Computerized Medical Imaging and Graphics 99 (2022) 102089

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

FeMA: Feature matching auto-encoder for predicting ischaemic stroke evolution and treatment outcome

Zeynel A. Samak [a,*,1,2,3], Philip Clatworthy [b,c], Majid Mirmehdi [a,*,**]

[a] Department of Computer Science, University of Bristol, Bristol, UK
[b] Translational Health Sciences, University of Bristol, Bristol, UK
[c] Stroke Neurology, Southmead Hospital, North Bristol NHS Trust, Bristol, UK

ARTICLE INFO

Keywords:
Stroke evolution
Follow-up prediction
Treatment outcome
Classification
Prognosis
CNN
Feature similarity
NCCT

ABSTRACT

Although, predicting ischaemic stroke evolution and treatment outcome provide important information one step towards individual treatment planning, classifying functional outcome and modelling the brain tissue evolution remains a challenge due to data complexity and visually subtle changes in the brain. We propose a novel deep learning approach, Feature Matching Auto-encoder (FeMA) that consists of two stages, predicting ischaemic stroke evolution at one week without voxel-wise annotation and predicting ischaemic stroke treatment outcome at 90 days from a baseline scan. In the first stage, we introduce feature similarity and consistency objective, and in the second stage, we show that adding stroke evolution information increase the performance of functional outcome prediction. Comparative experiments demonstrate that our proposed method is more effective to extract representative follow-up features and achieves the best results for functional outcome of stroke treatment.

1. Introduction

Ischaemic stroke, a condition caused by a blockage of a blood vessel in the brain due to a blood clot, is the most common type (80 %) of stroke disease (Renowden, 2014) which is a leading cause of death in the UK (Stroke Association, 2018) and worldwide (WHO, 2018). The most effective treatment in the most severe ischaemic stroke cases is mechanical thrombectomy where thrombi in the large intracranial arteries are removed via an intra-arterial catheter to restore blood flow. However, thrombectomy risks include brain haemorrhages and death. Hence, the treatment decision is patient-specific and determined by the physician after assessment of the potential risks and benefits of functional outcome. The purpose of clinicians is to utilise all information sources (e.g., imaging and/or clinical information) to decide the most optimal treatment option as soon as possible, as the effect of treatment highly depends on the time to treatment.

Modelling the evolution of changes in the brain in the first few weeks and months post-stroke is key to establishing methods for predicting the effects and outcome of immediate stroke treatment for future patients. The functional outcome of treatment is frequently measured in the clinic using the modified Rankin Scale (mRS) (Van Swieten et al., 1988), starting at 0 for no symptoms, through 1–5 for degree of severity of symptoms, and finally 6 for death. However, predicting brain tissue changes following a stroke is challenging as it is inherently difficult due to the irregularity in lesion shape, size and location. This is sometimes the case even for radiologists to spot the damaged tissue from the NCCT scan on admission – see examples in Figs. 1, 2.

Deep learning has shown unparalleled levels of success in various image analysis domains, e.g., action recognition, image segmentation and classification. In particular, convolutional neural networks (CNNs) have been extensively applied to medical image analysis tasks attaining state-of-the-art results, for example in stroke-related applications, such as *lesion detection* (Chen et al., 2020; Gautam and Raman, 2021), *segmentation* (Clèrigues et al., 2020; Pinto et al., 2021) and *severity prediction or prognosis* (Samak et al., 2020; Robben et al., 2020).

In stroke outcome applications, the majority of the studies have designed their task as either a segmentation of final stroke lesion (Scalzo et al., 2012; Lucas et al., 2018; Pinto et al., 2021) or a prediction of

* Corresponding authors.
** Principal corresponding author.
E-mail addresses: zeynel.samak@bristol.ac.uk (Z.A. Samak), phil.clatworthy@bristol.ac.uk (P. Clatworthy), m.mirmehdi@bristol.ac.uk (M. Mirmehdi).
[1] 0000-0002-3835-4811
[2] 0000-0002-1206-3573
[3] 0000-0002-6478-1403

https://doi.org/10.1016/j.compmedimag.2022.102089
Received 17 September 2021; Received in revised form 4 May 2022; Accepted 3 June 2022
Available online 11 June 2022

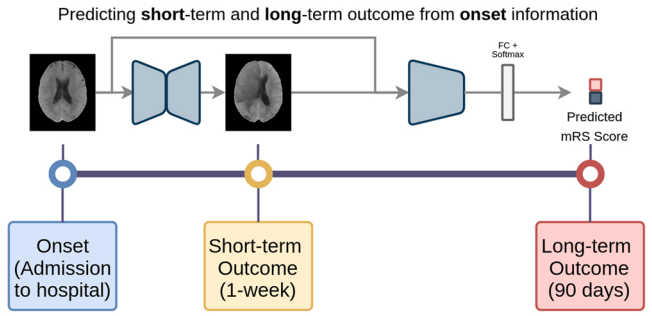

Fig. 1. FeMA predicts ischaemic stroke evolution at one week (short-term outcome) and functional outcome of treatment (long-term outcome) leveraging only the baseline NCCT scan on Stroke patient's admission to hospital.

functional outcome (via mRS scores) (Bacchi et al., 2019; Hilbert et al., 2019; Samak et al., 2020). Segmentation approaches require voxel-wise annotation, best performed by expert radiologists, which is laborious and costly. On the other hand, predicting mRS scores directly from the baseline scan and/or clinical information does not provide information about the progression of the disease and so is not enough for clinical use. Ernst et al. (2017) have shown the importance of the 1-week follow-up scan on mRS scores, by investigating the correlation between the lesion volume in mRS-relevant brain regions in 1-week follow-up NCCT scans with functional outcome. Only relatively few studies have investigated the combination or integration of final stroke lesion segmentation and mRS prediction, for example (Maier and Handels, 2016; Choi et al., 2016; Nishi et al., 2020). Although these methods provide information both about disease evolution and functional outcome, they still require annotations of segmentation maps.

This work aims to model stroke progression, without manual annotations, to build a model capable of predicting the outcome of thrombectomy treatment on new hospital admissions - with the immediate benefit that it would allow clinicians to decide whether to apply thrombectomy or not. In the first step, our model predicts the follow-up non-contrast computed tomography (NCCT) scan (1-week follow-up) from the baseline scan. Then, it combines the baseline scan with the predicted follow-up scan to predict the functional outcome of the stroke treatment - the mRS scores. Therefore, our model can assist physicians both qualitatively by the predicted scans and quantitatively by the predicted mRS score.

There is no universally accepted method of selecting patients for thrombectomy. As a minimum, patients should have an identified

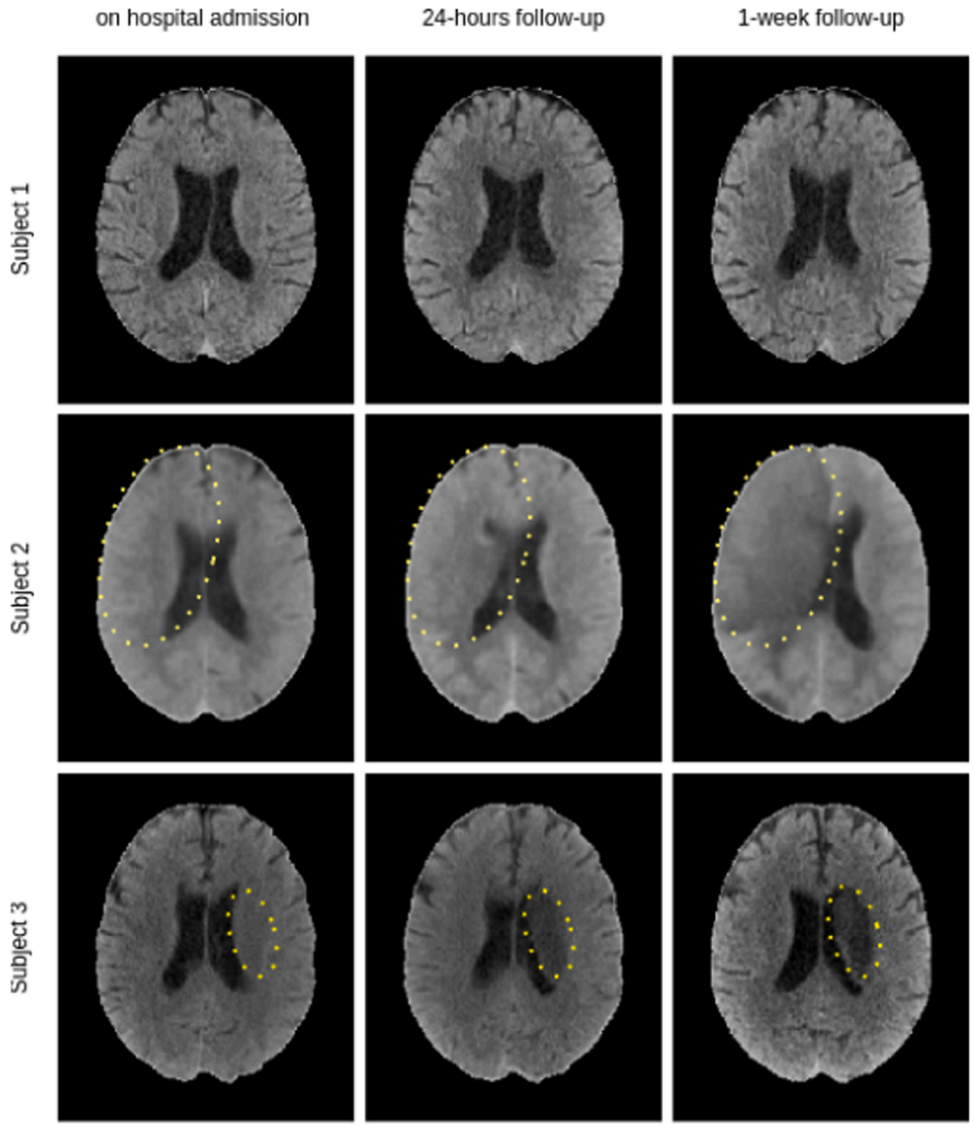

Fig. 2. Example of stroke NCCT scans (from MR CLEAN), first column is the first scan when the patient was admitted to hospital, middle column is the 24 h follow-up scan and last column is the 1-week follow-up scan. Affected regions are highlighted. Unaffected slices for Subject 1 are shown for comparative purposes.

occlusion of a large intracranial artery and no intra-cerebral haemorrhage on NCCT imaging of the brain. Potentially eligible patients therefore currently undergo NCCT or magnetic resonance imaging (MRI) to exclude intra-cerebral haemorrhage, followed by angiographic imaging, usually CT or MR angiography, to identify occlusion of a large intracranial artery (large vessel occlusion, LVO) to be targeted by thrombectomy treatment. Most patients who are eligible for thrombectomy are also eligible for intravenous thrombolysis. Both are time-critical treatments, therefore it is important that methods of assessing eligibility for thrombectomy do not unnecessarily delay intravenous thrombolysis. Furthermore, it has been estimated that a typical patient loses close to 2 million neurons for every minute that a stroke remains untreated (Saver, 2006). A large registry-based study found that every minute of delay to intravenous thrombolysis was associated with a measurable effect on risk of early post-treatment intra-cerebral haemorrhage, disability and death (Darehed et al., 2020).

NCCT brain imaging is extremely rapid, taking as little as 1–2 min for images to be acquired. CT angiography, CT perfusion and MRI take longer to perform; while scan time is only a matter of minutes, this imaging is not always readily available, leading to further delays. An imaging-based method of patient selection for thrombectomy that requires only NCCT scans therefore has the potential for significant patient benefit. Other benefits include: uncommon but occasionally severe reactions to intravenous contrast media used for CT and MR angiography; cost of intravenous contrast, scanner operator time and image reporting; and the fact that NCCT is the only brain imaging modality that is almost universally available in hospitals treating stroke across the world. Moreover, Weyland et al. (2022) recently showed that using only non-contrast CT are improving and may give useful information about clot constituents not provided by angiographic imaging, supporting the potential feasibility of identifying LVO using such an approach. Our contributions in this work are as follow: (i) we propose a self-supervised, annotation-free, voxel-wise approach to predict the follow-up scans in stroke patients from their baseline scan, (ii) our proposed network learns to predict mRS scores from embedded features that are a combined representation of the baseline scan and our predicted follow-up scans, and finally (iii) we present extensive comparative experiments to evaluate and validate the proposed model for relevant architectures, i.e. the auto-encoder (AE), perceptual AE (AE_{PL}) (Johnson et al., 2016), 3D U-Net (Çiçek et al., 2016) and Wasserstein generative adversarial network (WGAN) (Bowles et al., 2018; Kwon et al., 2019). Comprehensive empirical tests show that our proposed method, FeMA achieves competitive results compared to U-Net based methods and superior results compared to state-of-the-art non-U-Net based methods in follow-up scan prediction. Also, FeMA obtains state-of-the-art results compared to both U-Net and non-U-Net based methods in mRS score prediction, at 0.79 AUC.

The remainder of the paper is structured as follows. We address related works in Section 2 and review the MR CLEAN dataset and our preprocessing methods in Section 3. In Section 4, we develop the fundamental components of our proposed method. Our experimental set-up and comparative results are presented in Section 6. Finally, we provide conclusions and future work in Section 7.

2. Related works

Over the past decades, a variety of methods have been proposed for stroke lesion detection and segmentation, ranging from image thresholding (Chawla et al., 2009) to machine learning (Mckinley et al., 2017; Maier et al., 2014) and more recently, deep learning, such as (Clèrigues et al., 2020; Wang et al., 2020; Pinto et al., 2021). Most of the works that have attempted to estimate treatment success, given the patient's imaging and/or clinical information, have investigated either the evolution of the appearance of the final stroke lesion or the process to arrive at an mRS score. Only relatively few studies have attempted to investigate these together. In this section, we review previous works based on their approach to the estimation of stroke treatment outcome. In the main, these can be categorised as methods that *predicts by:* (i) final stroke lesion segmentation, (ii) mRS scores and (iii) combining or integrating the categories in (i) and (ii).

Final Stroke Lesion Segmentation – In this approach, stroke treatment outcome is determined by predicting what the final stroke lesion appearance would be (within a range of 5–90 days), given the baseline scan of a patient at the time of admission. In such models, an annotated dataset is required for training, where for each baseline scan, stroke lesions are manually segmented on the corresponding follow-up scan by expert neuroradiologists. The model is then trained on the baseline scan with a target of the final stroke lesion segmentation map. At inference for a new patient, the model visualises the 'final' stroke lesion map based on their admission baseline scan and this allows clinicians to select treatment.

There have been several methods that follow this strategy using linear regression models (Kemmling et al., 2015), decision trees (Boers et al., 2013; Mckinley et al., 2017), and CNNs-based deep learning (Scalzo et al., 2012; Pinto et al., 2018, 2021). Many of these works have participated in the ISLES[4] Challenge (Clèrigues, 2018; Winzeck et al., 2018; Pinto et al., 2018, 2021) which functions as a benchmark for the state-of-the-art in stroke segmentation. For example, Pinto et al. (2021) obtained the best result on the ISLES 2017 dataset benchmark via a two-stage method: in an unsupervised stage, two Restricted Boltzmann Machines (RBMs), each used a different subset of parametric MRI maps to establish lesion location and blood flow circulation features. Then, in a supervised stage, these were fed through CNN and Recurrent Neural Networks (RNNs), along with parametric MRI maps, to extract short and long distance spatial relationships to predict the stroke lesions. Robben et al. (2020) used spatio-temporal CT perfusion data and clinical information (e.g.time to recanalization, completeness of recanalization) as input to a CNN model based on DeepMedic (Kamnitsas et al., 2017) to predict 24-hour or 5-day follow-up NCCT scans.

This kind of approaches requires voxel-wise annotation of the 3D volumes for training, but our proposed method is annotation free and predicts whole follow-up scans instead of the final lesion segmentation map in order to model the tissue evolution.

Prediction of mRS Scores – In this category, the models predict mRS scores at 90 days by leveraging imaging and/or clinical information at the point of hospital admission. The dataset for such models typically consists of imaging and/or clinical information with the mRS scores acquired by interviewing patients or their carer with a structured questionnaire 90 days post-stroke treatment. The models are trained with the objective of predicting outcome as dichotomised (mRS scores 0–2 as favourable and mRS scores $3-6$ as unfavourable) or individual mRS score class ($0-6$).

The majority of such studies have applied machine learning to clinical information, such as logistic regression (Venema et al., 2017; Heo et al., 2018), SVMs (Asadi et al., 2014; Bentley et al., 2014), RFs (van Os et al., 2018; Heo et al., 2018) and artificial neural networks (ANNs) (Asadi et al., 2014; van Os et al., 2018). A few works have used imaging with or without clinical records to predict mRS scores (Bacchi et al., 2019; Hilbert et al., 2019; Samak et al., 2020; Osama et al., 2020). For example, Osama et al. (2020) utilised a Siamese architecture to predict mRS scores from the three middle slices of their multi-parametric MRI scans. To overcome data scarcity and class imbalance, they employed the same number of samples for each class and set the number of similar and dissimilar pairs to be equal during training. The mRS score of a test sample was then the class of the training data point it was closest to via feature similarity. Recently, Samak et al. (2020) combined 3D NCCT and clinical information to predict the dichotomised and individual scores of functional outcome of thrombectomy treatment. They proposed a multimodal CNN network with an attention module that captures global

[4] Ischaemic Stroke Lesion Segmentation, http://www.isles-challenge.org

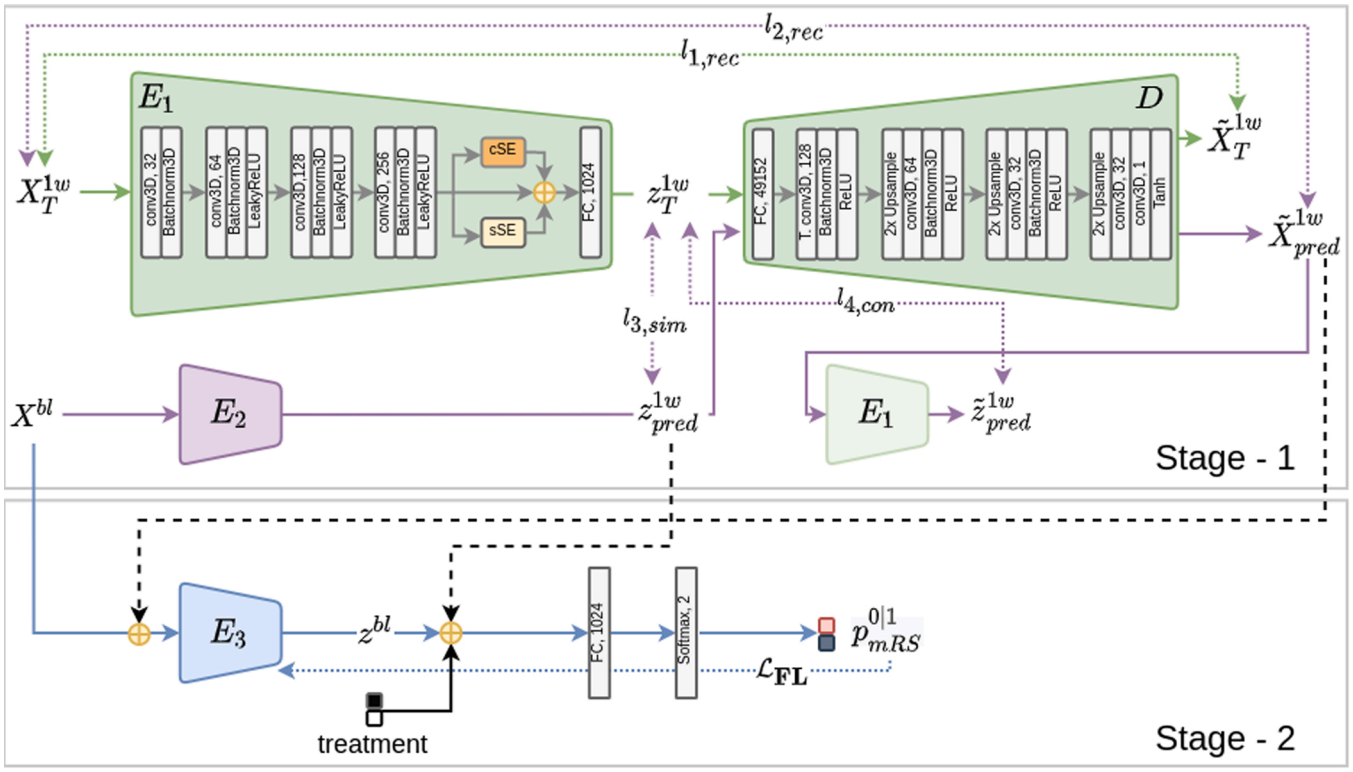

Fig. 3. FeMA for extracting the feature and predicting the volume of follow-up scans. X^{bl}: baseline scan, X_T^{1w}: target 1-week follow-up scan, $\tilde{X}_{pred}^{1w}$: predicted 1-week follow-up scan from X^{bl}, $\tilde{X}_{gt}^{fu}$: reconstructed target 1-week follow-up scan from X_{gt}^{1w}, 'z's are the features of the related encoder input. The dotted black lines show the optional information utilised in Stage-2 (see analysis in Section 6.2).

feature inter-dependencies both spatially and channel-wise. They provided the prediction of treatment outcome at patient level, whereas our proposed method predicts at both patient level and tissue level by visualising the evolution of the brain regions in the follow-up scan, including the stroke-affected regions.

Final Stroke Lesion Segmentation and Prediction of mRS Scores – Studies applying this approach are examples of the sequential combination or integration of the two previous categories. Choi et al. (2016), the winner of ISLES 2016 Challenge, proposed an ensemble of CNN models whose weighted average was used for the segmentation of final stroke lesion. They applied four 3D U-Nets to perform voxel-wise final lesion segmentation and two sets of Fully Connected Networks (FCNs) to perform patch-wise classification. The first FCN determined whether a patch includes any lesion voxels and the second FCN classified a patch if the centre voxel was a lesion. Then, for mRS score prediction, again the average weighted outcome of a linear regressor trained on clinical information and a CNN classifier trained to segment stroke lesions at patch level was used.

Nishi et al. (2020) integrated the final stroke lesion and mRS score prediction tasks into a U-Net model that trained on diffusion-weighted images of a dataset of 250 patients and validated on an external dataset of 74 patients who underwent mechanical thrombectomy. They used their encoder feature maps for mRS score prediction and the final output of their model for final stroke lesion segmentation, and compared their approach with a logistic regression model that trained on manually acquired neuroimaging biomarkers. For functional outcome, they used dichotomised mRS scores. This is the closest study to our work in terms of predicting mRS scores and lesion evolution on follow-up scans. However, using the final feature maps of their U-Net encoder for mRS score prediction could negatively impact their prediction performance as some information may be omitted due to the skip connections. For this reason, we use an auto-encoder based approach to encode NCCT volume information in latent code and perform each task separately for better focus on each specific task. As before, the methods in this category rely on annotated training maps for their segmentation.

Other Relevant Approaches – A few studies in various medical domains, such as Alzheimer's disease (AD) (Bowles et al., 2018; Wegmayr et al., 2019) and lung nodule growth detection (Rafael-Palou et al., 2021), have investigated modelling disease evolution by predicting follow-up scans directly from baseline scans. For example, Wegmayr et al. (2019) employed a generative adversarial network (GAN) to model disease progression to predict whether a person diagnosed with mild cognitive impairment (MCI) will convert to AD within five years, and similarly, Bowles et al. (2018) used a Wasserstein GAN (WGAN) on 3D MRI patches to model the progression of AD. To the best of our knowledge, there is no such disease evolution modelling study in the Stroke domain that has attempted to predict final stroke lesion without using segmentation maps. We believe the success and effectiveness of the GANs in medical imaging Yi et al. (2019), and particularly in modelling disease evolution (Bowles et al., 2018; Wegmayr et al., 2019) and image generation Kwon et al. (2019), can positively impact the Stroke domain. We present pioneering results in this area and compare against our implementation of other works (Bowles et al., 2018; Kwon et al., 2019) that were applied to other areas and data modalities.

Another approach of interest applied with significant success in other domains, such as image embedding (Pihlgren et al., 2020), CT denoising (Gholizadeh-Ansari et al., 2020) and medical image translation (Armanious et al., 2020), involves the application of a perceptual loss (PL) to encourage a network to capture high-level structure (perceptual quality) in the image by comparing the feature maps of a predicted image and a target image obtained from a pre-trained network. This allows a network to generate visually plausible images and representative image feature embeddings. We demonstrate the impact of PL on our mRS score prediction.

3. Data and preprocessing

In this work, we use one of the most comprehensive NCCT datasets of patients who underwent ischaemic stroke treatment from a multi-centre study involving several stroke clinics, i.e. the MR CLEAN Trial dataset.[5] This was a randomised, clinical trial of intra-arterial treatment versus usual care in patients with a proximal arterial occlusion in the anterior circulation treated within 6 h of symptom onset. Five hundred patients (233 assigned to EVT and 267 to usual care) were treated in 16 medical centres in the Netherlands. The dataset includes a baseline or onset (i.e., the first scan when the patient was admitted to hospital) NCCT (for all 500 patients), 24-hour follow-up NCCT (for 394 patients), and a 1-week follow-up NCCT scan (for 358 patients). An example of ischaemic changes following a stroke for this timeline (onset, 24-hours, 1 week) can be seen in Fig. 2. For more detailed information about the dataset, see the MR CLEAN study protocol (Berkhemer et al., 2015; Fransen et al., 2014).

As multiple clinics were involved, there are various acquisition protocols of the NCCT scans in the MR CLEAN dataset. Through preprocessing, we reduced some of this variation to allow our learning network to deal with the same standard input and a smaller image size. First, all the scans were re-sampled to the same voxel size of $3 \times 1 \times 1$ mm^3 (zyx orientation) followed by clipping the intensity range with window level 40HU and width 80HU which is standard for brain tissue. Then, as skull-stripping generates more non-brain regions, we removed the unnecessary non-brain regions from the volume (i.e., the black regions in Fig. 2). Next, the volumes were centre cropped with size $32 \times 192 \times 128$ (zyx orientation) – selected by visually inspecting the outputs – to standardize inputs for the models. Also, due to the relatively small size of the dataset, augmentations, such as horizontal/vertical flips, rotations, elastic deformations and Gaussian noise, were applied to help the network train more effectively. The image voxels were finally normalised to zero mean and one standard deviation. We also scaled all data into the ± 1 range. It should be noted that as the dataset contains only stroke cases, we assume that the NCCT scans fed into our model is a stroke case whether it has been determined by a physician or another automatic stroke detection model.

4. Proposed approach

An overview of our proposed method FeMA is presented in Fig. 3, which illustrates our process for predicting ischaemic stroke evolution, as well as mRS scores, given a baseline scan on hospital admission. In this section, we explain the components of our network architecture and loss functions in detail.

Ideally, if a follow-up scan can be predicted from a baseline scan one week after stroke treatment, it will provide significant information about the condition of patients in the future – and so would allow making a more robust prediction of a mRS score at 90 days. In the proposed method, there are two main stages. The first step extracts the 1-week follow-up scan features and volume from the baseline scan, as seen in Fig. 3 (Stage-1). The second combines the predicted follow-up scan information with the baseline scan to deduce the dichotomised mRS score at 90 days, Fig. 3 (Stage-2).

Let $Q = \{X_i^{bl}, X_{T,i}^{1w}, y_i\}_{i=1}^{N}$ be the training set of N patients, where X_i^{bl} and $X_{T,i}^{1w} \in \mathbb{R}^{1xDxWxH}$ denote a 3D baseline and 1-week follow-up NCCT scan target respectively, with D slices of height H and width W, and $y_i \in [0,1]$ is the dichotomised mRS score of the i^{th} patient.

4.1. Predicting 1-week follow-up scans

In Stage-1 of FeMA, the goal is to predict 1-week follow-up scan features and volume given an admission time baseline scan X^{bl}. This is performed in two steps: the first encodes and reconstructs a 1-week follow-up scan and the second predicts 1-week follow-up scan features and volume from a baseline scan. Note that these two steps are performed consecutively in each training update.

This stage consists of two identical encoders E_1 and E_2, and a decoder D that operate as follows (see top of Fig. 3). During training, in the first step, encoder E_1 *extracts features* z_T^{1w} *of the actual target follow-up scan* X_T^{1w}, and the decoder D processes these features to reconstruct a follow-up scan $\widetilde{X}_T^{1w}$. E_1 and D are trained to minimise the follow-up scan reconstruction loss, $l_{1,rec}$,

$$\begin{aligned} \mathscr{L}_{\mathbf{E_1,D}} &= l_{\mathbf{1,rec}}, \\ \mathscr{L}_{\mathbf{E_1,D}} &= \left\| \widetilde{X}_T^{1w} - X_T^{1w} \right\|_2 \end{aligned} \tag{1}$$

In the second step of Stage-1, still during training, the weights of E_1 and D are frozen and the focus is on E_2. Encoder E_2 *predicts the 1-week follow-up scan features* z_{pred}^{1w} *from the baseline scan* X^{bl}, and decoder D processes these features to predict a follow-up scan $\widetilde{X}_{pred}^{1w}$. To further improve a follow-up prediction from a baseline scan, $\widetilde{X}_{pred}^{1w}$ is re-routed back through E_1 to impose cyclic consistency of target follow-up features z_T^{1w} and extracted features $\widetilde{z}_{pred}^{1w}$ of predicted follow-up volume $\widetilde{X}_{pred}^{1w}$.

E_2 is regulated by $\mathscr{L}_{\mathbf{E_2}}$, comprising the reconstruction loss $l_{2,rec}$ of $\widetilde{X}_{pred}^{1w}$, the feature similarity loss $l_{3,sim}$ of z_{pred}^{1w} and the consistency loss $l_{4,con}$ of $\widetilde{z}_{pred}^{1w}$ with z_T^{1w}:

$$\begin{aligned} \mathscr{L}_{\mathbf{E_2}} &= \lambda_1 \cdot l_{\mathbf{2,rec}} + \lambda_2 \cdot l_{\mathbf{3,sim}} + \lambda_3 \cdot l_{\mathbf{4,con}}, \\ \mathscr{L}_{\mathbf{E_2}} &= \lambda_1 \cdot \left\| \widetilde{X}_{pred}^{1w} - X_T^{1w} \right\|_2 + \lambda_2 \cdot \left\| z_{pred}^{1w} - z_T^{1w} \right\|_2 \\ &\quad + \lambda_3 \cdot \left\| \widetilde{z}_{pred}^{1w} - z_T^{1w} \right\|_2 \end{aligned} \tag{2}$$

where $\lambda_1 = 0.1$, $\lambda_2 = 1$ and $\lambda_3 = 10$ modulate the contribution of each loss.

At the inference stage of follow-up volume prediction, encoder E_2 processes a baseline scan X^{bl} to estimate follow-up features z_{pred}^{1w} and these features pass into decoder D to predict follow-up volume $\widetilde{X}_{pred}^{1w}$.

4.2. Predicting mRS scores

In Stage-2 of FeMA (see lower part of Fig. 3), the aim is to make a robust prediction of mRS scores by utilising the baseline scan X^{bl} and the predicted follow-up features z_{pred}^{1w} acquired from X^{bl} in Stage-1. Hence, armed with (X^{bl}, z_{pred}^{1w}), encoder E_3 - which is identical to E_1 and E_2, but does not share weights - operates during training as follows. It processes X^{bl} to extract features z^{bl} which are then concatenated with z_{pred}^{1w} and treatment information are processed by a fully connected (FC) layer and a Softmax layer to generate a probability vector, p_{mRS}.

E_3 and the fully connected layers are updated with the objective of minimising loss $\mathscr{L}_{\mathbf{FL}}$. As the class distribution in the dataset is imbalanced, we apply focal loss (Lin et al., 2017; Abraham and Khan, 2019) to generate an adaptive weighting loss function, such that

$$\mathscr{L}_{\mathbf{FL}} = -\sum_{i=1}^{M} \alpha_i y_i (1 - p_i)^{\gamma} \log(p_i) \quad , \tag{3}$$

where M is number of training samples, α_i is a weighting factor of each class, where in the case of binary classification, α_i is set to a value between 0 and 1 to balance the positive and negative labelled samples. γ is the focal intensity factor, where the higher the value of γ, the lesser the cost contribution by well-classified samples.

At inference time, E_2 and E_3 are frozen and FeMA takes only a baseline scan X^{bl} to predict mRS scores.

[5] https://www.mrclean-trial.org/home.html

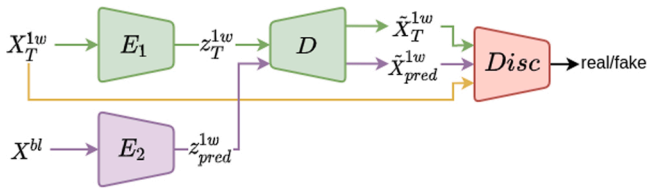

Fig. 4. FeMA$_{WGAN}$: the GAN extension for Stage-1 of FeMA.

4.3. GAN extension of FeMA (FeMA$_{WGAN}$)

To model disease evolution, we extend our method by adding a discriminator module to our Stage-1 architecture (see Fig. 4) to benefit from adversarial training and enable our model to generate visually plausible disease evolution volumes. As the vanilla GAN loss suffers from training instability and convergence issues (Kwon et al., 2019), we use the loss from WGAN (Arjovsky et al., 2017), which measures the distance between two probability distributions (training samples and generated examples), with a gradient penalty (gp) (Gulrajani et al., 2017).

The architecture of our discriminator *Disc* is identical to that of our encoders. During training, we consider both $\tilde{X}_T^{1w}$ and $\tilde{X}_{pred}^{1w}$ as fake, and X_T^{1w} as the real sample. Therefore, the adversarial loss contributes to the training of D when reconstructing the target follow-up scan, and E_2 when predicting the follow-up scan. Thus, this loss encourages D to generate visually better reconstruction and prediction of the follow-up scan, and also encourages E_2 to predict representative follow-up features. *Disc* is trained with loss $\mathscr{L}_{\text{Disc}}$ with gradient penalty $l_{gp-Disc}$ where $\alpha = 10$ is an empirically determined penalty coefficient,

$$\begin{aligned}\mathscr{L}_{\text{Disc}} \quad = \quad & \mathbb{E}_{predicted}\left[Disc\left(\tilde{X}_{pred}^{1w}\right)\right] + \mathbb{E}_{reconstructed}\left[Disc\left(\tilde{X}_T^{1w}\right)\right] \\ & -2\mathbb{E}_{real}\left[Disc\left(X_T^{1w}\right)\right] + \alpha \cdot l_{gp-Disc}.\end{aligned} \tag{4}$$

4.4. Perceptual loss

We also examine the potential of PL, described earlier in Section 2, as a potential means of improving results. A pretrained network is required to extract mid-layer feature maps of a predicted follow-up volume and 1-week follow-up target. We train an auto-encoder with the reconstruction objective of an input NCCT volume. Then, we use the encoder part of the model as a perceptual network P, with the loss $\mathscr{L}_{\text{PL}}$ computed as the squared and normalised Euclidean distance between the feature representation of the target volume X_T^{1w} and predicted volume $\tilde{X}_{pred}^{1w}$,

$$\mathscr{L}_{\text{PL}} = \sum_{i=1}^{L} \left\| P_i\left(X_T^{1w}\right) - P_i\left(\tilde{X}_{pred}^{1w}\right) \right\|_2, \tag{5}$$

where L is the number of hidden layers of P network.

5. Experimental setup

The architecture of the proposed method includes three encoders (E_1, E_2, E_3) which are identical, and each of them consists of four 3D convolutional layers, followed by batch normalisation (except the first layer) and a LeakyReLU activation to extract features. Then, channel-wise (cSE) and spatial (sSE) Squeeze and Excitation attentional blocks (Hu et al., 2018; Roy et al., 2018) are applied to recalibrate the high-level features. Finally, a FC layer in the encoder maps these recalibrated features to the latent space which becomes the output of the encoder.

Decoder D consists of a FC layer which projects latent code to the spatial space and four 3D convolutional blocks. The first block contains a transpose convolution layer followed by batch normalisation and a ReLU activation. Each of the following two blocks contains an upsample layer followed by a 3D convolution, batch normalisation and ReLU activation. The final block contains an upsample layer, two 3D convolution layers and tanh activation.

Data Registration – We registered our MRCLEAN data to a CT template (Rorden et al., 2012) in order to facilitate pixel-wise prediction. Following Muschelli (2019) and Rorden et al. (2012), we applied symmetric normalisation to perform affine and deformable transformations, with mutual information as the optimisation metric.

Implementation Details – We divided the data set into three subsets, training (70 %, 350 patients – 155 with EVT and 195 w/o EVT), validation (15 %, 75 patients – 40 with EVT and 35 w/o EVT) and testing (15 %, 75 patients – 38 with EVT and 37 w/o EVT). However, during the 1-week follow-up scan prediction stage, only the patients who have 1-week follow-up NCCT scans were used. During training, the Adam optimizer was used with a learning rate of 0.0001 and cosine annealing scheduler, and batch size was set to 24 to train over 200 epochs (in the GAN methods it was set to 500). In the second stage, the SGD optimizer applied a learning rate of 0.0003 and cosine annealing scheduler, and batch size was set to 16 to train over 100 epochs. The experiments were implemented in PyTorch on a single NVIDIA P100 GPU 16 GB.

We compare the performance of our proposed method against the following methods (often with some variances to allow for as close a direct comparison as possible): .

- Samak et al. (2020) – we retrain their 3D CNN model on our registered MR CLEAN dataset from scratch. This is the only existing study that performs mRS score prediction by using 3D NCCT data from MR CLEAN. This presents an apple-to-apple comparison to FeMA for the mRS score prediction task.
- Bacchi et al. (2019) – we implement our version of their 3D CNN model and train it on our MR CLEAN dataset from scratch. This is one of the closest works to ours that predicts dichotomised mRS scores from 3D NCCT volumes for patients who underwent thrombolysis treatment (rather than thrombectomy).
- *BE* – a basic encoder model used only for mRS score prediction as a baseline method.
- *AE* – a classic auto-encoder with a dense bottleneck, trained with the objective of follow-up volume prediction from a baseline scan. This allows us to show the impact of our feature similarity losses and two-step training of 1-week follow-up scan prediction.
- *AE* $_{WGAN}$ – a WGAN extension of the classic auto-encoder, such as in Bowles et al. (2018); Kwon et al. (2019). This becomes our baseline WGAN-based approach that enables us to observe the impact of using WGAN in U-Net$_{WGAN}$ and FeMA$_{WGAN}$.
- *AE*$_{PL}$ – an extension of the classic auto-encoder where a perceptual loss is added, similar to Johnson et al. (2016); Armanious et al. (2020).
- *U-Net* and *U-Net* $_{WGAN}$ – these are the original U-Net (Çiçek et al., 2016) and its extension with WGAN (Wegmayr et al., 2019) to predict a follow-up scan from a baseline. To enable these methods to predict the mRS score, Stage-2 of FeMA is added to them. The rationale behind comparing to U-Net models is their broad success in medical image analysis, such as in disease progression (Wegmayr et al., 2019).
- Nishi et al. (2020) – a 3D residual U-Net model that performs dichotomised mRS score classification and final lesion segmentation tasks. We implemented their model as close to what is described in their paper.
- *FeMA*, *FeMA* $_{PL}$ and *FeMA* $_{WGAN}$ – our proposed method and its extension to perceptual loss and WGAN respectively.

6. Results

6.1. Predicting 1-week follow-up scans

We first report in Table 1 the results for Stage-1 of our proposed

Table 1
The **test set** results of the models that can predict the 1-week follow-up scan. MAE: Mean Absolute Error, MSE: Mean Squared Error, SSIM: Structural Similarity, MS-SSIM: Multi-Scale Structural Similarity. Shaded background section indicates models that are based on U-Net. CI is confidence interval.

Methods	MAE↓ (95 % CI)	MSE↓ (95 % CI)	SSIM↑ (95 % CI)	MS-SSIM↑ (95 % CI)
AE	0.1453 (0.1392–0.1521)	0.0514 (0.0470–0.0565)	0.4852 (0.4691–0.5007)	0.8243 (0.8120–0.8361)
AE_{WGAN} (Kwon et al., 2019)	0.1269 (0.1187–0.1350)	0.0584 (0.0518–0.0648)	0.5078 (0.4889–0.5250)	0.8357 (0.8204–0.8490)
AE_{PL}(Johnson et al., 2016),(Armanious et al., 2020)	0.1259 (0.1185–0.1343)	0.0505 (0.0445–0.0573)	0.5185 (0.4990–0.5368)	0.8412 (0.8258–0.8555)
$FeMA_{WGAN}$	0.1240 (0.1163–0.1326)	0.0544 (0.0484–0.0611)	0.5148 (0.4958–0.5329)	0.8398 (0.8250–0.8535)
$FeMA_{PL}$	0.1274 (0.1198–0.1354)	0.0514 (0.0456–0.0579)	0.5182 (0.4993–0.5361)	0.8415 (0.8268–0.8549)
FeMA	**0.1226 (0.1153–0.1306)**	**0.0492 (0.0433–0.0558)**	**0.5229 (0.5039–0.5412)**	**0.8441 (0.8293–0.8577)**
(Nishi et al., 2020)	0.1234 (0.1157–0.1321)	0.0487 (0.0429–0.0552)	**0.5482 (0.5248–0.5691)**	0.8574 (0.8409–0.8723)
U-Net (Çiçek et al., 2016)	0.1240 (0.1168–0.1319)	**0.0477 (0.0422–0.0540)**	0.5447 (0.5227–0.5647)	0.8564 (0.8403–0.8707)
$U\text{-}Net_{WGAN}$ (Wegmayr et al., 2019)	**0.1187 (0.1111–0.1274)**	0.0530 (0.0467–0.0601)	0.5478 (0.5256–0.5681)	**0.8582 (0.8409–0.8732)**

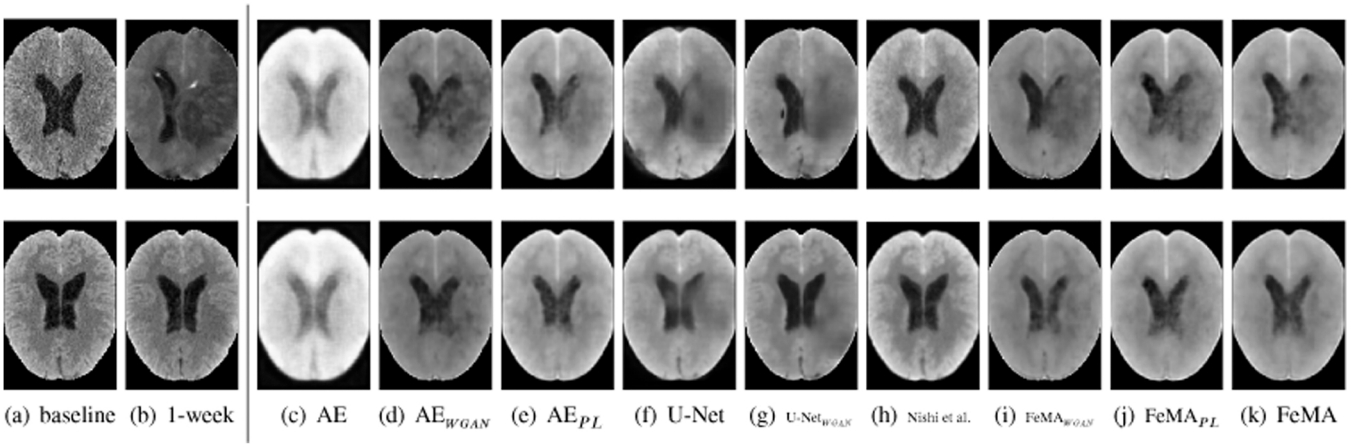

Fig. 5. Qualitative results of different models on predicting ischaemic stroke evolution from baseline to follow-up scan.

approach to measure the quality of the follow-up scan prediction from a baseline scan. Four metrics were applied for this assessment: the mean absolute error (MAE) and mean squared error (MSE) metrics quantify the error voxel-wise between the predicted and actual follow-up scans, and the structural similarity index (SSIM) and multi-scale structural similarity (MS-SSIM) metrics perceptually calculate the quality of the predicted follow-up volume compared to the 1-week follow-up scan target. Note, the differences between the error or similarity values may appear insignificant due to the limited ± 1 range of voxel values, however even such small differences represent large differences, which are manifested when observing the qualitative results (as in Fig. 5.

Table 1 shows U-Net based methods outperform other methods across the four metrics, for example at 0.1187 MAE and 0.8582 MS-SSIM for $U\text{-}Net_{WGAN}$. This can be attributed to the skip connections in U-Net architecture, which pass high-level information from encoder to decoder. This enables the U-Net to reconstruct better follow-up volumes, but it does not allow it to predict the informative follow-up features needed for mRS score prediction as demonstrated later in Section 6.2. Our proposed FeMA obtains competitive results compared to U-Net, and the best results compared to methods that do not have skip connections, at 0.1226 MAE, 0.0492 MSE, 0.5229 SSIM and 0.8441 MS-SSIM.

Interestingly, the WGAN option improves the performance of AE in all metrics and U-Net very marginally in MAE and SSIM metrics, however, it narrowly fails to have such an effect on FeMA. The reason is that the contribution of WGAN losses and FeMA losses to the model training need to be regulated for better performance. Similarly, $FeMA_{PL}$ is marginally below FeMA in all metrics, whereas it improves on AE_{PL}. Fig. 5 shows qualitative results for predictions of the 1-week scan for two example cases, one along each row. For each example, 5(a) is the baseline hospital admission scan, 5(b) is the actual 1-week scan and the remaining are the predicted 1-week follow-up volumes of the methods we compare against and our variations of the proposed method. Although the predicted follow-up volume of FeMA (Fig. 5(k)) successfully shows the stroke evolution, the U-Net models (Fig. 5(f) and 5(g)) appear to highlight the stroke lesion better. WGAN and PL improve the visual quality of the follow-up volume prediction, especially when we apply them as an extension of AE (Figs. 5(d) and 5(e)). Again note, even though there is a large difference between the predicted follow-up volume of AE and AE_{PL} compared to the actual 1-week follow-up volume Fig. 5(b) as seen in Fig. 5(c) and 5(e) respectively, the difference in MSE metric is only 0.0009. The result of Nishi et al. (2020) seems to indicate that the baseline scan is reproduced as the predicted 1-week follow-up scan Fig. 6.

6.2. Predicting mRS scores

For the prediction of mRS scores in Stage-2, three metrics were used as evaluation measures, Accuracy, *F1-score* and Area Under ROC Curve (AUC), to quantify different aspects of classification performance.

Our comparative results for different methods, including cases where different combinations of the information available were used (i.e. combinations of baseline scan X^{bl}, predicted follow-up features z^{1w}_{pred}, and volume $\widetilde{X}^{1w}_{pred}$) are listed in Tables 2 and 3.

Our proposed method achieves the best results and significantly improves the performance over non-GAN and GAN based methods, with 0.60 *F1-score* and 0.79 AUC for FeMA (row 17, Table 2), and 0.58 *F1-score* and 0.76 AUC for $FeMA_{WGAN}$ (row 7, Table 3). Although U-Net was the overall best in follow-up volume prediction, it does not reach the same accolade in this task due to the following reasons. Since the U-Net model utilises skip connections, it does not allow to compute feature similarity on embedded features and so it cannot make use of z^{1w}_{pred}, while FeMA benefits from using z^{1w}_{pred} and achieves a significantly better result at 0.79 AUC. Additionally, U-Net is only supervised by voxel-wise prediction loss which is not sufficient to learn the fine-grained details of the

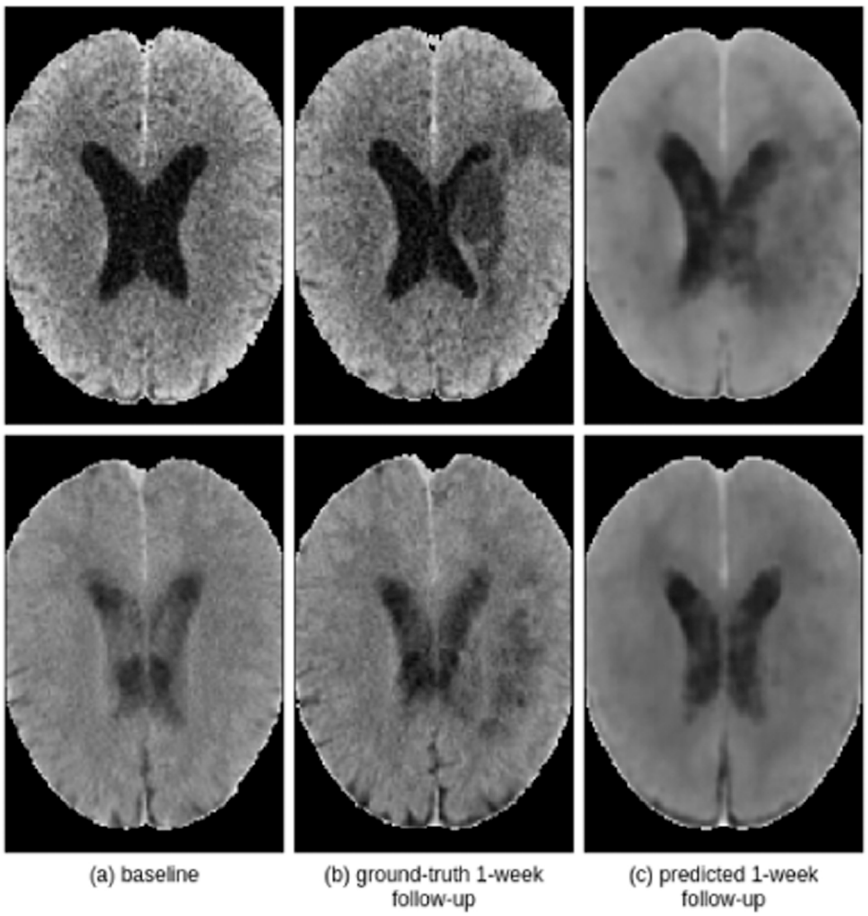

Fig. 6. Qualitative results of FeMA model on predicting ischaemic stroke evolution from baseline to follow-up scan. The first row is an example of a patient did not underwent EVT and the second row is a patient had EVT.

Table 2

Results for dichotomised mRS score classification for different combinations of information from X^{bl}, z^{1w}_{pred} and $\widetilde{X}^{1w}_{pred}$, where applicable. CI is confidence interval. The best and the second best results are shown in bold and underlined respectively.

Method	Row	X^{bl}	z^{1w}_{pred}	$\widetilde{X}^{1w}_{pred}$	Accuracy (95 % CI)	*F1-score* (95 % CI)	AUC (95 % CI)
(Samak et al., 2020)	1	✓	✗	✗	0.72 (0.62–0.82)	0.33 (0.09–0.53)	0.63 (0.44–0.81)
(Bacchi et al., 2019)	2	✓	✗	✗	0.75 (0.65–0.85)	0.40 (0.16–0.60)	0.66 (0.48–0.80)
BE	3	✓	✗	✗	0.66 (0.55–0.77)	0.37 (0.16–0.56)	0.68 (0.54–0.82)
AE	4	✓	✓	✓	0.56 (0.44–0.67)	0.36 (0.18–0.52)	0.63 (0.47–0.79)
	5	✓	✗	✓	0.71 (0.60–0.82)	0.49 (0.29–0.67)	0.65 (0.48–0.81)
	6	✓	✓	✗	0.65 (0.54–0.76)	0.44 (0.25–0.62)	0.67 (0.50–0.83)
AE_{PL} such as (Johnson et al., 2016), (Armanious et al., 2020)	7	✓	✓	✓	0.74 (0.62–0.83)	0.49 (0.26–0.67)	<u>0.77 (0.62–0.89)</u>
	8	✓	✗	✓	0.72 (0.61–0.83)	0.50 (0.29–0.68)	0.66 (0.48–0.82)
	9	✓	✓	✗	0.67 (0.56–0.78)	0.37 (0.16–0.56)	0.68 (0.52–0.82)
U-Net (Çiçek et al., 2016)	10	✓	✗	✓	0.69 (0.58–0.80)	0.35 (0.24–0.46)	0.65 (0.54–0.76)
(Nishi et al., 2020)	11	✓	✗	✗	0.62 (0.51–0.73)	0.13 (0.00–0.29)	0.45 (0.31–0.64)
FeMA $_{PL}$	12	✓	✓	✓	0.68 (0.56–0.79)	0.47 (0.26–0.65)	0.75 (0.60–0.87)
	13	✓	✗	✓	0.68 (0.56–0.79)	0.44 (0.23–0.62)	0.67 (0.51–0.82)
	14	✓	✓	✗	0.77 (0.68–0.87)	0.33 (0.09–0.56)	0.68 (0.51–0.83)
FeMA	15	✓	✓	✓	0.74 (0.62–0.83)	<u>0.51 (0.29–0.69)</u>	<u>0.77 (0.62–0.90)</u>
	16	✓	✗	✓	<u>0.78 (0.68–0.88)</u>	0.47 (0.22–0.69)	0.71 (0.54–0.86)
	17	✓	✓	✗	**0.79 (0.69–0.88)**	**0.60 (0.37–0.76)**	**0.79 (0.63–0.92)**

1-week follow-up scan i.e.FeMA (row 16, Table 2) and $FeMA_{WGAN}$ (row 6, Table 3) exceed the results of U-Net even when using the same configuration, only combining $\widetilde{X}^{1w}_{pred}$ with X^{bl}, like U-Net. This demonstrates the effectiveness of our follow-up feature prediction which contains valuable information about the follow-up scan even though it is not visually obvious in the predicted follow-up volume.

Although AE_{PL} and AE_{WGAN} improve on AE's results quite markedly, corresponding increases are not always observed for FeMA. This occurs because the AE model only utilises MSE loss on the follow-up scan volume during training, so the additional supervision offered by the PL and WGAN processes positively affect the follow-up volume prediction and so the mRS score prediction. However, FeMA is supervised by a feature consistency loss which performs better than the PL or WGAN losses. FeMA models are more effective when using z^{1w}_{pred} with X^{bl} (row

Table 3

Results of WGAN based methods for dichotomised mRS score classification for different combinations of information from X^{bl}, z^{1w}_{pred} and $\widetilde{X}^{1w}_{pred}$, where applicable. CI is confidence interval. The best and the second best results are shown in bold and underlined respectively.

Method	Row	X^{bl}	z^{1w}_{pred}	$\widetilde{X}^{1w}_{pred}$	Accuracy (95 % CI)	*F1-score* (95 % CI)	AUC (95 % CI)
U-Net$_{WGAN}$ (Wegmayr et al., 2019)	1	✓	✗	✓	<u>0.72 (0.62–0.83)</u>	0.33 (0.22–0.44)	0.63 (0.52–0.74)
AE$_{WGAN}$ (Kwon et al., 2019)	2	✓	✓	✓	0.72 (0.61–0.82)	0.50 (0.29–0.68)	**0.76 (0.60–0.89)**
	3	✓	✗	✓	0.65 (0.54–0.76)	0.42 (0.22–0.59)	0.70 (0.55–0.84)
	4	✓	✓	✗	<u>0.72 (0.62–0.82)</u>	0.41 (0.18–0.61)	0.74 (0.60–0.87)
FeMA $_{WGAN}$	5	✓	✓	✓	0.71 (0.60–0.81)	<u>0.55 (0.36–0.71)</u>	<u>0.76 (0.60–0.90)</u>
	6	✓	✗	✓	0.70 (0.61–0.81)	0.32 (0.09–0.53)	0.73 (0.57–0.87)
	7	✓	✓	✗	**0.82 (0.72–0.90)**	**0.58 (0.32–0.76)**	0.75 (0.59–0.89)

Table 4

Examining the impact of different losses on 1-week follow-up scan prediction. $l_{2,rec}$, $l_{3,sim}$ and $l_{4,con}$ are reconstruction loss, feature similarity loss and consistency loss respectively. CI is confidence interval.

Methods	MAE↓ (95 % CI)	MSE↓ (95 % CI)	SSIM↑ (95 % CI)	MS-SSIM↑ (95 % CI)
FeMA$_{l_{2,rec}}$	0.1251 (0.1177–0.1332)	0.0500 (0.0441–0.0565)	0.5205 (0.5011–0.5387)	0.8419 (0.8273–0.8556)
FeMA$_{l_{3,sim}}$	0.1237 (0.1161–0.1319)	0.0494 (0.0436–0.0559)	0.5228 (0.5039–0.5409)	0.8431 (0.8292–0.8575)
FeMA$_{l_{2,rec}+l_{3,sim}}$	0.1237 (0.1163–0.1318)	0.0493 (0.0435–0.0558)	0.5225 (0.5036–0.5407)	0.8437 (0.8290–0.8572)
FeMA$_{l_{2,rec}+l_{3,sim}+l_{4,con}}$	**0.1226 (0.1153–0.1306)**	**0.0492 (0.0433–0.0558)**	**0.5229 (0.5039–0.5412)**	**0.8441 (0.8293–0.8577)**

Table 5

Examining the impact of different losses on mRS score prediction. $l_{2,rec}$, $l_{3,sim}$ and $l_{4,con}$ are reconstruction loss, feature similarity loss and consistency loss respectively. CI is confidence interval. The best and the second best results are shown in bold and underlined respectively.

Methods	Accuracy (95 % CI)	*F1-score* (95 % CI)	AUC (95 % CI)
FeMA$_{l_{2,rec}}$	0.75 (0.65–0.85)	0.44 (0.21–0.65)	0.68 (0.52–0.82)
FeMA$_{l_{3,sim}}$	<u>0.79 (0.69–0.89)</u>	0.44 (0.17–0.67)	0.74 (0.57–0.88)
FeMA$_{l_{2,rec}+l_{3,sim}}$	0.76 (0.67–0.85)	<u>0.51 (0.29–0.70)</u>	<u>0.77 (0.59–0.90)</u>
FeMA$_{l_{2,rec}+l_{3,sim}+l_{4,con}}$	**0.79 (0.69–0.88)**	**0.60 (0.37–0.76)**	**0.79 (0.63–0.92)**

17, Table 2) than using $\widetilde{X}^{1w}_{pred}$ with X^{bl} (row 16, Table 2). These show that the prediction of follow-up features is more efficient than the prediction of follow-up volume because during the mapping process between the baseline scan and the follow-up feature prediction, the model focuses on predicting only informative features. On the other hand, the inclusion of $\widetilde{X}^{1w}_{pred}$ with z^{1w}_{pred} and X^{bl} (row 15) reduces the performance of FeMA (row 17) and AE (row 6) i.e.from 0.79 to 0.77 in AUC for FeMA (rows 17 and 15 respectively) while increasing the mRS score prediction capabilities of AE$_{PL}$, FeMA$_{PL}$, A$_{WGAN}$ and FeMA$_{WGAN}$ (i.e., see rows 5 and 7 in Table 3). For example the performance of FeMA$_{PL}$ (row 14, Table 2) increases from 0.68 to 0.75 AUC in FeMA$_{PL}$ (row 12, Table 2). The reason is that fusing the volumes and features is not optimal, so a better volume and future fusion method needs to be proposed to enhance the predictive performance of the models for mRS scores.

6.3. Ablation study

We conducted several experiments to assess the effects of our loss functions in FeMA - based on the combinations of the reconstruction loss $l_{2,rec}$, feature similarity loss $l_{3,sim}$ and feature consistency loss $l_{4,con}$.

The results are shown in Table 4 for follow-up volume prediction and in Table 5 for mRS score prediction. The importance of $l_{3,sim}$ and $l_{4,con}$ can be seen in both tables. For example, when $l_{2,rec}$ or $l_{3,sim}$ alone are applied, the model reaches 0.68 and 0.74 AUC respectively, but when we combine both $l_{2,rec}$ and $l_{3,sim}$, the results are improved to 0.77 AUC. Further, adding feature consistency loss alongside the feature similarity and reconstruction loss increase the performance of FeMA to 0.79 AUC. Although the difference in follow-up prediction metrics is marginal between different loss combinations in Table 4, the impact on the mRS score prediction is significant as observed in Table 5. For example, when the difference in MSE is 0.0008 between FeMA$_{l_{2,rec}}$ and FeMA$_{l_{2,rec}+l_{3,sim}+l_{4,con}}$, the resulting difference in corresponding AUC values is 0.11.

Furthermore, we investigated the effect of clinical records on prediction performance. Alongside the treatment information, we added clinical information, such as patient demographics (e.g., age, gender), medical records (e.g., hypertension, glucose level), and stroke-related information (e.g., baseline NIHS, symptom side) into our FeMA model training. This improved FeMA's performance from 0.79 to 0.82 (0.74–0.90) in accuracy and 0.79–0.83 (0.70–0.93) in AUC score, however, there was a slight decrease in F1-score from 0.60 to 0.58 (0.33–0.77). This result shows that clinical information can have a positive impact on the prediction of mRS scores.

7. Conclusions

In this paper, we presented a novel framework to estimate the 1-week follow-up scan and mRS score as ischaemic stroke treatment outcome. For the prediction of the follow-up scan, we trained our model to predict ischaemic stroke evolution without voxel-wise supervision for the 1-week follow-up. We added feature similarity and consistency supervision to obtain better follow-up scan representation from the baseline scan. For the estimation of the mRS score, we combined the predicted follow-up scan features and volume with the baseline scan to arrive at more accurate predictions.

Our results demonstrate that the proposed model obtains competitive results compared to U-Net based methods and best results compared to a number of non-U-Net based methods in the prediction of the follow-up volume. Further, our method significantly outperforms U-Net and several non-U-Net based methods in mRS score prediction.

For future work, we plan to extend the current study by developing an unsupervised approach, while additionally using both 24-hour follow-up scans and clinical information for improve mRS score prediction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the MR CLEAN Trial Principal Investigators: Prof Aad van der Lugt, Prof Diederik W.J. Dippel, Prof. Charles B.L.M. Majoie, Prof. Yvo B. W.E.M. Roos, Prof. Wim H. van Zwam and Prof. Robert J. van Oostenbrugge for providing the data. Zeynel Samak gratefully acknowledges funding from the Republic of Turkey Ministry of National Education Grant MoNE-1416/YLSY).

CRediT authorship contribution statement

Zeynel A. Samak: Conceptualization, Methodology, Software, Writing – original draft preparation, Formal analysis, Investigation. **Philip Clatworthy:** Writing – review & editing, Supervision. **Majid Mirmehdi:** Resources, Writing – original draft preparation, Writing – review & editing, Supervision, Project administration.

References

Abraham, N., Khan, N.M., 2019. A novel focal tversky loss function with improved attention U-Net for lesion segmentation, in: ISBI, IEEE. pp. 683–687.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks. In: Proceedings of the International Conference on Machine Learning, PMLR. pp.214–223.

Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., Yang, B., 2020. MedGAN: medical image translation using GANs. Comput. Med. Imaging Graph. 79, 101684.

Asadi, H., Dowling, R., Yan, B., Mitchell, P., 2014. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. PLOS One 9, e88225.

Bacchi, S., Zerner, T., Oakden-Rayner, L., Kleinig, T., Patel, S., Jannes, J., 2019. Deep learning in the prediction of ischaemic stroke thrombolysis functional outcomes: a pilot study. Acad. Radiol. 27, e19–e23.

Bentley, P., Ganesalingam, J., CarltonJones, A.L., Mahady, K., Epton, S., Rinne, P., Sharma, P., Halse, O., Mehta, A., Rueckert, D., 2014. Prediction of stroke thrombolysis outcome using CT brain machine learning. NeuroImage: Clin. 4, 635–640.

Berkhemer, O.A., Fransen, P.S., Beumer, D., van den Berg, L.A., Lingsma, H.F., Yoo, A.J., Schonewille, W.J., Vos, J.A., Nederkoorn, P.J., Wermer, M.J., van Walderveen, M.A., Staals, J., Hofmeijer, J., van Oostayen, J.A., Lycklama à Nijeholt, G.J., Boiten, J., Brouwer, P.A., Emmer, B.J., de Bruijn, S.F., van Dijk, L.C., Kappelle, L.J., Lo, R.H., van Dijk, E.J., de Vries, J., de Kort, P.L., van Rooij, W.J.J., van den Berg, J.S., van Hasselt, B.A., Aerden, L.A., Dallinga, R.J., Visser, M.C., Bot, J.C., Vroomen, P.C., Eshghi, O., Schreuder, T.H., Heijboer, R.J., Keizer, K., Tielbeek, A.V., den Hertog, H. M., Gerrits, D.G., van den Berg-Vos, R.M., Karas, G.B., Steyerberg, E.W., Flach, H.Z., Marquering, H.A., Sprengers, M.E., Jenniskens, S.F., Beenen, L.F., van den Berg, R., Koudstaal, P.J., van Zwam, W.H., Roos, Y.B., van der, A., van Oostenbrugge, R.J., Majoie, C.B., Dippel, D.W., 2015. A randomized trial of intraarterial treatment for acute ischemic stroke. N. Engl. J. Med. 372, 11–20.

Boers, A., Marquering, H., Jochem, J., Besselink, N., Berkhemer, O., Lugt, A., v.d., Beenen, L., Majoie, C., 2013. Automated cerebral infarct volume measurement in follow-up noncontrast ct scans of patients with acute ischemic stroke. Am. J. Neuroradiol. 34, 1522–1527.

Bowles, C., Gunn, R., Hammers, A., Rueckert, D., 2018. Modelling the progression of Alzheimeras disease in MRI using generative adversarial networks, Medical Imaging 2018: Image Processing, International Society for Optics and Photonics. p. 105741K.

Chawla, M., Sharma, S., Sivaswamy, J., Kishore, L., 2009. A method for automatic detection and classification of stroke from brain CT images. In: Proceedings of the IEEEMBS, IEEE. pp. 3581–3584.

Chen, H., Wang, Y., Zheng, K., Li, W., Chang, C.T., Harrison, A.P., Xiao, J., Hager, G.D., Lu, L., Liao, C.H., et al., 2020. Anatomy-aware siamese network: Exploiting semantic asymmetry for accurate pelvic fracture detection in x-ray images. In: Proceedings of the European Conference on Computer Vision. Springer, pp. 239–255.

Choi, Y., Kwon, Y., Lee, H., Kim, B.J., Paik, M.C., Won, J.H., 2016. Ensemble of deep convolutional neural networks for prognosis of ischemic stroke. In: Crimi, A., Menze, B., Maier, O., Reyes, M., Winzeck, S., Handels, H. (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, Cham, pp. 231–243.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Proceedings of the International Conference on Medical image Computing and Computer-Assisted Intervention, Springer. pp. 424–432.

Clèrigues, A., Valverde, S., Bernal, J., Freixenet, J., Oliver, A., Lladó, X., 2020. Acute and sub-acute stroke lesion segmentation from multimodal MRI. Comput. Methods Prog. Biomed. 194, 105521.

Clèrigues, A., et al., 2018. SUNet: a deep learning architecture for acute stroke lesion segmentation and outcome prediction in multimodal MRI. arXiv preprint arXiv: 1810.13304.

Darehed, D., Blom, M., Glader, E.L., Niklasson, J., Norrving, B., Eriksson, M., 2020. In-hospital delays in stroke thrombolysis: every minute counts. Stroke 51, 2536–2539.

Ernst, M., Boers, A.M., Aigner, A., Berkhemer, O.A., Yoo, A.J., Roos, Y.B., Dippel, D.W., van der Lugt, A., van Oostenbrugge, R.J., van Zwam, W.H., et al., 2017. Association of computed tomography ischemic lesion location with functional outcome in acute large vessel occlusion ischemic stroke. Stroke 48, 2426–2433.

Fransen, P.S., Beumer, D., Berkhemer, O.A., Van Den Berg, L.A., Lingsma, H.A., van Zwam, W.H., van Oostenbrugge, R.J., Roos, Y.B., Majoie, C.B., et al., 2014. MR CLEAN, a multicenter randomized clinical trial of endovascular treatment for acute ischemic stroke in the Netherlands: study protocol for a randomized controlled trial. Trials 15, 343.

Gautam, A., Raman, B., 2021. Towards effective classification of brain hemorrhagic and ischemic stroke using CNN.Biomedical. Signal Process. Control 63, 102178.

Gholizadeh-Ansari, M., Alirezaie, J., Babyn, P., 2020. Deep learning for low-dose ct denoising using perceptual loss and edge detection layer. J. Digit. Imaging 33, 504–515.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A., 2017. Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028.

Heo, J., Yoon, J., Park, H.J., Kim, Y.D., Nam, H.S., Heo, J.H., 2018. Machine learning-based model can predict stroke outcome. Stroke 49 (A194-A194).

Hilbert, A., Ramos, L., van Os, H., Olabarriaga, S., Tolhuisen, M., Wermer, M., Barros, R., van der Schaaf, I., Dippel, D., Roos, Y., et al., 2019. Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke. Comput. Biol. Med., 103516

Hu, J., Shen, L., Sun, G., 2018.Squeeze-and-excitation networks. In: Proceedings of the Computer Vision and Pattern Recognition, IEEE Computer Society. pp. 7132–7141.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the European Conference on Computer Cision, Springer. pp. 694–711.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78.

Kemmling, A., Flottmann, F., Forkert, N.D., Minnerup, J., Heindel, W., Thomalla, G., Eckert, B., Knauth, M., Psychogios, M., Langner, S., Fiehler, J., 2015. Multivariate dynamic prediction of ischemic infarction and tissue salvage as a function of time and degree of recanalization. J. Cereb. Blood Flow. Metab. 35, 1397–1405. PMID: 26154867.

Kwon, G., Han, C., Kim, D.s., 2019. Generation of 3d brain mri using auto-encoding generative adversarial networks. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp.118–126.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the Computer Vision and Pattern Recognition. pp. 2980–2988.

Lucas, C., Kemmling, A., Bouteldja, N., Aulmann, L.F., Madany Mamlouk, A., Heinrich, M.P., 2018. Learning to predict ischemic stroke growth on acute ct perfusion data by interpolating low-dimensional shape representations. Front. Neurol. 9, 989.

Maier, O., Handels, H., 2016. Predicting stroke lesion and clinical outcome with random forests. In: Proceedings of the International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer. pp. 219–230.

Maier, O., Wilms, M., vonderGablentz, J., Krämer, U., Handels, H., 2014. Ischemic stroke lesion segmentation in multi-spectral MR images with support vector machine classifiers. In: Proceedings of the Medical Imaging 2014: Computer-Aided Diagnosis, ISOP. p. 903504.

Mckinley, R., Häni, L., Gralla, J., El-Koussy, M., Bauer, S., Arnold, M., Fischer, U., Jung, S., Mattmann, K., Reyes, M., Wiest, R., 2017. Fully automated stroke tissue estimation using random forest classifiers (FASTER). J. Cereb. Blood Flow. Metab. 37, 2728–2741.

Muschelli, J., 2019. Recommendations for processing head CT data. Front. Neuroinformatics 13, 61.

Nishi, H., Oishi, N., Ishii, A., Ono, I., Ogura, T., Sunohara, T., Chihara, H., Fukumitsu, R., Okawa, M., Yamana, N., et al., 2020. Deep learning–derived high-level neuroimaging features predict clinical outcomes for large vessel occlusion. Stroke 51, 1484–1492.

van Os, H.J.A., Ramos, L.A., Hilbert, A., van Leeuwen, M., van Walderveen, M.A.A., Kruyt, N.D., Dippel, D.W.J., Steyerberg, E.W., van der Schaaf, I.C., Lingsma, H.F., Schonewille, W.J., Majoie, C.B.L.M., Olabarriaga, S.D., Zwinderman, K.H., Venema, E., Marquering, H.A., Wermer, M.J.H., the MR CLEAN Registry Investigators, 2018. Predicting outcome of endovascular treatment for acute ischemic stroke: potential value of machine learning algorithms. Front. Neurol. 9, 784.

Osama, S., Zafar, K., Sadiq, M.U., 2020. Predicting clinical outcome in acute ischemic stroke using parallel multi-parametric feature embedded siamese network. Diagnostics 10, 858.

Pihlgren, G.G., Sandin, F., Liwicki, M., 2020. Improving image autoencoder embeddings with perceptual loss. In: Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), IEEE. pp.1–7.

Pinto, A., Pereira, S., Meier, R., Wiest, R., Alves, V., Reyes, M., Silva, C.A., 2021. Combining unsupervised and supervised learning for predicting the final stroke lesion. Med. Image Anal. 69, 101888.

Pinto, A., Pereira, S., Meier, R., Alves, V., Wiest, R., Silva, C.A., Reyes, M., 2018. Enhancing clinical MRI perfusion maps with data-driven maps of complementary nature for lesion outcome prediction. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp.107–115.

Rafael-Palou, X., Aubanell, A., Bonavita, I., Ceresa, M., Piella, G., Ribas, V., GonzálezBallester, M.A., 2021. Re-Identification and growth detection of pulmonary

nodules without image registration using 3D siamese neural networks. Med. Image Anal. 67, 101823.
Renowden, S., 2014. Imaging in stroke and vascular disease—part 1: ischaemic stroke. Pract. Neurol. 14, 77–87.
Robben, D., Boers, A.M., Marquering, H.A., Langezaal, L.L., Roos, Y.B., van Oostenbrugge, R.J., van Zwam, W.H., Dippel, D.W., Majoie, C.B., van der Lugt, A., Lemmens, R., Suetens, P., 2020. Prediction of final infarct volume from native CT perfusion and treatment parameters using deep learning. Med. Image Anal. 59, 101589.
Rorden, C., Bonilha, L., Fridriksson, J., Bender, B., Karnath, H.O., 2012. Age-specific CT and MRI templates for spatial normalization. Neuroimage 61, 957–965.
Roy, A.G., Navab, N., Wachinger, C., 2018. Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In: Proceedings of the International Conference on Medical Image Computing and Computer-assisted intervention, Springer. pp. 421–429.
Samak, Z.A., Clatworthy, P., Mirmehdi, M., 2020. Prediction of thrombectomy functional outcomes using multimodal data. In: Proceedings of the Medical Image Understanding and Analysis, Springer International Publishing, Cham. pp. 267–279.
Saver, J.L., 2006. Time is brain—quantified. Stroke 37, 263–266.
Scalzo, F., Hao, Q., Alger, J.R., Hu, X., Liebeskind, D.S., 2012. Regional prediction of tissue fate in acute ischemic stroke. Ann. Biomed. Eng. 40, 2177–2187.
Stroke Association, 2018. State of the Nation: stroke statistics.⟨https://www.stroke.org.uk/resources/state-nation-stroke-statistics⟩ [Accessed Nov-2019].
Van Swieten, J., Koudstaal, P., Visser, M., Schouten, H., Van Gijn, J., 1988. Interobserver agreement for the assessment of handicap in stroke patients. Stroke 19, 604–607.
Venema, E., Mulder, M.J., Roozenbeek, B., Broderick, J.P., Yeatts, S.D., Khatri, P., Berkhemer, O.A., Emmer, B.J., Roos, Y.B., Majoie, C.B., Van Oostenbrugge, R.J., Van Zwam, W.H., Van Der Lugt, A., Steyerberg, E.W., Dippel, D.W., Lingsma, H.F., 2017. Selection of patients for intra-arterial treatment for acute ischaemic stroke: development and validation of a clinical decision tool in two randomised trials. BMJ 357.
Wang, G., Song, T., Dong, Q., Cui, M., Huang, N., Zhang, S., 2020. Automatic ischemic stroke lesion segmentation from computed tomography perfusion images by image synthesis and attention-based deep neural networks. Med. Image Anal. 65, 101787.
Wegmayr, V., Hörold, M., Buhmann, J.M., 2019.Generative aging of brain mr-images and prediction of alzheimer progression. in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer. pp.247–260.
Weyland, C.S., Papanagiotou, P., Schmitt, N., Joly, O., Bellot, P., Mokli, Y., Ringleb, P.A., Kastrup, A., Möhlenbruch, M.A., Bendszus, M., Nagel, S., Herweh, C., 2022. Hyperdense artery sign in patients with acute ischemic stroke-automated detection with artificial intelligence-driven software. Front. Neurol. 13.
WHO, 2018.The top 10 causes of death.⟨https://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death⟩ [Accessed Nov-2019].
Winzeck, S., Hakim, A., McKinley, R., Pinto, J.A.A.D.S.R., Alves, V., Silva, C., Pisov, M., Krivov, E., Belyaev, M., Monteiro, M., Oliveira, A., Choi, Y., Paik, M.C., Kwon, Y., Lee, H., Kim, B.J., Won, J.H., Islam, M., Ren, H., Robben, D., Suetens, P., Gong, E., Niu, Y., Xu, J., Pauly, J.M., Lucas, C., Heinrich, M.P., Rivera, L.C., Castillo, L.S., Daza, L.A., Beers, A.L., Arbelaezs, P., Maier, O., Chang, K., Brown, J.M., Kalpathy-Cramer, J., Zaharchuk, G., Wiest, R., Reyes, M., 2018. ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI. Front. Neurol. 9.
Yi, X., Walia, E., Babyn, P., 2019. Generative adversarial network in medical imaging: a review. Med. Image Anal. 58, 101552.