Acute Ischemic Stroke Onset Time Classification with Dynamic Convolution and Perfusion Maps Fusion

Peng Yang[1], Yuchen Zhang[2], Haijun Lei[2], Yueyan Bian[3], Qi Yang[3,4(✉)], and Baiying Lei[1(✉)]

[1] Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen 518060, China
leiby@szu.edu.cn

[2] Key Laboratory of Service Computing and Applications, Guangdong Province Key Laboratory of Popular High Performance Computers, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

[3] Department of Radiology, Beijing Chaoyang Hospital, Capital Medical University, Beijing, China
yangyangqiqi@gmail.com

[4] Laboratory for Clinical Medicine, Capital Medical University, Beijing, China

Abstract. In treating acute ischemic stroke (AIS), determining the time since stroke onset (TSS) is crucial. Computed tomography perfusion (CTP) is vital for determining TSS by providing sufficient cerebral blood flow information. However, the CTP has small samples and high dimensions. In addition, the CTP is multi-map data, which has heterogeneity and complementarity. To address these issues, this paper demonstrates a classification model using CTP to classify the TSS of AIS patients. Firstly, we use dynamic convolution to improve model representation without increasing network complexity. Secondly, we use multi-scale feature fusion to fuse the local correlation of low-order features and use a transformer to fuse the global correlation of higher-order features. Finally, multi-head pooling attention is used to learn the feature information further and obtain as much important information as possible. We use a five-fold cross-validation strategy to verify the effectiveness of our method on the private dataset from a local hospital. The experimental results show that our proposed method achieves at least 5% higher accuracy than other methods in TTS classification task.

Keywords: Computed tomography perfusion · Dynamic convolution · Multi-map

P. Yang and Y. Zhang—These authors contributed equally to this work.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 558–568, 2023.
https://doi.org/10.1007/978-3-031-43904-9_54

1 Introduction

Acute ischemic stroke (AIS) is a disease of ischemic necrosis or softening of localized brain tissue caused by cerebral blood circulation disturbance, ischemia, and hypoxia [1, 2]. Intravenous thrombolysis can be used to open blood vessels within 4.5 h. For patients with large vessel occlusion, the internal blood vessels can be opened by removing the thrombus within 6 h. Therefore, determining the time since stroke onset (TSS) is crucial for creating a treatment plan for AIS patients, with a TSS of less than 6 h being critical. However, approximately 30% of AIS occurs at unknown time points due to wake-up stroke (WUS) and unknown onset stroke (UOS) [3]. For such patients, determining the TSS accurately is challenging, and they may be excluded from appropriate treatment. Computed tomography perfusion (CTP) is processed with special software to generate perfusion maps: cerebral blood volume (CBF), cerebral blood volume (CBV), mean transit time (MTT), and peak response time (Tmax) [4]. These perfusion maps can provide sufficient information on cerebral blood flow, ischemic penumbra, and infarction core area.

There are some machine learning methods to determine the TSS of AIS by automatic discrimination [5–8]. For example, Ho *et al.* [5] first used auto-encoder (AE) to learn magnetic resonance imaging (MRI) and then put the learned and original features into the classifier for TSS classification. Lee *et al.* [7] analyzed diffusion-weighted imaging (DWI) and fluid-attenuated inversion recovery (FLAIR) using automatic image processing methods to obtain appropriate dimensional features and used machine learning to classify the TSS. The above machine learning-based TSS classification methods often use regions of interest (ROI) while ignoring the spatial correlation between neural images. Several researchers try to employ deep learning techniques to classify AIS TSS, considering the spatial correlation among neural images. Zhang *et al.* [9] designed a new intra-domain task adaptive migration learning method to classify the TSS of AIS. Polson *et al.* [10] designed the neighborhood and attention network with segmented weight sharing to learn DWI, apparent diffusion coefficient (ADC), and FLAIR, then used weighted softmax to aggregate sub-features and achieve TSS classification.

With the small samples and the high dimension of CTP, the convolution neural network (CNN) cannot effectively extract features, resulting in the problem of network non-convergence. Additionally, some existing TSS classification methods leverage multi-map mainly by simple linear connections, which do not thoroughly learn the supplementary information of CTP [6, 10]. Therefore, we design a classification model based on dynamic convolution and multi-map fusion to classify the TSS. Firstly, we replace the ordinary convolution in the feature extraction network with dynamic convolution (DConv) [11] to improve the network performance without increasing the network complexity. Secondly, low-order multi-map features are fused to enhance the acquisition of local information by multi-scale feature fusion (MFF). Then high-order features are fused to obtain the global association information by transformer fusion (Transfus). Finally, multi-head pooling attention (MPA) is used to emphasize the high-order features, and the most discriminative features are selected to classify TSS.

2 Methodology

The main framework of our proposed method is depicted in Fig. 1. Specifically, the feature extraction of each map is performed independently using four feature extraction networks, each consisting of five stages. Dconv [11] is used to improve network performance without increasing complexity. In the first three stages, MFF is used to capture and fuse the details of different scales of multi-map features. In the last two stages, Transfus is used to fuse the global correlation of the high-order features. The learned multi-map high-order features are put into the MPA to learn them further and merge the potential tensor sequence. Finally, the selected features are put into the full connection layer to achieve the classification of TSS.

2.1 Dynamic Convolution Feature Extraction Network

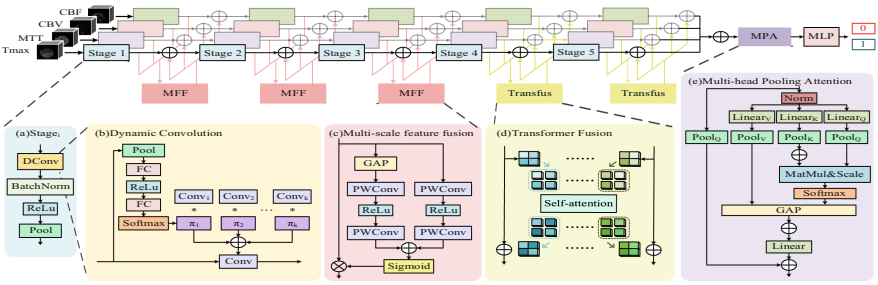

Fig. 1. The framework of our proposed method.

The feature extraction network of a single map consists of five stages. The structure of each stage is shown in Fig. 1 (a). Considering the high data dimension and the small number of samples, the network layers are too deep to cause over-fitting. We replace traditional convolution with dynamic convolution [11]. DConv improves the model expression ability by fusing multiple convolution cores, and its structure is shown in Fig. 1 (b). DConv will can not increase the network complexity and thus improve the network performance. It will not increase too many parameters and calculation amount while increasing the capacity of the model. Inspired by static perceptron [12], Dconv has k convolution cores, which share the same core size and input/output dimensions. An attention block is used [13] to generate the attention weight of k convolution cores, and finally aggregate the results through convolution. The formula is shown in Eq. (1):

$$y = \Sigma_{i=1}^{k} \pi_i(x) Conv(x), \ \ s.t\, 0 \leq \pi_i(x) \leq 1, \ \ \Sigma_{i=1}^{k} \pi_i(x) = 1 \tag{1}$$

where $\pi_i(x)$ is the weight of the i-th convolution, which varies with each input x. . The attention block compresses the features of each channel through global average pooling and then uses two fully connected layers (with ReLU activation function between them) and a Softmax function to generate the attention weight of k convolution cores.

2.2 Multi-map Fusion Module

For multi-map information fusion of low-order features, considering the small area of acute stroke focus, a multi-scale attention module is used to fuse multi-map features in the first three stages. Its structure is shown in Fig. 1 (c). Record the output feature of each stage as $x_{ij}(i = 1, 2, 3; j = 1, 2, 3, 4)$, where i represents the output feature of the i-th stage, and j represents the j-th map. The feature x_i of the input multi-scale channel attention module are denoted as:

$$x_i = x_{i1} \oplus x_{i2} \oplus x_{i3} \oplus x_{i4} \tag{2}$$

By setting different global average pooling (GAP) sizes, we can focus on the interactive feature information in channel dimensions at multiple scales, and aggregate local and global features. Through point-wise convolution (PWConv), point-wise channel interaction is used for each spatial location to realize the integration of local information. The local channel features are calculated as follows:

$$L(x_i) = PWConv(ReLu(PWConv(GAP(x_i)))) \tag{3}$$

The core size of $PWConv$ is $\frac{C}{r} \times C \times 1 \times 1 \times 1$ and $C \times \frac{C}{r} \times 1 \times 1 \times 1$. The global channel features are calculated as follows:

$$G(x_i) = PWConv(ReLu(PWConv(x_i))) \tag{4}$$

The final feature $x_i^{'}$ is calculated as follows:

$$x_i\prime = x_i \otimes \sigma(L(x_i) \oplus G(x_i)) \tag{5}$$

For the fusion of multi-map information of high-order features, considering the relevance of global information between different modes, the self-attention mechanism [14] in the transformer is used to learn multi-map information, and its structure is shown in Fig. 1 (d). Specifically, the output characteristic of each stage is $x_{ij}(i = 4, 5; j = 1, 2, 3, 4)$, where i represents the output feature of the i-th stage, and j represents the j-th mode. Similar to the previous work of some scholars [15–18], we think that the middle feature map of each mode is a set rather than a patch, and treat each element in the set as a token [19]. At this time, each token takes into account all the token information of the four branches. Finally, the fused features are superimposed on the branches for the next stage. Let the characteristic $x \in \mathbb{R}^{N \times D}$ of the input transformer block, where N is the number of tokens in the sequence and each token is represented by the feature vector of dimension D. It uses the scaling dot product between Query and Key to calculate the focus weight and aggregates the value of each Query. Finally, nonlinear transformation is used to calculate the fused output features x_{out}.

Table 1. Comparison of results of different methods (%). (p-value < 0.05)

Methods	Accuracy	Sensitivity	Precision	F1-score	Kappa	AUC
ResNet18	75.04 ± 4.39	92.42 ± 6.04	75.59 ± 3.22	83.07 ± 3.25	37.05 ± 10.17	69.57 ± 5.03
ResNet34	77.56 ± 4.57	91.74 ± 6.07	78.44 ± 4.07	84.43 ± 3.36	44.82 ± 11.65	77.07 ± 5.15
ResNet50	74.01 ± 2.63	91.62 ± 8.80	75.22 ± 3.46	82.31 ± 2.65	34.31 ± 6.70	67.60 ± 8.44
VGG11	73.52 ± 4.68	86.38 ± 7.62	76.92 ± 3.87	81.17 ± 3.90	36.71 ± 10.48	63.59 ± 5.28
AlexNet	74.02 ± 4.37	84.9 ± 7.28	78.31 ± 4.48	81.23 ± 3.51	38.93 ± 11.06	69.54 ± 6.24
CoAtNet0 [20]	76.00 ± 4.21	91.00 ± 3.25	77.38 ± 4.76	83.51 ± 2.30	40.17 ± 13.63	67.48 ± 8.29
C3D [21]	74.95 ± 2.98	87.09 ± 8.89	78.22 ± 3.54	82.11 ± 2.77	40.18 ± 7.28	67.78 ± 5.31
I3D [22]	73.09 ± 4.84	**92.48 ± 7.08**	74.07 ± 4.86	82.03 ± 3.17	30.28 ± 14.93	69.15 ± 7.35
MFNet [23]	74.55 ± 4.49	90.91 ± 6.97	75.99 ± 4.47	82.58 ± 3.23	36.11 ± 12.57	71.08 ± 4.78
SSFTTNet [24]	75.95 ± 4.16	89.32 ± 10.00	78.09 ± 3.60	83.00 ± 3.82	41.70 ± 8.58	74.74 ± 6.05
Slowfast [25]	76.05 ± 4.82	85.75 ± 4.70	79.79 ± 4.01	82.62 ± 3.73	44.16 ± 10.84	74.21 ± 5.58
Ours	**82.99 ± 4.14**	89.40 ± 5.77	**85.89 ± 4.61**	**87.45 ± 3.12**	**60.97 ± 9.65**	**81.48 ± 5.74**

2.3 Multi-head Pooling Attention

With the deepening of the network layers, the semantic information contained in the output features becomes higher and higher. After the post-fusion of the branch network, we use an MPA to learn the high-order semantic details further. Here, a smaller number of tokens is used to increase the dimension of each token to facilitate the storage of more information. Unlike the original multi-head attention (MHA) operator [14], the multi-head pooling attention module gathers the potential tensor sequence to reduce the length of the input sequence, and its structure is shown in Fig. 1 (e). Like MHA [14], Query, Key, and Value are obtained through the linear operation. Add the corresponding pooling layer to Query, Key, and Value to further sample it.

3 Experiments

3.1 Experimental Configuration

Dataset and Data Preprocessing. The dataset of 200 AIS patients in this paper is from a local hospital. The patients are divided into two categories: positive (TSS $<$ 6 h) and negative (TSS $\geq$ 6 h). Finally, 133 in the positive subjects and 67 in the negative subjects are included. Each subject contains CBF, CBV, MTT, and Tmax. The size of all CTP images is set to 256 $\times$ 256 $\times$ 32.

Experimental Setup. The network structure is based on PyTorch 1.9.0 framework and CUDA 11.2 Titan $\times$ 2. We use a five-fold cross-validation method to verify the effectiveness of our method. 80% of the data is used as a training set and 20% as a test set. During the training process, the Adam optimizer optimizes the parameters, and the learning rate is set to 0.00001. The learning strategy of fixed step attenuation is adopted, in which the step size is set to 15, γ is 0.8. The number of iterations of training is 50.

3.2 Experimental Results and Analysis

Comparative Study. We evaluate the effectiveness of our method by comparing it with other approaches on the same dataset [20–25]. The results in Table 1 demonstrate that our model achieves at least a 5% higher accuracy than the other methods and outperforms them in other evaluation indicators. Table 1 also shows each method's area under the curve (AUC). Our method achieves an 81% AUC in the TSS classification task, indicating its superiority. To provide a more intuitive comparison of the model's performance under different indicators, we created radar charts to represent the evaluation results of the comparative experiment, as shown in Fig. 2. These charts indicate that our method performs well in all evaluate indicators. To verify the reliability of our method, we conduct T-test verification on the comparison methods and find the p-value to be less than 0.05. Therefore, we believe that our method is valid.

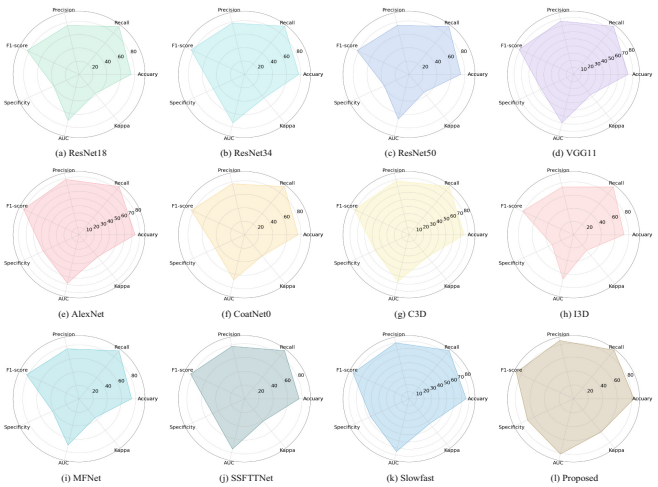

Fig. 2. Radar chart of different model performances.

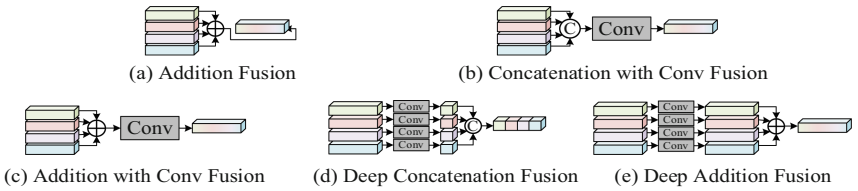

Fig. 3. The flow chart of different fusion methods.

Table 2. The results of different fusion methods (%).

Method	Accuracy	Sensitivity	Precision	F1-score	Specificity	Kappa
AF	79.44 ± 3.70	90.91 ± 5.19	80.68 ± 2.33	85.43 ± 2.80	56.59 ± 7.08	50.76 ± 8.54
CCF	79.41 ± 6.04	90.97 ± 7.77	81.05 ± 5.52	85.45 ± 4.06	56.37 ± 19.05	50.05 ± 17.31
ACF	79.45 ± 2.56	90.17 ± 4.44	81.19 ± 2.75	85.36 ± 1.91	58.02 ± 9.37	50.99 ± 6.59
DCF	77.99 ± 2.80	**95.50 ± 3.17**	77.19 ± 3.72	85.27 ± 1.32	43.19 ± 13.94	43.46 ± 10.32
DAF	80.48 ± 3.90	93.16 ± 5.08	80.62 ± 3.05	86.37 ± 2.90	55.27 ± 8.82	52.46 ± 9.26
Ours	**82.99 ± 4.14**	89.40 ± 5.77	**85.89 ± 4.61**	**87.45 ± 3.12**	**70.33 ± 11.18**	**60.97 ± 9.65**

Fusion Effectiveness. To effectively fuse the image features of CBV, CBF, MTT, and Tmax and realize the task of classifying TSS, This section verifies the fusion method in this paper. We mainly compare it with five different feature fusion methods. They are 1) Addition Fusion (AF); 2) Concatenation with Conv Fusion (CCF); 3) Addition with Conv Fusion (ACF); 4) Deep Concatenation Fusion (DCF); 5) Deep Addition Fusion (DAF). The details of these five fusion methods are shown in Fig. 3. Based on the feature extraction network used in this paper, the comparison results are shown in Table 2. It can be seen from the results that the addition fusion method can better fuse features than the concatenation fusion method. The method we proposed is also to fuse features with the addition method. Our method is better than 1) and 4) because we have further learned the fused features. Therefore, our fusion method is the best.

Ablation Study. To assess the efficacy of each module in the proposed method, a series of ablation experiments are conducted by gradually incorporating the four main modules, namely Dconv, MFF, TransF, and MPA, into the backbone network. The results of these experiments are presented in Table 3. The first four lines in Table 3 indicate that adding a single module is sufficient to enhance the performance of the backbone net-work. The CTP data is characterized by small size and high dimension, posing challenges to the deep learning model training. The network can extract more critical features without increasing the network depth by substituting the convolution with dynamic convolution on the backbone network. Map fusion on the backbone net-work improves the model accuracy by exploiting complementary information from multiple maps. Subsequently, MPA is added to extract more profound features from the learned features, ultimately improves the model's overall performance. In conclusion, the ablation experiments demonstrate that including the four modules in the proposed method positively impacts the model performance.

Map Combination Experiment. To investigate the impact of different modes on the time window of disease onset, a series of experiments are conducted on various mode combinations using the techniques proposed in this study. The results of different map combinations are presented in detail in Table 4. Comparing the results of the second to fourth rows in Table 4 shows that the CBV mode fusion exhibits a higher classification rate, which is superior to the fusion of other groups of modes. Furthermore, the experimental outcomes of the single map are slightly lower than the experimental results after fusion, indicating that multi-map fusion can assist the feature extraction network

Table 3. Ablation study (%) (D: DConv, M1: MFF, T: TransF, and M2: MPA).

Module	Accuracy	Sensitivity	Precision	F1-score	Kappa	AUC
D	77.05 ± 2.70	90.23 ± 3.32	78.54 ± 2.58	83.94 ± 2.01	44.33 ± 6.48	66.9 + 7.91
M1	77.48 ± 2.06	89.34 ± 9.20	79.72 ± 3.14	83.92 ± 2.60	46.78 ± 14.24	66.15 ± 23.87
T	76.91 ± 4.62	85.67 ± 5.76	80.9 ± 3.37	83.12 ± 3.51	43.51 ± 14.76	73.10 ± 7.71
M1 + T	78.01 ± 1.81	86.44 ± 4.39	81.98 ± 4.65	83.96 ± 0.73	54.60 ± 15.90	72.96 ± 3.60
M1 + M2	78.48 ± 2.48	90.94 ± 5.75	79.66 ± 1.50	84.83 ± 2.25	51.78 ± 14.24	76.04 ± 5.45
T + M2	78.00 ± 2.74	**91.60 ± 6.90**	78.96 ± 2.95	84.65 ± 2.40	56.54 ± 10.46	74.69 ± 3.08
M1 + T + M2	80.04 ± 5.07	90.04 ± 4.23	81.82 ± 4.57	85.74 ± 3.51	58.26 ± 6.18	77.82 ± 12.48
Ours	**82.99 ± 4.14**	89.40 ± 5.77	**85.89 ± 4.61**	**87.45 ± 3.12**	**60.97 ± 9.65**	**81.48 ± 5.74**

in obtaining more crucial disease information, thereby enhancing the effectiveness of our network. These observations demonstrate that our approach can improve the TSS classification result by learning the multi-map relationship.

Comparison with SOTA Methods. In Table 5, we have chosen relevant works for comparison. These studies aim to classify TSS based on brain images, with a time threshold of 4.5 h. The results demonstrate that our method performs relatively well in comparison. Specifically, our method achieves the best AUC and ACC among all methods, as reported in [8]. This may be attributed to the relatively large size of their dataset. Compared to studies such as [7, 26], our method achieves better results despite using the same amount of data.

Table 4. Map combination study (%) (V: CBV, F: CBF, M: MTT, and T: Tmax).

Map				Accuracy	Sensitivity	Precision	F1-score	Kappa	AUC
V	F	M	T						
√				77.01 ± 4.68	85.67 ± 7.26	81.02 ± 3.66	83.11 ± 3.77	46.90 ± 10.30	76.58 ± 4.74
√		√		78.99 ± 5.18	83.50 ± 10.65	85.22 ± 5.46	83.85 ± 5.05	53.23 ± 10.13	71.61 ± 11.49
	√	√		79.03 ± 3.94	90.14 ± 6.48	81.08 ± 5.16	85.11 ± 2.58	49.61 ± 11.83	75.46 ± 5.71
	√		√	78.03 ± 3.43	90.91 ± 6.46	79.49 ± 4.38	84.59 ± 2.41	46.54 ± 9.99	75.22 ± 7.24
√	√	√		80.99 ± 2.93	91.68 ± 5.66	81.89 ± 0.90	86.44 ± 2.53	54.77 ± 5.19	75.18 ± 3.86
√		√	√	80.50 ± 5.66	89.40 ± 5.09	83.41 ± 7.98	86.00 ± 3.71	53.90 ± 15.11	75.93 ± 5.73
	√	√	√	79.96 ± 3.79	**95.41 ± 5.02**	78.96 ± 3.04	86.34 ± 2.75	49.95 ± 8.97	76.94 ± 8.99
√	√	√	√	**82.99 ± 4.14**	89.40 ± 5.77	**85.89 ± 4.61**	**87.45 ± 3.12**	**60.97 ± 9.65**	**81.48 ± 5.74**

Table 5. Comparison of results of SOTA methods.

Ref.	Data	Subjects	AUC	sensitivity	specificity	Accuracy
[9]	DWI; FLAIR	422	0.74	0.70	0.81	0.758
[8]	DWI; FLAIR	342;245	0.896	0.823	0.827	0.878
[27]	DWI; FLAIR	404;368	–	0.777	0.802	0.791
[28]	DWI; MRI; ADC	25;26	0.754	0.952	0.500	0.788
[26]	DWI; FLAIR	173;95	–	0.769	0.840	0.805
[6]	DWI; FLAIR; MR perfusion	85;46	0.765	0.788	–	–
[7]	DWI;FLAIR;ADC	149;173	–	0.48	0.91	–
Ours	CTP	133;67	0.807	0.859	0.894	0.830

4 Conclusion

In this study, we propose a TSS classification model that integrates dynamic convolution and multi-map fusion to enable rapid and accurate diagnosis of unknown stroke cases. Our approach leverages the dynamic convolution mechanism to enhance model representation without introducing additional network complexity. We also employ a multi-map fusion strategy, consisting of MFF and TransF, to incorporate local and global correlations across low-order and high-order features, respectively. Furthermore, we introduce an MPA module to extract and incorporate as much critical feature information as possible. Through a series of rigorous experiments, our proposed method outperforms several state-of-the-art models in accuracy and robustness. Our findings suggest that our approach holds immense promise in assisting medical practitioners in making effective diagnosis decisions for TSS classification.

Acknowledgement. This work was supported National Natural Science Foundation of China (Nos. 62201360, 62101338, 61871274, and U1902209), National Natural Science Foundation of Guangdong Province (2019A1515111205), Guangdong Basic and Applied Basic Research (2021A1515110746), Shenzhen Key Basic Research Project (KCXFZ20201221173213036, JCYJ20220818095809021, SGDX202011030958020–07, JCYJ201908081556188–06, and JCYJ20190808145011259) Capital's Funds for Health Improvement and Research (No. 2022–1-2031), Beijing Hospitals Authority's Ascent Plan (No. DFL20220303), and Beijing Key Specialists in Major Epidemic Prevention and Control.

References

1. Phipps, M.S., Cronin, C.A.: Management of acute ischemic stroke. RMD Open **368**, l6983 (2020)
2. Paciaroni, M., Caso, V., Agnelli, G.: The concept of ischemic penumbra in acute stroke and therapeutic opportunities. Eur. Neurol. **61**(6), 321–330 (2009)
3. Peter-Derex, L., Derex, L.: Wake-up stroke: from pathophysiology to management. Sleep Med. Rev. **48**, 101212 (2019)

4. Konstas, A., Goldmakher, G., Lee, T.-Y., Lev, M.: Theoretic basis and technical implementations of CT perfusion in acute ischemic stroke, part 1: theoretic basis. Am. J. Neuroradiol. **30**(4), 662–668 (2009)
5. Ho, K.C., Speier, W., El-Saden, S., Arnold, C.W.: Classifying acute ischemic stroke onset time using deep imaging features. In: AMIA Annual Symposium Proceedings, pp. 892–901 (2017)
6. Ho, K.C., Speier, W., Zhang, H., Scalzo, F., El-Saden, S., Arnold, C.W.: A machine learning approach for classifying ischemic stroke onset time from imaging. IEEE Trans. Med. Imaging **38**(7), 1666–1676 (2019)
7. Lee, H., et al.: Machine learning approach to identify stroke within 4.5 hours. Stroke **51**(3), 860–866 (2020)
8. Jiang, L., et al.: Development and external validation of a stability machine learning model to identify wake-up stroke onset time from MRI. Eur. Radiol. **32**(6), 3661–3669 (2022)
9. Zhang, H., et al.: Intra-domain task-adaptive transfer learning to determine acute ischemic stroke onset time. Comput. Med. Imaging Graph. **90**, 101926 (2021)
10. Polson, J.S., et al.: Identifying acute ischemic stroke patients within the thrombolytic treatment window using deep learning. J. Neuroimaging **32**(6), 1153–1160 (2022)
11. Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z.: Dynamic convolution: attention over convolution kernels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11030–11039 (2020)
12. Rosenblatt, F.: The perceptron, a perceiving and recognizing automaton Project Para. Cornell Aeronautical Laboratory (1957)
13. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
14. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems. Vol. 30 (2017)
15. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
16. Qi, D., Su, L., Song, J., Cui, E., Bharti, T., Sacheti, A.: ImageBERT: cross-modal pre-training with large-scale weak-supervised image-text data. arXiv preprint arXiv:2001.07966 (2020)
17. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International Conference on Machine Learning, pp. 1691–1703. PMLR (2020)
18. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: VideoBERT: a joint model for video and language representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7464–7473 (2019)
19. Chitta, K., Prakash, A., Jaeger, B., Yu, Z., Renz, K., Geiger, A.: TransFuser: Imitation with transformer-based sensor fusion for autonomous driving. IEEE Trans. Pattern Anal. Mach. Intell. (2022). https://doi.org/10.1109/TPAMI.2022.3200245
20. Dai, Z., Liu, H., Le, Q.V., Tan, M.: CoatNet: marrying convolution and attention for all data sizes. Adv. Neural. Inf. Process. Syst. **34**, 3965–3977 (2021)
21. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
22. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
23. Chen, Y., Kalantidis, Y., Li, J., Yan, S., Feng, J.: Multi-fiber networks for video recognition. In: Proceedings of the European Conference on Computer Vision, pp. 352–367 (2018)
24. Sun, L., Zhao, G., Zheng, Y., Wu, Z.: Spectral–spatial feature tokenization transformer for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. **60**, 1–14 (2022)

25. Feichtenhofer, C., Fan, H., Malik, J., He, K.: SlowFast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202–6211 (2019)
26. Zhu, H., Jiang, L., Zhang, H., Luo, L., Chen, Y., Chen, Y.: An automatic machine learning approach for ischemic stroke onset time identification based on DWI and FLAIR imaging. NeuroImage: Clin. **31**, 102744 (2021)
27. Polson, J.S., et al.: Deep learning approaches to identify patients within the thrombolytic treatment window (2022).https://doi.org/10.1111/jon.13043
28. Zhang, Y.-Q., et al.: MRI radiomic features-based machine learning approach to classify ischemic stroke onset time. J. Neurol. **269**(1), 350–360 (2022)