Automated Functional Outcome Prediction in Stroke using Combined Imaging and Clinical Parameters

Samantha de Graaf[1]

[1]Technical University of Delft

Abstract

Predicting functional outcome after intra-arterial treatment (IAT) in acute ischemic stroke (AIS) patients is an important aspect of treatment decision making and prognostics. Standard methods for functional outcome prediction after stroke combine baseline clinical (and radiological) parameters.

In this study, we investigated to what extent baseline CTA images can be used for the prediction of functional outcome and how this relates to standard scoring methods. Furthermore, it was investigated whether combining baseline CTA images with clinical parameters improved the predictive accuracy compared to outcome prediction based on clinical parameters.

We proposed two network architectures, a convolutional neural network (CNN) for the processing of image data and a multilayer perceptron for the processing of clinical (and radiological) parameters. Various training strategies were applied for the fusion of image and clinical data.

The CNN processing CTA images achieved an average cross-validated area under the curve (AUC) score of 0.67, which was lower than for models processing clinical (and radiological) parameters. The best performing model combining CTA images and clinical parameters was trained end-to-end and applied weight initialization of the pre-trained CNN (AUC = 0.78). The DeLong test showed that the combined model performed significantly better than the model processing clinical parameters (AUC = 0.75). However, the difference is small and might not be clinically relevant. Compared to scoring methods processing clinical and radiological parameters the combined model achieved similar performance.

Index Terms

Predictive modelling, Stroke, Deep neural networks

I. INTRODUCTION

A. Clinical Background

Stroke is worldwide the second leading cause of death and the third leading cause of disability [1]. In 2019 the annual incidence of stroke was around 2 per 1000 in The Netherlands. Half of patients die or remain severely disabled [2]. A stroke occurs when blood flow is reduced in an artery that supplies blood to the brain, leading to damage or death of brain cells. Brain damage can be permanent when the circulation is not restored quickly, and therefore commencing treatment quickly is crucial for reducing the chances of disability and death [3]. Acute ischemic stroke (AIS) covers 80% of strokes in Western countries [4]. This type of stroke is caused by a blood clot blocking an artery within the brain. There are two treatment options for AIS patients. The first option is intravenous thrombolysis with tissue plasminogen activator (IV-tPA), which attempts to dissolve the blood clot by intravenous injection of medication. Up until 2015 this was the only therapy with proven efficiency in patients with AIS, provided that the treatment commenced within 4.5 hours after stroke onset [5]. However IV-tPA injection is less effective for the dissolution of blood clots in the proximal arteries than for blood clots in the distal arteries [6]. A second treatment option is available for occlusions in the proximal arteries, namely intra-arterial therapy (IAT). With this therapy a catheter is inserted and guided towards the occluded area under X-ray guidance, subsequently the thrombolytic agents can be delivered locally (thrombolysis) or the clot is removed mechanically (thrombectomy). A randomized trial by Berkhemer et al. [7] concluded that IAT performed within 6 hours after stroke onset was effective and safe for proximal occlusions. A meta-analysis in 2016 concluded that IAT thrombectomy led to an increase in functional outcome in a majority of patients, which included different age groups, varying stroke severity, sex and stroke localization [8]. However, this concerned an average benefit over a patient group. Treatment benefit is likely to vary for individual patients [9]. Patients with no treatment benefit are preferably withhold from IAT, because IAT carries a small risk of bleeding or movement of blood clots to previously unaffected parts of the brain due to insertion of catheters and guidewires. By distinguishing between patients who do and do not benefit from IAT, personalized stroke care may be improved.

In current practice, the physician relies on experience and several patient specific parameters to decide which treatment to perform. Predicting functional outcome after treatment from baseline parameters provides support for treatment decision making and assists physicians with informing patients on their prognosis after treatment. Functional outcome of stroke patients is generally measured 90 days after treatment with the modified Rankin Scale (mRS).

Several clinical and radiological parameters serve as predictors for the mRS. Machine learning methods have been developed to automate functional outcome prediction based on these parameters [10]. A recent study developed a convolutional neural network for functional outcome prediction based on CT images [11]. Their promising result raised the question whether baseline images that are routinely acquired when a patient is admitted in the hospital with stroke symptoms have predicted value and can be used to directly predict functional outcome. The aim of this study can be divided into two parts. First, we investigated to what extent baseline CTA images can be used for the prediction of functional outcome after IAT in stroke patients and how this relates to traditional scoring methods. Secondly, it was investigated whether combining baseline CTA images with clinical parameters improves the predictive accuracy compared to outcome prediction based on clinical parameters. This combined approach was compared to standard prediction methods based on both clinical and radiological parameters to investigate whether the CTA image removes the need of radiological parameters for outcome prediction.

B. Related Work

1) Clinical parameter analysis: The majority of existing work regarding automated functional outcome prediction in AIS patients is based on clinical information only, and investigated machine learning algorithms such as: logistic regression (LR) [12][13][14][15][16][17][18][19][20], random forest (RF) [12][13][17][18][19][20][21][22], support vector machine (SVM) [12][13][18][19][20][21][23], super learner [20] and artificial neural networks (ANN) [17][19][20][21][23]. Machine learning approaches based on baseline parameters performed more accurately than single pretreatment scores or clinical judgement alone [17][18]. Venema et al. [15][16] applied LR for (good) functional outcome prediction with 11 baseline clinical and radiological parameters. Their decision making tool, MR PREDICTS, was the first to be developed from a recent multicenter cohort within the Netherlands. The tool had modest discriminative potential, reaching an externally validated Area Under ROC Curve (AUC) of 0.73 for the prediction of good functional outcome. Studies comparing various machine learning algorithms for this task found insignificant difference in performance [13][17][18][19][20][21]. A study by Van Os et al. [20] performed on 1,383 subjects from the Netherlands compared LR, SVM, RF and Super Learner approaches and found a mean AUC range of 0.77-0.79.

Model performance depended highly on selecting the right parameters, and less on the selected machine learning algorithm. For example, Heo et al. [17] conducted two different experiments on three machine learning algorithms with a large cohort of 2604 subjects, one with 38 parameters and the second with 6 parameters used in a pretreatment scoring system (ASTRAL). ML algorithms did not outperform the ASTRAL scoring system when using the same parameters, however with the 38 parameters all ML algorithms had a higher performance than the ASTRAL score. Studies that included non-baseline parameters in their model predicted more accurately. For example, Lin et al. [21] found a mean AUC range of 0.94-0.95 using 30 day follow up data with no significant performance difference between RF, SVM and ANN. This information is valuable for prognosis, but it is not relevant for treatment decision-making [14][20][21].

TABLE 1: mRS after 90 days prediction methods combining imaging and clinical parameters

Study	AUC image	AUC clinical	AUC fused	Training strategy	Image modality	Clinical parameters
Bacchi et al. [31]	0.54	0.61	0.75	End-to-end	NCCT	Age, gender, time from stroke onset, NIHSS, blood pressure,blood glucose, temperature, past history of hypertension, diabetes mellitus, hypercholesterolaemia, or atrialfibrillation
Samak et al. [33]	0.67	0.70	0.75	End-to-end	NCCT	Clinical parameters used by Van Os et al. [19].
Zihni et al. [32]	0.68	0.75	0.76 0.75	End-to-end Extracted feature	TOF-MRA	Age, sex, initial NIHSS, cardiac history, diabetes, hypercholesterolemia and thrombolysis treatment

AUC = Area Under the Curve; NIHSS = National Institutes of Health Stroke Scale

2) Image data analysis: Recently, CNN's have become popular in stroke segmentation and image classification tasks. Deep learning has successfully been introduced to automate scoring of biomarkers such as ASPECTS and collateral score, reporting high accuracy [24][25][26]. These methods required expert annotation and are subject to inter-observer variability. Only a few studies have taken the prediction of functional outcome into consideration and whether imaging features can serve as clinically relevant prognostic biomarkers. This section elaborates on state-of-the-art convolutional neural networks for dichotomized functional outcome prediction. Hilbert et al. [11] reported that in their specific dataset deep learning methods outperformed models using radiological image biomarkers for acute ischemic stroke outcome prediction in three out of four cross-validation folds (average AUC of 0.71). Their dataset consisted of Maximum Intensity Projections (MIP) from CTA data in the axial plane of 1301 patients from the MR CLEAN Registry part 1 [27]. A Residual Neural Network (ResNet) was adapted with Structured Receptive Fields (RFNN), which redefined convolutional kernels as a fixed set of Gaussian derivative filters.

Nishi et al. [28] designed a multi-output CNN to segment ischemic core lesions and derived high level imaging features for the prediction of functional outcome. An U-Net design [29] was applied for the segmentation task. From the deepest convolutional layer of the encoder path high level image features are extracted, followed by a 2 layer neural network for the prediction of functional outcome. The model was trained with diffusion weighted MRI images of 146 patients. Their method showed moderate performance (AUC = 0.73) and this was significantly higher than predictions based on a standard neuroimaging biomarker (AUC = 0.64).

3) Imaging and clinical parameters combined: Previous work in the medical domain proved promising results for combined clinical parameters and image modalities in outcome prediction. For example, performance improved for the assessment of tumor response in breast cancer patients when imaging and clinical parameters were combined in a multimodal network [30]. Also in stroke outcome prediction, these approaches have been pursued. A summary of studies concerning automated functional outcome prediction in AIS patients from baseline image and clinical parameter data with their performance is provided in Table I.

In 2019 Bacchi et al. [31] conducted a pilot study over 204 patients who received IV-tPA therapy. They merged a CNN trained on non-contrast enhanced CT images with an Artificial Neural Network (ANN) trained on 11 clinical parameters for the prediction of good functional outcome and trained the entire network from scratch. The result was compared with single use of the CNN and ANN. Their best performing model was the combined network, reaching an accuracy and AUC of 0.74 and 0.75 respectively.

Zihni et al. [32] proposed two learning strategies for their combined network: extracted features and end-to-end. For the processing of 7 clinical parameters a Multilayer Perceptron (MLP) was used with a single fully connected layer. TOF-MRA data of 316 patients was processed with a 3D CNN consisting of 3 convolutional blocks. Their best performing model was the end-to-end model, reaching a mean AUC of 0.76 over 5 folds. The extracted features strategy had an equal performance as the MLP with a mean AUC of 0.75.

C. Contributions

Image data analysis for functional outcome prediction in stroke thus far yielded a moderate performance. Reported predictions from image data analysis were relatively poor compared to clinical parameter analysis. Possibly, the performance of image data analysis could improve to some extent, but for substantial improvement it might be necessary to include clinical parameters. Combining baseline imaging with clinical information showed improvement in performance compared to models processing images [31][32]. Combined models performed similar [32] or better [31][33][34] than clinical parameter analysis. Overall, the effect of combining images with clinical parameters is small but promising. To the best of my knowledge, no studies combined baseline clinical parameters and baseline CTA for (dichotomized) functional outcome prediction after IAT in stroke patients.

The contribution of this study is two-fold, we apply a CNN trained on baseline volumetric CTA data from a large dataset for the prediction of dichotomized functional outcome and investigated how this model relates to traditional clinical parameter analysis. Second, we apply baseline CTA data and clinical parameters in a combined model for the outcome prediction in stroke patients, which has not been done for a large dataset and for this image modality. Different network architectures and training strategies were adopted with the purpose of automating the process of functional outcome prediction in AIS patients.

The remainder of this report is organized as follows: Section II describes the data used in our experiments and provides an overview of our proposed networks. Section III presents the experimental design. Section IV presents the results. Section V provides further discussion, section VI concludes the report.

II. METHOD

In our method we first applied a preprocessing of the imaging data to address the variation in image resolution and enhance the contrast in brain tissue. Subsequently, the imaging data is processed by a CNN, after which imaging data is combined with baseline clinicalF parameters processed by a MLP. In the next section we describe the data used, followed by processing of image data. Finally, we present an overview of our proposed networks.

A. Data

The dataset used in this project was of patients from the MR CLEAN registry 1 and 2 [27]. This registry is an ongoing, prospective, observational multicenter study at 17 medical centers distributed across The Netherlands, containing AIS patients treated with IAT registered since March 2014. For each patient the dataset includes CTA image data acquired before IAT and clinical and radiological parameters, such as: patients demographics, medical history, and outcome data. Functional outcome is available as the mRS after 90 days.This 7 point scale measures the degree of disability and dependence in daily live activities of stroke patients. 0 indicates no symptoms and a score of 6 indicates the patient did not survive. Scores in between are linked to progressing disability, see Table 2. For the dichotomized mRS a good functional outcome corresponds a score between 0-2 an a poor outcome a score between 3-6.

TABLE 2: The modified Rankin Scale

score	description.
0	No symptoms.
1	No significant disability. Able to carry out usual activities despite some symptoms.
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.
4	Moderate 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	Dead.

B. Data Selection

Data were selected based on image quality, performed procedure, stroke location and availability of the mRS after 90 days. Images were excluded when (1) scans consisted of less than 50 slices, (2) slice spacing or slice thickness was >1.5 mm and (3) slice thickness < slice spacing, followed by a manual inspection to remove images with poor quality and to select one image per patient. The quality during manual inspection was assessed by brain coverage, visible artefacts, symmetry of the brain (whether a part of the frontal lobe or the occipital lobe is cropped from the image), contrast and sharpness. Patients with large vessel occlusions are generally eligible for IAT. Patients treated with IAT and an occlusion in the internal carotid artery, M1 segment or M2 segment were included, see Figure 1.

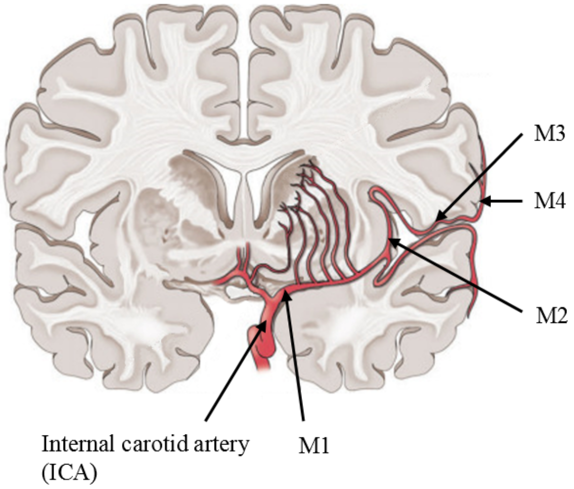


Fig. 1: Brain vasculature. [35]

The dataset consisted of 3280 patients, based on our image quality criteria 1569 were excluded. Of these 231 were excluded due to a failed registration, 111 were excluded due to missing mRS scores and 369 were excluded because they contained occlusion locations other than the ICA, M1 and M2 segment. The flowchart is provided in Supplementary Data A. Table 3 provides baseline characteristics of patients used in our experiments.

Two different sets of clinical predictors were selected as clinical input. One set adopted 11 clinical predictors used by Venema et al. [15] as input: age, known diabetes mellitus, systolic blood pressure, National Institutes of Health Stroke Scale (NIHSS), pre-mRS, glucose, ASPECT, use of IV alteplase, location of occlusion, collateral score and the estimated time from onset to groin puncture. The second set excluded radiological image biomarkers from the first set, leaving 9 clinical predictors. Ordinal parameters such as pre-mRS, NIHSS, ASPECT and collateral score were treated as linear continuous scores. Missing values (less than 2 percent) were imputed with the mean (see Supplementary Data C), hereafter all clinical features were normalized to [0,1].

C. Image Data Preprocessing

The raw CTA scans were of different slice spacing and axial extent and were acquired with various acquisition protocols. The main purpose of the preprocessing was to spatially align the images to an atlas, and to adapt the intensity range. It was assumed that such a preprocessing will be beneficial for the subsequent CNN. Therefore, the CTA scans were first registered to a reference brain atlas with no abnormalities [36], using rigid and affine registration with ANTs software [37]. In order to maintain high level brain structures, an additional elastic transformation was applied to the reference image, and the mask images that were defined in the reference image space. Intensity values were clipped between -40 and +260 Hounsfield Units and normalized to [0,1]. The skull was removed by multiplying the baseline CTA images with a brain mask which was available from the brain atlas. Finally, all occluded hemispheres were flipped to the same lateral side. To preserve spatial context information, we opted to keep the image data volumetric. The final input size of both hemispheres for the DL models was 80x160x112 pixels (voxel size of 1x1x1 mm). Figure 2 shows an example image after processing.

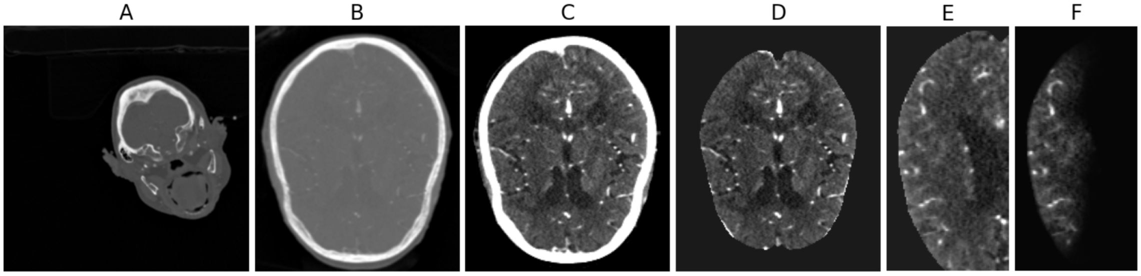


Fig. 2: (A) Original image, (B) spatially aligned to the reference scan (C) intensity range range normalized between -50 + 260 HU (D) brain mask applied (E) occluded hemisphere aligned (F) occluded hemisphere masked with MCA region.

D. Proposed Networks

We proposed two network architectures, one for the processing of image data and one for the processing of clinical (and radiological) parameters. Various training strategies were applied for the fusion of image and clinical data, leading to a total of four different approaches.

1) CNN: Image data were processed using a Siamese Neural Network, which was used for the classification of collateral score from CTA images. This network consisted of two branches that contained identical subnetworks and configuration of parameters and weights. The network expected three inputs: a volumetric image of the occluded hemisphere, a volumetric image of the healthy hemisphere and a mask containing the region of interest. We opted for a probability density mask of the MCA region. The network is based on voxelwise residual network (VoxResNet) [38]. Deep residual learning networks currently have state-of-the-art performance on 2D image recognition tasks and are best known for tackling the optimization degradation problem by approximating the objective with residual functions [39][40]. VoxResNet extended this concept to volumetric input data. Feature maps were multiplied with a probability density map of the MCA region followed by a global average pooling (GAP) layer. The network ended with a fully connected (FC) layer with sigmoid normalization, which output the predicted probability. See Figure 3. The CNN was used to evaluate the predictive value of baseline CTA for the prediction of dichotomized outcome. Subsequently, the CNN served as feature extractor for the network architectures combining imaging and clinical parameters.

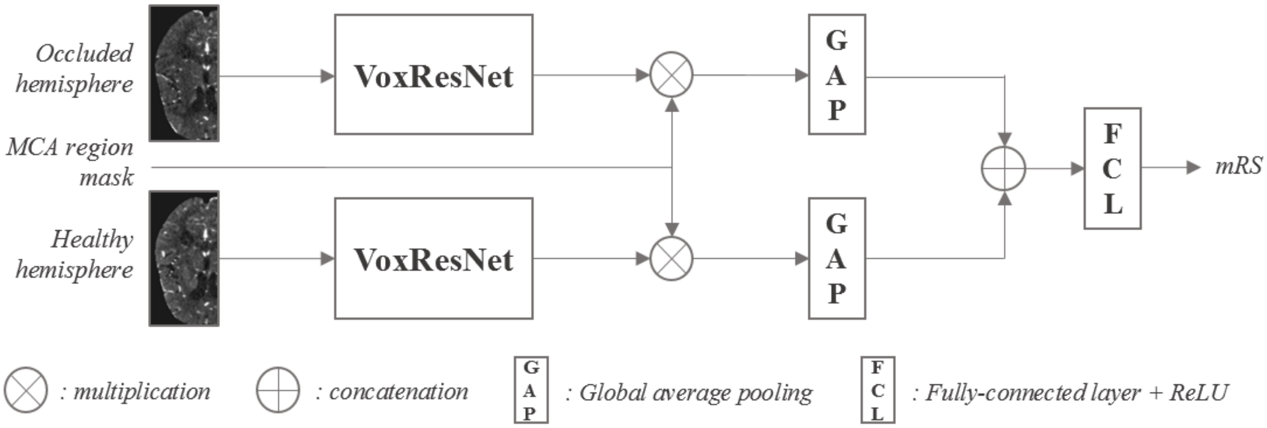


Fig. 3: Diagram of the CNN used for the processing of image data.

2) MLP A: Baseline clinical (and radiological) parameters were processed by MLP. The MLP consisted of a single fully connected (FC) hidden layer of 10 neurons. The hidden layer allowed the model to make non-linear relations between input parameters and output. This network was used to process 9 clinical parameters (MLP A.1) and 9 clinical and 2 radiological parameters (MLP A.2). Both networks served as baseline. MLP A.1 is used as building block for the combined network architectures.

3) MLP B: The third approach combined image features with clinical parameters. The network required 56 image features and 9 clinical parameters as input. Image features were extracted and frozen from the pre-trained CNN. Clinical parameters were processed by MLP A.1, followed by concatenation with the pre-trained image features. The first network (MLP B.1) ended with a fully connected (FC) layer with sigmoid normalization, which output the predicted probability. The second network (MLP B.2) added one FC layer with 10 neurons after concatenation, see Figure 4. Adding an extra layer increased the number of weights, i.e. the model complexity.

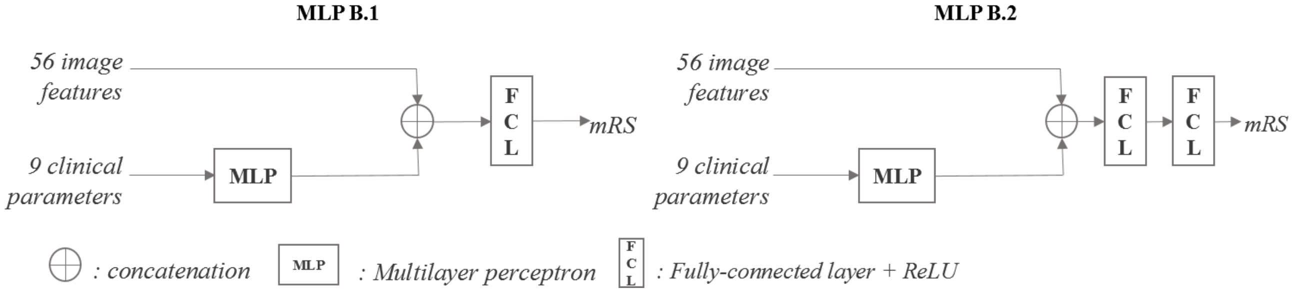


Fig. 4: Diagram of the MLP used for the processing of image features and clinical parameters.

4) MLP + CNN: The last approach combined baseline CTA images with 9 clinical parameters. Images and clinical parameters were processed by the CNN and MLP A.1 respectively. Image and clinical features were concatenated, and fed into two FC layers adopted from MLP B.2. The network is trained end-to-end, which means the network processed the images and clinical parameters simultaneously. Three training variants were applied to this network. The first variant applied random weight initialization at the beginning of training (CNN+MLP 1). The second and third variant loaded in weights from previously trained networks. The weights provided a starting point for the optimization of the model, they were not frozen. Weights of the CNN were applied in the second variant (CNN + MLP 2). The third variant (CNN + MLP 3) applied the weights of CNN and MLP B.2.

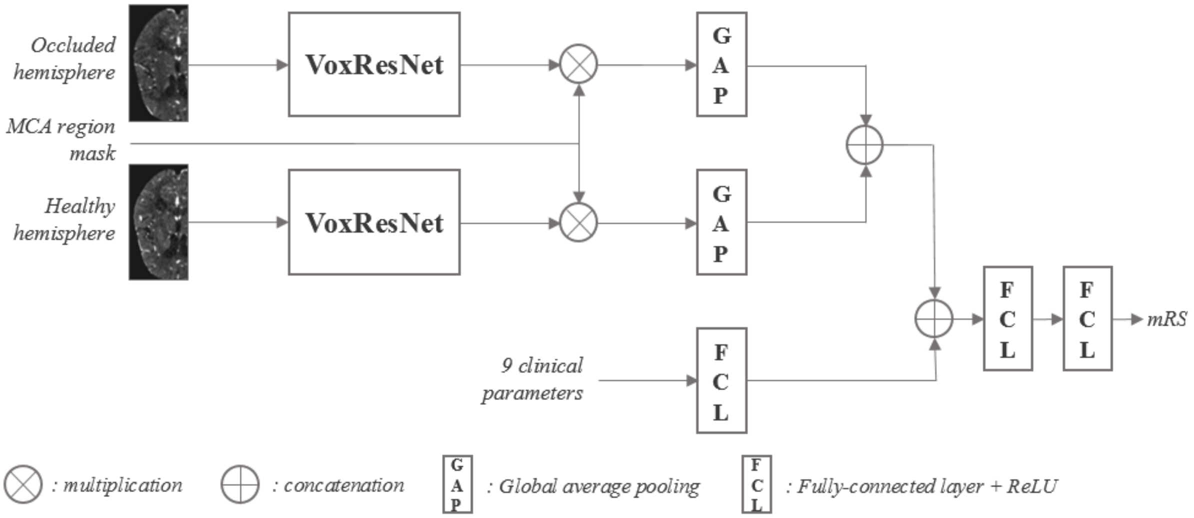


Fig. 5: Diagram of the MLP+CNN used for the processing of images and clinical parameters.

TABLE 3: Patients characteristics at baseline in MR CLEAN Registry and MR CLEAN Trial

Characteristics	MR CLEAN Registry N = 1000	MR CLEAN Registry mRS 0-2 N = 417	MR CLEAN Registry mRS 3-6 N= 583	MR CLEAN Trial N = 132	MR CLEAN Trial mRS 0-2 N = 42	MR CLEAN Registry mRS 3-6 N= 90
Median age (years) (IQR)	71 (62 – 79)	67 (55 – 74)	75 (66 – 83)	64 (54-76)	59 (48-69)	67 (58-79)
Male (%)	506 (50.6)	233 (55.9)	273 (46.8)	76 (58)	22 (52)	54 (60)
Median NIHSS (IQR)	16 (11 – 19)	14 (9 – 17)	17 (17 – 21)	17 (14-20)	15 (11-19)	18 (15-21)
Pre-stroke mRS (IQR)	0 (0 – 0)	0 (0 – 0)	0 (0 – 2)	0 (0-0)	0 (0-0)	0 (0-1)
Diabetes mellitus (%)	162 (16.2)	40 (9.6)	122 (20.9)	19 (14)	2 (5)	17 (19)
Mean systolic blood pressure (mmHg) (std)	150 (24.5)	147 (22.6)	152 (25.5)	143 (23.1)	137 (19.9)	146 (24.0)
Mean glucose (std)	7.4 (2.2)	6.9 (1.8)	7.7 (2.5)	7.6 (4.4)	6.6 (1.4)	8.1 (5.2)
IV alteplase (%)	702 (70.2)	310 (74.3)	392 (67.2)	110 (83.3)	38 (91)	72 (80)
Median ASPECT score (IQR)	9 (8 – 10)	9 (8 – 10)	9 (7 – 10)	9 (7-10)	9 (8-10)	8 (7-10)
Occlusion Location(%) •*ICA* •*M*1 •*M*2	235 (23.5) 530 (53.0) 235 (23.5)	64 (15.3) 245 (58.8) 108 (25.9)	171 (29.3) 285 (48.9) 127 (21.8)	0 (0) 107(81.1) 25 (19)	0 (0) 33 (79) 9 (21)	0 (0) 74 (82) 16 (18)
Collateral score(%) •*Absent* • $< 50\%$ • $> 50\% < 100\%$ • 100%	54 (5.4) 364 (36.4) 393 (39.3) 189 (18.9)	10 (2.4) 124 (29.7) 182 (43.6) 101 (24.2)	44 (7.5) 240(41.2) 211 (36.2) 88 (15.1)	6 (5) 35 (27) 55 (42) 36 (27)	0 (0) 7 (17) 17 (41) 18 (43)	6 (7) 28 (31) 38 (42) 18 (20)
Estimated time from onset to groin puncture (minutes) (IQR)	190 (140–254)	173 (134–235)	200 (152–268)	258 (210-307)	237 (196-299)	270 (221-313)

mRS = modified Rankin Scale; IQR = interquartile range; NIHSS = National Institutes of Health Stroke Scale; ASPECT = Alberta Stroke Program Early CT

E. External Dataset

External data were available from the MRCLEAN Trial [7]. After applying the exclusion criteria listed in II.B 134 patients were included in the dataset. Preprocessing of the external data were performed as described in section II.b and II.c. Baseline characteristics of patients used as external data are provided in Tabel 3.

III. EXPERIMENTS

A. Implementation

Each network was implemented in Python 3.7.4 and TensorFlow 2.2.0. Experiments were conducted on a computer installed with a NVIDIA GeForce RTX 2080 GPU. The ADAM was optimizer used to minimize the Binary Cross Entropy (BCE) loss. The BCE loss is a common choice for binary classification tasks. The loss function encompasses the performance of the model, a loss of 0 indicates that the model predicts all subjects correctly. Training lasted for 90 epochs, all training data were iterated once per epoch. A mini-batch size of 1 was set due to the limited GPU memory. See Supplementary Data B for the complete list of hyperparameter settings.

B. Hyperparameter Optimization

Learning rate (scheduler), number of image features and image data augmentation were tuned on the validation set of the CNN and applied to the remaining network architectures. Data augmentation was applied on image data on-the-fly and consisted of random translations (x-direction∈[5,5]; y-direction∈[-15,6]; z-direction ∈[-5,5]) and rotations (∈[15°,15°]). Transformation range was set while keeping the entire ipsilateral MCA region visible in the bounding box, as well as the MCA region of the contralateral hemisphere not appearing. We applied a learning rate scheduler with a warm-up of 20 epochs to a learning rate of 0.0002 followed by decay. Warmup removed the effect of early overfitting. By decreasing the learning rate in the scheduler we enabled the network to take smaller steps which might avoid overshooting low areas of the loss landscape. The number of image features was set at 28.

C. Evaluation Criteria

Model performance was evaluated on the test set using the Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. Poor functional outcome was defined as the positive class. We tested whether models combining imaging and clinical parameters differed significantly from models using clinical (and radiological) parameters using the DeLong test [41]. Difference was considered statistically significant at a p-value <0.05.

Furthermore, for the networks described in II.d we presented the following results:

- The learning curve of the loss function and AUC.
- ROC-curve.
- Confusion matrix.
- Violin plot.

The learning curves illustrated the training process and kept track of overfitting. The ROC curve, confusion matrix and violin plot were generated on the test set. The confusion matrix showed the relation between the predicted label against the actual label. The accuracy, specificity and sensitivity can be derived from the confusion matrix. The threshold that determined the predicted label is set by the Youden Index. The distribution of the predicted probability against the dichotomized and categorical class label is visualized as a violin plot.

D. Experimental Setup

Experiments were performed for each network approach and training variation described in section II.d. All followed the same pipeline. Model evaluation was performed on the test set for 1000 patients through cross-validation. The data were randomly split into 5 folds, 3 folds for training, one fold for validation and one fold for testing. All splitting was done in a stratified manner to preserve class balance. The validation set was used for selecting model weights (weights are saved at the lowest validation loss).

Models trained during cross-validation were evaluated against the external dataset for each network approach and training variation. At each fold the best performing model is saved. Hence, for each approach and training variation 5 models were created and tested against the external dataset.

IV. Results

Table 4 summarizes performances for the network approaches and training variations described in II.d. The highest average cross-validated result was found for the models combining imaging and clinical data trained end-to-end with weight initialization. For each experiment the learning curves of the loss function and AUC are provided in Supplementary Data D. Supplementary Data E shows the ROC-curve for each network architecture and training strategy. Confusion matrices of the predictions on the test sets are depicted in Figure 6 and Supplementary Data F. Violin plots of the predicted probability of CNN + MLP 3 against the dichotomized and categorical label are shown in Figure 7. The violin plot showed that the the predicted probability value tends to increase with increasing labels of the (dichotomized) mRS.

TABLE 4: Cross validation results

	Test performance AUC					
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
CNN	0.578	0.754	0.657	0.677	0.698	0.673
MLP A.1	0.740	0.777	0.716	0.754	0.741	0.746
MLP A.2	0.752	0.775	0.740	0.774	**0.803**	0.769
MLP B.1	0.763	0.802	0.717	0.770	0.736	0.758
MLP B.2	**0.774**	0.814	0.719	0.762	0.755	0.765
MLP + CNN 1	0.728	0.810	0.697	0.740	0.747	0.744
MLP + CNN 2	0.771	**0.820**	**0.740**	**0.781**	0.791	**0.781**
MLP + CNN 3	0.760	**0.820**	0.746	0.785	0.783	0.779

A. *CNN*

The CNN had an average cross-validated AUC score of 0.66 over the 5 folds on the test set. There was a strong overfitting behavior found from the learning curves of the loss function.

B. *MLP A*

MLP A.1 and MLP A.2 achieved an average cross validated AUC score of 0.75 and 0.77 respectively. The significance test showed that MLP A.2 performed significantly better than MLP A.1 (p-value = 2e-4).

C. *MLP B*

MLP B.1 achieved an average cross validated AUC score of 0.76. MLP B.2 performed slightly better achieving an average cross validated AUC of 0.77. Both MLP B.1 and MLP B.2 achieved a higher AUC score than the MLP that processed clinical parameters, though not statistically significant (p-value = 0.4 and 0.2 respectively).

D. *MLP + CNN*

MLP + CNN 1 was the only model combining imaging and clinical parameters that achieved a lower averaged cross validated AUC score than MLP A.1 (AUC = 0.74). The model performed significantly worse than MLP A.2 (p-value = 0.03).

MLP + CNN 2 achieved the highest average cross validated AUC score (AUC = 0.78). The statistical test showed that the end-to-end III approach performed significantly better than MLP A.1 (p-value = 2e-4). No significant difference was found with the MLP A.2 (p-value = 0.2).

MLP + CNN 3 achieved an average cross validated AUC of 0.78. The model performed significantly better than MLP A.1 (p-value = 8e-4). No significant difference was found with the MLP A.2 (p-value = 0.2).

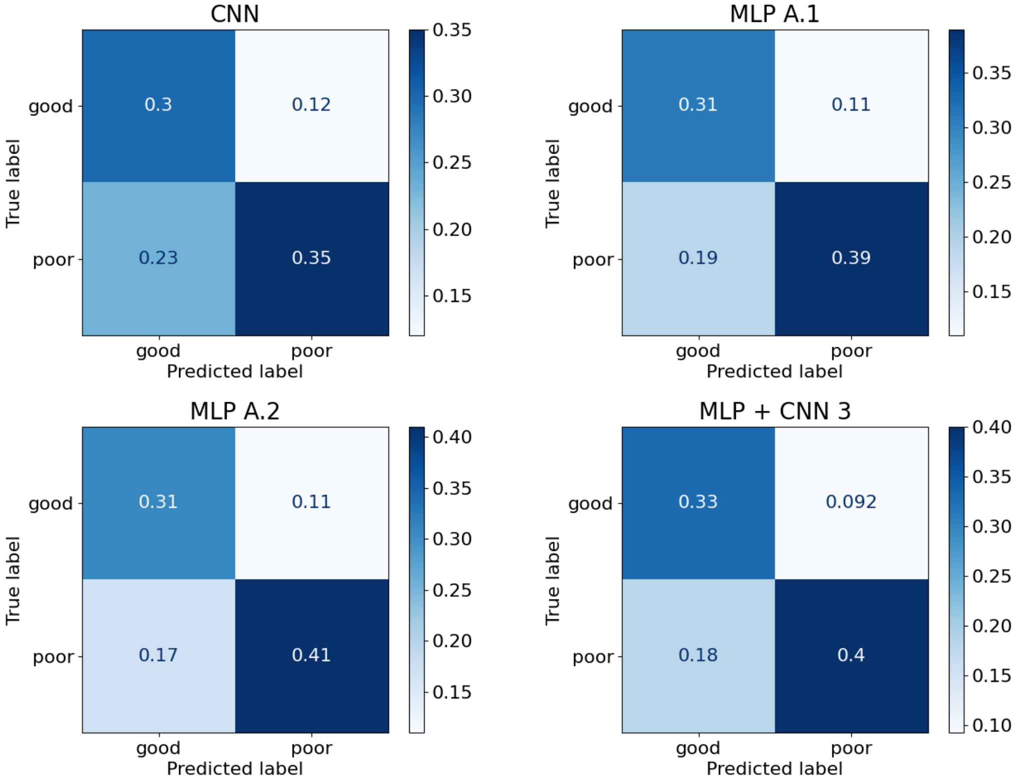


Fig. 6: Confusion matrices of the CNN, MLP A.1, MLP A.2 and MLP+CNN 3. Elements of the matrix are normalized.

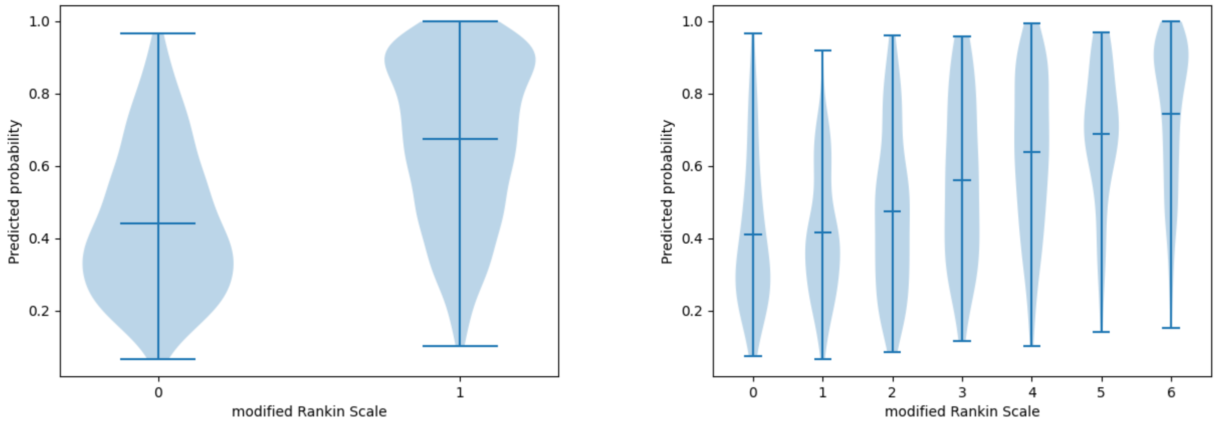


Fig. 7: MLP + CNN 3: Violin plots of (left) the predicted probability related to the dichotomized mRS and (right) the predicted probability related to the categorical mRS

E. *External Dataset*

Table 5 summarizes the results of the assessment of the external dataset on the models trained during cross validation. The average cross-validated AUC was highest for MLP A.2. Except CNN + MLP 1, all combined networks achieved a higher cross validated AUC than the MLP processing clinical parameters.

TABLE 5: External test set results

	Test performance AUC					
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
CNN	0.637	0.681	0.634	0.584	0.654	0.638
MLP A.1	0.673	0.698	0.743	0.735	0.736	0.717
MLP A.2	0.724	**0.747**	**0.761**	0.753	**0.740**	**0.745**
MLP B.1	0.715	0.693	0.740	0.731	0.711	0.718
MLP B.2	**0.727**	0.735	0.744	**0.763**	0.728	0.739
MLP + CNN 1	0.654	0.693	0.640	0.654	0.665	0.661
MLP + CNN 2	0.718	0.723	0.738	0.756	0.730	0.733
MLP + CNN 3	0.701	0.744	0.725	0.730	0.707	0.721

V. DISCUSSION

We have shown the potential predictive value of baseline CTA images for the prediction of dichotomized functional outcome after IAT in stroke patients. On the MR CLEAN Registry, pretrained end-to-end models combining CTA images with baseline clinical parameters significantly increases AUC compared to the model processing clinical parameters. Except for the end-to-end model with random weight initialization, all combined models had a similar AUC as the model combining clinical and radiological parameters.

We first investigated to what extent baseline CTA images can be used for the prediction of dichotomized functional outcome. To this end, an existing Siamese Neural Network once used for classification of collateral score was adopted. The confusion matrix demonstrates an accuracy of 0.65, which was lower than accuracies found for the MLPs and MLP + CNNs. This is in line with the expectations. It was hypothesized that image data might improve the performance of prediction models based on clinical parameters, rather than providing reliable functional outcome predictions by itself.

Secondly, it was investigated whether combining baseline image information with clinical parameters improved prediction accuracy compared to prediction methods based on clinical (and radiological) parameters. Various networks and training strategies were assessed.

MLP A.1 and MLP A.2 served as baseline for the combined networks. MLP A.2 added two radiological biomarkers to the clinical input. The significance test showed that this model had a significant increase in performance compared to the model using clinical parameters. The accuracy of MLP A.2 was slightly higher than MLP A.1, achieving a value of 0.72 and 0.70 respectively. The results suggests that the information derived from images has adds to the predictive value for the dichotomized functional outcome prediction after IAT in stroke patients.

The models with frozen image features achieved a higher average cross-validated AUC than the model processing clinical parameters, though not statistically significant. Furthermore, MLP B.2 performed better in terms of cross-validated AUC and accuracy than MLP B.1. For this reason, the added FC layer of MLP B.2 was adopted in CNN + MLP models.

MLP + CNN 1 achieved a lower average cross-validated AUC then the models processing clinical (and radiological) parameters. The network performed significantly worse than MLP A.2. It was hypothesized that adding weights of previously trained models might improve the performance by guiding the model in a certain direction in the loss landscape, which was adopted in MLP + CNN 2 and MLP + CNN 3.

The end-to-end models with weight initialization yielded a significantly better average cross-validated AUC score than the model processing clinical parameters. From the confusion matrices an accuracy of 0.72 and 0.73 was found for MLP + CNN 2 and MLP + CNN 3 respectively.

Following these results we found that combining baseline images with clinical parameters can improve the performance of models processing clinical parameters for the dichotomized functional outcome prediction of stroke patients. Though, the difference is small and therefore might not be relevant. Compared to the model processing clinical and radiological parameters a combined model can achieve similar performance. However, it might be advantageous to replace radiological parameters with CTA in a model combining imaging and clinical information as it improves assessment time and removes the inter-observer variability in radiological scoring.

The relation between predicted probability and actual class label was explored visually using a violin plot. The distribution of predicted probabilities is positively correlated with the dichotomized and categorical labels. Predicted probability tends to increase with increasing labels of the (dichotomized) mRS.

Results of the CNN were comparable to the study by Hilbert et al. [10]. They selected images from the same dataset but created MIPs on the volumetric data. Creating MIPs or slicing the images decreases training time and memory usage. We opted to keep data volumetric to preserve spatial context. This might not be necessary for functional outcome prediction as results are comparable

Results of the combined imaging and clinical parameter model are in line with recent literature [31][32][33]. They found that adding image data to clinical parameters leads to an increase in AUC compared to models processing images or clinical parameters for the dichotomized functional outcome prediction in stroke. However, in these studies different image modalities and clinical parameters were applied. For example, Zihni et al. [32] applied TOF-MRA image data. In The Netherlands, MRI is generally not the image modality used when patients with stroke symptoms are hospitalized. Commonly, CT images are generated. Two studies applied non-contrast CT [31][33]. This was the first study combining baseline CTA images and clinical parameters for the dichotomized outcome prediction. Furthermore, this study did not only compare the combined model against a model processing clinical parameters, but compared against a model processing clinical and radiological parameters as well. This allowed us to investigate to what extent radiological parameters can be replaced by adding the complete image to clinical parameter models.

Limitations

Training time was longer for the network variations that processed images. As a consequence, the number of folds was limited to 5. Furthermore, cross validation was not performed for the optimization of each model. Instead, hyperparameter settings were tuned on one fold of the CNN and used for remaining networks. Results between folds might be more stable when cross-validation is performed at each fold for hyperparameter settings. The external dataset gave insight to how each model performs at a different test set. For external validation each network and training variation must be trained on the complete MR CLEAN Registry, followed by an evaluation on the MR CLEAN Trial data.

Recommendations

Future work consists of incorporating categorical functional outcome prediction. We investigated the performance of our models on the prediction of dichotomized functional outcome. For dichotomized classification a score of 3 and 6 both belong to the same class. However, the difference between categorical labels is clinically very relevant.

Secondly, in this study we combined clinical parameters with baseline CTA. Information (derived) from non-contrast enhanced CT is not included, adding one channel in the model for non-contrast CT might make the model more complete.

Third, CTA scans were masked with MCA region probability density map. Hereby occlusions in the ICA are not visible in the region of interest. It was hypothesized that occlusions in the ICA would be visible in the M2 segment indirectly. However, using the full brain instead of the MCA region provides more information, which is possibly relevant for outcome prediction.

Furthermore, more insight in the predictions of the automated method can be provided by visualizing relevant regions in the CTA for the classification. Class Activation Mapping [42] is a suitable technique for this task.

Finally, our model could provide a first step into for further development into a treatment decision making tool. For this, data is required of patients who did not undergo IAT, of whom very limited data is available. We suggest to resolve this problem by comparing functional outcome of patients with successful reperfusion with patients that were not successfully treated, classifying them as treated and not treated with IAT respectively.

VI. Conclusion

We investigated to what extent baseline CTA images can be used for the prediction of dichotomized functional outcome after IAT in stroke patients. The CNN processing CTA images achieved a cross-validated AUC of 0.67, which was lower than for models processing clinical (and radiological) parameters. Furthermore, it was investigated whether combining baseline image information with clinical parameters improves performance compared to prediction methods based on clinical parameters. The best performing combined model performed significantly better than the model processing clinical parameters, achieving an average cross-validated AUC of 0.78 and 0.75 respectively. Performance is similar to traditional scoring methods combining clinical and radiological parameters.

References

[1] *Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016.* Geneva, World Health Organization, 2018.

[2] Volksgezondheidenzorg.info, "Beroerte,Cijfers Context,Huidige Situatie," 2021. https://www.volksgezondheidenzorg.info/onderwerp/beroerte/cijfers-context/huidigesituatiebron–node-tabel-bronnen-bij-de-cijfers-over-beroerte. Assessed: 2021-11-03.

[3] J. L. Saver, "Time is brain—quantified," *Stroke*, vol. 37, no. 1, p. 263–266, 2006.

[4] H. B. van der Worp and J. van Gijn, "Acute ischemic stroke," *New England Journal of Medicine*, vol. 357, no. 6, p. 572–579, 2007.

[5] J. Emberson, K. R. Lees, P. Lyden, L. Blackwell, G. Albers, E. Bluhmki, T. Brott, G. Cohen, S. Davis, G. Donnan, J. Grotta, G. Howard, M. Kaste, M. Koga, R. von Kummer, M. Lansberg, R. I. Lindley, G. Murray, J. M. Olivot, and M. Parsons, "Effect of treatment delay, age, and stroke severity on the effects of intravenous thrombolysis with alteplase for acute ischaemic stroke: a metaanalysis of individual patient data from randomised trials," *Lancet (London, England)*, vol. 384, no. 9958, p. 1929–1935, 2014.

[6] W. S. Smith, M. H. Lev, J. D. English, E. C. Camargo, M. Chou, S. C. Johnston, G. Gonzalez, P. W. Schaefer, W. P. Dillon, W. J. Koroshetz, and K. L. Furie, “Significance of large vessel intracranial occlusion causing acute ischemic stroke and tia,” *Stroke*, vol. 40, no. 12, p. 3834–3840, 2009.

[7] O. A. [7] Berkhemer, P. S. Fransen, D. Beumer, L. A. van den Berg, H. F. Lingsma, A. J. Yoo, W. J. Schonewille, J. A. Vos, P. J. Nederkoorn, M. J. Wermer, M. A. van Walderveen, J. Staals, J. Hofmeijer, J. A. van Oostayen, G. J. Lycklama a Nijeholt, J. Boiten, P. A. Brouwer, B. J. Emmer, S. F. de Bruijn, and L. C. M. C. I. van Dijk, “A randomized trial of intra-arterial treatment for acute ischemic stroke,” *The New England journal of medicine*, vol. 372, no. 1, p. 11–20, 2015.

[8] M. Goyal, B. K. Menon, W. H. van Zwam, D. W. Dippel, P. J. Mitchell, A. M. Demchuk, A. Davalos, C. Majoie, A. van der Lugt, M. A. de Miquel, G. A. Donnan, Y. B. Roos, A. Bonafe, R. Jahan, H. C. Diener, L. A. van den Berg, E. I. Levy, O. A. Berkhemer, V. M. Pereira, J. Rempel, and . . . HERMES collaborators, “Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials,” *Lancet (London, England)*, vol. 387, no. 10029, p. 1723–1731, 2016.

[9] P. M. [9] Rothwell, “Subgroup analysis in randomised controlled trials: importance, indications, and interpretation,” *The Lancet*, vol. 365, no. 9454, p. 176–186, 2005.

[10] X. [Li, X. Pan, C. Jiang, M. Wu, Y. Liu, F. Wang, X. Zheng, J. Yang, C. Sun, Y. Zhu, J. Zhou, S. Wang, Z. Zhao, and J. Zou, “Predicting 6-month unfavorable outcome of acute ischemic stroke using machine learning,” *Frontiers in Neurology*, vol. 11, 2020.

[11] A. Hilbert, L. Ramos, H. van Os, S. Olabarriaga, M. Tolhuisen, M. Wermer, R. Barros, I. van der Schaaf, D. Dippel, Y. Roos, W. van Zwam, A. Yoo, B. Emmer, G. Lycklama a Nijeholt, A. Zwinderman, G. Strijkers, . C. Majoie, and H. Marquering, “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*, vol. 115, no. 103516, 2019.

[12] S. Nusinovici, Y. C. Tham, M. Y. Chak Yan, D. S. Wei Ting, J. Li, C. Sabanayagam, T. Y. Wong, and C. Y. Cheng, “Logistic regression was as good as machine learning for predicting major chronic diseases,” *Journal of Clinical Epidemiology*, vol. 56, no. 122, pp. 56–69, 2020.

[13] T. van der Ploeg, D. Nieboer, and E. W. Steyerberg, “Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury,” *Journal of Clinical Epidemiology*, vol. 78, pp. 83–89, 2016.

[14] C. Weimar, A. Ziegler, I. R. König, and H. C. Diener, “Predicting functional outcome and survival after acute ischemic stroke,” *Journal of Neurology*, vol. 249, no. 7, pp. 888–895, 2002.

[15] E. Venema, M. Mulder, and H. F. Lingsma, “Letter by venema et al regarding article, ”validating a predictive model of acute advanced imaging biomarkers in ischemic stroke”,” *Stroke*, vol. 48, no. 8, 2017.

[16] E. Venema, B. Roozenbeek, M. Mulder, S. Brown, C. Majoie, E. W. Steyerberg, A. M. Demchuk, K. W. Muir, A. D´avalos, P. J. Mitchell, S. Bracard, O. A. Berkhemer, G. J. Lycklama ‘A Nijeholt, R. J. van Oostenbrugge, Y. Roos, W. H. van Zwam, A. van der Lugt, M. D. Hill, P. White, B. Campbell, HERMES collaborators, and MR CLEAN Registry Investigators, “Prediction of outcome and endovascular treatment benefit: Validation and update of the mr predicts decision tool,” *Stroke*, vol. 52, no. 9, pp. 2764–2772, 2021.

[17] J. Heo, J. G. Yoon, H. Park, Y. D. Kim, H. S. Nam, and J. H. Heo, “Machine learning–based model for prediction of outcomes in acute stroke,” *Stroke*, vol. 50, no. 5, pp. 1263–1265, 2019.

[18] H. Nishi, N. Oishi, A. Ishii, I. Ono, T. Ogura, T. Sunohara, H. Chihara, R. Fukumitsu, M. Okawa, N. Yamana, H. Imamura, N. Sadamasa, T. Hatano, I. Nakahara, N. Sakai, and S. Miyamoto, “Predicting clinical outcomes of large vessel occlusion before mechanical thrombectomy using machine learning,” *Stroke*, vol. 50, no. 9, pp. 2379–2388, 2019.

[19] X. Li, X. Pan, C. Jiang, M. Wu, Y. Liu, F. Wang, X. Zheng, J. Yang, C. Sun, Y. Zhu, J. Zhou, S. Wang, Z. Zhao, and J. Zou, “Predicting 6-month unfavorable outcome of acute ischemic stroke using machine learning,” *Frontiers in Neurology*, vol. 11, 2020.

[20] H. J. A. van Os, L. A. Ramos, A. Hilbert, M. van Leeuwen, M. A. A. van Walderveen, N. D. Kruyt, D. W. J. Dippel, E. W. Steyerberg, I. C. van der Schaaf, H. F. Lingsma, W. J. Schonewille, C. B. L. M. Majoie, S. D. Olabarriaga, K. H. Zwinderman, E. Venema, H. A. Marquering, and M. J. H. Wermer, “Predicting outcome of endovascular treatment for acute ischemic stroke: Potential value of machine learning algorithms,” *Frontiers in Neurology*, vol. 9, 2018.

[21] C. H. Lin, K. C. Hsu, K. R. Johnson, Y. C. Fann, C. H. Tsai, Y. Sun, L. M. Lien, W. L. Chang, P. L. Chen, C. L. Lin, and C. Y. Hsu, “Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry,” *Computer Methods and Programs in Biomedicine*, vol. 190, 2020.

[22] C. Fernandez-Lozano, P. Hervella, V. Mato-Abad, M. Rodr´ıguez-Y´a~nez, S. Su´arez-Garaboa, I. L´opez-Dequidt, A. Estany-Gestal, T. Sobrino, F. Campos, J. Castillo, S. Rodr´ıguez-Y´a~nez, and R. Iglesias-Rey, “Random forest-based prediction of stroke outcome,” *Scientific Reports*, vol. 11, no. 1, 2021.

[23] H. Asadi, R. Dowling, B. Yan, and P. Mitchell, “Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy,” *PLoS ONE*, vol. 9, no. 2, 2014.

[24] R. Sales Barros, M. L. Tolhuisen, A. M. Boers, I. Jansen, E. Ponomareva, D. Dippel, A. van der Lugt, R. J. van Oostenbrugge, W. H. van Zwam, O. A. Berkhemer, M. Goyal, A. M. Demchuk, B. K. Menon, P. Mitchell, M. D. Hill, T. G. Jovin, A. Davalos, B. Campbell, J. L. Saver, Y. Roos, and H. A. . . . Marquering, “Automatic segmentation of cerebral infarcts in follow-up computed tomography images with convolutional neural networks,” *Journal of neurointerventional surgery*, vol. 12, no. 9, pp. 848–852, 2020.

[25] R. A. Rava, S. E. Seymour, K. V. Snyder, M. Waqas, J. M. Davies, E. I. Levy, A. H. Siddiqui, and C. N. Ionita, “Automated collateral flow assessment in patients with acute ischemic stroke using computed tomography with artificial intelligence algorithms,” *World Neurosurgery*, vol. 155, 2021.

[26] L. N. Do, B. H. Baek, S. K. Kim, H. Yang, I. Park, and W. Yoon, “Automatic assessment of aspects using diffusion-weighted imaging in acute ischemic stroke using recurrent residual convolutional neural network,” *Diagnostics*, vol. 10, no. 10, 2020.

[27] G. R. Jansen IGH, Mulder MJHL and MR CLEAN Registry investigators, “Endovascular treatment for acute ischaemic stroke in routine clinical practice: prospective, observational cohort study (mr clean registry),” *BMJ*, 2018.

[28] H. Nishi, N. Oishi, A. Ishii, I. Ono, T. Ogura, T. Sunohara, H. Chihara, R. Fukumitsu, M. Okawa, N. Yamana, H. Imamura, N. Sadamasa, T. Hatano, I. Nakahara, N. Sakai, and S. Miyamoto, "Deep learning–derived high-level neuroimaging features predict clinical outcomes for large vessel occlusion," *Stroke*, vol. 51, no. 5, pp. 1484–1492, 2020.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation. in international conference on medical image computing and computer-assisted intervention," *Springer*, pp. 234–241, 2015.

[30] S. Mani, Y. Chen, X. Li, L. Arlinghaus, A. B. Chakravarthy, V. Abramson, S. R. Bhave, M. A. Levy, H. Xu, and T. E. Yankeelov, "Machine learning for predicting the response of breast cancer to neoadjuvant chemotherapy," *Journal of the American Medical Informatics Association*, vol. 20, no. 4, 2013.

[31] S. Bacchi, T. Zerner, L. Oakden-Rayner, T. Kleinig, S. Patel, and J. Jannes, "Deep learning in the prediction of ischaemic stroke thrombolysis functional outcomes," *Academic Radiology*, vol. 27, no. 2, pp. 19–23, 2020.

[32] E. Zihni., V. Madai., A. Khalil., I. Galinovic., J. Fiebach., J. Kelleher., D. Frey., and M. Livne., "Multimodal fusion strategies for outcome prediction in stroke," in *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF,*, pp. 421–428, INSTICC, SciTePress, 2020.

[33] Z. A. Samak, P. Clatworthy, and M. Mirmehdi, "Prediction of thrombectomy functional outcomes using multimodal data," 2020.

[34] S. Winzeck, A. Hakim, R. McKinley, J. A. A. D. S. R. Pinto, V. Alves, C. Silva, M. Pisov, E. Krivov, M. Belyaev, M. Monteiro, A. Oliveira, Y. Choi, M. C. Paik, Y. Kwon, H. Lee, B. J. Kim, J. H. Won, M. Islam, H. Ren, and M. . . . Reyes, "Isles 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral mri," *Frontiers in Neurology*, vol. 9, 2018.

[35] H. Blumenfeld, "Coronal view of the basal ganglia with the recurrent artery of heubner, a branch off the aca," 2010. [Illustration] https://www.sciencedirect.com/topics/neuroscience/recurrent-artery-of-heubner.

[36] R. Peter, B. J. Emmer, A. C. van Es, and T. 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)*, pp. 1141–1145, 2017.

[37] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ants similarity metric performance in brain image registration," *NeuroImage*, vol. 54, no. 3, pp. 2033–2044, 2011.

[38] L. Y. J. Q.] H. Chen, Q. Dou and P.-A. Heng, "Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images," *NeuroImage*, vol. 170, pp. 446–455, 2017.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

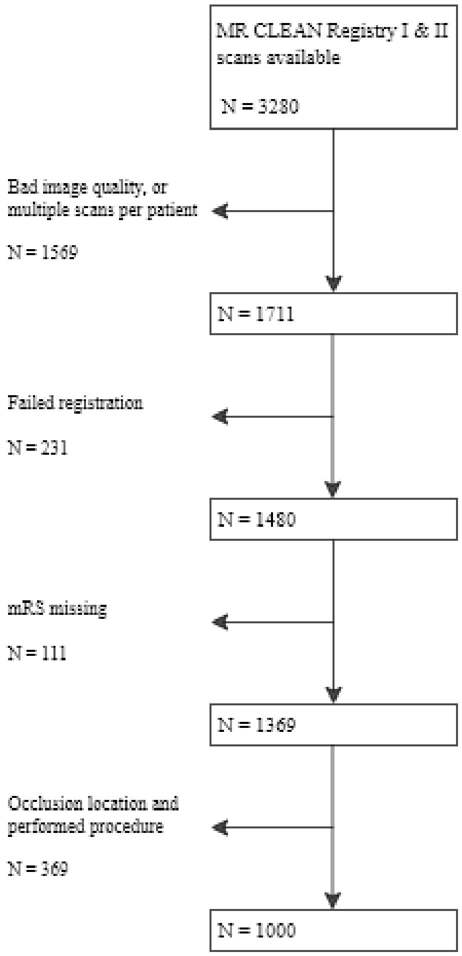
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," *CoRR*, vol. abs/1603.05027, 2016.

[41] "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[42] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *CoRR*, vol. abs/1512.04150, 2015.

Supplementary data

A. Flowchart of included patients from the MR CLEAN Registry

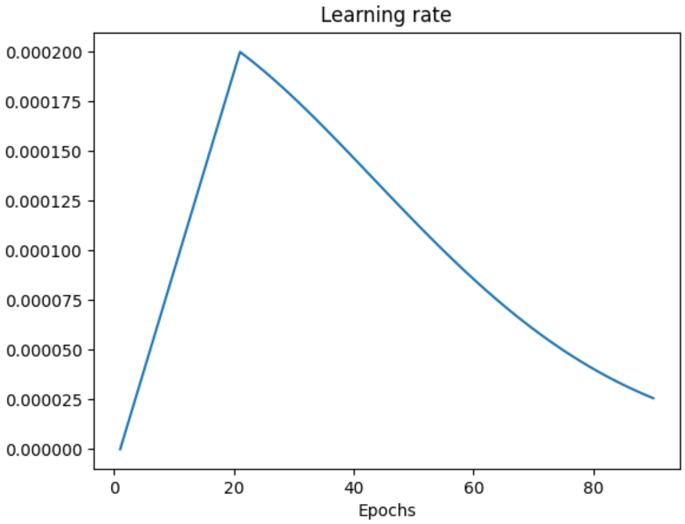

MR CLEAN Registry I & II scans available
N = 3280
Bad image quality, or multiple scans per patient
N = 1569
N = 1711
Failed registration
N = 231
N = 1480
mRS missing
N = 111
N = 1369
Occlusion location and performed procedure
N = 369
N = 1000

B. Imputed clinical and radiological variables of the MR CLEAN Registry

Variables	Missing in selection sheet	Missing in total sheet	Average in selection	Average in total	Imputed
Age	0	0	-	-	-
Baseline NIHSS	11	57	15,03	15,4	15
Pre-stroke mRs	21	72	0.68	0.7	1
Diabetes mellitus	6	24	No	No	No
Baseline systolic blood pressure (mmHg)	22	89	149.9	149.7	150
Baseline glucose	99	371	7.41	7.35	7.4
IV alteplase	4	12	Yes	Yes	Yes
ASPECT score	13	109	8.3	8.3	8
Location of occlusion	0	0	-	-	
CTA collateral score	14	207	⟩50⟨100	⟩50⟨100	⟩50⟨100
Estimated time from onset to groin puncture	4	15	214.5	213.8	214

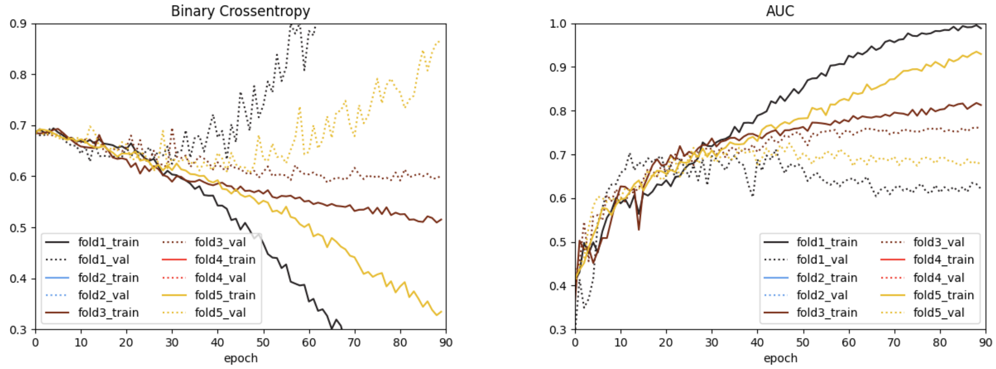
C. Hyperparameter values for network architectures of section II.d.

Hyperparameter	Value
Intensity normalization	Rescaled 0 to 1 for Hounsfield between -40 and +260 HU
Data augmentation: Translation x-direction Translation y-direction Translation z-direction Rotation x-y-z-direction	 [-5, 5] [-15, 6] [-5,5] [-15, 15]
Batch size	1
Epochs	90
Siamese features	28
Optimizer	Adam
Learning rate	0.0002
Learning rate scheduler	Warm-up decay

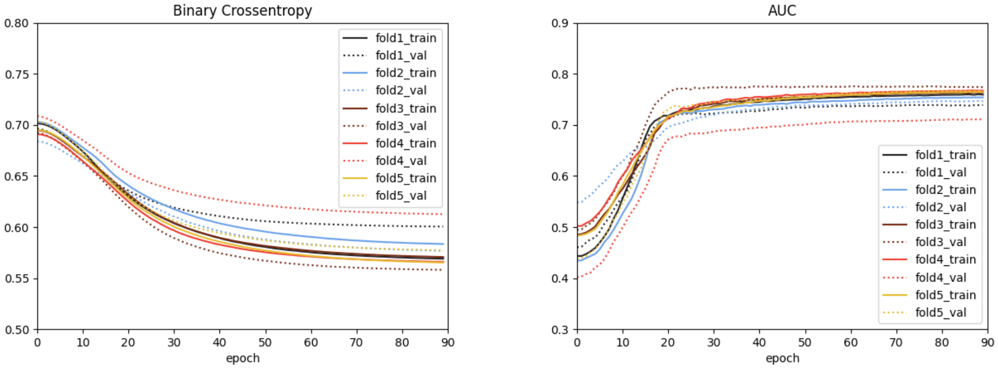


D. Learning curves of the loss function and AUC during training.

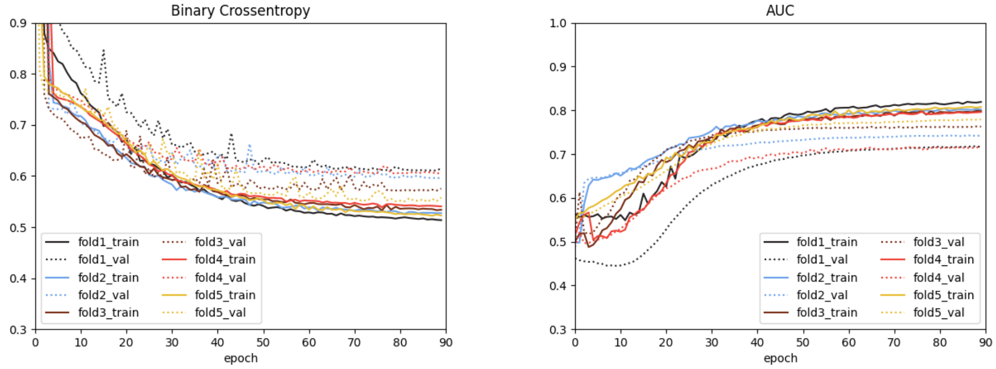
CNN



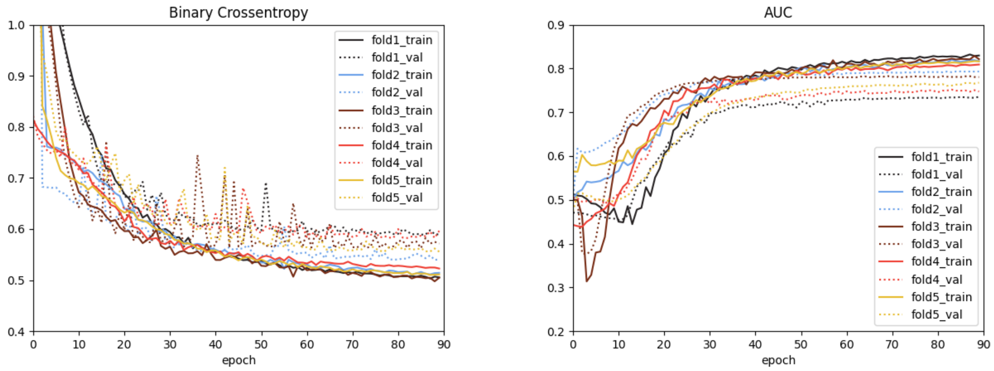
MLP A.1



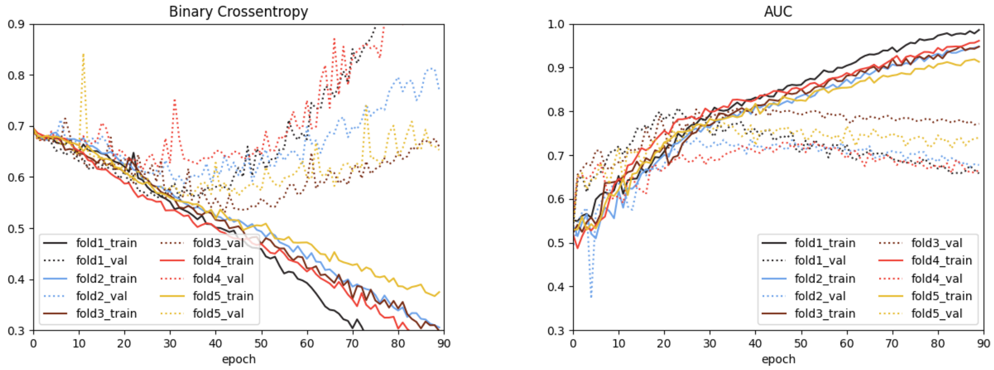
MLP B.1



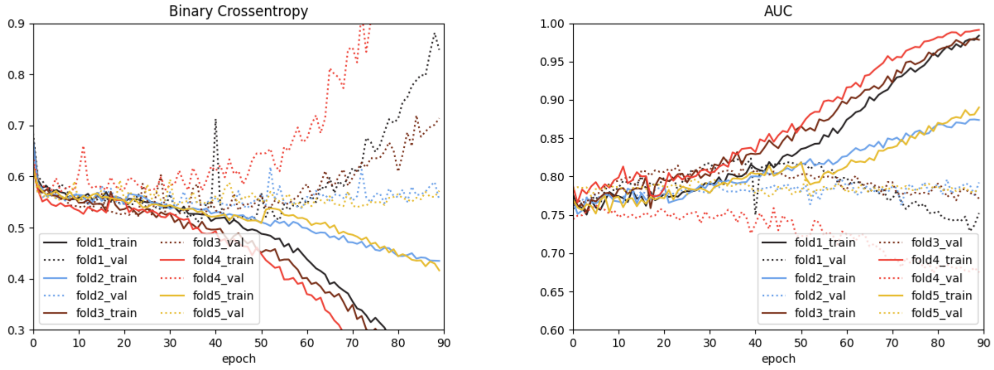
MLP B.2



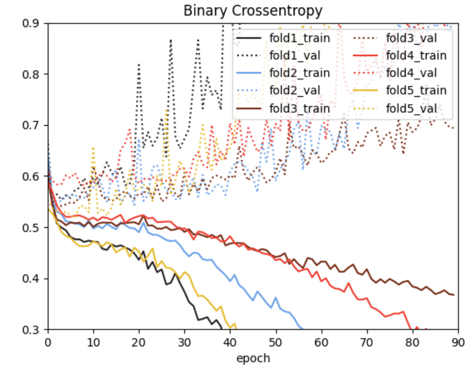
MLP + CNN 1

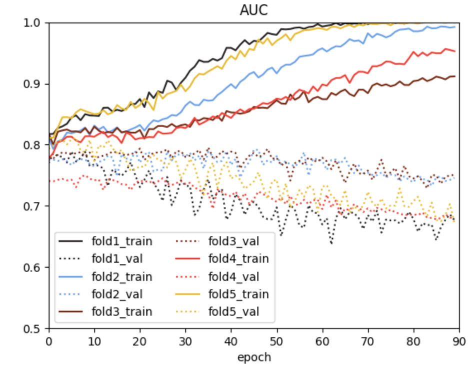


MLP + CNN 2



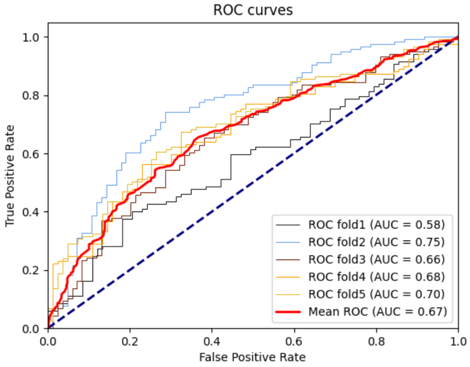
MLP + CNN 3

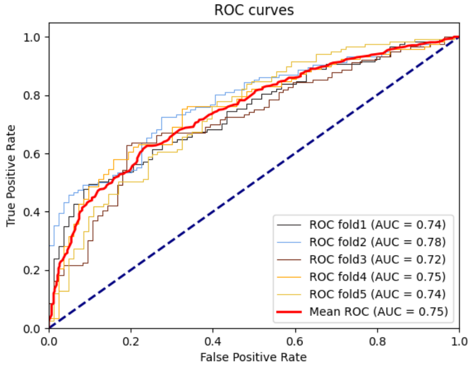




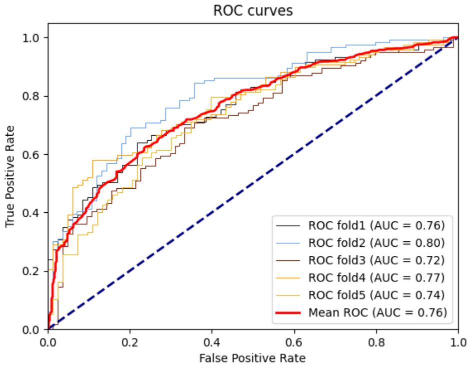
E. ROC-AUC of the cross-validated test set

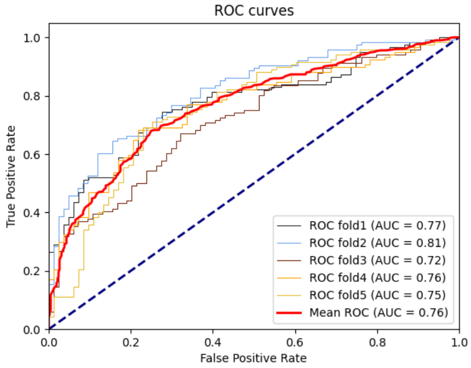
CNN and MLP A.1



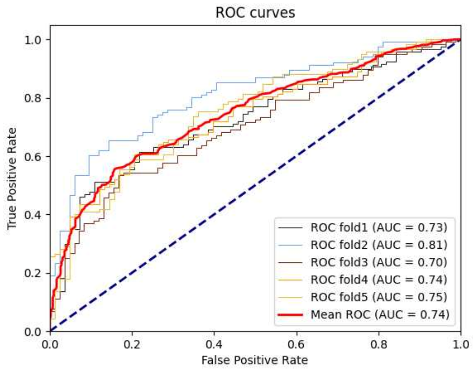


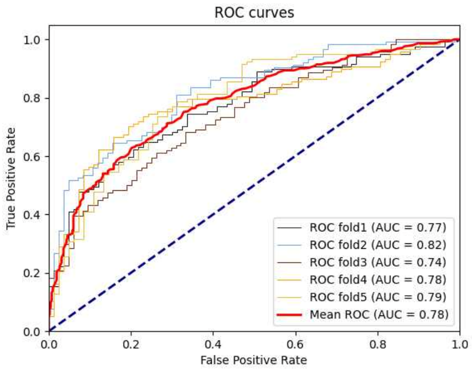
MLP B.1 and MLP B.2



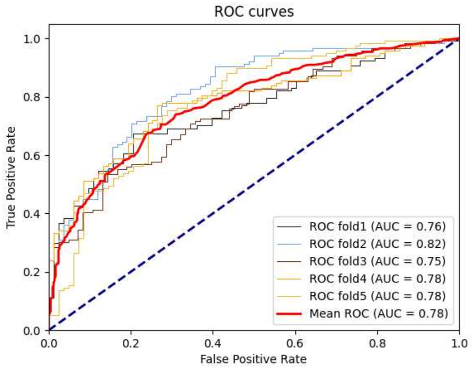


MLP + CNN 1 and MLP + CNN 2

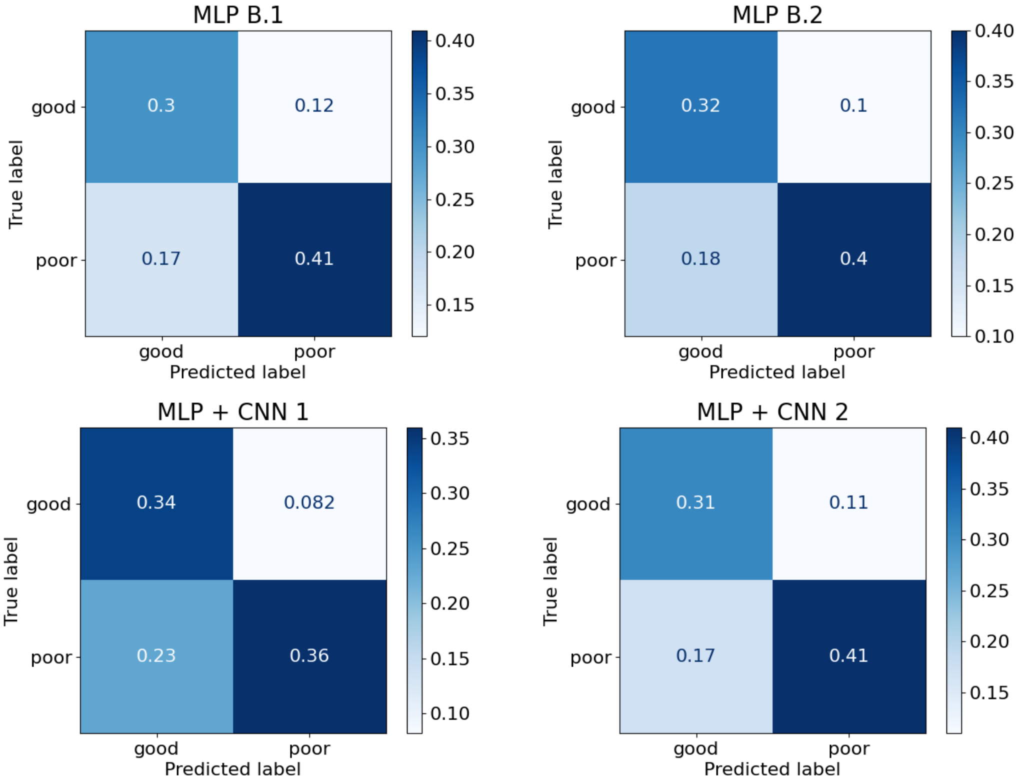




MLP + CNN 3



F. *Confusion matrices of MLP B.1, MLP B.2, MLP + CNN 1 and MLP + CNN 2*


MLP B.1
good
poor
True label
good
poor
Predicted label
MLP B.2
good
poor
True label
good
poor
Predicted label
MLP + CNN 1
good
poor
True label
good
poor
Predicted label
MLP + CNN 2
good
poor
True label
good
poor
Predicted label