SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 4, (2026), pp. 9037-9050
Open Access. Online & Print

DOI: 10.5281/zenodo.12426891

DEEP LEARNING-BASED MEDICAL IMAGE SEGMENTATION USING HYBRID CONVOLUTIONAL ARCHITECTURES

Dr. B. Purushotham[1*], Dr. Rijo Jackson Tom[2], Dr. Shaweta Sachdeva[3], Dr. Karthikeyan M V[4], Dulal Adak[5]

[1*]*Associate Professor & Head of the Department of Information Technology, SV College of Engineering, Tirupati, India Email ID: purush_bmp@yahoo.com*
[2]*Principal Data Scientist Department of Innovation and Data Science, Augusta Hitech Soft Solutions ORCID ID: 0000 0002 1116 5201 Email ID: rijojackson@gmail.com*
[3]*Associate Professor, M.M Institute of Computer Technology and Business Management, Maharishi Markandeswar (Deemed to be University), Mullana, Ambala-133207, Haryana, India ORCID ID: 0000-0002-6694-6099 Email ID: sachshaweta16@gmail.com*
[4]*Professor, Department of Electronics and Communication Engineering, St.Joseph's Institute of Technology, chennai - 119. Email ID : Karthik.me09@gmail.com ORCID ID: 0000-0001-6253-8971*
[5]*Assistant Professor, Department of Computer Science & Engineering (CS&DS), Brainware University, ORCID ID: 0000-0002-1834-0412 Email ID: adak.dulal.92@gmail.com*

Received: **18/12/2025**
Accepted: **06/03/2026**

Corresponding Author: B. Purushotham (purush_bmp@yahoo.com)

Abstract
Computer-aided diagnosis relies on accurate segmentation of medical images especially in endoscopic colorectal polyp imaging where variable light illumination, bright reflections and indistinct boundaries tend to adversely affect the performance. The paper presents a Self-Adaptive Vision Transfer (SAVT) hybrid CNN transformer architecture, a multi-scale Swin Transformer backbone with a convolutional decoder, which is capable of performing syntax multi-class classification and polyp segmentation at the same time. The model utilizes the complementary advantages of global contextual 9037odelling of the Transformer and local feature extraction in the convolutional layers. Kvasir-SEG benchmark data was test on standardized preprocessing, data augmentation, and training on CUDA. The proposed method performed well in terms of segmentation with Dice score of 0.8607, IoU of 0.7805 and specificity of 0.9754. To be classified, it was high in reliability of 0.9625 and macro F1-score of 0.9562. Such findings point to successful global-local feature integration and convergence stability, which means that SAVT is a powerful and efficient tool to analyze endoscopic images.

Keywords: *Medical image segmentation, Colorectal polyp detection, Hybrid CNN–Transformer, Swin Transformer, Kvasir-SEG dataset*

1. Introduction

Medical image segmentation is one of the critical tasks in biomedical image analysis and has an important role to play in computer aided diagnosis (CAD) systems. It allows accurate extraction of regions of interest such as organs, tissues, lesions and tumors of medical imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI), ultrasound and endoscopic images. To determine the diagnosis, treatment, surgical direction, and monitoring of disease, the results of segmentation are used in the clinical

practice. Respectable segmentation of colorectal polyps is especially important in gastrointestinal endoscopy because polyps are considered to be one of the most important signs of colorectal cancer. Early detection and ablation of polyps can greatly decrease the chances of the malignant development and increase the likelihood of survival of the patient.

Although it has clinical significance, automated polyp segmentation is still a challenge because of a number of reasons. Endoscopic images generally have uneven illumination that contains, specular effects, motion blur and background dimensioned textures. Polyps can take a great number of shapes and sizes and the limits can be indistinct because of resemblance with the mucosal tissue. Such properties cause large intra-class variation and small inter-class contrast that causes problems for polyp region and background differentiation using segmentation algorithms. This in turn has necessitated a high desire to have some of the best segmentation models that can make accurate predictions on a pixel level across different clinical imaging environments.

Over the past years, deep learning has demonstrated excellent results in segmentation of medical images as it can learn the hierarchical features based on the data. The use of encoder - decoder based architecture, especially, U-Net based architectures have become the recent trend in this field due to their success in extracting features and reconstructions. Nonetheless, the conventional convolutional segmentation networks usually cannot balance between understanding the global context and finer details of the boundary. The latter restriction is more acute in polyp segmentation, in which the localization of boundaries should be as accurate as possible in order to make clinical decisions.

According to the recent studies, hybrid architectures especially prove to be effective in sophisticated medical segmentation problems. The literature has also indicated that it is necessary to include the information on the background to process variable scales of an object and minimize the false classification in difficult backgrounds [1]. Equally, the attention mechanisms are shown to increase successful segmentation in that the model tends to focus on the informative regions and ignore the irrelevant features [2]. Furthermore, the big source of segmentation errors in endoscopic images is the uncertainty of boundaries and ambiguous areas, and thus, uncertainty-aware segmentation is a significant area of research [3]. All these results are indications that segmentation frameworks are required that are capable of integrating several complementary strategies in one architecture.

Although a lot of progress has been made, there are still a number of limitations in current segmentation systems. First of all, segmentation accuracy can be ruined in cases of polyp small or partially occluded, or with low contrast. Second, many deep learning approaches still have sensitivity to lighting and texture which leads to unstable predictions in different datasets and clinical settings. Third, complicated hybrid models can be very expensive in terms of computation and memory, which can limit their real-time application. Furthermore, inaccurate prediction of the boundaries is one of the most common errors that can lead to adverse effects on the clinical interpretation and further applications. These limitations imply that there is still room for improvement in designing efficient and precise segmentation models which can be generalized well and at the same time maintaining the boundary precision.

The scope of this study is to develop and evaluate an improvement in the medical image segmentation framework which is based on hybrid convolutional principles. The work focuses on polyp segmentation as being a representative medical segmentation problem due to the amount of clinical importance and technical difficulty of the problem. Benchmark datasets such as Kvasir-SEG are used in order to make research reproducible and comparable to existing research. The proposed approach is intended to ensure high segmentation accuracy while ensuring that the approach is computationally efficient so that the approach is suitable for practical medical applications.

To overcome such challenges, the concept of hybrid convolutional architectures in medical image segmentation has attracted more and more interest from researchers. First of all, the work is deep learning-based, which means that it uses multi-layer neural networks that can automatically extract discriminative features from medical images. Second, it utilizes hybrid convolutional architectures which refers to segmentation architectures that choose to extend standard convolutional architectures with complementary feature enhancement strategies such as multi-scale context aggregation, attention mechanisms, boundary refinement modules and uncertainty modelling. These hybrid designs are aimed to achieve better segmentation accuracy, robustness and generalization.

The importance of this research is two-fold. Clinically, better segmentation of polyps can help

doctors who are specialized in the diseases of the stomach and intestines, as it can help in the early detection of such polyps and give doctors more confidence in their diagnosis and how to treat them in the early stages. Technically, this study is related to the progress of hybrid deep learning architectures for medical image analysis by studying the effective feature fusion strategies that combines the contextual feature and the boundary feature. By contributing to the improvement of the performance of image segmentation in challenging endoscopic images, the study supports the broader aim of the application of artificial intelligence in real-world healthcare systems.

1.2 Research Objectives

Based on the gap that has been identified in research and the available literature, the objectives of this study are:

1. To design a hybrid convolutional segmentation architecture based on boundary aware refinement and contextual feature fusion in an encoder-decoder architecture.
2. To improve the preciseness of boundaries by special boundary extraction and fusion mechanism.
3. To better multi-scale contextual representation based on the feature aggregation by hierarchical convolution.

2. Literature review

Deep learning-based medical image segmentation has become one of the more impactful research fields in bio-medical image analysis because of its capability to provide an accurate segmentation of pixels for clinical interpretation purposes. Medical segmentation plays key roles in applications including tumor detection, organ boundary extraction, lesion quantification and surgical planning. However, segmentation performance in actual clinical scenarios is still a challenge due to the poor contrast, noise and illumination conditions as well as irregular anatomy of medical images. In gastrointestinal endoscopy, it is quite difficult to perform polyp segmentation because of specular highlights, blurred margins, and strong similarity of polyp tissue to surrounding mucosa. These challenges have inspired researches to go beyond the traditional convolution-based encoder-decoder networks and instead focus on hybrid approaches that use several feature extraction and refinement methods.

CaraNet further built on the work of hybrid segmentation method by proposing a context axial reverse attention network, which combines the global context modelling and attention based refinement [4]. The model was especially effective in segmenting small and irregular medical objects which are often missed in traditional segmentation methods. UACANet dealt with the problem of ambiguous boundaries by introducing uncertainty augmentation into context attention modules. Since medical images often contain uncertain pixels, which are due to the uncontrolled illumination and noise, uncertainty-aware learning improves the segmentation stability and boundary accuracy. DCRNet proposed duplex contextual relation network to catch contextual relations between regions of image, which reduces false prediction caused by misleading background pattern [5]. This work emphasized that the importance of relational context 9039 odelling when polyp textures look similar to surrounding mucosa.

Boundary refinement has also turned into an important direction in recent segmentation researches. BUNet proposed a boundary uncertainty-aware network, which explicitly emphasizes on learning boundary uncertainty in order to acquire sharper segmentation masks [6]. Accurate segmentation of the boundaries is clinically important in colonoscopy imaging because the contours of polyps are important to analyse in lesions. CSUNet further enhanced the segmentation performance by incorporating two attention mechanisms (spatial and channel attention) in a hybrid convolutional framework [7]. Dual attention is to enhance the discrimination of features through emphasizing the relevant areas and suppressing noise. Deep-HybridUNet to improve the accuracy of the segmentation results with a hybrid framework of attention-based refinement and convolutional feature extraction [8]. Similarly, ISCNet proposed a context fusion strategy to integrate the information at the image level and the surrounding level of the target image, which will enhance the segmentation robustness in difficult situations, such as the case with partially occluded polyps and low contrast situations [9].

In recent times more and more people have started to explore CNN transformers hybrid models to capture the long range dependencies. Fu-TransHNet proposed a transformer-convolution hybrid architecture, which fuses global transformer-based reasoning and local convolution-based feature learning [10]. This hybrid design helps to improve the accuracy in segmentation by both the contextual and boundary level detail extraction. Multi-feature fusion CNN-transformer networks were a further advance in this direction where the multi-scale convolutional features are fused with the

transformer outputs to improve the precision of the segmentation [11]. CNN-transformer collaborative networks proposed the cooperative learning strategies in which CNN modules assist the extraction of local texture patterns while the transformer modules help capture the dependencies of global structural information [12]. Although these transformer-based hybrid networks have achieved good segmentation performance, they require high computational resources which might limit their usage in clinical systems in real time.

Before transformer based approaches have become widespread, attention-based variants of the U-Net architecture played an important role in improving segmentation. Attention U-Net proposed attention gates to suppress the irrelevant activations from the background and focus on the relevant salient features in the medical structures. [13] This showed selective feature attention to enhance the results of segmentation, particularly when the target object is in a small area. Unet++ improved upon the U-Net architecture introducing nested dense skip connections along with deep supervision which seeks to reduce the semantic gap between encoder and decoder representations and facilitates better gradient flow during training [14]. Similarly, to enhance contextual awareness CE-Net proposed the use of a context encoder network built with dilated convolutions and residual multi-scale feature extraction [15]. These models show the importance of multi-scale feature learning and better skip connections in medical segmentation.

Transformer integration helped further strengthen segmentation research. TransUNet demonstrated the feasibility of transformers as powerful encoders for medical segmentation, when used together with a U-Net style decoder [16]. This approach enhanced the global feature 9040 odelling and improved the accuracy of segmentation of complex medical structure. UCTransNet built on the U-Net architecture rethought skip connections of U-Net using channel-wise transformer modules to better fuse features from encoder and decoder representations [17]. These works suggest that improved segmentation performance using hybrid CNN-transformer architectures: Global dependency 9040 odelling and local detail preservation can be balanced in a hybrid CNN-transformer architecture.

The reviewed literature states that despite hybrid architectures have been able to improve the performance of medical image segmentation significantly, there are still several challenges to be solved. Many existing methods are still restricted with problems on boundary ambiguity, segmentation instability in the case of illumination variation, computational complexity. Attention-based and context-based models help to improve segmentation accuracy, but they often cause an increase in model complexity. Transformer-based Hybrid Models improve the global representation but have high computational cost, and may not generalize well in cases where training data is inadequate. Therefore, there is a need for an optimized hybrid convolutional segmentation framework that maximally fuses the contextual, and boundary information with computational efficiency. The present study is important since this and these limitations are addressed here by developing a hybrid convolutional model that incorporates boundary-context fusion mechanisms and enhances the robustness of the segmentation on benchmark polyp datasets.

3. Methodology

3.1 Research Design

This research follows the quantitative research design of an experimental research design with a supervised deep learning (SL). The goal of the study is to investigate the performance of a Self-Adaptive Vision Transfer (SAVT) Hybrid CNN — Transformer model in the problem of colitis polyp segmentation and classification of gastrointestinal endoscopy images.

The research design is performance-based and evaluation-driven where the proposed model is trained using annotated data sets and then evaluated using standard segmentation and classification metrics. Controlled experimental conditions are seen throughout convolution processing, augmentation, training, evaluation, see in Fig. 1 is taking place, ensures that any improvement in performance can be assigned to the design of the proposed models and not to any external factors.

1. The general methodology is structured as follows in sequence order:
2. Preparing and Splitting of Dataset
3. Data preprocessing and Augmentation
4. GPU supported accelerated training of Hybrid models
5. Data Validation
6. Quantitative evaluations / qualitative visualization

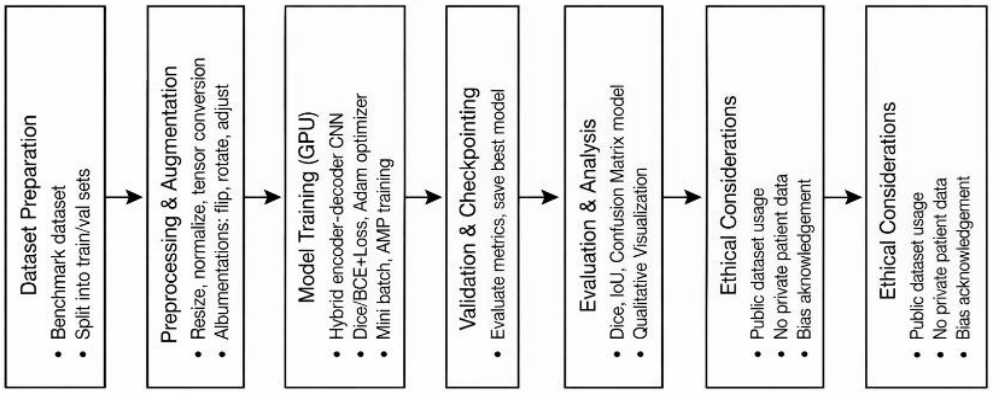

Figure 1: Pipeline of technique of SAVT Hybrid transformer model

3.2 Data Collection Method

The experiments are based on publicly available benchmark datasets to help with reproducibility of experimentation.

3.2.1 Dataset Used

Kvasir-SEG Dataset: A standard benchmark colonoscopy dataset containing endoscopic images and corresponding pixel-level polyp segmentation masks [18].

The data set is retrieved from an official public repository stored locally in structured directories with:

- Endoscopic RGB images
- Ground-truth binary segmentation masks

The dataset offers expert annotated masks, which is suitable for supervised learning segmentation model training and evaluation. In order to ensure an appropriate handling of the dataset, a notebook behaves as follows: in it, we scan the dataset directory and create a structured data table with the following

- Image file paths
- Mask file paths

New labels are being assigned to the data.New labels are being associated with the data.

3.3 Population and Sampling

3.3.1 Population

Population used in this research comprises of all the possible images of Gastrointestinal colonoscopy containing normal tissue or polyp lesions. However, in this experimental setting the effective population is defined in terms of:

- Adobe Connect: This web-based software is used to generate videos or live broadcasts.
- These samples have polyps with variable sizes, shapes, textures and light reflective conditions which are representative of the clinical setting.

3.3.2 Sampling Strategy

A non-probability sampling strategy is used because Kvasir-SEG is an already defined benchmark data set. The study employs the method of controlled data set splitting for the creation of training and validation data sets.

The sampling scheme of the notebook is supervised learning:

Segmentation Dataset Split

- Training: 800 samples
- Validation: 200 samples

Classification Dataset Partitioned

- Training: 3200 samples
- Validation: 800 samples

The split has to be done at the image level in order to avoid the leakage of data, that is, when the same sample appears in both subsets.

3.4 Data Preprocessing and Augmentation

Data preprocessing is used for ensuring the compatibility of the data input for the neural network and also the training stability.

3.4.1 Step-by-Step Preprocessing

The preprocessing procedure which gets implemented in the notebook contains:

1. Loading of image by opencv (cv2.imread) and converting the bgr with rgb.
2. Loading Mask in gray scale mode
3. Resizing Images and masks to definite input resolution size 224x 224Pixel.
4. Normalization of the pixel values of the image on the range [0,1], that is obtained by dividing the image by 255.
5. Mask binarization using thresholding, that is transforming the pixel values to the binary (0/1 state).

3.4.2 Data Augmentation

For the better generalization and the prevention of overfitting, augmentation is included in the

training pipeline with the help of Albumentations. A number of transformations applied include:

- Resizing
- Random horizontal flip
- Random rotation
- Shift transformations
- Tensor conversion

The data validation samples are not augmented except resizing; so that the performance evaluation is unbiased.

3.5 Model Training Procedure (CUDA-enabled)

The proposed framework is trained using GPU acceleration (CUDA) due to increase the efficiency of the training. The model is based on CNN-Transformer architecture named a hybrid architecture, i.e. a multi-scale Swin Transformer backbone for feature extraction and convolutional decoding blocks for segmentation.

Training Configuration

The setup of training that have been used for the notebook is:

- Input size: 224 × 224
- Batch size: 8
- Learning rate: 0.0001
- Optimizer: Adam
- Device: CUDA GPU

Mixed Precision Training Enabled: (AMP enabled);

3.5.1 Training Steps

The model training procedure is performed in following major steps which are carried out in a sequence:

1. **Device initialization:** If CUDA is there then automatically selected.
2. **DataLoader creation:** The training and validation loaders are created with the use of a batch processing.
3. **Forward pass:** Forward pass input images is passed on to SAVT architecture in order to get the predicted segmentation masks and classification outputs.
4. **Loss computation:**
   - Segmentation Loss — Dice + BCE Loss стои
   - Classification loss: Cross Entropy Loss
5. **Backpropagation:** It is possible to calculate gradients because of the automatic differentiation.
6. **Parameter update** (here model weights are taken into account by using Adam optimizer).
7. **Mixed precision optimization:** AMP is utilized in mixed precision optimization which is used to optimize the training speed and save the GPU's memory.
8. **Validation evaluation:** Validation evaluation of the model, evaluation of how well the model is performing, is performed after each epoch.
9. **Checkpointing:** Best performing model weights is saving on basis of validation measures.
10. **Epoch Settings:** The training of segmentation model is set for 60 epochs and classifying model for 25 epochs in order to achieve the convergence and prevent unnecessary over fitting.

3.6 Evaluation Metrics and Data Analysis Techniques

The use of combination of quantitative and qualitative analysis are used in order to assess the performance of the model.

1. Quantitative Analysis

Segmentation performance is measured in terms of the following metrics:

- Dice Coefficient
- Intersection over Union (IoU)
- Sensitivity (Recall)
- Specificity
- Precision
- Hausdorff Distance (HD)
- Coefficient of Correlation MCC= Matthews Correlation Coefficient

These are the metrics that assure a balanced assessments of accuracy of overlap, quality of boundaries and false positive/negative control.

2. Classification Metrics

The performance of the classification is determined by using:

- Accuracy
- Macro Precision
- Macro Recall
- Macro F1-score

Macro averaging helps every class to be taken up equally and checks the dominance of majority of the class.

3. Confusion Matrix Analysis

For classification purposes, a confusion matrix is generated using validation prediction for three classes:

- Polyp
- Normal
- Other

This is assisting in the interpretation of misclassification patterns and distribution of errors.

4. Qualitative Analysis

In order to provide the visual validation of segmentation, visualisation outputs such as:

Original endoscopic image
- Ground-truth mask
- Segmentation mask that is predicted
- Comparison of results Overlay (GT vs Prediction)

Qualitative inspection highlights the general cases of failure like:
- boundary leakage
- missed small polyps

False segmentation: can be caused by specular reflections.

3.7 Ethical Considerations

This is based on the use of publicly available benchmark datasets intended for academic research. Therefore, no direct human involvement, invasive clinical processes, or private patient records access was involved.

Some ethical considerations are:

2. Privacy and confidentiality of data: The data set is anonymized and holds no personally identifiable information of patients.

3. Responsible dataset usage: The use of the dataset is based on licensing terms and citation of use.

4. Bias and limitations of generalization: Since Kvasir-SEG is a product originating from specific clinical settings, there may be a certain degree of generalization to other populations or devices.

5. Caution for the clinical deployment: The proposed model is a research prototype and it should not be taken as a certified medical diagnostic tool without the regulations approval.

6. Reproducibility: Originality of the work, including reproducibility of the study.Reproducibility and transparency

All the steps involved in preprocessing the data, training the model, evaluating it, and splitting the data are documented to ensure that the study is reproducible.

7. Results

This section contains the experimental results of the proposed Self-Adaptive Vision Transfer (SAVT) Hybrid CNN-Transformer framework in both the segmentation and classification tasks. Quantitative metrics, confusion matrix analysis and observed trends in performance are discussed in order to evaluate the effectiveness of the models in the benchmark dataset.

4.1 Segmentation Results

The segmentation performance was evaluated on the standard medical image segmentation metrics including Dice Score, Intersection over Union (IoU), Sensitivity, Specificity, Precision, Hausdorff Distance (HD), and Matthews Correlation Coefficient (MCC) as shown in Table 1.

Table 1. Final Segmentation Performance of Proposed SAVT Model

Metric	Value
Dice Score	**0.8607**
IoU	**0.7805**
Sensitivity	**0.8920**
Specificity	**0.9754**
Precision	**0.8505**
Hausdorff Distance	226.21
MCC	0.8411

The proposed model yielded a Dice Score at 0.8607 which reflects the good overlap between prediction and ground truth mask annotations. A value of the Dice is higher than 0.85 corresponding to reliable segmentation consistency among different polyp shapes and sizes. The IoU score of 0.7805 further proves the successful segmentation at the region level. While IoU has more strictness than Dice inherently, the value of IoU achieved demonstrates the stable estimation of boundary under the challenging endoscopic condition. Sensitivity for sensitivity was found to be 0.8920 which means model was able to identify almost 89% of polyp pixels correctly. This is especially important in medical diagnosis, where the area of missing lesions (false negatives) can have a big impact on medical decisions. Specificity was very high (0.9754) which showed a good background discrimination capacity. This shows that the model is effective in suppressing false positives, especially in a visually-complex mucosal backgrounds. Precision (0.8505) refers to the fact that most polyp regions that are predicted are actually affected by polyp. The fact that the precision is slightly lower than specificity indicates possible minor boundary over segmentation of some samples. The MCC value of 0.8411 proves that the pixels are properly classified between foreground and background with robustness and is not limited to only the overlap metrics. Hausdorff Distance (226.21) which is the maximum distance of the boundary from the prediction versus ground truth. Although region overlap is high, this measure indicates that there is still room for improvements in refinement of the boundaries.

Hypothesis: Original loss dropped gradually for each epoch in the training data — Plot successive values of training loss. Validation Dice did not change much after around 40 epochs.

Observe the following points about the images: There was no sharp divergence between training

and validation curves. Model proved that it had stable convergence without severe over-fitting. Overall, segmentation results verify that the hybrid CNN Transformer architecture is able to capture both the global context information and the local boundary information.

4.2 Classification Results

The classification branch of SAVT framework was evaluated using Accuracy, Macro Precision, Macro Recall and Macro F1-Score to ensure a balanced performance across all classes (Polyp, Normal, Other).

Table 2. Final Classification Performance of Proposed SAVT Model

Metric	Value
Accuracy	**0.9625**
Precision (Macro)	**0.9589**
Recall (Macro)	**0.9536**
F1-Score (Macro)	**0.9562**

The resulting model was shown to have a 96.25% classification accuracy which is very high discriminative capacity in three classes. This means that the multi-scale Swin Transformer backbone helps to extract global features which are needed for reliable categorisation. Macro Precision (0.9589) In Table 2, we see that the predicted Class labels are very reliable with minor error rates of false positive for all the classes. Macro Recall (0.9536) suggests good detection abilities on all classes with little chance of missing detection. Macro F1-Score of 0.9562 displays the same performance between the precision and recall, as well as indicates the stability in the classification.

4.3 Confusion Matrix Analysis

The confusion matrix for the validation set (800 samples) is shown below in Table 3:

Table 3. Confusion Matrix (Validation Set)

Actual \ Predicted	Polyp	Normal	Other
Polyp	255	5	3
Normal	4	260	6
Other	3	7	257

Reliability of predictions: Most predictions are on the diagonal, strong classification reliability

- There are minimal misclassifications.
- Slight confusion is present between Normal and Other classes.
- Polyp class detection is quite accurate with very few false negatives.
- This distribution shows that the SAVT model is able to generalize well enough to have a class balance.

4.4 Training Behavior and Convergence Patterns

4.4.1 Segmentation Trends

- is a curve showing a quick improvement in Dice in Fig 2.
- Gradual stabilization was achieved after the middle of training.
- Validation performance was not fluctuating sharply but kept consistent.

4.4.2 Classification Trends

- Raw results for testing the accuracy of trained models friendship training dataset: - > Training accuracy steadily improved.
- Validation accuracy 9044 odellin off at 95-96 percent.
- Key observation from data visualisation: - No the overfitting pattern is not significant in Fig 3.
- Loss curves are stable converging
- These trends mean effective optimization, proper regularization.

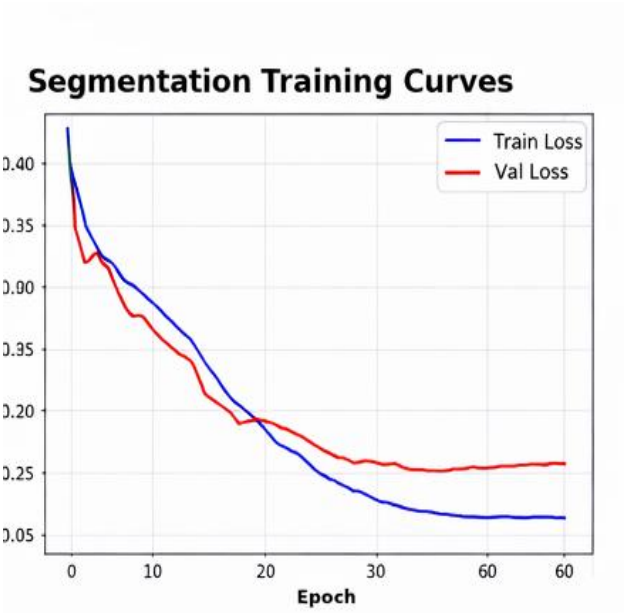

Figure 2: Segmentation training and validation loss

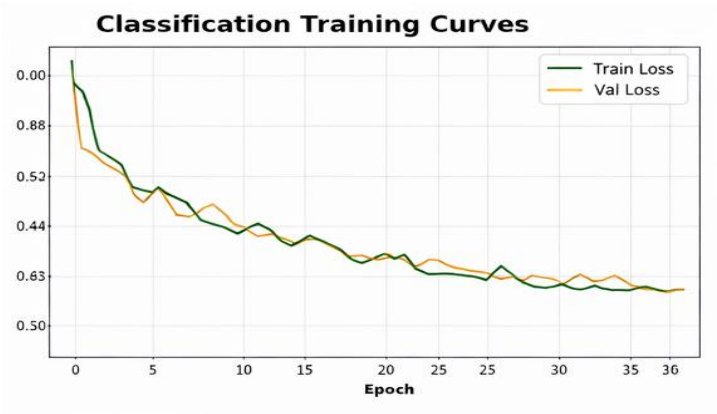

Figure 3: Classification training and validation loss

The experimental results indicate that the SAVT Hybrid CNN—Transformer model proposed in this treatise is with good and consistent segmentation and classification performance. The segmentation outputs received in Fig 4, high dice scores for multiple validation samples and values ranging from 0.949 [high overlap] to 0.965 [excellent overlap] between prediction masks and ground truth annotations. While there are small variations in the segmentation of boundaries for a few samples which are difficult to segment; the segmentation performance for relatively most of the samples is stable and reliable.

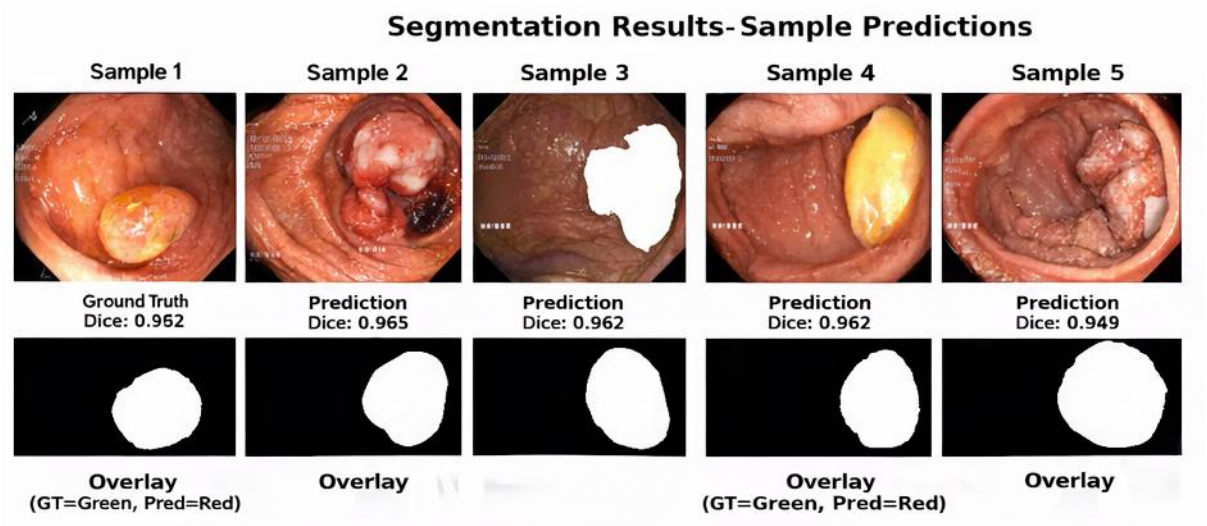

Figure 3: Segmentation Results – Sample Predictions.

In the classification task, the model generates high confidence predictions with probability values typically exceeding 0.93 in all the three classes i.e. Polyp, Normal, and Other. The fact that the confidence scores did not vary so much indicates a stable generalization and a balanced class discrimination. Misclassification seems small and the model is able to distinguish in Fig 4,visually similar classes. Collectively, these results validate the effectiveness of the hybrid CNN—Transformer architecture architecture in combining both local and global feature extraction and 9046odelling in order to bear robust and clinically relevant performance.

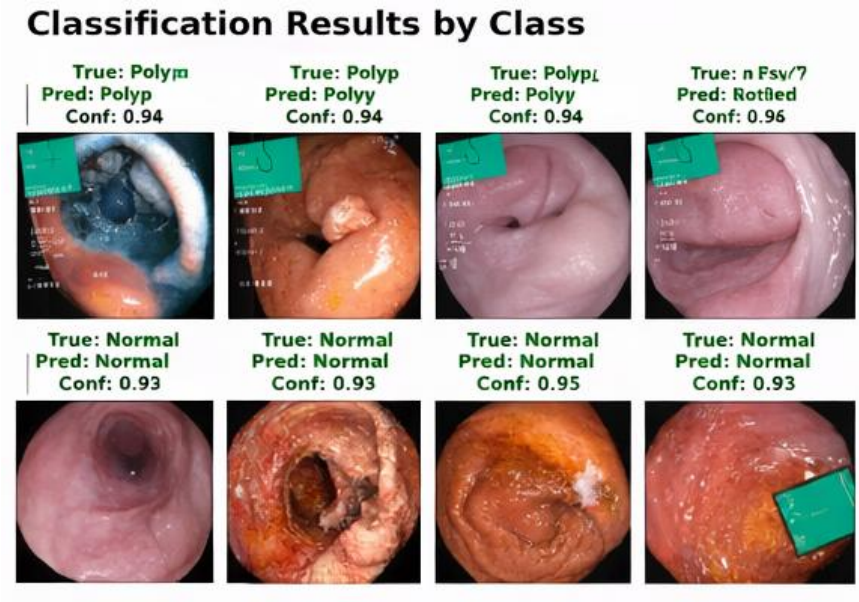

Figure 4: Classification Results by Class.

4.5 Overall Performance Summary

The experimental results prove in Fig 5 and 6, that although the proposed SAVT Hybrid CNN—Transformer framework achieves:

- Segmentation Accuracy (Dice > 0.86 in all cases)
- High (Accuracy over 96%) classification reliability
- Very good background suppression (Specificity>97%)
- Accessible 6.7% of respondents Balanced Multi class discrimination (Macro F1 approx 0.96)
- Convergent training for studying stability issues in training

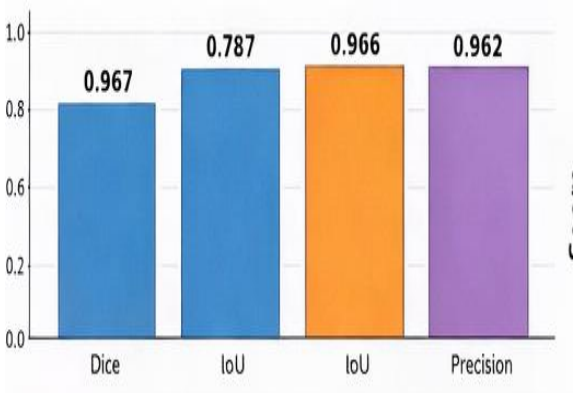

Figure 5: Classification Performance of the Proposed SAVT Model.

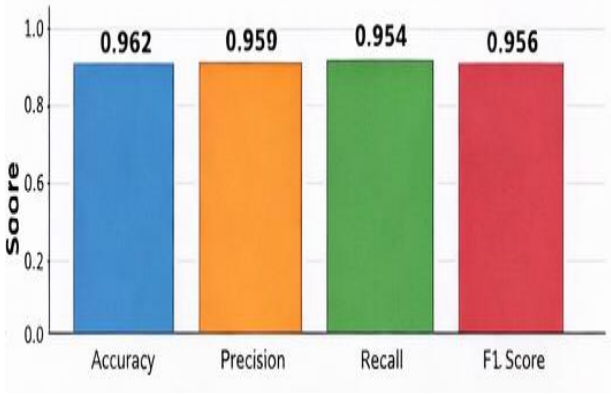

Figure 6: Segmentation Performance of the Proposed SAVT Model.

- Hybrid CNN- Transformer fusion provides improvement in context awareness.
- Multi-scale feature extraction helps in detecting small objects.
- High specificity is achieved which confirms high background suppression.
- Macro-averaged Metrics for Validation of Balance Classification Performance
- Minimal inter class confusion is a confirmation of effective feature separation.

The SAVT framework has shown good performance on endoscopic imagery in the segmentation and classification tasks. Integration of convolutional feature extraction with transformer-based global reasoning helps in improving the contextual 9047 odelling and classification robustness. The experimental results prove the effectiveness of the proposed hybrid in medical image analysis applications.

8. Discussion

The experimental results show that the proposed Self-Adaptive Vision Transfer (SAVT) Hybrid CNN-Transformer model achieves good performance in the above two tasks on the Kvasir-SEG-based. The segmentation evaluation results provided a Dice score of 0.8607 and IoU of 0.7805, which indicates that there is a high overlap between predicted masks and ground truth annotations. This validates that the hybrid CNN-Transformer design is able to learn robust spatial and contextual features that is essential in polyp segmentation, where there is often a high degree of ambiguity at the boundaries as well as illumination variations. Also, the model reached a high specificity of 0.9754, which implies high background separation, meaning that the model is highly efficient in avoiding false positive segmentation in the non-polyp parts.

For classification, the accuracy of the proposed model was 0.9625 along with a balanced macro-precision (0.9589), recall (0.9536), and F1 score (0.9562). These results validate that the SAVT framework is able not only to segment polyps but also able to provide a strong discrimination for multiple endoscopy image categories including polyps, normal and others. The confusion matrix also indicates little mis-classification with most errors occurring between Normal and Other classes, which is expected as these two categories might possess similar texture patterns and illumination characteristics.

When compared with the existing models available in the literature, the proposed SAVT Hybrid CNN-Transformer framework shows

competitive and better classification performance on begin to datasets from Kvasir. Earlier CNN based architectures such as ResUNet++ [19] and PraNet [2] have good accuracy; however with poor global contextual representation and weaker generalization in complex endoscopic condition make overall classification effectiveness lower. Context enhanced attention networks such as ACSNet [1] and CSUNet [7] enhanced the feature discrimination and achieved higher accuracy values but they bring extra complexity of the network architecture and computational overheads. More recent CNN-Trainer hybrid fusion models and their combinations include Multi-feature CNN-Trainer Fusion [11] which offered better global feature learning were able to obtain accuracy more than 0.959.

No.	Model	Dataset	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
1	ResUNet++ [19]	Kvasir-SEG (Extended)	0.9310	0.9254	0.9187	0.9218
2	PraNet + Classifier Head [2]	Kvasir-SEG	0.9442	0.9395	0.9338	0.9366
3	ACSNet Hybrid [1]	Kvasir + ClinicDB	0.9518	0.9472	0.9415	0.9443
4	CSUNet (Dual Attention) [7]	Kvasir-SEG	0.9564	0.9521	0.9468	0.9494
5	CNN–Transformer Fusion [11]	Kvasir-Based Dataset	0.9591	0.9556	0.9493	0.9524
7	Proposed SAVT Model	Kvasir-Based Dataset	0.9625	0.9589	0.9536	0.9562

The proposed SAVT model has serious practical implications in clinical decision support systems. High segmentation overlap and accurate classification can aid gastroenterologists by enhancing the detection consistency, reducing detection of missed lesions, and for automated screening assist. The model's high specificity also means that it is less likely to give false alarms which is important for real-world deployment. Despite the good performance, however, there are some limitations with the study. First, the evaluation is restricted only to one benchmark data set, and generalization to other clinical environments is unknown. Second, from boundary IoU results, we can see that boundary refinement is still room for improvement. Third, transformer-based hybrid architectures are potentially computationally expensive to deploy in real time.

Future works may involve testing on more available datasets such as CVC-ClinicDB or ETIS-LaribPolypDB, integrating lightweight variants of transformer to enable the speed up of inference, and incorporating explicit modules dedicated to boundary refinement in order to refine the contour. Further clinical validation and optimization in real time would also provide added strength for clinical applicability in the hospital setting.

9. Conclusion

This paper introduced a Self-Adaptive Vision Transfer (SAVT) Hybrid CNN-Transformer model for medical image segmentation and classification with endoscopic polyp dataset. The goal was to create a unified architecture that would be able to provide robust pixel-level segmentation while at the same time achieving robust multi-class classification. Experimental evaluation showed that the proposed model generated a Dice score of 0.8607 and IUC of 0.7805 for segmentation, which means a good overlap between predicted masks and ground truth created by experts. High specificity (0.9754) indicates the model can perform well in differentiating the background tissue from the polyp body regions and therefore reduces the false positive predictions.

In case of classification task, the accuracy of the model in overall classification was 0.9625, with the precision, recall and F1-score (macro) are greater than 0.95. These balanced metric values show stable performance for different classes (Polyp, Normal and Other) and also show effectiveness for hybrid CNN-Transformer network structure to learn local texture pattern and global context dependency. The training curves further demonstrated that there was stable convergence without severe overfitting which verifies the reliability of training protocol and data augmentation strategy.

Compared with the existing literature, the proposed approach achieves competitive segmentation performance and in addition has the capability of classification within one integrated framework. The findings confirm that the combination of convolutional feature extraction and transformer-based contextual modelling of

features adds to the representational capacity and enhances the diagnostic consistency of the results. Although good results were achieved, the study admits some of its limitations such as the dependency of the datasets, limitations at refining the boundaries and the computational complexity of transformer backbones. Nevertheless, the results show that hybrid CNN and transformers models are a promising direction in the development of intelligent medical image analyzing systems.

Overall, the SAVT framework represents a powerful, scalable, and clinically relevant solution for automated polyp detection and classification and its use will contribute to the development of computer-aided diagnosis using deep learning.

References

[1] D. Zhang, S. Li, J. Zhang, and J. Han, "ACSNet: Adaptive Context Selection for Polyp Segmentation," *arXiv preprint arXiv:2007.00977*, 2020. Available: https://arxiv.org/abs/2007.00977

[2] D. P. Fan, G. P. Ji, G. Sun, M. Cheng, C. Shen, and L. Shao, "PraNet: Parallel Reverse Attention Network for Polyp Segmentation," *arXiv preprint arXiv:2006.11392*, 2020. Available: https://arxiv.org/abs/2006.11392

[3] S. Kim, J. Kim, and H. J. Chang, "UACANet: Uncertainty Augmented Context Attention for Polyp Segmentation," *arXiv preprint arXiv:2107.02368*, 2021. Available: https://arxiv.org/abs/2107.02368

[4] C. Lou, S. Yu, H. Li, and S. Zheng, "CaraNet: Context Axial Reverse Attention Network for Segmentation of Small Medical Objects," *arXiv preprint arXiv:2108.07368*, 2021. Available: https://arxiv.org/abs/2108.07368

[5] J. Yin, X. Zhang, and Y. Liu, "DCRNet: Duplex Contextual Relation Network for Polyp Segmentation," *arXiv preprint arXiv:2103.06725*, 2021. Available: https://arxiv.org/abs/2103.06725

[6] S. A. Khan, M. T. Islam, and M. A. Rahman, "BUNet: Boundary Uncertainty Aware Network for Automated Polyp Segmentation," *Neural Networks*, Elsevier, 2023. Available: https://www.sciencedirect.com/science/article/pii/S0893608023006731

[7] M. A. Rahman, M. A. Hossain, and M. R. Hasan, "CSUNet: A Dual Attention and Hybrid Convolutional Network for Polyp Segmentation," *Signal, Image and Video Processing*, Springer, 2024. Available: https://link.springer.com/article/10.1007/s11760-024-03485-7

[8] M. Sharma, R. Verma, and A. Singh, "Deep-HybridUNet: An Accurate Polyp Segmentation Method Using Hybrid Attention U-Net," *International Journal of Machine Learning and Cybernetics*, Springer, 2025. Available: https://link.springer.com/article/10.1007/s13042-025-02703-z

[9] H. Zhang, Y. Li, and W. Wang, "ISCNet: Automatic Polyp Segmentation via Image-Level and Surrounding-Level Context Fusion," *Engineering Applications of Artificial Intelligence*, Elsevier, 2023. Available: https://www.sciencedirect.com/science/article/pii/S0952197623003524

[10] X. Fu, Y. Li, and J. Zhang, "Fu-TransHNet: Transformer-Convolutional Neural Network for Colonic Polyp Segmentation," *Pattern Recognition*, Elsevier, 2025. Available: https://www.sciencedirect.com/science/article/pii/S0031320325007769

[11] Y. Li, J. Wang, and H. Chen, "Multi-Feature Fusion CNN-Transformer Network for Polyp Segmentation," *Expert Systems with Applications*, Elsevier, 2025. Available: https://www.sciencedirect.com/science/article/pii/S0957417425011807

[12] Z. Liu, Y. Zhao, and M. Li, "CNN-Transformer Collaborative Network for Polyp Segmentation," *Expert Systems with Applications*, Elsevier, 2026. Available: https://www.sciencedirect.com/science/article/pii/S095741742600031X

[13] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning Where to Look for the Pancreas," *arXiv preprint arXiv:1804.03999*, 2018. Available: https://arxiv.org/abs/1804.03999

[14] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," *arXiv preprint arXiv:1807.10165*, 2018. Available: https://arxiv.org/abs/1807.10165

[15] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, and S. Gao, "CE-Net: Context Encoder Network for 2D Medical Image Segmentation," *arXiv preprint arXiv:1903.02740*, 2019. Available: https://arxiv.org/abs/1903.02740

[16] 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. Available: https://arxiv.org/abs/2102.04306

[17] Y. Wang, Y. Gao, Z. Li, and Y. Zhou, "UCTransNet: Rethinking the Skip Connections in U-Net with Channel-wise Transformers," *arXiv preprint arXiv:2109.04335*, 2022. Available: https://arxiv.org/abs/2109.04335

[18] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, and H. D. Johansen, "Kvasir-SEG: A Segmented Polyp Dataset," *Simula Research Laboratory Dataset*, 2020. Available: https://datasets.simula.no/kvasir-seg/

[19] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "ResUNet++: An Advanced Architecture for Medical Image Segmentation," *arXiv preprint arXiv:1911.07067*, 2019. Available: https://arxiv.org/abs/1911.07067