PINet: A Multimodal Network for Pontine Infarction Segmentation and Early Neurological Deterioration Prediction

Jinshuo Liu[1], Ruiquan Ge[1✉], Ning Zhang[1], Beining Wu[1], Xinyan Yao[1], Zhiyao Wang[1], Changmiao Wang[2✉], Gangyong Jia[1], Xiang Wan[2], Ruyue Huang[3], and Yi Pan[4,5]

[1] Hangzhou Dianzi University, Hangzhou, China
gespring@hdu.edu.cn
[2] Shenzhen Research Institute of Big Data, Shenzhen, China
cmwangalbert@gmail.com
[3] Third Affiliated Hospital of Wenzhou Medical University, Wenzhou, China
[4] Shenzhen Institute of Advanced Technology, Shenzhen, China
[5] Faculty of Computer Science and Artificial Intelligence Shenzhen University of Advanced Technology, Shenzhen, China

Abstract. Pontine infarction is a high-risk subtype of ischemic stroke, which can easily lead to severe neurological dysfunction and even death, posing a high risk of disability and death. Accurate infarct segmentation and early prediction of Early Neurological Deterioration (END) are critical for treatment planning. However, prior work often treats these tasks separately or lacks explicit inter-task information transfer in multi-task settings. This results in the inability of segmentation-learned evidence of lesion structure to be used in classification tasks, while the risk discrimination signals of classification tasks cannot inversely constrain segmentation, limiting its reliable use in clinical decision-making. We propose PINet, a multimodal multi-task framework that jointly performs infarct lesion segmentation and END prediction with explicit cross-task feature interaction. PINet integrates Discrete Wavelet Transform (DWT) and its inverse into an encoder–decoder to replace pooling/upsampling, preserving high-frequency boundary details and improving small-lesion representation. A dual-task guided fusion module is designed to deeply fuse MRI features with clinical variables and learn task-specific representations. To model slice-wise continuity and long-range non-local dependencies in brain MRI, we further introduce a Mamba-based state-space global aggregation module that accumulates global context via sequential state propagation. Finally, a multimodal interaction classifier captures imaging–clinical correlations to enhance END prediction. Experiments on an in-house dataset of 386 patients with five-fold cross-validation show PINet consistently outperforms state-of-the-art methods on both tasks, demonstrating the benefit of explicit inter-task interaction and multimodal global-context modeling for stroke risk assessment. Our code is available at https://github.com/LDS151/PINet.

Keywords: Multi-Task Learning · Pontine Infarction · Cross-modal.

1 Introduction

Brainstem stroke is one of the deadliest types of stroke, with pontine infarction (PI) being the main subtype, accounting for approximately 60%–85% of brainstem strokes [12, 19]. Despite advances in reperfusion therapies and multimodal imaging, outcomes for a substantial proportion of patients remain poor. Early Neurological Deterioration (END) is a key contributor, occurring in approximately 20%–40% of patients [23] and strongly associated with long-term disability [20, 26]. Owing to the dynamic and heterogeneous nature of infarct progression, subtle imaging cues and fluctuating clinical indicators are difficult to identify reliably by manual assessment. Current evaluation of infarct volume and END risk still relies on subjective bedside judgment and manual ROI annotation, which is time-consuming and prone to inter-observer variability [4], while failing to exploit high-dimensional cross-modal interactions for precise intervention. Therefore, there is an urgent need for automated, integrated frameworks that can simultaneously segment lesions and deliver robust END predictions.

Existing PI-related models still suffer from two critical limitations. First, most approaches decouple lesion segmentation from END prediction [3, 10, 13, 16, 22, 24], thereby breaking the clinical structure–function evidence chain. Clinically, END is often associated with dynamic infarct expansion and involvement of eloquent functional territories; without joint modeling, predictors cannot fully exploit lesion-specific evidence such as lesion volume, topography, and territorial patterns [17]. Conversely, without prognostic supervision, segmentation models tend to optimize only voxel-wise similarity, may fail to emphasize END-relevant regions, and are more likely to be confused by similar imaging patterns, ultimately degrading both segmentation accuracy and clinical credibility [9]. Second, while various multimodal methods [7,11,14] can fuse heterogeneous information, many still pursue a single, uniform fusion representation, lacking task-specific tuning, leading to modal dominance and gradient imbalance. Segmentation relies more on local boundaries and texture cues, while end-prediction depends more on lesion morphology, spatial distribution, and global context. Without task-guided fusion, the fused features may fail to meet the diverse information needs of segmentation and prediction, thus limiting effective cross-modal interaction and downstream performance.

To address these issues, we propose PINet to jointly perform lesion segmentation and END prediction by deeply integrating imaging and structured clinical data. The main contributions of this work are summarized as follows. First, we replace average pooling in U-Net with a bijective discrete wavelet transform (DWT) and inverse wavelet transform (IWT) invertible scale transform, which reduces spatial resolution while avoiding irreversible information loss. Second, we propose a Dual-Task Guided Fusion Module (DGFM), which achieves deep fusion of clinical information and imaging features and produces task-specific feature representations for different tasks. Third, considering that END depends on whole-brain lesion distribution, slice-wise continuity, and remote inter-regional associations, we introduce a Mamba Global Feature Aggregator (MGFA) that leverages the recursive modeling property of state-space models to accumulate

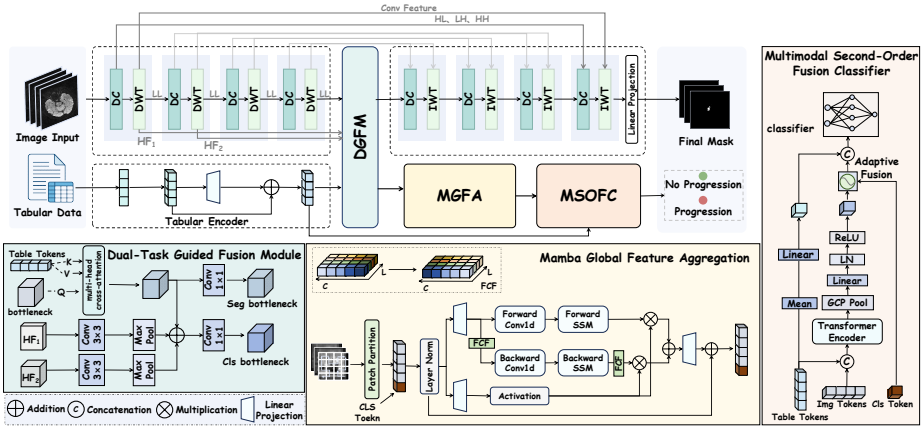

Fig. 1. Overview of PINet. The architecture integrates DWT/IWT for mitigating the irreversible loss of subtle lesion cues and a DGFM for cross-modal integration. A MGFA is employed to capture long-range dependencies, while a MSOFC performs the final prediction. DC denotes DoubleConv.

context at the whole-brain scale for capturing long-range non-local dependencies, while maintaining controllable computational and memory costs. Fourth, we propose a Multimodal Second-Order Fusion Classifier (MSOFC), which explicitly models pairwise interactions between imaging and clinical features via second-order statistics, thereby enhancing cross-modal synergistic information utilization to improve END prediction performance.

2 Proposed Method

2.1 Overview

As shown in Fig. 1, PINet has three stages and takes a 3D MRI volume as input, which it treats as an ordered set of 2D slices. In the first stage, a wavelet-based encoder extracts features from each slice, while a separate tabular encoder converts the clinical variables into embeddings. Next, the DGFM integrates the imaging and clinical features to produce compact, task-specific representations. In the final stage, the segmentation pathway reconstructs slice features using the IWT and generates a 2D lesion map for each slice. For classification, the MGFA captures patient-level patterns by modeling all slices from the same individual, and the MSOFC uses the fused information to predict END progression.

2.2 Wavelet-based Pooling

Due to the morphological diversity and small size of pontine infarction lesions, conventional downsampling easily causes irreversible loss of subtle lesion fea-

tures and spatial information. To address this issue, we introduce a deterministic wavelet-based mechanism that uses the DWT and IWT. This design preserves essential structural content and frequency details when downsampling each slice. In the encoder, DWT is implemented using stride-2 grouped convolutions with fixed Haar wavelet filters, yielding an invertible decomposition of $\boldsymbol{F}_l^{out}$ into four sub-bands: $\boldsymbol{LL}_l, \boldsymbol{LH}_l, \boldsymbol{HL}_l, \boldsymbol{HH}_l$. The low-frequency component $\boldsymbol{LL}_l$, which captures coarse structural patterns, is forwarded to subsequent stages. The high-frequency components are concatenated as a residual, $\boldsymbol{H}_l = \mathbf{Concat}(\boldsymbol{LH}_l, \boldsymbol{HL}_l, \boldsymbol{HH}_l)$, and stored for the decoder, as they contain fine-scale information such as lesion boundaries and surrounding texture. By retaining $\boldsymbol{H}_l$, the model avoids suppressing weak micro-lesion signals during deep feature extraction and maintains the full frequency content needed for accurate reconstruction. The resulting bottleneck representation is $\boldsymbol{F}_b \in \mathbb{R}^{H_b \times W_b \times C_b}$.

The decoder restores spatial resolution using multi-stage IWT. At each stage, features are first projected to a consistent channel dimension and refined using a 1×1 convolution, after which IWT is applied as $\hat{\boldsymbol{D}}_l = \mathbf{IWT}(\boldsymbol{D}_l, \boldsymbol{H}_l)$. This operation recombines the low-frequency features with the stored high-frequency coefficients $\boldsymbol{H}_l$ in the frequency domain to recover fine details. The reconstructed features $\hat{\boldsymbol{D}}_l$ are then concatenated with the corresponding encoder features $\boldsymbol{F}_l^{out}$, integrating high-level representations with high-resolution spatial detail.

2.3 Dual-Task Guided Fusion Module

To simultaneously achieve deeper cross-modal fusion between imaging and clinical data, and to provide targeted discriminative representations for downstream multi-task learning, we propose a Dual-Task Guided Fusion Module (DGFM) that combines visual and tabular features and produces task-specific bottleneck representations for segmentation and classification. Given the visual bottleneck features $\boldsymbol{F}_b$ for each slice and the patient's tabular representation $\boldsymbol{F}_t$, within this shared space, visual features serve as queries, and tabular features provide the keys and values. We compute the fused representation using multi-head cross-attention: $\boldsymbol{F}_{fusion} = \mathbf{MHCA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$. We then add residual connections to preserve the original visual information and to stabilize optimization. Because classification is typically more sensitive than segmentation to subtle texture cues and local structural differences, we incorporate intermediate features $\boldsymbol{HF}_1$ and $\boldsymbol{HF}_2$ to provide additional high-frequency and local structural information. Using convolution and downsampling, we align their spatial resolutions and channel dimensions and compress them into enhanced representations $\boldsymbol{HF}'_1$ and $\boldsymbol{HF}'_2$. Finally, we construct task-specific bottleneck representations as follows:

$$\boldsymbol{F}_{seg} = \phi_s(\boldsymbol{F}_{fusion}), \quad \boldsymbol{F}_{cls} = \phi_c\big([\boldsymbol{F}_{fusion}; \boldsymbol{HF}'_1; \boldsymbol{HF}'_2]\big), \tag{1}$$

where $\phi_s(\cdot)$ and $\phi_c(\cdot)$ are 1×1 convolutional projection heads.

2.4 Mamba Global Feature Aggregator

END prediction depends not only on local lesion properties but also on how lesions are distributed across the whole brain and how they co-vary globally.

To capture this information, we introduce MGFA, which performs patient-level global modeling using the bottleneck feature maps produced by the DGFM. For each patient, we split the bottleneck feature maps from all slices into fixed-size patches and project each patch into a vector embedding. Then, we sort these patch embeddings according to the Hilbert curve within each slice to obtain a slice sequence, and directly concatenate the slice sequences according to the anatomical order of the slices to form a single patient-level sequence. To retain spatial structure within each slice and depth relationships across slices, we incorporate three-dimensional positional encoding. A learnable `CLS` token is prepended to the sequence to summarize global information for downstream prediction. Building on this global sequence, we further introduce Feature Channel Flipping (FCF) to preserve anatomical continuity while strengthening bidirectional dependency modeling. Unlike traditional Bi-Mamba [27], which reverses the token order, FCF keeps the token order unchanged and instead reverses the feature channels within each token to obtain a complementary representation. Let $\boldsymbol{Z}_0 = [\boldsymbol{z}(1), \ldots, \boldsymbol{z}(L)] \in \mathbb{R}^{L\times d}$ denote the input token sequence, where $\boldsymbol{z}(i) \in \mathbb{R}^d$ is the i-th token embedding, $\boldsymbol{z}(i) = [z_1(i), z_2(i), \ldots, z_d(i)]$. The feature-channel flip is defined as:

$$\mathbf{FCF}(\boldsymbol{z}(i)) = [z_d(i), z_{d-1}(i), \ldots, z_1(i)]. \tag{2}$$

Then, by calculating and connecting the residuals, we obtained result $\boldsymbol{Z}$:

$$\boldsymbol{Z} = \boldsymbol{Z}_0 + \mathbf{SSM}(\boldsymbol{Z}_0) + \mathbf{FCF}(\mathbf{SSM}(\mathbf{FCF}(\boldsymbol{Z}_0))). \tag{3}$$

The `CLS` token and image token sequence in $\boldsymbol{Z}$ are separated for subsequent prediction tasks.

2.5 Multimodal Second-Order Fusion Classifier

To better exploit cross-modal interactions for END prediction, we propose a Multimodal Second-Order Fusion Classifier (MSOFC). We first concatenate the image token sequence with the clinical tokens to form a unified token set and apply a self-attention layer to model dependencies between the two modalities. Let $\boldsymbol{T} \in \mathbb{R}^{N\times d}$ denote the output token features after self-attention. We then summarize the attended tokens using Global Covariance Pooling (GCP) [21] to capture second-order cross-token statistics. Specifically, we project $\boldsymbol{T}$ to obtain $\boldsymbol{U} = \phi(\boldsymbol{T}) \in \mathbb{R}^{N\times d'}$, and compute the covariance matrix followed by Matrix Power Normalization (MPN) for numerical stability:

$$\boldsymbol{\Sigma} = \frac{1}{N-1}(\boldsymbol{U} - \bar{\boldsymbol{U}})(\boldsymbol{U} - \bar{\boldsymbol{U}})^\top, \qquad \boldsymbol{g}_{\text{gcp}} = \text{MPN}(\boldsymbol{\Sigma}). \tag{4}$$

Next, we adaptively fuse $\boldsymbol{g}_{\text{gcp}}$ with the classification token from the MGFA module, denoted by $\boldsymbol{g}_{\text{cls}}$:

$$\boldsymbol{g}_{\text{fuse}} = \alpha\, \boldsymbol{g}_{\text{cls}} + \beta\, \boldsymbol{g}_{\text{gcp}}, \tag{5}$$

where α and β are learnable scalar weights. Raw clinical markers are mean-pooled and linearly projected to obtain $\boldsymbol{g}_{\text{cli}}$. The fused representation and $\boldsymbol{g}_{\text{cli}}$ are then concatenated and fed into an MLP for END prediction.

3 Experiment

3.1 Data and Implementation

Data Preparation. Due to the scarcity of multimodal, multitask datasets, we used a partner-hospital clinical dataset of 386 patients diagnosed with pontine infarction. For each patient, volumetric diffusion-weighted imaging (DWI) scans (slice thickness 5 mm; inter-slice gap 1.5 mm) were paired with structured admission clinical data. From the initial 32 variables, 11 features were selected based on prior clinical evidence [1, 15, 26] and Spearman rank-correlation analysis: age, admission NIHSS score, systolic blood pressure, admission random glucose, neutrophil count, lymphocyte count, platelet count, vertebrobasilar artery stenosis, international normalized ratio, prothrombin time, and D-dimer. Patients were labeled as END-positive if the NIHSS score increased by at least 2 points during hospitalization compared with admission, and END-negative otherwise. The class distribution was approximately 1:2. For preprocessing, DWI scans were intensity-normalized, resampled to a uniform voxel size, and resized to $224 \times 224 \times 20$. Data augmentation (random rotations, flips, and intensity perturbations) was applied during training only. Continuous clinical variables were standardized using training-set statistics, categorical variables were binary-encoded, and missing values were imputed with training-set medians. Identical preprocessing parameters were applied to validation and test sets. The dataset was strictly partitioned according to patients to prevent data leakage.

Implementation Details. All experiments were conducted on an NVIDIA V100 GPU. The model was trained for 50 epochs using Adam with an initial learning rate of 1×10^{-4}, a batch size of 1, and a cosine annealing schedule down to 1×10^{-6}. To mitigate class imbalance, we used a weighted random sampler ($w_c = \frac{1}{\sqrt{N_c}}$). The segmentation branch employed a combined Dice and BCE loss (DiceBCELoss, positive weight = 10.0), while the classification branch used BCEWithLogitsLoss. Segmentation and classification were jointly optimized using homoscedastic uncertainty–based automatic loss weighting, with initial loss weights set to 1.0 for segmentation and 0.5 for classification. The optimal weights for model selection were determined based on a comprehensive index, score (F1 + AUC + Dice). Model performance was evaluated using five-fold stratified cross-validation. The best-performing model from each fold was selected, and results were reported as the mean and standard deviation across the five folds.

3.2 Experimental Results

Comparison. Table 1 provides a quantitative comparison of PINet with representative methods [2, 5–8, 11, 14, 18, 22, 24, 25] on our self-built dataset. Overall, PINet delivers consistently strong performance on both segmentation and classification. For segmentation, PINet achieves a Dice score of 90.9, slightly below nnU-Net [6], which is designed exclusively for segmentation. This gap is expected: nnU-Net [6] prioritizes pixel-level optimization, whereas PINet is built

Table 1. Quantitative comparison of the PINet against state-of-the-art methods on a self-built dataset. The best and second-best results are indicated by **bold** and underline, respectively. I denotes imaging modality, and T denotes tabular clinical data.

Method	Modal	Segmentation		Classification		
		Dice↑	HD95↓	Rec↑	Pre↑	AUC↑
Multimodal single task						
XGBoost [2]	T	–	–	56.7 ± 17.6	65.8 ± 11.1	85.2 ± 9.3
MMGL [24]	I+T	–	–	64.3 ± 9.3	82.7 ± 7.4	84.5 ± 6.8
IRENE [25]	I+T	–	–	63.9 ± 12.8	80.9 ± 16.3	81.7 ± 7.8
MMCL [5]	I+T	–	–	66.3 ± 9.2	84.9 ± 11.7	89.5 ± 5.6
PI-MMNet [22]	I+T	–	–	$\underline{67.5 \pm 7.1}$	83.7 ± 7.4	88.9 ± 9.3
nnU-Net [6]	I	$\mathbf{91.3 \pm 1.3}$	$\underline{2.8 \pm 0.5}$	–	–	–
MedSAM [8]	I+T	84.1 ± 2.3	3.7 ± 0.5	–	–	–
TextBrats [18]	I+T	77.2 ± 3.6	6.2 ± 0.6	–	–	–
Multimodal multitasking						
SelfmedMAE [11]	I+T	86.9 ± 1.8	6.8 ± 0.4	67.9 ± 17.4	80.2 ± 16.7	88.8 ± 6.3
ResGaNet [7]	I+T	87.5 ± 1.2	4.8 ± 0.7	66.3 ± 16.2	84.2 ± 19.7	83.4 ± 5.6
MTANet [14]	I+T	88.7 ± 1.4	3.3 ± 0.4	62.6 ± 16.0	$\underline{85.1 \pm 12.0}$	$\underline{89.3 \pm 6.7}$
PINet (Ours)	I+T	$\underline{90.9 \pm 1.4}$	$\mathbf{2.6 \pm 0.3}$	$\mathbf{68.0 \pm 11.2}$	$\mathbf{87.3 \pm 7.3}$	$\mathbf{90.9 \pm 4.9}$

to learn shared features that support both anatomical delineation and clinically relevant predictive cues. Even under this setting, PINet remains highly competitive and reduces HD95 by approximately 7.1% relative to nnU-Net [6]. For classification, PINet attains state-of-the-art results in Recall, Precision, and AUC. Compared with MTANet [14], PINet improves the Precision by about 2.6% and the AUC by about 1.8%. It also achieves approximately 0.7% higher Recall than PI-MMNet [22], indicating stronger robustness and better class discrimination. Notably, PINet contains 25.15M parameters, fewer than MTANet with 29.7M parameters and SelfMedMAE with 86.6M parameters, suggesting a more compact model scale while maintaining strong overall performance. Fig. 2 shows qualitative results on the test set.

Ablation Study. Table 2 summarizes the ablation results. Compared with the single-task baselines (seg-single and cls-single), the multitask framework delivers clear improvements across both tasks: the Dice score increases from 87.9 to 90.9 (+3.4%), HD95 decreases from 3.1 to 2.6 (-16.1%), and AUC rises from 89.1 to 90.9. Together, these gains indicate that learning segmentation and prediction jointly allows the model to better integrate imaging structure with clinical information, improving both lesion delineation and outcome prediction.

To identify which modules drive these benefits, we remove each component in turn. MGFA contributes most strongly: removing it leads to the largest degradation (Dice -4.7%, HD95 +65.4%, and AUC -7.3%). Excluding the DWT/IWT module also reduces segmentation performance, suggesting that preserving information across multiple frequency bands during downsampling is important

Table 2. Ablation study of PINet. Gray row represents the default full model configuration. When removing a module, a replacement structure with equivalent parameter counts was used, including basic modules such as average pooling, MLP and CNN.

Setting	Segmentation		Classification		
	Dice↑	HD95↓	Rec↑	Pre↑	AUC↑
seg-single	87.9 ± 1.7	3.1 ± 0.3	–	–	–
cls-single	–	–	64.8 ± 15.2	85.4 ± 14.1	89.1 ± 5.8
w/o DWT/IWT	90.1 ± 1.4	2.9 ± 0.4	67.3 ± 11.4	85.2 ± 8.1	90.2 ± 5.1
w/o MGFA	86.6 ± 2.7	4.3 ± 0.5	65.6 ± 13.7	84.1 ± 11.0	84.3 ± 6.71
w/o DGFM	87.5 ± 1.8	3.1 ± 0.7	66.3 ± 16.2	84.2 ± 11.7	86.4 ± 5.6
w/o MSOFC	88.9 ± 1.6	3.8 ± 0.4	66.9 ± 17.4	82.2 ± 13.9	86.8 ± 6.3
Full model	$\mathbf{90.9 \pm 1.4}$	$\mathbf{2.6 \pm 0.3}$	$\mathbf{68.0 \pm 11.2}$	$\mathbf{87.3 \pm 7.3}$	$\mathbf{90.9 \pm 4.9}$

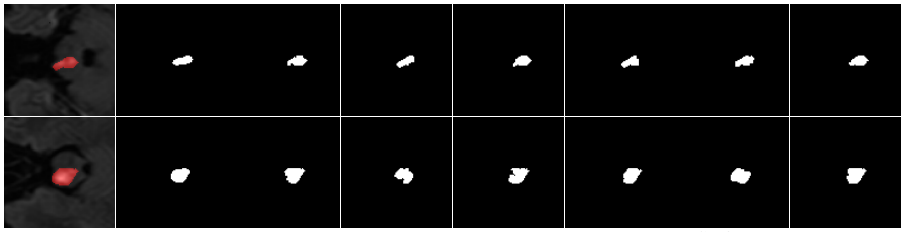

Fig. 2. Comparisons of segmentation results by different methods and our network.

for retaining fine lesion detail. Removing DGFM lowers performance on both tasks, highlighting the need for effective alignment between imaging and clinical signals in multitask learning. Likewise, omitting MSOFC weakens prediction accuracy, supporting the role of richer cross-modal interactions in improving END discrimination and robustness. Overall, each component meaningfully supports cross-modal integration and stable multitask performance.

4 Conclusion

This paper proposes a unified multimodal framework, termed PINet, for jointly segmenting pontine infarction lesions and predicting END. The framework incorporates a DWT/IWT mechanism to better represent complex lesion boundaries and subtle pathological patterns. To integrate imaging and clinical information, PINet uses a lightweight Dual-Task Guided Fusion Module that aligns these feature types at the semantic level while preserving representations tailored to each task. Building on this fusion, the framework introduces an MGFA module to capture global contextual dependencies efficiently. Finally, PINet employs an MSOFC component to model higher-order relationships across modalities and improve the stability of predictions. Experiments on real-world clinical dataset

demonstrate the effectiveness and reliability of the proposed framework. However, the current experiments are limited to a single-center cohort, and multi-center validation is needed to assess generalizability. Future work will explore a more lightweight and efficient PINet design and extend the framework to broader stroke types and other cerebrovascular disease analysis tasks.

Ethics Approval. This study was approved by the Ethics Committee of the Third Affiliated Hospital of Wenzhou Medical University (approval No. 2023053).

Acknowledgments. This work was supported by the Guangxi Science and Technology Program (No. FN2504240022), the Guangxi Key R&D Project (No. AB24010167), the Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011617), the National Natural Science Foundation of China (Nos. 61702146, 62076084, U22A2033, and U20A20386), the Zhejiang Provincial Natural Science Foundation of China (No. LY21F020017), and the Open Project Program of the State Key Laboratory of CAD&CG, Zhejiang University (No. A2410).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bamodu, O.A., Chan, L., Wu, C.H., Yu, S.F., Chung, C.C.: Beyond diagnosis: Leveraging routine blood and urine biomarkers to predict severity and functional outcome in acute ischemic stroke. Heliyon **10**(4) (2024)
2. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
3. Di Matteo, A., Mahé, Y., Leplaideur, S.S., Bonan, I., Bannier, E., Galassi, F.: Deep learning and multi-modal mri for the segmentation of sub-acute and chronic stroke lesions. Pattern Recognition Letters (2025)
4. Gómez, S., Rangel, E., Mantilla, D., Ortiz, A., Camacho, P., de la Rosa, E., Seia, J., Kirschke, J.S., Li, Y., El Habib Daho, M., et al.: Apis: a paired ct-mri dataset for ischemic stroke segmentation-methods and challenges. Scientific Reports **14**(1), 20543 (2024)
5. Hager, P., Menten, M.J., Rueckert, D.: Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23924–23935 (2023)
6. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
7. J, C., S, T., L, Y., C, G., X, K., X, M., W, W., S, L., H, L.: Resganet: Residual group attention network for medical image classification and segmentation. Medical Image Analysis **76**, 102313 (2022)
8. J, M., Y, H., F, L., L, H., C, Y., B, W.: Segment anything in medical images. Nature Communications **15**, 654 (2024)

9. Kamiri, J., Wambugu, G.M., Oirere, A.M.: Multi-task deep learning in medical image processing: a systematic review. International Journal of Computing Sciences Research **9**, 3418–3440 (2025)
10. Kulkarn, A., Harshit, M., Trivedi, V., Rathna, R.: Deep learning framework for stroke detection and lesion segmentation on brain ct scans. In: 2025 Innovations in Power and Advanced Computing Technologies (i-PACT). pp. 1–5. IEEE (2025)
11. L, Z., H, L., J, B., J, H., D, S., P, P.: Self pre-training with masked autoencoders for medical image classification and segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging. pp. 1–6. IEEE (2023)
12. Lei, W.: Bridging neuroscience and clinical practice: Accelerating stroke recovery in a pontine infarction case through neuroplasticity-based rehabilitation. Cureus **16**(10) (2024)
13. Li, M., Kristjánsdóttir, I., van Eimeren, T., Giehl, K., Ellingsen, L.M., Initiative, A.N., et al.: A hybrid cnn and ml framework for multi-modal classification of movement disorders using mri and brain structural features. arXiv preprint arXiv:2602.05574 (2026)
14. Ling, Y., Wang, Y., Dai, W., Yu, J., Liang, P., Kong, D.: Mtanet: Multi-task attention network for automatic medical image segmentation and classification. IEEE Transactions on Medical Imaging **43**(2), 674–685 (2024)
15. Liu, H., Liu, K., Zhang, K., Zong, C., Yang, H., Li, Y., Li, S., Wang, X., Zhao, J., Xia, Z., et al.: Early neurological deterioration in patients with acute ischemic stroke: a prospective multicenter cohort study. Therapeutic Advances in Neurological Disorders **16**, 17562864221147743 (2023)
16. Olmos, J.A., Manzanera, A., Martínez, F.: A geometric multimodal foundation model integrating bp-mri and clinical reports in prostate cancer classification. arXiv preprint arXiv:2602.00214 (2026)
17. Sakamoto, Y., Aoki, J., Nishi, Y., Shoda, S., Kimura, R., Saito, T., Kanamaru, T., Suzuki, K., Katano, T., Kutsuna, A., et al.: Acute dwi volume is a strong imaging predictor of favorable outcomes in patients with acute stroke and treated with mechanical thrombectomy. Journal of the Neurological Sciences **468**, 123334 (2025)
18. Shi, X., Jain, R.K., Li, Y., Hou, R., Cheng, J., Bai, J., Zhao, G., Lin, L., Xu, R., Chen, Y.w.: Textbrats: Text-guided volumetric brain tumor segmentation with innovative dataset development and fusion module exploration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 638–648. Springer (2025)
19. Vlašković, T., Brkić, B.G., Stević, Z., Vukićević, M., Đurović, O., Kostić, D., Stanisavljević, N., Marinković, I., Kapor, S., Marinković, S.: Anatomic and mri bases for pontine infarctions with patients presentation. Journal of Stroke and Cerebrovascular Diseases **31**(8), 106613 (2022)
20. Wang, J., Fu, K., Wang, Z., Wang, N., Wang, X., Xu, T., Li, H., Han, X., Wu, Y.: Mri-based clinical-radiomics nomogram to predict early neurological deterioration in isolated acute pontine infarction: a two-center study in northeast china. BMC Neurology **24**(1), 39 (2024)
21. Wang, Q., Zhang, L., Wu, B., Ren, D., Li, P., Zuo, W., Hu, Q.: What deep cnns benefit from global covariance pooling: An optimization perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10771–10780 (2020)
22. Yu, D., Du, W., Lin, K., Li, X., Ye, Y., Luo, C., Chen, X.: Pimmnet: Introducing multi-modal precipitation nowcasting via a physics-informed perspective. In:

Proceedings of the 33rd ACM International Conference on Multimedia. pp. 11522–11531 (2025)
23. Zhang, J., Luo, Z., Zeng, Y.: Predictive modeling of early neurological deterioration in patients with acute ischemic stroke. World Neurosurgery **191**, 58–67 (2024)
24. Zheng, S., Zhu, Z., Liu, Z., Guo, Z., Liu, Y., Yang, Y., Zhao, Y.: Multi-modal graph learning for disease prediction. IEEE Transactions on Medical Imaging **41**(9), 2207–2216 (2022)
25. Zhou, H.Y., Yu, Y., Wang, C., Zhang, S., Gao, Y., Pan, J., Shao, J., Lu, G., Zhang, K., Li, W.: A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nature Biomedical Engineering **7**(6), 743–755 (2023)
26. Zhu, J., Zhou, H., Mo, J., Zhou, T., Jiang, J., Zhang, Z.: Impact of low-grade inflammation on neurological outcomes in patients with acute pontine infarction: a retrospective cohort study. Frontiers in Immunology **16**, 1744943 (2025)
27. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. In: Forty-first International Conference on Machine Learning. vol. 235, pp. 62429–62442 (2024)