An Explainable Hybrid CNN–Transformer Framework for Multiclass Brain Tumor Classification Using MRI Imagery

J. Chinna Babu[1,2], Sai Kiran Oruganti[3]

[1]Post-Doctoral Researcher, Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan, Malaysia. **(Student ID: LUC/CPGS/PDF/2025291003)** (**Application Number: LUC/CPGS/PDF/2025291003)**

[2]Associate professor, Department of Electronics and Communication Engineering, Annamacharya University, Rajampet, Andhra Pradesh, India

[3]Associate Professor, Lincoln University College, Malaysia

[1,2]jchinnababu@gmail.com, [1,2]pdf.jchinnababu@lincoln.edu.my, [3]saisharma@lincoln.edu.my, [3]saikiran.oruganti@gmail.com

[1,2]0000-0001-9005-0615

Abstract: Brain tumor analysis is a tough challenge, because there is a lot of variation in Magnetic Resonance Imaging (MRI) and the requirement for global context reasoning for accurate multiclass classification. This paper introduces a hybrid model for multiclass tumor classification named IntegIVRV-Net, which combines InceptionV3, ResNet50, and a Vision Transformer (ViT-L16) to address the challenges of extracting multi-scale features, conducting deep residual structural learning, and modeling global context. IntegIVRV-Net, a hybrid model combines InceptionV3 for multi-scale feature extraction, ResNet50 for deep residual structural learning and ViT-L16 for global context modeling in a single network was proposed in this paper. This work emphasizes that the current literature on CNN, Vision Transformer and hybrid segmentation/classification methods on the BraTS, Kaggle, and Figshare benchmarks sheds light on a series of limitations that hinder the reliability and deployability of these methods in clinical practice, such as the heavy dependency on computational power and data availability in transformer-based pipelines, single-metric evaluation failing to capture clinically relevant failure modes, and a lack of studies incorporating rotational invariance with global context modeling, along with a long-standing disconnect between interpretability and deployment readiness. To overcome these deficiencies, the proposed methodology includes a complete pipeline: intensity normalization and skull-striping preprocessing, three-branch feature extraction (InceptionV3, ResNet50, ViT-L16), feature fusion with the help of refinement head, explainability methods (Grad-CAM, attention), and edge–cloud deployment with encrypted transmission (ONNX/TFLite). The framework is designed to be evaluated on benchmark datasets, and the paper highlights the knowledge gaps and presents the entire proposed methodology, experimental validation, and results will be presented in a later stage of this study.

Keywords: Brain Tumor; Hybrid CNN-Transformer; InceptionV3; Vision Transformer; Edge-Cloud Deployment.

Introduction

Diagnosing brain tumors is crucial since they are among the most dangerous neurological conditions. MRI is more effective distinction of image types of brain tumors providing high-quality soft tissue-based imaging [1]. However, manual interpretation is time-consuming, and the heterogeneous form, irregular morphological characteristics, and overlapping characteristics of the tumors can lead to variability among observers in MRI characterization.

Deep learning has made a great contribution to the automatic brain tumor classification by accomplishing discriminative features. InceptionV3 and ResNet50 are effective to model local and multi-scale spatial features, but they usually fail to adopt appropriate relationships. Similarly, vision Transformers are extensive computational resources which impede their real-time clinical application [2-3].

Hybrid CNN–Transformer architectures have recently been developed that take advantage of both models, resulting in better classification accuracy. Despite this, the current approaches have several drawbacks: they are computationally complex, they only measure the model's accuracy, they have few integration tools for explainability, and they lack support for their deployment on the edge, especially in resource-constrained settings [4].

To overcome these challenges, IntegIVRV-Net is proposed with ViT-L16, combining their features in a unified feature fusion strategy. The framework also includes intensity normalization, skull stripping, explainability through Grad-CAM and secure edge–cloud deployment via ONNX/TFLite. The identified research gaps are presented in this paper, and experimental validation has been left for later work and is not included in this paper.

Related Work

Hatamizadeh et al. [5] introduced a Transformer-based brain tumor MRI analysis approach in 2026, which successfully utilized the self-attention mechanism to extract long-range contextual information. While this approach has yielded high Dice scores, it has also introduced challenges in terms of the extensive use of annotated large-scale datasets and the significant computational resources required, which are not ideal for practical, real-time, and resource-efficient clinical applications.

In 2026, Chen et al. [6] presented a multiclass brain tumor segmentation model, where the CNN was used to capture local spatial patterns, and the Vision Transformer to model the global context. Although the framework achieved high accuracy in classification of benchmark MRI datasets, it was shown to be inefficient in terms of computation load, requirement for extensive training data, and the limited scope of comprehensive experiments conducted under various clinical imaging conditions.

Khan et al. [7] proposed the deep learning model with EfficientNet for multiclass brain tumours classification in MRI images in 2026. Efficient feature extraction and transfer learning enabled the proposed model to achieve a very high level of classification accuracy. However, the CNN-only architecture mainly learnt local representations, so could not model long-range dependencies and was not robust to morphological and orientation changes in the tumor across datasets.

Wang et al. [8] established a EAI technique in diagnose brain tumours in 2026. The model achieved satisfactory classification results and visual explanations using attention maps. The framework, however, was still rather computationally heavy, didn't get a quantitative measure of explainability, and had no consideration of deployment requirements like lightweight inference or edge-based implementation.

Research Gap: Recent work has made great progress in automated brain tumor analysis but there are several issues which are not yet addressed. Existing Transformer-based models demand high computational resources and large annotated datasets, making it difficult to implement. CNN-based approaches tend to be ineffective in incorporating full global context information and many hybrid approaches are still high in terms of computational cost. Moreover, most of the studies focus on overall accuracy without considering the important aspect of clinically relevant failure cases and explainability is not always coupled with deployment-ready edge intelligence.

Proposed Methodology

The proposed IntegIVRV-Net is a hybrid CNN–Transformer network which uses a multi-scale feature extraction technique, deep residual learning and global contextual modelling with a single network as depicted in Figure 1. The entire workflow involves the preprocessing, three parallel feature extraction branches, feature fusion, classification, explainability, and secure deployment on edge–cloud. The methodology is tailored to tackle the key research gaps identified in the literature, namely: computational efficiency, contextual feature learning, explainability and deployment readiness.

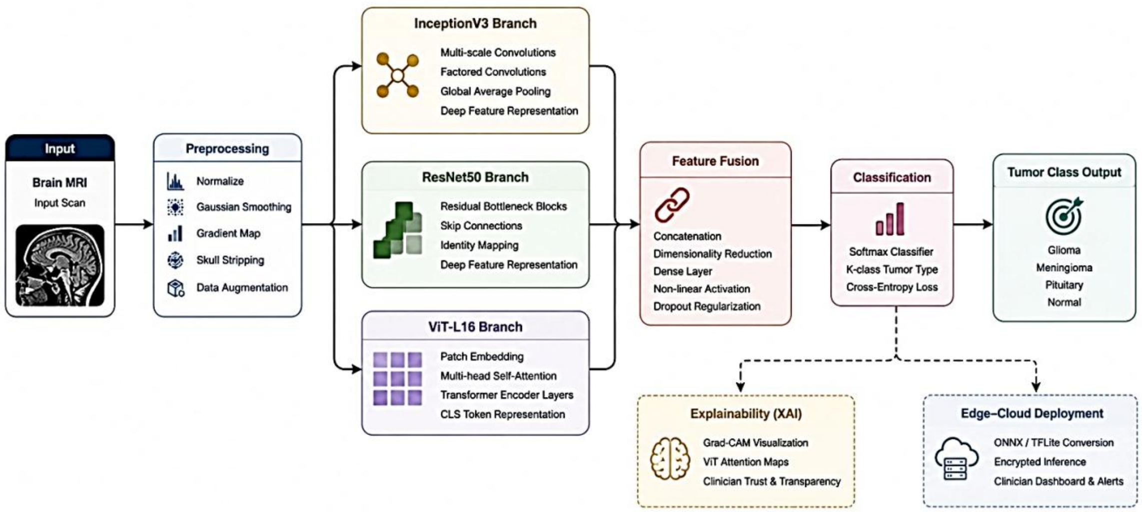

Figure 1: Proposed Methodology Architecture

A. MRI Image Preprocessing

Prior to feature learning, the input brain MR images are pre-processed to enhance image quality and to remove unwanted variations in the images. Intensity normalization makes it possible to normalize the pixel values from various MRI scanners to reduce intensity differences. Skull stripping: Remove non-brain tissues (skull and scalp) and keep only brain region to discard unnecessary information. The Gradient Map (GM) enhancement helps to better discriminate the different tumor classes by improving the visibility of the tumor boundaries and structure

B. Multi-Scale Feature Learning with InceptionV3

The first feature extraction branch is based on the InceptionV3 network which is used to extract representative features from the pre-processed MRI images. It has several parallel convolutional filters of different sizes to learn fine local features and larger structure features of the tumor. With this multi-scale processing, tumors of different shapes, sizes and textural properties can be effectively characterized while maintaining important spatial information. Then, a Global Average Pooling (GAP) layer is used to combine the feature maps extracted from each of the previous layers into a feature vector that contains only the most discriminative information.

C. Deep Residual Structural Learning using ResNet50

The second branch adopts ResNet50 to learn deep structural representation of brain tumor by using residual learning. They can learn very discriminative features without losing on accuracy thanks to the residual connections which facilitate the efficient propagation of gradients. This branch deals with the recognition of complex tumor textures, anatomical structures and boundaries characteristics which are crucial for multiple tumor categorization. These deep residual features are extracted from the main branch of the InceptionV3 network, complementing the multi-scale information gained from the different branches, giving a richer representation of features without any instability in terms of computation. The Global Context Modeling module is implemented using Vision Transformer (ViT-L16).

D. Vision Transformer (ViT-L16)

The third branch uses the Vision Transformer Large (ViT-L16) model to learn long-range spatial relations, which are difficult to model using conventional CNNs. The MRI image is split into fixed-size patches that are embedded into feature tokens and then processed with multi-head self-attention. This allows the network to form connections between remote areas of tumor and gain a global context learning the overall context. Combining contextual reasoning based on transformer models with local features generated by CNN alleviates one of the limitations of traditional CNN architecture, namely the limited receptive field, while providing a comprehensive understanding of the characteristics of the tumor [9].

E. Organization of Groups and Stages

The feature vectors obtained from the 3 branches (InceptionV3, ResNet50, and ViT-L16), are combined to form a single vector with complementary local, structural and global contextual information. A dense feature refinement module further learns the interaction between these heterogeneous representations, while pruning the representations. This fusion approach allows the framework to leverage the best capabilities of each of the three branches without depending solely on the potentially high computation costs of transformer processing and achieves a trade-off between classification performance and computation efficiency.

F. Multiclass Brain Tumor Classification

The refined feature representation is given to the classification layer where the Softmax classifier classifies the brain MRI into one of 4 categories. The global context is used to enhance the feature representation, which enhances discrimination of visually similar tumor classes. Therefore, the proposed framework can reliably predict multiclass data, and it minimizes the misclassification risk due to morphological similarity between tumors.

G. Explainable Artificial Intelligence

To make clinical decisions more transparent, the proposed method includes Grad-CAM and Transformer attention visualization techniques for explainable decision making. In contrast to Grad-CAM which identifies the regions of tumors that most drive the classification result, the attention maps highlight the wider contextual regions emphasized by the Vision Transformer. The proposed approach is compatible with the traditional trade-off between classification accuracy and clinically relevant interpretation, which is a time-tested problem as highlighted in earlier works [10].

H. Edge–Cloud Deployment

The proposed model is then converted into lightweight deployment formats such as ONNX and TensorFlow Lite (TFLite) for efficient deployment on resource constrained edge devices. Secure communication ensures that patient MRI data remains confidential and secure during transmission, maintaining data integrity and confidentiality. Edge inference helps make quick decisions to diagnose patients with minimal latency, and the cloud infrastructure facilitates centralized monitoring, clinical data management and periodical model updates. The hybrid edge–cloud deployment architecture provides greater scalability, decreases the workload on the centralized servers, and makes the architecture more practical to be applied to real-world clinical environments.

Conclusion

In this paper, a hybrid CNN–Transformer network called IntegIVRV-Net was proposed. A thorough review of previous research identified common problems such as high computational cost, requirement of large datasets with annotations, lack of integration of local and global feature representations, lack of model interpretability, and no consideration of deployment in real clinical settings. To overcome these drawbacks, we propose to use a multi-scale feature extractor InceptionV3, a residual feature learning ResNet50 and a global contextual information extractor ViT-L16 in the proposed framework. In addition, Grad-CAM and Transformer attention visualization are added to enhance transparency of model predictions, and ONNX and TensorFlow Lite (TFLite) provide an efficient edge–cloud deployment framework that is appropriate for resource-constrained healthcare systems. Proposed approach will be validated on benchmark datasets. The results of the experiment in the form of quantitative performance analysis, class-wise evaluation, comparative analysis with the existing approaches, explainability analysis, and deployment performance analysis will be presented in the next step of this research.

REFERENCES

1. S. Bauer, R. Wiest, L.-P. Nolte, and M. Reyes, "A survey of MRI-based medical image analysis for brain tumor studies," Physics in Medicine & Biology, vol. 58, no. 13, pp. R97–R129, Jul. 2013. https://doi.org/10.1088/0031-9155/58/13/R97
2. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, Dec. 2017. https://doi.org/10.1016/j.media.2017.07.005
3. K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, **"A Survey on Vision Transformer,"** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, Jan. 2023, https://doi.org/10.1109/TPAMI.2022.3152247
4. Aburass, S., Dorgham, O., Al Shaqsi, J. *et al.,* "Vision Transformers in Medical Imaging: a Comprehensive Review of Advancements and Applications Across Multiple Diseases", J Digit Imaging. Inform. med. **38**, 3928–3971 (2025). https://doi.org/10.1007/s10278-025-01481-y
5. A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D Medical Image Segmentation," Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 574–584, 2022. https://doi.org/10.1109/WACV51458.2022.00181
6. O. Dalmaz, M. Yurt and T. Çukur, "ResViT: Residual Vision Transformers for Multimodal Medical Image Synthesis," in IEEE Transactions on Medical Imaging, vol. 41, no. 10, pp. 2598-2614, Oct. 2022, https://doi.org/10.1109/TMI.2022.3167808
7. S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," ACM Computing Surveys, vol. 54, no. 10s, pp. 1–41, 2022. https://doi.org/10.1145/3505244
8. G. Wang, W. Li, S. Ourselin, and T. Vercauteren, "Automatic Brain Tumor Segmentation Using Cascaded Anisotropic Convolutional Neural Networks," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Lecture Notes in Computer Science, vol. 10435, 2017. https://doi.org/10.1007/978-3-319-75238-9_16
9. Kelei He, Chen Gan, Zhuoyuan Li, Islem Rekik, Zihao Yin, Wen Ji, Yang Gao, Qian Wang, Junfeng Zhang, Dinggang Shen, "Transformers in medical image analysis," Intelligent Medicine, Volume 3, Issue 1, 2023,
   DOI: 10.1016/j.imed.2022.07.002.
10. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," International Journal of Computer Vision, vol. 128, no. 2, pp. 336–359, 2020, https://doi.org/10.1007/s11263-019-01228-7.