A Survey on Federated Multi-Modal Hybrid CNN-Transformer with Optimized Random Graph Dilated Diffusion Attention for Brain Tumor Segmentation and Early Classification in MRI

Analp Pathak[1], Dr. Nitish Pathak[2]
[1]Lincoln Global Postdoctoral Research (LGPR) Program, Lincoln University College, Petaling Jaya 47301, Selangor Darul Ehsan, Malaysia 1Department of Information Technology,
[1]Krishna Institute of Engineering & Technology (KIET), Ghaziabad, Delhi-NCR, Uttar Pradesh, India
[2]Associate Professor, CSE/IT, Bhagwan Parshuram Institute of Technology, GGSIPU, New Delhi, India
[1] pdf.analp@lincoln.edu.my, [1]analp.pathak@gmail.com, [2]nitishpathak@bpitindia.com

Abstract— Accurate and privacy-preserving segmentation and early categorization of brain cancers from multi-modal magnetic resonance imaging (MRI) is an important topic in computational neuro-oncology. CNNs excel in local feature extraction but not long-range context, while Transformers capture long-range dependencies but at much higher computational cost. Thus, hybrid CNN-Transformer designs have become the de facto standard for lesion delineation. Graph neural networks (GNNs) and diffusion-based generative refinement can model irregular tumor topology and denoise coarse predictions into sharper masks, while dilated and multi-scale attention mechanisms increase the effective receptive field without increasing cost. Privacy regulations have led to clinical MRI data being compartmentalized in hospitals, and federated learning (FL) has become the preferred training paradigm to construct generalizable models without centralizing raw scans. In this review, we survey recent literature (2023–2026) for five constituent technologies: hybrid CNN–Transformer segmentation, graph-attention modeling, diffusion-based refinement, dilated multi-scale attention and federated multi-modal learning. We discuss how their integration could plausibly result in a federated, multi-modal hybrid CNN–Transformer architecture with an optimized random-graph dilated diffusion attention module. We summarize representative approaches, datasets, assessment protocols, open difficulties (communication overhead, non-IID data, interpretability, computational cost) and point out possible future paths.

Index Terms—brain tumor segmentation, federated learning, hybrid CNN–Transformer, graph attention network, diffusion refinement, dilated attention, multi-modal MRI.

INTRODUCTION

There is no single sequence which can consistently distinguish enhanced tumor, edema and necrotic core, and so multi-modal MRI (T1, T1-contrast, T2, and FLAIR) is the dominant modality for the diagnosis and monitoring of gliomas and other primary brain tumors [8]. Manual demarcation of these zones is time-consuming and suffers from inter-reader variability, so automated segmentation has been a topic of research for over a decade, most famously during the yearly BraTS challenge [8].

Convolutional networks were the dominant model in this arena thru the late 2010s, owing to their local inductive bias. But a fixed receptive field makes it difficult to reason about a diffusely infiltrating tumor across a volume. Vision Transformers bridge this gap with global self-attention, losing part of the fine boundary precision that CNNs provide. The response in recent literature has been to merge the two into hybrid encoders rather than pick one [1,2]. We further extend this hybrid backbone with two additional ideas . Graph neural networks that treat the tumor as a set of connected regions instead of a regular grid, which naturally handles irregular shape [9], [10]. Diffusion-based refinement that treats an initial coarse mask as noisy and learns to denoise it to a sharper boundary [11]. Since both dense attention and diffusion sampling scale poorly with 3D volume size, dilated and multi-scale attention versions have been proposed to retain the receptive-field advantages while reducing computing [12].

This is not about data governance. Not even close. The MRI archives are stored in the local hospitals and cannot be pooled into one training set, due to rules such as HIPAA and GDPR. Federated learning, instead of exchanging images [5], trains a shared model via exchanging parameter updates. Recent works on brain tumor segmentation have shown that it can achieve the accuracy of centralized training even with non-IID and class-imbalanced client data [6], [7]. This review connects these five strands—hybrid encoders, graph attention, diffusion refinement, dilated attention, and federated training—and explores what it would take to bring them together in one design.

The stakes in the clinic makes this combination worthwhile, rather than treating each strand in isolation. A few dice points or a few percentage points of accuracy are not abstract benchmark gains, but adjustments to treatment decisions, for segmentation accuracy immediately influences radiation planning margins and classification accuracy directly affects how quickly a patient is sent to biopsy or surveillance. No single hospital has enough labeled multi-modal scans to train a large hybrid model from scratch, so the federated, multi-modal nature of this problem cannot be an afterthought tacked onto a segmentation backbone designed for a single, centrally held dataset.

RELATED WORK

A. Hybrid CNN–Transformer Encoders

CKD-TransBTS divides the four MRI sequences into two groups according to clinical knowledge, encodes each group using a CNN-Transformer branch, merges them via a modality-correlated cross-attention block, and decodes them thru a Trans&CNN calibration stage. The method achieves better performance than six CNN and six Transformer baselines on BraTS 2021 [2]. A similar approach combines a CNN encoder with a ViT branch and overlays saliency maps to maintain clinical interpretability [1]. HUT and HTTU-Net use dual CNN-Transformer or dual-CNN encoders expressly to preserve boundary accuracy for tiny lesions [3], [4]. A common pattern of these systems is that the CNN branch provides local texturing, the Transformer provides long-range context, and a learnt fusion block reconciles the two (rather than naively concatenating features). The cross-attention block in CKD-TransBTS, for instance, lets features from one modality group query the other, so the network can weight, say, FLAIR-derived edema cues against T1-contrast enhancement patterns instead of averaging them uniformly, which is closer to how a radiologist reads sequences against each other.

B. Graph-Attention Networks

Representing a tumor as a graph—voxels, super-voxels, or patches as nodes, spatial or feature similarity as edges—lets a network reason about irregular, non-convex tumor shape directly rather than through a regular convolutional grid. A graph attention network built over a region-adjacency graph reports a whole-tumor Dice of 0.91 on BraTS 2021, a clear gain over its GNN baseline [9]. VSA-GCNN pairs a variational spatial-attention graph convolutional segmenter with a Bi-GRU classifier and reports combined segmentation and classification accuracy above 99% on BraTS 2019–2021 [10]. The attention weights on each edge are what distinguish these models from an ordinary graph convolution: rather than aggregating a node's neighbors with fixed or degree-normalized weights, the network learns how much each neighboring region should influence a given node, which lets it down-weight a spurious adjacency (for instance, two regions that happen to touch but belong to different tissue types) instead of blending them by default. A random-graph construction, in which edges are sampled or optimized rather than fixed to a lattice, is a natural extension of this idea for balancing local detail against long-range topological reasoning, though we did not find a study that isolates this specific variant for brain tumor MRI.

C. Diffusion Refinement and Dilated Attention

Diffusion probabilistic models can be repurposed for segmentation by learning to reverse a noising process applied to the label mask rather than the image. NNDM couples an nnU-Net backbone with such a refinement stage, sharpening ambiguous boundaries relative to the CNN backbone alone [11]. In practice this means the network is trained to predict the noise added to a ground-truth mask at a given step and, at inference time, starts from the coarse nnU-Net prediction and iteratively removes noise toward a sharper boundary, rather than generating a mask from pure noise the way an unconditional diffusion model would. Because attention and diffusion sampling both scale poorly with 3D volume size, dilation offers a cheaper route to the same multi-scale context: EfficientMedNeXt widens its receptive field with a dilated multi-receptive-field convolutional block instead of larger kernels or denser attention, matching or beating its baseline's Dice score while cutting parameters by over 95% [12]. This pairing—diffusion for boundary sharpening, dilation for affordable context—is complementary rather than competing.

D. Federated Training Across Institutions

FedAvg, the original federated averaging algorithm, established that a shared model can be trained by exchanging local weight updates instead of raw data [5]. Applied to brain tumor segmentation, FedHG adds virtual adversarial training and a validation-weighted aggregation strategy to counter data scarcity and class imbalance across sites, and reports a 2.2-point DSC improvement over a naive federated baseline while closing most, though not all, of the gap to centralized training [6]. The validation-weighted step matters because plain FedAvg weights each client's update by its local dataset size, which can let one large but non-representative site dominate the global model; weighting instead by performance on a shared held-out validation set lets smaller, higher-quality sites still shape the aggregate. Because hospitals rarely acquire the same set of MRI sequences, federated modality-specific encoders with a shared multimodal anchor have been proposed so that a client missing one or two sequences can still contribute to, and benefit from, the global model [7]. The MICCAI FeTS challenge continues to benchmark aggregation strategies under exactly this kind of realistic, non-IID hospital data [8].

Table1. Sources surveyed in this review

Ref.	Technique	Dataset	Task
[1]	Explainable hybrid ViT+CNN	Multimodal MRI, BraTS 2019	Segmentation
[2]	CKD-TransBTS dual-branch hybrid encoder	Multimodal MRI, BraTS 2021	Segmentation
[3]	HUT hybrid UNet-Transformer	Brain lesion MRI	Segmentation
[4]	HTTU-Net dual-track hybrid CNN	Multimodal MRI	Segmentation
[5]	FedAvg federated averaging	General (foundational)	Federated training
[6]	FedHG validation-weighted aggregation	Multi-client MRI	Segmentation
[7]	Federated modality-specific encoders	Multimodal MRI, non-IID	Segmentation
[8]	FeTS federated challenge	Multi-institution MRI	Benchmark
[9]	Graph attention network (region-adjacency graph)	BraTS 2021	Segmentation
[10]	VSA-GCNN + Bi-GRU	BraTS 2019–2021	Segmentation + classification
[11]	NNDM diffusion refinement	MRI, BraTS	Segmentation
[12]	EfficientMedNeXt dilated block	BTCV, FeTA, MSD	Segmentation efficiency

REPORTED QUANTITATIVE RESULTS

This paper is a review, not a fresh empirical investigation, and hence the numbers in Table 2 are reproduced as reported in the original studies, rather than as the result of a controlled re-run. Because different studies use different datasets, splits, and preprocessing, the numbers are more a relative direction than a precise head-to-head comparison.

Table2. SELECTED RESULTS AS REPORTED IN THE SURVEYED PAPERS

Ref.	Method	Reported Result
[2]	CKD-TransBTS (hybrid)	Dice: WT 86.9%, TC 80.1%, ET 77.8% (BraTS 2021)
[9]	Graph attention network	Dice: WT 91%, TC 86%, ET 79%; HD95: 5.9/6.1/9.5 mm (BraTS 2021)
[10]	VSA-GCNN + Bi-GRU	Dice 98.6%, accuracy 99.8%, sensitivity 99.4% (BraTS 2021)
[6]	FedHG (federated)	+2.2 DSC points over baseline FL; ≈3% below centralized accuracy
[12]	EfficientMedNeXt-L	Dice 87.0% (BTCV), +1.0 pt over baseline, 96.5% fewer parameters

Hybrid and graph-based encoders both clear 85–91% Dice on the whole-tumor region, ahead of the CNN and Transformer-only baselines each paper reports internally. Federated training loses only a few Dice points to its centralized counterpart once the aggregation strategy accounts for client heterogeneity, rather than using plain averaging. And the dilated EfficientMedNeXt block shows that a large parameter reduction does not have to cost accuracy, which matters for any architecture meant to run inside hospital-grade, federated infrastructure.

A caveat applies to all three points: enhancing-tumor Dice is consistently the weakest sub-region across every method in Table II, which matches a well-known pattern in the wider BraTS literature rather than a weakness specific to any one architecture—the enhancing region is small, low-contrast against surrounding edema, and disproportionately affected by scanner and contrast-timing variation between institutions. Any of the reviewed techniques that improves whole-tumor Dice without a matching gain on enhancing tumor has therefore only solved the easier half of the segmentation problem, and a fair comparison of future federated hybrid systems should report all three sub-regions rather than an averaged score.

TOWARD A UNIFIED ARCHITECTURE

No single paper in this review combines all five components, but their individual results suggest how they would fit together. Modality-specific CNN stems, following the two-branch split used in CKD-TransBTS, would extract local features per sequence at each client site [2]. A Transformer branch would model global context, with dilation applied to its key and value projections to keep 3D compute manageable, in the spirit of the dilated block in EfficientMedNeXt [12]. A random-graph attention module, built over patch or super-voxel nodes with sparsely and adaptively sampled edges as in graph attention networks generally [9], would add non-Euclidean reasoning about tumor topology on top of the CNN–Transformer features. Then, a lightweight diffusion head refines the fused coarse prediction to get a final mask and classification score based on NNDM [11]. A validation-weighted aggregation technique in the style of FedHG would enable the federated training of the complete pipeline such that each client contributes but raw scans are not shared [6].

The composition naturally emerges from the existing strengths of each individual component: hybrid encoders balance local and global features, graph attention handles the irregular tumor topology, dilation keeps multi-scale context affordable, diffusion sharpens boundaries, and federation preserves privacy across sites. The process of constructing and testing this particular combination is still in progress.

CONCLUSION

Here is a concise integrated summary of both sections, keeping the key datasets, metrics, findings, challenges, and future directions clear and easy to follow:

1. Problem Statement / Motivation
    - Brain tumor segmentation requires accurate analysis of multiple MRI modalities, while federated learning enables hospitals to collaborate without sharing sensitive patient data.

- However, differences in modalities, scanners, patient populations, and data distributions make reliable multi-institutional segmentation challenging.

2. Method Used
    - BraTS is the main benchmark, using co-registered T1, T1ce, T2, and FLAIR images with expert tumor labels; performance is mainly measured using Dice score and 95% Hausdorff distance, with classification and federated metrics added where relevant.
    - Current approaches combine CNN–Transformer encoders, graph attention, dilated multi-scale attention, diffusion refinement, and federated training to improve segmentation while preserving data privacy.
3. Key Findings / Current Best Practices
    - Hybrid CNN–Transformer and attention-based models generally outperform simpler single-technique approaches, while federated training can achieve performance close to centralized learning.
    - A promising direction is to integrate graph attention, dilated context, and diffusion refinement within a federated multi-modal framework, but this combined architecture has not yet been systematically validated end to end.
4. Limitations and Future Work
    - Major challenges include high client-side computation, missing MRI modalities, privacy–accuracy trade-offs, heterogeneous/non-IID data, and limited validation beyond adult glioma datasets such as BraTS.
    - Future research should use standardized datasets and evaluation protocols, test pediatric tumors, metastases, and diverse scanners, and perform controlled component-by-component ablation studies to identify the actual contribution of each architectural element.

REFERENCES

[1] R. A. Zeineldin, M. E. Karar, Z. Elshaer, et al., "Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI," Sci. Rep., 2024.

[2] J. Lin, J. Lin, C. Lu, H. Chen, H. Lin, B. Zhao, Z. Shi, B. Qiu, X. Pan, Z. Xu, B. Huang, C. Liang, G. Han, Z. Liu, and C. Han, "CKD-TransBTS: Clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation," arXiv:2207.07370, 2022.

[3] W. K. Soh, H. Y. Yuen, and J. C. Rajapakse, "HUT: Hybrid UNet transformer for brain lesion and tumour segmentation," Heliyon, vol. 9, no. 12, 2023.

[4] N. M. Aboelenein, P. Songhao, A. Koubaa, A. Noor, and A. Afifi, "HTTU-Net: Hybrid two track U-Net for automatic brain tumor segmentation," IEEE Access, vol. 8, pp. 101406–101415, 2020.

[5] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. Artif. Intell. Statist. (AISTATS), 2017, pp. 1273–1282.

[6] J. Wen, X. Li, X. Ye, et al., "A highly generalized federated learning algorithm for brain tumor segmentation (FedHG)," Sci. Rep., vol. 15, art. 21053, 2025.

[7] Q. Dai, D. Wei, H. Liu, J. Sun, L. Wang, and Y. Zheng, "Federated modality-specific encoders and multimodal anchors for personalized brain tumor segmentation," in Proc. AAAI Conf. Artif. Intell., vol. 38, 2024, pp. 1445–1453.

[8] S. Pati, U. Baid, M. Zenk, B. Edwards, M. Sheller, et al., "The federated tumor segmentation (FeTS) challenge," arXiv:2105.05874, 2021.

[9] "Multi-class brain tumor segmentation using graph attention network," arXiv:2302.05598, 2023.

[10] K. Pratap Joshi, V. B. Gowda, P. Bidare Divakarachari, P. Siddappa Parameshwarappa, and R. K. Patra, "VSA-GCNN: Attention guided graph neural networks for brain tumor segmentation and classification," Big Data Cogn. Comput., vol. 9, no. 2, p. 29, 2025.

[11] S. Makanaboyina, "NNDM: NN_UNet diffusion model for brain tumor segmentation," arXiv:2510.09681, 2025.

[12] "EfficientMedNeXt: Multi-receptive dilated convolutions for medical image segmentation," in Proc. MICCAI, 2025, pt. IV.

[13] Pathak, N., Ather, D., Buhari, A., Sudhakar, A., Ramanujam, S. (2026). A Machine Learning–Based Approach for Eye State Classification Using Multichannel EEG Signals. In: Hassanien, A.E., Anand, S., Jaiswal, A., Kumar, P. (eds) Innovative Computing and Communications. ICICC 2026. Lecture Notes in Networks and Systems, vol 2039. Springer, Cham. https://doi.org/10.1007/978-3-032-30005-8_14