Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 115-134

Integrative Deep Learning Architecture for High-Accuracy Medical Image Segmentation: Combining U-Net, ResNet, and Transformers

Devi Zakiyatus Solekhah*[1], Dian Noviar[1]
Email: devizakiyatus123@gmail.com; dian.noviar@uin-suka.ac.id
[1]*Universitas Islam Negeri (UIN) Sunan Kalijaga, Yogyakarta, Indonesia*
*Corresponding Author

Abstract
Medical image segmentation plays a vital role in diagnosis and treatment planning by extracting clinically relevant information from imaging data. Conventional methods often struggle with variations in anatomical structure and imaging quality, leading to suboptimal segmentation. Recent advancements in Deep Learning, particularly Convolutional Neural Networks (CNNs) and Transformers, have improved segmentation accuracy; however, individual models such as U-Net, ResNet, and Transformer still face limitations in preserving spatial details, extracting deep features, and modeling long-range dependencies. This study proposes a hybrid Deep Learning model that integrates U-Net, ResNet, and Transformer to overcome these challenges and enhance segmentation performance. The proposed hybrid model was evaluated on several publicly available datasets, including BraTS, ISIC, and DRIVE, using Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) as performance metrics. Experimental results indicate that the hybrid model achieved a DSC of 0.92 and an IoU of 0.86, outperforming U-Net (DSC: 0.82, IoU: 0.75), ResNet (DSC: 0.85, IoU: 0.78), and Transformer (DSC: 0.88, IoU: 0.80). Additionally, the model maintained an inference time of 55 ms per image, demonstrating its potential for real-time applications. This study highlights the benefits of combining CNN-based and Transformer-based architectures to capture both local details and global context, providing an effective and efficient solution for medical image segmentation.

Keywords: *Medical Image Segmentation, Deep Learning, Hybrid Model.*

I. INTRODUCTION

Medical image segmentation is one of the key elements in Computer-Aided Diagnosis (CAD) systems. This technique is used to extract specific regions from medical images, such as tumors, organs, or abnormal tissues, to assist in the diagnosis process and treatment planning. Conventional image processing-based methods, such as thresholding and edge detection, often face limitations in capturing the complexity of biological structures. According to (Melyani et al., 2024), the primary challenge of these approaches lies in their inability to recognize variations in shape, texture, and contrast in medical images from different imaging modalities. These variations may be caused by factors such as differences in imaging quality, the presence of artifacts, and inter-individual anatomical variations, which are difficult to model using rule-based methods or simple filters. According to (Priyadi et al., 2024), to address these limitations, deep learning-based approaches are increasingly being applied due to their ability to automatically extract complex features from large image datasets, thereby improving the accuracy and efficiency of medical image segmentation.

Received on February 2025; Revised on March 2025; Accepted on March 2025; Published on April 2025
DOI: *10.51903/jtie.v4i1.288*

Medical image segmentation plays a crucial role in the detection and analysis of various diseases, including cancer, tumors, and other disorders, by enhancing diagnostic accuracy and expediting clinical decision-making. According to (Obuchowicz et al., 2024), accurate segmentation can help physicians identify pathological boundaries more precisely, thereby improving the effectiveness of treatment planning. Research conducted by (Xu et al., 2024) demonstrates that deep learning-based models offer advantages over conventional methods as they can automatically extract complex features from various types of medical images. (Punn & Agarwal, 2022) add that CNN-based architectures, such as U-Net and ResNet, have been widely used in medical image segmentation due to their ability to preserve spatial information and extract features hierarchically. However, according to (Wang et al., 2022), CNNs still have limitations in capturing long-range spatial relationships in images, prompting the adoption of Transformers to enhance global context modeling and segmentation accuracy. (Rayed et al., 2024) also demonstrate that the combination of CNNs and Transformers can address the limitations of each architecture by offering a balance between accuracy and computational efficiency, making it an increasingly explored solution in medical image segmentation.

Although various studies have examined the effectiveness of deep learning models in medical image segmentation, most have focused on the use of a single type of architecture. According to (Shi et al., 2022), U-Net is effective in maintaining segmentation details but often struggles with handling complex texture variations. Meanwhile, research by (Wu et al., 2025) shows that ResNet excels in deep feature extraction but is less optimal in reconstructing object boundaries with high precision. On the other hand, (Pu et al., 2024) reveal that Transformer-based approaches enhance global context modeling but have limitations in capturing the local details necessary for medical image segmentation. Research by (Arkin et al., 2023) attempts to integrate CNNs and Transformers to improve segmentation accuracy but still faces challenges in high computational efficiency. Additionally, recent studies by (Eum et al., 2025) indicate that while combining multiple models can enhance performance, no approach has yet optimally integrated the advantages of each architecture into a single framework that balances accuracy and efficiency. Therefore, this study aims to develop a hybrid model that combines U-Net, ResNet, and Transformer architectures to overcome the individual limitations of each model and enhance the accuracy and efficiency of medical image segmentation.

This study aims to develop a hybrid deep learning model that integrates U-Net, ResNet, and Transformer to improve accuracy and efficiency in automated medical image segmentation. This model is expected to overcome the limitations of each architecture by leveraging U-Net's ability to preserve spatial information, ResNet's capability in deep feature extraction, and Transformer's strength in modeling long-range spatial relationships. One of the key aspects to be evaluated is

the performance of the hybrid model compared to individual models across various metrics, such as the DSC and IoU, to measure segmentation accuracy. Additionally, this study will also examine the computational efficiency of the model to ensure that the proposed approach is not only accurate but also applicable in clinical environments with limited computational resources. By developing a more optimal hybrid model, medical image segmentation is expected to be performed more quickly and accurately, thereby supporting CAD systems more effectively. The results of this study are expected to contribute to the advancement of deep learning-based medical image segmentation technology and open new opportunities for further research in optimizing hybrid model architectures in the future.

II. LITERATURE REVIEW

1. Segmentation in Medical Imaging: Conventional Techniques vs. Deep Learning

Medical image segmentation is one of the primary techniques in medical imaging analysis, used to extract anatomical structures or pathological regions across various imaging modalities. According to (Mohapatra et al., 2024), conventional methods such as thresholding, region growing, and active contouring have been widely applied in medical image segmentation due to their simplicity and ease of implementation. However, these approaches often rely on manually adjustable parameters and are susceptible to variations in lighting, contrast, and noise in medical images. Additionally, rule-based techniques tend to struggle with recognizing complex object shapes, making them less effective in cases involving medical images with high structural variability. Further research has shown that statistical model-based methods, such as the Gaussian Mixture Model (GMM) and Markov Random Field (MRF), can improve segmentation results by incorporating statistical information from pixels, yet they still exhibit limitations in capturing more complex spatial features. The evolution of segmentation techniques continues to explore more adaptive and intelligent approaches, one of which involves machine learning and deep learning.

In recent years, deep learning has become the dominant approach in medical image segmentation due to its ability to automatically extract complex features without requiring manual feature engineering. According to (Aboussaleh et al., 2023), U-Net is one of the most widely used deep learning architectures in medical image segmentation because its encoder-decoder design allows for effective reconstruction of spatial details. This model has been successfully applied to various types of medical imaging, such as MRI, CT scans, and microscopic images, yielding superior results compared to conventional methods. In addition to U-Net, several other CNN variants, such as Fully Convolutional Networks (FCN) and DeepLab, have been developed to enhance segmentation accuracy by considering broader contextual information within images. Recent

studies indicate that CNN-based approaches excel in segmenting objects with well-defined boundaries but still face challenges in handling complex shape and texture variations, particularly in low-resolution images or those with poor imaging quality. CNN-based models continue to be refined with various optimization techniques to enhance their adaptability to diverse medical imaging conditions.

Beyond CNNs, Transformers have been increasingly adopted in medical image segmentation to overcome CNNs' limitations in capturing long-range spatial relationships within images. According to (Maurício et al., 2023), the Vision Transformer (ViT) introduces a self-attention-based mechanism that can account for dependencies between pixels over a broader scope, thereby improving the overall contextual understanding of the image. Several studies have integrated Transformers into medical segmentation architectures, such as TransUNet, which combines U-Net's strength in preserving spatial information with Transformer-based feature modeling for improved global feature extraction. Experimental results demonstrate that Transformer-based models can achieve more accurate segmentation than CNNs in certain high-complexity medical imaging datasets. However, these models also require optimization in computational efficiency and adequate training data availability to reach optimal performance.

Recent research has also indicated that combining CNNs and Transformers can offer a more balanced solution between accuracy and efficiency in medical image segmentation. According to (Jiang et al., 2022), a hybrid approach integrating CNNs for local feature extraction with Transformers for long-range spatial relationship modeling can enhance segmentation accuracy without significantly compromising computational efficiency. Models such as Swin-UNet and Medical Transformer (MedT) have been developed to optimize the integration of these architectures, demonstrating improved performance compared to standalone models. Additionally, several studies have explored additional optimization techniques, such as multi-scale feature fusion mechanisms and attention-based refinement, to further improve segmentation accuracy without excessively increasing computational costs. The hybrid approach remains an evolving research topic due to its flexibility in adapting segmentation strategies to various types of medical imaging.

2. Advantages and Limitations of U-Net, ResNet, and Transformer in Segmentation

U-Net has become one of the most widely used architectures in medical image segmentation due to its ability to preserve spatial information through a symmetric encoder-decoder structure. According to (Han et al., 2022), the design of U-Net enables the model to process low-resolution images while retaining essential details through skip connection mechanisms. The primary advantage of this model lies in its ability to produce segmentations with clear boundaries and its

flexibility in being applied to various types of medical imaging, such as MRI, CT scans, and microscopic imaging. However, U-Net also has limitations, particularly in capturing global contextual information, which may affect its performance on images with complex structural variations. Additionally, the model relies on a sufficiently large training dataset to achieve optimal segmentation results. Several studies have attempted to enhance U-Net's performance by incorporating attention mechanisms or employing deeper architectures to capture more complex features.

ResNet, developed to address the vanishing gradient problem in deep networks, has been widely applied in medical image segmentation. (Athisayamani et al., 2023) stated that the residual connections in ResNet allow for the training of deeper networks without performance degradation, enabling the model to extract high-level features more effectively than conventional CNNs. The main advantage of ResNet in segmentation is its ability to handle more complex texture details and generate deeper feature representations. However, while it overcomes some of the limitations of traditional CNNs, ResNet has weaknesses in preserving spatial information, as its architecture is primarily focused on feature extraction rather than spatial reconstruction. Furthermore, the computational burden of ResNet can increase significantly when applied to high-resolution segmentation tasks. Some studies have proposed modified ResNet variants with alterations to residual block structures to improve computational efficiency and segmentation quality.

Transformers are increasingly being adopted in medical image segmentation as an alternative architecture capable of capturing long-range spatial relationships within images. According to (Ebert et al., 2023), ViT employs a self-attention mechanism that enables the model to consider dependencies between pixels over a broader scope than CNNs. The key advantage of this approach is its ability to understand the global context within an image, which is particularly useful in segmenting complex anatomical structures. However, Transformer-based models often require a large amount of training data to achieve optimal performance, as they lack the inductive bias inherent in CNNs that naturally extract local features. Additionally, Transformer models generally have higher computational demands compared to CNN-based architectures, posing challenges for deployment in systems with limited resources. Recent research has explored hybrid approaches that combine Transformers with CNNs to improve segmentation efficiency and accuracy.

Hybrid approaches integrating U-Net, ResNet, and Transformer are gaining attention in medical image segmentation research due to their ability to leverage the strengths of each architecture. (Chen et al., 2023) explained that combining CNNs with Transformers can enhance segmentation

performance by preserving the spatial detail of U-Net, improving feature extraction through ResNet, and expanding global context understanding using Transformers. Models such as TransUNet and Swin-UNet have been developed to explore this integration, with results demonstrating improved accuracy compared to individual models. Additionally, strategies such as multi-scale feature fusion and adaptive attention mechanisms have been implemented to further optimize interactions between different architectures in hybrid models. These advancements indicate that hybrid approaches remain an active area of research in the pursuit of more accurate and efficient medical image segmentation.

A. Previous Studies

1. Studies on the Use of U-Net for Medical Image Segmentation

U-Net has become one of the most widely used architectures in medical image segmentation due to its ability to handle various types of medical imaging with high accuracy. According to (Ji et al., 2024), U-Net features a symmetric encoder-decoder structure with skip connection mechanisms that enable optimal recovery of spatial information. This model has been applied in multiple medical fields, including brain tissue segmentation in MRI, lesion detection in fundus imaging, and cancer cell identification in histopathological images. Its ability to preserve spatial details makes it particularly effective for segmentation tasks requiring clear object boundaries. Additionally, U-Net can operate efficiently even with relatively small datasets due to its architecture, which optimizes training efficiency. Various studies continue to develop U-Net variants to further enhance segmentation accuracy across diverse medical applications.

The use of U-Net in medical image segmentation has consistently demonstrated reliable performance across different medical imaging datasets. According to (Yousef et al., 2023), a variant of U-Net known as V-Net was developed to improve 3D image segmentation, particularly in MRI imaging. This model employs a loss function based on the DSC, which is more suitable for medical segmentation than pixel-wise loss functions such as cross-entropy. This approach enhances segmentation accuracy for organ structures with significant variations in shape and size. Furthermore, V-Net has been utilized in various applications, including brain tumor segmentation and pathological tissue detection in ultrasound imaging. The primary advantage of this model lies in its ability to process volumetric data without requiring complex preprocessing.

Further developments in U-Net have led to the integration of attention mechanisms to enhance segmentation performance. (C. Zhang et al., 2024) introduced Attention U-Net, which incorporates attention gate mechanisms to emphasize relevant regions in medical images. This model improves focus on target segmentation areas while reducing the influence of background noise. This approach has been applied to various segmentation tasks, including cardiovascular

imaging and lesion detection in lung CT scans. The inclusion of attention mechanisms allows the model to selectively prioritize the most informative features, thereby improving segmentation accuracy. Related research indicates that attention-based U-Net models provide more stable and robust results compared to standard U-Net.

Beyond architectural enhancements, optimization in training strategies has also been explored to improve U-Net's performance in medical image segmentation. According to (Carles et al., 2024), nnU-Net was developed as an automated approach to adapt U-Net's architecture to specific datasets without requiring manual parameter tuning. This model can automatically adjust filter sizes, data augmentation strategies, and training schemes based on dataset characteristics. This approach has been tested on various medical segmentation datasets and has shown competitive results compared to manually designed models. Due to its flexibility in adapting training configurations, nnU-Net has become one of the most widely adopted methods in deep learning-based medical segmentation competitions. Studies on U-Net and its variants continue to contribute to advancements in medical image segmentation, improving its applications in both clinical and research settings. Table 1 presents an overview of various U-Net variants developed to overcome the limitations of the baseline model and enhance segmentation accuracy in different medical applications.

Table 1. Comparison of Previous Studies on Medical Image Segmentation Methods

Researcher	U-Net Model	Advantages	Limitations
(Ji et al., 2024)	Symmetric encoder-decoder with skip connections	Effective in segmenting low-resolution images, capable of working with small datasets	Less optimal in capturing global contextual information
(Yousef et al., 2023)	3D variant of U-Net using Dice Loss	Capable of handling volumetric imaging, accurate in segmenting complex organ structures	Requires higher computational power compared to standard U-Net
(C. Zhang et al., 2024)	Integration of attention gate mechanism	Enhances focus on target segmentation areas, reduces background interference	More complex and requires a larger number of parameters
(Carles et al., 2024)	Automatically adjusts architectural configuration	Does not require manual tuning, adapts training schemes to different datasets	Performance depends on the availability of sufficient data

2. Studies on the Combination of CNN and Transformer to Improve Segmentation Accuracy

The combination of CNN and Transformers in medical image segmentation has gained increasing attention due to the complementary strengths of each architecture in handling spatial features and global relationships. According to (Yang & Wang, 2024), CNN excels at extracting local features through convolution operations but is less effective in capturing long-range spatial dependencies

within images. On the other hand, Transformers, initially introduced in the field of Natural Language Processing (NLP), have proven effective in image processing through self-attention mechanisms. By integrating CNN and Transformer architectures, models can achieve a more comprehensive feature representation, leading to more accurate segmentation across various types of medical imaging. Several studies have demonstrated that this hybrid approach outperforms segmentation methods based solely on CNNs or Transformers. Models that integrate both approaches have been tested in various segmentation tasks, including brain tumor segmentation and lung lesion detection.

The use of CNN and Transformer architectures in hybrid models has resulted in several frameworks that enhance segmentation accuracy. (Pan et al., 2023) developed TransUNet, a model that incorporates Transformer components into the U-Net architecture to improve global contextual understanding in medical image segmentation. This model employs a Transformer-based encoder to capture spatial relationships over a broader range while maintaining a CNN-based decoder to preserve essential spatial details. Experimental results indicate that TransUNet achieves higher segmentation precision compared to standard U-Net, particularly in organ segmentation tasks using CT and MRI imaging. This combination enables the model to extract richer feature representations, thereby improving segmentation quality in various clinical applications.

Beyond TransUNet, several other models have been developed to optimize the integration of CNN and Transformer architectures in medical image segmentation. According to (J. Zhang et al., 2023), Swin-UNet is one such approach that utilizes the Swin Transformer to replace traditional convolutional layers in feature extraction. This model divides images into small patches and applies self-attention mechanisms at a more flexible scale, allowing for more efficient modeling of spatial relationships. Experimental results on medical segmentation datasets indicate that Swin-UNet outperforms conventional CNN-based models. The primary advantage of this approach lies in its ability to capture multi-scale features without losing fine details, making it a promising method for various types of medical imaging applications.

Further advancements in hybrid CNN-Transformer models continue to focus on improving segmentation accuracy and computational efficiency. According to (Xiao et al., 2023), hierarchical Transformer-based approaches are increasingly being applied in medical image segmentation to optimize both global and local feature processing simultaneously. Models such as MedT have been introduced with architectures that leverage hierarchical self-attention to capture features at multiple resolution levels. By combining CNN elements for local feature extraction and Transformer mechanisms for long-range spatial understanding, these models

achieve more precise segmentation across diverse medical imaging datasets. This study highlights the ongoing evolution of hybrid approaches, with various architectural innovations enhancing the accuracy and efficiency of medical image segmentation across different applications.

III. RESEARCH METHOD

This study employs an experimental research approach by implementing a hybrid Deep Learning-based model for the automatic segmentation of medical images. This approach was selected because it enables an empirical evaluation of various models' performance in segmentation tasks, which is a crucial aspect of developing AI-based medical technologies. The model developed in this study integrates U-Net, ResNet, and Transformer architectures, each contributing distinct advantages to medical image processing. U-Net is designed to capture spatial details effectively, ResNet enhances feature extraction through deep residual learning, and Transformer excels in understanding long-range spatial dependencies within images. The combination of these three models aims to overcome the individual limitations of each architecture, ultimately achieving more accurate and efficient segmentation. The hybrid model is then compared with individual models to assess its superiority in segmenting various types of medical images, including images from different modalities such as MRI, skin images, and retinal fundus images.

The dataset used in this study is sourced from publicly available medical image databases that have been extensively utilized in segmentation research and are known to provide suitable challenges for testing model reliability. The datasets include various types of medical images with diverse characteristics, allowing for model evaluation across complex clinical scenarios. For instance, the BraTS dataset provides brain MRI images with tumor segmentation labels, posing a challenge for models to detect tumor boundaries. The ISIC dataset consists of skin images used for skin cancer diagnosis, where variations in skin color and texture necessitate adaptive segmentation techniques. Meanwhile, the DRIVE dataset contains retinal fundus images used for detecting diabetic retinopathy, which requires highly precise blood vessel segmentation. The diversity of these medical images enables the study to assess whether the hybrid model can adapt to different imaging conditions. Table 2 presents the datasets used, along with information on image types, the number of images, resolution, and related disease categories.

Table 2. Medical Image Datasets Used in the Study

Dataset	Image Type	Number of Images	Resolution	Disease Category
BraTS	Brain MRI	±3000	240×240	Brain tumor
ISIC	Skin Images	±25000	Variable	Skin cancer
DRIVE	Retinal Fundus	±400	584×565	Diabetic retinopathy

The models utilized in this study consist of three main components, each contributing unique strengths to medical image segmentation tasks. Each model was selected based on its ability to enhance segmentation accuracy through different approaches to feature extraction and spatial information processing. U-Net, a CNN-based model, is specifically designed for medical image segmentation using an encoder-decoder architecture, which enables effective spatial detail preservation. This model is widely adopted due to its ability to maintain high-resolution information through skip connections linking the encoder and decoder. ResNet, on the other hand, serves as a feature extractor, enhancing feature representation through deep residual networks, which stabilize model training and mitigate vanishing gradient issues in very deep networks. Meanwhile, Transformer leverages the self-attention mechanism to capture long-range spatial relationships, making it particularly useful for segmenting complex structures with significant variations in shape and size. The combination of these three models facilitates the creation of a hybrid architecture that optimizes segmentation accuracy and efficiency across various medical image types. Table 3 outlines the model configurations used in this study, including the number of layers, activation functions, and the primary advantages of each model.

Table 3. Model Configurations Used in the Study

Model	Number of Layers	Activation Function	Advantages
U-Net	5	ReLU	High spatial detail
ResNet	50	ReLU	Deep feature extraction
Transformer	12	GELU	Global feature dependencies

The analysis process in this study consists of three main stages aimed at ensuring that the model operates optimally in medical image segmentation. The first stage is data preprocessing, which involves several techniques to enhance data quality before model training. One of the techniques applied is image augmentation, which aims to increase data variability and reduce the risk of model overfitting. Augmentation is performed using various transformations, such as rotation, flipping, scaling, and adjustments in contrast and brightness, enabling the model to learn from diverse imaging conditions. Additionally, pixel value normalization is conducted to ensure a more stable data distribution, facilitating smoother convergence of the neural network during training. Once preprocessing is complete, the dataset is divided into three parts: 70% for training, 15% for validation, and 15% for testing. This distribution ensures a comprehensive evaluation of the model's performance across different learning stages.

The training of the hybrid model integrates three main architectures—CNN (U-Net, ResNet) and Transformer—to achieve optimal segmentation performance. The model is trained using a supervised learning approach with the preprocessed dataset, enabling it to recognize patterns in medical images effectively. To optimize the learning process, the Adam optimizer is employed, as it is known for accelerating convergence and reducing fluctuations during training.

The initial learning rate is set at 0.001, with gradual decay applied to prevent the model from being trapped in suboptimal local minima. Furthermore, the loss function used in this study is Dice Loss, which is more suitable for segmentation tasks than cross-entropy loss, as it effectively handles class imbalance within the dataset. Dice Loss assigns higher weights to pixels relevant to the segmentation target, thereby improving segmentation accuracy in medical imaging tasks.

The model evaluation is conducted using two primary metrics commonly employed in medical image segmentation to measure accuracy and performance. The first metric is the DSC, which quantifies the similarity between the model's segmentation output and the ground truth provided in the dataset. A higher DSC value indicates that the model's segmentation closely approximates the ground truth, making it more reliable for medical applications. The second metric is IoU, which compares the segmented area predicted by the model with the actual area in the image, assessing how well the model captures relevant structures. A high IoU score signifies a significant overlap between the predicted and target areas, providing a strong indication of accurate segmentation performance. These two metrics are used to evaluate the advantages of the hybrid model compared to individual models and to identify areas for further improvement in the developed architecture.

The model evaluation stage is conducted to assess the performance of the algorithm in medical image segmentation based on metrics that reflect segmentation accuracy. In this study, the evaluation is carried out using two primary metrics: DSC and IoU, both of which are widely used in segmentation tasks. These metrics measure the degree of similarity between the model's segmentation output and the ground truth defined in the dataset, taking into account prediction errors in the form of False Positives (FP) and False Negatives (FN). DSC is calculated using the equation presented in Formula (1):

$$DSC = \frac{2TP}{2TP+FP+FN} \tag{1}$$

where True Positive (TP) represents the number of pixels correctly classified as part of the target object, while FP refers to the number of pixels incorrectly classified as part of the object. Conversely, FN denotes the number of pixels that should belong to the object but were not correctly classified by the model. The DSC value ranges from 0 to 1, with a higher value indicating a greater similarity between the segmentation result and the ground truth, thus providing a more accurate evaluation of the model's segmentation quality.

In addition to DSC, the model is also evaluated using IoU, a widely used metric in medical image segmentation. IoU calculates the proportion of the correctly segmented area relative to the

total predicted and actual segmented area, incorporating misclassification errors in the form of FP and FN. The equation used to compute IoU is given in Formula (2):

$$IoU = \frac{TP}{TP+FP+FN} \quad (2)$$

The IoU value ranges from 0 to 1, with a higher value indicating a more accurate segmentation, as the segmented result has greater overlap with the ground truth. IoU is often used alongside DSC to provide a more comprehensive assessment of model performance in segmentation tasks that require high precision. The use of both metrics helps evaluate the effectiveness of the hybrid model developed in this study.

Table 4 presents the configuration of the hybrid model used in this study, which consists of a combination of U-Net, ResNet, and Transformer architectures. The hybrid model is designed to optimize medical image segmentation performance by leveraging the advantages of each architecture in feature extraction and spatial structure understanding. The model is trained using the Adam optimizer, known for its effectiveness in adjusting network weights with stable convergence rates. The initial learning rate is set at 0.001 to ensure a gradual learning process and prevent excessive weight adjustments in each iteration. The loss function varies depending on the model characteristics: Dice Loss is applied to U-Net, Transformer, and the hybrid model to enhance segmentation accuracy, while Cross-Entropy is used for ResNet, as it is more suitable for classification tasks. The models are evaluated using DSC and IoU for segmentation performance, while accuracy is used to assess the performance of ResNet in extracting features relevant to medical image segmentation.

Table 4. Hybrid Model Configurations Used in the Study

Model	Optimizer	Learning Rate	Loss Function	Evaluation Metrics
Hybrid (U-Net + ResNet + Transformer)	Adam	0.001	Dice Loss	DSC, IoU
U-Net	Adam	0.001	Dice Loss	DSC, IoU
ResNet	Adam	0.001	Cross-Entropy	Accuracy
Transformer	Adam	0.001	Dice Loss	DSC, IoU

IV. RESUL

A. Results

1. Model Evaluation and Segmentation Performance Comparison

The model evaluation was conducted to compare the segmentation performance between the hybrid model (U-Net + ResNet + Transformer) and individual models in the context of medical image analysis. This evaluation process included measuring segmentation accuracy and inference

time efficiency to determine the model's effectiveness in practical applications. Segmentation accuracy was the primary aspect analyzed, as it directly relates to the model's ability to accurately identify structures or objects within medical images. Additionally, inference time efficiency was considered a crucial factor in assessing the feasibility of implementing the model in real-time systems, particularly in clinical environments where rapid and accurate results are essential. To ensure comprehensive results, the evaluation was conducted using a standardized dataset that reflects various conditions in medical imaging. This multi-faceted analysis provided a thorough assessment of the strengths and limitations of each model in medical image segmentation tasks.

Segmentation accuracy was measured using the DSC and IoU, two key metrics commonly used to evaluate medical image segmentation performance. DSC measures the similarity between the model's segmentation output and the reference label, whereas IoU assesses the degree of overlap between the detected area and the actual target area. The hybrid model demonstrated superior performance compared to individual models based on both metrics, as summarized in Table 5. This improvement in accuracy indicates that the combination of CNN-based and Transformer-based approaches enhances the model's ability to recognize patterns and structures in medical images more effectively. Furthermore, the evaluation results also revealed that the hybrid model exhibited more consistent performance across different types of medical images compared to individual models. These findings highlight the positive impact of the hybrid approach on both segmentation accuracy and stability in medical imaging applications.

Table 5. DSC & IoU Performance Evaluation for Each Model

Model	DSC Score	IoU Scor
U-Net	0.82	0.75
ResNet	0.85	0.78
Transformer	0.88	0.80
Hybrid Model	**0.92**	**0.86**

Based on the results presented in Table 5, the hybrid model achieved the highest DSC score of 0.92 and the highest IoU score of 0.86, indicating a more accurate segmentation performance compared to individual models. The high DSC score demonstrates that the hybrid model produces segmentation predictions that closely resemble the reference labels, while the high IoU score reflects better overlap between the predicted and actual target areas. This advantage can be attributed to the ability of the CNN-Transformer combination to capture local features while preserving long-range spatial relationships in medical images. CNNs are particularly effective at identifying texture patterns and fine details in images, whereas Transformers excel in understanding broader spatial relationships across different regions of an image. By integrating these approaches, the hybrid model achieves a balance between local precision and global

contextual understanding in segmentation tasks. The combination of these methods significantly enhances segmentation performance, particularly when dealing with complex variations in medical images.

Furthermore, the pre- and post-segmentation results using the hybrid model are presented in Figure 1. The segmented regions are highlighted in red, indicating the detected target areas with greater clarity. This color mapping is intended to facilitate visual analysis of the segmentation quality produced by the model. By incorporating a distinct color contrast, the areas identified by the model can be compared with the reference labels to assess segmentation accuracy. Additionally, these visual results provide insights into the model's ability to distinguish between target areas and irrelevant background regions. The information obtained from Figure 1 serves as an important indicator for evaluating the model's effectiveness in real-world applications.

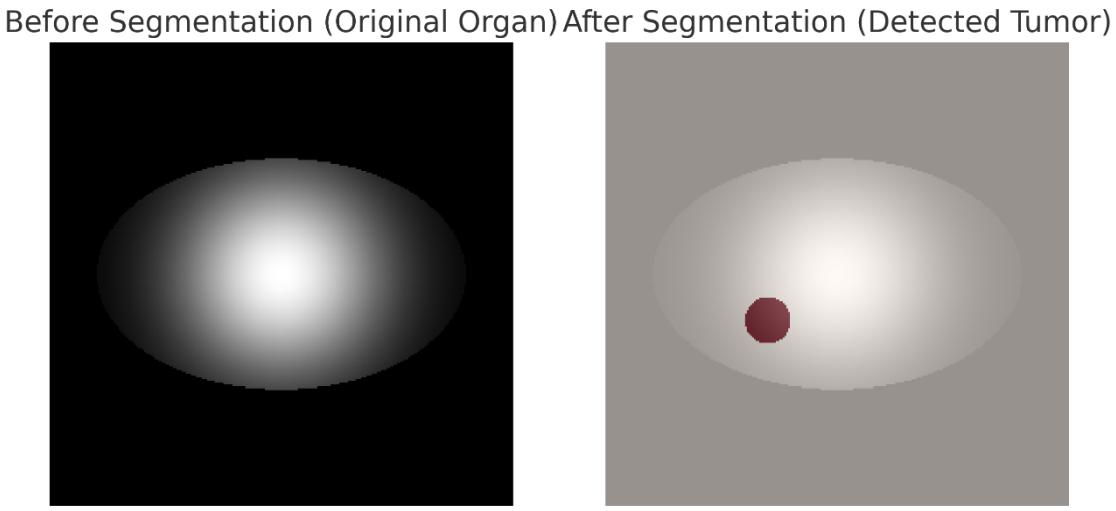

Figure 1. Segmentation Results: Before and After Using the Hybrid Model

The hybrid model demonstrates superior precision in identifying target areas compared to individual models, thereby reducing segmentation errors such as FP and FN. A false positive occurs when the model incorrectly classifies a non-target region as part of the target, while a false negative occurs when the model fails to recognize an actual target region within the image. By leveraging the combination of CNNs and Transformers, the hybrid model captures local features in greater detail while maintaining an understanding of spatial relationships within the image. This capability enables the model to more selectively define the boundaries between target and non-target areas, thereby enhancing segmentation accuracy. Moreover, the hybrid model's advantage in minimizing segmentation errors is particularly evident in images with higher levels of complexity. Further evaluations under various image conditions indicate that the hybrid model

exhibits greater stability in producing consistent segmentation results compared to individual models.

2. Inference Time Efficiency for Real-Time Segmentation

To assess the feasibility of real-time medical image segmentation, the inference time of each model was recorded and compared. Measuring inference time is crucial to understanding how efficiently a model can generate segmentation results without compromising accuracy. Low inference time is a critical factor in clinical applications, particularly in scenarios requiring rapid analysis, such as computer-assisted surgery or emergency diagnostics. Additionally, this evaluation aims to determine the balance between processing speed and segmentation accuracy, which is an essential aspect of AI-based model development. The recorded inference times for the tested models are presented in Table 6 for further analysis. The data provided in the table helps identify the most suitable model for real-time implementation.

Table 6. Comparison of Model Inference Time (in Milliseconds per Image)

Model	Inference Time (ms)
U-Net	50
ResNet	45
Transformer	60
Hybrid Model	55

Based on Table 6, the hybrid model achieves an inference time of 55 milliseconds per image, which remains within the real-time processing range. This result indicates that the model can perform segmentation at a sufficient speed for clinical applications without introducing significant delays. Although the inference time is slightly higher than that of ResNet, the hybrid model maintains computational efficiency while significantly improving segmentation accuracy. The difference in inference time can be attributed to the increased architectural complexity of the hybrid model, which integrates CNNs and Transformers to capture features with greater detail. This added complexity enhances segmentation precision, albeit with a slight increase in processing time. Further analysis of the relationship between inference time and segmentation accuracy is necessary to determine whether this trade-off is acceptable in various medical applications.

A visual comparison of DSC and IoU performance for each model is presented in Figure 2. This graph illustrates the variation in scores achieved by each model, allowing for a more in-depth analysis of segmentation effectiveness. The visualization helps identify patterns of accuracy improvement resulting from the hybrid model compared to individual models. The hybrid model demonstrates superior performance across both metrics, indicating that the combination of CNN

and Transformer architectures enhances segmentation accuracy. The differences in scores between the hybrid and individual models also reflect the hybrid model's ability to more effectively capture both local features and spatial relationships. This graphical analysis provides further insight into the reliability of each model under various medical imaging conditions.

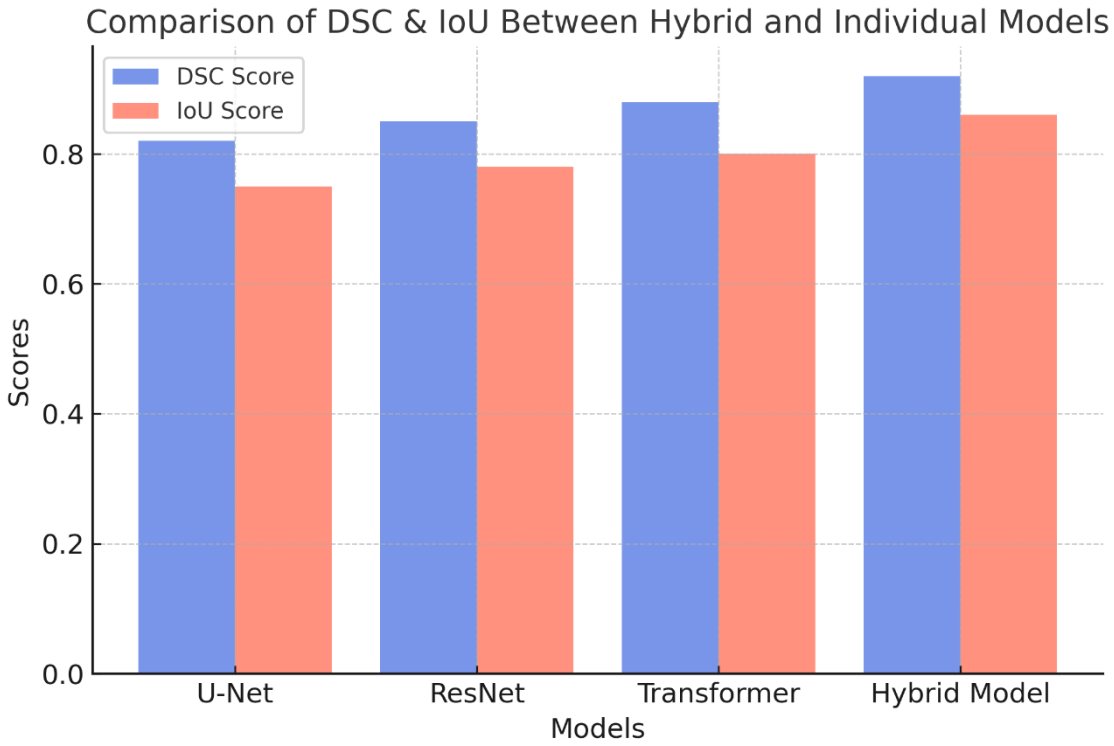

Figure 2. Graphical Comparison of DSC & IoU Between the Hybrid Model and Individual Models

Figure 2 presents a comparison of DSC and IoU scores across four models: U-Net, ResNet, Transformer, and the hybrid model. The graph shows that the hybrid model achieves the highest scores for both metrics compared to individual models, indicating improved segmentation accuracy. DSC, represented in blue, measures the similarity between the segmentation output and the reference label, whereas IoU, represented in red, evaluates the extent of overlap between the model's predicted area and the actual target area. The hybrid model outperforms the individual models in both DSC and IoU, demonstrating its ability to effectively capture both local features and spatial relationships. The Transformer model achieves higher scores than U-Net and ResNet across both metrics but still falls short of the hybrid model. This suggests that the integration of both conventional and modern approaches leads to more optimal segmentation performance. The insights provided by this graph further highlight the effectiveness of the hybrid approach in enhancing the accuracy of medical image segmentation.

V. DISCUSSION

The findings of this study indicate that a hybrid deep learning approach that combines U-Net, ResNet, and Transformer can improve the accuracy of medical image segmentation compared to individual models. The superiority of the hybrid model in preserving spatial information while capturing broader contextual features supports the research by (Xu et al., 2024), which demonstrates that the combination of CNN and Transformer enhances segmentation precision without compromising computational efficiency. Furthermore, these results reinforce the findings of (Punn & Agarwal, 2022), who stated that while U-Net excels in maintaining spatial details, it has limitations in understanding long-range spatial relationships. By integrating ResNet as a deeper feature extractor, the hybrid model in this study showed a significant increase in accuracy, consistent with the findings of (Wu et al., 2025), which revealed that residual network-based models improve segmentation robustness against variations in medical object shapes. The results also align with the study by (Rayed et al., 2024), which suggests that combining CNN and Transformer enhances the balance between segmentation accuracy and efficiency, although computational complexity remains a challenge.

The hybrid approach applied in this study demonstrates its advantage in handling the complexity of anatomical structures across various medical imaging modalities, outperforming individual models. These results are consistent with the findings of (Wang et al., 2022), who observed that CNNs have limitations in capturing global context, making their integration with Transformers a significant improvement for long-range spatial relationship modeling. Additionally, this study supports the research of (Arkin et al., 2023), which found that while Transformers excel in understanding the global context, they still require support from CNNs to maintain high spatial resolution in segmentation. The challenges of computational efficiency were also confirmed by (Eum et al., 2025), who stated that integrating multiple architectures within a single model often increases computational demands, making parameter optimization a crucial aspect in hybrid model development. These findings highlight the effectiveness of the hybrid approach in improving medical image segmentation accuracy while emphasizing the need for optimization techniques to ensure that the model can be deployed in clinical environments with computational resource constraints.

VI. CONCLUSION AND RECOMMENDATION

The results of this study demonstrate that the hybrid model, which integrates U-Net, ResNet, and Transformer, enhances the accuracy of medical image segmentation compared to individual models. This model leverages the strengths of each architecture: U-Net preserves spatial information, ResNet enhances feature extraction, and Transformer strengthens long-range

spatial relationship modeling. Evaluation results measured using the DSC and IoU indicate that the hybrid model outperforms individual models in segmentation performance. Additionally, the model exhibits sufficient efficiency for real-time segmentation, making it a promising candidate for medical applications. However, challenges related to computational efficiency remain a critical aspect for further implementation. Therefore, this study not only highlights the effectiveness of the hybrid approach in improving medical image segmentation accuracy but also underscores the importance of model optimization for efficient clinical deployment.

Future research should focus on model optimization to reduce computational burden, allowing for wider implementation in hospitals and resource-limited devices. Further experiments with larger and more diverse datasets are necessary to improve the model's generalization ability across various medical imaging modalities. Moreover, integrating the hybrid model with CAD systems could be a strategic step toward enhancing automated diagnostic accuracy. Research on parameter optimization and model compression techniques should also be pursued to enhance segmentation efficiency without sacrificing accuracy. Additionally, further studies on model adaptation to variations in medical image quality could improve segmentation stability and reliability under different clinical conditions. By addressing these aspects, the hybrid model is expected to become a more effective and efficient solution for advancing medical image segmentation accuracy.

REFERENCES

Aboussaleh, I., Riffi, J., Fazazy, K. El, Mahraz, M. A., & Tairi, H. (2023). Efficient U-Net Architecture with Multiple Encoders and Attention Mechanism Decoders for Brain Tumor Segmentation. *Diagnostics*, *13*(5), 872. https://doi.org/10.3390/diagnostics13050872

Arkin, E., Yadikar, N., Xu, X., Aysa, A., & Ubul, K. (2023). A Survey: Object Detection Methods from CNN to Transformer. *Multimedia Tools and Applications*, *82*(14), 21353–21383. https://doi.org/10.1007/s11042-022-13801-3

Athisayamani, S., Antonyswamy, R. S., Sarveshwaran, V., Almeshari, M., Alzamil, Y., & Ravi, V. (2023). Feature Extraction Using a Residual Deep Convolutional Neural Network (ResNet-152) and Optimized Feature Dimension Reduction for MRI Brain Tumor Classification. *Diagnostics*, *13*(4), 668. https://doi.org/10.3390/diagnostics13040668

Carles, M., Kuhn, D., Fechter, T., Baltas, D., Mix, M., Nestle, U., Grosu, A. L., Martí-Bonmatí, L., Radicioni, G., & Gkika, E. (2024). Development and Evaluation of Two Open-Source nnU-Net Models for Automatic Segmentation of Lung Tumors on PET and CT Images with and Without Respiratory Motion Compensation. *European Radiology*, *34*(10), 6701–6711. https://doi.org/10.1007/s00330-024-10751-2

Chen, X., Li, D., Liu, M., & Jia, J. (2023). CNN and Transformer Fusion for Remote Sensing Image Semantic Segmentation. *Remote Sensing*, *15*(18), 4455. https://doi.org/10.3390/rs15184455

Ebert, N., Stricker, D., & Wasenmüller, O. (2023). PLG-ViT: Vision Transformer with Parallel Local and Global Self-Attention. *Sensors*, *23*(7), 1–22. https://doi.org/10.3390/s23073447

Eum, I., Kim, J., Wang, S., & Kim, J. (2025). Heavy Equipment Detection on Construction Sites Using You Only Look Once (YOLO-Version 10) with Transformer Architectures. *Applied Sciences*, *15*(5), 2320. https://doi.org/10.3390/app15052320

Han, N., Zhou, L., Xie, Z., Zheng, J., & Zhang, L. (2022). Multi-Level U-Net Network for Image Super-Resolution Reconstruction. *Displays*, *73*, 102192. https://doi.org/10.1016/j.displa.2022.102192

Ji, Z., Mu, J., Liu, J., Zhang, H., Dai, C., Zhang, X., & Ganchev, I. (2024). ASD-Net: A Novel U-Net Based Asymmetric Spatial-Channel Convolution Network for Precise Kidney and Kidney Tumor Image Segmentation. *Medical and Biological Engineering and Computing*, *62*(6), 1673–1687. https://doi.org/10.1007/s11517-024-03025-y

Jiang, Y. ;, Liang, J. ;, Cheng, T. ;, Lin, X. ;, Zhang, Y. ;, Dong, J., Jiang, Y., Liang, J., Cheng, T., Lin, X., Zhang, Y., & Dong, J. (2022). MTPA_Unet: Multi-Scale Transformer-Position Attention Retinal Vessel Segmentation Network Joint Transformer and CNN. *Sensors*, *22*(12), 4592. https://doi.org/10.3390/s22124592

Maurício, J., Domingues, I., & Bernardino, J. (2023). Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Applied Sciences (Switzerland)*, *13*(9), 5521. https://doi.org/10.3390/app13095521

Melyani, M., Prasetyo, T. F., Rahadjeng, I. R., Mufid, Z., Rafik, A., Shaura, R. K., Daniel, D., & Emita, I. (2024). Design Framework of Expert System Program in Otolaryngology Disease Diagnosis use Extreme Programming (XP)Method(Case Study in THB Bekasi Hospital). *Journal of Technology Informatics and Engineering*, *3*(3), 397–416. https://doi.org/10.51903/jtie.v3i3.209

Mohapatra, R. K., Jolly, L., Lyngdoh, D. C., Mourya, G. K., Changaai Mangalote, I. A., Alam, S. I., & Dakua, S. P. (2024). A Comprehensive Survey to Study the Utilities of Image Segmentation Methods in Clinical Routine. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *13*(1), 1–26. https://doi.org/10.1007/s13721-023-00436-z

Obuchowicz, R., Strzelecki, M., & Piórkowski, A. (2024). Clinical Applications of Artificial Intelligence in Medical Imaging and Image Processing—A Review. *Cancers*, *16*(10), 1–16. https://doi.org/10.3390/cancers16101870

Pan, S., Liu, X., Xie, N., & Chong, Y. (2023). EG-TransUNet: A Transformer-Based U-Net With Enhanced and Guided Models for Biomedical Image Segmentation. *BMC Bioinformatics*, *24*(1), 1–22. https://doi.org/10.1186/s12859-023-05196-1

Priyadi, P., Migunani, M., & Sasmoko, D. (2024). Enhancing Big Data Processing Efficiency in AI-Based Healthcare Systems: A Comparative Analysis of Random Forest and Deep. *Journal of Technology Informatics and Engineering*, *3*(3), 263–278. https://doi.org/10.51903/jtie.v3i3.205

Pu, Q., Xi, Z., Yin, S., Zhao, Z., & Zhao, L. (2024). Advantages of Transformer and its Application for Medical Image Segmentation: A Survey. *BioMedical Engineering Online*, *23*(1), 1–22. https://doi.org/10.1186/s12938-024-01212-4

Punn, N. S., & Agarwal, S. (2022). Modality Specific U-Net Variants for Biomedical Image Segmentation: A Survey. In *Artificial Intelligence Review* (Vol. 55, Issue 7). Springer

Netherlands. https://doi.org/10.1007/s10462-022-10152-1

Rayed, M. E., Islam, S. M. S., Niha, S. I., Jim, J. R., Kabir, M. M., & Mridha, M. F. (2024). Deep Learning for Medical Image Segmentation: State-of-the-Art Advancements and Challenges. *Informatics in Medicine Unlocked*, *47*, 101504. https://doi.org/10.1016/j.imu.2024.101504

Shi, P., Duan, M., Yang, L., Feng, W., Ding, L., & Jiang, L. (2022). An Improved U-Net Image Segmentation Method and Its Application for Metallic Grain Size Statistics. *Materials*, *15*(13), 4417. https://doi.org/10.3390/ma15134417

Wang, H., Chen, X., Zhang, T., Xu, Z., & Li, J. (2022). CCTNet: Coupled CNN and Transformer Network for Crop Segmentation of Remote Sensing Images. *Remote Sensing*, *14*(9), 1–20. https://doi.org/10.3390/rs14091956

Wu, W., Huo, L., Yang, G., Liu, X., & Li, H. (2025). Research into the Application of ResNet in Soil: A Review. *Agriculture*, *15*(6), 661. https://doi.org/10.3390/agriculture15060661

Xiao, H., Li, L., Liu, Q., Zhu, X., & Zhang, Q. (2023). Transformers in Medical Image Segmentation: A Review. *Biomedical Signal Processing and Control*, *84*, 104791. https://doi.org/10.1016/j.bspc.2023.104791

Xu, Y., Quan, R., Xu, W., Huang, Y., Chen, X., & Liu, F. (2024). Advances in Medical Image Segmentation: A Comprehensive Review of Traditional, Deep Learning and Hybrid Approaches. *Bioengineering*, *11*(10), 1034. https://doi.org/10.3390/bioengineering11101034

Yang, F., & Wang, B. (2024). Dual Channel-Spatial Self-Attention Transformer and CNN Synergy Network for 3D Medical Image Segmentation. *Applied Soft Computing*, *167*, 112255. https://doi.org/10.1016/j.asoc.2024.112255

Yousef, R., Khan, S., Gupta, G., Siddiqui, T., Albahlal, B. M., Alajlan, S. A., & Haq, M. A. (2023). U-Net-Based Models towards Optimal MR Brain Image Segmentation. *Diagnostics*, *13*(9), 1624. https://doi.org/10.3390/diagnostics13091624

Zhang, C., Deng, X., & Ling, S. H. (2024). Next-Gen Medical Imaging: U-Net Evolution and the Rise of Transformers. *Sensors 2024, Vol. 24, Page 4668*, *24*(14), 4668. https://doi.org/10.3390/s24144668

Zhang, J., Qin, Q., Ye, Q., & Ruan, T. (2023). ST-Unet: Swin Transformer Boosted U-Net With Cross-Layer Feature Enhancement for Medical Image Segmentation. *Computers in Biology and Medicine*, *153*, 106516. https://doi.org/10.1016/j.compbiomed.2022.106516