Computerized Medical Imaging and Graphics 90 (2021) 101926

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

Intra-domain task-adaptive transfer learning to determine acute ischemic stroke onset time

Haoyue Zhang [a,b,1], Jennifer S Polson [a,b,1], Kambiz Nael [c], Noriko Salamon [c], Bryan Yoo [c], Suzie El-Saden [d], Fabien Scalzo [e], William Speier [a,c], Corey W. Arnold [a,b,c,f,*]

[a] Computational Diagnostics Lab, University of California, Los Angeles, CA 90024, USA
[b] Department of Bioengineering, University of California, Los Angeles, CA 90024, USA
[c] Department of Radiology, University of California, Los Angeles, CA 90024, USA
[d] Department of Radiology, VA Phoenix Healthcare system, AZ 85012, USA
[e] Departments of Neurology and Computer Science, University of California, Los Angeles, CA 90024, USA
[f] Department of Pathology, University of California, Los Angeles, CA 90024, USA

ARTICLE INFO

Keywords:
Deep learning
Structural MRI
Acute ischemic stroke

ABSTRACT

Treatment of acute ischemic strokes (AIS) is largely contingent upon the time since stroke onset (TSS). However, TSS may not be readily available in up to 25% of patients with unwitnessed AIS. Current clinical guidelines for patients with unknown TSS recommend the use of MRI to determine eligibility for thrombolysis, but radiology assessments have high inter-reader variability. In this work, we present deep learning models that leverage MRI diffusion series to classify TSS based on clinically validated thresholds. We propose an intra-domain task-adaptive transfer learning method, which involves training a model on an easier clinical task (stroke detection) and then refining the model with different binary thresholds of TSS. We apply this approach to both 2D and 3D CNN architectures with our top model achieving an ROC-AUC value of 0.74, with a sensitivity of 0.70 and a specificity of 0.81 for classifying TSS < 4.5 h. Our pretrained models achieve better classification metrics than the models trained from scratch, and these metrics exceed those of previously published models applied to our dataset. Furthermore, our pipeline accommodates a more inclusive patient cohort than previous work, as we did not exclude imaging studies based on clinical, demographic, or image processing criteria. When applied to this broad spectrum of patients, our deep learning model achieves an overall accuracy of 75.78% when classifying TSS < 4.5 h, carrying potential therapeutic implications for patients with unknown TSS.

1. Introduction

Acute ischemic stroke (AIS) is a cerebrovascular disease accounting for 2.7 million deaths worldwide every year (Benjamin et al., 2019). Treatment of AIS is heavily dependent on the time since stroke onset (TSS); current clinical guidelines recommend thrombolytic therapies for AIS patients presenting within 4.5 h and endovascular thrombectomy for those presenting up to 24 h after onset. AIS without a clear TSS is relatively common, accounting for up to 25% of all AIS (Thomalla et al., 2014; Urrutia et al., 2018). Some reasons for unclear TSS include unwitnessed strokes, wake-up strokes, or unreliable reporting by patients. For this patient population, the most recent AHA guidelines recommend using MRI sequences to assess patient eligibility for thrombolytics (Powers et al., 2019).

Following the WAKE UP trial (Thomalla et al., 2018), which used DWI-FLAIR mismatch to select patients for extending the time window for intravenous thrombolysis, the use of MRI (FLAIR-DWI mismatch) is now recommended (level IIa) to identify unwitnessed AIS patients who may benefit from thrombolytic treatment (Powers et al., 2019). Specifically, diffusion-weighted imaging (DWI) displays increased signal in ischemic areas within minutes of stroke occurrence, while fluid-attenuated inversion recovery (FLAIR) imaging can show fluid accumulation after a few hours (Etherton et al., 2018), as shown in Fig. 1. A DWI-positive, FLAIR-negative mismatch can identify stroke lesions that could benefit from administration of thrombolytics. However, assessing this mismatch is subject to high variability compared across multiple readings and/or radiologists (Thomalla et al., 2011). Thus, determining stroke onset using imaging alone could increase the

* Corresponding author at: Computational Diagnostics Lab, University of California, Los Angeles, CA 90024, USA.
[1] Haoyue Zhang and Jennifer S Polson contributed equally.

https://doi.org/10.1016/j.compmedimag.2021.101926
Received 7 October 2020; Received in revised form 21 January 2021; Accepted 5 April 2021
Available online 24 April 2021

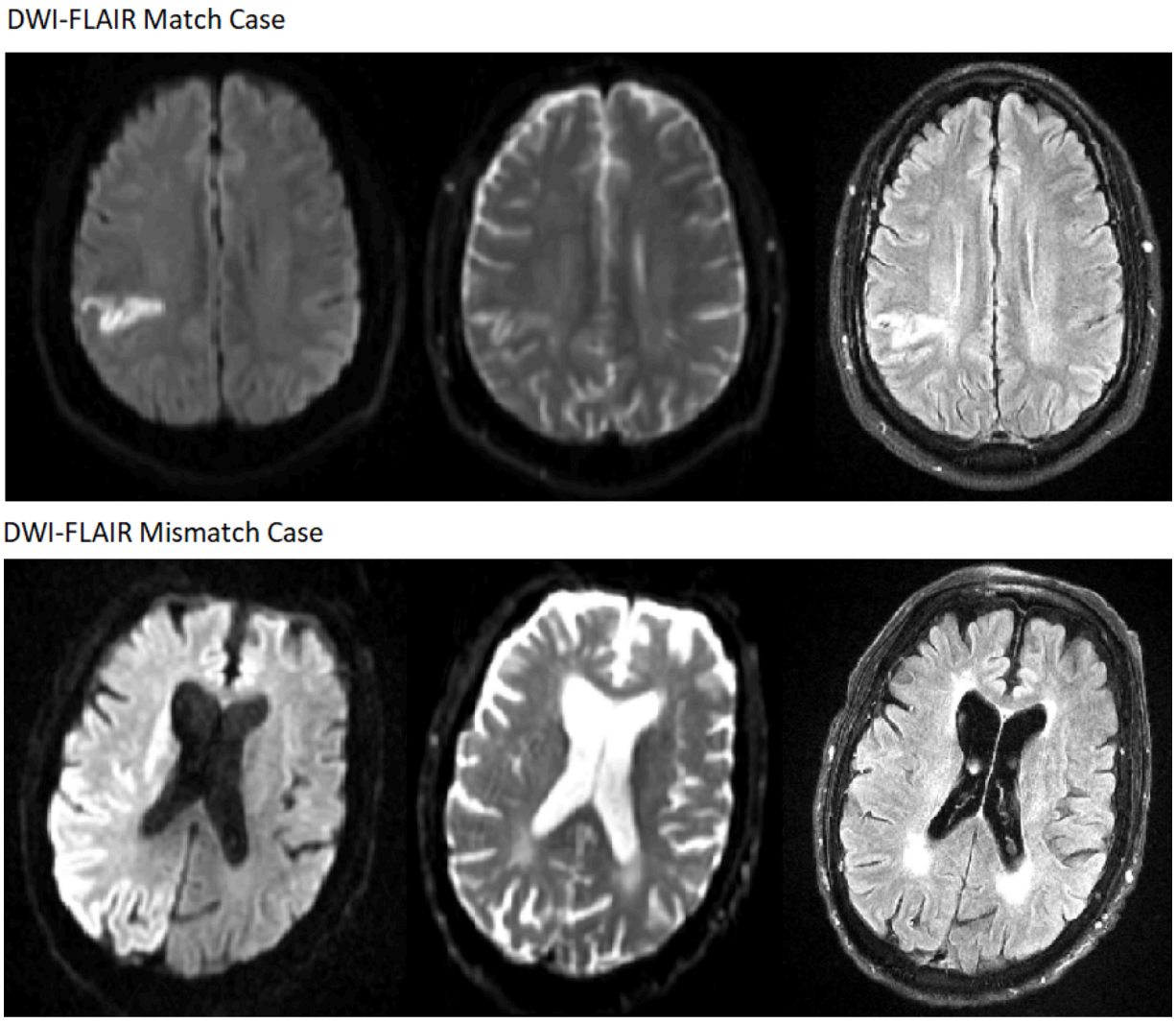

Fig. 1. Sample cases of DWI-FLAIR mismatch. Sequences from left to right: DWI b1000, DWI B0, FLAIR.

number of patients eligible to receive thrombolytic treatments, possibly improving their outcomes.

Several machine learning approaches have been used to determine stroke onset time in an automated fashion. These involve generating hand-crafted, radiomic, or deep learning-derived features from either clinical reports or images and then using these features as inputs to a variety of machine learning models (Ho et al., 2019, 2017; Lee et al., 2020). This feature extraction has typically relied on defined regions of interest, determined by applying an image threshold or using a parameter map. Limiting features to this immediate region may fail to capture imaging characteristics in the surrounding region, which could be crucial to informing TSS given the interconnected nature of cerebral blood flow (Bang et al., 2011). Moreover, previously published approaches have applied meticulous exclusion criteria, either by stroke location or imaging factors related to preprocessing; for these studies, as many as 40% of patients were ineligible for assessment (Lee et al., 2020).

Deep learning models have excelled in medical imaging for segmentation and classification tasks (Shin et al., 2016; Milletari et al., 2016; Chan et al., 2020; Nie et al., 2016; Winzeck et al., 2018). Specifically, convolutional neural networks (CNNs) have produced state-of-the-art results even in small datasets common in medical imaging research (Litjens et al., 2017). Convolutions, which aggregate pixel neighborhoods across layers, may occur in either two or three dimensions. While there has been a wide range of 2D CNNs applied to medical image tasks, 3D CNNs offer the added advantage of integrating information along the z-axis as well. The potential advantages of 3D convolutions come with a cost of increased model complexity, which generally requires a higher amount of data and computation power to train.

Due to the large number of parameters in a deep neural network, a high volume of data is typically required for training. For particularly complex classification tasks, transfer learning has been shown to achieve model convergence using less computation and boost performance in less time compared to training models from scratch (Pan and Yang, 2010). Transfer learning traditionally involves training a model on one dataset, then refining the model on another set of data for a different task. Cross-domain transfer learning involves training on data from a source domain, and using those learned weights in a model trained on data from a different target domain (Weiss et al., 2016), e.g., from the natural image domain to the medical image domain or from the CT image domain to the MR image domain. Many deep learning approaches applied to medical images have used established architectures pre-trained on large natural image datasets such as ImageNet (Russakovsky et al., 2015) and refined the model to the domain-specific task. This is thought to improve model convergence, and use the low-level features learned on a high volume dataset for a smaller dataset, which is usually the case for medical image models given the high cost to acquire sufficient data. However, the differences in natural images and those in the medical domain limit the wide applicability of this method, likely due to over-parameterization of the original models (Carneiro et al., 2019).

Table 1
Patient cohort demographics. Numbers are n (%) or median (interquartile ranges). MRI indicates magnetic resonance imaging; NIHSS, National Institutes of Health Stroke Scale.

	Training set ($n = 340$)	Test set ($n = 82$)
Age (years)	70 (55–80)	68 (57–79)
Female	176 (52%)	46 (56%)
NIHSS	8 (4–16)	6.5 (2–18)
Onset to MRI (min)	210 (105–683)	230 (107–661)

Efforts have been made to pretrain models on public medical datasets, but access to such medical datasets is still limited. Moreover, higher-level features of medical images vary significantly for different medical domains. To combat the limitations of cross-domain transfer learning and increase features reuse across models, intra-domain transfer learning has been implemented for both natural image and medical image tasks (Raghu et al., 2019). Commonly, a model is initialized in a self-supervised or unsupervised fashion. The advantage of this approach is that it does not require outside datasets or labels. However, even intra-domain pretraining may result in limited feature reuse beyond the first convolutional layer (Verenich et al., 2020). A task-adaptive approach, which uses the same data set for pretraining and then refines the model using two different label sets, has been demonstrated to increase feature reuse and enhance performance (Elman, 1993; Bengio et al., 2009). However, this has not yet been applied in the medical image domain.

We propose an intra-domain task-adaptive transfer learning approach and implement it for TSS classification. The approach uses a multi-stage training schema, leveraging features learned by training on an easier task (stroke detection) to refine the model for a more difficult task (TSS classification). We developed both 2D and 3D CNN models to classify TSS, and we demonstrated our proposed transfer learning approach enhanced classification performance for both architectures when compared to other pretraining schemas, with our 2D model achieving the best performance for classifying TSS < 4.5 h. We also showed that adding soft attention mechanisms during latter stages further improved the performance. To offer clinical insight, we compared our model performance to both previously published methods and radiologist assessment of DWI-FLAIR mismatch. Our deep learning models were able to achieve greater classification sensitivity while maintaining specificity achieved by expert neuroradiologists. By visualizing network gradients via Grad-CAM (Selvaraju et al., 2019), we illustrated that our pretrained models were able to localize the stroke infarct more precisely than the models trained from scratch. To our knowledge, this is the first end-to-end, deep learning approach to classify TSS on a patient dataset with minimal exclusion criteria; moreover, our model exceeds the performance of previously reported state-of-art machine learning models.

2. Material and methods

2.1. Dataset and preprocessing

A total of 422 patients treated for AIS at the UCLA Ronald Reagan Medical Center from 2011 to 2019 were included in this study. This work was performed under the approval of the UCLA Institutional Review Board (#18-000329). A patient was included if they were diagnosed with AIS, had a known stroke onset time, and underwent MRI prior to any treatment, if given. Clinical parameters were gathered from imaging reports and the patient record, with demographic data summarized in Table 1. The study cohort had a median age of 70 (55–80) years, a mean National Institutes of Health Stroke Scale (NIHSS) score of 8(4–15), and were 56% female. The median onset to MRI was 222 (105–715.25) minutes. For performance evaluation, we used 64% for training (272), 16% for validation (68), and 20% (82) as a hold-out test

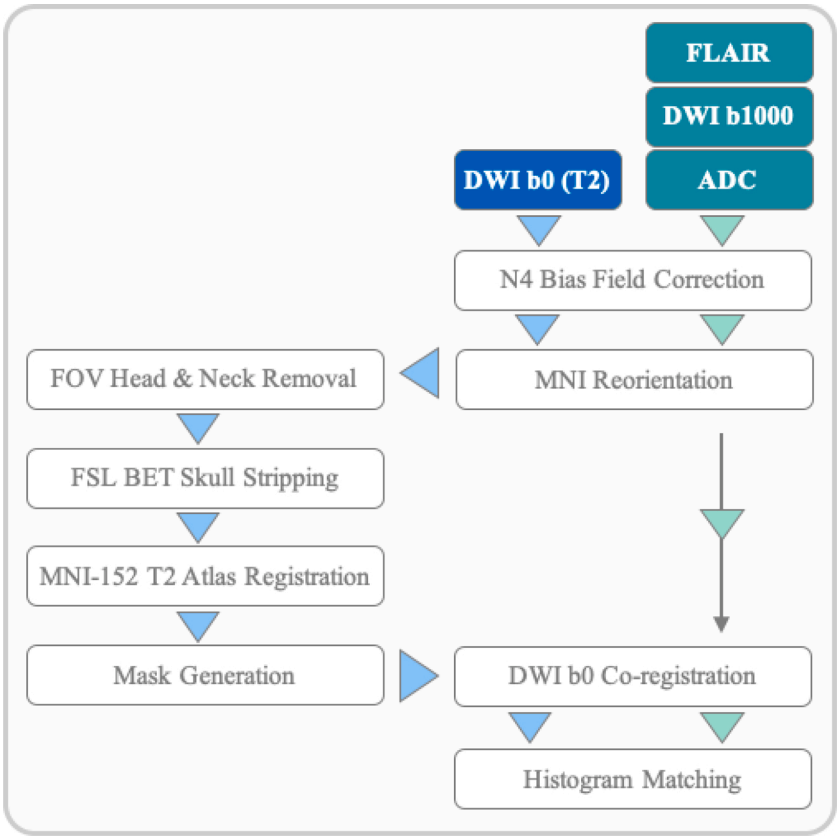

Fig. 2. Preprocessing pipeline for patient series.

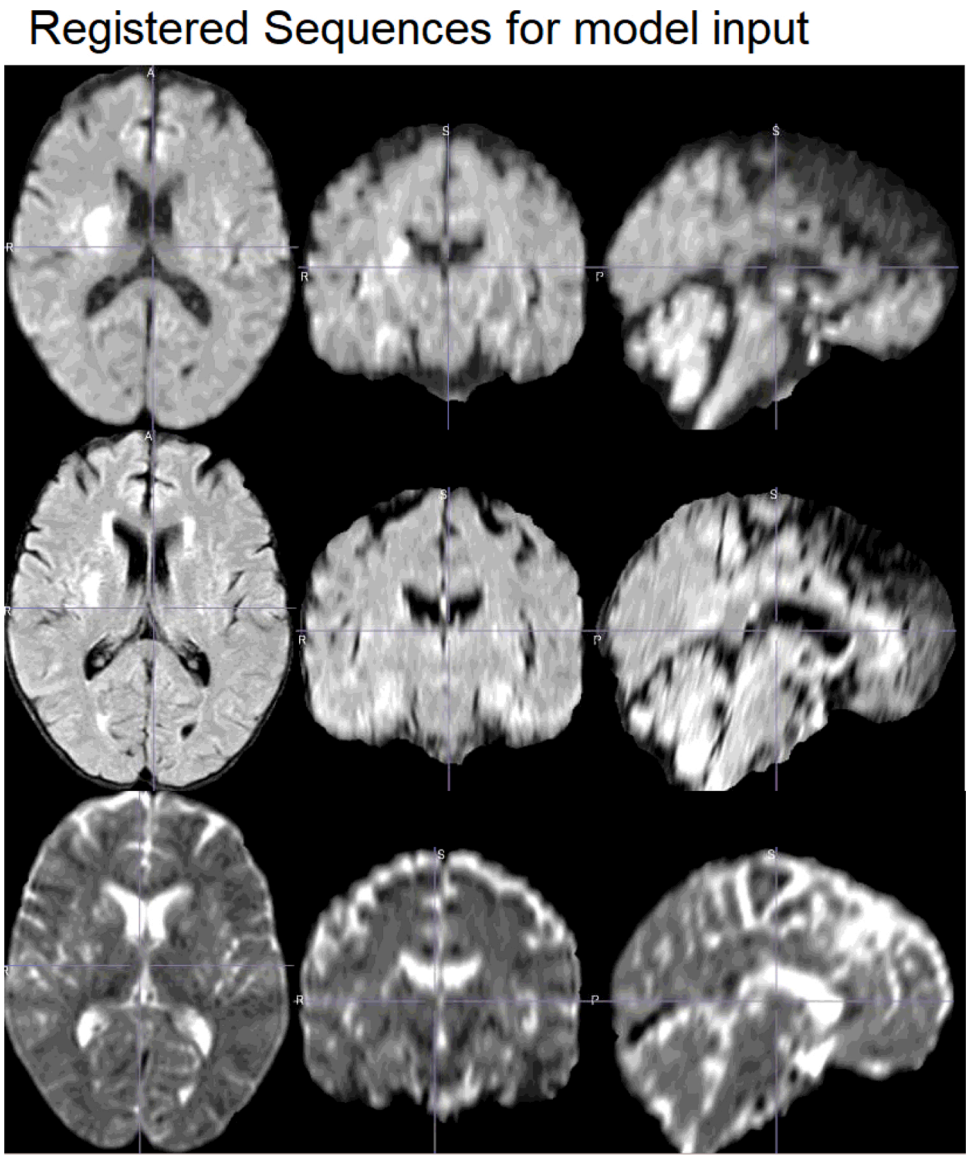

Fig. 3. Sample case of registered output. Sequences from top to bottom: DWI(b1000), FLAIR, T2w(DWI b0).

set. In order to prevent information leakage across tasks, the same test set was used across a set of experiments. The training and testing sets had similar distributions of these clinical factors and TSS. For each patient in the test cohort, DWI-FLAIR mismatch was assessed independently by three senior neuroradiologists with full access to all sequences used in our model.

For each patient, the T2w(DWI b0), DWI(DWI b1000) and FLAIR imaging sequences were retrieved from the institutional picture archiving and communication system (PACS). All patients underwent MRI using a 1.5 T or 3 T echo-planar Siemens MR imaging scanner, performed with 12-channel head coils. The FLAIR images were acquired using a TR range of 8000–9000 ms and a TE range of 88–134 ms. The pixel dimension varied from 0.688 × 0.688 × 6.000 to 0.938 × 0.938 × 6.500 mm. The DWI images were acquired using a TR range of 4000–9000 ms and a TE range of 78–122 ms. The corresponding pixel dimensions varied from 0.859 × 0.859 × 6.000 to 1.850 × 1.850 × 6.500 mm. The DWI b0 sequence was used as a T2w proxy, as it denotes the first step of DWI acquisition with no diffusion attenuation, and the DWI here represents the sequence with b value equal to 1000. The rationale for using these sequences was: (1) T2w represents the anatomical image, so we theorized it might provide contrast information when input along with DWI and FLAIR sequences; (2) since our goal is to classify TSS, and the DWI-FLAIR mismatch is only a surrogate for this goal, extra anatomical imaging information could provide more features related to TSS; and (3) we used three sequences to mimic the RGB channels used in many image classification models, enabling us to compare our training schema to other pretraining approaches. After image retrieval, the sequences were fed into our automated preprocessing pipeline. First, N4 bias field correction (Tustison et al., 2010) was applied to all sequences. Then, each image series was reoriented to the T2w MNI-152 atlas (Fonov et al., 2009). Next, the neck and skull were removed using FSL BET (Jenkinson et al., 2012). The T2w sequence was registered using FSL FLIRT to a version of the T2w MNI-152 atlas that was resized to 224 × 224 × 26 using linear interpolation in order to match the z dimension of the stroke sequences. After a second run of FSL BET was performed to remove remnant artifacts, the remaining sequences were co-registered to the T2w volume. Finally, intensity was normalized, and histogram matching was performed using a reference study. A visual quality check was manually performed for all cases before the experiments. This data preprocessing pipeline is summarized in Fig. 2 and a sample output is shown in Fig. 3.

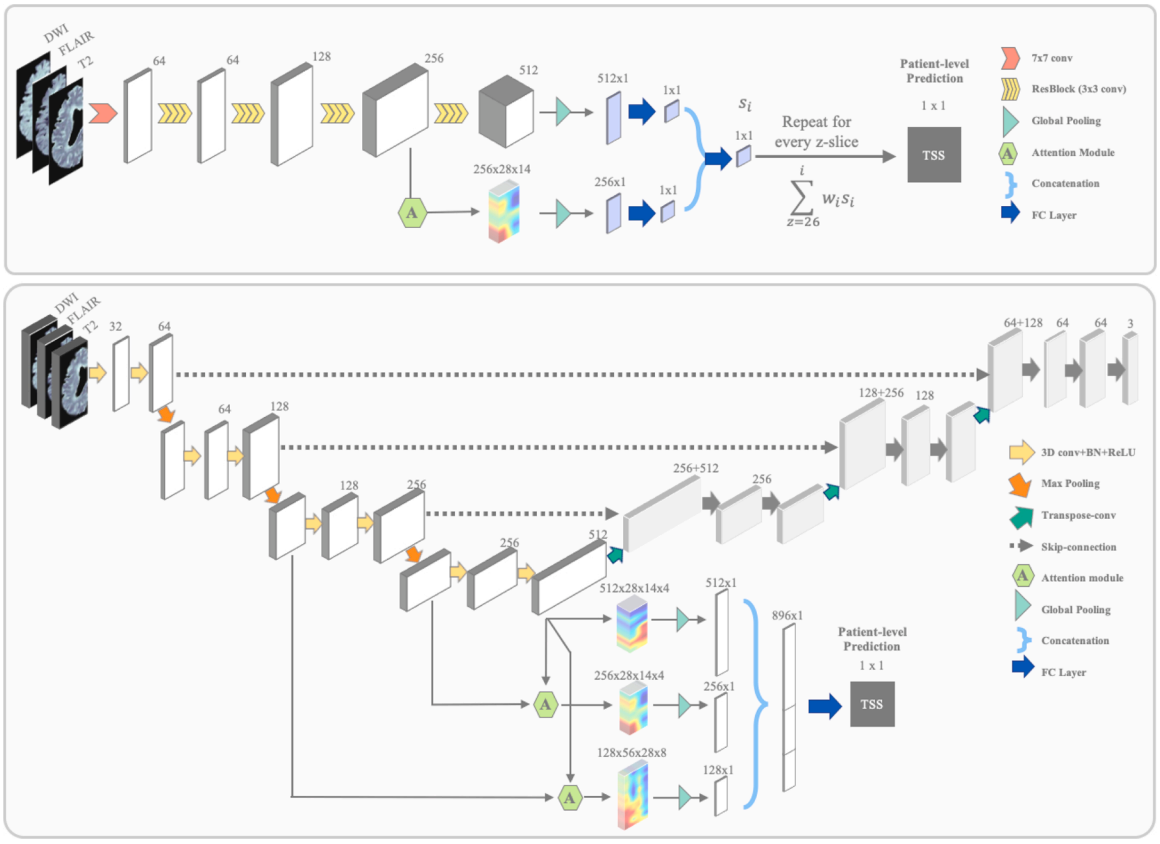

Fig. 4. Architectures for 2D (top) and 3D (bottom) models. Our 2D Self-weighted Slice-wise Attention model took DWI b1000, T2w(b0), and FLAIR as a 3-channel input to a feature extraction backbone. Each slice of the brain was individually fed through four Resblocks of ResNet-18 to generate a 512 × 7 ×4 feature map, then pooled to a 512 × 1 feature vector (He et al., 2016). A soft attention module at the 256-channel convolutional layer was added to generate a 256 × 28 × 14 attention feature map and then pooled to a 256 × 1 feature vector. The feature map and attention feature map were aggregated for each slice with a learnable weighting factor for final classification. Our 3D model first used the entire structure of a 3D U-Net to train an initial weight using Models Genesis. Then volumetric DWI, T2w and FLAIR were directly fed into the encoder part of the network. Two soft attention modules were added at 128 and 256-channel convolution layers. Feature maps from the original network and the two attention modules were pooled globally and concatenated for classification.

2.2. Model architectures

We tested our intra-domain transfer learning schema on custom 2D and 3D architectures. The 2D network takes individual slices as input and feeds them through a convolutional backbone (ResNet-18) adapted from (He et al., 2016) for feature extraction. To account for the large pixel input of an individual MRI slice, we also incorporated a soft attention gate into the architecture (Schlemper et al., 2018). This module uses the final and penultimate convolutional outputs to generate individual pixel weights which identify the most salient regions for the task. This attention module was refined during the TSS tasks later in training to avoid the possibility of convergence at a local minimum and precluding further optimization during model refinement (Oktay et al., 2018). The attention module output and convolutional output were concatenated into a feature vector, which was then fed into a fully-connected layer to generate a single, slice-level output. To aggregate these slice-level predictions into an image-level prediction, we implemented a trainable weighting factor, ranging from 0 to 1, to assign a weight to each slice, and the slice-level outputs were summed in a weighted fashion, resulting in one probability label. The attention module and trainable weight factor ascribe pixel-level and slice-level importance that can be trained and optimized, which enables the model to localize to salient regions.

Given the 3D anatomical information in our dataset, we also evaluated a 3D model architecture. Our 3D model used the encoder part of the 3D U-Net as the model backbone (Çiçek et al., 2016). U-Net (Ronneberger et al., 2015), like ResNet, uses connections between layers for model training and also has been widely used in medical image research. Our 3D approach also used soft attention modules at the 128- and 256-channel intermediate outputs in the encoder part of 3D U-Net in order to allow the network to capture relevant information in early stages of classification. Training a 3D CNN model from scratch does not necessarily yield better performance than 2D models due to the higher number of parameters and the potential for over-fitting. To address these challenges, we first adapted a self-supervised learning approach, known as Models Genesis (Zhou et al., 2019), to train a full 3D U-Net in order to generate initial weights for the stroke detection task. Using Models Genesis, we first modified the original images using non-linear transformation, local shuffling, in-painting, and out-painting and then trained the model to restore the original image, enabling the model to learn important high level features in the original image. We then used the encoder component of the 3D U-Net network, along with two soft attention modules, to train this classification model to detect stroke side and classify TSS. Fig. 4 illustrates the 2D Self-weighted Slice-wise Attention Model structure, and the 3D Attention Model structure. The Models Genesis and soft attention modules bolstered 3D model performance.

2.3. Training schema

To train the models, each brain volume was split into hemispheres along the midsagittal plane on the registered volume. For each hemisphere, three imaging series, T2w, DWI, and FLAIR, were concatenated and input as channels with values normalized to a range of 0 to 1 and

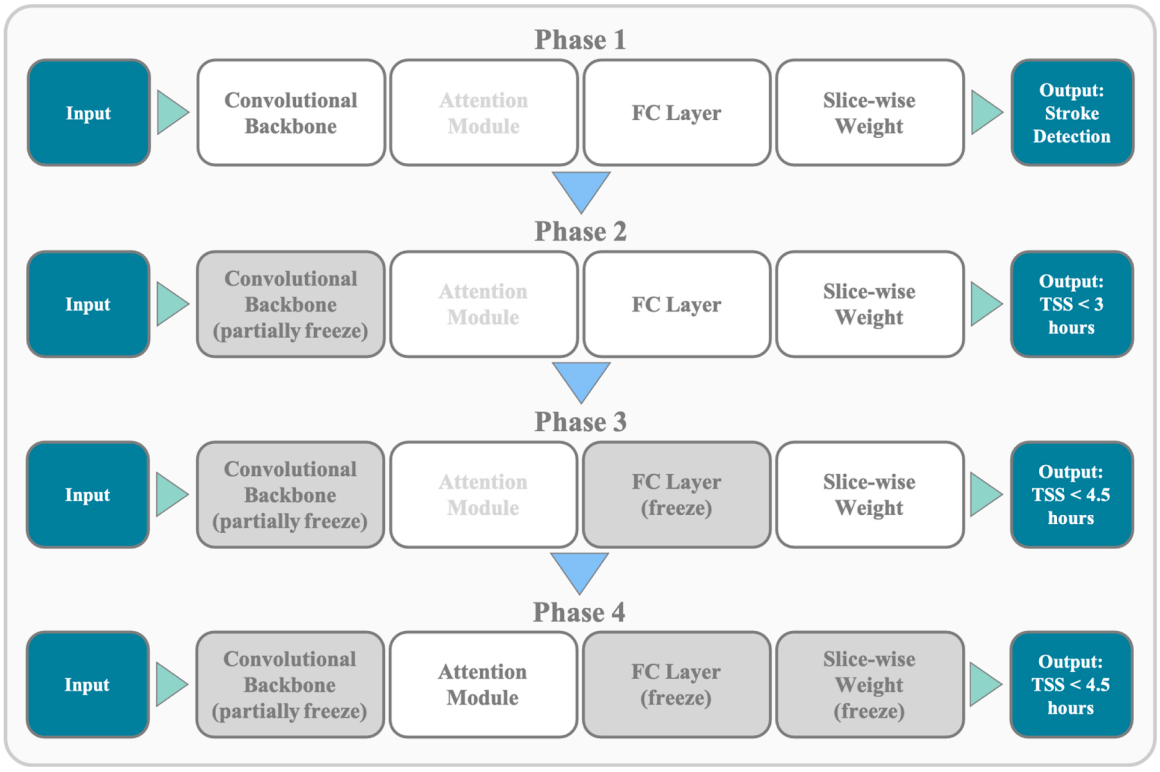

Fig. 5. A summary of our training schema. Each phase utilized a unique classification label, as enumerated in the Outputs boxes for each phase. At the end of each training phase, the weights of certain components were frozen; these frozen weights were then initialized for the model at the start of the following phase.

input dimension of 112 × 224 × 26. The right hemispheres were flipped on the vertical axis in order to spatially align with the left hemispheres for inputs. Our models used a multi-phase training regimen. The first phase consisted of stroke detection, where hemispheres were fed into the model separately and labeled as positive (1) if they had a stroke lesion in the hemisphere and as negative (0) if they had no stroke lesion in the hemisphere. The 2D model was trained from random initialization on this task. For our 3D model, initial weights were generated in a self-supervised fashion before the stroke side detection task for more rapid convergence. Once the model finished training, the first two convolutional layers/blocks were frozen. Specifically, for 2D models, in the ResNet-18 backbone, we froze the first 7 × 7 convolutional layer as well as the following two Resblocks, where the 7 × 7 convolutional layer is denoted conv1 and the two Resblocks each contains two 3 × 3 convolution layers denoted conv2_x from Table 1 of (He et al., 2016). For 3D models, we froze the two layers in the downward path of the 3D U-Net backbone. As described in Çiçek et al. (2016), each layer represents two 3 × 3 ×3 convolutions each followed by a ReLU, then a 2 × 2 ×2 max

Table 2
Performance metrics across tasks and architectures. Double lines separate models with different outputs. Sens = Sensitivity, Spec = Specificity, Acc = Accuracy, AUC = Receiver Operating Characteristic Area Under Curve, Rad = Radiologist, Agg Rad = Aggregate Radiologist.

Stage	Model	Weights	Sens.	Spec.	Acc.	AUC
Phase 1	2D	Random	0.7347	0.9286	0.8316	0.8905
Stroke detection	3D	Random	0.7732	0.9579	0.8646	0.9460
Phase 2	2D	Random	0.2444	0.9310	0.5135	0.6720
TSS < 3 h		ImageNet	0.7879	0.5510	0.6463	0.6733
		Medical	0.6970	0.7142	0.7073	0.7297
		Phase 1	0.8222	0.6552	0.7568	**0.7648**
	3D	Random	0.7143	0.4848	0.6220	0.6129
		Medical	0.5952	0.7750	0.6829	0.7173
		Phase 1	0.8904	0.6000	**0.7724**	0.7452
Phase 3	2D	Random	0.2162	0.9189	0.5676	0.6311
TSS < 4.5 h		ImageNet	0.8789	0.4285	0.6098	0.6054
		Medical	0.6666	0.6939	0.6829	0.6684
		Phase 2	0.5405	0.7838	0.6622	**0.7392**
	3D	Random	0.3750	0.6429	0.5122	0.5863
		Medical	0.8788	0.4489	0.6220	0.6619
		Phase 2	0.6279	0.7895	**0.7037**	0.7087
Phase 4	ML		0.6522	0.7143	0.6363	0.7174
TSS < 4.5 h	2D	Phase 3	0.7027	0.8108	**0.7568**	**0.7407**
Attention+fine-tune	3D	Phase 3	0.5405	0.8378	0.6892	0.7370
DWI-FLAIR mismatch	Rad 1		0.5476	0.8500	0.6951	
	Rad 2		0.4286	0.9250	0.6707	
	Rad 3		0.5714	0.6500	0.6098	
	Agg Rad		0.5730	0.8750	0.7195	

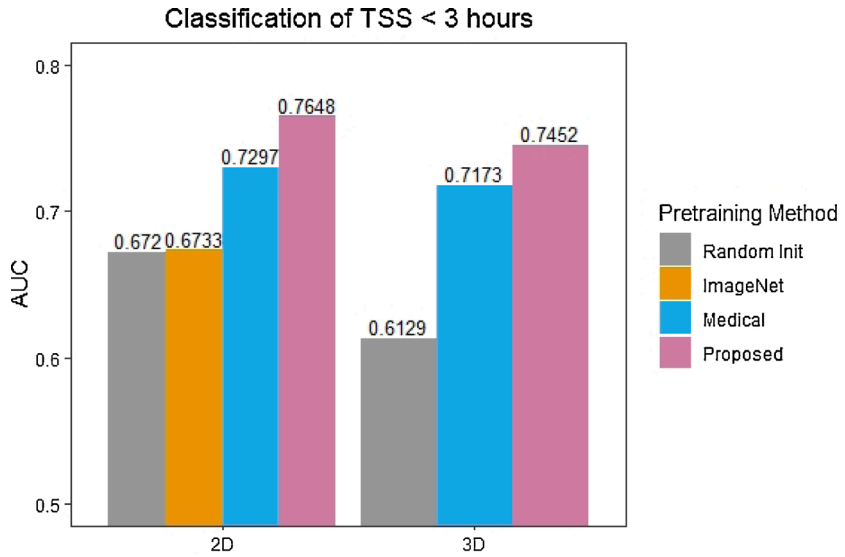

Fig. 6. On second phase task TSS < 3 h, for 2D model, our proposed transfer learning approach has a 5.1% increase, whereas for the 3D model, there is a 8.3% increase in ROC-AUC score.

pooling with strides of two. This pretrained network was then utilized in a second phase of training, whereupon only hemispheres with stroke lesions (positive cases in the stroke side detection task) were used as input. In the second phase, we froze early convolutional weights to refine later layers and trained our model on TSS < 3 h, given the clinical correlation of DWI-FLAIR mismatch to this binarization. For the third phase, we used the pretrained weights of the TSS < 3 h model to train on the TSS < 4.5 h task. The last phase of our training schema (Fig. 5) involved fine-tuning the soft attention modules to further enhance performance. We compared this multi-phase training regimen to training on TSS labels from scratch, pretraining on natural images, and pretraining on external datasets of brain MRIs (Cheng et al., 2015; Mateusz Buda and Saha, 2019).

2.4. Performance evaluation

We trained the stroke detection algorithm for 100 epochs with early stopping, minimizing binary cross-entropy loss functions. All models were trained with the AdaBound optimizer (Luo et al., 2019), which used bounds on a dynamic learning rate to transition smoothly from an adaptive method to the more traditional stochastic gradient descent. This approach allowed the model to maintain a higher rate of convergence in early training epochs. Hyperparameters were selected using a validation set during training. The batch size was 16 for the stroke detection task and 8 for the TSS classification tasks. The early stopping criteria was based on the validation AUC during training with a patience of 10. For Adabound, in stroke side detection task, the initial learning rate was 0.0005 and the final learning rate was 0.01; in TSS classification task, the initial learning rate was 0.00001 and the final learning rate was 0.001. The code was written in PyTorch, and experiments were run on an NVIDIA DGX-1. Our memory usage during training for the 2D models was 4 GB VRAM with batch size 8 and 6 GB VRAM with batch size 16; for the 3D models, memory usage was 7 GB VRAM with batch size 8 and 12 GB VRAM with batch size 16.

3. Results

The performance metrics for all of our training phases are summarized in Table 2. For stroke detection, the 2D and 3D architectures achieved ROC-AUC values of 0.8905 and 0.9460, respectively. This

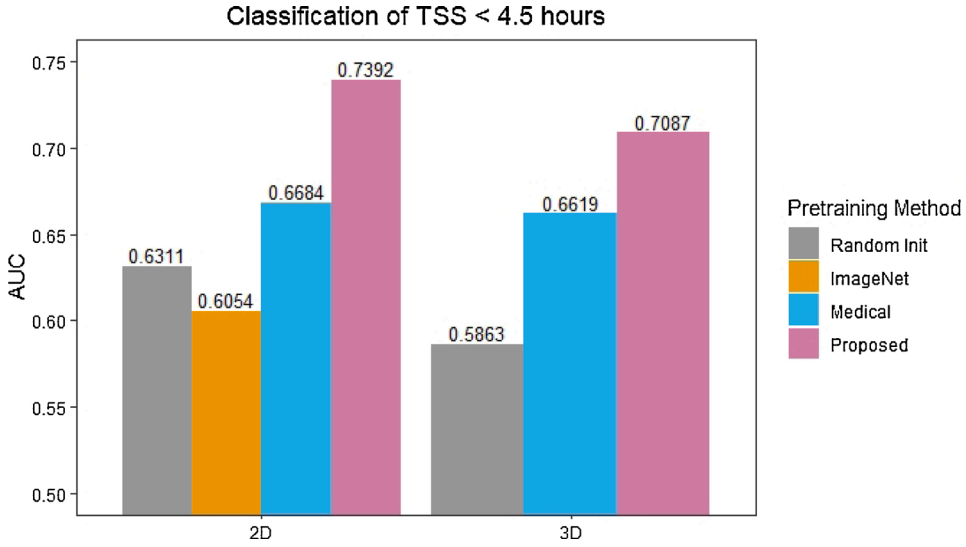

Fig. 7. On third phase task TSS < 4.5 h, for 2D model, our proposed transfer learning approach has a 22.1% increase in AUC; for 3D model, there is a 20.9% increase.

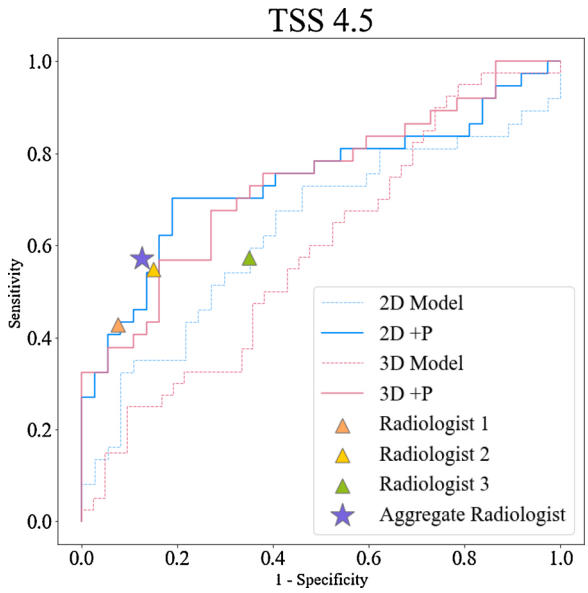

Fig. 8. ROC curves for classifying TSS $<$ 4.5 h. +P = with pretraining.

indicates that the models were able to reliably identify stroke at both the slice and volume level, which aligns with intensity differences usually observed for stroke lesions on DWI and FLAIR series. For the second training phase, classifying TSS $<$ 3 h, our pretraining approach improved the performance of 2D model by 14.0% and our 3D model by 21.6% when compared to random initialization or to pretraining on natural images (2D model only). For both models, we also examined TSS classification performance with weights pretrained on medical image datasets. We used models trained for brain tumor classification and segmentation to initialize our 2D and 3D models, respectively, given that these tasks are in the same domain and use the same medical imaging modalities (Cheng et al., 2015; Mateusz Buda and Saha, 2019). We froze the weights from earlier layers for both models, and we compared the effect of this pretraining to frozen weights learned from our stroke detection task. While performance improvement was observed using medical image pretraining, our pretraining approach was able to achieve higher performance for both models when compared to both natural image and domain-specific pretraining, with the 2D and 3D models achieving 76.48% and 74.52% increase in AUC, respectively (Fig. 6).

In the third phase, we train the models to classify TSS $<$ 4.5 h using weights from the second phase. As shown in Fig. 7, both the 2D and 3D models improved classification performance by 22.1% and 20.9%. For the 2D model, pretraining on natural images reduced performance, which has been observed for other medical-image specific tasks (Raghu et al., 2019). As in Phase 2, We also show the results from ImageNet, Tumor detection and segmentation weight transfer for comparison. As expected, due to the similarity of the dataset, the performance improvement is high, from AUC 0.6311 to 0.6684 and from 0.58 to 0.66 for the 2D and 3D models, respectively. However, the performance improvement (12.9% and 5.9%) is still lower than our proposed method (17.1% and 20.9% to AUC 0.7392 and 0.7087). For both tasks, the 2D model achieves higher performance than the 3D model, even with random initialization.

In the last of our proposed training phases, fine tuning the attention

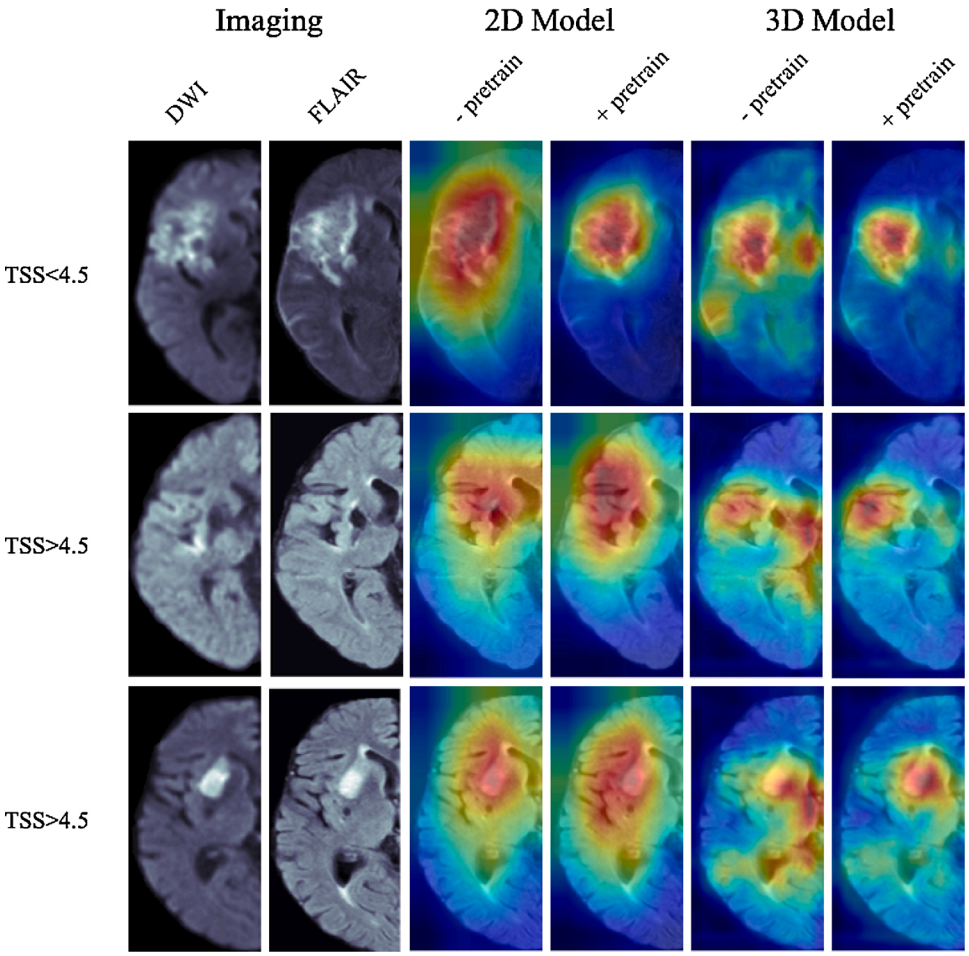

Fig. 9. Grad-CAM visualizations of the penultimate convolutional layer for 2D and 3D models, both from scratch and with pretraining.

modules yields improved performance for both the 2D and 3D models, though the improvement was more notable for the 3D model. The optimal ROC-AUC scores for classification of TSS < 4.5 h are 0.7407 and 0.7370 for 2D and 3D respectively with 17.4% and 25.7% performance gain compared to training from scratch.

For each model, we computed Youden's *J* statistic and reported the sensitivity, specificity, accuracy, and ROC-AUC score. We compared our model to the performance metrics of each radiologist's DWI-FLAIR mismatch assessments, which served as a proxy for TSS. We also compared our model to the previously-published model with the highest performance metrics by applying this model to our own dataset (Lee et al., 2020); these metrics are included in Table 2. Of note, the inter-reader agreement (Fleiss' kappa) was 0.46 among all three radiologists, which is typically regarded as a moderate level of agreement and aligns with previous findings of high variability among reader assessments. We also reported the ROC-AUC curves for each of our models in Fig. 8.

We generated GradCAMs to visually assess model activation. To evaluate the utility of GradCAMs in a clinical context, an expert radiologist evaluated the overlap of activation map and stroke lesion. The radiologist found that, for slices most representative of stroke lesion, 96% of cases evaluated had substantial overlap (>50%) between the lesion and activation, while the remainder of cases had moderate overlap. This indicates that Grad-CAM can qualitatively localize to stroke lesions when trained on the TSS tasks.

4. Discussion

Among the models tested, the pretrained 2D model achieved the highest performance metrics with a sensitivity of 0.70 and a specificity of 0.81 in classifying TSS < 4.5 h. Our model was more sensitive than the DWI-FLAIR assessments performed by the neuroradiologists, which we treated as a surrogate for determining a TSS < 4.5 h. We also compared our model to the previously published state-of-the-art method. The threshold method implemented in Lee et al. (2020), which was used to create the ROI, was very stringent, in that only 221 of our original 422 patients had large enough ROI from which features could be extracted. Thus, their performance metrics represent a subset of our larger dataset. We also tested our model performance on this subset and achieved an ROC-AUC of 0.76. Nevertheless, on the entire dataset, the optimal 2D model with pretraining was able to outperform the previous model. From a clinical perspective, these results indicate that our model may be able to correctly identify more patients within the 4.5 h window and therefore eligible to receive thrombolytic therapy when compared to both DWI-FLAIR mismatch assessment and the threshold-based machine learning method. There are many tasks within the medical image domain to which our proposed task-adaptive pretraining schema can be applied. For example, this schema could be used for brain tumor classification, where brain tumor detection is the pretraining task.

The optimal 2D model has a ROC-AUC comparable to that of the 3D model; however, the sensitivity (0.54) and specificity (0.84) of the 3D model are less balanced, indicating that while the rate of true negatives is high, there are less true positives identified by that model. In total, our model metrics illustrate that the progressive pretraining schema enhances performance for our task considerably, for both our proposed 2D and 3D architecture. For both models, attention modules enhance the performance. The use of GradCAM for our models highlights regions of the brain that impact decisions, as illustrated in Fig. 9. The GradCAMs illustrate that the pretrained model is able to more precisely localize to the stroke infarct and highlight other regions outside of the infarct that may inform this classification task.

Our model performance metrics are comparable to previous approaches in TSS classification. However, this study has a few factors that increase its potential clinical applicability. The patients in our dataset comprise a wider range of stroke locations and other clinical demographics than in previously assessed datasets. Additionally, our model leverages the entire brain hemisphere, which may contain more relevant information among this broader patient cohort. This has the potential to reduce bias in our model, and with the convolutional architecture, allows this information to be incorporated in decision-making.

That said, deep learning generally requires a high volume of data. While many medical image-related tasks have used deep learning with a comparable amount of patient data used here, a higher volume of data would greatly enhance the model performance. This model only uses diffusion-based imaging, as these are the images used in current clinical practice. Incorporating perfusion-based imaging and its derivatives such as perfusion maps may better inform TSS. There is a substantial body of work using perfusion imaging parameters for stroke outcomes (Scalzo et al., 2013; d'Esterre et al., 2017; Ho et al., 2019, 2017). Based on our examination, registration quality was not affected by the ischemic lesion in the T2w images, as the lesion was not apparent in T2w sequence. While this type of registration failure was not a concern in our dataset, it could possibly affect other neuroimaging studies. Finally, the use of clock time as a label for TSS may not fully encompass the physiology underlying ischemia in the brain; for example, cerebral collateral flow may compensate for a hypoperfused area within the brain and reduce the amount of ischemia that tissue is experiencing during a stroke (Bang et al., 2011), which may be the biological reason for DWI-FLAIR mismatch.

5. Conclusion

This approach uses 2D and 3D CNN models to classify TSS for 422 patients and compares model performances to DWI-FLAIR mismatch readings performed by three expert neuroradiologists. We demonstrate that our 2D model outperforms the 3D model when classifying TSS < 4.5 h, which is the current clinical guideline. We show that pretraining the model on stroke detection, then refining the model on TSS classification yields better performance than training on TSS classification labels alone; the incorporation of soft attention modules also enhances performance of both the 2D and 3D when compared to CNNs without them. By visualizing network gradients via Grad-CAM, we show that our pretrained models localize to stroke infarcts and surrounding regions. We demonstrate that our both our 2D and 3D model is able to generalize to an inclusive dataset comprising multiple types of ischemic stroke, and that this model may be able to inform TSS for patients with unknown symptom onset.

Authors' contribution

Haoyue Zhang: conceptualization, methodology, software, validation, formal analysis, visualization, contribution, writing – original draft. Jennifer Polson: conceptualization, methodology, software, visualization, contribution, validation, formal analysis, writing – original draft. Kambiz Nael, Noriko Salamon and Bryan Yoo: data curation, validation, writing – review & editing. Fabien Scalzo: writing – review & editing. William Speier: conceptualization, formal analysis, writing – review & editing. Corey Arnold: conceptualization, formal analysis, writing – review & editing, supervision, project administration, funding acquisition.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgments

This work was supported by the United States National Institutes of Health (NIH) grants R01NS100806 and T32EB016640, and an NVIDIA Academic Hardware Grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the

National Institutes of Health.

Conflict of interest: None declared.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.compmedimag.2021.101926.

References

Bang, O.Y., et al., 2011. Collateral flow predicts response to endovascular therapy for acute ischemic stroke. Stroke 42 (3), 693–699.

Bengio, Y., Louradour, J., Collobert, R., Weston, J., 2009. Curriculum learning. Proceedings of the 26th Annual International Conference on Machine Learning, ICML'09. Association for Computing Machinery, New York, NY, USA, pp. 41–48. https://doi.org/10.1145/1553374.1553380.

Benjamin, E.J., et al., 2019. Heart disease and stroke statistics-2019 update: a report From the American Heart Association. Circulation 139 (10), e56–e528.

Carneiro, G., Tavares, J.M.R.S., Bradley, A.P., Papa, J.P., Nascimento, J.C., Cardoso, J.S., Lu, Z., Belagiannis, V., 2019. Editorial. Comput. Methods Biomech. Biomed. Eng.: Imaging Visualization 7, 241.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation. International Conference on Medical Image Computing and Computer-Assisted Intervention 424–432.

Chan, H.P., Samala, R.K., Hadjiiski, L.M., Zhou, C., 2020. Deep learning in medical image analysis. Adv. Exp. Med. Biol. 1213, 3–21.

Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., Wang, Z., Feng, Q., 2015. Enhanced performance of brain tumor classification via tumor region augmentation and partition. PLOS ONE 10.

d'Esterre, C.D., et al., 2017. Regional comparison of multiphase computed tomographic angiography and computed tomographic perfusion for prediction of tissue fate in ischemic stroke. Stroke 48 (4), 939–945.

Elman, J.L., 1993. Learning and development in neural networks: the importance of starting small. Cognition 48 (1), 71–99.

Etherton, M.R., Barreto, A.D., Schwamm, L.H., Wu, O., 2018. Neuroimaging paradigms to identify patients for reperfusion therapy in stroke of unknown onset. Front. Neurol. 9, 327.

Fonov, V., Evans, A., Mckinstry, R., Almli, C., Collins, L., 2009. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. Neuroimage 47. https://doi.org/10.1016/S1053-8119(09)70884-5.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 770–778. https://doi.org/10.1109/CVPR.2016.90.

Ho, K.C., Speier, W., El-Saden, S., Arnold, C.W., 2017. Classifying acute ischemic stroke onset time using deep imaging features. AMIA Annu. Symp. Proc. 2017, 892–901.

Ho, K.C., Speier, W., Zhang, H., Scalzo, F., El-Saden, S., Arnold, C.W., 2019. A machine learning approach for classifying ischemic stroke onset time from imaging. IEEE Trans. Med. Imaging 38 (7), 1666–1676.

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M., 2012. Fsl. NeuroImage 62, 782–790.

Lee, H., et al., 2020. Machine learning approach to identify stroke within 4.5 hours. Stroke 51 (3), 860–866.

Litjens, G., et al., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005.

Luo, L., Xiong, Y., Liu, Y., Sun, X., 2019. Adaptive gradient methods with dynamic bound of learning rate. Proceedings of the 7th International Conference on Learning Representations. New Orleans, Louisiana.

Mateusz Buda, M.A.M., Saha, Ashirbani, 2019. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. Comput. Biol. Med. 109, 218–225.

Milletari, F., Navab, N., Ahmadi, S., 2016. V-net: fully convolutional neural networks for volumetric medical image segmentation. 2016 Fourth International Conference on 3D Vision (3DV) 565–571. https://doi.org/10.1109/3DV.2016.79.

Nie, D., Zhang, H., Adeli, E., Liu, L., Shen, D., 2016. 3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. Medical image computing and computer-assisted intervention: MICCAI... International Conference on Medical Image Computing and Computer-Assisted Intervention 9901 212–220.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M.C.H., Heinrich, M.P., Misawa, K., Mori, K., McDonagh, S.G., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-net: Learning Where to Look for the Pancreas. arXiv:1804.03999.

Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22, 1345–1359.

Powers, W.J., et al., 2019. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 50 (12), e344–e418.

Raghu, M., Zhang, C., Kleinberg, J.M., Bengio, S., 2019. Transfusion: understanding transfer learning for medical imaging. NeurIPS.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234–241.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Fei-Fei, L., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vision 115, 211–252.

Scalzo, F., et al., 2013. Multi-center prediction of hemorrhagic transformation in acute ischemic stroke using permeability imaging features. Magn. Reson. Imaging 31 (6), 961–969.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention gated networks: learning to leverage salient regions in medical images. Med. Image Anal. 53 https://doi.org/10.1016/j.media.2019.01.012.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2019. Grad-cam: visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vision 128 (2), 336–359. https://doi.org/10.1007/s11263-019-01228-7.

Shin, H.C., et al., 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging 35 (5), 1285–1298.

Thomalla, G., et al., 2011. DWI-FLAIR mismatch for the identification of patients with acute ischaemic stroke within 4–5 h of symptom onset (PRE-FLAIR): a multicentre observational study. Lancet Neurol. 10 (11), 978–986.

Thomalla, G., et al., 2014. A multicenter, randomized, double-blind, placebo-controlled trial to test efficacy and safety of magnetic resonance imaging-based thrombolysis in wake-up stroke (WAKE-UP). Int. J. Stroke 9 (6), 829–836.

Thomalla, G., Simonsen, C.Z., Boutitie, F., Andersen, G.I., Berthèzene, Y., Cheng, B., Cheripelli, B., Cho, T.-H., Fazekas, F., Fiehler, J., Ford, I., Galinovic, I., Gellissen, S., Golsari, A., Gregori, J., Günther, M., Guibernau, J., Haeusler, K.G., Hennerici, M., Kemmling, A., Marstrand, J., Modrau, B., Neeb, L., de la Ossa, N.P., Puig, J., Ringleb, P.A., Roy, P., Scheel, E., Schonewille, W.J., Serena, J., Sunaert, S., Villringer, K., Wouters, A., Thijs, V.N., Ebinger, M., Endres, M., Fiebach, J.B., Lemmens, R., Muir, K.W., Nighoghossian, N., Pedraza, S., Gerloff, C., 2018. Mri-guided thrombolysis for stroke with unknown time of onset. N. Engl. J. Med. 379, 611–622.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4itk: improved n3 bias correction. IEEE Trans. Med. Imaging 29, 1310–1320.

Urrutia, V.C., Faigle, R., Zeiler, S.R., Marsh, E.B., Bahouth, M.N., Trevino, M.C., Dearborn, J.L., Leigh, R., Rice, S., Lane, K., Saheed, M.O., Hill, P.M., Llinás, R.H., 2018. Safety of intravenous alteplase within 4.5 hours for patients awakening with stroke symptoms. PLOS ONE 13.

Verenich, E., Velasquez, A., Murshed, M.G.S., Hussain, F., 2020. The Utility of Feature Reuse: Transfer Learning in Data-Starved Regimes. arXiv:2003.04117.

Weiss, K., Khoshgoftaar, T.M., Wang, D.D., 2016. A survey of transfer learning. J. Big Data 3. https://doi.org/10.1186/s40537-016-0043-6.

Winzeck, S., Hakim, A., McKinley, R., Pinto, José 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, 679. https://doi.org/10.3389/fneur.2018.00679.

Zhou, Z., et al., 2019. Models genesis: generic autodidactic models for 3d medical image analysis. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham, pp. 384–393.