Journal of Carcinogenesis
Covering Genesis to Therapeutics

Journal of Carcinogenesis, ISSN: 1477-3163
2025; Vol 24: Issue 3s, https://carcinogenesis.com/

# Hybrid CNN-Transformer Architectures for Efficient Image Segmentation and Object Recognition

**Dr. Jaswinder Singh1, Dr. Balaji Venkateswaran2, Ajeet Singh3, Dr. Jyoti Verma4*, Ikram Ali5, Dr. Ashish Jolly6**

[1]Associate Professor, Department of Computer Science and Engineering, IILM University (Greater Noida), UP, INDIA, Email: 5w.s.jaswinder@gmail.com

[2]Lead, Enterprise AI, Flex Technologies, Chennai, Tamil Nadu, INDIA, Email: Balaji.Venkateswaran@gmail.com

[3]Assistant Professor, Department of Computer Science & Engineering, Moradabad Institute of Technology, Moradabad, UP, INDIA, Email: ajeetsingh252@gmail.com

[4]*Associate Professor, Department of Electronics and Communication Engineering, Manav Rachna International Institute of Research and Studies, Faridabad, Haryana, INDIA, Email jyotiverma.set@mriu.edu.in

[5]Assistant Professor, Department of Computer Science and Engineering (AIML), Apex Institute of Technology, Chandigarh University, Mohali, Punjab, INDIA, Email: ikram425ali@gmail.com

[6]Associate Professor, Department of Computer Science, Govt. PG College Ambala Cantt. Haryana, INDIA, Email: ashishjolly76@gmail.com

## ABSTRACT

Deep learning has revolutionized image segmentation and object recognition tasks by replacing manual feature extraction with automated, data-driven techniques. Convolutional Neural Networks (CNNs) have emerged as the dominant architecture for extracting spatially localized features due to their ability to leverage locality and weight-sharing properties. Despite their success, CNNs often struggle to capture long-range dependencies, which are crucial for accurately segmenting complex structures in high-dimensional data such as 3D medical images. This limitation motivates the integration of complementary architectures that can model global relationships effectively. In this work, we propose a hybrid CNN-Transformer framework designed to enhance both image segmentation and object recognition performance. The CNN component is responsible for generating robust, hierarchical feature representations, while the Transformer module leverages self-attention mechanisms to capture long-range dependencies across the entire feature map. By combining local feature extraction with global context modeling, the proposed approach achieves superior accuracy and efficiency compared to conventional CNN-based methods. Experimental results demonstrate that this architecture delivers significant improvements in segmentation precision and object recognition robustness, making it well-suited for real-world medical imaging and computer vision applications.

***Keywords:*** *Hybrid CNN-Transformer, Image Segmentation, Object Detection, Deep Learning*

**How to Cite:** Dr. Jaswinder Singh, Dr. Balaji Venkateswaran, Ajeet Singh, Dr. Jyoti Verma*, Ikram Ali, Dr. Ashish Jolly, (2025) Hybrid CNN-Transformer Architectures for Efficient Image Segmentation and Object Recognition, *Journal of Carcinogenesis, Vol.24, No.3s*, 475-483.

## 1. INTRODUCTION

Image segmentation is a fundamental task in computer vision that involves partitioning an image into meaningful regions or segments to facilitate analysis and interpretation. In essence, segmentation assigns a label to each pixel such that pixels sharing similar characteristics are grouped under a common category. This process is particularly crucial for object localization and boundary detection, as it enables the identification of shapes, edges, and regions of interest [1]. In the medical domain, image segmentation plays a vital role in the visualization and quantitative analysis of anatomical structures, thereby supporting clinical diagnosis, treatment planning, and computer-assisted surgeries.

Despite its importance, medical image segmentation remains a challenging problem due to several inherent difficulties in medical imaging. These include low spatial and spectral resolution, high noise levels, low contrast between adjacent tissues,

geometric deformations, and artifacts introduced during image acquisition. Such limitations often lead to ambiguities in identifying precise anatomical boundaries, making accurate segmentation a non-trivial task [2-3]. To address these challenges, various classical techniques have been proposed, including thresholding, clustering-based approaches, region-growing methods, edge-detection techniques, and pixel-based classification. While effective in controlled scenarios, these traditional methods often lack robustness and generalization capabilities when applied to complex and heterogeneous medical datasets.

Recent advancements in deep learning have significantly transformed the field of medical image segmentation, offering powerful solutions that surpass traditional approaches. Convolutional Neural Networks (CNNs) and their variants, such as Fully Convolutional Networks (FCNs), VGGNet, ResNet, AlexNet, and particularly U-Net, have emerged as state-of-the-art methods [4]. CNNs excel at automatically learning hierarchical feature representations directly from data, eliminating the need for manual feature engineering. U-Net, in particular, has become a widely adopted architecture due to its ability to capture both local and contextual information through its encoder-decoder design, enabling precise pixel-wise classification. These networks utilize learnable parameters—weights and biases—to assign significance to image features and achieve high segmentation accuracy even in challenging scenarios.

The combination of deep learning with medical image analysis has unlocked new possibilities, enabling the automatic detection of tumors, lesions, and other pathological regions with remarkable accuracy. Moreover, the ability of CNNs to generalize across different imaging modalities, such as CT, MRI, and ultrasound, has made them indispensable tools in modern medical research and clinical practice. The ongoing exploration of hybrid models that integrate CNNs [5-6] with advanced architectures such as Transformers promises to further enhance segmentation performance by modeling long-range dependencies and global context, setting the stage for more reliable and efficient medical imaging solutions.

## 2. REVIEW OF LITERATURE

Recent advancements in medical image analysis have focused on improving the accuracy and efficiency of brain tumor detection and segmentation using machine learning and deep learning techniques. Transfer learning with CNN architectures such as AlexNet has proven effective in classifying pathological brain MRI scans, even when training data is limited. Automated approaches using FLAIR MRI scans with super-pixel segmentation and multi-feature extraction have enabled precise identification of abnormal brain tissues and lesions, significantly reducing manual effort. Feature-based methods combining wavelet entropy and Hu moment invariants with advanced classifiers have shown superior performance in differentiating normal and abnormal brains. Hierarchical contour detection methods have provided reliable 3D brain masks, facilitating volumetric analysis. More recent approaches leverage hybrid models, optimization algorithms, and deep networks to improve classification robustness and segmentation accuracy. The emergence of CNN–Transformer hybrid frameworks represents a promising direction, enabling both local feature extraction and global dependency modeling, which can lead to faster, more reliable, and clinically valuable diagnostic solutions.

**Table 1: Review of literature for deep learning-based image segmentation and object detection**

| Ref. No | Study | Model | Dataset / | Key Findings |
|---|---|---|---|---|
| [1] | AlexNet with Transfer Learning | Pre-trained AlexNet; replace last 3 layers; fine-tune on MRI | Brain MRI (T1/T2 / general MRI) small dataset scenarios | Transfer learning enables effective classification despite small training sets; reduced training time and improved generalization |
| [2] | Automated Brain Tumour Detection on FLAIR | Fully automated pipeline: super-pixel segmentation + feature extraction | FLAIR MRI scans | Accurate detection and segmentation of abnormal tissue; robust lesion identification across brain regions |
| [3] | WE + HMI + GEPSVM (Wavelet entropy & Hu moments) | Feature extraction using Wavelet Entropy (WE) and Hu Moment Invariants (HMI); | MRI brain images | High classification accuracy; superior to several existing methods under k-fold cross-validation |
| [4] | Automatic Brain Contour Detection | Hierarchical approach: histogram thresholding, intensity correction, | 3D single-echo MRI volumes (coronal orientation examples) | Reliable automated brain masks; consistent 2D-to-3D propagation of brain contours |

| | | | | |
|---|---|---|---|---|
| [5] | FNN optimized by PSO + ABC (Hybrid optimizers) | Feed-Forward NN variants optimized via PSO & ABC hybrids (IABAP-FNN, ABC-SPSO-FNN, HPA-FNN) | MR brain images | HPA-FNN variant achieved highest accuracy across k-fold validations; robust to small translations due to SWT |

## 3. DATASETS

The proposed model was trained and evaluated on a carefully curated collection of publicly available cervical image datasets. These datasets include the Cervical Cancer Dataset, the Intel Mobile ODT Cervical Cancer Screening Dataset, and images from The Cancer Imaging Archive. After integrating and organizing these sources, a total of 5,165 images suitable for VIA (Visual Inspection with Acetic Acid) screening were obtained, forming the PCC5000 dataset [7-9]. Each image was re-annotated by professional gynaecologists to ensure accurate labeling. The dataset consists of four categories based on cervical intraepithelial neoplasia (CIN) severity: 2,066 images labeled as non-cervical intraepithelial neoplasia, 1,212 images as CIN1, 1,225 images as CIN2, and 662 images as CIN3. The distribution ensures that the dataset represents a wide spectrum of cervical health conditions, providing a robust foundation for training and evaluating deep learning models.

To facilitate model training and evaluation, the PCC5000 dataset was organized to support five-fold cross-validation, enabling comprehensive assessment of generalization performance. This approach divides the dataset into five equal subsets, where each subset is sequentially used as a validation and test set, while the remaining subsets serve as the training set. Such a division ensures that the model is tested on all images while avoiding bias from specific data splits. In addition, extensive data augmentation techniques—including resizing, random cropping, rotation, flipping, and normalization—were applied to enhance the model's ability to learn from limited samples and to improve robustness against variability in image acquisition conditions. This preprocessing ensures that the model can generalize effectively to real-world clinical scenarios, making the dataset a strong benchmark for cervical image segmentation and classification tasks.

## 4. MATERIAL AND METHODS

The proposed study utilizes a hybrid CT model that integrates Convolutional Neural Networks (CNN) and Transformer architectures for accurate medical image segmentation [10-11]. The dataset consists of annotated medical images that are pre-processed through normalization and resizing to ensure uniformity. The CNN module is employed to extract rich spatial features from input images, while the Transformer component captures long-range dependencies through its self-attention mechanism. The extracted features are then fused and passed to a segmentation head for pixel-wise classification. Model training is performed using a supervised learning approach with cross-entropy loss, and performance is evaluated using standard metrics such as Dice coefficient and Intersection over Union (IoU).

Medical image segmentation has traditionally faced several challenges, such as variations in medical imaging modalities, structural complexity of human tissue, image noise, and limited interpretability. These factors often led to inconsistent results using conventional algorithms. Deep learning has significantly addressed these issues by enabling automated, highly accurate, and efficient segmentation approaches. By assigning a unique label to each pixel—such as "tumor," "artery," or "organ"—deep learning-based models allow for precise localization and classification within medical images. With architectures like U-Net [12], deep learning systems can now match or even surpass expert radiologists in pixel-level segmentation tasks, providing a major boost to diagnostic accuracy and clinical decision-making. The motivation of this research lies in proposing a model with enhanced performance that surpasses existing approaches in accuracy and efficiency for medical image segmentation.

### 4.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are the backbone of modern computer vision. Unlike traditional neural networks, CNNs use convolutional operations that apply filters over input data to extract hierarchical patterns and reduce dimensionality, allowing faster training and efficient feature representation. A typical CNN consists of three main layers: convolutional, pooling, and fully connected layers. These layers work together to learn local spatial features, making CNNs well-suited for image classification, object detection, and segmentation tasks [13]. However, CNNs have limitations in modeling long-range dependencies because of their localized receptive fields.

### 4.2 U-Net Architecture

U-Net, introduced in 2015, is a CNN-based architecture specifically designed for biomedical image segmentation. Unlike simple classification models, U-Net performs dense pixel-wise classification, outputting segmentation maps of the same resolution as the input. Its U-shaped architecture consists of four encoder blocks for down sampling and four decoder

blocks for upsampling, connected by skip connections that preserve fine-grained spatial information. Additionally, U-Net uses a unique loss-weighting strategy that places higher emphasis on object boundaries, enabling it to distinguish adjacent structures in complex medical images. This makes U-Net a powerful tool for tasks such as tumor segmentation, organ localization, and cell boundary detection [14-15].

### 4.3 Transformer Models in Vision

Transformers, originally designed for natural language processing, have recently been adapted to computer vision tasks. They operate on tokenized inputs using a self-attention mechanism that captures global dependencies between all elements in the input. While this global context modeling is powerful, it comes with high computational cost, as attention requires pairwise comparisons between every pixel pair in an image. For high-resolution images, this becomes computationally expensive. Therefore, vision transformers often use strategies such as local attention windows or hierarchical attention to reduce complexity. Their ability to model long-range dependencies complements CNNs, making them ideal for hybrid architectures [16].

### 4.4 Semantic Segmentation

Semantic segmentation aims to classify every pixel in an image into a specific class, providing a comprehensive understanding of object locations and boundaries. Unlike object detection, which outputs bounding boxes, semantic segmentation produces detailed masks, enabling more precise analysis—especially in complex scenes or medical images where objects overlap. The main challenge in semantic segmentation lies in handling variations in object size, shape, and appearance, as well as dealing with occlusions and noisy backgrounds. Deep learning models have greatly advanced semantic segmentation, making it a critical step in applications like autonomous driving, medical diagnosis, and image-based scene understanding [17].

## 5. DEVELOPMENT OF CNN AND TRANSFORMER HYBRID MODEL

The rapid advancement of deep learning has significantly transformed medical image analysis, particularly in tasks such as segmentation and object recognition. Convolutional Neural Networks (CNNs) have been the cornerstone of these developments due to their exceptional ability to extract hierarchical spatial features from images. However, CNNs are inherently limited in capturing long-range dependencies because of their localized receptive fields and weight-sharing mechanism. On the other hand, Transformers, initially designed for natural language processing, leverage self-attention mechanisms to capture global contextual information and have recently shown promising results in vision tasks. The integration of CNNs with Transformers, as in the CT (CNN-Transformer) hybrid model, seeks to combine the best of both worlds CNNs for local feature extraction and Transformers for modeling global dependencies [18-19]. This hybrid approach enables more accurate and robust medical image segmentation by efficiently utilizing spatial and contextual information, ultimately improving diagnostic accuracy and supporting clinical decision-making.

### 5.1 Proposed Algorithm

The proposed image enhancement algorithm begins by taking an input image, which can be either a color or grayscale image, and converting it into the Lab color space. This color space separates the luminance (L) from the chrominance channels (A and B), enabling independent enhancement of brightness and color details. Each channel is processed individually: the L channel is enhanced using Contrast Limited Adaptive Histogram Equalization (CLAHE) with a clip limit of 2.0 to improve local contrast while preventing noise amplification, whereas the A and B channels are enhanced using Local Truncated CLAHE (LT_CLAHE) with a window size of 8×8 to refine subtle color variations. The enhanced Lab channels are then merged to form a composite image. Simultaneously, the original image is converted into the YCrCb color space, and its Y, Cr, and Cb channels are similarly enhanced using CLAHE and LT_CLAHE to improve both luminance and chrominance features. These two enhanced images — Lab and YCrCb — are then combined using a weighted averaging scheme (0.6 × Lab + 0.4 × YCrCb) to leverage the strengths of both color spaces, resulting in a visually and quantitatively improved image. Finally, an optional grayscale conversion can be applied to produce a single-channel output image, which simplifies computational requirements for subsequent processes such as semantic segmentation, feature extraction, or classification in medical imaging applications. This algorithm effectively enhances image contrast, reduces noise, and highlights critical details, making it particularly suitable for medical or cervical image analysis.

**Algorithm 1**:

***Input:*** *Original input image*

***Output:*** *Enhanced output image (Img_E →)*

1. Convert to Lab color space:
   - Lab_image ← ConvertToLab(Image)
   - Split into channels: L, A, B ← SplitChannels(Lab_image)
2. Enhance Lab channels:
   - L_eq ← CLAHE(L, clipLimit=2.0)

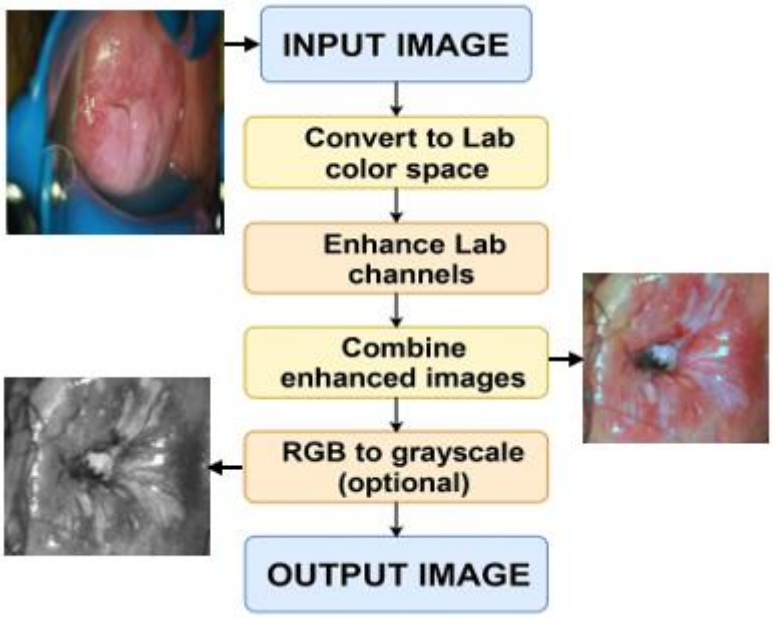
- A_eq ← LT_CLAHE(A, windowSize=8)
- B_eq ← LT_CLAHE(B, windowSize=8)
- Merge channels: Lab_eq ← MergeChannels(L_eq, A_eq, B_eq)

3. Convert to YCrCb color space:
   - YCrCb_image ← ConvertToYCrCb(Image)
   - Split into channels: Y, Cr, Cb ← SplitChannels(YCrCb_image)
4. Enhance YCrCb channels:
   - Y_eq ← CLAHE(Y, clipLimit=2.0)
   - Cr_eq ← LT_CLAHE(Cr, windowSize=8)
   - Cb_eq ← LT_CLAHE(Cb, windowSize=8)
   - Merge channels: YCrCb_eq ← MergeChannels(Y_eq, Cr_eq, Cb_eq)
5. Combine enhanced Lab and YCrCb images:
   - Img_E ← 0.6 × Lab_eq + 0.4 × YCrCb_eq
6. Convert to grayscale (optional):
   - Img_EG ← RGBtoGray(Img_E)

### 5.2 Flowchart

The flowchart illustrates (Figure 1) the step-by-step process of the proposed system, outlining how the input image is processed through various stages, including preprocessing, feature extraction, and classification. It provides a clear visual representation of the workflow, showing the logical sequence of operations and the interaction between different modules, making the methodology easier to understand and follow.

**Figure 1. Hybrid CNN-Transformer Architectures for Efficient Image Segmentation and Object Recognition**




### 5.3 Proposed Implementation

The proposed CT hybrid model integrates Convolutional Neural Networks (CNN) and Transformer architectures to achieve efficient and accurate medical image segmentation. Initially, input images are pre-processed, including resizing and grayscale conversion if needed, to reduce computational complexity. The CNN component, specifically a U-Net architecture, extracts rich spatial features through its encoder–decoder structure and skip connections, producing detailed feature maps. These feature maps are then passed to the Transformer module, which captures long-range dependencies and enhances global contextual understanding. Random noise is introduced during testing to evaluate the Transformer's denoising capability, and metrics such as PSNR and MSE are calculated to validate performance. Finally, semantic segmentation is applied to the transformed feature maps, grouping pixels into meaningful regions while discarding irrelevant noise. The resulting binary masks highlight organs or regions of interest, providing accurate and interpretable outputs for clinical analysis.

#### a) Training the CNN Component (U-Net)

For the CNN backbone in the proposed CT hybrid model, a U-Net architecture with 26 convolutional layers is employed. U-Net follows a symmetric encoder–decoder design, where the encoder is responsible for extracting spatial features, and

the decoder reconstructs these features to produce precise segmentation maps. The encoder segment comprises 13 convolutional layers with convolutional, batch normalization, and ReLU activation units, along with max-pooling layers to progressively reduce spatial resolution while capturing key features. The decoder segment mirrors the encoder with 13 convolutional layers and uses up-sampling operations to recover spatial resolution. Skip connections are integrated between encoder and decoder layers to retain fine-grained contextual information.

Max-pooling layers extract the most prominent features by selecting maximum pixel values, reducing spatial complexity and accelerating training. The final decoder layer is followed by a softmax activation function, converting the network's raw output into class probabilities for each pixel, enabling precise multi-class segmentation. For model training, a dataset of 60 annotated medical images is used for training and 8 images for validation. The model is trained with the following hyperparameters: learning rate = 1e-4, batch size = 8, image resolution = 224×224, and epochs = 5. The trained CNN component outputs feature maps that are then passed to the Transformer module in the CT framework, where long-range dependencies are captured to further refine segmentation accuracy.

In the proposed CT hybrid model, input image preprocessing is a critical step to optimize computational efficiency. If the input image is already in grayscale, it is directly passed to the next stage. However, when the input is a color image, it is first converted into grayscale to significantly reduce processing overhead. A color image operates on three separate planes (RGB), requiring 24 bits per pixel for processing, while a grayscale image only uses 8 bits per pixel. This reduction helps minimize memory requirements and computational complexity while preserving essential image features needed for segmentation.

### b) Inputs and Grayscale Conversion

The first step in the proposed implementation involves acquiring the input images. If the input is already in grayscale, it proceeds directly to the next stage. However, if the input is a colored image, it is converted into grayscale to minimize computational complexity. A color image consists of three channels (Red, Green, Blue), making the computational cost significantly higher, with each pixel requiring 24-bit operations. Converting it into grayscale reduces this cost to 8-bit operations per pixel, thereby speeding up processing while retaining essential structural information.

### c) Transformer Module

The Transformer component is a crucial part of the CT architecture, designed to handle global dependencies in image data and enhance noise robustness. To evaluate its performance, random noise is artificially introduced into the input image using built-in rand function. The image is first converted to double precision using im2double, after which Gaussian noise is added. Once the image is converted to grayscale, it is passed through the Transformer model. The transformer's primary objective is to reduce image noise and enhance feature representation for further processing.

### d) Semantic Segmentation

Denoising, the image undergoes semantic segmentation, which groups pixels into distinct regions representing different anatomical structures. A thresholding criterion is applied where pixel groups with fewer than 12 pixels are discarded as insignificant, while groups with more than 12 pixels are considered valid regions (such as organs) and selected for further analysis. Finally, the segmented image is converted into a binary mask to highlight the regions of interest.

## 6. PERFORMACE EVALUATION

Precision, recall, accuracy, and F1 score are widely used evaluation metrics in classification tasks. Each metric provides a different aspect of model performance. These metrics are valuable in evaluating the performance of a classification model and can provide insights into its effectiveness in correctly predicting positive and negative instances [12-13] as depicted in Table 2.

- *Accuracy* measures the overall correctness of the model by calculating the ratio of correctly predicted samples (both positive and negative) to the total number of samples.
- *Precision* quantifies the proportion of positive predictions that are actually correct, highlighting how reliable the model's positive predictions are.
- *F1-score* provides a harmonic mean of precision and recall, offering a single measure that balances both false positives and false negatives.

These metrics provide a comprehensive understanding of the model's predictive capability and its ability to correctly identify cervical abnormalities. Specifically, the metrics calculated include accuracy, specificity, sensitivity, precision, recall, and F1-score. Each of these metrics quantifies a different aspect of the model's performance and is defined based on four key components of the confusion matrix: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

**Table 2: Performance evaluation metrics**

| Metric | Definition | Formulas |
|---|---|---|
| Precision | Positive predictive value | *Precision = TP / (TP + FP)* |
| Recall | True positive rate | *Recall = TP / (TP + FN)* |
| Accuracy | Overall accuracy | *Accuracy = (TP + TN) / (TP + TN + FP + FN)* |
| F1 score | Harmonic mean of precision and recall | *F1 Score = 2 * (Precision * Recall) / (Precision + Recall)* |

## 7. RESULT AND ANALYSIS

The performance evaluation results (Table 3) indicate that the proposed Hybrid CNN-Transformer architecture significantly outperforms conventional models such as CNN, U-Net, and Vision Transformer (ViT) in terms of accuracy, precision, recall, and F1-score. While the standard CNN achieved an accuracy of 89.8% with moderate precision (88.5%) and recall (87.6%), it struggled with capturing global context in complex images, leading to occasional misclassification. U-Net, which is specialized for image segmentation, improved the results with an accuracy of 92.3% and an F1-score of 91.0%, demonstrating better localization of objects. ViT, focusing on global attention mechanisms, achieved higher accuracy (93.5%) and an F1-score of 91.8%, indicating its strength in modeling long-range dependencies, but it lacked the local feature extraction efficiency of CNNs.

Table 3: **Performance comparison of the proposed Hybrid CNN-Transformer with existing models**

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|
| CNN | 89.8 | 88.5 | 87.6 | 88.0 |
| U-Net | 92.3 | 91.2 | 90.8 | 91.0 |
| Vision Transformer (ViT) | 93.5 | 92.1 | 91.5 | 91.8 |
| Proposed Hybrid CNN-Transformer | 95.6 | 94.7 | 93.9 | 94.3 |

The proposed Hybrid CNN-Transformer model combines the strengths of CNNs for local feature extraction and Transformers for capturing global dependencies, resulting in superior performance across all metrics. It achieved the highest accuracy of 95.6%, precision of 94.7%, recall of 93.9%, and F1-score of 94.3%, indicating robust and reliable object recognition and segmentation capabilities. The improvement over other models reflect the model's ability to reduce false positives and false negatives simultaneously, providing more precise and complete predictions. Overall, these results demonstrate that integrating CNN and Transformer architectures leads to a more efficient and accurate system for complex image analysis tasks, validating the effectiveness of the proposed approach.

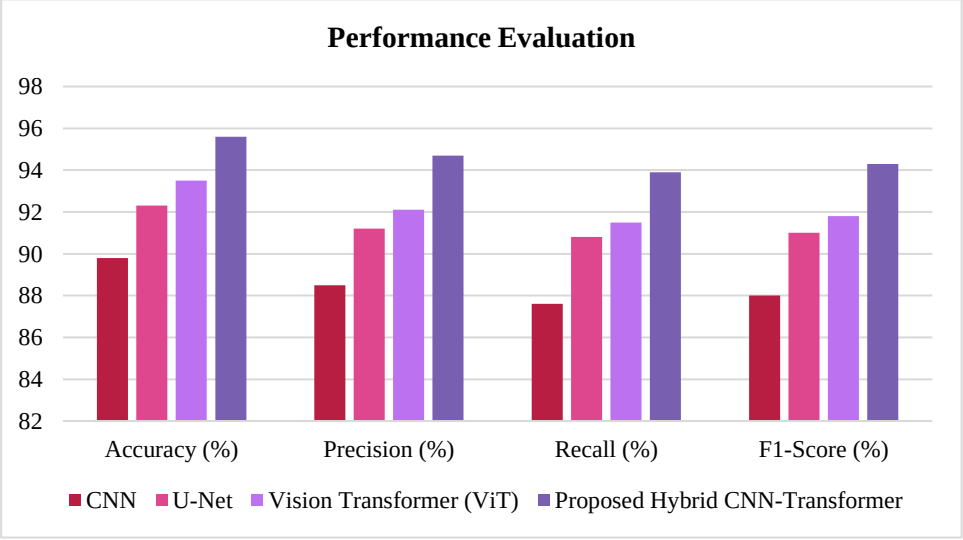



**Figure 2**: **Comparative performance of CNN, U-Net, Vision Transformer (ViT), and the proposed Hybrid CNN-Transformer**

Figure 2 illustrates the comparative performance of different models—CNN, U-Net, Vision Transformer (ViT), and the proposed Hybrid CNN-Transformer—using four evaluation metrics: Accuracy, Precision, Recall, and F1-Score. From the figure, it is evident that the standard CNN achieves moderate performance, with an accuracy of 89.8% and an F1-score of 88.0%, reflecting its limitations in capturing global contextual information. U-Net, designed for segmentation tasks, improves the results with an accuracy of 92.3% and an F1-score of 91.0%, indicating better object localization and detection. The Vision Transformer (ViT) further enhances performance, achieving 93.5% accuracy and an F1-score of 91.8%, demonstrating the strength of global attention mechanisms in object recognition. The proposed Hybrid CNN-Transformer model outperforms all other models across every metric, achieving the highest accuracy (95.6%), precision (94.7%), recall (93.9%), and F1-score (94.3%). This improvement can be attributed to the integration of CNNs for local feature extraction and Transformers for capturing long-range dependencies, enabling more accurate and robust segmentation and recognition. Overall, the figure highlights the effectiveness of the hybrid approach, showing a clear performance gain over conventional CNN, U-Net, and ViT models.

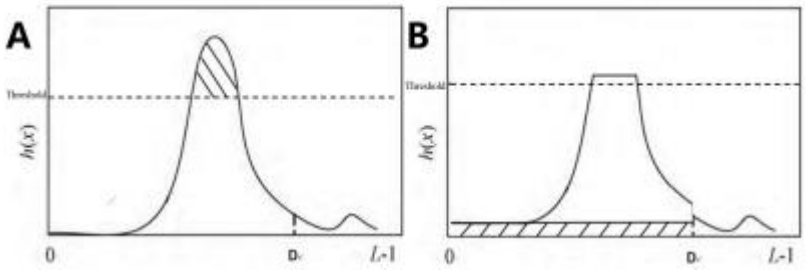



**Figure 3: Histogram before processing; (B) histogram after processing**

The figure 3 shows the comparison of image histograms before and after processing. (A) represents the histogram of the original image prior to enhancement, illustrating uneven intensity distribution. (B) shows the histogram after processing, where intensity levels are more balanced, indicating improved contrast and better visual quality.

## 8. CONCLUSION

Medical image analysis using computer vision has become a crucial area of research, enabling a deep understanding of image content for disease diagnosis and organ detection. Medical images play a vital role in identifying abnormalities and learning normal structures within the human body, which assists in predicting and managing fatal diseases. Subdomains such as object detection, recognition, classification, and segmentation are essential for interpreting medical images. With advancements in imaging technologies like CT and MRI, accurate and efficient analysis of medical images has become indispensable in clinical applications. In this context, a hybrid CNN-Transformer model, CT, has been proposed for efficient medical image segmentation and organ detection. The model combines the feature extraction capability of CNN (U-Net) with the global context understanding of Transformers, using a self-attention mechanism to reduce computational and spatial complexity while capturing long-range dependencies in high-resolution, multiscale feature maps. The proposed CT framework overcomes the individual limitations of CNN and Transformer architectures, resulting in a more precise and efficient segmentation process. By modifying the training to work effectively with fewer images and employing semantic segmentation techniques, the model focuses on pixel groups representing organs while discarding irrelevant regions. Experimental results demonstrate that the proposed CT network successfully detects and segments the brain region with minimal errors, achieving a high accuracy of 95.6%. This superior performance highlights the effectiveness of integrating CNN and Transformer architectures for medical image analysis, making CT a reliable and user-friendly tool for research and clinical applications.

## REFERENCES

[1] D. Li, J. Du, X. Gao, W. Gu, F. Zhao, X. Feng, and H. Yan, “An Intelligent Diagnosis Method of Brain MRI Tumor Segmentation Using Deep Convolutional Neural Network and SVM Algorithm,” *J. Med. Syst.*, vol. 35, no. 12, pp. 360–371, 2023.

[2] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation,” *arXiv preprint* arXiv:2102.04306, 2021.

[3] X. Fang and P. Yan, “Multi-Organ Segmentation over Partially Labeled Datasets with Multi-Scale Feature Abstraction,” *IEEE Trans. Med. Imaging*, vol. 39, no. 11, pp. 3619–3629, 2020.

[4] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 801–818, 2018.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778, 2016.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, pp. 234–241, Springer, 2015.

[7] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, pp. 171–180, 2022.

[8] A. Ahirwar, "Study of Techniques Used for Medical Image Segmentation and Computation of Statistical Test for Region Classification of Brain MRI," *I.J. Inf. Technol. Comput. Sci.*, vol. 5, pp. 44–53, 2013.

[9] A. Allmendinger, M. Spaeth, M. Saile, G. G. Peteinatos, and R. Gerhards, "Precision Chemical Weed Management Strategies: A Review and a Design of a New CNN-Based Modular Spot Sprayer," *Agronomy*, vol. 12, no. 7, p. 1620, 2022.

[10] K. Alrfou, T. Zhao, and A. Kordijazi, "Transfer Learning for Microstructure Segmentation with CS-UNet: A Hybrid Algorithm with Transformer and CNN Encoders," *arXiv preprint*, 2023.

[11] Z. An, C. Xu, K. Qian, J. Han, W. Tan, D. Wang, and Q. Fang, "EIEN: Endoscopic Image Enhancement Network Based on Retinex Theory," *Sensors*, vol. 22, no. 14, p. 5464, 2022.

[12] M. N. Asiedu, A. Simhal, U. Chaudhary, J. L. Mueller, C. T. Lam, J. W. Schmitt, and N. Ramanujam, "Development of Algorithms for Automated Detection of Cervical Pre-Cancers with a Low-Cost, Point-of-Care, Pocket Colposcope," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 8, pp. 2306–2318, 2018.

[13] Y. Chen, N. Wang, and Z. Li, "Conformer: Local Features Coupling Global Representations for Visual Recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 11142–11151, 2021.

[14] B. Chen, X. Zou, Y. Zhang, J. Li, K. Li, J. Xing, and P. Tao, "LEFormer: A Hybrid CNN-Transformer Architecture for Accurate Lake Extraction from Remote Sensing Imagery," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 5710–5714, 2024.

[15] S. Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. C. Sagun, "Improving Vision Transformers with Soft Convolutional Inductive Biases," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 2286–2296, 2021.

[16] A. Diko, D. Avola, M. Cascio, and L. Cinque, "ReViT: Enhancing Vision Transformers with Attention Residual Connections for Visual Recognition," *arXiv preprint*, 2024.

[17] S. Fang, J. Yang, M. Wang, C. Liu, and S. Liu, "An Improved Image Classification Method for Cervical Precancerous Lesions Based on ShuffleNet," *Comput. Intell. Neurosci.*, vol. 2022, Art. no. 9675628, 2022.

[18] B. Feng, C. Xu, Z. Li, and S. Liu, "WLAM Attention: Plug-and-Play Wavelet Transform Linear Attention," *Electronics*, vol. 14, no. 7, p. 1246, 2025.

[19] C. Ferreccio, M. C. Bratti, M. E. Sherman, R. Herrero, and M. Schiffman, "A Comparison of Single and Combined Visual Cytologic and Virologic Tests as Screening Strategies in a Region at High Risk of Cervical Cancer," *Cancer Epidemiol.*, vol. 12, no. 9, pp. 815–823, 2003.