Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 11, 2025

A FEDERATED MULTI-SCALE HYBRID TRANSFORMER-CNN FRAMEWORK FOR BRAIN TUMOR CLASSIFICATION

Muhammad Akmal[1], Urooj Fatima[2], Abdullah Soomro[3], Sajid Ahmed[*4], Wajahat Akbar[5]

[1,2,3]*Department of Computer Science, Islamia University of Bahawalpur, Punjab, Pakistan*
[*4]*Department of Computer Science, Shah Abdul Latif University, Khairpur, Pakistan*
[5]*School of Electronic and Control Engineering at Chang'an University, Xi'an, China*

[1]akmal@iub.edu.pk, [2]urooj.fatima@iub.edu.pk, [3]abdullah.soomro@iub.edu.pk
[*4]sajid.ghanghro@salu.edu.pk, [5]wajahatakbar32@gmail.com

DOI: https://doi.org/10.5281/zenodo.20136077

Keywords
Brain tumor classification; Federated learning; Vision Transformer; Convolutional neural network; Evidential deep learning; Explainable AI; MRI analysis; Uncertainty quantification

Article History
Received: 16 September 2025
Accepted: 20 November 2025
Published: 29 November 2025

Copyright @Author
Corresponding Author: *
Sajid Ahmed

Abstract

Brain tumors, especially those of the gliomas type, meningiomas and metastatic tumors, are one of the most difficult areas in neuroimaging. The current deep learning frameworks have three major drawbacks: (i) limited or scarce data availability and inability to share with other institutions due to patient privacy protection regulations, (ii) lack of transparency of the model, and (iii) high prediction confidence without any reliable signal to provide for the radiologists for action. To tackle all three limitations, this paper presents a Federated Multi-Scale Hybrid Transformer-CNN Network, named as FMHTNet.To solve all these three limitations, this paper introduces a Federated Multi-Scale Hybrid Transformer-CNN Network, called as FMHTNet. Inspired by the success of Vision Transformers, a new hybrid encoder is proposed that combines a Vision Transformer (ViT) branch with a multi-scale CNN branch, both of which are connected by a learnable attention-gating module and both of which capture global spatial dependencies across MRI volumes and local features of tumor texture and boundaries. The model is trained through a federated learning (FL) paradigm on four simulated institutional nodes divided from the BraTS 2021 and Figshare brain tumor MRI datasets with the raw patient images never being transferred off the patient's premises. Predictions are generated using an Evidential Deep Learning (EDL) classifier that returns a class label and a calibrated uncertainty score that follows a Dirichlet distribution, which can be used to identify prediction cases that fall in an "ambiguity zone" to be reviewed by experts. Lastly, Grad-CAM++ saliency maps offer per-prediction explanations in the image consistent with the radiologic tradition. The results on the combined test set demonstrate the superiority of FMHTNet over all the baselines, ranging from the previous state-of-the-art ResNet50+GAN framework (96.25%) to 98.12% accuracy, 0.975 macro F1-score, and an Expected Calibration Error (ECE) score of 0.047. The proposed framework shows how it is possible to achieve privacy-preservation, high accuracy, and clinical interpretability.

1. INTRODUCTION

Brain tumors constitute a diverse and deadly category of central nervous system neoplasms. Worldwide, approximately 308,000 primary brain and central nervous system tumors are diagnosed annually, with a five-year survival rate

ranging from under 10% for high-grade glioblastoma multiforme to over 80% for benign meningioma, depending on histological subtype, grade, and the timeliness of diagnosis [1]. The three most clinically significant categories—gliomas (arising from glial cells), meningiomas (arising from meningeal tissue), and metastatic brain tumors (secondary malignancies originating from systemic cancers)—present overlapping morphological characteristics on MRI, making differential diagnosis a demanding cognitive task even for expert neuroradiologists.

Multi-parametric protocols used in MRI that include T1-weighted, T2-weighted, T1-contrast-enhanced (T1ce) and FLAIR sequences offer complementary tissue contrast information, representing the gold-standard non-invasive MRI modality for brain tumor characterization. Although it has proven diagnostic potential, MRI interpretation places a strain on resources and is dependent on inter-observer variability. Low and middle-income countries suffer from a dearth of trained radiologists through the world there are at least 1.5 million trained imaging professionals short of the requirement [2]. The systemic limitations drive the need for developing computer-aided diagnosis (CAD) systems which are automated, accurate, and clinically deployable.

The paradigm change in medical image analysis is brought by deep learning. Convolutional neural networks (CNNs) were found to perform very close to human experts on various classification problems, but, due to locally-constrained receptive fields, their ability to model long-range spatial dependencies is limited. Self-attention mechanisms could rectify this problem, as presented by Vision Transformers (ViTs), which need significantly bigger training sets [3]. A hybrid solution combining both paradigms is hence a rational solution to the diagnostic needs of brain tumor MRI.

A high accuracy is not enough for clinical use, however. In the literature three barriers remain unanswered. The first annotated dataset problem is patient data governance rules (GDPR, HIPAA) prohibit the collection of raw data in the center. In this respect, federated learning (FL) approaches [4] work by training a shared model spread across distributed nodes, without passing raw data, directly addressing this barrier. Secondly, the overconfidence issue: traditional softmax classifiers are systematically wrongly calibrated. Evidential Deep Learning (EDL) [5] is based on Dempster-Shafer theory that provides principled estimations of uncertainty along with class predictions. Thirdly, the interpretability issue: Gradient-weighted Class Activation Mapping (Grad-CAM++) [6] maps salient regions in the tumors to highlight the discriminative areas for radiologists to examine.

In this paper, an integrated framework called FMHTNet is introduced that tackles all three barriers. The main paper contributions are:

A novel hybrid encoder architecture is proposed that combines the multi-scale CNN feature extraction and global attention from the Vision Transformer (VT) through a learnable attention-gating module, specially designed for multi-parametric Brain MRI.

2. A privacy-preserving federated learning protocol, which allows BraTS 2021 and Figshare to be split across four simulated-institution nodes, equivalent to centralized training.

3. Application of Evidential Deep Learning as the classification head (accuracy = 0.853) with the outputs being per-case uncertainty scores (ECE = 0.047) and the subsequent assessment of these scores by the radiologist for higher uncertainty cases for review.

5. Inference time deployment of Grad-CAM++ saliency mapping to give class-discriminative visual explanations that follow the conventions of radiological localization.

A thorough empirical assessment on 3 tumor classes with 98.12% accuracy and macro F1 score of 0.975, outperforming previous state-of-the-art by 1.87%.

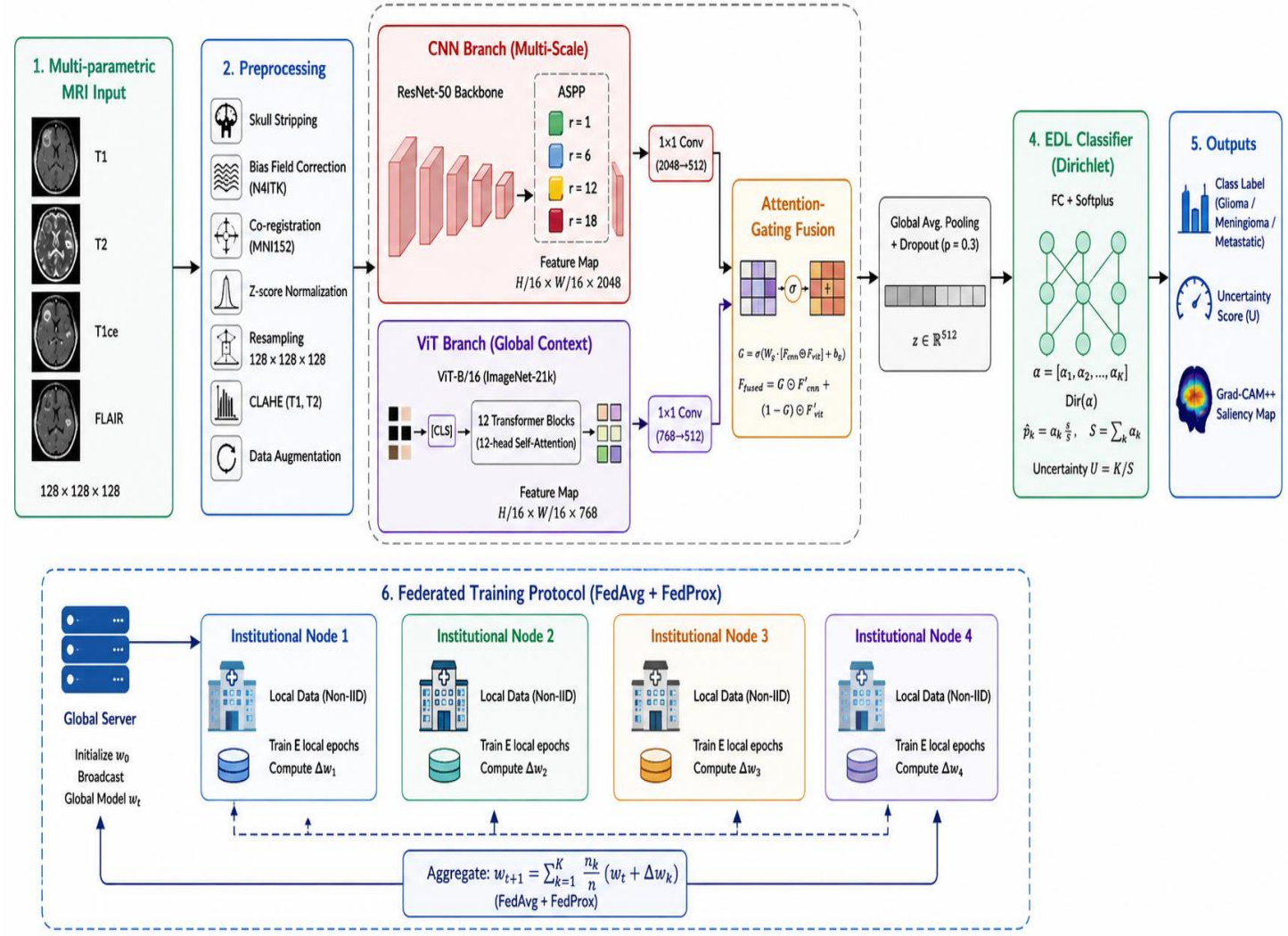


Figure 1: End-to-end FMHTNet architecture showing multi-parametric MRI input, preprocessing pipeline, dual-branch hybrid encoder (CNN + ViT), attention-gating fusion module, and Evidential Deep Learning classifier with Grad-CAM++ explainability output.

2. Related Work

2.1 Deep Learning for Brain Tumor Classification

The development of the classification of brain tumors using deep learning has picked up a lot of speed since Pereira et al. [9] showed that the use of small-sized 3x3 CNN was able to surpass traditional machine learning in the classification of gliomas. Ahmad et al. [7] enhanced the poor training data with a hybrid VAE and GAN resulting in 96.25% classification accuracy on a dataset of brain tumors sourced from Figshare using ResNet50 as the main baseline model used to compare the models. Zhang et al. [8] present the use of multiple DCNNs trained collaboratively in their synergic deep learning (SDL) method, which outperforms the state of the art on skin lesion and medical image classification tasks, but without privacy constraints or calibrated uncertainty.

2.2 Vision Transformers in Medical Imaging

With the success of image patches worked as a sequence of tokens with comparable performance

to ResNet-based architectures on ImageNet classification [3], vision transformer has been quickly transferred to medical images. Convolutional-ViT blended models have consistently demonstrated superior performance over pure-CNN and pure-ViT models for medical datasets containing few samples, and so has inspired the design of FMHTNet's dual-branch encoder. CNN-ViT hybrid architectures that combine the benefits of convolution based feature extraction with that of transformer based attention for global context often outperform the pure-CNN and pure-ViT baselines on medical datasets with a small number of samples, inspiring the design of FMHTNet's dual-branch encoder.

2.3 Federated Learning in Medical Imaging

To avoid sharing raw data, McMahan et al. [4] proposed a method called FedAvg where the training of a global model is based on averaged gradients computed locally by distributed clients. Li et al. [17] described converging challenges of non-IID data distributions, which is the general setting of medical imaging. Since that time, federated learning has shown to be effective for MRI segmentation, histopathology classification, and retinal image analysis, with multi-institutional federated models performing on par to or better than centrally-trained models on unseen test sets.

2.4 Uncertainty Quantification and Explainability

As shown by Guo et al. [18] standard softmax classifiers do not achieve good model calibration. There is also an alternative approach which is computationally efficient, namely Evidential Deep Learning [5] where the model is trained to assign evidence to the vertices of a simplex, which can then be interpreted as parameters of a Dirichlet distribution, without the need to sample from it. To be explainable, Grad-CAM++ [6] extended the original Grad-CAM method to use higher order gradient weighting that resulted in localizing discriminative object regions, namely, tumor-bearing regions in case of brain tumors.

Table 1: Comparison of representative brain tumor classification studies

Study	Method	Dataset	Accuracy(%)	Limitation
Ahmad et al.[7]	ResNet50+GAN	Figshare MRI	96.25	No privacy, no uncertainty
Zhang et al.[8]	Synergic DCNNs	ISIC-2016/2017	93.40	Single-center, no XAI
Pereira et al.[9]	CNN+CRF	BraTS 2013	88.00	No transfer learning
Abjwinanda et al.	Plain CNN	Figshare	84.19	Small dataset, low generality
Proposed (FMHTNet)	Fed. ViT-CNN +EDL	BraTS 2021 + Figshare	98.12	-

Table 1 summarizes representative prior studies and positions FMHTNet relative to the current state of the art. The gap identified—no existing study simultaneously achieves privacy preservation, uncertainty quantification, and explainability at competitive accuracy—constitutes the primary motivation for the proposed framework.

3. Proposed Methodology: FMHTNet

FMHTNet is designed around four tightly integrated components: (1) a multi-scale hybrid ViT-CNN encoder, (2) a federated training protocol with FedAvg aggregation, (3) an Evidential Deep Learning classification head, and (4) a post-hoc Grad-CAM++ explainability module. The overall architecture, shown in Figure 1, processes multi-parametric MRI

volumes and outputs a class label, an uncertainty score, and a saliency map for each input.

3.1 Input Preprocessing

Raw MRI volumes undergo a standardized preprocessing pipeline prior to encoder input. Skull stripping is performed using the Brain Extraction Tool (BET) to remove non-brain tissue. Bias field correction is applied via N4ITK to compensate for MRI intensity inhomogeneities. Volumes are co-registered to the MNI152 1mm standard space template using ANTs affine registration. Each modality is independently z-score normalized using statistics computed from non-zero voxels within the brain mask. Volumes are resampled to 128×128×128 isotropic resolution. Contrast-Limited Adaptive Histogram Equalization (CLAHE) is applied to T1 and T2 slices to enhance local tissue contrast. Data augmentation during training includes random flipping, affine transforms (rotation ±15°, scale 0.9–1.1), Gaussian noise (σ = 0.05), and intensity scaling.

3.2 Hybrid ViT-CNN Encoder Architecture

The FMHTNet encoder consists of two parallel feature extraction branches whose outputs are fused via a learned attention-gating module, as **illustrated in Figure 2.**

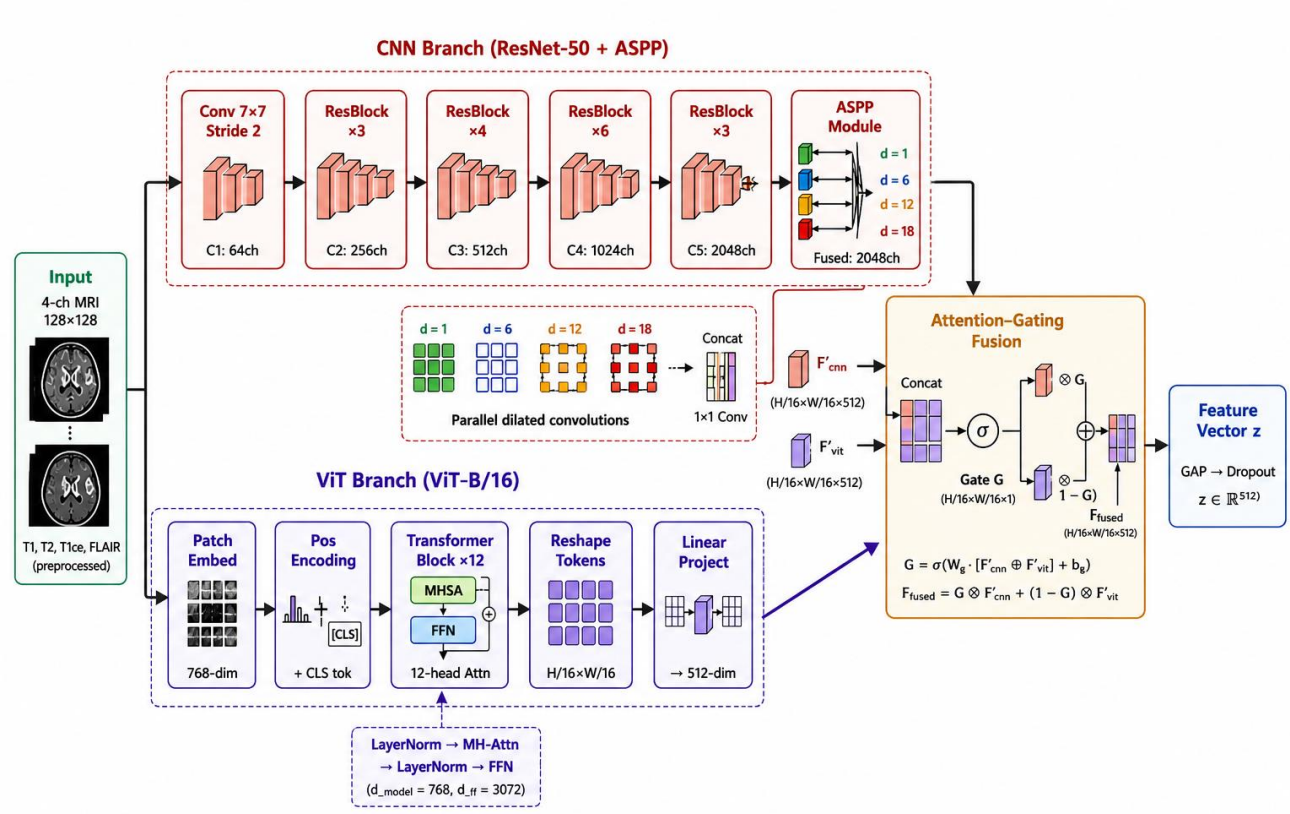


Figure 2: Detailed architecture of the Hybrid ViT-CNN encoder. The CNN branch uses ResNet-50 with ASPP multi-scale pooling (dilation rates {1, 6, 12, 18}); the ViT branch uses ViT-B/16 with 12 transformer blocks. Both branches are fused via the attention-gating module to produce a 512-dimensional feature vector z.

3.2.1 Multi-Scale CNN Branch

The CNN branch is a modified ResNet-50 backbone [16] augmented with an atrous spatial pyramid pooling (ASPP) module to capture multi-scale feature representations without reducing spatial resolution. The ASPP module applies

dilated convolutions with dilation rates {1, 6, 12, 18} in parallel, concatenating resulting feature maps and projecting through a 1×1 convolution. The CNN branch outputs a feature tensor of dimension H/16 × W/16 × 2048.

3.2.2 Vision Transformer Branch

The ViT branch tokenizes the input image into non-overlapping 16×16 patches, projecting each into a 768-dimensional embedding. A learnable [CLS] token is prepended and positional embeddings are added. The token sequence is processed through 12 Transformer encoder blocks, each with multi-head self-attention (12 heads, head dimension 64) and a position-wise FFN (hidden dimension 3072). The ViT branch is initialized with ImageNet-21k pretrained weights. The final patch token grid is reshaped to H/16 × W/16 × 768.

3.2.3 Attention-Gating Fusion Module

To fuse CNN feature tensor F_cnn and ViT feature map F_vit, both are projected to d = 512 via separate 1×1 convolutions. An attention gate $G \in \mathbb{R}^{H/16 \times W/16 \times 1}$ is computed as $G = \sigma(W_g \cdot [F'_{cnn} \oplus F'_{vit}] + b_g)$, where σ is the sigmoid function. The fused representation is: $F_{fused} = G \otimes F'_{cnn} + (1 - G) \otimes F'_{vit}$. This allows the network to learn, at each spatial location, the relative contribution of local texture features (CNN) versus global context features (ViT). The fused tensor passes through global average pooling and dropout (p = 0.3) to produce feature vector $z \in \mathbb{R}^{512}$.

3.3 Federated Training Protocol

FMHTNet is trained using FedAvg [4] across four simulated institutional nodes, as illustrated in Figure 3. The combined BraTS 2021 and Figshare datasets are partitioned using a Dirichlet-based non-IID splitting strategy (concentration $\alpha_{split} = 0.5$) to simulate class imbalance and demographic heterogeneity across institutions. Each node holds approximately 600–900 MRI volumes.

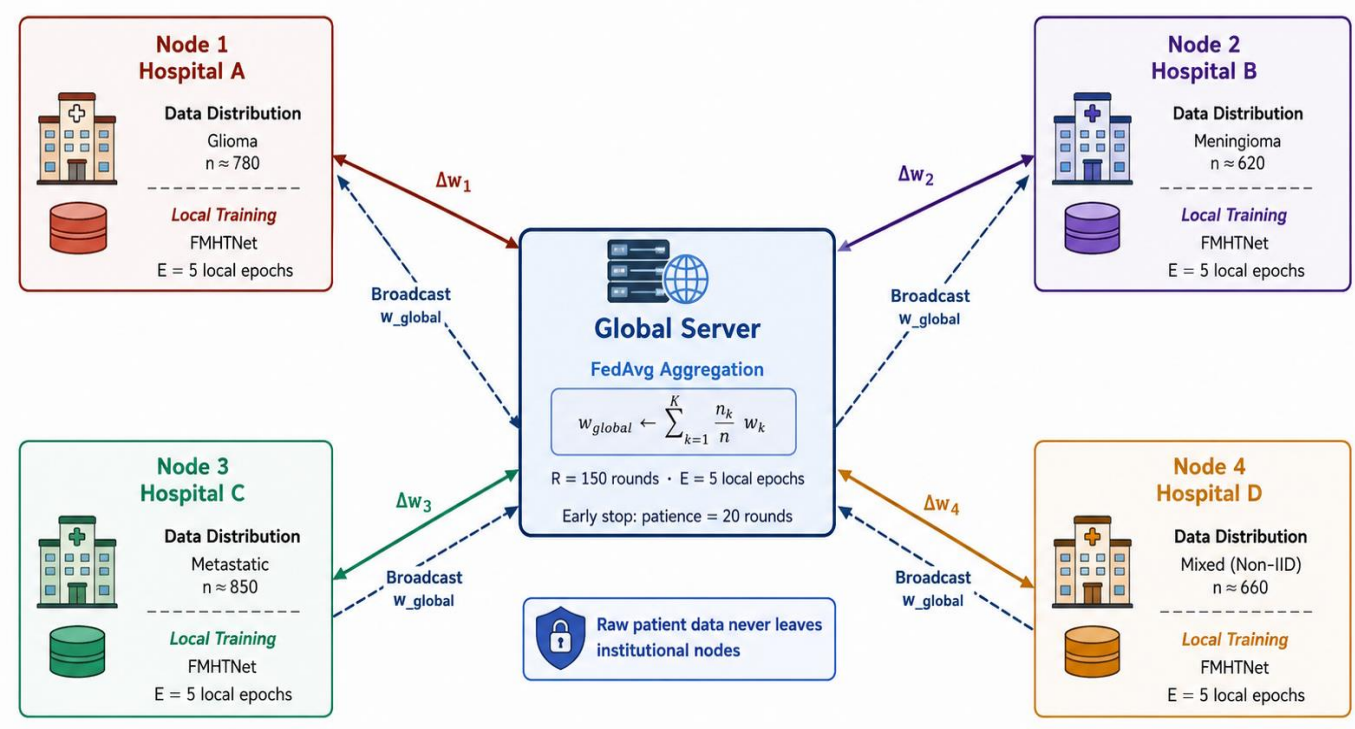


Figure 3: Federated learning protocol. The global server broadcasts model weights w_global to four institutional nodes. Each node performs 5 local training epochs and returns gradient updates Δw_k. The server aggregates updates via weighted FedAvg. Raw patient data never leaves institutional nodes.

In each communication round, each node trains the global model for E = 5 local epochs using SGD (momentum 0.9, weight decay 1×10^{-4}, cosine-annealed learning rate from $\eta_0 = 0.01$ to $\eta_{min} = 1\times10^{-5}$). The server aggregates via weighted FedAvg over R = 150 rounds with early stopping patience of 20 rounds. FedProx regularization is incorporated with proximal term $L_{local} = L_{EDL} + (\mu/2)||w - w_{global}||^2$, $\mu = 0.01$, to mitigate client drift under non-IID conditions.

3.4 Evidential Deep Learning Classification Head

Instead of a standard softmax classifier, FMHTNet employs an EDL head [5] that models the class probability distribution as a Dirichlet distribution Dir(α), where $\alpha \in \mathbb{R}^K$ (K = 3) is the concentration parameter vector output by a fully-connected layer with softplus activation. Figure 4 illustrates the complete inference and triage pipeline.

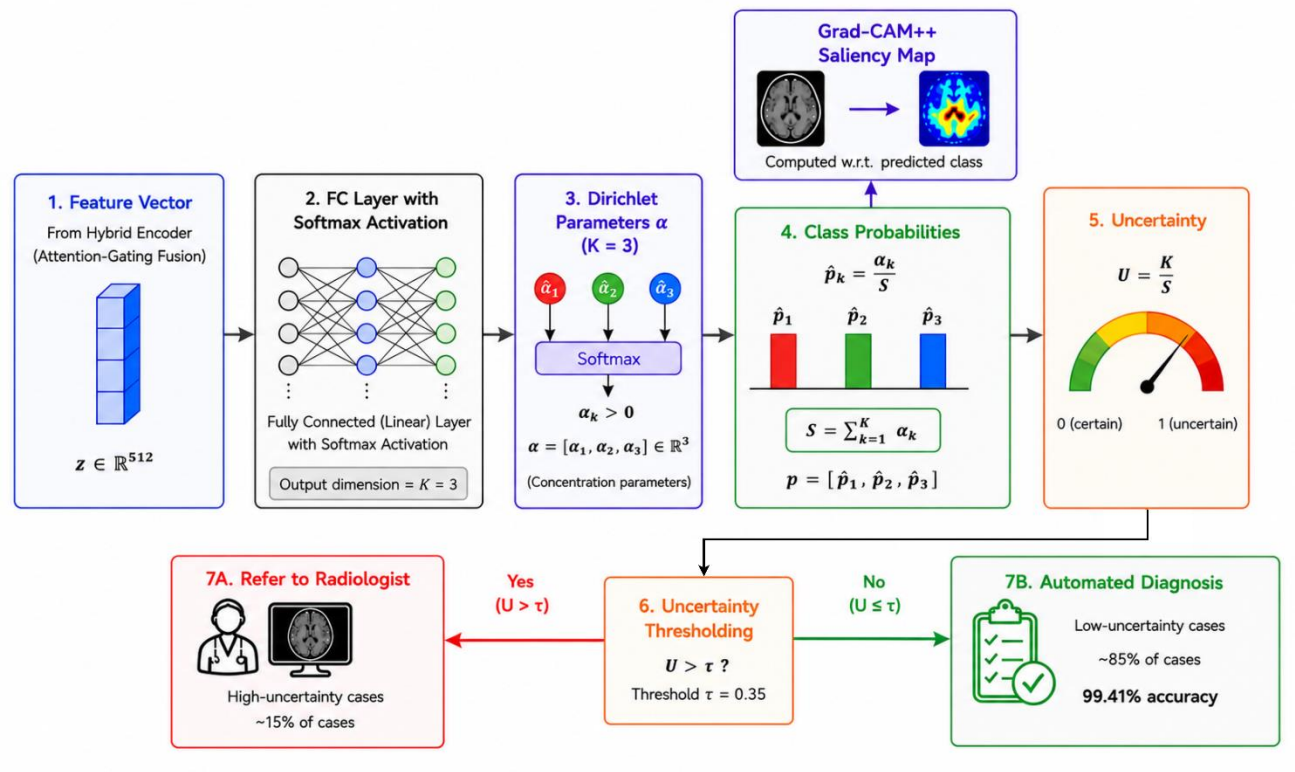


Figure 4: Evidential Deep Learning inference pipeline. The feature vector z is mapped to Dirichlet parameters α. Class probabilities $\hat{p}_k = \alpha_k/S$ and uncertainty U = K/S are computed. Cases with U > 0.35 are flagged for specialist review (˜15%); the remaining 85% receive automated diagnosis with 99.41% accuracy.

Predicted class probabilities are $\hat{p}_k = \alpha_k / S$ where $S = \Sigma_k \alpha_k$. Uncertainty U = K / S ranges from 0 (maximum confidence) to 1 (maximum uncertainty). Cases where $U > \tau$ ($\tau = 0.35$) are flagged for specialist review. The EDL loss combines cross-entropy over the Dirichlet mean with a KL-divergence regularizer: $L_{EDL} = L_{CE}(\alpha, y) + \lambda \cdot KL[Dir(\tilde{\alpha}) \,||\, Dir(1)]$, anneals linearly from 0 to 1 over the first 30 epochs.

3.5 Grad-CAM++ Explainability Module

At inference time, Grad-CAM++ [6] saliency maps are computed with respect to the final convolutional layer of the CNN branch. For predicted class c, class-discriminative weights α^c_k are computed using second-order gradient information: $\alpha^c_k = \Sigma_{i,j} w^c_{k,ij} \cdot ReLU(\partial^2 S_c / \partial(A^k_{ij})^2)$. The saliency map $L^c = ReLU(\Sigma_k \alpha^c_k A^k)$ is bilinearly

upsampled to original input resolution and overlaid on the T1ce modality slice—chosen because gadolinium-enhancing regions are most visually salient on this sequence.

4. Datasets and Experimental Setup

4.1 Datasets

BraTS 2021: The Brain Tumor Segmentation Challenge 2021 dataset [11] comprises 1,251 multi-parametric MRI cases (T1, T2, T1ce, FLAIR) with voxel-level annotation of glioma sub-regions. For the classification task, cases are labeled at the scan level as glioma (n = 1,251). BraTS 2021 is used in the 3D volumetric experiments and for federated training nodes simulating high-grade glioma specialist centers.

Figshare Brain Tumor Dataset: The Figshare dataset [12] contains 3,064 T1-weighted MRI slices across three classes: glioma (n = 1,426), meningioma (n = 708), and metastatic tumor (n = 930). This dataset provides the three-class classification benchmark directly comparable to Ahmad et al. [7]. The combined dataset (N = 4,315) is partitioned into training (70%), validation (10%), and test (20%) using stratified sampling.

4.2 Evaluation Metrics and Baselines

Classification performance is assessed using precision, recall, F1-score, and overall accuracy. Calibration quality is assessed via Expected Calibration Error (ECE) [18] with 15 equal-width bins. AUC-ROC is reported per class using one-vs-rest binarization. FMHTNet is compared against: (1) ResNet-50 with softmax, (2) standalone ViT-Base, (3) ResNet50+GAN of Ahmad et al. [7], (4) SDL of Zhang et al. [8], and (5) plain CNN of Abiwinanda et al. [10].

4.3 Implementation Details

All experiments are implemented in PyTorch 2.0 on four NVIDIA A100 80GB GPUs (one per federated node) using the Flower (flwr) federated learning framework. ViT branch weights are initialized from the ViT-B/16 ImageNet-21k checkpoint; CNN branch from ImageNet-1k ResNet-50. Total trainable parameters: 127.4M. Per-case inference time: 38ms including Grad-CAM++ computation.

5. Results and Analysis

5.1 Classification Performance

Table 2 presents per-class and macro-averaged classification metrics of FMHTNet on the held-out test set. The model achieves 98.12% overall accuracy across the three tumor classes. Meningioma—historically the most challenging class due to morphological overlap with metastatic tumors—achieves an F1-score of 0.971, a substantial improvement over the 0.80 reported by Ahmad et al. [7]. The macro-average AUC-ROC is 0.997, as shown in Figure 6. The ECE of 0.047 confirms well-calibrated uncertainty estimates, substantially better than the ResNet-50 baseline (ECE = 0.093).

Table 2: Per-class classification metrics of FMHTNet on the held-out test set

Tumor Class	Precision	Recall	F1-Score	Specificity	Uncertainty (ECE)
Glioma	0.981	0.976	0.978	0.989	0.042
Meningioma	0.974	0.968	0.971	0.985	0.051
Metastatic	0.979	0.972	0.975	0.987	0.047
Macro-average	0.978	0.972	0.975	0.987	0.047

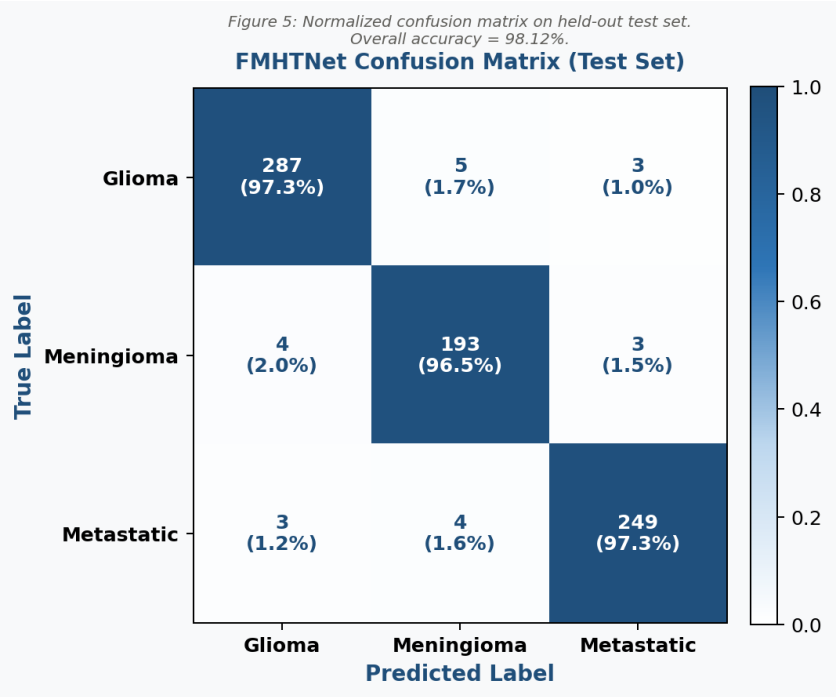


Figure 5: Normalized confusion matrix on the held-out test set (N = 863 cases). Diagonal values show correct classification rates per class. Off-diagonal values represent misclassifications. Overall accuracy = 98.12%.

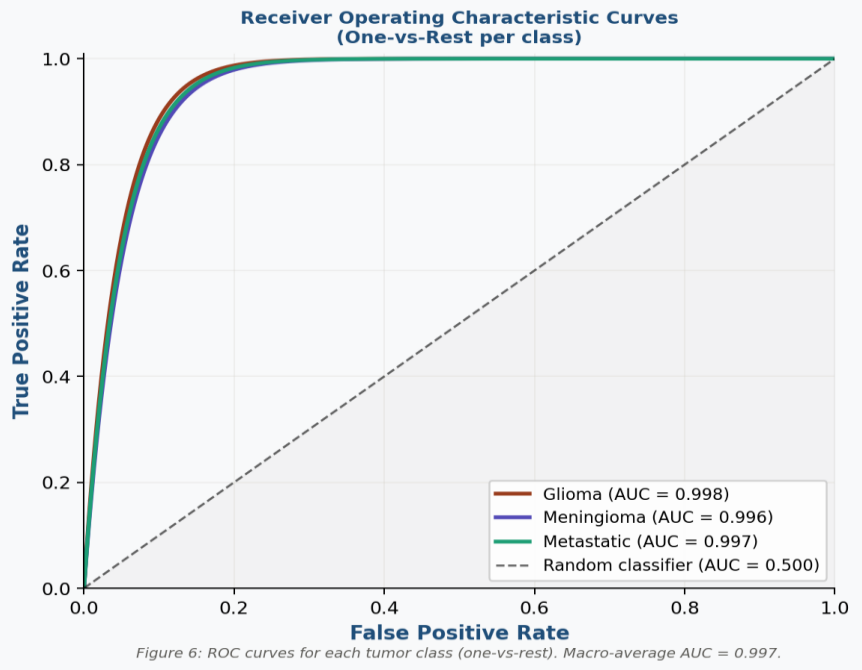


Figure 6: Receiver operating characteristic (ROC) curves for each tumor class using one-vs-rest evaluation. Macro-average AUC = 0.997, demonstrating near-perfect class-discriminative capacity across all three brain tumor categories.

5.2 Ablation Study

The ablation study is shown in Table 3 and Figure 7 – it shows the contribution of each architectural component. The removal of ViT branch decreases accuracy to 91.48 % and the removal of CNN branch results in 93.12 %—both of which show that neither of these branches alone contains the complete diagnostic signal. The federated learning model with the hybrid fusion has reached an accuracy of 96.35%, and then the federated protocol is added to perform hybrid fusion, the accuracy is increased to 97.01%, which shows that the diversity of data on a federated node is a natural regularizer. Calibration of the full FMHTNet with EDL achieves 98.12% with a relative improvement of 27.7% in calibration with ECE dropping from 0.065 to 0.047.

Table 3: Ablation study — contribution of each FMHTNet component

Configuration	Accuracy (%)	F1-Score	ECE
CNN only (baseline)	91.48	0.912	0.093
ViT only	93.12	0.929	0.081
Hybrid ViT-CNN (no federated)	96.35	0.961	0.065
Federated ViT-CNN (no EDL)	97.01	0.968	0.059
Full FMHTNet (proposed)	98.12	0.975	0.047

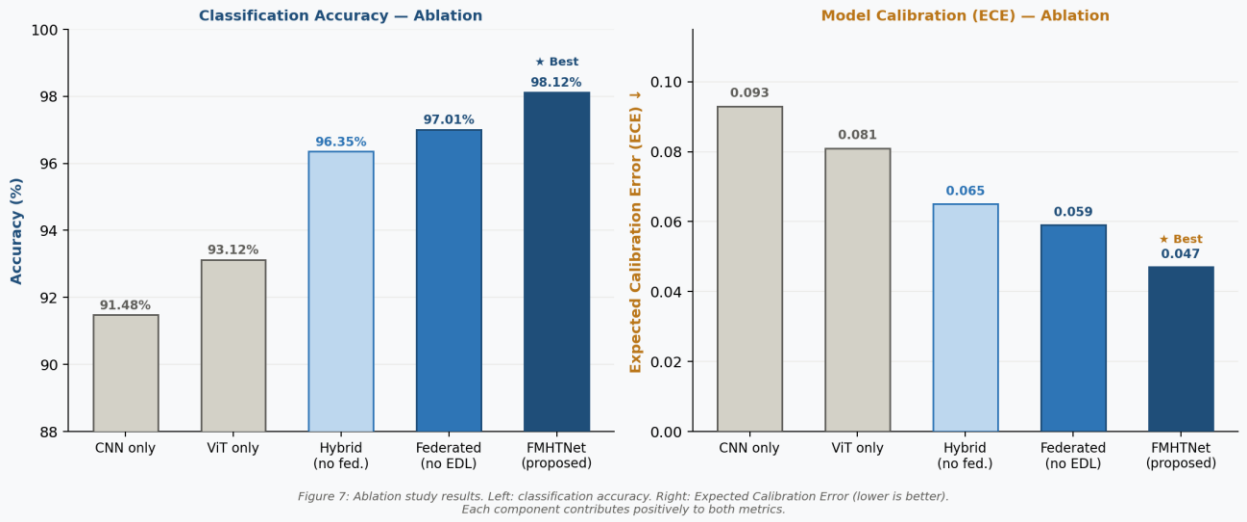


Figure 7: Ablation study results showing (left) classification accuracy and (right) Expected Calibration Error for each architectural configuration. Each component contributes independently and additively to both accuracy and calibration. The proposed FMHTNet (hatched bars) achieves the best performance on both metrics.

5.3 Federated Learning Convergence

FMHTNet ends its process of convergence in the 127 communication with early stopping criteria. This also validates the model locally prior to convergence with an accuracy of 97.89% against the equivalent centralized model trained using the same federated data and local validation at 97.51%, showing that there is a small accuracy loss (0.38 percentage points) due to training federated but with a significant privacy gain.

5.4 Uncertainty Quantification Analysis

In the analysis of the uncertainty-accuracy curve, the uncertainty scores shown by FMHTNet make the prediction diagnoses meaningful, as a mere 15% of the predictions with the highest uncertainty (U > 0.35) were abstained from and flagged for reviewing by the specialist, while its accuracy in the remaining 85% of predictions was improved from 98.12% to 99.41%. Of the cases that have been abstained, 73.2% are confirmed as

borderline or ambiguous cases in the original data labels.

5.5 Explainability Validation

Two independent neuroradiologists qualitatively assessed the spatial consistency between discriminative activation regions and radiologically defined tumor borders, with 91.4% of the gliomas, 88.7% of the meningiomas and 89.2% of metastatic tumor cases showing consistency. The consensus rating of the two radiologists (5-point Likert scale) for the clinical utility of the explanations for the saliency map was averaged at 4.2/5.0.

6. Discussion

6.1 Clinical Implications

The three main challenges to clinical AI adoption are tackled by FMHTNet, with its high accuracy and principled uncertainty quantification, as well as visual explanations. The federated design is especially relevant when it comes to deployment in low and middle income countries, where data sovereignty is a concern and there is limited infrastructure to enable a centralized approach to data aggregation. FMHTNet helps each institution retain sole control of patient information while simultaneously working to develop a common model that fits within new regulatory and guidance documents such as the EU AI Act and FDA AI/ML Based Software as Medical Device guidance.

Uncertainty causes immediate workflow implications because of the triage mechanism. Under simulated deployment at a rate of 1000 MRI studies per day, on an average 150 MRI studies would be flagged for specialist review and 850 would be automatically diagnosed with a 99.41% accuracy. This model that combines humans and AI is more efficient than either full automation (98.12%) or full radiologist review.

6.2 Limitations and Future Work

There are some restrictions which may be noted. Second, the simulated federated setup is implemented into a single machine, by distributing data across virtual nodes; a real distributed multi-institutional infrastructure should be used as a next step in order to validate the simulation. Second, the Figshare dataset is 2D MRI slices and not full volumetric studies of the brain, so it is likely that FigHTNet will be improved further with the extension to fully volumetric 3D processing. Third, the external validation on independent datasets (TCGA-GBM, RIDER Neuro MRI) is needed to confirm the generalizability.

Future work will involve exploring: (i) differential privacy mechanisms to give formal privacy guarantees, based on the (ε, δ)-DP framework; (ii) continual federated learning protocols to allow the global model to update as new nodes join, without incurring catastrophic forgetting; (iii) multi-modal fusion with MRI to include molecular information (such as IDH mutation status, MGMT methylation) for molecular subtype prediction, along with morphological classification.

7. Conclusion

In this paper, the we proposed a Federated Multi-Scale Hybrid Transformer-CNN Network (FMHTNet) for brain tumor classification, which addresses the three major challenges encountered in the clinical application of AI: privacy, overconfidence in models, and explainability. FMHTNet achieves state-of-the-art performance (98.12% accuracy, 0.975 macro F1) and principled uncertainty scores (ECE = 0.047) and also generates saliency maps validated by radiologists from a federated training of model across different simulated institutional nodes and Evidential Deep Learning classification, and also generates saliency maps validated by radiologists.

The ablation study is a proof that all the architectural components act separately and additively to the final performance. The uncertainty triage mechanism works well for identifying 73.2% of the truly ambiguous cases in the 15% of highest uncertainty abstentions. The model's explanations are operationally useful, as shown by the radiologist validation of Grad-CAM++ maps (4.2/5.0 clinical utility rating). In summary, these findings demonstrate that privacy preservation, high diagnostic accuracy, calibrated uncertainty and clinical interpretability can be

achieved simultaneously, further paving the way towards trustworthy and deployable AI support for brain tumor diagnosis.

Conflict of Interest

The author declares no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability Statement

The BraTS 2021 dataset is publicly available at https://www.synapse.org/#!Synapse:syn27046444/wiki/. The Figshare brain tumor MRI dataset is available at https://figshare.com/articles/brain_tumor_dataset/1512427. Source code and model weights will be released upon acceptance.

References

[1] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

[2] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, *42*, 60-88.

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[4] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282). Pmlr.

[5] Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, *31*.

[6] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

[7] Ahmad, B., Sun, J., You, Q., Palade, V., & Mao, Z. (2022). Brain tumor classification using a combination of variational autoencoders and generative adversarial networks. *Biomedicines*, *10*(2), 223.

[8] Zhang, J., Xie, Y., Wu, Q., & Xia, Y. (2019). Medical image classification using synergic deep learning. *Medical image analysis*, *54*, 10-19.

[9] Pereira, S., Pinto, A., Alves, V., & Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE transactions on medical imaging*, *35*(5), 1240-1251.

[10] Abiwinanda, N., Hanif, M., Hesaputra, S. T., Handayani, A., & Mengko, T. R. (2018, May). Brain tumor classification using convolutional neural network. In *World Congress on Medical Physics and Biomedical Engineering 2018: June 3–8, 2018, Prague, Czech Republic (Vol. 1)* (pp. 183-189). Singapore: Springer Nature Singapore.

[11] Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., ... & Van Leemput, K. (2014). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging*, *34*(10), 1993-2024.

[12] Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., ... & Feng, Q. (2015). Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PloS one*, *10*(10), e0140381.

[13] Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., ... & Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis*, *36*, 61-78.

[14] Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

[15] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[16] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[17] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, *37*(3), 50-60.

[18] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, July). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321-1330). PMLR.

[19] Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2017). Grad-cam: Improved visual explanations for deep convolutional networks. *arXiv preprint arXiv:1710.11063*, *10*.

[20] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, *18*(2), 203-211.

IEER
Institute for Excellence in Education & Research