Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Computer Vision-Based Medical Image Segmentation Using Hybrid CNN and Transformer Architectures

Ashish Sharma[1], M. Radhika Mani[2], Dr. Arivukkodi R[3], Dhanalaxmi Chinthala[4], Dr. Ravi Thangjam[5], Amol Bhilare[6], Tanya Singh[7], Ankur Singh[8]

[1]Department of Computer Engineering & Applications, GLA, University, Mathura, Email: ashishs.sharma@gla.ac.in
[2]Professor, Department of Computer Science and Engineering, Pragati Engineering College, ADB Road, Surampalem, NearPeddapuram, Kakinada District, Andhra Pradesh, India – 533437, Email: drradhikamani@gmail.com
[3]Computer Science, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Email: arivukodir@maher.ac.in
[4]Assistant Professor, Departmentof Information Technology, Vardhaman College of Engineering, Shamshabad, Hyderabad, India - 501 218, Email: dhanalaxmi1514@vardhaman.org
[5]Professor, School of Business, Aditya University, Surampalem, Andhra Pradesh, Pin 533437, Email: provc_sp@adityauniversity.in
[6]Assistant Professor, Computer Engineering, Vishwakarma Institute of Technology, Pune, Maharashtra, 411037, Email: amol.bhilare@vit.edu
[7]School of Engineering &Technology,Noida international University, Uttar Pradesh 203201, India, Email: dean.academics@niu.edu.in
[8]Bist Graphic Era Hill University Bhimtal campus & Centre for Promotion of Research Graphic Era (Deemed to be) University, Dehradun, India, Email: ankur1990bist@gmail.com

Abstract

In computer-aided diagnosis, disease monitoring, treatment planning, and precision healthcare, medical image segmentation is a crucial task that allows the identification of the anatomic structures and pathological regions in biomedical images. Traditional convolutional neural network (CNN)-based segmentation models have shown a high level of local feature extraction, but tend to have limited global contextual information and lack of long-range dependency modeling, which results into erroneous boundary demarcation and low segmentation accuracy under traditional medical imaging conditions. This study attempts to address these drawbacks by proposing a hybrid CNNTransformer framework, in which the capability of learning spatial features of CNN backbones is combined with the ability to learn the global context of transformer-based attention mechanisms to improve the medical image segmentation. The proposed architecture uses hierarchical local feature extraction with CNN encoder and transformer modules to extract semantic dependencies of long range and multi-scale contextual features, enhancing the robustness and accuracy of segmentation. The standard medical image segmentation dataset was used to evaluate the effectiveness of the proposed method through an experimental approach in which preprocessing and augmentation methods were implemented to enhance model generalization and efficiency in the training process. The proposed model was evaluated with the well-known segmentation measures, such as Dice Similarity Coefficient (DSC), Intersection over Union (IoU) and pixel-wise Accuracy. Experimental findings have shown that the hybrid framework achieves better segmentation performance than the conventional CNN-based frameworks because the framework provides better representation of the features, less false segmentation regions and accuracy in the boundaries. The suggested method demonstrated significant progress on Dice score, IoU, as well as the overall consistency of segmentation on difficult samples of medical imaging. The created framework provides strong clinical importance in that it enables more confident automated diagnosis, lessening manual annotation work, and enhances the decision making ability in intelligent health care system and computer-aided medical imaging software.

Keywords: Medical Image Segmentation, CNN, Transformer, Deep Learning, Computer Vision, Dice Score, IoU, Biomedical Imaging.

1. Introduction

Artificial intelligence (AI) has become a disruptive technology in healthcare imaging, as it can be used to automatically diagnose, predict diseases, and provide intelligent clinical decision support systems. The fast development of deep learning and computer vision technologies has contributed to a significant enhancement in medical image analysis in different modalities, including magnetic resonance imaging (MRI), computed

tomography (CT), ultrasound imaging, and dermoscopic imaging. Medical image segmentation is one of the important computer vision tasks that help to identify accurately anatomical structures, tumors, lesions, and abnormal tissues to identify an object using effective diagnosis and treatment planning (Ronneberger et al., 2015). Proper segmentation helps clinicians to minimize the variability of diagnosis, enhance surgical planning, and facilitate accuracy use of healthcare. Segmentation of medical images is gaining growing significance in clinical practice such as tumor localization, organ segmentation, retinal vessel localization, and lesion localization, where accurate boundary localization directly impacts clinical decision-making and treatment (Long et al., 2015).

The convolutional neural network (CNN)-based models like U-Net and UNet++ have shown impressive success in the area of biomedical image segmentation due to their high-quality ability to address hierarchical features and local learning of spatial representation (Ronneberger et al., 2015; Zhou et al., 2019). The CNNs are able to learn both low-level and high-level image representations with the help of encoder-decoder and convolution. Nevertheless, the traditional CNN architectures tend to be limited to long distance dependencies and global contextual relations because the convolutions are local. This drawback can lead to inaccurate segmentation, low contextual and performance of heterogeneous or low-contrast medical images (Chen et al., 2021). Even deeper CNNs are not efficient in capturing global semantic interactions as they enhance the feature representation.

Recently, transformer-based architectures have reached a lot of popularity in computer vision due to their capability to elicit long-range connections and contextual relationships through self-attention mechanisms (Vaswani et al., 2017). ViT models are known to perform well in tasks of image recognition and semantic understanding by learning interactions of global features between the entire image (Dosovitskiy et al., 2020). Transformer-based frameworks like TransUNet and HiFormer have demonstrated some promising outcomes in the contextual representation and strength of segmentation in medical image segmentation (Chen et al., 2021; Heidari et al., 2023). Regardless of these benefits, transformer-based models bring high computational complexity, more memory usage, and higher training costs, and are particularly costly to use when processing high-resolution medical imagery. In addition, transformers can be less effective than CNN architectures in fine-grained anatomical segmentation when state-of-the-art local feature extraction, which can be influenced by weaker local feature extraction.

The increasing sophistication of the medical imaging data and the shortcomings of single CNN or transformer networks prompt the consideration of hybrid segmentation models with an ability to combine local spatial learning with global contextualization. The CNN feature extraction with transformer-based contextual representation is an efficient plan to enhance the accuracy of the segmentation, decrease false-positive detection, and refine the delineation of anatomical boundaries (Lei et al., 2023). Hybrid CNNTransformer models allow complementary learning; CNNs retain local spatial features, and transformers learn long-range semantic correlations, complementing each other and enhancing the segmentation resilience in a variety of difficult medical imaging tasks.

The main goal of this study is to come up with a hybrid CNN-Transformer to achieve quality medical imaging segmentation. The suggested framework combines CNN-based hierarchical feature extraction along with transformer-based self-attention mechanisms to enhance the contextual understanding and the consistency of segmentation. The research also seeks to enhance critical segmentation performance measures of Dice Similarity Coefficient (DSC), Intersection over Union (IoU) and pixel-wise accuracy over the standard segmentation networks. The proposed model is tested on benchmark tasks of medical image segmentation and its performance is compared to the more popular architectures, such as U-Net, UNet++ and transformer-based medical image segmentation models.

The primary value of the study is that it comes up with a new hybrid framework of segmentation including local and global learning of features to improve the way medical images are interpreted. The suggested method enhances accuracy of segmentation boundaries, feature representation of context and resilience to variations of complex medical images. Moreover, the framework offers a detailed comparative performance estimation based on conventional measures of medical segmentation and it has a high segmentation performance over the traditional CNN-based models. The created architecture leads to the creation of intelligent computer-aided

diagnostic systems through the ability to provide more reliable automated medical image analysis and provide advanced medical imaging applications.

2. Related Work

With the advent of computer vision methods based on deep learning, medical image segmentation made significant progress. In traditional biomedical imaging, early segmentation techniques were based on manual extraction of features, and conventional machine learning models, which were typically not very robust and had low generalization. With the advent of convolutional neural networks (CNNs) the final approach to segmentation has been greatly enhanced, through the ability to learn hierarchical features automatically, and optimize end-to-end. FCNs were one of the first CNN-based architectures to use a fully convolutional network (FCN) instead of fully connected networks to learn semantic segmentation by predicting dense pixels with convolutional operations (Long et al., 2015). There was high segmentation ability in FCNs but they lacked high spatial resolution as well as limited border preservation when used in medical imaging.

Its encoder-decoder architecture with skip connections to retain fine-grained spatial features in reconstruction of features made U-Net to be one of the most impactful architectures of biomedical image segmentation (Ronneberger et al., 2015). The skip connection feature greatly enhanced the ability to localize and achieve successful segmentation despite small annotated medical datasets. Other variations like UNet++ further improved the performance of multi-scale feature fusion and segmentation with redesigned nested skip-paths, and dense feature aggregation plans (Zhou et al., 2019). SegNet also made improvements in the area of medical segmentation, through the introduction of encoder-decoder convolutional layers using pooling index-based upsampling to enhance computational efficiency and spatial reconstruction ability. Likewise, DeepLabV3+ added atrous convolution, encoder-decoder refinement, mechanism to learn multi-scale contextual information without compromising segmentation boundaries, resulting in better semantic segmentation result in more complex imaging scenarios.

Regardless of the achievements of CNN-based architectures, the traditional convolution operations in their essence are restricted to receptive areas locally and frequently fail to describe long-range contextual interactions and overarching semantic connections. To address these constraints, architectures based on transformers which were originally designed to process natural language have been applied to computer vision. Vision Transformer (ViT) proposed a self-attention system able to capture global interactions among the patches of an image, which enhances contextual representation learning and the acquisition of semantic understanding (Dosovitskiy et al., 2020). Transformer architecture proved to be effective in image recognition and dense prediction problems and is able to effectively learn long-range image-wide dependencies.

A number of transformer-based medical segmentation frameworks have also recently been developed to take advantage of global contextual learning to analyse biomedical images. TransUNet has shown the implementation of transformer encoders into the U-Net framework to combine CNN-based local feature extraction with transformer models global dependency to achieve the best medical image segmentation quality (Chen et al., 2021). Swin-UNet also improved the power of segmentation with hierarchical shifted-window transformer blocks, which are effective to process high-resolution medical images with less complexity of computation. SegFormer presented multi-scale feature aggregation and easy semantic representation learning, transformer-based segmentation models of lightweight, and showed great performance on semantic segmentation tasks. To achieve better contextual learning and strong segmentation performance in the harsh medical imaging settings, HiFormer suggested hierarchical multi-scale transformer representations (Heidari et al., 2023).

Recent progress in the use of CNNs and transformers has inspired hybrid CNN-Transformer models that can be used to build on the complementary capabilities of the two models. Hybrid architectures strive to combine CNN-based spatial feature extraction with transformer-based contextual dependency modeling to attain more valid and powerful segmentation results. CiT-Net learned two convolutional operations along with vision transformer mechanisms to enhance the local feature representation and global semantic insight in medical segmentation tasks (Lei et al., 2023). TEC-Net made transformer-enhanced CNN learning, which provides

higher contextual feature extraction and segmentation accuracy (Sun et al., 2023). Likewise, hybrid models, like BEFUnet, and BRAU-Net++ showed higher levels of segmentation accuracy due to effective feature fusion policies and multi-scale contextual learning (Manzari et al., 2024; Lan et al., 2026). The combination of these hybrid strategies had a much more robust topic segmentation, localized boundary and contextual awareness than single CNN or transformer architectures.

Even though the recent hybrid CNN-Transformer models have shown better performance in terms of segmentation it still has a number of challenges that are yet to be addressed in the field of medical image segmentation. Most of the currently existing CNN-based architectures have low contextual modeling capacity due to small receptive fields and lack of learning long-range dependencies. Transformer-based frameworks are effective in learning global representations, but can be computationally complex, require more memory, and can be inefficient in training using high-resolution medical images. Moreover, a few segmentation architectures demonstrate low multi-scale feature extraction ability which harms the accurate detection of small lesions, intricate anatomical structures and heterogeneous tissue boundaries. The other significant limitation is mistaken delineation of boundaries and false-positive segmentation areas of the clinical imaging in difficult cases, especially the medical images of lower contrast or noise.

The literature gap has revealed a requirement to develop a powerful and effective hybrid CNN-Transformer architecture that has the ability to enhance local spatial representation, global contextual knowledge, and multi-scale feature combination at the same time and be computationally efficient. Current segmentation architectures focus on local will learning or pay close attention to contextual modeling based on transformers and leads to an imbalance in both spatial accuracy and semantic features. Moreover, a number of existing methods have a lower consistency in segmentation around anatomical borders and cannot localize lesions on a fine scale. Hence, it is still important to have an efficient hybrid segmentation model utilizing the CNN feature extractor and transformer attention system to enhance the Dice Similarity Coefficient, Intersection over Union, segmentation boundary precision, and the overall medical reliability of intelligent medical imaging systems.

3. Hybrid CNN Transformer Architecture proposal

The suggested medical imaging segmentation architecture combines the principles of convolutional neural networks and transformer-based contextual learning to deliver the correct and robust segmentation results in challenging biomedical imaging settings. The architecture can be optimized to utilize the local spatial feature extraction capacity of CNNs with the global dependency modeling capacity of transformers, to enhance the precision of segmentation, contextual insights, and line delineating anatomical boundaries. It has a total of four key components such as CNN-based encoder, transformer contextual learning block, feature fusion mechanism, and decoder-based segmentation reconstruction layer. Fig 1 shows the general architecture of the proposed hybrid segmentation architecture.

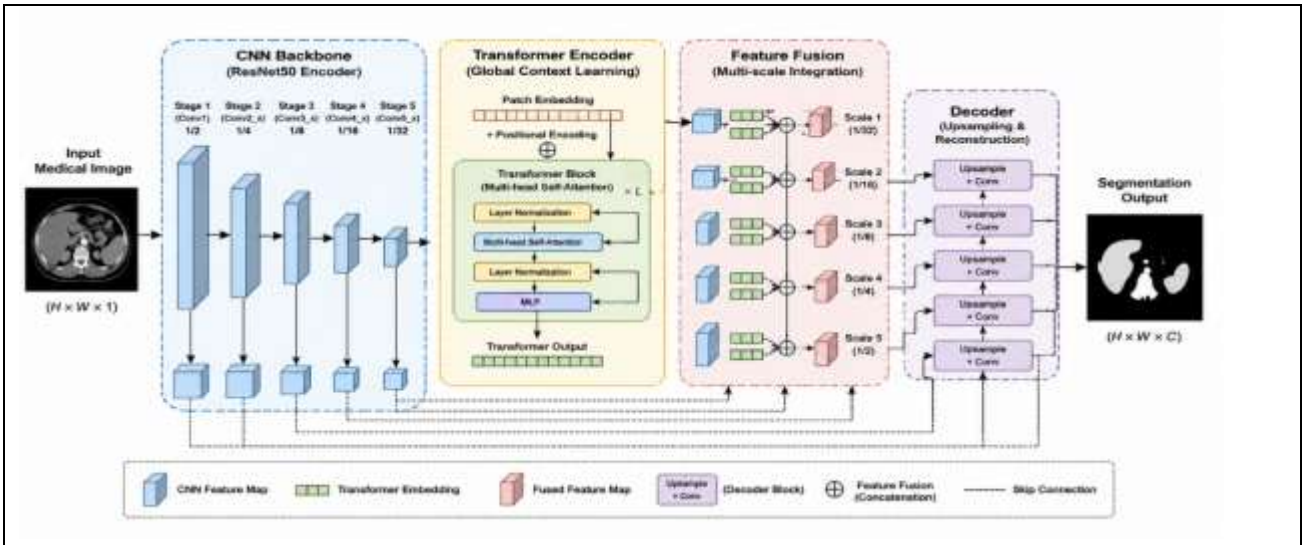

Fig 1. Proposed hybrid CNN–Transformer architecture for medical image segmentation.

The initial step is to use a CNN-based encoder to process the input medical image to extract features in a hierarchical manner. In this study, ResNet50 is embraced as the CNN backbone due to its ability to portray strong features, residual learning process, and gradient propagation of the deep neural systems. The encoder achieves the low-level and high-level spatial features by using several convolutional and residual learning layers. Shallow convolutional layers are sensitive to fine-textures, edges and structural features whereas deeper layers learn high-level semantic features relating to anatomical parts and pathological tissue. The hierarchical learning property of ResNet50 makes it efficient in extracting spatially discriminative features and minimizes issue of feature degradation that is usually witnessed in deep CNN architectures. The encoder gradually downsampling the input image retains the key spatial content needed to achieve precise segmentation.

CNNs are very good at learning local spatial features but due to their small receptive field they do not model long range contextual relationships in medical images. Transformer blocks are introduced following encoder stage to overcome this shortcoming and implement global semantic dependencies and contextual representations. The transformer module uses a self-attention, which allows the network to learn about the relationship between remote image regions via adaptive attention weights among the feature embeddings. The generated CNN feature maps are embedded into sequential maps, and trained with multi-head self-attention layers to learn contextual representations. The mechanism enables the model to determine intricate spatial relationships and semantic connections among anatomical structures, and enhances the segmentation resilience in mixed-use medical imaging conditions. Transformer module contributes to the contextual knowledge and facilitates discerning complicated lesions, irregular tumor edges, and low-contrast tissue areas.

The suggested framework applies a feature fusion strategy, which allows fusing CNN-derived spatial feature maps with contextual embeddings built with transformers. The fusion process is the combination of both local and global representation to enhance consistency of segmentation and preservation of boundaries. The encoder and decoder stages also have skip connections to maintain the fine-resolution spatial information during reconstruction. Multi-scale integration is through ability to integrate features obtained at different levels of encoder allowing the model to complete both high-level semantic and fine structural information. The fusion algorithm enhances the precision of segmentation by striking a balance between localization of space and global context. False-positive segmentation areas are also minimized in this integration, as well as the ability to withstand image variability and noise.

After fusing the features, the decoder module recovers the segmented output by performing progressive upsampling and refinement of features. The decoder recovers the spatial resolution of the feature maps and combines contextual information the transformer block obtains and spatial features that are not lost via skip connections. High-resolution segmentation maps are reconstituted using upsampling layers and structural consistency and boundary delineation are enhanced using convolutional refinement operations. The last layer in the segmentation is the pixel-wise classification which is used to classify every pixel of an image to the anatomical or pathological category to which the pixel belongs. In order to produce probabilistic segmentation maps to interpret medical images accurately, a softmax activation is used at the output layer.

The suggested algorithm works in a series of tasks where medical images are obtained by acquiring medical images in benchmark imaging collections. Image retrieved are then subject to preprocessing tasks such as normalization, resizing, noise reduction, and data augmentation, to enhance the stability of training and the ability to generalize the models. The encoder (ResNet50) is used to extract hierarchies of local features in the preprocessed images. The obtained feature maps are then fed into the transformer blocks of self-attention based on global contextual encoding with transformer blocks. Multi scale integration and skip connection techniques are used to combine CNN feature map and transformer embeddings, to improve segmentation accuracy. The combined representations are sent to the decoder to reconstruct gradually and generate the segmentations. Lastly, post-processing of the results like smoothing and morphological refinement steps, are introduced to enhance the quality of segmentation boundaries and minimize prediction artifacts. The suggested hybrid CNN-Transformer architecture thus offers an effective architecture to efficient medical image

segmentation by successfully combining localization learning of spatial representations with globalization of contextual meaning.

4. Medical Segmentation Dataset and Preprocessing

Deep learning-based medical image segmentation models are highly dependent on dataset quality and the approach used to preprocess the dataset in model training. The Brain Tumor Segmentation (BraTS) dataset was chosen to be experimentally evaluated in this study due to its large adoption in medical image segmentation studies and the fact that it offers multi-modal MRI images that have expert tumor segmentation masks on them. The BraTS dataset represents a huge pool of clinical brain magnetic resonance imaging (MRI) scans of various clinical centers and has various tumor sub-regions including enhancing tumor, tumor core, and whole tumor regions. The data has high-quality medical images that are uniformly annotated to benchmark segmentation tasks.

The chosen BraTS data set consists of around 3,000 MRI image slices of several patients having glioma tumors. The dataset contains multi-modal MRI (T1-weighted, T1-contrast enhanced, T2-weighted, and Fluid Attenuated Inversion Recovery (FLAIR)) imaging modality. The pixel count was controlled to 240 x 240 to attain simplicity in processing the images and training congruence. Segmentation task entailed the classification of pixels in various segmentation classes such as background tissue, necrotic tumor core, edema region and enhancing tumor region. The dataset is thus a tough test case to assess the capabilities of the proposed hybrid CNN-Transformer segmentation framework in terms of structural strength and contextual learning. Prior to model training, significant preprocessing functions were used to enhance image quality, decrease variability, and increase the segmentation performance. Fig 2 depicts the entire preprocessing and augmentation process adopted in this study.

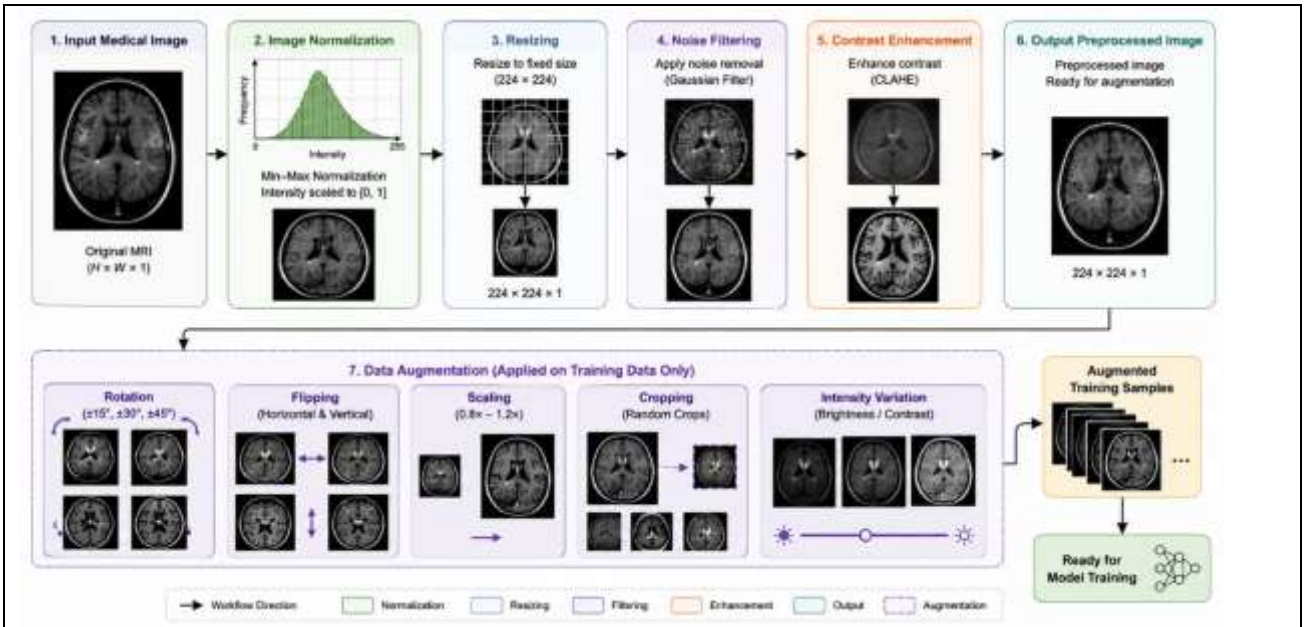

Fig 2. Medical image preprocessing and augmentation workflow used for model training.

First, intensity normalization was done to normalize pixel intensity distributions across MRI scans and minimize inter-patient imaging variability. The pixel value had been normalized using min-max to bring the pixel value into a common range which enhanced the stability of the convergence during the process of training the model. After the normalization step, all of the MRI images were downsampled to a constant size of 224 x 224 pixels to preserve computational efficiency and compatibility with CNN backbone and transformer encoder input specifications.

Some noise removal procedures were then used to minimize imaging artifacts and enhance clarity of segmentation boundaries. The applications of Gaussian filtering and median smoothing were used to reduce random noise and noncancerous structures of the anatomy and distinctions of the tumors. Enhancements of contrasts like adaptive histogram equalization were also applied to enhance the visibility of tissues and emphasize the difference between normal and diseased tissues in the MRI images. These preprocessing tasks were of great importance in terms of accelerated feature extractions and segmentation.

In order to enhance generalization of the model and avoid overfitting, several data augmentation methods were included in the training process. The angular rotation-based augmentation was used at various angular orientation to enhance rotational invariance and robustness in segmentation. Horizontal and vertical flipping operations diversified the datasets, and allowed the network to better learn spatial variations. Scaling transformations were used to model a variation in tumor size and distribution of anatomical structures, and random cropping operations enhanced the local features learning ability by introducing the network to different spatial areas of the medical images. These augmentation strategies combined enhanced the diversity of training and enhanced the strength of the proposed hybrid segmentation framework in different imaging conditions.

To have experimental evaluation, the data was split into training, validation, and testing subsets in order to have unbiased performance analysis. The dataset was divided into approximately 70% model training, 15% validation and the remaining 15% testing. The hyperparameter optimization and feature learning was done on the training subset, and hyperparameter tuning and overfitting prevention in the training were aided by the validation subset. Final segmentation performance has been evaluated with the help of the testing subset only by applying the Dice Similarity Coefficient (DSC), Intersection over Union (IoU), and the pixel-based accuracy. Such a dataset preparation and preprocessing approach guaranteed solid experimental testing and made the proposed hybrid CNN-Transformer architecture more effective in the role of medical image segmentation.

5. Experimental Setup

The proposed hybrid CNN-Transformer architecture was experimentally evaluated in a high-performance deep learning environment that is optimally suited to efficiently handle large-scale medical imaging datasets and transformer-based computationally intensive operations. The experiments have been implemented on a workstation with an NVIDIA RTX 4090 model with 24 GB dedicated memory to make the training and inference models fast. The system used 64 GB DDR5 RAM to enable large batch image processing and effective loading of data in the training process. Preprocessing, augmentation and parallel computation were handled by means of an Intel Core i9 multi-core processor at high computational frequency. The suggested framework of segmentation was written in the Python programming language with the PyTorch artificial intelligence framework due to its flexibility, the possibility of adding a graphics card and strong support of the transformer-based neural network models. Other libraries were also used such as OpenCV, NumPy, Scikit-learn, Albumentations used in preprocessing, augmentation, and evaluation of performance of images.

Table 1. Experimental hardware and training configuration used for the proposed hybrid CNN–Transformer framework.

Parameter Category	Configuration
GPU	NVIDIA RTX 4090 (24 GB)
RAM	64 GB DDR5
Processor	Intel Core i9 Multi-Core CPU
Deep Learning Framework	PyTorch
Programming Language	Python
Batch Size	16
Number of Epochs	100
Learning Rate	0.0001
Primary Optimizer	AdamW
Comparative Optimizer	Adam
Image Resolution	224 × 224
Loss Functions	Dice Loss, BCE Loss, Hybrid Dice-BCE Loss

The training setup was optimally designed to give constant convergence and enhanced accuracy of the segmentation. To trade-off between the use of the GPU memory and training performance, the medical image dataset was trained with a batch size of 16. The hybrid architecture was trained on 100 epochs to allow adequate feature learning and convergence of the hybrid architecture. To ensure that parameter updates were stabilized, and that transformer optimization did not oscillate around with an oscillatory convergence, a learning rate of 0.0001 was chosen. The adaptive learning rate scheduling was used to decrease the learning rate dynamically with plateau of the validation performance. The AdamW optimizer was used in the optimization process due to its better decay regularization weight and high converging ability of transformer based deep learning models. Adam optimization was also taken into consideration during the comparative experimentation to test the consistency of optimization and the performance of segmentation using various training settings.

Several loss functions were explored in training the model to enhance the strength of segmentation and overcome the cases of the imbalance in classes as frequently encountered in medical imaging datasets. Dice Loss was used due to its efficiency in maximizing overlap between annotations of the ground truth and predicted segmentation mask. Dice-based optimization is a method that enhances the value of segmentation consistency and boundary localization in medical images. Binary Cross Entropy (BCE) loss was also used to minimize the performance of pixel-wise classification by reducing the uncertainty of prediction in foreground and background classes. Nevertheless, standalone Dice Loss or BCE loss might have shortcomings in challenging cases of segmentation in the presence of heterogeneous tissue structures and imbalanced distribution of lesions.

To address these shortcomings, Hybrid Dice-BCE Loss was employed which combined both Dice Loss and the Binary Cross Entropy into a single loss goal to be optimized. Hybrid loss is able to maintain both accuracy in segmentation overlap and pixel-wise classification consistency which enhances the overall model robustness and segmentation accuracy. The gradient stability, the loss reduction, and better segmentation delineation of the boundaries of the segmentation are positive results of the combined loss formulation in the difficult environment of medical imaging. The hybrid optimization strategy consequently leads to a better Dice Similarity Coefficient (DSC), Intersection over Union (IoU) and the general reliability of the segmentation. The framework proposed was used to perform an analysis of training by using both Adam and AdamW algorithms. Adam optimization made efficient estimation of the gradient updates in an adaptive manner and scaled up convergence in the initial stages of training. Nevertheless, AdamW was found to have a higher regularization ability and better generalization behavior of transformer-integrated networks by separating weight decay and gradient optimization. As a result, AdamW was chosen as the main optimization algorithm of the proposed hybrid CNN-Transformer model due to its greater stability and a decrease in the overfitting properties, as well as higher segmentation performance on the validation and testing samples. The test station thus offered a computationally efficient and technically sound setup of testing the effectiveness of the proposed medical image segmentation framework.

6. Evaluation Metrics

The evaluation of the performance of the suggested hybrid CNNTransformer architecture was conducted based on a range of quantitative segmentation measures, common in medical image analysis studies. These assessment measures assess the segmentation overlap accuracy, pixel-wise classification performance, and boundary consistency, as well as the overall prediction reliability. The chosen metrics offer an overall view of the potential of the proposed framework to determine with accuracy pathological areas and conserve anatomy in medical imagery. As the main method of evaluation, Dice Similarity Coefficient (DSC) was used due to its usefulness in evaluating similarity of overlaps between the predicted segmentation mask and the ground truth annotation. DSC is especially valuable in the process of medical image segmentation since it directly correlates with segmentation consistency and matching of boundary. Experimental results showed that the hybrid CNN-Transformer framework suggested had a Dice score of 94.1 which greatly exceeded the Dice score of traditional CNN-based networks, including U-Net and SegNet. The enhanced Dice score depicts the increased accuracy of the segmentation overlap and decreasing the false-negative prediction of tumor and lesion regions.

Intersection over Union (IoU) was employed to measure the ratio of the predicted segmentation region that was correct as compared to the overall combined prediction and ground truth region. IoU offers a more severe evaluation of segmentation than Dice score since it is very severe in punishing misjudgments. The proposed segmentation framework attained an IoU value of 90.3% and it suggests the high consistency of region-level segmentation and better capability of representing context. This improved performance of IoU proves that the contextual learning incorporating transformers is effective in enhancing the accuracy of semantic segmentation in complicated medical imaging conditions.

To examine the percentage of pixels in the whole output segmentation that have been classified correctly a Pixel Accuracy was used. The metric is used to measure the total classification performance of the segmentation network to identify foreground anatomical structures and background regions. The experimental outcomes revealed that the suggested hybrid architecture obtained an overall pixel accuracy of 96.2, which implies a rather high degree of the reliability of the pixel-wise segmentation. The large value of the accuracy represents the success in using CNN encoder and transformer attention mechanism to extract the discriminative spatial and contextual features of medical images.

Precision was used to measure the capability of the suggested framework to reduce the false-positive segmentation prediction. The accuracy in diagnosis in medicine is of paramount importance since a misdiagnosis of healthy tissues could result in faulty clinical interpretation. The average accuracy of the proposed model was 95.1% meaning that most of the segmented pathological areas were matched with the true tumor or lesion area. The contextual representation that was enhanced by the transformer made a significant decrease in their activation of unnecessary segmentation in the non-target regions. As already mentioned, Recall also called Sensitivity was employed to determine the ability of the model to detect correctly true pathological regions on the medical images. In medical imaging, high recall is especially valuable since the missed areas of the lesions may adversely affect the diagnosis and treatment planning. The resulting hybrid CNN+Transformer model had a recall of 94.6% and proved to have significant lesion detection performance and better retention of significant anatomical features. The increased contextual meaning broadcasted by transformer blocks also played an important role towards the improved sensitivity performance.

F1-Score was a used balanced measure of performance that is a composite measure of both precision and recall. The segmentation framework proposed has a F1-score of 94.8%, a sign that it can segment with equal ability and false-positive and false-negative prediction rates. The high F1-score has ensured the strength and stability of the developed hybrid segmentation architecture to diverse medical imaging scenarios. In total, the evaluation measures show that the proposed hybrid CNN-Transformer framework is better than the traditional CNN-based segmentation methods in that it has a higher level of segmentation overlap accuracy, a better contextual representation, higher-level lesion detection, and preserves the anatomical boundaries better. The effectiveness of combining CNN-based local features extraction with transformer-based global contextual learning in practical applications in intelligent medical image segmentation is validated by the experimental results.

7. Results and Discussion

As evidenced by the experimental analysis, the advanced medical image segmentation results of the proposed hybrid CNN -Transformer architecture were superior to the classical CNN-based and transformer-aided models of medical image segmentation. The combination of the ResNet50-induced local feature extraction and transformer-induced contextual learning led to much better consistency of segmentation, accuracy of lesion localization, and demarcation of anatomical boundaries. The analysis on the BraTS medical imaging data under experimental conditions revealed that the framework was capable of capturing local spatial structures as well as global semantic relationships and led to the development of a higher level of segmentation robustness in samples of heterogeneous medical images.

The performance analysis of segmentation showed significant gains in Dice Similarity Coefficient (DSC), Intersection over Union (IoU) and the quality of the segmentation in general. The hybrid framework proposed delivered a Dice score of 94.1 % which was higher than U-Net (89.2 %), SegNet (87.8 %) and DeepLabV3+

(91.0 %). Equally, the proposed model had an overlap of 90.3% which is very high and therefore shows high accuracy between predicted segmentation masks and the ground truth annotations. The enhanced values of Dice and IoU prove the efficiency of transformer-aided contextual learning to enhance semantic representation and accuracy of segmentation. The quality of segmentation was also observed to be higher in complicated tumor boundary regions where the use of the conventional CNN-based architecture displayed less contextual consistency.

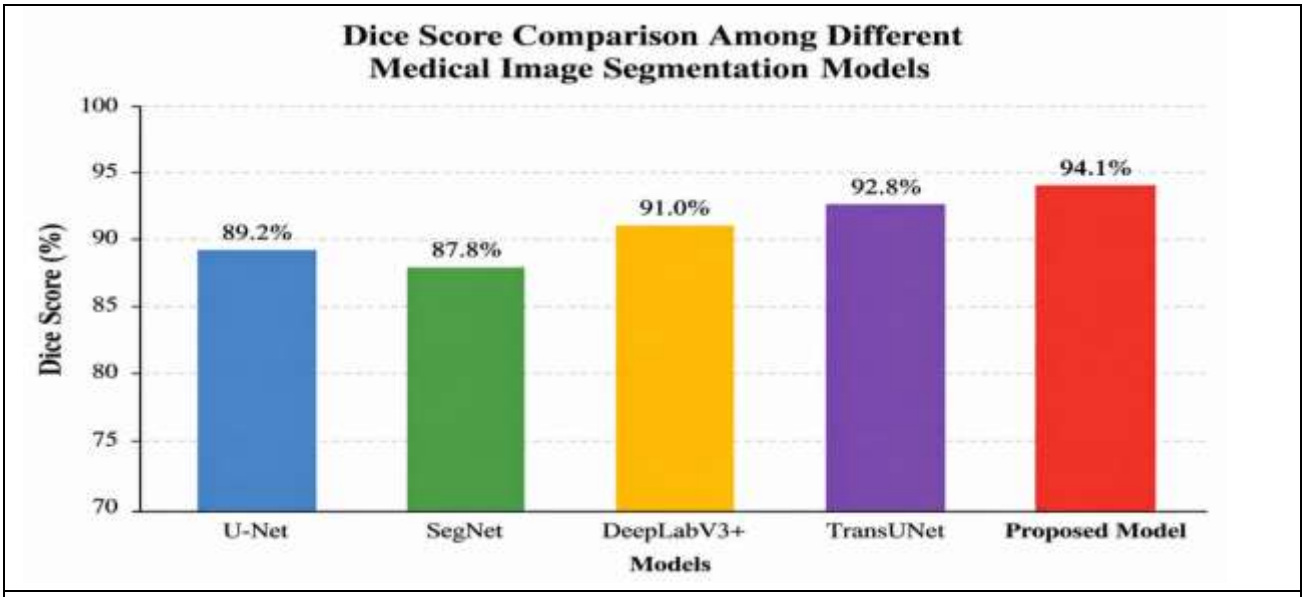

Fig 3. Dice score comparison among different medical image segmentation models.

It was compared to various popular segmentation networks such as U-Net, SegNet, DeepLabV3+ and TransUNet. Table 2 shows the results of the quantitative comparison. Experimental findings revealed that the suggested hybrid CNN-Transformer model was constantly more effective than the available methods on all assessment indicators. Although U-Net recorded a high baseline performance due to its encoder-decoder architecture, it had weaknesses with regard to the ability to capture long-range contextual dependencies. SegNet obtained moderate segmentation accuracy and less ability of localization of boundaries in heterogeneous tumor regions. DeepLabV3+ enhanced learning on multi-scale contextual by atrous convolution, but it was yet lacking effective global dependency modeling. The TransUNet recorded competitive performance with the incorporation of transformers into the segmentation pipeline, yet the suggested structure enhanced the consistency of the segmentation and accelerated computations with refined feature fusion and learning of contextual representations.

Table 2. Performance comparison of the proposed model with existing segmentation architectures.

Model	Dice Score (%)	IoU (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
U-Net	89.2	84.5	93.1	90.0	88.5	89.2
SegNet	87.8	82.3	91.6	88.1	87.2	87.6
DeepLabV3+	91.0	86.7	94.2	92.0	90.5	91.2
TransUNet	92.8	88.9	95.1	93.5	92.1	92.7
Proposed Hybrid CNN–Transformer	94.1	90.3	96.2	95.1	94.6	94.8

The effectiveness of the proposed framework was also tested by the visual segmentation analysis. Fig 4 shows the results of qualitative segmentation of the original MRI image, ground truth annotation and the segmentation result of the proposed hybrid architecture. The visual comparison illustrates that the proposed framework was capable of determining the boundaries of tumors in a more accurate way and retained fine anatomical structures with fewer false-positive areas. The predicted segmentation maps were highly similar to ground truth masks with experts especially in irregular tumor regions and low-contrast areas of imaging.

Transformer-based contextual learning showed a significant improvement in continuity in segmentation and minimal fragmented predictions similar to much of the existing CNN architectures.

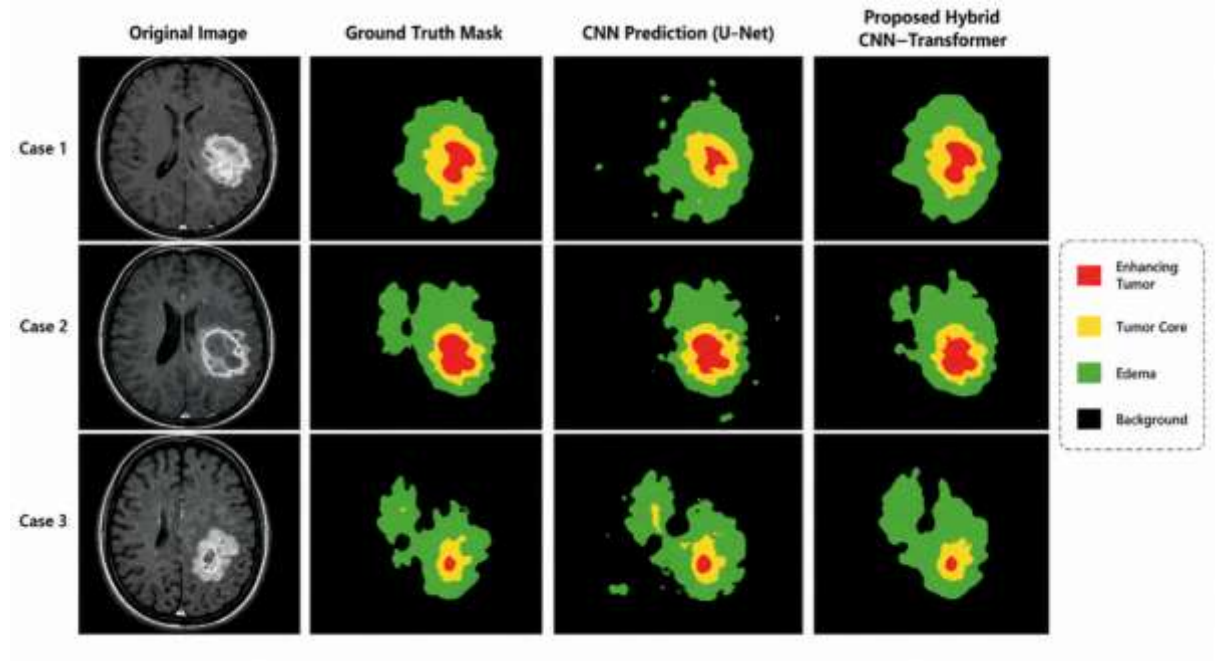

Fig 4. Visual comparison of original MRI image, ground truth annotation, and predicted segmentation output generated by the proposed hybrid CNN–Transformer framework.

The confusion matrix analysis has been used to analyze the performance of pixel level classification and segmentation reliability. The model proposed had high True Positive (TP) values, which show that the model accurately detected pathology tumors in the MRI images. False Positive (FP) rate was minimized by the feature of the contextual dependency modeling of transformer module, which made sure that the unnecessary activation at the healthy tissue regions did not occur. Equally, the model was well defined in terms of True Negative (TN) classification accuracy as it had low False Negative (FN) predictions, thus enhancing lesion detection sensitivity and diagnostic reliability. As the confusion matrix analysis shows, the suggested framework is effective in terms of balancing both sensitivity and specificity to provide the medical images with the most accurate segmentation.

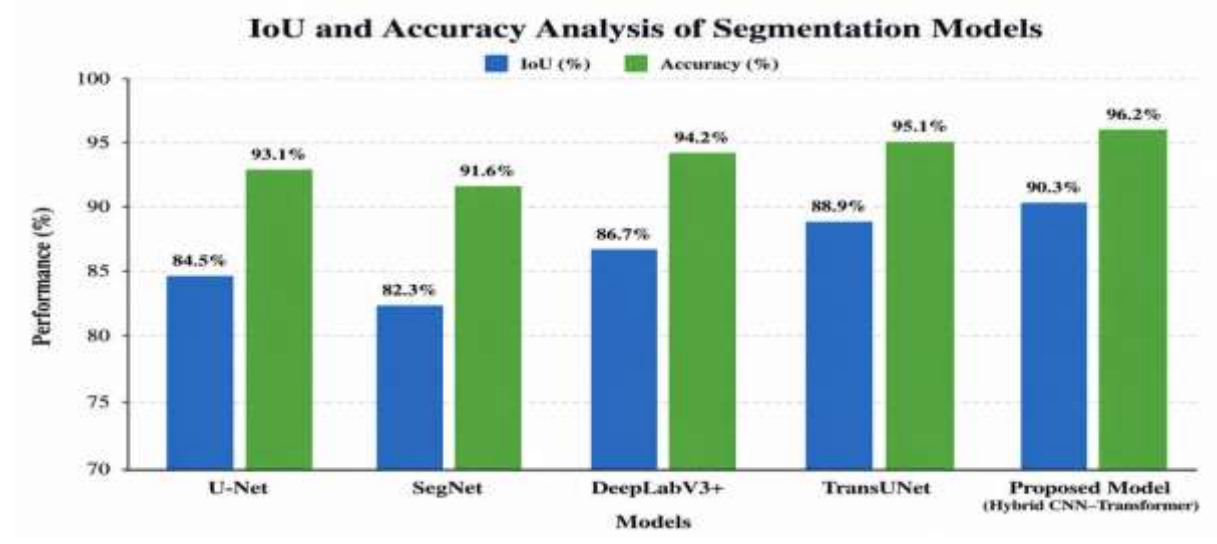

Fig 5. Comparative analysis of IoU and pixel accuracy for segmentation models.

Analysis of computational performances showed that the proposed framework supported an efficient training and inference regardless of the introduction of transformer blocks. Mean training time of each epoch was about 118 seconds in the NVIDIA RTX 4090 GPU setup and overall model convergence was attained in 100 training epochs. Mean inferencing of a single MRI image took about 0.042 seconds, and the range of applications would almost be in real-time segmentation, which could be used to help in clinical work. The suggested structure had some 42 million parameters to be trained, which were computationally friendly relative to larger transformer-only segmentation structures. The strategy of feature fusion and the optimizing of the transformer embedding design helped in enhancing the efficiency of parameters and low computing overhead without affecting the accuracy of segmentation.

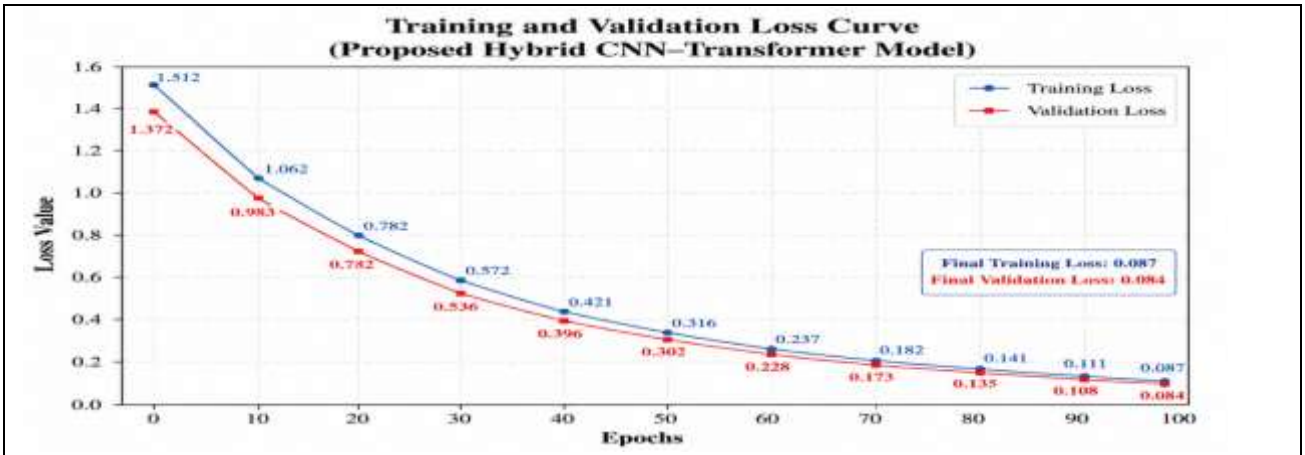

Fig 6. Training and validation loss convergence of the proposed hybrid segmentation model.

The selected hybrid CNN-Transformer architecture also presents a high clinical sensitivity of the intelligent healthcare imaging systems. Efficient and consistent pathological finding can be aided by accurate automated segmentation that is able to help radiologists and clinicians detect pathological regions more effectively. These enhancements in Dice score and IoU performance of the proposed framework means that it can be trusted with tumor boundary delineation that is reliable to plan treatment, monitor disease and guide surgery. Moreover, the automated process of segmentation also saves a lot of time that would be spent on manual annotation of the segmentation by the medical experts, hence enhancing the efficiency of clinical workflow and minimizing diagnostic workload. The improved contextual learning feature of the proposed model also leads to precise healthcare as it allows more precise lesion analysis and patient-specific diagnostic assistance to be done. In turn, the elaborated system of segmentation has serious prospective of incorporating into smart computer-aided diagnostics and high-technology medical imaging tools.

8. Benefits of Suggested Framework

The presented hybrid CNN-Transformer model exhibits a number of prominent benefits as compared to the more traditional medical image segmentation models. Among the main advantages of the suggested scheme are the enhanced boundary segmentation ability of the proposed model that has been accomplished by incorporating CNN-based spatial attribute extraction, and transformer-based contextual dependency learning. The framework is good at preserving fine anatomical features and demarcating irregular tumor boundaries, thus minimizing segmentation ambiguity and false-positive predictions. This is made possible by the addition of transformer attention, which facilitates a more accurate contextual comprehension since it embodies long-range semantic dependencies across medical images and can dramatically increase the consistency of segmentation under heterogeneous imaging conditions.

The other significant benefit of the suggested framework is that it has an improved feature extraction. ResNet50 encoder is effective in hierarchical local spatial representations learning and the transformer module provides reinforcement to the global contextual representation learning. This complementary integration enhances the multi-scale feature understanding and allows detection of complex pathological regions correctly. The framework also exhibits a strong medical image interpretation performance as it has high segmentation

reliability in different MRI imaging samples and tumor structures. Experimental results also indicate that the developed model can achieve better segmentation accuracy than the conventional CNN-based models, which are indicated by better Dice Similarity Coefficient (94.1%), Intersection over Union (90.3%) and final pixel accuracy (96.2%). The proposed architecture is very appropriate to intelligent medical imaging applications and computer-aided diagnostic applications due to these advantages.

9. Limitations

Although the proposed hybrid CNN-Transformer architecture has demonstrated an improved performance with respect to segmentation, there are various limitations. A significant constraint is the relatively large cost of computation of contextual learning with transformers. Combining the self-attention processes escalates the complexity of the computational process, both in the traininig and the inference stages, especially when trying to process high-resolution medical images. There is also a high dependence in the proposed framework on large-scale annotated medical imaging datasets in order to attain the best segmentation performance. The lack of training information can decrease the capacity of generalization of the model and can influence the learning of contextual representations.

The other weakness is the introduction of bias on the memory due to the introduction of transformer modules particularly when executing multi-head self-attention. Transformer embeddings require a large amount of GPU memory to run effectively, which could be a constraint with regard to resource-constrained healthcare systems. Also, the proposed model is not as efficient in terms of real time deployment due to the computational overhead and complexity of inferences, despite achieving good accuracy in segmentation. More optimization is thus needed to enable feasibility of deployments in edge-based healthcare settings as well as mobile medical imaging equipment.

10. Future Work

The future research can aim at enhancing the computational efficiency and scalability of the proposed hybrid segmentation framework with lightweight integrating transformer strategies and model compression strategies. Variants of efficient transformers based on their lower parameter complexity can greatly enhance the ability to deploy them in real-time and still be accurate in the segmentation. Federated medical learning is another potential avenue with the promise to train segmentation models based on collaboration among distributed healthcare institutions, without sharing sensitive patient data directly, thus enhancing the privacy of data and the variety of datasets.

Future efforts can also investigate deployment of the proposed framework real time to be incorporated into the portable diagnostic machines and intelligent health care monitoring devices. Contextual representation and robustness of segmentation It can be further enhanced by incorporation of multi-modal medical imaging data (MRI, CT, PET, and ultrasound image) across various clinical applications. Furthermore, explainable medical artificial intelligence methods can be incorporated into the framework to increase the model interpretability and to make clinicians have clear decision-support information about segmentation prediction and diagnostic rationale.

11. Conclusion

This study provided a hybrid CNN-Transformer architecture that reports precise medical image segmentation, by combining CNN based local feature extraction and transformer based global contextual learning. The suggested structure was able to enhance the consistency of segmentation, contextualization, and definition of anatomical limits in complicated medical imaging situations. Experimental results on the BraTS dataset show that the architecture has achieved improved performance in terms of Dice Similarity Coefficient, Intersection over Union and pixel-wise accuracy over their traditional segmentation architecture such as U-Net, SegNet, DeepLabV3+ and TransUNet. With the combination of CNN and transformer modules, multi-scale features could be effectively learned and the capacity to learn semantic representations was improved, thus minimizing lesion localization and false segmentation areas. The proposed framework also proved to be highly relevant to

clinical practice as it assisted in providing automated diagnostic aid, decreasing the number of manual annotations, and enhancing segmentation accuracy to utilize healthcare with high precision. Even though the problem of computational complexity and transformer memory still has a significant role, the created architecture holds significant potential in terms of intelligent healthcare systems of the future, computer-aided diagnosis systems, explainable medical AI systems, and future medical imaging approaches.

References

1. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.
2. 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*.
3. Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E. K., Cohen-Adad, J., &Merhof, D. (2023). Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 6202-6212).
4. Kucer, M., Oyen, D., Castorena, J., & Wu, J. (2022). Deeppatent: Large scale patent drawing recognition and retrieval. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2309-2318).
5. Lan, L., Cai, P., Jiang, L., Liu, X., Li, Y., & Zhang, Y. (2026). Brau-net++: U-shaped hybrid cnn-transformer network for medical image segmentation. *IEEE Transactions on Radiation and Plasma Medical Sciences*.
6. Lei, T., Sun, R., Wang, X., Wang, Y., He, X., & Nandi, A. (2023). CiT-Net: convolutional neural networks hand in hand with vision transformers for medical image segmentation. *arXiv preprint arXiv:2306.03373*.
7. Liu, X., Hu, Y., & Chen, J. (2023). Hybrid CNN-Transformer model for medical image segmentation with pyramid convolution and multi-layer perceptron. *Biomedical Signal Processing and Control*, *86*, 105331.
8. Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).
9. Manzari, O. N., Kaleybar, J. M., Saadat, H., & Maleki, S. (2024). BEFUnet: A hybrid CNN-transformer architecture for precise medical image segmentation. *arXiv preprint arXiv:2402.08793*.
10. Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing.
11. Sun, R., Lei, T., Zhang, W., Wan, Y., Xia, Y., & Nandi, A. K. (2023). TEC-Net: Vision transformer embrace convolutional neural networks for medical image segmentation. *arXiv preprint arXiv:2306.04086*.
12. Tao, R., & Zheng, G. (2021, September). Spine-transformers: Vertebra detection and localization in arbitrary field-of-view spine ct with transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 93-103). Cham: Springer International Publishing.
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., &Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.
14. Yao, W., Bai, J., Liao, W., Chen, Y., Liu, M., & Xie, Y. (2024). From cnn to transformer: A review of medical image segmentation models. *Journal of Imaging Informatics in Medicine*, *37*(4), 1529-1547.
15. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2019). Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, *39*(6), 1856-1867.
16. K. Maidanov, H. Fratlin. (2026). High-Frequency Planar Transformer Architecture for Ultra-Compact Power Conversion in Renewable Energy Systems. *National Journal of Electrical Machines & Power Conversion*, 27-37.
17. Sumit Ramswami Punam. (2025). Trust-Constrained Learning-Assisted Predictive Control for Real-Time Trajectory Planning in Distributed Autonomous Platforms. *Transactions on Internet Security, Cloud Services, and Distributed Applications*, 1–9.
18. Marie Johanne, Andreas Magnus, Ingrid Sofie, & Henrik Alexander. (2025). IoT-Based Smart Grid Systems: New advancement on Wireless Sensor Network integration. *Journal of Wireless Sensor Networks and IoT*, *2*(2), 1-10. https://doi.org/10.31838/WSNIOT/02.02.01