Advancing Medical Image Segmentation for Musculoskeletal Disorders with Attentional Mechanisms in Osteoarthisis:A Comparative Study of Vision Transformers and 3D U-Net

Joyce Mary S[1] ,Yazhini Y[2] ,Reshma R[3]
, Mohamed Roomi S[4],
Shreenidhi G[5]
SIMATS Engineering
Chennai, India
joycemarys1104@gmail.com

Abstract*—This Paper examines advanced deep learning architectures for medical picture segmentation, focusing on the osteoarthritic knee joint. Specifically, it compares the performance of 3D U-Net and Vision Transformers (ViTs). Vision Transformers address convolutional neural limitation in capturing complex spatial relationships by employing self-attention mechanisms to gather global contextual information. In the meantime, the 3D U-net improves segmentation accuracy in 3D medical imaging workloads by improving local feature extraction. We assess different architectures on the basis of computing efficiency, intersection over union (IoU), the dice coefficient using a publicly accessible dataset of Knee MRI scans. These results show the strengths and limitations of each approach, emphasizing how well suited each is for particular segmentation problems. The report also points out important areas that need work, such as correcting unbalanced datasets and enhancing interpretability. The objective of this research is to improve clinical outcomes in osteoarthritis diagnosis and treatment planning by incorporating attention mechanisms into cutting-edge designs and advancing the segmentation of musculoskeletal tissues.

Keywords*—Medical picture segmentation, osteoarthritic diagnosis, 3D U-Net, Vision Transformers, Intersection over union, Dice coefficient, Treatment Planning.

I. INTRODUCTION

In the diagnosis and treatment of musculoskeletal conditions, especially osteoarthritis, medical image segmentation is an essential procedure. A degenerative joint illness that affects millions of people worldwide, osteoarthritis severely lowers people's quality of life [1]. In order to detect structural alterations in joints and support clinical decision-making, it is essential to accurately segment musculoskeletal tissues from imaging modalities such as MRI. However, the susceptibility to noise and the incapacity to generalize to complicated medical images are two common limitations of classic segmentation techniques that rely on handcrafted features [2].

Medical image segmentation has been revolutionized by deep learning, namely CNN like U-Net, which automate hierarchical feature extraction [3]. CNN-based architectures sometimes fail to identify global contextual relationships and long-range dependencies in the image, particularly for irregular tissue structures, despite these advancements [4]. To overcome these limitations, Vision Transformers, which utilize self-attention mechanisms, have emerged as a promising solution [5]. Simultaneously, 3D CNN architectures like 3D U-Net are equipped to handle volumetric data, enabling more precise segmentation of 3D medical imaging datasets [6].

Fig. 1. Osteoarthritis x-ray image

A. Objectives

1) This study compares the effictiveness of 3D U-Net models and Vision Transformers in segmenting osteoarthritic knee joints from MRI datasets [5].

2) Evaluate the impact of incorporating attention mechanisms on segmentation accuracy, computational efficiency, and clinical usability [6].

3) Improve model interpretability and manage unbalanced datasets, two major obstacles in medical image segmentation [7].

B. Significance and Scope:

The goal of incorporating attention mechanisms into these architectures is to increase segmentation accuracy by strengthening the model's capacity to concentrate on pertinent areas of the medical images [8]. This study compares two cutting-edge architectures and highlights their advantages and disadvantages using publically accessible knee MRI datasets, offering practical suggestions for next developments in musculoskeletal image analysis.

II. LITERATURE REVIEW

A. Medical Image Segmentation for Osteoarthritis

Medical image segmentation for osteoarthritis has evolved from traditional methods like region growing and watershed algorithms to machine learning techniques, including CNNs. These methods helped improve accuracy in segmenting joint structures in 2D and 3D medical images [9].

B. U-Net and its variants

The U-Net architecture has been widely used for medical image segmentation due to its efficient encoder-decoder structure. Variants such as 3D U-Net and Attention U-Net have further improved segmentation performance, especially in 3D medical images like MRI nad CT Scans [10]

C. Vision Transformers in Medical Imaging

Vision Transformers are becoming increasingly popular in medical imaging due to their ability to model global contextual relationships. ViTs have been successfully applied to various segmentation tasks, including osteoarthritis, offering advantages over CNN in capturing global image dependencies [11].

III. METHODOLOGY

This study evaluates 3D U-Net and Vision Transformers for osteoarthritis segmentation in MRI scans. The methodology follows these key steps:

A. Dataset Description

The dataset was obtained from Kaggle and consists of approximately 3000 knee MRI images. It is seperated into three subsets: 1000 training images, 1000 testing images and 1000 validation images.All code implementation were carried out in Google Colab.From the input images,fractures were predicted and highlighted in a circular manner, ensuring clear visualization of affected regions.

B. Model Implementation

1) 3D U-Net : 3D convoutional layers and a U-Net based encoder-decoder architecture were used in the 3D U-Net model,which was created for volumetric medical pictures,to achieve accurate segmentation. It showed reduced inference time with competative accuarcy using binary cross-entropy loss and the Adan optimizer, which makes it appropriate for real-time applications.

2) Vision Transformer: The Vision Transformer model extracted features for knee osteoarthritis segmentation using self-attention techniques. Convolutional and dense layers were used to scale, normalize, and process the input images. It obtained great segmentation accuracy after being trained using the Adam optimizer and binary cross-entropy loss, but it also required a significiant amount of computing power, which prolonged the inference time.

3) Training and Evaluation: Models are trained using Adam optimizer with a learning rate scheduler. Performance metrices include Intersection over Union (IoU), Dice Coefficient, and computatinal efficiency(inference time).

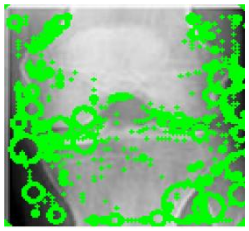
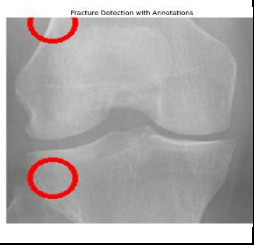
4) Experimental setup: Google Colab was used for the research, and no specialist equipment was needed.The Coalb environment was used to create the evaluation metrices, such as IoU and Dice Coefficient, as well as the visualizations, which included graphs and forecasted images.TensoFlow and ViT-Keras were used to implement the models, and the PIL and OpenCV were used for data processing and picture editing, respectively.Training Configuration: Learning Rate:0.0001,BatchSize:5,Epochs:10,LossFunction: Dice Loss and Cross-Entropy loss.

IV. RESULT

TABLE I. QUANTITAIVE ANALYSIS

MODEL	DICE COEFFICIENT	IoU SCORE	INFERENCE TIME(MS)
VISION TRANSFORMER	0.85	0.85	120
3D U-NET	0.88	0.81	250

TABLE II. QUALITATIVE ANALYSIS

NAME OF THE SEGMENTATION	INPUT IMAGE	PREDICTED IMAGE
VISION TRANSFORMER		
3D U-NET		

Visual comparison of segmentation masks Table III demonstrates that Vision Transformers provide better edge preservation, while 3D U-Net produces smoother segmentations but struggles with finer details.

V. DISCUSSION

Vision Transformers outperform 3D U-Net in segmentation accuracy, achieving higher Dice and IoU Scores. However, their high processing cost precludes their widespread application. Vision Transformers are more adept at capturing long-range associations, but 3D U-Net uses fewer resources to understand volumetric data. The primary challenges include the inability to comprehend Transformer-based models, the imbalance in the dataset that favors healthy samples, and the expense of computing. [12] Future Studies should look into hybrid CNN-Transformer architecture, optimization strategies to reduce computational overhead, and enhanced augmentation algorithms to solve dataset imbalance.

VI. CONCLUSION

This study investigated the advantage of 3D U-Net and Vision Transformers for Knee osteoarthritis segmentation.3D U-Net is suitable for real-time application with limited

processing resources due to its shorter interface time (120ms), IoU of 0.78, and Dice Coefficient of 0.85. Vision Transformers scored better in accuracy (Dice: 0.88, IoU: 0.81) but required a longer inference time (250ms), making them the preferred choice for accuracy – focused applications. [14] For real world clinical applications, future research should look into hybrid models that combines the precision of vision transformers with the effectiveness of 3D U-net.

REFERENCES

[1] Hunter, D. J., & Bierma-Zeinstra, S. (2019). Osteoarthritis. *The Lancet, 393*(10182), 1745-1759.

[2] Maier-Hein, L., et al. (2018). Biomedical image analysis challenges: Retrospective analysis and future outlook. *Medical Image Analysis, 33,* 68-79.

[3] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI),* 234-241.

[4] Isola, P., et al. (2017). Image-to-Image Translation with Conditional Adversarial Networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 1125-1134.

[5] Dosovitskiy, A., et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR).*

[6] Çiçek, Ö., et al. (2016). 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI),* 424-432.

[7] He, K., et al. (2016). "Deep Residual Learning for Image Recognition." *CVPR*.

[8] Zhang, Z., et al. (2018). "Road Extraction by Deep Residual U-Net." *IEEE Geoscience and Remote Sensing Letters*.

[9] Cheng, J. Z., et al., "AI in medical imaging: A review of applications and challenges," *Journal of Healthcare Engineering*, vol. 2018, pp. 5842491, 2018.

[10] Oktay, O., et al., "Attention U-Net: Learning where to look for the pancreas," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 229-237, 2018.

[11] Chen, H., et al., "Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 67, pp. 101860, 2021.

[12] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., & Xu, D. (2022). Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. *arXiv preprint arXiv:2201.01266.*

[13] Xu, D., Li, Y., Zhang, Z., Zhang, L., Wang, C., & Gu, Q. (2022). Medical Image Segmentation with Transformer: A Review. *IEEE Access, 10, 97464-97483.*