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 (2019) 103516

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke

A. Hilbert [a,1], L.A. Ramos [a,b,*,1], H.J.A. van Os [c], S.D. Olabarriaga [b], M.L. Tolhuisen [a,i], M.J. H. Wermer [c], R.S. Barros [a], I. van der Schaaf [d], D. Dippel [e], Y.B.W.E.M. Roos [f], W.H. van Zwam [g], A.J. Yoo [h], B.J. Emmer [i], G.J. Lycklama à Nijeholt [j], A.H. Zwinderman [b], G.J. Strijkers [a], C.B.L. M. Majoie [i], H.A. Marquering [a,i]

[a] Department of Biomedical Engineering and Physics, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands
[b] Department of Clinical Epidemiology and Biostatistics, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands
[c] Department of Neurology, Leiden University Medical Center, Leiden, the Netherlands
[d] Department of Radiology and Nuclear Medicine, University Medical Center Utrecht, Utrecht, the Netherlands
[e] Department of Neurology, Erasmus MC - University Medical Center, Rotterdam, the Netherlands
[f] Department of Neurology, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands
[g] Department of Radiology and Nuclear Medicine, Maastricht University Medical Center, Maastricht, the Netherlands
[h] Neurointervention, Texas Stroke Institute, Dallas-Fort Worth, Texas, USA
[i] Department of Radiology and Nuclear Medicine, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands
[j] Radiology, Haaglanden Medical Center, The Hague, the Netherlands

ARTICLE INFO

Keywords:
Acute ischemic stroke
Radiological images
Deep learning
Prognostics
ResNet
RFNN
Structured receptive fields
Gradient-weighted class activation mapping

ABSTRACT

Treatment selection is becoming increasingly more important in acute ischemic stroke patient care. Clinical variables and radiological image biomarkers (old age, pre-stroke mRS, NIHSS, occlusion location, ASPECTS, among others) have an important role in treatment selection and prognosis. Radiological biomarkers require expert annotation and are subject to inter-observer variability. Recently, Deep Learning has been introduced to reproduce these radiological image biomarkers. Instead of reproducing these biomarkers, in this work, we investigated Deep Learning techniques for building models to directly predict good reperfusion after endovascular treatment (EVT) and good functional outcome using CT angiography images. These models do not require image annotation and are fast to compute. We compare the Deep Learning models to Machine Learning models using traditional radiological image biomarkers. We explored Residual Neural Network (ResNet) architectures, adapted them with Structured Receptive Fields (RFNN) and auto-encoders (AE) for network weight initialization. We further included model visualization techniques to provide insight into the network's decision-making process. We applied the methods on the MR CLEAN Registry dataset with 1301 patients. The Deep Learning models outperformed the models using traditional radiological image biomarkers in three out of four cross-validation folds for functional outcome (average AUC of 0.71) and for all folds for reperfusion (average AUC of 0.65). Model visualization showed that the arteries were relevant features for functional outcome prediction. The best results were obtained for the ResNet models with RFNN. Auto-encoder initialization often improved the results. We concluded that, in our dataset, automated image analysis with Deep Learning methods outperforms radiological image biomarkers for stroke outcome prediction and has the potential to improve treatment selection.

* Corresponding author. Meibergdreef 9, L0-148, Amsterdam, 1105AZ, the Netherlands.
E-mail address: l.a.ramos@amsterdamumc.nl (L.A. Ramos).
[1] Authors contributed equally.

https://doi.org/10.1016/j.compbiomed.2019.103516
Received 25 June 2019; Received in revised form 15 October 2019; Accepted 16 October 2019
Available online 22 October 2019

1. Introduction

Stroke is ranked among the leading causes of death and permanent disability in the last 15 years worldwide [1,2]. Approximately 80% of all stroke patients with untreated large vessel occlusion in the anterior circulation do not regain functional independence or die within 90 days after stroke onset [3]. The Multicenter Randomized Clinical Trial of Endovascular Treatment for Acute Ischemic Stroke in the Netherlands (MR CLEAN Registry) has shown that this patient population can be effectively treated with endovascular treatment (EVT) [4].

Accurate prediction of reperfusion and functional outcome has the potential to improve stroke care, as it could lead to selecting the most beneficial treatment option for the individual patient: to perform or to withhold EVT. Recent studies on outcome prediction strategies in ischemic stroke patients after EVT utilized clinical variables and radiological image biomarkers [5,6]. In favor of standardized prognosis, various radiological stroke imaging biomarkers have been defined by specific, visually observable phenomena that imply stroke severity and functional outcome. These biomarkers include the extent of tissue damage characterized by edema (e.g. ASPECTS [7]) and extent of blood flow through the collateral circulation (e.g. Collateral Score [8]), and they have been proven to be associated with functional outcome. The number of proposed radiological image biomarkers for prognosis in acute ischemic stroke is quite large. In the MR CLEAN Registry, for example, 20 biomarkers have been assessed. These biomarkers are commonly scored manually, may demand a considerable time effort and suffer from observer variability. For the collateral score, observer agreement as low as 50% with kappa's ranging from 0.49 to 0.60 has been reported [9], and for the ASPECTS score a mean deviation close to one point has been found, with above 25% of the cases deviating more than two points [10]. Details about the other image biomarkers are shown in Supplemental Table 1.

Machine Learning (ML) enables the discovery of empirical patterns and linear/non-linear relationships in data through automated algorithms. Regarding imaging data, Deep Learning (DL) algorithms are particularly able to learn important predictive patterns, which may lead to increased prediction accuracy [11]. For example, an encoder-decoder CNN (inspired by the SegNet [12]) was developed to predict final lesion volume and outcome in acute ischemic stroke patients using magnetic resonance imaging, with an AUC of 0.88 (10% higher than linear models) [13]. E-ASPECTS, a machine learning-based commercial software developed to automate ASPECTS scoring in CT scans, has recently been used to predict functional recovery and adverse outcome in acute ischemic stroke patients [14]. In the stroke lesion segmentation challenge (ISLES), a multi-scale 3D-CNN was the best performing model, with an average Dice score of 0.59 [15]. This multi-scale 3D-CNN (named DeepMedic [15]) was also successfully applied to CTA to detect acute ischemic stroke (and segment the lesion) with an AUC of 0.93 [16]. However, these specific DL approaches are generally limited by the manual determination of many of the biomarkers that are considered ground truth but suffer from high inter-observer variability. Besides, DL methods generally come at the cost of high complexity models with low interpretability [17], hampering the applicability in clinical settings.

Due to the recent success of DL approaches in stroke medical imaging, we hypothesize that data-efficient DL methods trained on CT Angiography (CTA) imaging data might outperform well-known radiological image biomarkers in predicting good reperfusion after EVT and good functional outcome in patients with acute ischemic stroke. Next to assessing the accuracy, we adopt visualization techniques, since these prediction systems can be of more assistance when they provide direct insights into their decision-making process beyond generating a probability distribution.

2. Methods

2.1. Clinical data and pre-processing

We included 1526 ischemic stroke patients registered between March 2014 and June 2016 in the MR CLEAN Registry part1 [18]. The MR CLEAN Registry is an ongoing, prospective, observational, multicenter study at 16 intervention hospitals in the Netherlands. Imaging data (CTA scans) available before EVT were used to develop the DL models and to determine radiological image biomarkers by expert radiologists.

The raw CTA scans were of size (512x512xS), where S was the number of axial slices. Due to limited computational resources, we opted to reduce sparsity and dimensionality of images by computing Maximum Intensity Projections (MIPs) from the CTA data in the axial plane. First, CTA scans were co-registered to a reference scan (a scan with no abnormalities) using rigid registration with the Elastix software [19] and the skull was removed from the images with a region growing algorithm since it is of high intensity and can hamper the quality of the MIPs [20]. The attenuation of the MIPs was clipped between +50 and + 400 Hounsfield Units (HU) and normalized to the interval of [0,1]. The surrounding air was removed to reduce image size. The final image data size used as input for the DL models was 368x432 pixels (voxel size of 0.52,0.52 mm). After the pre-processing steps, 225 patients were excluded due to failure during registration, poor image quality, and noise or artifacts, leaving a total of 1301 patients to be used for model development. Table 1 contains the baseline characteristics of the patients used in our models.

We created models to predict two outcome measures. First, good functional outcome after ischemic stroke - defined by the dichotomized modified Rankin Scale (mRS $\leq$ 2) at 90 days - mRS is a scale commonly used to assess the disability of stroke patients in daily activities. mRS ranges from 0 to 6, where zero means no disability, progressing to five (severe disability) and six (death).

Second, good reperfusion - defined by the dichotomized modified

Table 1
Characteristics of patients in MR CLEAN Registry. Values correspond to the percentages of participants unless stated otherwise.

Characteristics	MR CLEAN Registry N = 1301	MR CLEAN Registry mRS 0–2 N = 463	MR CLEAN Registry mRS 3–6 N = 838	MR CLEAN Registry mTICI 0-2a N = 552	MR CLEAN Registry mTICI 2b-3 N = 749
Age (years) (median/IQR)	71 (59 - 79)	66 (55 - 74)	74 (63 - 82)	72 (60 - 80)	70 (59 - 78)
Men (%)	695 (53.4)	262 (56.6)	433 (51.7)	290 (52.5)	405 (54.1)
NIHSS at baseline (median/IQR)	16 (11 - 20)	13 (9 - 17)	17 (13 - 21)	16 (12 - 20)	15 (11 - 19)
Onset to groin puncture time (mins) (IQR)	210 (160 - 270)	190 (145 - 253)	220 (170 - 279)	215 (157 - 281)	205 (160 - 265)
Systolic blood pressure (mm Hg) (Mean/STD)	150 (1.89)	146 (4.86)	152 (3.03)	152 (4.63)	148 (3.16)
Intravenous alteplase treatment (%)	1014 (77.9)	380 (82.1)	634 (75.7)	411 (74.5)	603 (80.5)
ASPECTs at Baseline (subgroups)					
0–4	81 (6.2)	16 (3.5)	65 (7.8)	36 (6.5)	45 (6.0)
5–7	310 (23.8)	95 (20.5)	215 (25.7)	131 (23.7)	179 (23.9)
8–10	880 (67.6)	340 (73.4)	540 (64.4)	370 (67.0)	510 (68.1)

Thrombolysis In Cerebral Infarction score (mTICI ≥ 2b). mTICI is a score that ranges from 0 (no antegrade reperfusion of the occluded vascular territory) to 3 (complete reperfusion). mTICI was assessed by 20 neuroradiologists and one neurologist at an imaging core laboratory. The observers were blinded to all clinical findings, except occlusion location. mTICI was scored on digital subtraction angiography images [18].

2.2. Structured Receptive Field Neural Networks (RFNNs)

Conventional CNNs (Convolutional Neural Networks) hardly excel in the presence of relatively small datasets, which presents a common challenge for many medical applications. To this end, in this work we explore a data-efficient CNN formulation that builds on the structure of biological receptive fields.

Structured Receptive Field Neural Networks (RFNNs) were proposed in Ref. [21] and have been shown to outperform CNNs on small- and medium-sized datasets. RFNNs redefine convolutional kernels as linear combinations of Gaussian derivative filters. Contrary to traditional kernels, only the combination weights are trained, whereas the set of Gaussian derivatives is fixed. In this way, the number of parameters to train is potentially decreased, and prior knowledge about the spatial properties of local features is introduced. Gaussian filters and the Scale-space theory in computer vision [22] have been broadly explored in the medical imaging domain. Scale-space approaches have been successfully applied to medical imaging classification and segmentation with great performance improvements since they often assist the classifiers by revealing low and high-level features without introducing artifacts [23–25]. Furthermore, the interpretability of CNNs is enhanced by explicitly connecting classical image processing methods with the data-driven paradigm.

Fig. 1 illustrates the computation of $I*N$ kernels of a RFNN convolutional layer l, where I is the number of input feature maps and N is the number of output feature maps. This RFNN formulation can be used in any convolutional layer to replace the conventional convolutional kernels while keeping the architecture of a CNN intact.

2.3. Unsupervised pre-training

A random initialization scheme can place the parameters of a CNN in regions that do not generalize well, while the limitations in training data and computational resources create a burden in improving generalization during training (e.g. increasing batch size). These problems make the training of deep architectures unstable and can lead to lower model accuracy [26]. Moreover, supervised training of CNNs is influenced by the ground truth labels, even though learning effective image features does not necessarily rely on image annotation.

To face these challenges, we included unsupervised pre-training in the experiments using stacked denoising convolutional Auto-Encoders (AE) [26]. AEs learn a feature representation by compressing the input into a latent space and subsequently reconstruct the input using this representation. By optimizing the reconstruction from the input data, the AE is able to learn features that best represent the image. We constructed AEs from each CNN model by using their convolutional layers as the encoder part and extending with a corresponding sequence of transposed convolutional blocks as the decoder. Transposed convolutional blocks are comprised of the same Batch-Normalization, ReLU, convolutional sequence but convolutions are replaced by up-sampling transposed convolutions. Using the training data, we trained an AE until the loss between the output and the input images stopped decreasing (depicted by the dashed lines in Fig. 2). The learned encoding part of the network was subsequently used to train a dense layer using the labels (in a transfer learning fashion) [27,28]. The weights from the encoding part were used in two approaches: keeping them frozen during the training or fine-tuning them during the training of the dense layer [27].

2.4. Baseline models

To assess the added value of DL methods compared to existing radiological image biomarkers, we created ML-based models using radiological image biomarkers that have shown state-of-the-art results on the same dataset [5]. In these baseline prediction models, we used 20 radiological imaging biomarkers, which have been manually scored by designated experts of the core-lab of the MR CLEAN Registry. These radiological image biomarkers have shown predictive value for functional outcome and thus are commonly considered in clinical practice [4,29,30]. Supplemental Table 1 lists all included radiological imaging biomarkers. We developed two clinical baseline prediction models for both outcome measures using only radiological image biomarkers. The first is a Logistic Regression (LR) model, following the most common approach in clinical research. The second is a Random Forest Classifier (RFC), which has earlier been successfully used for the same patient population [5].

To assess the benefits of the proposed application of RFNN convolutional layers and unsupervised pre-training, we compared them to a standard DL model. We developed and optimized a ResNet architecture, and used it as the standard DL model. The ResNet architecture was composed of four blocks with two consecutive convolutions in each block, and skip-connections connecting the input and the output of blocks [31]. Inspired by Ref. [28], we followed the Batch-Normalization, ReLU, convolution sequence in each layer. Further details about the ResNet architecture can be found in Supplemental Tables 2–3 and Supplemental Fig. 2. Details of the ResNet-AE implementation for unsupervised pre-training are shown in Supplemental Table 4. RFNN models had the same ResNet (ResNet-AE in case of unsupervised pre-training) architecture, but the conventional convolutional kernels

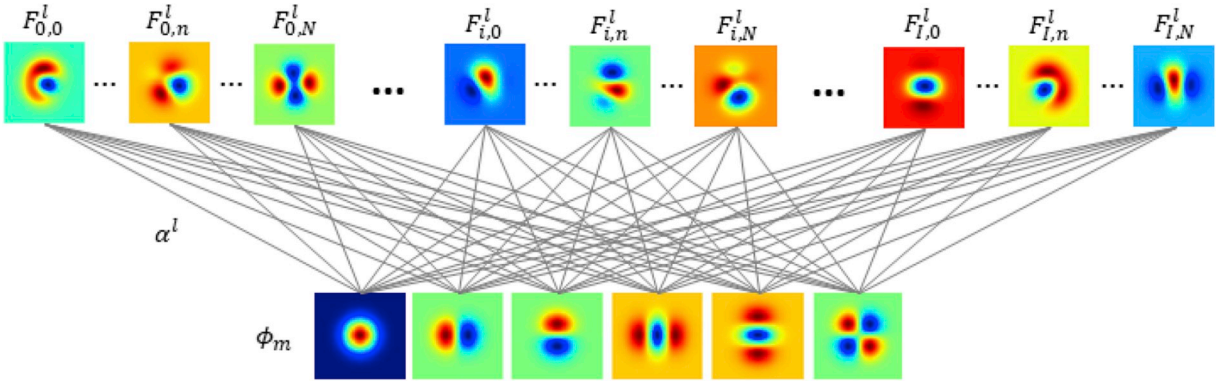

Fig. 1. Construction of RFNN convolutional kernels. φ_m denotes the fixed set of Gaussian derivatives, α^l the combination weights in the l^{th} convolutional layer, and $F^l_{i,n}$ the convolutional kernel producing the n^{th} output feature map of l^{th} layer from i^{th} input feature-map. (Should be printed in color).

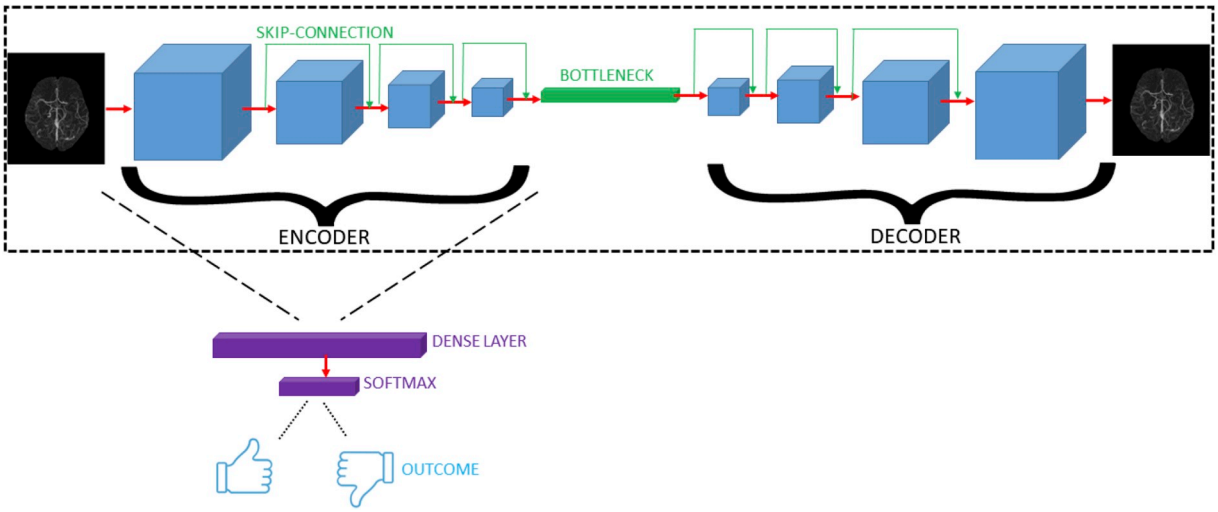

Fig. 2. Unsupervised and supervised training pipeline. First the AE is trained based on the reconstruction loss and then the trained encoder is used to train a dense layer for prediction.

were replaced with structured receptive field kernels.

2.5. Experimental setup

The 1301 patients were split into four balanced folds for cross-validation. In our data, class imbalance for functional outcome (mRS) was 463 good outcome (35.6%) and 838 (64.4%) unfavorable outcome. For reperfusion (mTICI), class imbalance was smaller, with 749 (60.9%) as good reperfusion and 552 (39.1%) as poor reperfusion. Since our data was slightly imbalanced, we opted for balancing the classes using random under-sampling. For each iteration, three folds were used to train and optimize the models, and one fold was used for testing the models. Area Under the Curve (AUC) was used to assess model accuracy. For each CNN model, we created three training schemes: (1) training models from scratch; (2) using unsupervised pre-training to initialize convolutional weights (encoder part of the AE) and keeping them unchanged during supervised training; and (3) keeping pre-trained encoder weights unchanged for 50 epochs, then releasing and fine-tuning the whole architecture. We selected the cut-off of 50 epochs (from 25, 50 or 75) by monitoring convergence of convolutional filters and training loss. Further details about the experimental setup and the hyper-parameters used for optimization can be found in Supplemental Table 5. All experiments were run on a PC with a single Titan X Pascal GPU, AMD Ryzen 7 1700X CPU, 16 GB of RAM memory and Windows 10.

2.6. Model visualization

Deep Neural Networks are commonly referred to as black boxes because of their complex structure utilizing millions of parameters, in contrast to classical image processing techniques. In medical applications, a good prediction system, in addition to high accuracy, also needs to deliver interpretable predictions. Even though RFNN convolutional layers increase interpretability, here we further investigate the explanation of neural predictions of our models. We hypothesize that the best way to explain outcome predictions is to visualize traits of input scans that led the model to the prediction. Various visualization techniques that analyze prediction models have been developed [32,33]. Here, we explored the visualizations with Gradient-weighted Class Activation Mapping (Grad-CAM) [33]. Grad-CAM is a popular technique for generating visual interpretations of CNN-based networks, which fuses the localization and class-discriminative properties of Class Activation Mapping [34] and the precision of Guided Backpropagation [35]. Grad-CAM explains predictions by unveiling the gradient-weighted contribution of convolutional feature maps in the input space. In practice, we used the implementation of a slightly improved version of the technique, namely Grad-CAM++ [36].

We created two visualizations for each of the best mRS and mTICI prediction models. The first, coined GCAM, was created with the gradient-weighted CAM method. It reveals the parts of a certain input scan that were the most influential in making a prediction as a heat-map. The second, coined GWGBP (gradient weighted guided back-propagation), was created using the output of Guided Backpropagation throughout the whole network multiplied pixel-wise by the output of GCAM. GWGBP shows how the network interprets an input scan in terms of the most relevant imaging features utilized for a prediction. We thresholded GCAM heat-maps at 0.5 significance level to facilitate interpretation by highlighting the most contributing areas only.

3. Results

The average and range of AUC values for predicting good functional outcome (mRS $\leq$ 2) and good reperfusion (mTICI $\geq$ 2b) are reported in Table 2. The LR and RFC methods used with the radiological image biomarkers for predicting good functional outcome resulted in an AUC of 0.68 for LR and 0.66 for RFC. For predicting reperfusion, the AUC was 0.52 for both methods. The best average AUC for mRS prediction was obtained using the *RFNN-ResNet* model without AE pre-training (trained from scratch). The best average AUC for mTICI prediction was obtained with *RFNN-ResNet-AE fine-tuned* (with AE initialization and fine-tuning). Note that all models benefit from the AE pre-training for mTICI prediction. However, this is not the case for the prediction of mRS, where

Table 2
AUC using 4-fold cross-validation. Standard *ResNet* and *RFNN-ResNet* trained with scheme (1), *ResNet-AE* and *RFNN-ResNet-AE* with scheme (2) and *ResNet-AE fine-tuned* and *RFNN-ResNet-AE fine-tuned* with scheme (3).

Method	mRS AUC - Avg (range)	mTICI AUC - Avg (range)
LR Baseline	0.68 (0.66 – 0.69)	0.52 (0.51 – 0.54)
RFC Baseline	0.66 (0.64 – 0.69)	0.52 (0.50 – 0.55)
Standard ResNet	0.56 (0.54 – 0.58)	0.51 (0.41 – 0.56)
ResNet-AE	0.58 (0.53 – 0.61)	0.57 (0.55 – 0.58)
ResNet-AE fine-tuned	0.57 (0.51 – 0.66)	0.57 (0.54 – 0.60)
RFNN-ResNet	**0.71 (0.62 – 0.75)**	0.57 (0.55 – 0.59)
RFNN-ResNet-AE	0.65 (0.60 – 0.69)	0.55 (0.53 – 0.57)
RFNN-ResNet-AE fine-tuned	0.67 (0.59 – 0.73)	**0.65 (0.55 – 0.72)**

RFNN-ResNet yielded the best result. Also, there was no difference in AUC between LR and RFC, as shown in previous studies [5]. Most importantly, the best performing data-efficient models outperformed the radiological image biomarkers baseline as well as standard CNN models for both mRS and mTICI outcome predictions. The AUC results for each fold are shown in Supplemental Tables 6 and 7

In Fig. 3 and Fig. 4 we present the model visualization for the best models, *RFNN-ResNet* and *RFNN-ResNet-AE fine-tuned*, respectively. We can observe in the center column (GCAM), that the affected side of the brain (right in this case) contributes the most for the predictions (Fig. 3). Even though the relevant regions are relatively large, the most important regions (depicted in red), are usually more specific. In the right column (GWGBP), we can see that the arteries are highlighted as important features learned by the model in Fig. 3, while in Fig. 4 we observe a noisier pattern. Additional model visualization examples are shown in Supplemental Fig. 1.

4. Discussion

We have shown that data-efficient DL analysis of CTA images outperformed prediction models with commonly used radiological image biomarkers in predicting reperfusion and functional outcome for patients with acute ischemic stroke. With model visualization tools, we have shown that the arteries are amongst the most common and influential features extracted by the DL models when predicting outcomes.

Recent studies on DL learning applied to stroke focused mostly on reproducing radiological image biomarkers and image segmentation tasks. In Ref. [37], SegNet [12] (an encoder-decoder architecture for image segmentation) was applied to MRI stroke imaging focused on predicting tissue outcome after acute ischemic stroke. DeepMedic [38], an open-source 3D CNN, was applied to CTA to detect ischemic stroke and segment lesions with high sensitivity and specificity in Refs. [14, 16], showed that e-ASPECTS, a commercially available artificial intelligence software for ASPECTS scoring, is statistically non-inferior to neuroradiologists in scoring ASPECTS. A method using 3D CNNs was proposed in Ref. [39] for automatic assessment of DWI-ASPECTS with high accuracy. Finally [13], presents a CNN named Stroke U-Net (SUNet) that was developed to segment and predict outcome of acute stroke lesions, and showed better results than 3D U-Net and uResNet.

Our results indicate an improvement by using structured convolutional kernels and unsupervised pre-training to predict good reperfusion (mTICI $\geq$ 2b) when compared to baseline models. The performance improvement derived from the use of Gaussian filters confirms the effect that has already been previously seen in previous medical imaging classification tasks [24]. The best performing method was the *RFNN-ResNet-AE fine-tuned* model (with auto-encoder initialization and fine-tuning after 50 epochs). For good functional outcome prediction (mRS $\leq$ 2), the *RFNN-ResNet* model (trained from scratch) achieved the highest AUC, slightly outperforming the baseline prediction models, with higher AUC scores for three out of four testing folds (Supplemental Tables 6 and 7). Interestingly, for good mRS prediction, the unsupervised pre-training and supervised fine-tuning strategy resulted in a lower AUC when compared to the RFNN-ResNet. This could be caused by the nature of the output labels. mTICI is derived directly from imaging thus more general image features (most effectively learnt by unsupervised learning) appeared to have the potential to be more predictive. One of the strongest predictors of mRS is age [5,18,40], which naturally reflects on many attributes of the brain detectable on CTA, e.g. atherosclerosis, structural abnormalities of the vasculature, changes in white matter and brain volume [41]. Recognition of any of these properties directly or indirectly on images can potentially lead to more confident mRS prediction in contrast to more general image features learnt during unsupervised training. We believe, for mRS, that supervised training from scratch enabled RFNN-ResNet models to grasp such complex features and discriminate subsets of patients better than more general image features learned in an unsupervised fashion. Our experiments suggest that image features learned directly from MIP images using *RFNN-ResNet* models can predict patients with good mTICI and mRS with higher accuracy than prediction models using well-known radiological image biomarkers. However, it should be mentioned that the predictive performance of the models is still limited.

MIP images are either present in organized databases or can be computed quickly and efficiently in seconds, making our method suitable for clinical practice. A prediction from our DL models takes only a few seconds, although the pre-processing steps (registration, skull-stripping and MIP computation) might take up to a couple of minutes (around 2 min for a scan with 400 slices). Consequently, an important advantage of our approach is that it is orders of magnitude faster and it does not require any manual image annotation, even during pre-processing, while delivering comparable prediction accuracy as

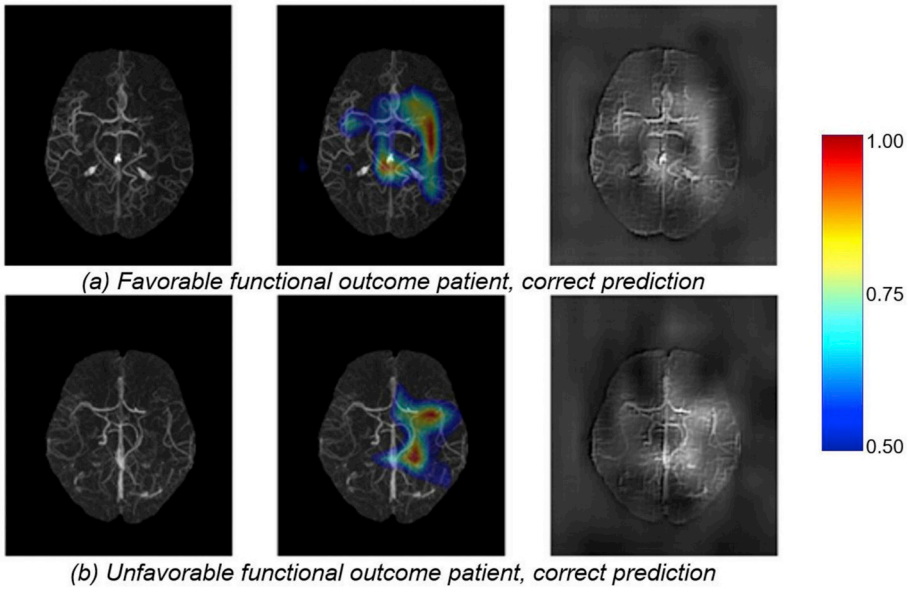

Fig. 3. Visualization of predictions for mRS using the RFNN-ResNet model. Original MIP scans shown in the left column, the GCAM heat-map in overlay in the center and the GWGBP visualization on the right. Colors indicate the level of contribution of each region (GCAM). Most contributing regions (1.0) are represented in red, less contributing (0.5) in blue. *(Should be printed in color).*

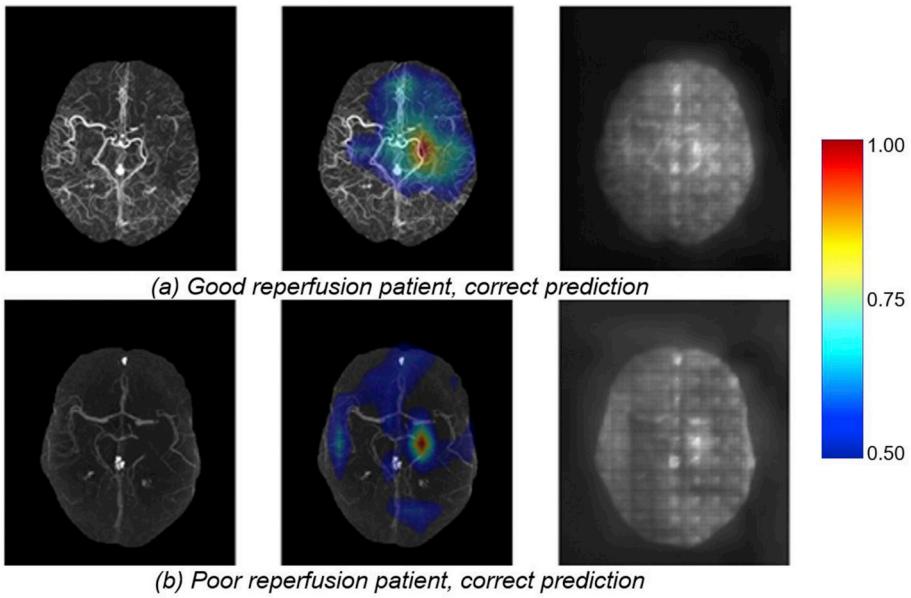

Fig. 4. Visualization of the predictions for mTICI using the RFNN-ResNet-AE fine-tuned model. Order of visualizations corresponds to Fig. 3. (Should be printed in color).

existing radiological imaging biomarkers.

We selected the ResNet architecture based on state-of-the-art performance on natural image classification and refined it for our task. Our main aim was to evaluate the advantages of RFNN kernels over the standard ones. Despite optimizing some hyper-parameters, deeper and wider architectures (such as DenseNets) could potentially yield better results in our experiments and should be explored in the future. CNNs intrinsically contain many parameters that need to be optimized. In determining these parameters, we were restricted by limited computational resources – single GPU –, thus deeper models trained with larger batches or with 3D images might further improve predictions. Also, note that we did not explore transfer learning from another domain (e.g. ImageNet [42]), because the input images would have to be scaled down to a lower resolution. This could lead to the loss of relevant information and the depths of layers would be restricted by the chosen architecture. Another potential limitation is the use of MIPs. Even though a lot of information is lost in using MIPs to represent the 3D images, the MIPs retain important artery structures, while keeping the input size feasible for training the DL models and reducing sparsity.

Even though k-fold cross-validation was applied, validation on an external dataset should be considered in future studies. Furthermore, we opted for a small number of folds for cross-validation due to the limited number of samples and class imbalance, as a high number of folds would lead to a test set with few samples, causing severe variance in our results. Also, due to the high number of experiments and hyper-parameters (from the optimizers and the RFNN), increasing the number of folds would have increased the duration of experiments radically. Besides, 4-fold cross-validation does not provide enough AUC values to compute the statistical significance of differences in accuracy between models. With more data available, more cross-validation folds could be performed to assess if the difference in AUC between models is statistical significant.

Other important clinical factors that are predictive for good mRS and mTICI, such as age, National Institutes of Health Stroke Scale (NIHSS), time from stroke onset to groin puncture, among others, should also be included in future prediction models [5,6]. Finally, we selected the cut-off of ≤ 2 for mRS to make our models comparable to previous mRS prediction modeling research [5,6]. However, other cut-offs should be explored, for instance mRS$\leq$5, where the patients are severely disabled. Besides, provided that more data is available in the future, experiments comparing models trained on the full dataset and models trained on smaller portions could be used to quantify the extent of RFNN improvements over standard ResNets.

Understanding predictions is of utmost importance to improve reliable decision support for individual prospective patients and to further assist in discovering relevant-yet-unseen image features. From our visualizations, one can observe that the highest contribution for good mRS prediction comes mostly from one – the occluded – side of the brain. For predicting reperfusion, however, information from both sides of the brain is taken into account. For good mRS prediction, it is clear from the GWGBP, that arteries – i.e., the extent of blood flow – were important for prediction, since such patterns were extracted in all cases. The important role of arteries is well known in clinical practice [45]. For example, the collateral score (which is highly dependent on the visualization of the arteries in the brain) is commonly used to assess the alternative blood flow and is strongly associated with the size of infarction. The occlusion location is also an important predictor of functional outcome and reperfusion. Regarding mTICI, in cases of poor reperfusion, little is known about the reasons despite a successful recanalization after EVT, though many aspects of the artery have an effect on the efficacy of EVT treatment [4,43,44]. Further study is necessary to evaluate and properly quantify the visual explanations of the networks in more depth, since the most important regions are diverse, arteries are not always at the same location and stroke can occur in various locations of both sides of the brain. Given the important role of arteries and the feature pattern extracted by the networks, we suggest that future research on mTICI and mRS prediction should include the artery pattern as a feature. Also, one could create quantitative measures of interpretability of models, sensitivity and specificity of detection of certain features could greatly facilitate the understanding and improvement of DL models, which can help to identify new relevant image regions and patterns.

5. Conclusion

We have shown that, in our dataset, automated radiological image analysis with data-efficient DL methods outperforms the combination of multiple radiological image biomarkers for good stroke outcome prediction. Our approach does not require image annotation and is faster to compute than any radiological image biomarker considered in this study. We also improved the interpretability of our models using model

visualization tools, which is valuable in clinical practice. Even though DL has shown improvement for outcome prediction, the predictive value is still relatively low and clinical characteristics should be included in future prediction models.

Declaration of competing interest

All authors have contributed to this work. All authors have read and approved the submitted manuscript.

The manuscript has not been submitted nor published elsewhere in whole or in part. All authors have no conflicts of interest.

Acknowledgement

This work was supported by ITea3 grant number 14003 Medolution

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2019.103516.

References

[1] World Health Organization, WHO - the Top 10 Causes of Death, 2018, 24 Maggio.
[2] O. Website, Global Burden of Disease Study 2017 , Country Profile , 2018, pp. 1–7.
[3] S.A. Alqahtani, et al., Endovascular management of stroke patients with large vessel occlusion and minor stroke symptoms, Cureus 9 (6) (2017) 6–11.
[4] M. Goyal, et al., Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials, Lancet 387 (10029) (2016) 1723–1731.
[5] H.J.A. Van Os, et al., Predicting outcome of endovascular treatment for acute ischemic stroke: potential value of machine learning algorithms, Front. Neurol. 9 (SEP) (2018) 1–8.
[6] E. Venema, et al., Selection of patients for intra-arterial treatment for acute ischaemic stroke: development and validation of a clinical decision tool in two randomised trials, BMJ (2017) j1710.
[7] P.A. Barber, A.M. Demchuk, J. Zhang, A.M. Buchan, Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy, Lancet 355 (9216) (2000) 1670–1674.
[8] I.Y.L. Tan, et al., CT angiography clot burden score and collateral score: correlation with clinical and radiologic outcomes in acute middle cerebral artery infarct, Am. J. Neuroradiol. 30 (3) (2009) 525–531.
[9] O.A. Berkhemer, et al., Collateral status on baseline computed tomographic angiography and intra-arterial treatment effect in patients with proximal anterior circulation stroke, Stroke 47 (3) (2016) 768–776.
[10] B.C. Stoel, et al., Automated brain computed tomographic densitometry of early ischemic changes in acute stroke, J. Med. Imaging 2 (1) (2015), 014004.
[11] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[12] V. Badrinarayanan, A. Kendall, R. Cipolla, S. Member, SegNet : a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[13] A. Clèrigues, S. Valverde, J. Bernal, J. Freixenet, A. Oliver, X. Lladó, SUNet: a Deep Learning Architecture for Acute Stroke Lesion Segmentation and Outcome Prediction in Multimodal MRI, 2018.
[14] S. Nagel, et al., e-ASPECTS software is non-inferior to neuroradiologists in applying the ASPECT score to computed tomography scans of acute ischemic stroke patients, Int. J. Stroke 12 (6) (2017) 615–622.
[15] K. Kamnitsas, et al., Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, Med. Image Anal. 36 (2017) 61–78.
[16] O. Oman, T. Makela, E. Salli, S. Savolainen, M. Kangasniemi, 3D convolutional neural networks applied to CT angiography in the detection of acute ischemic stroke, Eur. Radiol. Exp. 3 (1) (2019) 8.
[17] Q. Zhang, S.-C. Zhu, Visual Interpretability for Deep Learning: a Survey, 19 (1423305) (2018) 27–39.
[18] I.G.H. Jansen, M.J.H.L. Mulder, R.J.B. Goldhoorn, Endovascular treatment for acute ischaemic stroke in routine clinical practice: prospective, observational cohort study (MR CLEAN Registry), BMJ 360 (2018).
[19] S. Klein, M. Staring, K. Murphy, M.a. Viergever, J. Pluim, Elastix: a toolbox for intensity-based medical image registration, IEEE Trans. Med. Imaging 29 (1) (2010) 196–205.
[20] J.K. Berge, R.A. Bergman, Variations in size and in symmetry of foramina of the human skull, Clin. Anat. 14 (6) (2001) 406–413.
[21] J.-H. Jacobsen, J. van Gemert, Z. Lou, A.W.M. Smeulders, Structured receptive fields in CNNs, 2016.
[22] T. Lindeberg, Scale-Space Theory in Computer Vision, 1994.
[23] R. Manniesing, M.A. Viergever, W.J. Niessen, Vessel enhancing diffusion. A scale space representation of vessel structures, Med. Image Anal. 10 (6) (2006) 815–825.
[24] R. Zhang, J. Shen, F. Wei, X. Li, A.K. Sangaiah, Medical image classification based on multi-scale non-negative sparse coding, Artif. Intell. Med. 83 (2017) 44–51.
[25] W.S. Oliveira, J.V. Teixeira, T.I. Ren, G.D.C. Cavalcanti, J. Sijbers, Unsupervised retinal vessel segmentation using combined filters, PLoS One 11 (2) (2016) 1–21.
[26] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, S. Bengio, "Why does unsupervised pre-training help deep Learning ? J. Mach. Learn. Res. 11 (2010) 625–660.
[27] B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang, D. Tao, Stacked convolutional denoising auto-encoders for feature representation, IEEE Trans. Cybern. 47 (4) (2017) 1017–1027.
[28] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9908, LNCS, 2016, pp. 630–645.
[29] O. Ozdemir, S. Giray, Z. Arlier, D.F. Baş, Y. Inanc, E. Colak, Predictors of a good outcome after endovascular stroke treatment with stent retrievers 2015 (Iv) (2015).
[30] V. Nambiar, et al., CTA collateral status and response to recanalization in patients with acute ischemic stroke, 2014 no. May 2004.
[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proc IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (2016) 770–778, 2016-Decem.
[32] M.T. Ribeiro, S. Singh, C. Guestrin, "'Why should I trust you?': explaining the predictions of any classifier, 2016, pp. 1135–1144.
[33] M. Cogswell, A. Das, and D. Batra, "Visual explanations from deep networks via gradient-based localization.".
[34] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," pp. 2921–2929.
[35] J.T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for Simplicity: the All Convolutional Net, 2014, pp. 1–14.
[36] A. Chattopadhyay, A. Sarkar, P. Howlader, Grad-cam ++ : improved visual explanations for deep convolutional networks, IEEE Winter Conf. Appl. Comput. Vis. (2018) 839–847.
[37] A. Nielsen, M.B. Hansen, A. Tietze, K. Mouridsen, Prediction of tissue outcome and assessment of treatment effect in acute ischemic stroke using deep learning, Stroke 49 (6) (2018) 1394–1401.
[38] K. Kamnitsas, et al., Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, Med. Image Anal. 36 (2017) 61–78.
[39] O. Öman, T. Mäkelä, E. Salli, S. Savolainen, M. Kangasniemi, 3D convolutional neural networks applied to CT angiography in the detection of acute ischemic stroke, European Radiology Experimental 3 (8) (2019).
[40] M.J.H.L. Mulder, et al., Towards personalised intra-arterial treatment of patients with acute ischaemic stroke: a study protocol for development and validation of a clinical decision aid, BMJ Open 7 (3) (2017), e013699.
[41] A. Gupta, et al., Neuroimaging of cerebrovascular disease in the aging brain, Aging Dis. 3 (5) (2012) 414–425.
[42] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (1–9) (2012).
[43] H. Leischner, et al., Reasons for failed endovascular recanalization attempts in stroke patients, J. Neurointerventional Surg. 11 (5) (2019) 439–442.
[44] E. Qazi, F.S. Al-Ajlan, M. Najm, B.K. Menon, The role of vascular imaging in the initial assessment of patients with acute ischemic stroke, Curr. Neurol. Neurosci. Rep. 16 (4) (2016) 1–9.
[45] U. Jensen-Kondering, Hyperdense artery sign on computed tomography in acute ischemic stroke, World J. Radiol. 2 (9) (2010) 354.