Lesion-centered vision transformer for stroke outcome prediction from image and clinical data

Mingtian Liu[1], Nima Hatami[1], Laura Mechtouff[2,3], Tae-Hee Cho[2,3], Carole Lartizien[1], and Carole Frindel[1,4⋆]

[1] CREATIS, CNRS UMR5220, INSERM U1294, Université Lyon 1, INSA-Lyon, France
[2] Stroke Department, Hospices Civils de Lyon, France
[3] CarMeN, INSERM U1060, INRA U1397, Université Lyon 1, INSA-Lyon, France
[4] Institut Universitaire de France (IUF)

Abstract. Accurate prediction of stroke functional outcome, particularly the 3-month modified Rankin Scale (mRS), is crucial for personalized treatment. Vision Transformers excel in medical imaging and multimodal fusion but struggle with stroke MRI due to data scarcity and rigid tokenization, which may miss subtle anomalies. In response, we propose the Lesion-Centered Vision Transformer (LC-ViT), integrating lesion-focused MRI preprocessing, adaptive token merging, and multimodal fusion. LC-ViT extracts axial, coronal and sagittal views centered on ischemic lesions to optimize visibility and employs a pretrained TC-Former (token-clustering transformer) for adaptative token generation. A mutual cross-attention mechanism further integrates imaging and clinical data. Evaluated on a retrospective private cohort comprising DWI MRI and 62 clinical variables (e.g. demographics, neurological assessments.) of 119 stroke patients treated with thrombectomy (65% favorable outcome), LC-ViT achieves a new state-of-the-art performance (AUC:0.80 ± 0.03, Accuracy: 0.77 ± 0.02) significantly outperforming single modality based deep architectures. Our results highlight the potential of lesion-focused tokenization for stroke outcome prediction and interpretability and broader applications in lesion-localized multimodal analysis. Our code is available at `https://github.com/mingtian12345/LC-VIT`.

Keywords: Outcome Prediction · Multimodal fusion · Adaptive Token.

1 Introduction

Stroke is a leading cause of long-term disability, with ischemic stroke accounting for nearly 87% of all cases [1]. While the advent of reperfusion therapies has significantly improved ischemic stroke management [2], patient outcomes remain highly variable. This underscores the critical need for enhanced predictive models to assist clinicians in better anticipating recovery trajectories. In this study, we focus on predicting the modified 3-month Rankin Scale (mRS) metric, a

⋆ Corresponding author: carole.frindel@creatis.insa-lyon.fr

widely recognized clinical measure of stroke outcomes which serves as a key benchmark for evaluating both recovery and the effectiveness of interventions [4]. We develop a deep prognostic model that combines clinical data from patient EHRs with 3D MRI to predict mRS, a 7-point disability scale binarized in this study, with 0–2 indicating a favorable outcome.

Recent deep learning advances, especially Vision Transformers (ViTs) [5], have demonstrated potential for medical image analysis [8]. ViTs typically require large training datasets, a challenge in stroke outcome prediction. Additionally, 2D ViT models pre-trained on large datasets like ImageNet are not directly applicable to 3D MRI due to dimensional mismatches. Triamese-ViT [10] addresses this limitation by transforming 3D MRI scans into three anatomically relevant 2D views (axial, sagittal, and coronal) centered on the brain, preserving spatial information. This architecture has demonstrated strong performance in brain age estimation. Traditional ViT tokenization, which segments images into fixed-size patches, may be suboptimal for medical imaging, as different regions hold varying clinical significance [11]. To address this, recent works have refined tokenization strategies [11,12,13], with Token Clustering Transformer (TCFormer) [19,20] achieving state-of-the-art results through adaptive token merging.

Traditional machine learning models have been used to predict stroke outcomes from clinical data [3]. However, combining clinical and imaging features remains challenging. Hatami et al. [9] improved predictions by using a CNN combined with an LSTM that incorporated single clinical variables. Both [14] and [15] leveraged XTab [16]– a pre-trained tabular transformer – to fuse image descriptors with clinical data. While [14] was not specific to stroke, [15] applied it to sroke outcome prediction by integrating radiomics-derived quantitative features from DWI MRI with clinical data. Additionally, attention-based multimodal fusion models have shown promising results across various medical applications [21,22,23,24,25,26,27].

Building on this state-of-the-art, we propose the lesion-centered ViT (LC-ViT) for stroke outcome prediction. LC-ViT maximizes lesion visibility, enhances lesion-specific feature extraction, and efficiently fuses structured clinical data with 3D MRI data. Our contributions are: (1) LC-ViT, which integrates lesion-centered views and adaptive vision transformers for improved stroke lesion feature extraction, (2) TCFormer's adaptive token merging, repurposed for MRI, dynamically adjusting token boundaries to better represent lesion structures and (3) a framework combining LC-ViT's imaging features with mutual cross-attention mechanism for clinical data fusion, improving both accuracy and interpretability of stroke outcome prediction.

2 Method

2.1 Framework Overview

As shown in Fig. 1, our framework consists of three key components: (a) an enhanced Triamese-ViT based image encoder referred to as LC-ViT for *lesion-*

centered ViT, (b) a clinical tabular data encoder based on simple MLP, and (c) a classification module encompassing a mutual cross-attention module followed by a simple classification head. In (a), 3D MRI images are first preprocessed to manually segment the lesion, then automatically compute the lesion mask centroid and extract the three anatomical planes centered on the lesion in 3D Slicer [7]. These views are inputted to a pretrained TCFormer model. In (b), clinical data first undergo preprocessing steps including missing values imputation and one-hot encoding for categorical variables, then are mapped into a embedding space by a two-layers MLP for alignment with image features through cross-attention fusion. In(c), a mutual cross-attention module integrates imaging and clinical embeddings by capturing global relationships. The fused features are then processed by a classifier with two linear layers, a LeakyReLU activation, and a final logit output for prediction.

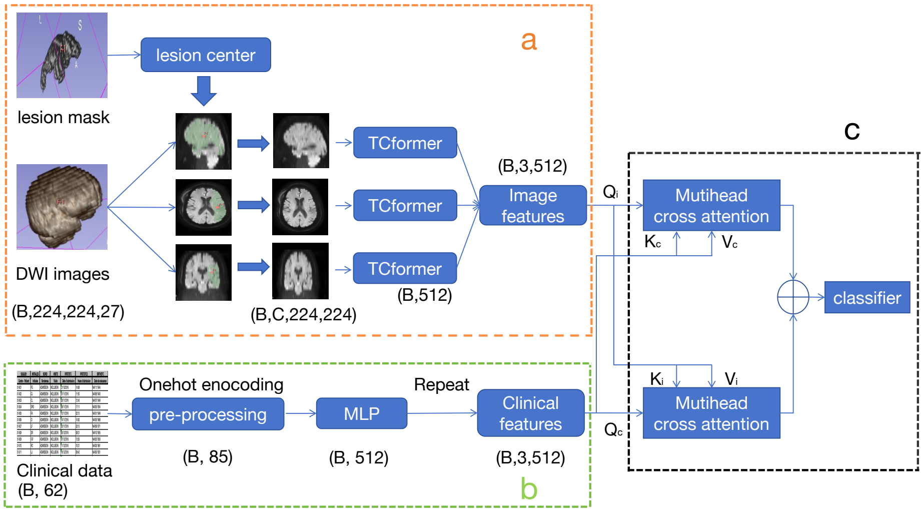

Fig. 1: Overview of the proposed framework (a) LC-ViT encoder based on the pretrained TCformer-light (b) Clinical MLP encoder (c) Mutual cross-attention fusion. The red point refers to lesion centroid and green mask refers to lesion.

2.2 LC-ViT

Enhanced Triamese-ViT Triamese-ViT [10] addresses the computational challenges of encoding high-dimensional 3D MRI by extracting three orthogonal 2D planes (axial, coronal, sagittal) intersecting at the brain center, enabling the use of pre-trained 2D ViTs while preserving essential spatial information. Inspired by this architecture, we propose to adapt it to our task and center the three anatomical views on the stroke lesion centroid, instead of the brain center

(Fig. 1-a). This ensures lesions remain maximally visible across all slices, allowing the model to capture lesion-specific features critical for stroke prognosis.

TCformer TCFormer [19,20], as shown in Fig. 2, is a novel vision transformer model which employs a progressive clustering strategy to merge tokens, allowing flexible token shapes and sizes based on image regions. Each stage contains stacked transformer blocks followed by a token merging block. Using a KNN-based density peaks clustering algorithm, it merges similar tokens while preserving critical details through a Multi-stage Token Aggregation (MTA) head. This enables TCFormer to prioritize clinically relevant areas by dynamically adjusting tokenization. For stroke outcome prediction, we leverage this mechanism to focus on lesion regions while merging background tokens, reducing noise and enhancing feature extraction.

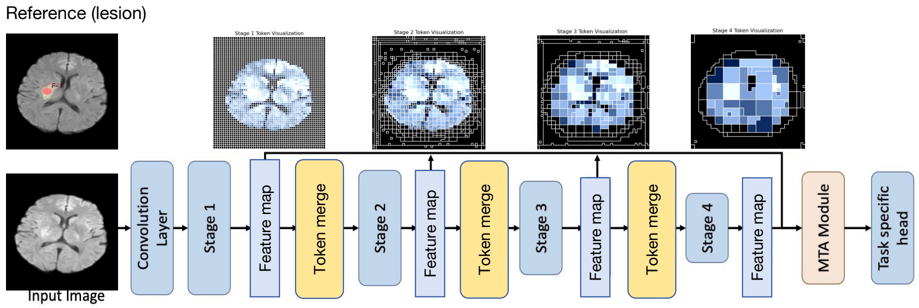

Fig. 2: The architecture of the TCFormer model (inspired by TCFormer [19,20]).

2.3 Mutual Cross-Attention

Cross-attention enables interaction between two modalities by learning their relationships through attention weights and has been widely used in multimodal fusion tasks [21,22,23,24,25,26,27]. In simple cross-attention, one modality serves as the query (Q), while the other provides the keys (K) and values (V). The attention mechanism computes the relevance between Q and K, then aggregates V to generate the query representation. Mutual cross-attention [22,24,25,26] extends simple cross-attention by enabling bidirectional interactions between modalities. Instead of a single-directional query-key relationship, both modalities serve as queries and key/value pairs for each other, capturing richer interdependencies. This approach is particularly useful in tasks where modalities provide complementary information, such as aligning clinical data with image features for stroke outcome prediction. In this study, clinical features act as Q, while image features provide K and V in the simple cross-attention mechanisms, and serve alternatively as the query and key/value pairs in the mutual cross-attention scenario.

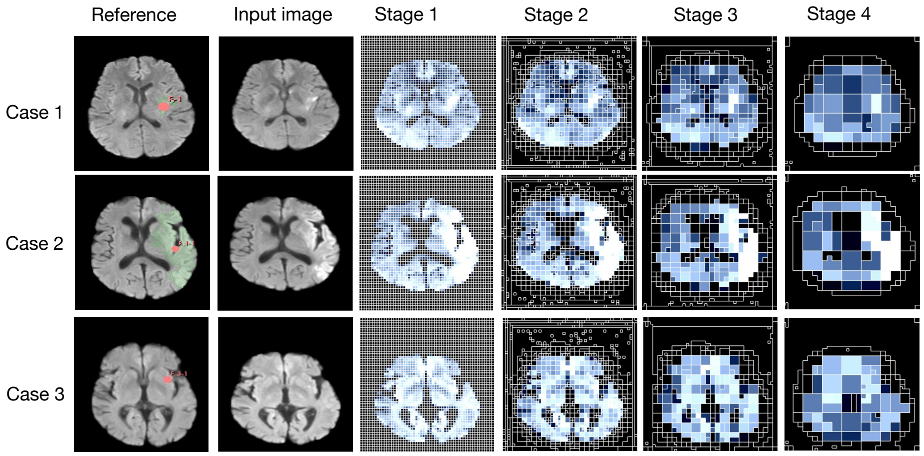

Fig. 3: Visualization of token distribution at each stage for different cases (small lesion, large lesion, and invisible lesion), with the same number of token for all images, equal to 3136, 784, 196, and 49, for stage 1 to 4 respectively.

3 Experiments

3.1 Dataset

Clinical data Our private cohort comprises 119 ischemic stroke patients enrolled in an ongoing single-center observational study approved by the local ethics committee (IRB number: 00009118). All patients were treated with thrombectomy. Clinical data, consisting of 62 variables selected by clinicians for their relevance to patient outcomes, were extracted from hospital records. These variables encompass demographics (9 variables), neurological assessments (e.g., NIHSS), medical history (21 variables), treatments details (3 variables). It also includes numerical clinical indicators (17 variables), such as blood test results and radiological scores, as well as time-related data from admission to treatment stages. Missing data were handled with tailored imputation: mean substitution for continuous variables (e.g., weight), most frequent value for binary categorical variables and a "missing" category for treatment-related data. Categorical variables were one-hot encoded, yielding 85 features after preprocessing. All features were normalized to [0, 1] to ensure comparability for model training.

Image data All patients underwent a standardized protocol, including MR DWI acquisition at the acute stage. The original 3D volumes wereand skull stripping was performed using HD-BET [29]. Expert neuroradiologists manually annotated lesions on the early scans. Subsequent preprocessing was performed, during which the excess black background was cropped that preserved a 20-pixel margin on both the left and right sides for all volumes and each view was resized from the original 192×192×27 to 224×224x27 to fit the input dimension of the TCFormer network.

3.2 Experiments and implementation details

We evaluated the performance of our proposed LC-ViT architecture and compared it to baseline approaches using image-only, clinical-only or multimodal (image and clinical) inputs. For image-based models, we assessed 3D ViT [17] and a radiomics-based model [15]. For structured clinical data-only models, we compared standard machine learning methods including Random Forest, Logistic Regression, XGBoost [18], and MLP as these models are known to mitigate overfitting on small datasets and perform well on tabular data. We also included a comparison with an hybrid model proposed in [15], which encodes image features by using radiomics and uses XTab for clinical data fusion.

Then, we conducted an ablation study to assess the contributions of LC-ViT's components (Fig. 1). For image encoding block (a), we analyzed the prognostic value of each orthogonal view in the triamese network and evaluated lesion-centered vs. brain-centered views. We also replaced the TCFormer network with standard ViT and ResNet backbones. For fusion block (c), we compared mutual vs. simple cross-attention for multimodal fusion.

We performed 10 independent iterations of 10-fold cross-validation. Each iteration used 8 folds for training, 1 for validation, and 1 for testing. Probability outputs from test folds were aggregated for final predictions. A threshold value of 0.36, reflecting the positive sample ratio, was applied for binary mRS prediction. To ensure reproducibility, we conducted experiments with 10 different random seeds and reported the mean and standard deviation across runs. We assessed models using AUC-ROC, specificity, sensitivity, F1 score, mean absolute error (MAE), and accuracy. Since stroke interventions carry inherent risks, specificity was prioritized over sensitivity.

LC-ViT was implemented in PyTorch and trained on a NVIDIA RTX A4000 GPU using the Adam optimizer (learning rate = 0.0001, batch size = 16, 100 epochs with early stopping at 10). 3D ViT followed hyperparameters from [17]. Machine learning models were trained using scikit-learn [28]. The MLP for clinical data consists of two hidden layers and a final linear layer for classification.

4 Results

4.1 Performance of LC-ViT and comparison with baseline models

Our fusion strategy integrates both imaging and clinical information, to leveraging complementary features from each modality. As shown in Table 1, our fusion model consistently outperforms individual image- and clinical-based models. It achieves the highest AUC (0.80 ± 0.03), accuracy (0.77 ± 0.02), and specificity (0.86 ± 0.03) while maintaining a low mean absolute error (0.23 ± 0.02). It is worth mentioning that our hybrid model's probability outputs were predominantly near 0 and 1, making it robust to threshold variations. These results

Table 1: Evaluation of different models for stroke outcome prediction across three modalities: image, clinical, and fusion.

Input	**Methods**	**AUC(↑)**	**ACC(↑)**	**SENS(↑)**	**SPEC(↑)**	**F1(↑)**	**MAE(↓)**
Image	3D ViT	0.51 ± 0.01	0.57 ± 0.02	0.29 ± 0.07	0.73 ± 0.07	0.32 ± 0.05	0.43 ± 0.02
	Radiomics[15]	0.64 ± 0.02	0.60 ± 0.03	0.57 ± 0.05	0.62 ± 0.05	0.50 ± 0.03	0.40 ± 0.03
Clinical	Random Foreast	0.71 ± 0.02	0.65 ± 0.03	0.62 ± 0.07	0.67 ± 0.02	0.56 ± 0.04	0.35 ± 0.03
	Logistic Regression	0.77 ± 0.03	0.73 ± 0.03	0.64 ± 0.05	0.78 ± 0.04	0.63 ± 0.05	0.27 ± 0.03
	XGBoost	0.68 ± 0.04	0.66 ± 0.04	0.53 ± 0.05	0.72 ± 0.05	0.52 ± 0.04	0.34 ± 0.04
	XTab	0.73 ± 0.04	0.69 ± 0.03	0.62 ± 0.07	0.73 ± 0.04	0.58 ± 0.05	0.31 ± 0.03
	MLP	0.75 ± 0.02	0.69 ± 0.04	0.65 ± 0.06	0.72 ± 0.02	0.60 ± 0.05	0.31 ± 0.04
Fusion	XTab[15]	0.75 ± 0.02	0.72 ± 0.02	**0.68 ± 0.05**	0.73 ± 0.05	0.63 ± 0.02	0.28 ± 0.02
	Ours	**0.80 ± 0.03**	**0.77 ± 0.02**	0.62 ± 0.06	**0.86 ± 0.03**	**0.66 ± 0.04**	**0.23 ± 0.02**

highlight the advantages of multimodal learning, as our fusion approach effectively combines different data sources to improve stroke outcome prediction. Furthermore, the visualization of token distribution in Fig 3 demonstrates that our approach effectively captures clinically relevant regions.

4.2 Ablation study

Table 2 evaluates the impact of using multi-view lesion-centered Triamese-ViT and cross attention mechanisms on model performance. In the image-only configuration, incorporating lesion information from multiple views (axial, coronal, and sagittal) improves both AUC and accuracy (AUC = 0.68 ± 0.02; ACC = 0.66 ± 0.03) compared to single-view inputs and to the standard Triamese-ViT implementation based on brain-centered views (AUC = 0.63 ± 0.04; ACC = 0.61 ± 0.03). The fusion model, which combines image and clinical features, further enhances results. Adding, a simple cross attention mechanism improves performance (AUC = 0.78 ± 0.02; ACC = 0.73 ± 0.02) compared with single modality models, but replacing it with a mutual cross attention strategy leads the highest scores (AUC = 0.80 ± 0.03; ACC = 0.77 ± 0.02). These results highlight the benefit of multi-view lesion integration and the importance of effective information exchange across modalities.

Table 2: Evaluation of lesion-centered multi-view inputs and cross-attention mechanisms on stroke outcome prediction.

Input	**Lesion**	**Triamese-VIT**			**Cross Attention**		**Results**	
		Axial	**Coronal**	**Sagittal**	**Simple**	**Mutual**	**AUC(↑)**	**ACC(↑)**
Image	✓	✓					0.66 ± 0.02	0.63 ± 0.02
	✓		✓				0.67 ± 0.03	0.64 ± 0.03
	✓			✓			0.57 ± 0.02	0.59 ± 0.01
		✓	✓	✓			0.63 ± 0.04	0.61 ± 0.04
	✓	✓	✓	✓			0.68 ± 0.02	0.66 ± 0.03
Fusion	✓	✓	✓	✓	✓		0.78 ± 0.02	0.73 ± 0.02
	✓	✓	✓	✓		✓	**0.80 ± 0.03**	**0.77 ± 0.02**

Table 3 compares the proposed TCFormer architecture against baseline models (ViT-Tiny, ResNet-18, and ResNet-50). We kept other settings unchanged and only swapped the TCFormer backbone, while ensuring that all baselines were pre-trained on ImageNet-1K. TCformer outperforms all baselines, achieving the highest AUC (AUC = 0.80±0.03) and accuracy (ACC = 0.77±0.02) while maintaining a reasonable parameter count (14.2 M). We performed a Wilcoxon two-sided test to evaluate the statistical significance of the AUC. The Wilcoxon test confirms that TCFormer's improvements are significant, with p-values below 0.05 for all baseline models. These ablation results show that both the multi-view Triamese-VIT design and the mutual cross-attention mechanism, along with the architectural choices in TCformer, contribute to the performance gains observed.

Table 3: Comparison of the proposed TCFormer backbone with ViT-Tiny, ResNet-18, and ResNet-50.

Model	**#params**	**AUC(↑)**	**ACC(↑)**	**Sens(↑)**	**Spec(↑)**	**F1(↑)**	**MAE(↓)**	**P-value**
ViT-Tiny	5.5M	0.77 ± 0.03	0.72 ± 0.03	0.60 ± 0.04	0.78 ± 0.04	0.60 ± 0.04	0.28 ± 0.03	0.0137
ResNet-18	11.7M	0.77 ± 0.02	0.74 ± 0.03	0.62 ± 0.04	0.81 ± 0.04	0.63 ± 0.04	0.26 ± 0.03	0.0098
ResNet-50	25.6M	0.76 ± 0.02	0.74 ± 0.03	0.60 ± 0.06	0.81 ± 0.03	0.62 ± 0.05	0.26 ± 0.03	0.0020
TCFormer	14.2M	**0.80 ± 0.03**	**0.77 ± 0.02**	**0.62 ± 0.06**	**0.86 ± 0.03**	**0.66 ± 0.04**	**0.23 ± 0.02**	**-**

5 Conclusion

In this study, we introduced LC-ViT, a lesion-centered vision transformer designed to enhance the extraction of lesion-specific features from 3D MRI data. By leveraging dynamic token clustering, it effectively captures clinically significant regions. Additionally, the integration of a mutual cross-attention mechanism for fusing imaging and clinical data further improves predictive performance for stroke outcomes, particularly in forecasting the 3-month mRS. Our extensive experiments demonstrate that the proposed framework not only outperforms conventional deep learning models and standard ViT architectures but also achieves statistically significant improvements in key evaluation metrics. These results validate the efficacy of our lesion-centric approach and highlight the potential of adaptive tokenization and mutual cross attention in addressing the challenges posed by limited training data and heterogeneous clinical information.

Nonetheless, our study is subject to certain limitations. The relatively small dataset size constrained our ability to fully optimize the TCFormer architecture. Future work will focus on larger, more diverse datasets to refine our model further. Additionally, as TCFormer is capable of effectively capturing lesion tokens, we plan to investigate using lesion-specific tokens as the final classification feature, potentially improving lesion representation. Furthermore, as our approach only requires lesion center input, we aim to incorporate automated segmentation models, such as Medical SAM [30], for more efficient lesion extraction, enabling broader dataset optimization.

Overall, LC-ViT shows strong potential as a robust tool for stroke outcome prediction and could be adapted for broader medical imaging applications where precise lesion characterization is critical.

Acknowledgments. This work was supported by the RHU MARVELOUS (ANR-16-RHUS-0009) of Universite Claude Bernard Lyon-1 and by the RHU BOOSTER (ANR-18-RHUS-0001), within the program "Investissements d'Avenir" operated by the French National Research Agency (ANR). The authors also acknowledge the support of the ANR, under grant ANR-20-CE23-0013 (project PMR)

Disclosure of Interests. The authors have no competing interests.

References

1. Feigin, V. et al.: Global, regional, and national burden of stroke and its risk factors, 1990–2019: A systematic analysis for the global burden of disease study 2019. *The Lancet Neurology* 20(10), 795–820 (2021)
2. Goyal, M. et al.: Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. The Lancet, 387(10029), 1723–1731 (2016)
3. van Os, H.J.A. et al.: Predicting outcome of endovascular treatment for acute ischemic stroke: Potential value of machine learning algorithms. Frontiers in Neurology, 9, 784 (2018).
4. Saver, J.L., Filip, B., Hamilton, S., Yanes, A., Craig, S., Cho, M., Conwit, R., Starkman, S., and FAST-MAG Investigators and Coordinators: Improving the Reliability of Stroke Disability Grading in Clinical Trials and Clinical Practice: The Rankin Focused Assessment (RFA). Stroke, 41(5), 1025–1031 (2010).
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (ICLR) (2021).
6. Samak, Z.A., Clatworthy, P., Mirmehdi, M.: TranSOP: Transformer-based multimodal classification for stroke treatment outcome prediction. In: IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2023)
7. 3D Slicer. `https://www.slicer.org/`
8. Han, K., et al.: A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence. 45(1), 87–110 (2022)
9. Hatami, N., et al.: CNN-LSTM Based Multimodal MRI and Clinical Data Fusion for Predicting Functional Outcome in Stroke Patients. In: IEEE Engineering in Medicine & Biology Society (EMBC), pp. 3430–3434. IEEE (2022)
10. Zhang, Z., Jiang, R.: Triamese-ViT: A 3D-aware method for robust brain age estimation from MRIs. arXiv preprint arXiv:2401.09475 (2024)
11. Playout, C., Legault, Z., Duval, R., Boucher, M.C., Cheriet, F.: A region-based approach to diabetic retinopathy classification with superpixel tokenization. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024, vol. 15005, pp. 36–45. Springer (2024)
12. Lin, X., Wang, Z., Yan, Z., Yu, L.: Revisiting self-attention in medical transformers via dependency sparsification. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024, vol. 15011, pp. 555–566. Springer (2024)

13. Yang, Z., Chen, H., Qian, Z., Zhou, Y., Zhang, H., Zhao, D., Wei, B., Xu, Y.: Region attention transformer for medical image restoration. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024, vol. 15007, pp. 603–613. Springer (2024)
14. Painchaud, N., Stym-Popper, J., Courand, P.-Y., Thome, N., Jodoin, P.-M., Duchateau, N., Bernard, O.: Fusing echocardiography images and medical records for continuous patient stratification. arXiv preprint arXiv:2401.07796 (2024)
15. Liu, M., Hatami, N., Mechtouff, L., Cho, T.-H., Lartizien, C., Frindel, C.: Fusion of DWI image and clinical variables for stroke outcome prediction using tabular transformer. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2025).
16. Zhu, B., Shi, X., Erickson, N., Li, M., Karypis, G., Shoaran, M.: XTab: Cross-table pretraining for tabular transformers. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 43181–43204 (2023)
17. 3D-ViT: Vision Transformer 3D with one MRI type. Kaggle. Available at: `https://www.kaggle.com/super13579/vit-vision-transformer-3d-with-one-mri-type/`
18. 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. ACM, NY, USA (2016).
19. Liu, W., Qian, C., Luo, P., Ouyang, W., Zeng, W., Jin, S., Wang, X.: Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In: Proceedings of CVPR 2022
20. Zeng, W., Jin, S., Xu, L., Wang, X., et al.: TCFormer: Visual recognition via token clustering transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP**(99), 1–14 (2024).
21. Zhou, Q., Zou, H., Wang, Z., Jiang, H., Wang, Y.: Refining intraocular lens power calculation: A multi-modal framework using cross-layer attention and effective channel attention. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024, vol. 15001, pp. 754–763. Springer(2024)
22. Bui, P.-N., Le, D.-T., Choo, H.: Visual-textual matching attention for lesion segmentation in chest images. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024, vol. 15009, pp. 702–711. Springer(2024)
23. Zhang, X., Shi, E., Yu, S., Zhang, S.: DTCA: Dual-Branch Transformer with Cross-Attention for EEG and Eye Movement Data Fusion. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024, vol. 15002, pp. 141–151. Springer (2024)
24. Zhao, Y., Gu, J.: Feature fusion based on mutual cross-attention mechanism for EEG emotion recognition. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024, vol. 15011, pp. 276–285. Springer(2024)
25. Zhao, J., Li, S.: Center-to-edge denoising diffusion probabilistic models with cross-domain attention for undersampled MRI reconstruction. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024, vol. 15007, pp. 171–180. Springer(2024)
26. Fang, Y., Wang, W., Wang, Q., Li, H.-J., Liu, M.: Attention-enhanced fusion of structural and functional MRI for analyzing HIV-associated asymptomatic neurocognitive impairment. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024, vol. 15011, pp. 113–123. Springer(2024)
27. Zhang, J., Xiong, H., Jin, Q., Feng, T., Ma, J., Xuan, P., Cheng, P., Ning, Z., Ning, Z., Li, C., Wang, L., Cui, H.: A multi-information dual-layer cross-attention model for esophageal fistula prognosis. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024, vol. 15005, pp. 25–35. Springer(2024)

28. Scikit-learn developers: scikit-learn: Machine Learning in Python. Available at: `https://scikit-learn.org/stable/`.
29. Isensee, F., Schell, M., Tursunova, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.P., Heiland, S., Wick, W., Bendszus, M., Maier-Hein, K.H., Kickingereder, P.: Automated brain extraction of multi-sequence MRI using artificial neural networks. Hum. Brain Mapp. 2019, 1–13 (2019).
30. Ma, J., He, Y., Li, F. et al.: Segment anything in medical images. Nature Communications, 15, 654 (2024).