TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY

Citation (APA)
te Nijenhuis, F., van der Sluijs, M., van Doormaal, P. J., van Zwam, W., Hofmeijer, J., Zhang, X., Cornelissen, S., Ruijters, D., Su, R., & van Walsum, T. (2026). Integrating cross-sectional imaging data into functional outcome prediction models for acute ischemic stroke of the anterior circulation. *Neuroscience Informatics*, *6*(1), Article 100260. https://doi.org/10.1016/j.neuri.2026.100260

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

6 (2026) 100260

Contents lists available at ScienceDirect

Neuroscience Informatics

journal homepage: www.elsevier.com/locate/neuri

Original article

Integrating cross-sectional imaging data into functional outcome prediction models for acute ischemic stroke of the anterior circulation

Frank te Nijenhuis [a,*], Matthijs van der Sluijs [a], Pieter Jan van Doormaal [a], Wim van Zwam [b], Jeannette Hofmeijer [c], Xucong Zhang [d], Sandra Cornelissen [a], Danny Ruijters [e,f], Ruisheng Su [a,g], Theo van Walsum [a]

[a] Department of Radiology & Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands
[b] Department of Radiology and Nuclear Medicine, Maastricht UMC+, Maastricht, The Netherlands
[c] Faculty of Science and Technology, University of Twente, Enschede, The Netherlands
[d] Faculty of Electrical Engineering, Mathematics and Computer Science, TU Delft, Delft, The Netherlands
[e] Philips Healthcare, Best, The Netherlands
[f] Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands
[g] Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

ARTICLE INFO

Keywords:
Ischemic stroke
Multimodal imaging
Deep learning
Artificial intelligence
Endovascular procedures
Predictive modelling

ABSTRACT

In acute ischemic stroke, large vessel occlusions of the anterior circulation are increasingly treated with endovascular therapy (EVT). The efficacy of this therapy depends on adequate treatment selection. Treatment decisions can be based on predictions of functional outcome. Most existing studies predict functional outcomes using clinical parameters. We set out to study functional outcome prediction performance by integrating imaging in a multimodal setting. Using a multi-center dataset containing 2927 patients, we compare the functional outcome prediction performances of clinical baseline models, including the clinically validated MR PREDICTS decision tool, image-based models with deep learning networks, and a multimodal approach combining clinical and imaging information. The predicted outcome measure is dichotomized modified Rankin Scale score 90 days after EVT. We perform sanity checks, hyperparameter optimization, and comparisons of effectiveness of using CTA, NCCT, or both images as input. Our experiments show that information extracted from CTA or NCCT images does not significantly improve the performance, as quantified using AUC, of functional outcome prediction methods compared to a baseline model. The multimodal approach may replace radiologically derived biomarkers, as its performance is non-inferior.

1. Introduction

In recent years, mechanical thrombectomy, also referred to as endovascular therapy (EVT), has emerged as an effective procedure for the treatment of acute ischemic stroke (AIS) in patients with a large vessel occlusion (LVO) of the anterior circulation [1–5].

Although recent developments indicate that EVT is feasible and worthwhile in most patients, adequate functional outcome prediction methods remain relevant especially in a delayed treatment setting [6,7]. These methods can provide patients with personalized prognostic information based on their unique clinical and demographic parameters. Accurate predictions may help clinicians tailor treatment and rehabilitation plans to individual needs, optimizing resource allocation, and improving recovery outcomes [8]. Furthermore, investigating functional outcome prediction offers insight into the interaction between stroke imaging features and therapeutic outcomes after EVT.

Multiple scoring methods have been developed to prognosticate functional outcome after EVT, using 90-day modified Rankin Scale (mRS_{90}) as the outcome variable [9]. These methods are based on traditional statistical techniques and are generally not equipped to extract information directly from radiological images. Additionally, they often require radiological image biomarkers, which necessitates an arduous process of expert annotation, complicated by the oftentimes high degree of inter-observer variability [10]. Passing the radiological images

* Corresponding author.
E-mail addresses: f.tenijenhuis@erasmusmc.nl (F. te Nijenhuis), p.vandersluijs@erasmusmc.nl (M. van der Sluijs), p.j.vandoormaal@erasmusmc.nl (P.J. van Doormaal), w.van.zwam@mumc.nl (W. van Zwam), j.hofmeijer@utwente.nl (J. Hofmeijer), xucong.zhang@tudelft.nl (X. Zhang), s.cornelissen@erasmusmc.nl (S. Cornelissen), d.ruijters@tue.nl (D. Ruijters), r.su@tue.nl (R. Su), t.vanwalsum@erasmusmc.nl (T. van Walsum).

https://doi.org/10.1016/j.neuri.2026.100260
Received 7 August 2025; Received in revised form 30 December 2025; Accepted 11 January 2026
Available online 14 January 2026
2772-5286/© 2026 The Authors. Published by Elsevier Masson SAS. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

through an AI based decision support system might circumvent the need for expert annotation. Alternatively, these decision support systems may be of benefit by bypassing the annotation process altogether, to directly predict functional outcome.

We hypothesize that information encoded in baseline imaging, as acquired for stroke patients prior to EVT, can be extracted using deep learning-based methods. A deep learning model trained to effectively predict functional outcome after EVT can be utilized in the clinical decision-making process. The purpose of this study is to investigate whether integration of medical imaging with tabular data improves functional outcome prediction performance compared to more conventional models based on tabular data alone.

1.1. Related work

Several recent studies have investigated the prediction of functional outcome using a combination of clinical features as well as imaging features extracted using deep learning-based methods. Table 1 summarizes previous studies in this area. Zihni et al. used a combined model with a Convolutional Neural Network (CNN) to extract imaging features, as well as a Multilayer Perceptron (MLP) to process clinical data. They compared a frozen weights based multimodal fusion approach to an end-to-end based approach where both arms of the model are trained in parallel. The imaging data consisted of 3D volumes of TOF-MRA images. They showed that an end-to-end multimodal model integrating neuroimaging and clinical data leads to the best performance when predicting dichotomized (0–2 versus 3–6) mRS_{90} [11]. The difference in AUC on the test set, however, while statistically significant, is minor and unlikely to affect clinical practice. Bacchi et al. predicted dichotomized mRS_{90} using several CNN and MLP based models, focusing on a combination of clinical and imaging data. For the imaging data, Non-Contrast CT (NCCT) scans were used. Similar to Zihni et al., the best-performing model was a combination of a CNN and an MLP [12]. The authors show superior dichotomized mRS_{90} prediction performance when combining a CNN and an MLP, compared to Totaled Health Risks in Vascular Events (THRIVE, [13]), Houston IAT (HIAT, [14]), and Stroke Prognostication using Age and NIHSS (SPAN100, [15]). In a comparative study it was shown that these clinical scoring methods have been superseded by MR PREDICTS in terms of discriminative performance [9], and as such the method proposed by Zihni et al. should be compared with MR PREDICTS as well, to achieve a fair comparison. Samak et al. also used NCCT imaging data, in combination with clinical information [16]. They used a custom deep learning architecture incorporating a convolutional encoding branch with squeeze and excitation mechanisms followed by global average pooling and a fusion block, which uses fully connected layers to combine image features and clinical data. Their multimodal method has an AUC of 0.75, outperforming an MLP trained solely on the clinical parameters, which achieved an AUC of 0.70. The proposed model was not compared to externally validated clinical scores. Hilbert et al. [10] and De Graaf [17] both successfully used deep learning models trained on CT Angiography (CTA) images, showing that these images also contain relevant information with regards to functional outcome prediction. In the work of Hilbert et al., a ResNet modified with Structured Receptive Fields outperforms machine learning models that used radiological biomarkers, achieving an AUC of 0.71 on the functional outcome prediction task. In their study, a multimodal outcome prediction approach is not explored, instead the focus is on investigating whether a deep learning model trained on imaging data can replace radiologically derived features. Ramos et al. used atlas-based registration of CTA scans to extract radiomics data [18]. Combining the imaging and clinical features did not significantly improve the predictive performance, with an AUC of 0.81 for the combined approach versus 0.80 for the clinical data only approach. In a similar vein, Jabal et al. [19] extracted predefined imaging features, such as e-ASPECTS score and CSF volume, from NCCT and CTA images and combined these with clinical features using several machine learning classifier methods. They also performed feature selection using a Shapley value-based approach. The selected clinical features were baseline NIHSS, age, occlusion side, and the time interval from symptoms onset to admission. In addition to the clinical features, local M5 infarct volume, local lentiform infarct volume, brain volume, percentage of lateral ventricle volume and collateral vessel deficit volume were selected as imaging derived features, indicating that a combination of imaging features and clinical features leads to the best predictive performance. eXtreme Gradient Boosting (XGB) on a selected subset attained the best performance. Combining all imaging and clinical features did not significantly outperform a model using only clinical features.

Park et al. have adapted the well-known U-Net architecture as a multi-task model, combining stroke lesion segmentation on Diffusion Weighted Imaging with functional outcome prediction. Interestingly, the authors show a correlation between their extracted DWI imaging features and the clinical features, indicating an overlap of information between the MRI scan and the clinical data. The authors note an improvement from the AUC of 0.69 to 0.77 when comparing clinical to combined clinical and imaging based classification models [20].

Jo et al. similarly compared clinical, image-based and combined clinical and image-based analysis models. They perform a multi-stage deep learning approach in which the image-based model outputs a single predicted probability, which is included as an input to the combined model. Their integrated clinical and image-based model outperforms the established THRIVE and HIAT scoring methods, achieving an AUC of 0.78, which was also significantly higher than the image-only or clinical-only models [21].

In recent years different approaches for outcome prediction is ischemic stroke have been tried based on multimodal deep learning techniques. These prior approaches differ in their data integration strategies. End-to-end fusion jointly optimizes imaging and clinical pathways, whereas feature-level fusion combines pre-extracted features from each modality before classification. Both these approaches have been tried in literature. So far, although many works show promise, no publication has convincingly demonstrated improved clinical performance compared to baseline tabular models.

1.2. Contributions

This manuscript is a follow-up and extension of previous work, where we attempted to predict functional outcome using CTA images, tabular data or a combination of both [22]. In this extended work, we investigate a multimodal framework, combining the processing of clinical features with the output of an image analysis backbone, which processes NCCT and CTA images together. We compare the performance of state-of-the-art deep learning-based medical image processing models with clinical baseline models, to investigate whether functional outcome prediction performance can be improved by incorporating images directly into the model. Our work adopts an end-to-end approach, enabling simultaneous learning of complementary representations. We perform hyperparameter optimization and train the models with differing dataset sizes to verify that we have sufficient data. We assess the difference in performance of models trained on only NCCT or CTA images with those trained on both. Training is performed on a dataset containing NCCT and CTA images, as well as tabular features of 2927 patients, which, to the best of our knowledge, is the largest combined CTA and NCCT dataset on which such an effort has been undertaken so far.

2. Methods

In this study, we investigate three models, with multiple backbones, for the prediction of dichotomized mRS_{90}. The mRS_{90} was dichotomized to functional independence ($mRS_{90} \leq 2$) versus functional dependence ($mRS_{90} \geq 3$) as this represents a meaningful threshold for treatment outcome evaluation. Each model takes different input types, but the output is always a prediction of dichotomized mRS_{90}. A schematic overview of these processing models is provided in Fig. 1.

Table 1
Previous work in mRS prediction after 90 days. TOF-MRA: Time Of Flight Magnetic Resonance Angiography, NCCT: Non-Contrast CT, CTA: CT Angiography, DWI: Diffusion Weighted Imaging, ADC: Apparent Diffusion Coefficient. †: Subgroup contains mRS_{90} score of 0–1.

Author	Image Modality	AUC Multimodal	Total Dataset Size	Size of mRS_{90} 0–2 Subgroup (%)
Zihni et al. [11]	TOF-MRA	0.76	313	87 (27.8 %)
Bacchi et al. [12]	NCCT	0.75	204	113 (55.4 %)†
Samak et al. [16]	NCCT	0.75	500	127 (25.4 %)
Hilbert et al. [10]	CTA	0.71	1526	463 (35.6 %)
De Graaf [17]	CTA	0.78	1000	417 (41.7 %)
Ramos et al. [18]	CTA	0.81	3279	1241 (37.8 %)
Jabal et al. [19]	NCCT & CTA	0.84	293	101 (34.4 %)
Park et al. [20]	DWI	0.81	5429	3575 (65.9 %)
Jo et al. [21]	DWI & ADC	0.78	4147	2879 (69.4)

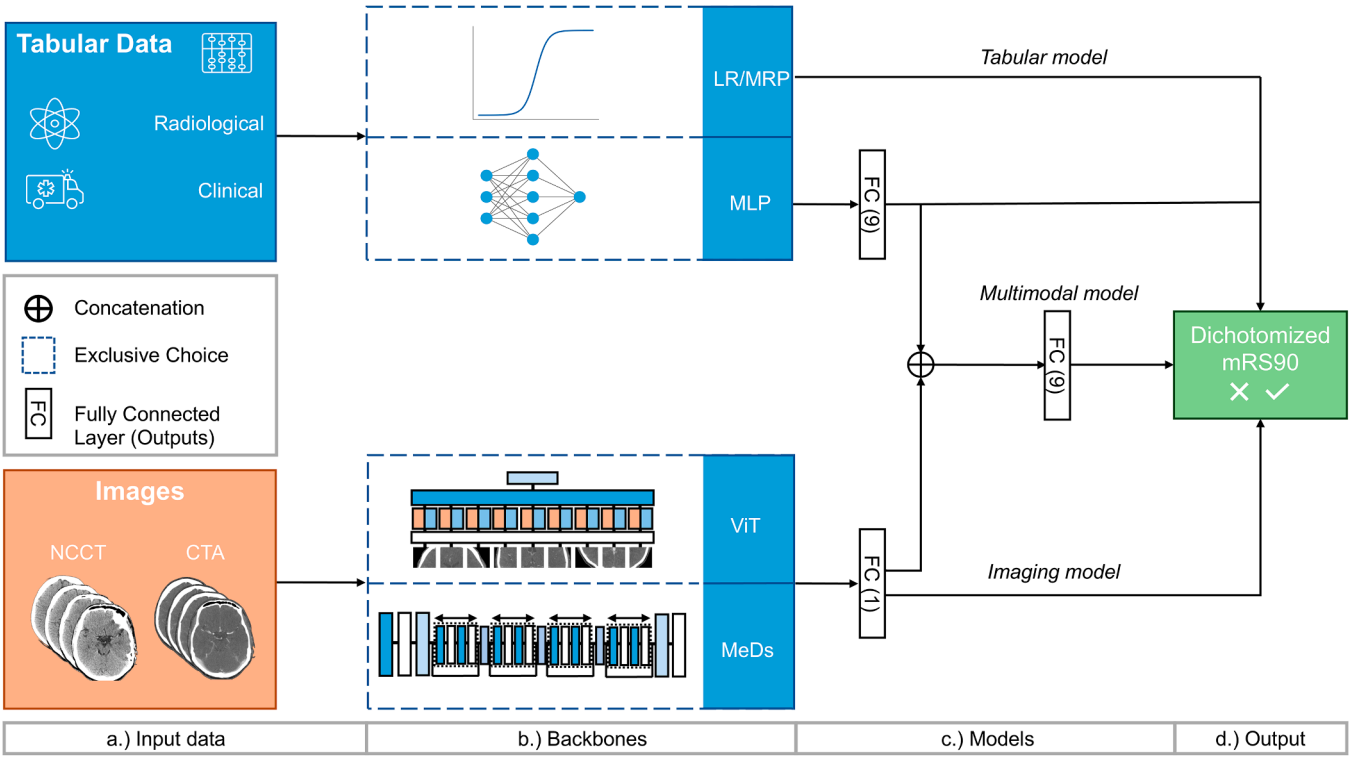

Fig. 1. Schematic overview of the different models. Panel *a.*: The tabular data contains radiological (RD) and clinical (CL) features, defined in Table A.3. The images refer to NCCT and CTA scans of the same patient. Panels *b.* and *c.*: There are three different models using different backbones. The *tabular model* uses tabular data. The *imaging model* exclusively uses an imaging backbone. The *multimodal model* concatenates the outputs from both models. Panel *d.*: Finally, each model predicts dichotomized mRS_{90} as a binary outcome. For the fully connected (FC) layers, the number of output nodes is defined in parentheses. Training and validation were performed using five-fold cross-validation on 90 % of the dataset ($n = 2927$), with 10 % held out for testing. NCCT: Non-Contrast CT, CTA: CT Angiography, ViT: Vision Transformer, MeD: Med3D model, LR: Logistic Regression, MRP: MR PREDICTS model, MLP: Multilayer Perceptron.

The most straightforward model is the *tabular model*, which takes tabular data as input. We subdivide the tabular variables into clinical and radiological features, based on the steps required to obtain these variables. Clinical features can be obtained by examining the electronic health records. In contrast, the derivation of radiological features requires a radiologist to inspect images. Age, baseline NIHSS, pre-stroke mRS, diabetes mellitus, baseline systolic blood pressure, baseline glucose, intravenous alteplase, and time from onset to groin puncture are defined as clinical features. Collateral score, as assessed on CTA, occlusion location (intracranial internal carotid artery, M1 or M2 branch of the middle cerebral artery) and Alberta stroke programme early CT score (ASPECTS) are considered radiological features. Table A.3 contains an overview of these features. In the following, let us introduce a notation for these feature sets. CL denotes the clinical features, RD denotes the radiological features, and RD + CL denotes the combined radiological and clinical features, also referred to as tabular features.

The tabular data consists of the predictors used by the MR PREDICTS functional outcome prediction model [23], displayed in Table A.3. In the *tabular model*, the tabular data are passed through a Multilayer Perceptron (MLP) classifier, the MR PREDICTS outcome prediction model (MRP) [23], or a Logistic Regression (LR) model. Note that MRP is a specific instance of a Logistic Regression model, using the weights from the original MR PREDICTS manuscript.

The *imaging model* takes preprocessed CTA, NCCT, or both images as input, depending on configuration. Again, we introduce notation for this. Let NC + CA denote the full image dataset (NCCT and CTA), whereas CA will refer to only the CTA images, and NC will indicate using only the NCCT images. The backbone of this model is either a modified Med3D (MeD) deep learning model [24], or a Vision Transformer (ViT) [25].

The third framework we consider is the *multimodal model*, which concatenates the output of the *clinical* model with the output of the

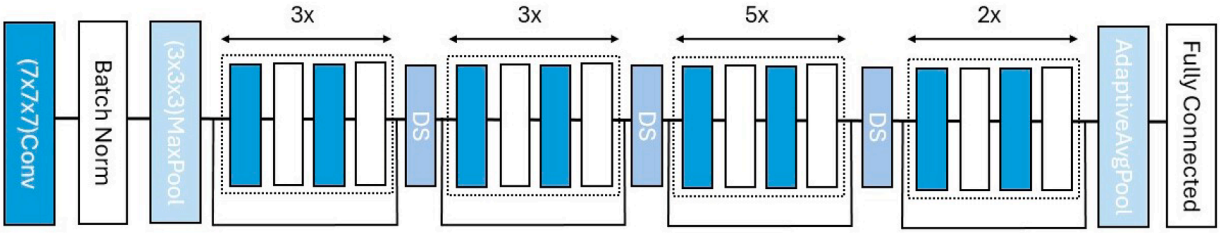

Fig. 2. Schematic overview of the modified Med3D architecture with ResNet50 backbone. The dark blue blocks represent three-dimensional convolutional layers. White blocks represent batch normalization. The light blue block represents a Max Pooling operation. After multiple repetitions of these blocks, an Adaptive Average Pooling layer is applied, followed by a final Fully Connected layer. Note that the blocks are repeating a different number of times, with residual connections encompassing each block. DS: Downsampling module, which is a convolutional layer with an increased stride of 2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

imaging model and concatenates the combined output into a fully connected layer, allowing the output of both models to be combined in a multimodal fashion. In our notation, $\text{MeD}_{\text{NC+CA,CL}}$ refers to the MeD model trained on images and clinical but not radiological features. Note that ViT and MeD models always require a subscript NC + CA, CA or NC as they are inherently image processing architectures, whereas MLP, LR and MRP will never process images.

2.1. Tabular data models

One of the three models evaluated in the *tabular model* is the Multilayer Perceptron. The model contains four fully connected layers. The number of nodes per layer has been determined using hyperparameter optimization. The first layer takes the tabular features as input, so either RD, CL or RD + CL. The subsequent layers have $18, 9, 10, 9$ outputs, respectively. For each Fully Connected layer, the LeakyReLU activation function is used to mitigate potential issues related to vanishing gradients [26], which might occur in the multimodal approach. In the *tabular model*, the output from the final layer is mapped to a single node with sigmoidal activation function, which predicts dichotomized mRS_{90}. In the *multimodal model*, the results are concatenated with the output of the imaging backbones, which consist of either MeD or ViT.

The other two tabular models we investigated were Logistic Regression (LR) and MR PREDICTS (MRP). For MRP, we take the weights as defined in the paper, and as such, we can only use this model when comparing to RD + CL, the full tabular set of features [23]. For the LR model, we fit a logistic regression on either the CL, RD or RD + CL feature sets.

2.2. Imaging models

MedicalNet, also referred to as Med3D (MeD) is a residual CNN specifically designed for medical image analysis. It is intended for segmentation purposes, but here we reconfigure it as a classification network. It consists of a modified ResNet backbone with an upsampling branch. The ResNet backbone is modified by changing the number of input channels from three to either two or one, depending on whether we input a single image type (NCCT or CTA) or both images. The combination of imaging modalities through multiple channels allows for mixing of the different imaging modalities. It is further modified by expanding the 2D convolution operations to 3D convolutions, to handle the volumetric CTA or NCCT inputs. Additionally, the stride in layers three and four of the network is set to one, to prevent downsampling in the first block. Finally, dilated convolutions are used in the downstream convolutional layers to decrease the computational complexity. We employ a transfer learning approach by initializing the Med3D architecture using the weights that were stored after training the modified architecture on the 3D segmentation dataset, as described in [24]. We further modify the Med3D network by replacing the final segmentation layer with an average pooling layer, followed by a linear layer mapping to 64 features. As part of the *imaging model*, these 64 features are then mapped to a single output node with a sigmoidal activation. In the *multimodal model*, the imaging features are concatenated with the final layer of the MLP, which contains nine features, which are then altogether mapped to a single output node with sigmoidal activation function. In both cases, the output prediction is dichotomized mRS_{90}. Fig. 2 provides a schematic overview of the MeD model.

The Vision Transformer (ViT) is a modification of the Transformer natural language processing model, such that it can be used to handle visual tasks [25]. While the original ViT is described as a model that handles 2D input images, a three-dimensional extension of the ViT model is provided by the MONAI consortium [27]. The ViT relies on the self-attention mechanism as an alternative to the convolutional layer. An important benefit of the ViT model is that it requires less computational resources to train compared to conventional CNNs. Because the ViT employs the self-attention mechanism instead of convolutional layers, it lacks the inductive biases commonly seen with CNNs, and as such it requires a relatively large training dataset to attain satisfactory performance [25]. Transformers are sequence-to-sequence models, but they can be modified for classification tasks. To this end, a learnable classification token is added at the beginning of the patch embeddings sequence. The final hidden representation of this token is fed to an MLP classification head. In the *imaging model*, the single output from the ViT is again passed to a single node with sigmoid activation function to mimic the structure of the MeD. For the *multimodal model*, the output of the ViT is concatenated with the final 9 features of the MLP, which are then altogether mapped to the sigmoidal output node to predict dichotomized mRS_{90}. See Fig. 3 for a schematic overview of the ViT model.

2.3. Multimodal models

The *multimodal model* fuses tabular inputs with the imaging input. This is done by concatenating the output of an imaging backbone model (ViT or MeD) with an MLP layer. An additional layer containing 9 features takes the concatenated imaging and tabular features and maps them to a single dichotomized mRS score. This allows for end-to-end training of both the imaging and the tabular pathways in a combined manner.

3. Data

We use imaging and tabular data from the MR CLEAN Registry, which is a prospective observational study involving seventeen centers in the Netherlands. Data collection started in March 2014 and was finished in 2018. These Registry data represent the standard of care for stroke patients in the Netherlands. Patients over 18 years of age undergoing arterial puncture with intention to perform EVT for AIS were included in the Registry. Patients were excluded if imaging showed signs of intracranial hemorrhage. Intervention consisted of arterial puncture and catheterization with or without thrombus removal, with or without administration of intra-arterial thrombolytics. The follow-up period was three months. The data were acquired in a secondary care setting. Sociodemographic features were not explicitly registered in the Registry,

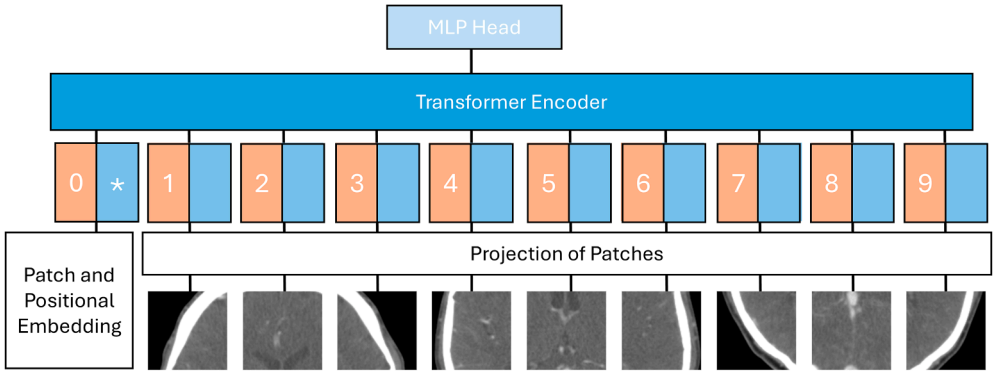

Fig. 3. Schematic overview of the Vision Transformer (ViT) architecture. Three-dimensional patches from the CT images are linearly embedded, and positional embeddings are added so the model can keep track of the patches in the sequence. The projected patches are passed as embedded vectors through the encoder module. The first token in the sequence is passed to a Multilayer Perceptron, after aggregrating information from the other patches through the encoder. The illustration is inspired by the original ViT manuscript by Dosovitskiy et al., [25].

which reflects the population of Dutch stroke patients eligible for EVT. The primary outcome of the Registry is mRS_{90}, which was assessed by research nurses, either in-person or through a telephonic interview. Outcome assessors were not blinded to the baseline characteristics of the patients. The study protocol for the MR CLEAN Registry was approved by the ethics committee of the Erasmus MC University Medical Centre, Rotterdam, the Netherlands (MEC-2014-235) [28].

The MR CLEAN Registry data is used for training as well as internal testing. For each patient the dataset contains demographic information, information about clinical parameters and patient outcomes. The patient features we use as predictors are derived from the MR PREDICTS clinical decision tool, displayed in Table A.3. The accompanying preoperative NCCT and CTA scans are also available for each included patient, as well as the mRS_{90} score.

Initially, we obtained 3577 patients from the MR CLEAN Registry, with mRS_{90} score, NCCT and CTA scans available. An illustration of the preprocessing model is provided in Fig. A.7. We started the preprocessing process by registering each scan to a brain atlas ([29]) using the ANTs software [30].

After registration, the quality of the scans was assessed using visual inspection. Images were rejected (n=537) if NifTI conversion failed (n=14), if there were insufficient image slices in CTA or NCCT (n=176), if there were artifacts (n=151), if brain mask registration (n=146) or clipping the image failed (n=44) due to insufficient brain coverage, and if there were relevant DICOM tags missing (n=34). From the remaining images (n=3040), some were discarded due to limited image quality of the final registered scans in either the NCCT or CTA (n=113). Images were then normalized, from the standard range of Hounsfield units, between -1000 and 3000, to between zero and one. After registration and normalization, we stripped the skull from the images by using the brain mask. We then conditionally mirrored the images along the sagittal plane, such that the affected side is always on the patient's left side. This was done so that we can crop the image to improve computational and memory efficiency. The final image has a width and a depth of $[112, 80]$, respectively, which is cropped from the original image size of $[256, 256]$. The height of the original images is variable depending on the number of slices included in each scan as well as the slice spacing. On NCCT, the median (IQR) slice thickness is 3 mm $(1 - 5)$. On CTA, the median (IQR) slice thickness is 1 mm $(0.75 - 1.25)$. After atlas registration, 80 slices were selected in a standard range to ensure optimal brain coverage. In total, 2927 cases were left for our experiments. Ten percent of the data (n=292) was used as a held-out test set, and the remaining data was used for five-fold cross-validation, such that each training and validation set contain n=2050 and n=585 cases, respectively. Clinical data were normalized between zero and one across the entire dataset. Of the independent variables, at most 2 % of the data was missing, which was assumed to be missing at random, meaning the probability of missingness depends only on the observed data and not on the missing values themselves. Missing data were imputed for all independent variables using Multiple Imputation by Chained Equations (MICE, [31]) with a Gaussian Mixture estimator. We obtain AUC values and binary accuracy scores on each validation dataset. For each model, the fold with the best validation performance in terms of AUC was selected for subsequent statistical analysis. For each of the best-validated classifiers, we obtained a Receiver Operating Characteristic (ROC) curve on the held-out test data. We compared these curves using DeLong's test [32] to quantify the difference in performance.

4. Implementation

All experiments were conducted on a single GPU node on the Snellius Dutch national supercomputer [33]. The compute node is equipped with a 72-core Intel Xeon Platinum 8360Y CPU, 512 GB of RAM, and an Nvidia A100 GPU. All software was written using Python [34]. Machine learning code was written using the Pytorch framework [35]. Registering and logging the performance of the deep learning models was done using the MLFlow framework [36]. Each model was trained for 100 epochs, with an initial learning rate of $1e^{-1}$. An exponential learning rate scheduler was used, with $\gamma = 0.91$ such that the learning rate after 100 epochs is $\sim 1e^{-}5$. Early stopping was employed, such that training stopped when no noticeable decrease in validation loss was detected after 50 consecutive epochs. Batch size was set to 16 for the imaging-based models due to limitations on the available GPU memory, and 40 for the clinical-only models. For the images, data augmentation was applied consistently across both NCCT and CTA modalities, with a probability of 0.5. The augmentation consists of random rotations and translations. Rotations were applied along each axis with a range $\in [-15°, 15°]$. Translations were applied at different settings per axis, such that the affected area would always remain visible. The translation settings were x-direction $\in [-5, 5]$ voxels, y-direction $\in [-15, 6]$ voxels and z-direction $\in [-5, 5]$ voxels. The translations were selected to be in these ranges by visually assessing their effects along each axis, ensuring that no cerebral structures are cut out of the 3D image. Binary cross entropy was used as a loss function. This loss is calculated using the raw output logits. A sigmoid operation was applied to the logits to obtain output probabilities. For optimization, the AdamW optimizer was used, with $\lambda = 1e^{-2}$, $\beta_1 = 0.9$ and $\beta_2 = 0.99$ [37]. Losses, AUC scores, and binary accuracies were registered for the training and validation set during every epoch using the MLFlow framework [36].

5. Experiments and results

We performed five experiments:

1. Hyperparameter Tuning.

2. Functional Outcome Prediction.
3. Sex Prediction Sanity Check.
4. Impact of Dataset Size.
5. Ablation Study of Image Modalities.

Each experiment is described below in detail, combined with its results.

5.1. Hyperparameter tuning

Hyperparameter tuning was performed on the full dataset excluding test-set, with the test-set again being used for evaluation. We used the *Optuna* framework [38], with a Gaussian process based sampling algorithm. For the MeD model we searched for the optimum number of ResNet layers, $l \in \{10, 18, 34, 50, 101\}$ by comparing overall mRS_{90} prediction performance in the *imaging model* using different backbones.

For the MLP model, we optimized the number of hidden layers, $h \in \{1 \dots 6\}$ as well as the number of hidden units per layer, $u \in \{4 \dots 128\}$, and dropout $d \in [0.1, 0.5]$. For the AdamW optimizer we tuned the weight decay, $\lambda \in [0.0001, 0.1]$, the learning rate $\gamma \in [0.0001, 0.01]$ and $(\beta_1, \beta_2) \in [(0.5, 0.9), (0.99, 0.9999)]$ during training of the MeD backbone in the *imaging model*. We also investigated the effect of a different learning rate in the setting of ViT backbone optimization for the *imaging model*.

Hyperparameter tuning experiments empirically provided a configuration of $l = 50$ for $MeD_{NC+CA,RD+CL}$, which we also adopted for the other models using MeD as backbone (i.e. $MeD_{NC+CA,CL}$ and MeD_{NC+CA}). Similarly, for the MLP model we found $h = 3$ and $u = 20$ to be optimal. For the AdamW optimizer settings, we obtained $\gamma = 0.0001$, $\lambda = 0.001$ and $(\beta_1, \beta_2) \in [(0.9, 0.9999)]$, which we applied for all subsequent experiments. See Table A.4 for an overview of all hyperparameter settings.

5.2. Functional outcome prediction

To assess the added value of a multimodal pipeline, we compare the *multimodal model* with both the *imaging model* and the *tabular model*. For each variation of each model, average 5-fold cross-validation performance and held-out test set performance is shown in Table 2. The model corresponding to the fold with the best validation performance was evaluated using the test set. DeLong's test for comparison of ROC curves on the test set reveals that there is no statistically significant difference between an MLP trained on combined tabular features (MLP_{RD+CL} AUC=0.84), which we consider to be a baseline model, and the multimodal ViT ($ViT_{NC+CA,RD+CL}$ AUC=0.83, $p = 0.671$) or MeD models ($MeD_{NC+CA,RD+CL}$ AUC=0.83, $p = 0.158$) trained on combined clinical features and images.

All models trained on combined clinical features perform significantly better than the MLP trained on radiological features alone (MLP_{RD} AUC=0.69, $p<0.0001$ for $MeD_{NC+CA,RD+CL}$, $ViT_{NC+CA,RD+CL}$ and MLP_{RD+CL}).

The MLP_{CL}, with AUC=0.80, also performs significantly better than MLP_{RD} ($p = 0.003$), but significantly worse than the multimodal models ($p<0.0001$ for $MeD_{NC+CA,RD+CL}$, $ViT_{NC+CA,RD+CL}$ and MLP_{RD+CL}).

MLP_{RD}, with AUC=0.69, performs similarly to MeD_{NC+CA}, (AUC=0.69, $p = 0.915$), and both perform significantly worse than the aforementioned models. ViT_{NC+CA} AUC=0.56, significantly worse than MLP_{RD} and MeD_{NC+CA}, with $p < 0.0001$ for both comparisons.

5.3. Sex prediction sanity check

Using ViT_{NC+CA} and MeD_{NC+CA}, we investigated sex-prediction performance. We trained on the entire dataset excluding test-set and evaluated once on the test-set, to save computational resources. We trained both models for 100 epochs, with early stopping and the AdamW optimizer. We find sex prediction AUCs of 0.73 and 0.75, for ViT_{NC+CA} and MeD_{NC+CA} respectively.

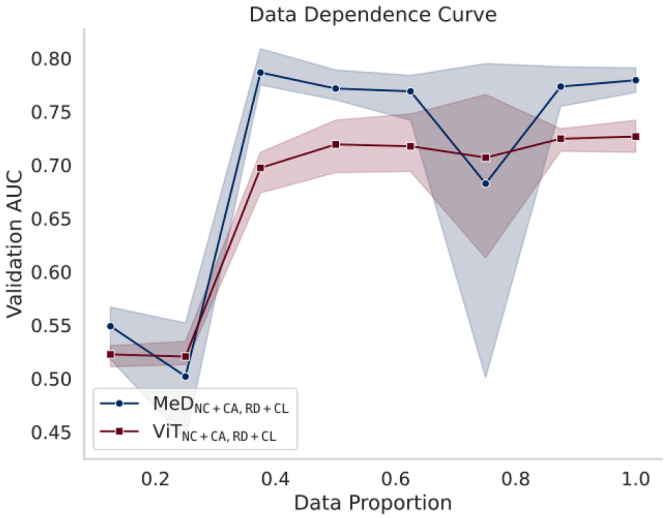

Fig. 4. Model performance (expressed in AUC on the validation set) as a function of training dataset size. For the $MeD_{NC+CA,RD+CL}$ model, performance no longer substantially improves after 37.5 % ($n = 769$) of the data is included. For the $ViT_{NC+CA,RD+CL}$ model, we see a similar plateau, which starts when 50 % ($n = 1026$) of the data is used in training.

5.4. Impact of dataset size

To verify that we have a sufficient amount of data, we trained the imaging models on different dataset sizes, going from 12.5 % to 100 % in increments of 12.5 %. sizes. A performance plateau when increasing the training set size is an indication that there is enough data available for the models. Dataset size experiments for the $MeD_{NC+CA,RD+CL}$ model show a plateau of the validation AUC value when 37.5 % (n=768) of the total training dataset is used, with a maximum average validation AUC of 0.77. For the ViT backbone, we see a similar pattern, where the average validation AUC attains a similar plateau around the 37.5 % mark. The maximum value average validation AUC for the ViT model is 0.73 (Fig. 4).

5.5. Ablation study of image modalities

We performed functional outcome prediction experiments comparing the performance of the imaging model when trained on NCCT and CTA in a combined fashion versus the models trained on NCCT or CTA individually. We did this to investigate whether there is a synergistic effect of combining NCCT and CTA imaging versus using either modality alone Figs. 5 and A.6.

The MeD_{NC+CA} model, with a test-set AUC=0.60, shows similar performance to the MeD_{CA}, where AUC=0.62, with (p=0.233), whereas the MeD_{NC}, with AUC=0.56 is significantly worse, with (p=0.034). We observe a similar effect when we compare the ViT_{NC+CA} model, with AUC=0.59 to the ViT_{CA} model, with AUC=0.56, $p = 0.313$, whereas ViT_{NC} again performs significantly worse with AUC=0.53, $p = 0.04$.

6. Discussion

We investigated whether adding NCCT and CTA imaging information directly in a multimodal model improves functional outcome prediction performance after acute ischemic stroke, when compared to using only tabular data, which is standard practice in the field of functional outcome prediction. We compared multiple deep learning frameworks and included conventional statistical approaches.

We found that the addition of NCCT and CTA imaging in a multimodal end-to-end fashion does not lead to statistically significant improvements in dichotomized functional outcome prediction

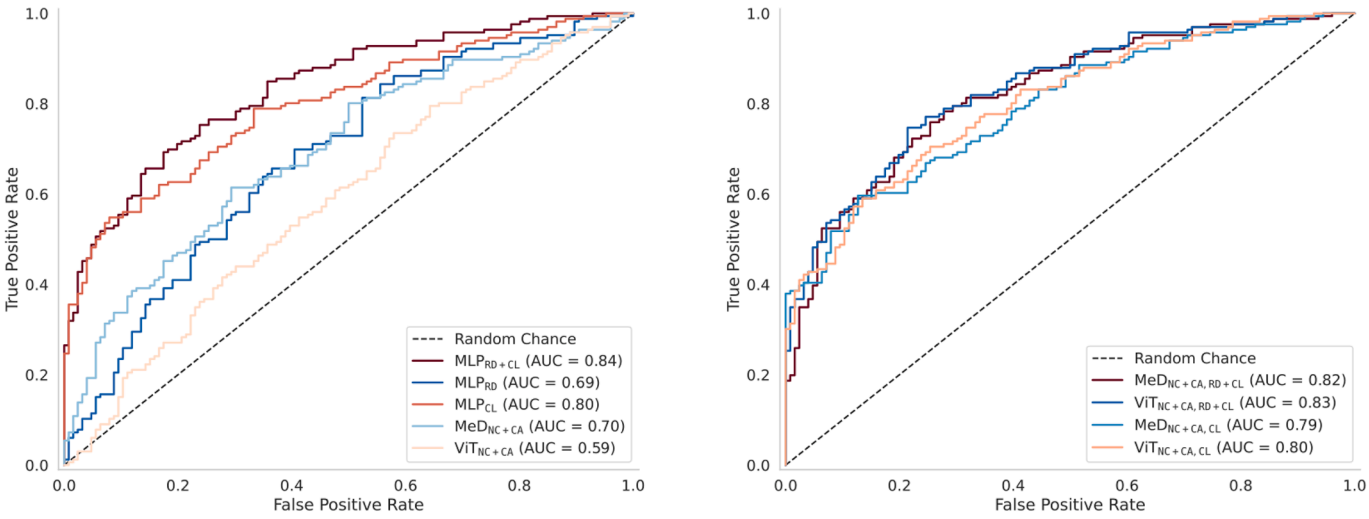

Fig. 5. ROC curves of the best validated models evaluated on the test set. On the left are the *imaging* or *tabular* only models, and on the right we see the performances of the *multimodal* model category. For the image or tabular only models, the MLP_{RD+CL} model performs best with an AUC of 0.84, whereas for the combined models, $ViT_{NC+CA,RD+CL}$ model performs best with an AUC of 0.83. This implies that there is no added benefit of including imaging features.

Table 2
Performance of each classifier expressed in terms of Area Under the receiver operating characteristic Curve (AUC) and binary accuracy (ACC). *Val* indicates 5-fold cross-validation performance, with standard deviation in parentheses. *Test* indicates performance on the test set. The top table shows results for tabular features: clinical (CL), radiological (RD), and combined (RD + CL). The bottom table contains results for imaging features (NC + CA) combined with clinical and radiological features.

		CL		RD		RD + CL	
Model	**Metric**	Val-	Test	Val-	Test	Val-	Test
MRP	AUC	-	-	-	-	0.70 (±0.01)	0.71
	ACC	-	-	-	-	0.41 (±0.01)	0.33
LR	AUC	0.70 (±0.01)	0.73	0.56 (±0.01)	0.60	0.79 (±0.01)	0.84
	ACC	0.72 (±0.02)	0.74	0.61 (±0.01)	0.64	0.74 (±0.02)	0.80
MLP	AUC	0.78 (±0.02)	0.79	0.62 (±0.06)	0.69	0.78 (±0.01)	0.83
	ACC	0.71 (±0.02)	0.71	0.60 (±0.04)	0.67	0.69 (±0.02)	0.74
		NC + CA, CL		NC + CA		NC + CA, RD + CL	
Model	**Metric**	Val-	Test	Val-	Test	Val-	Test
MeD	AUC	0.60 (±0.10)	0.70	0.60 (±0.10)	0.70	0.78 (±0.03)	0.82
	ACC	0.59 (±0.04)	0.65	0.59 (±0.04)	0.65	0.72 (±0.02)	0.75
ViT	AUC	0.77 (±0.02)	0.75	0.57 (±0.01)	0.59	0.76 (±0.01)	0.83
	ACC	0.72 (±0.03)	0.72	0.51 (±0.04)	0.52	0.70 (±0.01)	0.75

performance when comparing to models based on only tabular (clinical and radiological) patient data.

These findings are comparable to related work, where similar deep learning methods are applied on different imaging modalities to predict dichotomized mRS_{90}. Studies performing outcome prediction using only information available prior to endovascular therapy demonstrate an AUC in the range of 0.71 - 0.84, [10–12,16,17,19–21,45]. These studies integrate imaging either through an end-to-end deep learning pipeline or through an intermediate feature extraction step. Various imaging modalities are used, including CTA, NCCT and MRI sequences. The size of datasets that were used also varies from 204 to 5429 patients included, see Table 1. The fact that multiple authors employing different deep learning methods on distinct imaging modalities, have not been able to convincingly demonstrate significant improvement over conventional methods suggests that current functional outcome prediction models do not benefit from multimodal data integration.

When comparing the performance of the models trained using CTA or NCCT images only with the combined pipeline, we see that the performance for models trained using only CTA is non-inferior to models trained with CTA and NCCT combined. This implies that it is possible to use only the CTA images and achieve the same performance as when using both modalities combined with clinical features. The fact that $MeD_{CA,CL}$ attains similar performance compared to the full tabular models indicates that CTA images can replace the radiologically derived ASPECTS and collateral features, alleviating the burden for the radiologist, who manually derives these scores in clinical practice.

The fact that we are unable to demonstrate any substantial benefit when adding an image processing model to the tabular data model is most likely due to the lack of complementary information in the images compared to the tabular data. There are two potential explanations for these findings.

Firstly, dichotomized mRS_{90} is a rather noisy outcome measure, and as such, the irreducible error is potentially high. Even with perfect information at the time of the event, the outcome is unpredictable due to the 90-day waiting period between EVT and the measurement of the outcome, as well as the course of the procedure itself. We note that authors who incorporate post-EVT variables, such as Liu et al., reach an AUC of 0.90 [44]. This indicates that the outcome of the EVT procedure is highly predictive of the long-term outcome.

Secondly, it is possible that the information encoded in the images is to a large extent overlapping with the most salient clinical features. In a feature importance investigation, Ramos et al. show that age, NIHSS at baseline and pre-stroke mRS are the three most relevant predictive features among 50 variables [18]. We hypothesize that deep learning models learn to assess this information in the images, although this is challenging to verify due to the black-box nature of these methods. This would explain why the *imaging model* is able to predict dichotomized functional outcome moderately well (best AUC=0.60 for the MeD model), while there is no additive effect when adding clinical and radiological features to the imaging models. The added value of imaging may be limited in this context due to redundancy of the clinical and imaging features, and the relatively low image contrast for certain prognostic markers. For instance, it may be challenging for the model to derive age from the NCCT or CTA scan due to the inherent variability in atrophy levels and other age related imaging biomarkers.

We performed hyperparameter optimization for the MeD and ViT models in both the *imaging model*, as well as the *multimodal model* settings. Due to the number of models involved in the experiments, exhaustively investigating hyperparameter settings is computationally prohibitive. Instead, we attempted to identify sensible hyperparameters to optimize and to specify sensible ranges for these parameters.

Nevertheless it is possible that we are introducing bias by tuning the *multimodal model* models more than the tabular feature based MLP models. The fact that the performance of the *multimodal model* still does not significantly surpass the tabular models shows that this potential bias is practically not an issue.

To verify that our deep learning pipelines were working and that the cross-sectional images contain predictive information, we performed a sex prediction task using the CTA and NCCT images as inputs. The performance, with an AUC of 0.73, seems relatively low for this simple task. Most models trained for sex prediction on head CT are focused on cranial features, which are removed during our brain-masking stage [39–43]. This might explain the lower performance of our methods compared to other approaches, which are looking at the shape of the brain without access to skull features. Nevertheless, these results indicate that the models are able to extract useful information for a subsequent classification task directly from the images.

Deep learning for medical image analysis can often be a challenging field due to the relative data-scarcity. Indeed, the training datasets used in previous work are often relatively small, particularly when multiple imaging modalities are involved. Jo et al. are an exception, analyzing DWI and ADC MRI sequences of a total of 4147 patients, achieving an AUC of 0.78. The largest combined investigation of NCCT and CTA images, by Jabal et al. contains 293 patients, and an AUC of 0.84. In our dataset size impact experiment we saw that the performance reaches a plateau once approximately 50 % ($n = 1026$) of the total size of the training dataset is used for the ViT model, while MeD attains a slightly better performance earlier. This indicates that adding more training data is unlikely to increase the performance and that the models are approaching an irreducible error limit. When comparing the *imaging model* trained with both NCCT and CTA with models trained with either modality separately, we find similar performance for the CTA and full models in both the MeD and ViT case, whereas the performance of models trained on NCCT is again significantly worse, just as in the combined case. This indicates that most prognostic information on stroke outcome is found in the CTA. Models trained on NCCT do not seem to contain additional information, as we did not observe a synergistic, complementary effect when combining both modalities.

A limitation of this work is that we only used data from the MR CLEAN Registry. The dataset reflects the Dutch stroke population, but the models lack validation on international data and may be affected by temporal distribution shifts as stroke treatment continues to develop.

In this study, we investigated the applicability of neural networks when using NCCT and CTA images combined with tabular data in functional outcome prediction for stroke patients undergoing EVT. One potential new avenue of research would be to investigate whether the addition of CT-perfusion or MRI (DWI) information, for instance as a different input channel to the classifier, would further improve results.

The results also show that the $\text{MeD}_{\text{CA,CL}}$ model may be used as a replacement for the radiologically derived parameters, and could therefore be used to implement a functional outcome prediction model in clinical practice without increasing the radiological workload. Future work should aim to evaluate and improve the clinical applicability of this method.

Numerous approaches have been proposed with varying degrees of reported performance. Future work should focus on comparing the different proposed methods to see which methods truly outperform the baseline models when compared on an external validation set. While there is no statistically significant benefit of combining images with tabular data, there is a potential use case for automated biomarker extraction. Future work should additionally investigate how automated biomarker extraction can be used to facilitate the inclusion of outcome prediction models directly in the stroke care pathway.

7. Conclusion

We have compared multimodal deep learning techniques on the largest combined NCCT and CTA dataset of stroke patients so far. Our findings show non-inferiority of multimodal deep learning in combination with clinical features to the well-validated MR PREDICTS logistic regression model, with clinical and radiological features. While multimodal deep learning cannot be used for fully automated functional outcome prediction, there is a role for it in replacing radiologically derived features, potentially speeding up the outcome prediction process.

Ethics approval

The MR CLEAN Registry was approved by the Ethics Committee of the Erasmus MC, Rotterdam, The Netherlands (MEC-2014-235). The need for individual patient consent has been waived.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Microsoft Copilot in order to assist with syntax editing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Source of funding

This publication is part of the project ROBUST: Trustworthy AI-based Systems for Sustainable Growth with project number KICH3.LTP.20.006, which is (partly) financed by the Dutch Research Council (NWO), Philips Medical Systems Nederland B.V., and the Dutch Ministry of Economic Affairs and Climate Policy (EZK) under the program LTP KIC 2020-2023.

Data availability

Data are available upon reasonable request. Please contact `dawc.contrast@contrast-consortium.nl`.

Declaration of competing interest

We have no Conflicts of Interest to declare.

Acknowledgements

We thank the MR CLEAN Registry investigators for their contribution. A list of all investigators is given in the supplementary material. This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-10609.

Appendix A

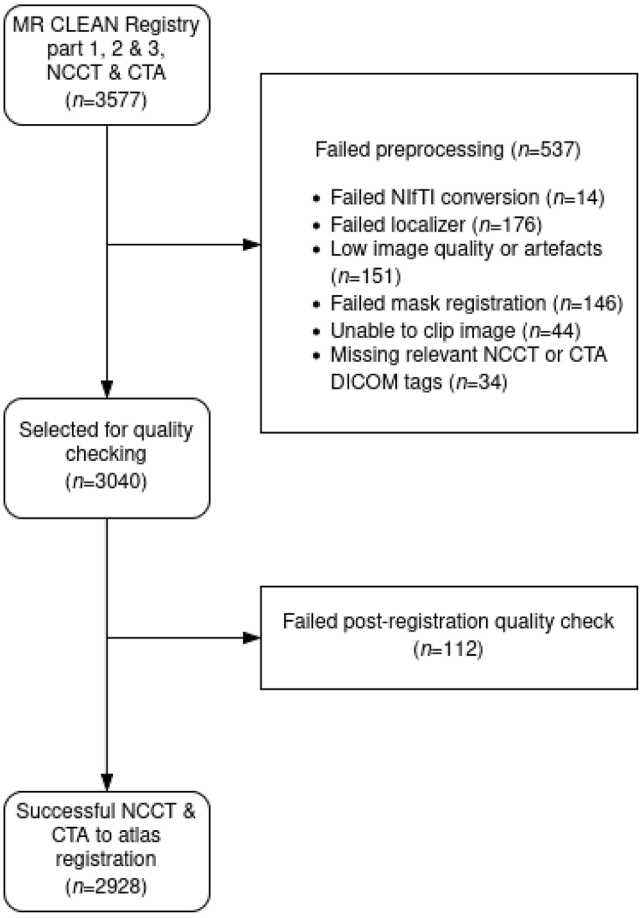

Fig. A.6. Data inclusion flowchart.

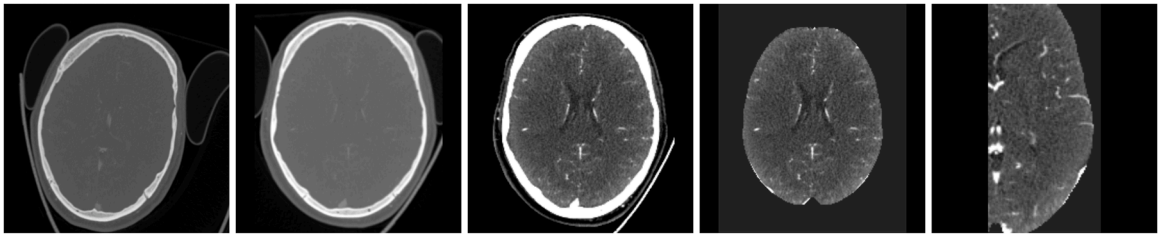

Fig. A.7. Illustration of steps in the image preprocessing model. A: input CTA. B: registration to atlas. C: normalization of Hounsfield Units to [0, 1]. D: skull stripping. E: affected hemisphere cropping.

Table A.3
Baseline characteristics of the unimputed, unnormalized clinical data. Glucose in mmol/l, BP: Blood Pressure in mmHg. ICA: internal carotid artery. M1-3: branches of the middle cerebral artery. †: radiologically derived features.

Feature	MR CLEAN Registry 1,2&3 (n = 2927)	MR CLEAN Registry mRS_{90} 0-2 (n = 1261, 43.09 %)	MR CLEAN Registry mRS_{90} 3-6 (n = 1666, 56.91 %)
Median Age (IQR)	72 (62-80)	68 (57-75)	76 (66-84)
Male % (n)	52.27 % (1530)	57.57 % (726)	48.25 % (804)
Median ASPECTS (IQR)†	9 (8-10)	9 (8-10)	9 (7-10)
DM % (n)	16.06 % (470)	11.10 % (140)	19.81 % (330)
Mean Glucose (std)	7.38 (2.50)	6.97 (2.05)	7.71 (2.76)
Median baseline NIHSS (IQR)	15 (11-20)	13 (8-17)	17(13-21)
Median pre-stroke mRS (IQR)	0 (0-1)	0 (0-0)	0 (0-2)
Intravenous Alteplase % (n)	70.21 % (2055)	76.36 % (963)	65.55 % (1092)
Occlusion location % (n)†			
ICA	17.66 % (517)	12.69 % (160)	21.43 % (357)
M1	45.27 % (1325)	47.89 % (604)	43.28 % (341)
M2	22.45 % (657)	25.06 % (316)	20.46 % (341)
M3 or other	4.24 % (124)	5.07 % (64)	3.60 % (60)
Collateral score % (n)†			
Absent	5.09 % (149)	2.06 % (26)	7.38 % (60)
< 50 %	35.36 % (1035)	28.31 % (357)	40.69 % (678)
> 50 < 100 %	38.33 % (1122)	41.63 % (525)	35.83 % (597)
100 %	20.09 % (588)	27.04 % (341)	14.83 % (247)
Median Systolic BP (IQR)	150 (132-167)	147 (130-163)	150 (134-170)
Median time-to-groin (IQR)	190 (145-262)	175 (134-240)	205 (154-275)

Table A.4
Summary of Hyperparameter Tuning and Optimal Values.

Model/Component	Hyperparameter	Search Space	Optimal Value
MeD (ResNet)	Number of residual layers l	$\{10, 18, 34, 50, 101\}$	50
	Hidden layers h	$\{1 \dots 6\}$	3
MLP	Hidden units u	$\{4 \dots 128\}$	20
	Dropout d	$[0.1, 0.5]$	0.1
	Learning rate γ	$[0.0001, 0.01]$	0.0001
AdamW Optimizer	Weight decay λ	$[0.0001, 0.1]$	0.001
	(β_1, β_2)	$[(0.5, 0.9), (0.99, 0.9999)]$	(0.9, 0.9999)

References

[1] D.J. Mccarthy, A. Diaz, D.L. Sheinberg, B. Snelling, E.M. Luther, S.H. Chen, D.R. Yavagal, E.C. Peterson, R.M. Starke, in: Long-term Outcomes of Mechanical Thrombectomy for Stroke: A Meta-analysis, 2019. https://doi.org/10.1155/2019/7403104

[2] O.A. Berkhemer, P.S. Fransen, D. Beumer, L.A. V.D. Berg, H.F. Lingsma, A.J. Yoo, W.J. Schonewille, J.A. Vos, P.J. Nederkoorn, M.J. Wermer, M.A.V. Walderveen, J. Staals, J. Hofmeijer, J.A.V. Oostayen, G.J. L.A. Nijeholt, J. Boiten, P.A. Brouwer, B.J. Emmer, S.F.D. Bruijn, L.C.V. Dijk, L.J. Kappelle, R.H. Lo, E.J.V. Dijk, J. Vries, P.L.D. Kort, W.J. J.V. Rooij, J.S. V.D. Berg, B.A.V. Hasselt, L.A. Aerden, R.J. Dallinga, M.C. Visser, J.C. Bot, P.C. Vroomen, O. Eshghi, T.H. Schreuder, R.J. Heijboer, K. Keizer, A.V. Tielbeek, H.M.D. Hertog, D.G. Gerrits, R.M. V.D. Berg-Vos, G.B. Karas, E.W. Steyerberg, H.Z. Flach, H.A. Marquering, M.E. Sprengers, S.F. Jenniskens, L.F. Beenen, R.V.D. Berg, P.J. Koudstaal, W.H.V. Zwam, Y.B. Roos, A.V.D. Lugt, R.J.V. Oostenbrugge, C.B. Majoie, D.W. Dippel, A randomized trial of intraarterial treatment for acute ischemic stroke, New Engl. J. Med. 372(2015) 25517348. https://doi.org/10.1056/NEJMoa1411587

[3] J.P. Broderick, Y.Y. Palesch, A.M. Demchuk, S.D. Yeatts, P. Khatri, M.D. Hill, E.C. Jauch, T.G. Jovin, B. Yan, F.L. Silver, R. Kummer, C.A. Molina, B.M. Demaerschalk, R. Budzik, W.M. Clark, O.O. Zaidat, T.W. Malisch, M. Goyal, W.J. Schonewille, M. Mazighi, S.T. Engelter, C. Anderson, J. Spilker, J. Carrozzella, K.J. Ryckborst, L.S. Janis, R.H. Martin, L.D. Foster, T.A. Tomsick, in: Endovascular Therapy after Intravenous T-pa versus T-pa Alone for Stroke, 2013. https://doi.org/10.1056/nejmoa1214300

[4] B.C. Campbell, P.J. Mitchell, T.J. Kleinig, H.M. Dewey, L. Churilov, N. Yassi, B. Yan, R.J. Dowling, M.W. Parsons, T.J. Oxley, T.Y. Wu, M. Brooks, M.A. Simpson, F. Miteff, C.R. Levi, M. Krause, T.J. Harrington, K.C. Faulder, B.S. Steinfort, M. Priglinger, T. Ang, R. Scroop, P.A. Barber, B. Mcguinness, T. Wijeratne, T.G. Phan, W. Chong, R.V. Chandra, C.F. Bladin, M. Badve, H. Rice, L. Villiers, H. Ma, P.M. Desmond, G.A. Donnan, S.M. Davis, Endovascular therapy for ischemic stroke with perfusion-imaging selection, 2015. https://doi.org/10.1056/nejmoa1414792

[5] A. Ciccone, L. Valvassori, M. Nichelatti, A. Sgoifo, M. Ponzio, R. Sterzi, E. Boccardi, Endovascular treatment for acute ischemic stroke, 2013. https://doi.org/10.1056/nejmoa1213701

[6] S.G.H. Olthuis, F.A.V. Pirson, F.M.E. Pinckaers, W.H. Hinsenveld, D. Nieboer, A. Ceulemans, R.R. M.M. Knapen, M.M.Q. Robbe, O.A. Berkhemer, M.A. A.V. Walderveen, G.J.L. Nijeholt, M. Uyttenboogaart, W.J. Schonewille, P.M. V.D. Sluijs, L. Wolff, H.V. Voorst, A.A. Postma, S.D. Roosendaal, A.V.D. Hoorn, B.J. Emmer, M.G.M. Krietemeijer, P.-J.V. Doormaal, B. Roozenbeek, R.-J.B. Goldhoorn, J. Staals, I.R.D. Ridder, C.V.D. Leij, J.M. Coutinho, H.B. V.D. Worp, R.T.H. Lo, R.P.H. Bokkers, E.I.V. Dijk, H.D. Boogaarts, M.J.H. Wermer, A.C. G. M.V. Es, J.H.V. Tuijl, H.G.J. Kortman, R.A.R. Gons, L.S.F. Yo, J.-A. Vos, K.F.D. Laat, L.C.V. Dijk, I.R. V.D. Wijngaard, J. Hofmeijer, J.M. Martens, P.J. A.M. Brouwers, T. Bulut, M.J.M. Remmers, T.E. A.M. Jong, H.M.D. Hertog, B.A. A. M.V. Hasselt, A.D. Rozeman, O.E.H. Elgersma, B.V.D. Veen, D.R. Sudiono, H.F. Lingsma, Y.B. W. E.M. Roos, C.B. L.M. Majoie, A.V.D. Lugt, D.W.J. Dippel, W.H.V. Zwam, R.J.V. Oostenbrugge, Endovascular treatment versus no endovascular treatment after 6–24 h in patients with ischaemic stroke and collateral flow on ct angiography (mr clean-late, The Lancet 401 (2023) 1371–1380. In the Netherlands: a multicentre, open-label, blinded-endpoint, randomised, controlled, phase 3 trial, https://doi.org/10.1016/S0140-6736(23)00575-5

[7] J.M. Wardlaw, Even more benefit with endovascular treatment for patients with acute ischaemic stroke: mr clean-late, Lancet 401 (2023) 1317–1319. https://doi.org/10.1016/S0140-6736(23)00803-6

[8] A. Elhabr, J.M. Katz, J. Wang, M. Bastani, G. Martinez, M. Gribko, D.R. Hughes, P.C. Sanelli, Predicting 90-day modified rankin scale score with discharge information in acute ischaemic stroke patients following treatment,, BMJ Neurol. Open 3 (2021) 177. https://doi.org/10.1136/bmjno-2021-000177

[9] F. Kremers, E. Venema, M. Duvekot, L. Yo, R. Bokkers, G.L. Nijeholt, A.V. Es, A.V.D. Lugt, C. Majoie, J. Burke, B. Roozenbeek, H. Lingsma, D. Dippel, Outcome prediction models for endovascular treatment of ischemic stroke: systematic review and external validation, Stroke 53(3) (2022) 825–836.

[10] A. Hilbert, L. Ramos, H.V. Os, S. Olabarriaga, M. Tolhuisen, M. Wermer, R. Barros, I.V.D. Schaaf, D. Dippel, Y. Roos, W.V. Zwam, A. Yoo, B. Emmer, G.L. Nijeholt, A. Zwinderman, G. Strijkers, C. Majoie, H. Marquering, Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke, Comput. Biol. Med. 115 (2019) 103516. https://doi.org/10.1016/j.compbiomed.2019.103516

[11] E. Zihni, V. Madai, A. Khalil, I. Galinovic, J. Fiebach, J.D. Kelleher, D. Frey, 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, the 13th International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF, SciTePress, 2020, pp. 421–428. https://doi.org/10.5220/0008957304210428

[12] S. Bacchi, T. Zerner, L. Oakden-Rayner, T. Kleinig, S. Patel, J. Jannes, Deep learning in the prediction of ischaemic stroke thrombolysis functional outcomes: a pilot study, Acad. Radiol. 27(2) (2020) 19–e23. https://doi.org/10.1016/j.acra.2019.03.015

[13] A.C. Flint, S.P. Cullen, B. Faigeles, V. Rao, V.A. Rao, Predicting long-term outcome after endovascular stroke treatment: the totaled health risks in vascular events score, American Journal of Neuroradiology, 2010. https://doi.org/10.3174/ajnr.a2050

[14] H. Hallevi, A.D. Barreto, D.S. Liebeskind, M.M. Morales, S.B. Martin-Schild, A.T. Abraham, J. Gadia, J.L. Saver, Identifying patients at high risk for poor outcome after intra-arterial therapy for acute ischemic stroke, Stroke 40(2009)

1780–1785. The UCLA Intra-Arterial Therapy Investigators, https://doi.org/10.1161/STROKEAHA.108.535146
[15] G. Saposnik, A.K. Guzik, M.J. Reeves, B. Ovbiagele, S.C. Johnston, in: Stroke Prognostication Using Age and Nih Stroke Scale: Span-100, 2013. B013e31827b1ace. https://doi.org/10.1212/wnl.0
[16] Z.A. Samak, P. Clatworthy, M. Mirmehdi, Prediction of thrombectomy functional outcomes using multimodal data, 2020.
[17] S.D. Graaf, Automated functional outcome prediction in stroke using combined imaging and clinical parameters, Technical Report, TU Delft, the Netherlands, 2022. Master's thesis, https://resolver.tudelft.nl/uuid:b6b126e7-b589-428a-a86a-f3d04cf9f85c.
[18] L.A. Ramos, H.V. Os, A. Hilbert, S.D. Olabarriaga, A.V.D. Lugt, Y.B. W. E.M. Roos, W.H.V. Zwam, M.A. A.V. Walderveen, M. Ernst, A.H. Zwinderman, G.J. Strijkers, C.B. L.M. Majoie, M.J.H. Wermer, H.A. Marquering, Combination of radiological and clinical baseline data for outcome prediction of patients with an acute ischemic stroke, Front. Neurol. 13 (2022). https://doi.org/10.3389/fneur.2022.809343
[19] M.S. Jabal, O. Joly, D. Kallmes, G. Harston, A. Rabinstein, T. Huynh, W. Brinjikji, Interpretable machine learning modeling for ischemic stroke outcome prediction, Front. Neurol. 13 (2022) 884693. https://doi.org/10.3389/fneur.2022.884693
[20] I.-S. Park, S. Kim, J.-W. Jang, S.-W. Park, N.-Y. Yeo, S.Y. Seo, I. Jeon, S.-H. Shin, Y. Kim, H.-S. Choi, C. Kim, Multi-modality multi-task model for mrs prediction using diffusion-weighted resonance imaging, Sci. Rep. 14(1) (2024). https://doi.org/10.1038/s41598-024-71072-4
[21] H. Jo, C. Kim, D. Gwon, J. Lee, J. Lee, K.M. Park, S. Park, Combining clinical and imaging data for predicting functional outcomes after acute ischemic stroke: an automated machine learning approach, Sci. Rep. 13(1) (2023). https://doi.org/10.1038/s41598-023-44201-8
[22] F. Nijenhuis, R. Su, P.J.V. Doormaal, J. Hofmeijer, J. Martens, W.V. Zwam, A.V.D. Lugt, X. Zhang, T.V. Walsum, Multimodal deep learning for functional outcome prediction in endovascular therapy, in: U. Baid, R. Dorent, S. Malec, M. Pytlarz, R. Su, N. Wijethilake, S. Bakas, A. Crimi (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Cham, Springer Nature Switzerland, 2024, pp. 144–153. https://doi.org/10.1007/978-3-031-76160-7_14
[23] E. Venema, B. Roozenbeek, M.J. H.L. Mulder, S. Brown, C.B. L.M. Majoie, E.W. Steyerberg, A.M. Demchuk, K.W. Muir, A, P.J. Mitchell, S. Bracard, O.A. Berkhemer, G.J.L. Nijeholt, R.J.V. Oostenbrugge, Y.B. W. E.M. Roos, W.H.V. Zwam, A.V.D. Lugt, M.D. Hill, P. White, B.C.V. Campbell, F. Guillemin, J.L. Saver, T.G. Jovin, M. Goyal, D.W.J. Dippel, H.F. Lingsma, A.V.D. Lugt, J. Boiten, J.A. Vos, I.G.H. Jansen, R.B. Goldhoorn, K.C.J. Compagne, M. Kappelhof, J. Brouwer, S.J. Hartog, W.H. Hinsenveld, A.C. G. M.V. Es, B.J. Emmer, J.M. Coutinho, W.J. Schonewille, J.A. Vos, M.J.H. Wermer, M.A. A.V. Walderveen, J. Staals, J. Hofmeijer, J.M. Martens, J. Boiten, S.F.D. Bruijn, L.C.V. Dijk, H.B. V.D. Worp, R.H. Lo, E.J.V. Dijk, H.D. Boogaarts, J. Vries, P.L. M.D. Kort, J.V. Tuijl, J.P. Peluso, P. Fransen, J.S. P. V.D. Berg, B.A. A. M.V. Hasselt, L.A.M. Aerden, R.J. Dallinga, M. Uyttenboogaart, O. Eschgi, R.P.H. Bokkers, T.H. C. M.L. Schreuder, R.J.J. Heijboer, K. Keizer, L.S.F. Yo, H.M.D. Hertog, T. Bulut, P.J. A.M. Brouwers, M.A. A.V. Walderveen, M.E.S. Sprengers, S.F.M. Jenniskens, R.V.D. Berg, A.J. Yoo, L.F.M. Beenen, A.A. Postma, S.D. Roosendaal, B.F. W. V.D. Kallen, I.R. V.D. Wijngaard, A.C. G. M.V. Es, B.J. Emmer, J.M. Martens, L.S.F. Yo, J.A. Vos, J. Bot, P.J.V. Doormaal, A. Meijer, E. Ghariq, R.P.H. Bokkers, M.P.V. Proosdij, G.M. Krietemeijer, J.P. Peluso, H.D. Boogaarts, R. Lo, D. Gerrits, W. Dinkelaar, A.P.A. Appelman, B. Hammer, S. Pegge, A.V.D. Hoorn, S. Vinke, J. Boiten, J.A. Vos, W.J. Schonewille, J. Hofmeijer, J.M. Martens, H.B. V.D. Worp, R.H. Lo, J. Hofmeijer, H.Z. Flach, N.E. Ghannouti, M. Sterrenberg, W. Pellikaan, R. Sprengers, M. Elfrink, M. Simons, M. Vossers, J. Meris, T. Vermeulen, A. Geerlings, G.V. Vemde, T. Simons, G. Messchendorp, N. Nicolaij, H. Bongenaar, K. Bodde, S. Kleijn, J. Lodico, H. Droste, M. Wollaert, S. Verheesen, D. Jeurrissen, E. Bos, Y. Drabbe, M. Sandiman, N. Aaldering, B. Zweedijk, J. Vervoort, E. Ponjee, S. Romviel, K. Kanselaar, D. Barning, V. Chalos, R.R. Geuskens, T.V. Straaten, S. Ergezen, R.R.M. Harmsma, D. Muijres, A. Jong, A.M.M. Boers, J. Huguet, P.F.C. Groot, M.A. Mens, K.R.V. Kranendonk, K.M. Treurniet, M.L. Tolhuisen, H. Alves, A.J. Weterings, E.L.F. Kirkels, E.J. H.F. Voogd, L.M. Schupp, S.L. Collette, A.E.D. Groot, N.E. Lecouffe, P.R. Konduri, H. Prasetya, N. Arrarte-Terreros, Stroke 52(2021) 2764–2772. Prediction of Outcome and Endovascular Treatment Benefit: Validation and Update of the MR PREDICTS Decision Tool, https://doi.org/10.1161/STROKEAHA.120.032935
[24] S. Chen, K. Ma, Y. Zheng, Med3d: Transfer learning for 3d medical image analysis, arxiv:1904.00625 edition, 2019. http://arxiv.org/abs/1904.00625.
[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2020. https://doi.org/10.48550/ARXIV.2010.11929
[26] B. Xu, N. Wang, T. Chen, M. Li, Empirical evaluation of rectified activations in convolutional network, 2015. https://doi.org/10.48550/ARXIV.1505.00853
[27] T.M. Consortium, Project monai, 2020. https://doi.org/10.5281/zenodo.4323059
[28] 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). https://doi.org/10.1136/bmj.k949
[29] R. Peter, B.J. Emmer, A.C.V. Es, T.V. Walsum, Cortical and vascular probability maps for analysis of human brain in computed tomography images, in: IEEE 14Th International Symposium on Biomedical Imaging, 2017, pp. 1141–1145. https://doi.org/10.1109/ISBI.2017.7950718
[30] B. Avants, N. Tustison, G. Song, Advanced normalization tools (ants), 11, 2008, pp. 1–35. https://doi.org/10.54294/uvnhin
[31] S.V. Buuren, K. Oudshoorn, Flexible multivariate imputation by MICE, TNO, Leiden, 1999.
[32] E.R. Delong, D.M. Delong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics 44(3) (1988) 837–845.
[33] Surf, Snellius: the national supercomputer, 2021. https://www.surf.nl/en/services/snellius-the-national-supercomputer.
[34] G.V. Rossum, F.L. Drake, Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, 2009.
[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A.K. "opf, E. Yang, Z. Devito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, arxiv:1912.01703 edition, 2019. http://arxiv.org/abs/1912.01703.
[36] A. Chen, A. Chow, A. Davidson, A. Dcunha, A. Ghodsi, S.A. Hong, A. Konwinski, C. Mewald, S. Murching, T. Nykodym, P. Ogilvie, M. Parkhe, A. Singh, F. Xie, M. Zaharia, R. Zang, J. Zheng, C. Zumar, Proceedings of the Fourth International Workshop on Data Management for End-to-End Machine Learning, DEEM'20, the Fourth International Workshop on Data Management for End-to-End Machine Learning, DEEM'20New York, NY, USA, Association for Computing Machinery, 2020. Developments in mlflow: A system to accelerate the machine learning lifecycle, https://doi.org/10.1145/3399579.3399867
[37] I. Loshchilov, F. Hutter, Fixing weight decay regularization in adam, arxiv:1711.05101 edition, 2017. http://arxiv.org/abs/1711.05101.
[38] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna, A next-generation hyperparameter optimization framework, in: Proceedings of the 25Th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.
[39] S. Toy, Y. Secgin, Z. Oner, M.K. Turan, S. Oner, D. Senol, A study on sex estimation by using machine learning algorithms with parameters obtained from computerized tomography images of the cranium, Sci. Rep. 12 (2022) 4278. https://doi.org/10.1038/s41598-022-07415-w
[40] R. Çiftçi, E. Dönmez, A. Kurtoğlu, O. Eken, N.A. Samee, R.I. Alkanhel, Human gender estimation from ct images of skull using deep feature selection and feature fusion, Sci. Rep. 14(1) (2024) 16879. https://doi.org/10.1038/s41598-024-65521-3
[41] H. Kondou, R. Morohashi, S. Kimura, N. Idota, R. Matsunari, H. Ichioka, R. Bandou, M. Kawamoto, D. Ting, H. Ikegaya, Artificial intelligence-based forensic sex determination of east asian cadavers from skull morphology, Sci. Rep. 13(1) (2023) 21026. https://doi.org/10.1038/s41598-023-48363-3
[42] R. Lye, H. Min, J. Dowling, Z. Obertová, M. Estai, N.A. Bachtiar, D. Franklin, Deep learning versus human assessors: forensic sex estimation from three-dimensional computed tomography scans, Sci. Rep. 14(1) (2024) 30136. https://doi.org/10.1038/s41598-024-81718-y
[43] T. Seo, Y. Yoon, Y. Kim, Y. Usumoto, N. Eto, Y. Sadamatsu, R. Tadakuma, J. Morishita, Sex estimation using skull silhouette images from postmortem computed tomography by deep learning, Sci. Rep. 14(1) (2024) 22689. https://doi.org/10.1038/s41598-024-74703-y
[44] Y. Liu, Y. Yu, J. Ouyang, B. Jiang, G. Yang, S. Ostmeier, M. Wintermark, P. Michel, D.S. Liebeskind, M.G. Lansberg, G.W. Albers, G. Zaharchuk, Functional outcome prediction in acute ischemic stroke using a fused imaging and clinical deep learning model, Stroke 54(2023) 2316–2327. https://doi.org/10.1161/STROKEAHA.123.044072
[45] L.A. Ramos, M. Kappelhof, H.J. A.V. Os, V. Chalos, K.V. Kranendonk, N.D. Kruyt, Y.B. W. E.M. Roos, A.V.D. Lugt, W.H.V. Zwam, I.C. V.D. Schaaf, A.H. Zwinderman, G.J. Strijkers, M.A. A.V. Walderveen, M.J.H. Wermer, S.D. Olabarriaga, C.B. L.M. Majoie, H.A. Marquering, Predicting poor outcome before endovascular treatment in patients with acute ischemic stroke, Front. Neurol. 11 (2020) 580957. https://doi.org/10.3389/fneur.2020.580957