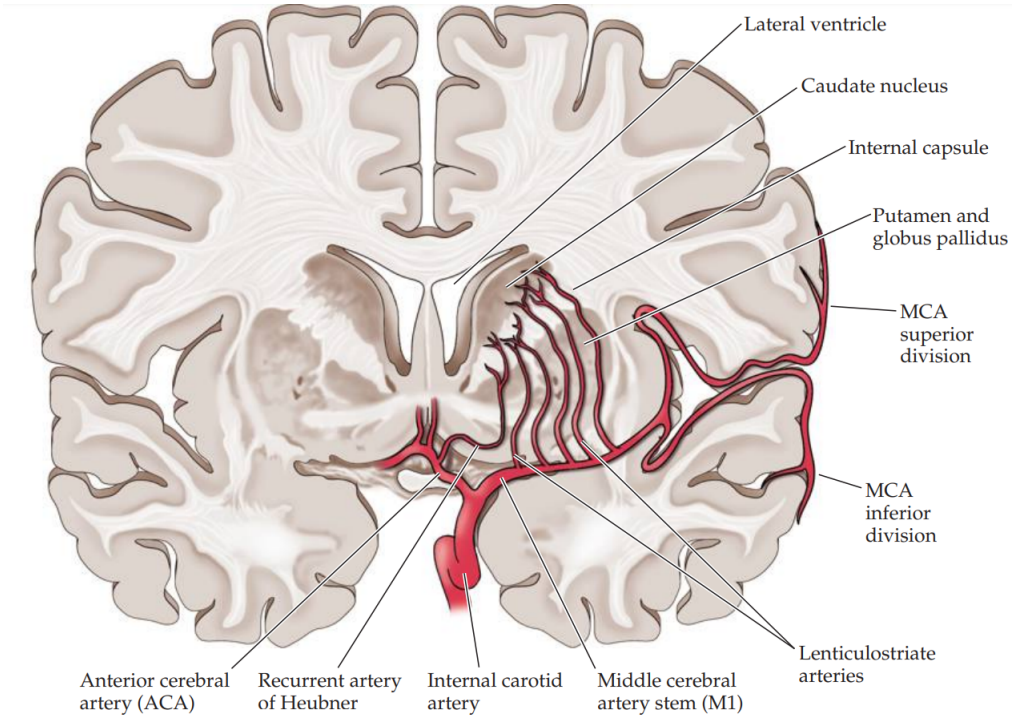


Figure 2.2: Schematic drawing of a coronal section of the brain at the level of the middle cerebral artery. From the stem (M1) segment it courses through the brain, giving off branches known as the lenticulostriate arteries, which provide blood to the basal ganglia. The M2 segment starts at the bifurcation, where a superior and an inferior division are formed. Image from [4].

Tissue or Substance	HU
Air	-1000
Lung	-600 to -400
Fat	-100 to -60
Water	0
Muscle	10 to 40
Blood	30 to 80
Soft tissue, contrast	100 to 300
Bone	400 to 3000

Table 2.2: Approximate Hounsfield Unit (HU) values for different tissues and substances on CT. From [6] and [9].

To enhance the visibility of blood vessels on CT, radio-opaque (X-ray absorbing) contrast can be injected, which will appear dense (white) on the scan. When contrast is present in the blood-vessels the scan is known as a CT-angiography (CTA) scan. A scan where no intravascular contrast is administered will be referred to as a non-contrast CT (NCCT).

Figure 2.3 shows a single slice from a NCCT scan. This case shows the devastating results of stroke when treatment is unsuccessful. Due to an occlusion of the left MCA (on the right side of the image, one should, by international convention, view medical imaging as if looking up at the patient from their feet), the entire vascular territory of the MCA has been damaged. This can be seen on the scan as the hypodense area on the right side, which indicates cytotoxic edema (swelling due to damage at the cellular level) and cell death. For the patient, this results in hemiparesis of the right side of the body.

2.3. Interventional Neuroradiology & Thrombectomy

The optimal treatment of stroke is an area of ongoing research. There are two main treatment approaches, endovascular therapy (commonly referred to as IAT, intra-arterial therapy) or intravenous administration of a potent thrombolytic agent. Intravenous thrombolysis constitutes administering med-

Figure 2.3: This non-constrast CT scan shows the outcome of unsuccessful stroke treatment. The patient presented with a right-sided hemiparesis (weakness of one side of the body). This scan shows the situation 24 hours after the initial symptoms. The vascular territory of the left MCA (right side of the picture) is irreparably lost. This can be seen on the CT as a hypodense (darker) signal compared to the other side. Case courtesy of dr. Ian Bickle, found on Radiopaedia.org, stroke progression on CT.

ication to dissolve the blood clot. This therapy does not work well for larger vessels, which reasonably contain larger clots.

To perform IAT, the femoral artery (the main artery in the groin) is punctured and a catheter system is introduced. This catheter is advanced through the arterial system under image guidance from the groin to the site of occlusion, usually the carotid artery or the MCA. Now the thrombus (blood clot) can be removed either mechanically using a stent retriever (mechanical thrombectomy), through suction (thrombosuction), or by administering a potent thrombolytic agent through the catheter (local thrombolysis). Throughout this text we are specifically interested in mechanical thrombectomy.

Mechanical thrombectomy proceeds by deploying a retrievable stent inside the thrombus. The stent embeds itself in the blood clot. After firm embedding, the stent is pulled back, hopefully with some of the clot attached to it. Often, multiple passes are required to restore blood flow. The procedure is not without risk of complications, such as perforation of the vessel or dissection (damage to and loosening of the layers) of the vessel wall.

Figure 2.6 depicts a thrombectomy procedure in a 63 year old male. The patient presented to the hospital with complaints of left sided weakness. Initial non-contrast CT of the head showed no ischemic changes, even though subsequent CTA imaging revealed a complete occlusion of the right internal carotid artery. Mechanical thrombectomy therapy was performed, leading to a successful clot removal. The patient left the hospital 12 days after the presentation with only minimal sequelae.

Panel A shows a noncontrast computed tomographic (CT) scan of the head (transverse section) revealing slight hypodensity in the left insular cortex (arrow). Panel B shows a CT angiogram (transverse section) revealing an occlusion of the first segment of the left middle cerebral artery (arrow). Panel C shows a cerebral arteriogram (anterior projection) revealing an occlusion of the first segment of the middle cerebral artery before mechanical thrombectomy (arrow). Panel D shows a cerebral arteriogram (anterior projection) revealing recanalization of the left middle cerebral artery after thrombectomy (arrows).

2.4. The MR CLEAN trial

MR CLEAN (Multicenter Randomized Clinical trial of Endovascular treatment for Acute ischemic stroke in the Netherlands) was a multicenter randomized clinical trial performed in the Netherlands. In the 16

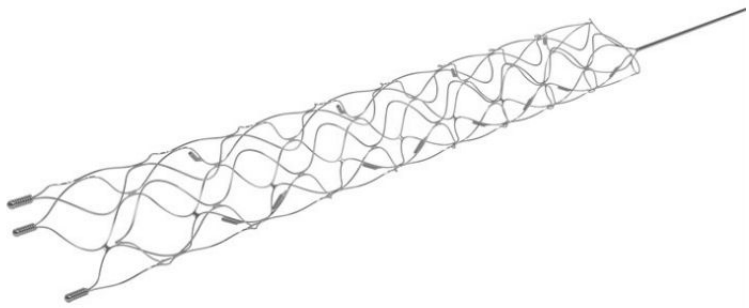


Figure 2.4: An example of a stent retriever device, consisting of a metal cage attached to a wire. Initially, the metal cage is folded. The device is threaded into the thrombus through the occluded artery, after which it is expanded using a balloon. After this expansion the device embeds itself in the blood clot. The stent retriever is subsequently pulled back, hopefully removing the clot. Image available under Creative Commons license from [20].

Clinical & Radiological features
Age
Baseline NIHSS
Pre-Stroke mRS
Diabetes Mellitus
Baseline Systolic Blood Pressure
Baseline Glucose
Intravenous Alteplase
ASPECT score†
Location of occlusion†
CTA collateral score†
Time from onset to groin puncture

Table 2.3: List of all clinical and radiological features used in the dataset. Note that these features coincide with the features validated for the MR PREDICTS clinical decision making tool [21].† Radiological features, which means that they can only be determined after imaging has been acquired.

participating centers, stroke patients were randomized to either receive usual care or usual care and intra-arterial therapy. Patients were only included if treatment was feasible within 6 hours of onset, with a confirmed proximal occlusion in the anterior cerebral circulation. The primary outcome measure was modified Rankin scale at 90 days. Clinical features were registered for each patient, the most relevant features are listed in 2.3. In total, 500 patients were included. Research based on this trial concluded that intra-arterial treatment improves outcome if administered within 6 hours after stroke onset [3].

As a follow up to the trial, the MR CLEAN registry was created. This is an ongoing, prospective, observational cohort study. This registry demonstrated a further significant improvement of functional outcome compared to the outcomes, both in the control as well as the intervention arms, of the MR CLEAN trial [11]. These differences are most likely due to improvements in clinical pipelines, leading to shorter time from onset of symptoms to reperfusion.

2.5. Functional Outcome Prediction

The modified Rankin Scale (mRS) was designed to quantify the outcome of a stroke. It is a seven point scale of increasing disability. In literature, it is often measured 90 days after the event. This outcome is also recorded in the MR CLEAN trial and registries so we can compare outcomes after therapy. Being able to predict the functional outcome using mRS is helpful in selecting those patients who will benefit from the therapy. Models have been developed to predict functional outcome using clinical parameters, such as the MR PREDICTS decision tool [21]. The models that have been described use conventional statistical methods. Functional outcome prediction performance might be improved, potentially directly

Figure 2.5: Example setup of an interventional suite. The patient is on the table, covered in sterile drapes. An X-ray machine (sometimes referred to as C-arm due to its characteristic shape) allows for visualization of the internal anatomy of the patient during the procedure. The radiological team is on the left, working with catheters, guidewires and other interventional tools to treat the condition of the patient under image guidance. Image from [19], available under Creative Commons license.

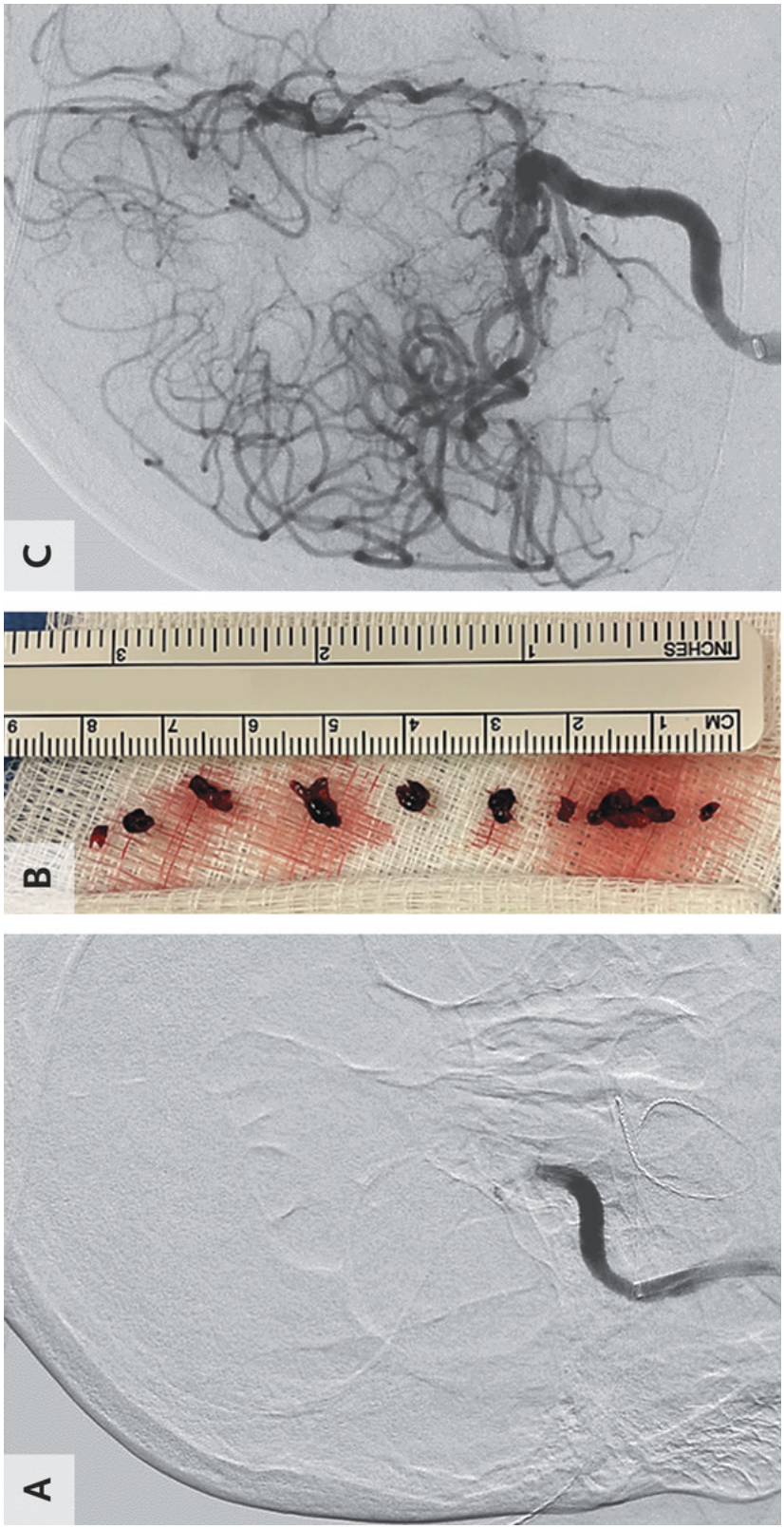


Figure 2.6: Images from the thrombectomy procedure. Panel A shows DSA images of contrast being injected in the right carotid artery, showing a sudden, suggesting an occlusion of the vessel. Panel B shows thrombotic material retrieved from the occlusion site through aspiration using a suction device. Panel C shows DSA after contrast injection at the same location, showing successful restoration of perfusion after removal of the clot. Reproduced with permission from [14], Copyright Massachusetts Medical Society.

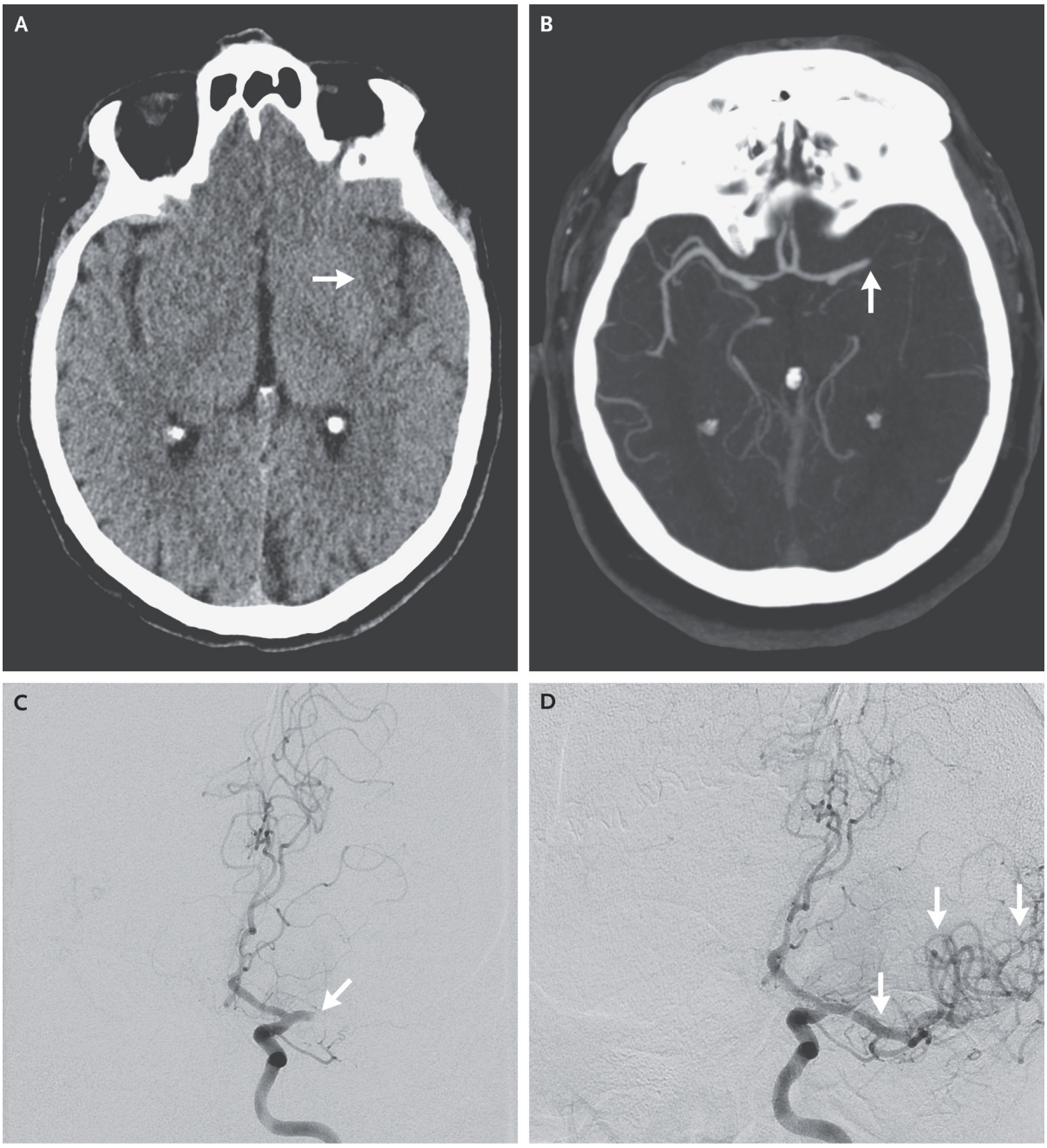


Figure 2.7: Imaging of acute ischemic stroke. Panel A shows NCCT of the head, where the white arrow points to subtle hypodense attenuation of the left insular cortex, indicating potential ischemic changes. The easiest way to see this is to compare to the other side, where the same structure is more clearly visible. Admittedly, this is a subtle finding. Panel B shows CTA imaging in the same patient. The middle cerebral artery is occluded at the location of the arrow. Panels C and D show DSA imaging of the arterial tree before and after revascularization, respectively. Reproduced with permission from [18], Copyright Massachusetts Medical Society.

leading to clinical benefit, using more complex, machine learning based models.

3

Machine Learning

The field of Machine Learning (ML) studies algorithms capable of ”learning” to detect patterns in data. ML is a subfield of the more general scientific field of Artificial Intelligence (AI). There are many different paradigms within the ML field, such as semi-supervised or unsupervised learning; in this work we are concerned with supervised learning. Supervised learning refers is a type of ML where the correct output labels are available to the algorithm during training, such that the algorithm can correct itself immediately after an incorrect prediction. Supervised ML algorithms are ”trained” using a training dataset. This training process involves improving the performance of the algorithm by presenting it with progressively more data, the algorithm is ”learning”. ”Training” and ”learning” are anthropomorphic terms used to describe an underlying optimization process, in which the performance of the algorithm is improved by minimizing a loss function using mathematically derived steps. After the training phase, we assess the performance by presenting the algorithm with new, unseen data in the form of a so-called test set. This is done because checking the efficacy of the method on the dataset it was trained on will overestimate its true performance, as models tend to overfit to the dataset they are trained on. Great care must be taken to avoid leakage of the training data into the validation set.

Deep learning is a sub-field within machine learning, which involves the use of artificial neural networks. Artificial neural networks are inspired by biological information processing systems, even though they significantly differ from real brains in multiple important ways. The term ”deep” refers to the fact that any arbitrary mathematical function can be approximated using a neural network with at least one hidden layer (of arbitrary width) of non-polynomial activation functions [10], whereas the simple linear perceptron model, which is the historical precursor of artificial neural networks, cannot. The network must therefore be ”deep”, either in the sense of the layer width or in the amount of layers, and not ”shallow”.

3.1. Medical Machine Learning

Medical ML can be defined as the application of ML methods to medical problems. There are certain difficulties which are more common in medical ML as compared to other application areas of machine learning. One such issue is the reduced availability of data. There are multiple factors contributing to the sparsity of data in the medical sectors, such as stringent privacy laws that preclude data-sharing, and the fact that obtaining new medical data is relatively expensive. One of the main predictors of ML algorithm success is the amount of data available. Care must therefore be taken, for instance by using appropriate data augmentation techniques, to ensure that enough data is available to the learning system. Luckily, there are multiple ongoing initiatives to share medical data between hospitals.

Another issue which is perhaps more relevant in medical ML as opposed to other application domains is the interpretability or explainability of the results. In the medical domain, this is very relevant as clinicians should be able to justify and explain their reasoning about certain treatment decisions with patients. Many ML methods are problematic in this regard, as they do not have a human readable way to show how the results are derived. In artificial neural networks, for instance, the representations of the data inside the network do not have a clear interpretation. Instead, artificial neural networks must be considered to be black box models, where only the input and output are amenable to interpretation by human observers. In recent years, the concepts of shared decision making and informed consent

have become increasingly relevant in the medical field, as such, development of explainable ML is even more pressing.

3.2. The Multilayer Perceptron

The simplest type of deep neural network is the Feedforward Neural Network (FNN), sometimes also referred to as Multilayer Perceptron (MLP). The network consists of an input layer, followed by one or more hidden layers and finally an output layer. In between the layers, information flows. Each layer except for the input layer consists of nodes, where information from the previous layers is aggregated and passed through a nonlinear activation function. The input layer is special in that it does not contain nonlinear activation functions. The network receives an input vector x, which is propagated through the nodes in the next layers. Each non-input node receives a weighted sum of the outputs of all the nodes in the preceding layer. A simple example of an MLP is provided in Figure 3.2.

The output of a layer in the network can be described as

$$y = f(x; \theta, w, b) = \Phi(x; \theta)^T w + b$$

which shows that the output of a layer, y, can be seen as a nonlinear function f taking a vector x as input. The function f is parameterized by the parameters θ of the nonlinear activation function Φ, as well as by weight vector w and bias term b (adapted from [7]). In this way, the MLP as a whole can be seen as a composition of arbitrary nonlinear functions.

3.2.1. Activation Functions

The nonlinear functions which process the inputs of each node are called activation functions. There are many different types of activation functions, their most important property is that they are nonlinear, such that the network can learn to combine them to approximate arbitrary functions, as opposed to a linear combination of linear functions, which would be unable to learn arbitrary nonlinear functions. Traditionally, the sigmoid function, defined as

$$f(x) = \frac{1}{1 + e^{-x}}$$

was used as an activation function, though it is less popular nowadays. Currently, the most popular activation function is the ReLU (Rectified Linear Unit) function. This is a piecewise linear function defined as

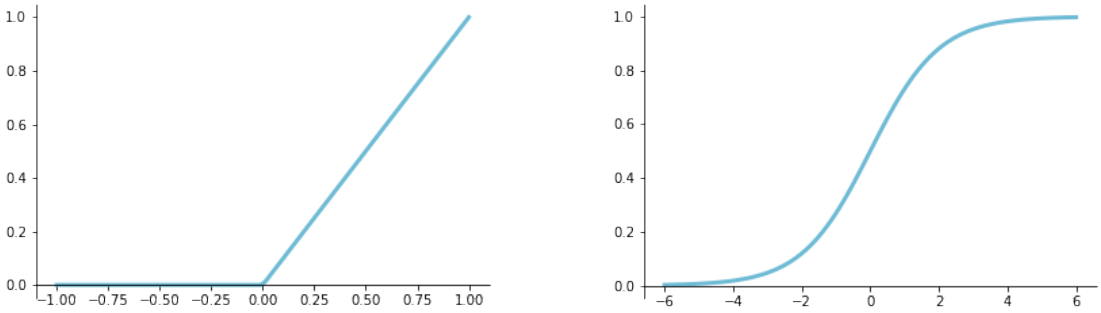
$$f(x) = \max(0, x)$$

such that the neuron is only activated for inputs with a positive sum. The main benefit of such a simple activation function is the fact that there are fewer problems with vanishing gradients since the function does not saturate. The vanishing gradient problem is a well known issue where gradients which are being backpropagated through the neural network become increasingly smaller, until the value is zero. This phenomenon obstructs the learning process, as nonzero gradients are required to update the weights during the learning steps. ReLU negates this problem by being linear, preventing decay of gradient values. Another advantage of ReLU is that it is simple to compute, and that it typically leads to sparse activations since negative inputs will result in no information being propagated to the next layer. A potential issue is the fact that this function is non-differentiable at 0. This problem can be fixed using smooth approximations of the ReLU which are differentiable everywhere, such as GELU (Gaussian Error Linear Unit). The aforementioned activation functions are displayed in Figure 3.1. A full treatment of the different activation functions and their properties is outside the scope of this work.

3.3. Stochastic Gradient Descent

In the context of ML, learning is defined as a process in which performance is improved over time, by allowing the model to adjust its weights based on examples. Now that we have a basic understanding of the structure of artificial neural networks we are equipped to reconsider learning in a more formal way. Let $\theta \in \mathbb{R}^d$ define the parameters of a neural network. We define the training dataset as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, containing N samples indexed by i, where x represents an input vector and y represents the desired model output. We define the loss function as

$$\mathcal{L}(\theta) : \mathbb{R}^d \to \mathbb{R}$$



(a) ReLU

(b) Sigmoid

Figure 3.1: ReLU and Sigmoidal activation functions.

We do not impose any restrictions on the loss function other than that it must be differentiable. As the training dataset does not change, the loss depends only on the parameters of the network. ”Learning” proceeds in two steps. Firstly, we iterate over the samples in $\mathcal{D}$ to calculate the loss as a function of the current parameterization θ. Secondly, we update θ using the update rule

$$\theta_m \leftarrow \theta_{m-1} - \eta \nabla \mathcal{L}(\theta_{m-1})$$

Here, m indicates the current iteration of the learning process, and $\eta \in (0, 1)$ is an update factor called the learning rate. Updating θ involves calculating the gradient of each weight with respect to the loss. This gradient calculation is implemented in an algorithm called backpropagation. The name refers to the fact that weights in layers closer to the input need gradient information from the deeper layers to be calculated. Partial gradient information must therefore be propagated backwards, from the deep to the shallow layers.

A single iteration through the entire training dataset is called an epoch. Often, it is computationally expensive to iterate through the entire dataset in order to calculate the gradient, and so, subsets of the dataset are used to calculate a partial gradient. These subsets are called batches, and the procedure is known as stochastic gradient descent (SGD). Stochasticity follows from the fact that we do not calculate the gradient using the entire dataset, instead relying on a small batch. As a result, SGD is faster to compute, at the expense of less precise gradient information. SGD is a simple optimization algorithm used in neural network training, in practice, more sophisticated optimizers, such as the Adam algorithm [13], are often used. These methods often incorporate additional information, using for instance an exponentially weighted moving average of the previous batch gradients, to stabilize the optimization trajectory.

3.4. Binary Cross Entropy

We have provided a broad definition of loss functions, restricting them only to be differentiable. Many loss functions exist, and different loss functions are more suited to different machine learning tasks. The problem studied in this thesis is dichotomized functional outcome prediction, which is an example of a binary classification problem in which $y \in \{0, 1\}$. For this problem, binary cross entropy (BCE) is often introduced as a loss function. Define $\hat{y}$ to be the output of the model, which is a probability. BCE is then defined as

$$L(\theta) = -\frac{1}{N} \sum_{i=1}^{N} (y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i))$$

Here, θ is the parametrization of the ML algorithm, y_i is the true class of the i-th sample, and $\hat{y}_i$ is the predicted probability of this sample. To obtain the total loss, we average over the losses of the individual samples. BCE loss is useful in binary classification because it especially penalizes those classifications which are confident and wrong.

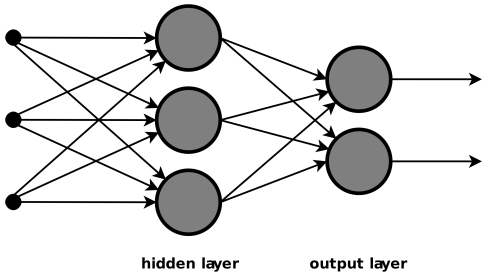


Figure 3.2: Schematic illustration of the multilayer perceptron. On the left side, an input is provided. The three nodes in the hidden layer each combine the values of the input using their own weights. The values produced by the hidden nodes are again combined in different ways by the two nodes in the output layer, leading to two output values. Image available under Creative Commons license from [16]

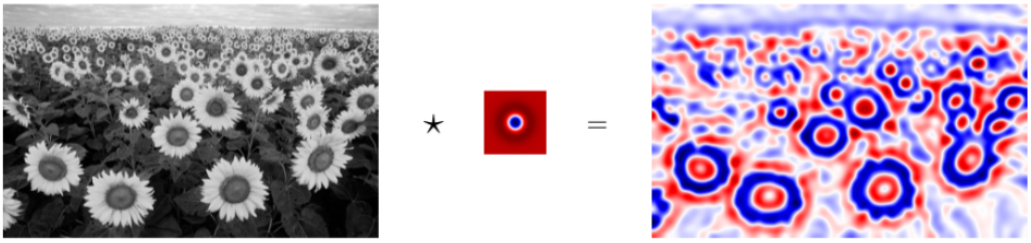


Figure 3.3: The image on the left is convolved with a circular kernel, leading to the activation map on the right hand side. Red indicates positive value, blue negative. The shape of the kernel has an effect on the type of features being detected. In this case, the round shape of the kernel leads to detection of the round shape of the sunflowers. Image adapted from the TU Delft Deep Learning course (CS4180).

3.5. CNN & Med3D

While the simple MLP described in the previous sections is an appropriate architecture for learning from independent data, it is less suitable when data is involved which has some kind of inherent, combined structure, such as images. There are multiple reasons for this, the most important being simply the number of parameters needed for a fully connected layer. Already, a relatively small image of 400×400 pixels would require 160000 parameters to represent the connections to a single neuron in the first hidden layer alone. A potential solution for this problem is a convolutional neural network (CNN), which reduces the required number of parameters by performing a discrete convolution operation.

3.5.1. Convolution

The one dimensional discrete convolution operation can be defined as follows

$$z[n] = (f * g)[n] = \sum_{m=-\infty}^{\infty} f[m]g[n-m]$$

extending this operation to more dimensions is straightforward. Essentially, this operation represents the dot product of the filter with a specific element in the signal, as well as its neighboring elements in the spatial sense. The convolutional operation is translation equivariant, meaning that the result changes predictably with translations of the input. The convolution operation is not rotation invariant, meaning that rotations of the input (in 2D or higher dimensions) do affect the output. An illustration of the effects of the discrete convolution operation is provided in Figure 3.3.

If we define an image (referring to a CT scan) $g \in \mathbb{R}^{h \times w \times d}$ and a filter $f \in \mathbb{R}^{i \times j \times k}$ then performing a convolution will result in an output image $z \in \mathbb{R}^{h' \times w' \times d'}$ that represents the activation map of the filter. Each convolutional layer in a CNN consists of multiple of these filters, leading to multiple activation maps. The filters are trainable, meaning that the parameters of the filter can be adjusted, such that the network ”learns what to look at” during the training process. Each element in the activation map is based on an area the size of the convolutional kernel from the preceding layer. This means that the later layers in the network represent larger receptive fields, and it has been demonstrated that they

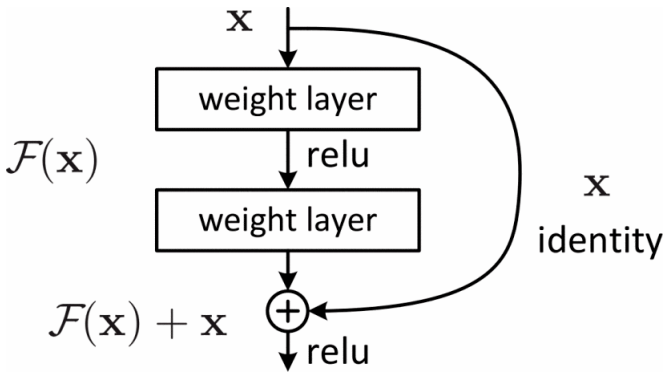


Figure 3.4: Schematic overview of a residual block. Image from [8]

come to represent more abstract features during the training process. In other words, the model learns to extract and process increasingly abstract features from the input as information flows from input to output.

Another important layer in the CNN is the pooling layer, which can perform subsampling by combining elements of the activation map. Multiple different types of pooling layers exist, such as the max pooling layer, which simply subsamples an area in the layer input by propagating only the maximum value of this area, for each area of a certain size in the layer input, and discarding the rest. Another important pooling method is average pooling, where instead of the maximum, the average is taken. Pooling layers can increase processing speed by discarding less relevant information.

There are multiple specific parameter settings for the convolutional layers, such as the stride, which refers to the step size between applications of the filter in the input image, or the padding, which refers to the method of dealing with missing values in the input image by adding additional values. Different CNN architectures use different settings with regard to these parameters.

In the classification setting the final activation map, which results from applying convolutional and downsampling layers in alternating fashion, usually feeds into one or more fully connected layers. The output of these fully connected layers is often passed through a softmax operation, which leads to a class probability as model output.

3.5.2. ResNet

A well known issue in deep networks is the vanishing gradient problem. Due to the depth of large feed-forward networks, the gradient becomes negligible at some point point in the backpropagation process. This means that certain weights in the network are not being updated. One potential solution is to introduce skip connections. A residual (or skip) block is a part of a neural network where information bypasses at least one layer. Figure 3.4 shows an example of a residual block. Because of this information bypass, referred to as a skip connection, information can flow freely through the skip block in addition to being processed. Mathematically, the residual connection can be expressed as:

$$\mathcal{F}(x) + x$$

where x is the layer input and $\mathcal{F}(\cdot)$ is the transformation applied by the layer. Residual networks are capable of achieving state-of-the-art performance in image recognition tasks because of the use of these residual connections.

The most well-known neural network architecture that employs residual blocks is the ResNet architecture [8]. There are many different variations of the ResNet architecture. The ResNet architecture follows the design philosophy of the famous VGG architecture, but it adds residual connections. Med3D is a three-dimensional version of the ResNet architecture which has been pretrained on a combined dataset of medical challenges called 3DSeg-8. In the Med3D architecture, the output is a segmentation, but one can easily move from performing segmentation to classification by appending a global average pooling layer followed by a softmax layer. Another relevant, ResNet based architecture is the VoxResNet, which is specifically designed to handle brain imaging data.

3.6. Attention mechanisms

Attention mechanisms were originally developed in the field of natural language processing (NLP). In NLP, one often works with sequence to sequence models, which are supposed to translate one sentence to another. An issue that occurred here was that for long sentences, state information would get lost in the Recurrent Neural Networks (RNNs) that were used to translate these sentences. One proposed solution to this problem was to use a weighting mechanism to combine state information based on how relevant words are to each other. This is the attention mechanism.

Attention is essentially a learnable soft weighted information retrieval mechanism, which can be used to detect salient features in some input. In one of the earlier definitions by Bahdanau et al. [2], the mechanism was used to weight encoder outputs using weights derived from the decoder inputs. In self-attention, which was developed at a later point in time, the attention mechanism uses inputs from the same layer as the weights.

To retrieve information, we need three tokens, the query, key and value token. The query represents ”what” we are looking at in the output. The key represents the relative importance of the other words in the input for that particular output. The value represents the relevance of the output words compared to the particular output. The self attention mechanism works based on these equations:

$$\begin{aligned}
\mathbf{q}_i &= \mathbf{W}_q \mathbf{x}_i + b_q \\
\mathbf{k}_i &= \mathbf{W}_k \mathbf{x}_i + b_k \\
\mathbf{v}_i &= \mathbf{W}_v \mathbf{x}_i + b_v \\
\mathbf{y}_i &= \sum_j \text{softmax}(\mathbf{q}_i^T \mathbf{k}_j) \cdot \mathbf{v}_j
\end{aligned}$$

In these equations, $\mathbf{x}_i$ is the embedded input. To obtain the query $\mathbf{q}_i$, the key $\mathbf{k}_i$ and the value $\mathbf{v}_i$ vectors we perform linear projections of the embedded input vectors with the relevant matrix and bias vector. For the query we project with $\mathbf{W}_q$ and add a bias term b_q, and we do the same with the key and value. To obtain the output $\mathbf{y}_i$ we sum over the weighted similarities between keys and the query, using the value vector to weight the similarity. In our equations, we use the softmax operation to normalize the similarity between key and query, but other similarity operations can be used. The softmax operation can be defined as $\text{softmax}(z_i) = \frac{\exp z_i}{\sum_{j=1}^{K} \exp z_j}$. The time complexity of the aforementioned operations is $\mathcal{O}(N^2 d)$, where N is the size of the database and d is the dimensionality of the feature space.

The attention mechanism provides an alternative to convolutional methods for processing information. The mechanism is inspired by the human visual system, where the field of view is only focused on certain aspects of an input at a time. As an example, when reading a book the brain is actually only looking at a few words at a time, it is not possible to process the entire page at once.

Multi-Head Attention is an extension of the aforementioned attention mechanism where multiple query, key and value triples are used, such that elements can attend to different concepts using the multiple different heads. The outputs of the different heads are subsequently concatenated.

3.6.1. Transformer models

The focus of current deep learning literature is shifting from the CNN to attention based models, such as the Transformer [5]. The Transformer model is a sequence-to-sequence model that was originally derived in the context of natural language processing (NLP), but it has been successfully modified so that it can be applied in other fields as well. Transformers are based solely on self-attention mechanisms, the convolutional mechanism is no longer used. Because transformers do not have a recurrent structure, they can be readily parallelized. The basic building blocks of the Transformer model consist of encoder and decoder blocks. The input sequence is first tokenized. After this, positional information is added using a positional encoding. Then, the input is passed through one or more encoder blocks in sequence. Each encoder block contains a Multi-Head Attention (MHA) mechanism, followed by a feed forward layer. Residual connections across the MHA and feed forward layers facilitate unconstrained propagation of information, by adding the input of each layer to the output and normalizing. After the encoder blocks, a decoding structure is also present. The decoder block takes as input both

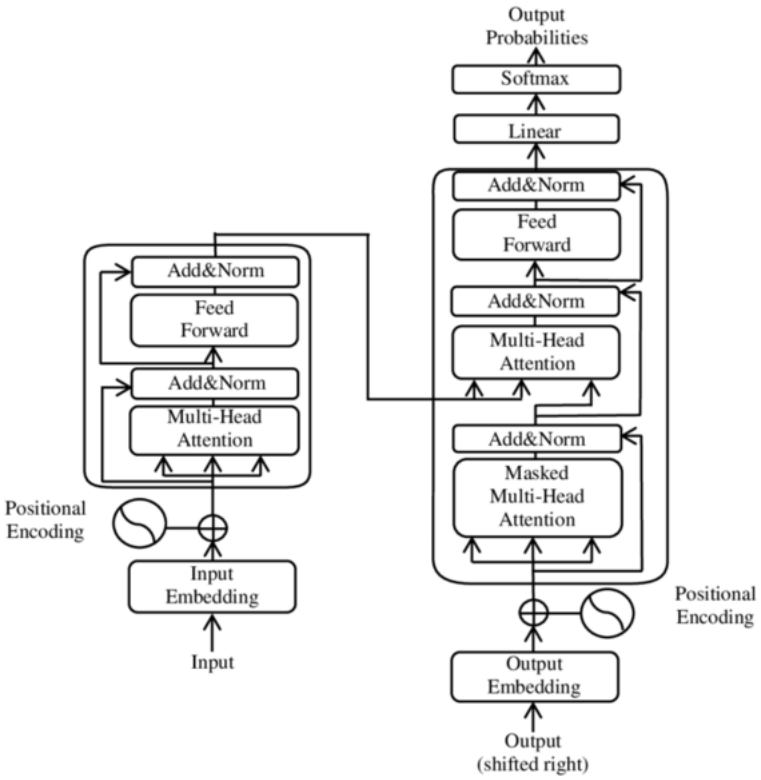


Figure 3.5: Example of the Transformer architecture. Input is embedded and a positional encoding is added. The input is then passed through a Multi-Head Attention (MHA) mechanism, followed by a feed forward connection. This structure encompasses the encoder. The decoder (on the right) has a similar structure, with the addition of a masked MHA mechanism to receive the masked output. After the decoded block, the outputs are passed through a linear layer followed by a softmax operation. Note that in all operations apart from the final output, residual connections are provided, allowing information to skip each operation. Image available under Creative Commons license from [12].

the output from the encoder, as well as a shifted version of the embedded output. The desired output is embedded using positional encoding, and fed to a masked version of the MHA module to prevent data peeking. The output of this first MHA module is combined with the output from the encoder module in another MHA module, after which it is passed through a feed-forward network. The output of the decoder block is passed through a linear layer followed by a softmax operation to predict classes. A schematic overview of the general structure of the Transformer model is provided in Figure 3.5.

Bibliography

[1] Another-anon-artist-234. *Stroke facial droop*. [Online, accessed 20 Aug 2022]. URL: https://commons.wikimedia.org/wiki/File:Stroke-facial-droop.jpg.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2014. DOI: 10.48550/ARXIV.1409.0473. URL: https://arxiv.org/abs/1409.0473.

[3] Olvert A. Berkhemer et al. "A Randomized Trial of Intraarterial Treatment for Acute Ischemic Stroke". In: *New England Journal of Medicine* 372.1 (2015). PMID: 25517348, pp. 11–20. DOI: 10.1056/NEJMoa1411587. eprint: https://doi.org/10.1056/NEJMoa1411587. URL: https://doi.org/10.1056/NEJMoa1411587.

[4] H. Blumenfeld. *Neuroanatomy through clinical cases*. Third. Sunderland, Mass.: Sinauer Associates., 2021.

[5] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2020. DOI: 10.48550/ARXIV.2010.11929. URL: https://arxiv.org/abs/2010.11929.

[6] Omnia Elsayed et al. "Automatic detection of the pulmonary nodules from CT images". In: Nov. 2015, pp. 742–746. DOI: 10.1109/IntelliSys.2015.7361223.

[7] Jan van Gemert. "Deep Learning course, lecture 1". [Online, accessed 14 Sept 2022]. 2021.

[8] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. DOI: 10.48550/ARXIV.1512.03385. URL: https://arxiv.org/abs/1512.03385.

[9] A. O. Hebb and A. V. Poliakov. "Imaging of deep brain stimulation leads using extended Hounsfield unit CT". In: *Stereotact Funct Neurosurg* 87.3 (2009), pp. 155–160.

[10] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5 (1989), pp. 359–366. ISSN: 0893-6080. DOI: https://doi.org/10.1016/0893-6080(89)90020-8. URL: https://www.sciencedirect.com/science/article/pii/0893608089900208.

[11] Ivo G H Jansen, Maxim J H L Mulder, and Robert-Jan B Goldhoorn. "Endovascular treatment for acute ischaemic stroke in routine clinical practice: prospective, observational cohort study (MR CLEAN Registry)". In: *BMJ* 360 (2018). Ed. by. ISSN: 0959-8138. DOI: 10.1136/bmj.k949. eprint: https://www.bmj.com/content/360/bmj.k949.full.pdf. URL: https://www.bmj.com/content/360/bmj.k949.

[12] Yuening Jia. "Attention Mechanism in Machine Translation". In: *Journal of Physics: Conference Series* 1314.1 (Oct. 2019), p. 012186. DOI: 10.1088/1742-6596/1314/1/012186. URL: https://doi.org/10.1088/1742-6596/1314/1/012186.

[13] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2014. DOI: 10.48550/ARXIV.1412.6980. URL: https://arxiv.org/abs/1412.6980.

[14] David C. Lauzier and Akash P. Kansagra. "Thrombectomy in Acute Ischemic Stroke". In: *New England Journal of Medicine* 386.14 (2022), pp. 1351–1351. DOI: 10.1056/NEJMicm2116727. eprint: https://doi.org/10.1056/NEJMicm2116727. URL: https://doi.org/10.1056/NEJMicm2116727.

[15] Lynda D. Lisabeth et al. "Acute Stroke Symptoms". In: *Stroke* 40.6 (2009), pp. 2031–2036. DOI: 10.1161/STROKEAHA.109.546812. eprint: https://www.ahajournals.org/doi/pdf/10.1161/STROKEAHA.109.546812. URL: https://www.ahajournals.org/doi/abs/10.1161/STROKEAHA.109.546812.

[16] *Multilayer Perceptron*. [Online, accessed 22 Jun 2022]. URL: https://commons.wikimedia.org/wiki/File:Multi-Layer_Neural_Network-Vector.svg.

[17] Y. S. Ng et al. “Comparison of clinical characteristics and functional outcomes of ischemic stroke in different vascular territories”. In: *Stroke* 38.8 (Aug. 2007), pp. 2309–2314.

[18] William J. Powers. “Acute Ischemic Stroke”. In: *New England Journal of Medicine* 383.3 (2020), pp. 252–260. DOI: 10.1056/NEJMcp1917030. eprint: https://doi.org/10.1056/NEJMcp1917030. URL: https://doi.org/10.1056/NEJMcp1917030.

[19] Shakeel Moideen. *Airway Management Outside the Operating Room - Scientific Figure*. [Online, accessed 14 Jun 2022]. URL: https://www.researchgate.net/figure/A-dimly-lit-crowded-interventional-radiology-room_fig1_343855237.

[20] *Stent retriever*. [Online, accessed 16 Jun 2022]. URL: https://operativeneurosurgery.com/doku.php?id=stent_retriever&rev=1647281872.

[21] Esmee Venema et al. “Prediction of Outcome and Endovascular Treatment Benefit: Validation and Update of the MR PREDICTS Decision Tool”. In: *Stroke* 52.9 (2021), pp. 2764–2772. DOI: 10.1161/STROKEAHA.120.032935. eprint: https://www.ahajournals.org/doi/pdf/10.1161/STROKEAHA.120.032935. URL: https://www.ahajournals.org/doi/abs/10.1161/STROKEAHA.120.032935.

[22] R.H. Wimmers Y. Koop. *Hart- en vaatziekten in Nederland, 2021*. [Online, accessed 15 Jun 2022]. URL: https://www.hartstichting.nl/getmedia/06fb9c92-a1f7-4135-a635-ff73680bfaa6/cijferboek-hartstichting-hart-vaatziekten-nederland-2016.pdf.