International Journal on Engineering Technology and Infrastructure Development (InJET-InDev)
www.lec.edu.np/journal
Volume 2, Issue no. 1, Page 1-13
Received Date: 2/4/2024
Revised Date: 5/10/2025
Accepted Date: 6/26/2025
ISSN: 3021-999X (Print)

Skin Lesion Segmentation using Vision Transformer and UNet

Bibat Thokar[1]
[1]*Department of Computer Engineering, Lalitpur Engineering College, Lalitpur, Nepal, bibatthokar@lec.edu.np*

Abstract

Medical images play a crucial role in diagnosing and analyzing serious illnesses. To make the typically lengthy process of reviewing these images more efficient, an automated approach for segmenting abnormal features is essential. Due to the limited availability of medical image data, deep learning frameworks for multi-class image segmentation have been developed. However, many current deep learning frameworks lack flexibility. To address the problem, advanced architectures have been added to improve segmentation performance. In particular, the UNet model has been enhanced with a Vision Transformer (ViT), enabling it to better capture structural features in medical images. The ISIC dataset has been preprocessed through methods like image augmentation, contrast limited adaptive histogram equalization (CLAHE), and normalization. The dataset has been then split into training, validation, and testing sets for optimal use. The training set has been used to train the model, with the adaptive moment estimation (Adam) optimizer aiding in optimization. The model performance has been evaluated using categorical cross-entropy to assess loss. The model has shown a*ccuracy* of 0.9378, a *precision* of 0.8713, a *sensitivity* of 0.8345, and an *F1-Score* of 0.8525 for the ISIC dataset.

***Keywords*:** Medical image, Segmentation, UNet, Vision Transformer, Adam Optimization

1. Introduction

Medical image segmentation, the process of distinguishing pixels representing organs or lesions from the background in medical scans like CT or MRI images, is a complex task in medical image analysis. This process is essential for extracting critical information regarding the shapes and sizes of organs. Over time, numerous automated segmentation systems have been developed using advancements in technology. Initially, these systems relied on traditional techniques, including edge detection filters and mathematical algorithms. Later, machine learning methods utilizing manually crafted features became widely used, although designing and extracting these features was often complex and presented challenges for practical deployment. Deep learning has emerged as the preferred approach for image segmentation, especially within medical applications. In recent years, deep learning-based image segmentation has gained significant attention, underscoring the need for a thorough review of developments in this area. To date, there appears to be no comprehensive review specifically focused on the application of deep learning in medical image segmentation (Hesamian, M.H., Jia, W., He, X. and Kennedy, P., 2019). Modern semantic segmentation methods generally employ convolutional encoder-decoder architectures.

The encoder extracts low-resolution features from the input image, while the decoder up-samples these features to produce segmentation maps with class labels for each pixel.Leading-edge techniques, such as Fully Convolutional Networks (FCN), have demonstrated remarkable performance on challenging segmentation benchmarks by using layered, learnable convolutions that capture rich semantic information, a key factor in their success within computer vision. However, the localized scope of convolutional filters restricts access to global image context, which is crucial for segmentation tasks since the labeling of local patches often depends on the overall image context (Strudel, R., Garcia, R., Laptev, I. and Schmid, C., 2021). Medical images are typically high-resolution and feature numerous intricate, interconnected structures. A key challenge lies in managing long-range dependencies in these images efficiently, minimizing computational resource demands. Additionally, precise boundary segmentation is vital for accurate diagnosis and effective treatment, making it more critical than general semantic segmentation. Consequently, a major focus of our approach is preserving detailed information

and achieving clearer boundaries in the segmentation process. To address these needs, an advanced Vision Transformer-based UNet model has been developed, aiming to lower computational costs while improving segmentation accuracy.

2. Literature Review

Several bio-medical image segmentation approaches have been introduced for medical research along with disease diagnosis and treatment processes. A U-shaped hybrid Transformer Network (UTNet) introducing both self-attention and convolution methods for medical image segmentation (Gao, Y., Zhou, M. and Metaxas, D.N., 2021) has been used. The approach has harnessed self-attention to capture long-range relationships while using convolution layers to gather local intensity features, thus avoiding the need for extensive transformer pre-training. The self-attention mechanism optimizes efficiency, reducing complexity from *O(n2)* to approximately *O(n)* in both time and space. The performance of UTNet has been evaluated using multi-label, multi-vendor cardiac MRI challenge data, covering segmentation tasks for the myocardium (MYO). The Dice scores achieved for UTNet were 93.1 for LV, 88.2 for MYO, and 83.5 for RV, with an average dice score of 88.3.

A query-informed calibration (QIC) strategy model called QNet, inspired by the learning approach of expert clinicians, has been implemented to enhance a recent state-of-the-art network, ADNet (Shen, Q., Li, Y., Jin, J. and Liu, B., 2023). The approach allows the network to dynamically adjust its thresholds and prototypes during inference, unlike ADNet, which relies on fixed thresholds and static prototypes to reduce distribution discrepancies between the query image and support set. The model has been evaluated on two well-known MRI datasets for few-shot segmentation (FSS): ABD, sourced from the ISBI 2019 CHAOS Challenge, which includes 20 3D T2-SPIR MRI scans averaging 36 slices each of the liver, kidneys, and spleen, and CMR, derived from the MICCAI 2019 Multi-sequence Cardiac MRI Segmentation Challenge (bSSFP fold), comprising 35 3D cardiac MRI scans with an average of 13 slices each. Q-Net has achieved a Dice score coefficient of 81.02±8.08 on the ABD dataset and 78.15±10.22 on the CMR dataset.

A hierarchical U-shaped transformer architecture, MISSFormer (Huang, X., Deng, Z., Li, D. and Yuan, X., 2021) that does not rely on positional encoding, has been employed specifically for medical image segmentation. An improved version of the Mix-FFN, a powerful feed-forward network, has been redesigned to enhance feature discrimination, capture long-range dependencies, and incorporate local context. This has been expanded into an Enhanced Transformer Block (ETB), forming the core of the Enhanced Transformer Context Bridge, which is structured to capture global and local correlations across hierarchical multi-scale features. Testing was conducted on the Automated Cardiac Diagnostic Challenge (ACDC) and Synapse multi-organ segmentation datasets. The Synapse dataset includes 30 abdominal CT scans, comprising 3779 axial clinical CT images, randomly divided into 18 scans for training and 12 for testing. The ACDC dataset achieved a dice score coefficient of 90.86, underscoring the model effectiveness.

A hybrid Transformer model, Medical TransFormer or Med-Former (Gao, Y., Zhou, M., Liu, D., Yan, Z., Zhang, S. and Metaxas, D.N., 2022), has been utilized to advance medical image segmentation. Med-Former demonstrates superior performance on smaller datasets without requiring pre-trained weights and offers scalability advantages on larger datasets, highlighting its efficiency, scalability, and adaptability across data sizes. Comprehensive testing has been conducted on both 2D and 3D scales using a large cardiac MRI dataset (containing 3,176 3D images) and three widely-used public datasets that feature various modalities and target structures. Med-Former achieved a dice score coefficient (DSC) of 89.05 across all datasets. Convolutional Swin-Unet (CS-Unet) is a low-complexity Transformer model that is built on entirely convolutional Transformer blocks. It was developed to improve the way Transformers model local information and segment organ borders. Tests conducted on CT and MRI datasets demonstrate that CS-Unet (24M parameters) trained from scratch achieves state-of-the-art performance, outperforming pre-trained Swin-Unet (27M) on ImageNet by around 3 dice scores. The model has generated 91.37 of DSC on the ACDC dataset.

For segmentation tasks, two types of 2D medical images were used: polyp segmentation in colonoscopy images and optic disc/cup segmentation in fundus images from the REFUGE20 challenge, employing a transformer-based model known as Segtran (Li, S., Sui, X., Luo, X., Xu, X., Liu, Y. and Goh, R.). Additionally, the model was evaluated on a 3D segmentation task, specifically brain tumor segmentation using MRI data from the BraTS19

challenge. When compared to U-Net and its variants (UNet++, UNet3+, PraNet, and nnU-Net), as well as deepLabv3+, Segtran consistently showed superior performance, achieving an average dice score of 0.817.

The diagnostic effectiveness of a Transformer-based Residual network (TransNetR) (Jha, D., Tomar, N.K., Sharma, V. and Bagci, U., 2024) has been evaluated for colon polyp segmentation. The architecture is structured as an encoder-decoder network with three decoder blocks, an upsampling layer at the network's output, and a ResNet50 encoder that is pre-trained. On the Kvasir-SEG dataset, TransNetR demonstrates a dice coefficient of 0.8706, a mean Intersection over Union (IoU) of 0.8016, and achieves real-time processing speeds of 54.60 frames per second.

To segment colonoscopic images, a refined ResUNet architecture known as ResUNet++ [Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P. and Johansen, H.D., 2019] has been utilized. It has achieved a dice coefficient of 81.33% and a mean Intersection over Union (mIoU) of 79.27% on the Kvasir-SEG dataset, along with a dice coefficient of 79.55% and an mIoU of 79.62% on the CVC-612 dataset, demonstrating strong evaluation performance. The ResUNet++ architecture incorporates components such as squeezing and excitation blocks, residual blocks, attention blocks, and Atrous Spatial Pyramidal Pooling (ASPP). When compared to other advanced techniques, the segmentation results for colorectal polyps using ResUNet++ were significantly superior. This architecture is particularly effective with a smaller number of images.

A Boundary Distribution Guided Network (BDG-Net) (Qiu, Z., Wang, Z., Zhang, M., Xu, Z., Fan, J. and Xu, L., 2022) has been employed for polyp segmentation. To create the Boundary Distribution Map (BDM), the Boundary Distribution Generate Module (BDGM) combines high-level characteristics. This BDM is then supplied to the Boundary Distribution Guided Decoder (BDGD) as supplemental spatial information to direct the polyp segmentation. Additionally, in BDGD, a multi-scale feature interaction technique is used to enhance the segmentation of polyps of various sizes. BDG-Net has shown 0.916, 0.864, 0.804, 0.725, 0.756, 0.679, 0.899, 0.831 of mDice, mIoU, mDice, mIoU, mDice, mIoU, mDice, mIoU for CVC-ClinicDB, ColonDB, ETIS, CVC300 dataset respectively. To detect ground glass opacities at the voxel level, a TV-Unet architecture similar to the Unet model was employed. A specific regularization term based on 2D-anisotropic total variation was developed and incorporated into the loss function, as the infected areas often exhibit related structures. Experimental results from a large CT segmentation dataset consisting of approximately 900 images indicated that this new regularization term enhances overall segmentation performance by 2% compared to a Unet model trained from scratch. The model achieved a mean Intersection over Union (mIoU) rate exceeding 99% and a Dice score of around 86%, demonstrating its exceptional capability in segmenting lung regions associated with COVID-19.

The best method for identifying and localizing discoveries for medical image analysis has been proposed, and it involves the generalization of models and limited datasets in terms of size and sample equality. The method works well with many datasets and doesn't require negative instances to be trained (Pogorelov, K., Ostroukhova, O., Jeppsson, M., Espeland, H., Griwodz, C., de Lange, T., Johansen, D., Riegler, M. and Halvorsen, P., 2018). Using only 356 training and 6000 test samples recorded by various devices, a detection specificity of 94% and an accuracy of 90.9% have been achieved with the generative adversarial network (GAN) approach.

3. Methodology

3.1. Vision transformer

The Transformer architecture, which was initially created for natural language processing (NLP), is extended to the computer vision domain by a vision transformer (ViT). A deep learning model called a "Vision Transformer" has an attention-based transformer architecture that is appropriate for pattern identification in images (Alexey, D., 2020). The vision transformer has an encoder-only architecture as opposed to the original transformer's encoder and decoder. The schematic diagram of the Vision transformer is shown in figure 1.

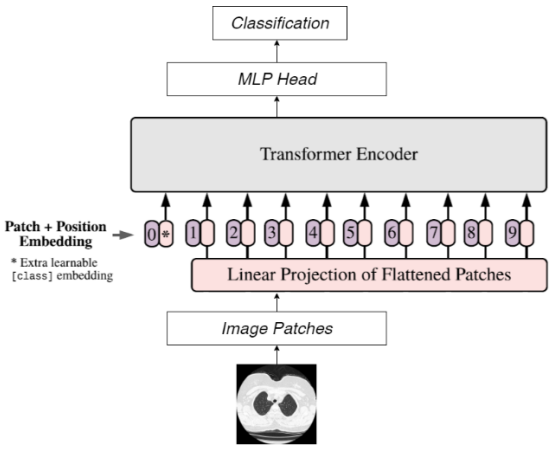

Figure 1. Vision Transformer

The Vision Transformer (ViT) model works by representing an image as a sequence of flattened patches, which are then processed by a transformer network. This network consists of sequential blocks containing a multi-head self-attention (MSA) module and a feed-forward neural network (FFNN). The self-attention module enables the model to focus on various regions of the input, capturing dependencies between image patches, while the FFNN applies non-linear transformations, capturing complex inter-patch relationships. Before entering the transformer network, image patches are embedded into a higher-dimensional space to preserve spatial information. The final transformer output is processed through a classification head that produces class probabilities. During pre-training, a Multi-Layer Perceptron (MLP) represents classification data, which is later replaced by a single linear layer during fine-tuning. The ViT architecture is composed of alternating MLP and MSA layers with residual connections and Layer Normalization (LN) before each block. Unlike traditional convolutional neural networks (CNNs), ViT captures long-range dependencies across image regions without relying solely on spatial proximity, allowing it to learn more abstract features suitable for complex image classification. ViT architecture also makes it effective on smaller datasets by facilitating learning from visual concepts.

3.2. UNet

UNet is a convolutional neural network architecture crafted specifically for segmenting medical images, featuring a U-shaped design with an encoder-decoder configuration (Liu, X., Song, L., Liu, S. and Zhang, Y., 2021). In the encoder, or contracting path, the model progressively reduces the input image's resolution, extracting key features through convolution and pooling layers, while the bottleneck compresses essential information. The decoder, or expansive path, reconstructs the image's resolution by up sampling and incorporating skip connections that directly pass high-resolution details from the encoder, maintaining crucial spatial information. The input image is processed through a series of convolutional layers with ReLU activation. During this process, the image dimensions reduce progressively, from 572×572 to 570×570 and finally to 568×568, due to the use of valid unpadded convolutions, which naturally decrease dimensionality. In the encoder block, the image size is further reduced through max-pooling layers with a stride of 2, while the convolutional layers increase the number of filters as the image moves deeper into the architecture. In the decoder block, however, the number of filters decreases as up sampling layers progressively restore the image's dimensions. Skip connections are also employed to link encoder outputs to decoder layers, preserving critical information from earlier layers and aiding in faster model convergence and improved results. The final convolution block has a series of layers ending in a convolution layer with a filter of 2, generating the output (Siddique, N., Paheding, S., Elkin, C.P. and Devabhaktuni, V., 2021).

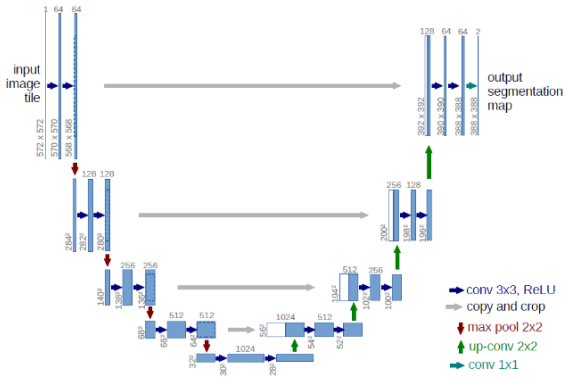

Figure 2. UNet Model (Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N. and Terzopoulos, D., 2021)

3.3. Combination of ViT with UNet

The UNet architecture has been enhanced with a Vision Transformer (ViT) backbone to improve segmentation performance, as depicted in figure 3. In this model, the Vision Transformer serves as the foundational architecture, initially capturing local features that aid in enhancing the structural feature visualization within the UNet (Shaker, A.M., Maaz, M., Rasheed, H., Khan, S., Yang, M.H. and Khan, F.S., 2024). This revised model follows a contracting-expanding structure, where an encoder, composed of a stack of transformers, is linked to a decoder via skip connections. Transformers operate on 1D input embedding sequences, splitting the input image into uniformly sized, non-overlapping patches. Each patch is linearly projected into a K-dimensional embedding space, retaining this dimensionality throughout transformer layers. A 1D positional embedding is added to preserve the spatial information of patches. Because this transformer backbone is intended for segmentation, it omits the [class] token from the embedding sequence.

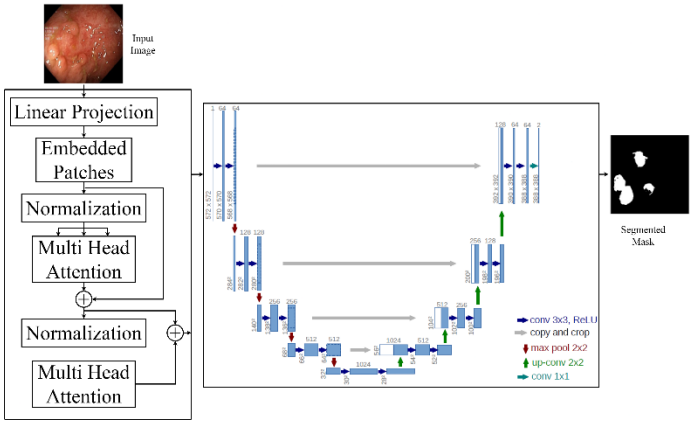

Figure 3. Combination of Vision Transformer with UNet Model

Transformer blocks include multi-head self-attention (MSA) and multilayer perceptron (MLP) with GELU activations, with each MSA sublayer containing n parallel self-attention heads. Self-attention (SA) aligns *query (Q)*, *key (K)*, and *value (V)* representations. After the transformer, sequence representations are reshaped as tensors, similar to the U-Net approach, and combined with features at various encoder resolutions. At each resolution, reshaped tensors from the embedding space are projected back to input space via 3×3×3 convolutional layers followed by normalization. A deconvolutional layer upscales the modified feature map by a factor of two. After concatenation with the previous transformer output, the up sampled map undergoes further 3×3×3 convolutions and is upscaled again using deconvolution. For voxel-wise semantic prediction, the final output passes through a 1×1×1 convolution with softmax activation, repeating this process across layers until reaching the original input resolution. The U-Net architecture includes a contracting (left) and expanding path (right). The contracting path uses consecutive 3×3 convolutions with rectified linear unit (ReLU) activation and 2×2 max pooling with stride 2 for down sampling, doubling feature channels at each stage. The expanding path involves upsampling the feature map, concatenating it with a cropped map from the contracting path, halving the feature channels via 2×2 convolution, and then applying two 3 × 3 convolutions with ReLU. Cropping is necessary due

to boundary pixel loss during convolution. At the last layer, a 1 × 1 convolution converts the 64-component feature vector to the desired class count, totaling 23 convolutional layers (Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W. and Wu, J., 2020). Selecting an input tile size that aligns with even x and y dimensions ensures continuous segmentation map tiling.

3.4 Data Collection

The medical image dataset was gathered from the International Skin Imaging Collaboration (ISIC) which has compiled a large public dataset of dermoscopy images, comprising over 20,000 images obtained from leading clinical centers, utilizing diverse technologies across institutions. This collection aims to provide a reliable dataset snapshot to aid in developing automated algorithms for melanoma diagnosis through three distinct lesion analysis tasks: segmentation, identification of dermoscopic features, and classification. A summary of the total data utilized is provided in Table 1, which shows the complete quantity of image data along with the associated masks used research work.

Table 1. Data Summary

Source	Images	Masks
International Skin Imaging Collaboration (ISIC)	4048	4048

3.5 Data Preprocessing

3.5.1 Image Augmentation

The collected image dataset has been preprocessed using series of data preprocessing steps (Thokar, 2024). The training dataset has been augmented by using multiple transformations, such as translation, rotation, scaling, flipping, intensity adjustment, deformation, and noise addition. These techniques help improve the model's ability to generalize and perform well on real-world medical images that exhibit different characteristics. Geometric augmentations, like horizontal flipping and Gaussian noise addition, have been applied to further enhance model performance. The horizontal flipping of the image can be expressed by equation 1.

$$x' = width - x \tag{1}$$

Here *width* and *height* are the relative width and height of the image. Flipping images allows the model to learn mirror invariant features, increasing its adaptability to images presented in different orientations. The Gaussian blur is applied as an image augmentation technique. It entails putting a Gaussian filter on the original image to blur it and eliminate high-frequency noise. Let the input image be $I(x, y)$ of size $W \times H$, then the mathematical formulations of the Gaussian blur operation are represented by equation 2.

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\pi\sigma^2}} \tag{2}$$

Here x and y are the pixel coordinates of the image, σ is the standard deviation of the Gaussian distribution to control the amount of blur applied to the image. The Gaussian blur operation has been implemented by convolving the input image $I(x,y)$ with the Gaussian kernel $G(x,y)$ using a convolution operation. The resulting blurred image $\hat{I}(x, y)$ is obtained as represented by equation 3.

$$\hat{I}(x,y) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} [I(x+i, y+j) \times G(i,j)] \tag{3}$$

Here i and j are the indices of the Gaussian kernel $G(x,y)$ ranging from $-k$ to k and k is an integer that determines the size of the Gaussian kernel. The larger value of k generates a stronger blur effect. $I(x+i,y+j)$ is the pixel value of the input image at the coordinates $(x+i, y+j)$. The convolution process is carried out using a sliding window technique, where a Gaussian kernel is centered on each pixel of the input image. Each pixel value is multiplied by the corresponding value in the Gaussian kernel, and the results are summed to produce the blurred pixel value in the output image.

3.5.2 Intensity Normalization

Normalization scales the pixel intensities of images to a consistent range, helping to minimize the impact of brightness and contrast variations due to differences in lighting or imaging conditions. Let I be an image of resolution $(H \times W \times C)$. Let H is the *height*, W is the width, and C is the channel number of the image. The pixel intensities of I are represented as a tensor of shape (H, W, C). To normalize the pixel intensities of I, the mean and standard deviation of the pixel intensities are first computed across all pixels and channels of the image. If μ and σ are the mean and standard deviation respectively, these values can be computed as represented by equation 4 and 5.

$$\mu = \frac{1}{N}\sum I \tag{4}$$

The above equation computes the mean across all pixels and channels

$$\sigma = \sqrt{(\frac{1}{N}\sum((I-\mu)^2)} \tag{5}$$

Here σ is the standard deviation across all pixels and channels where $N = H \times W \times C$ is the total number of pixels in the image. The image can be normalized by subtracting the mean and dividing by the standard deviation following the computation of the mean and variance of the pixel intensities as represented by equation 6.

$$I_{Normalize} = \frac{I-\mu}{\sigma} \tag{6}$$

The resulting image $I_{Normalize}$ has pixel intensities centered on zero and unit variance. In some cases, the normalized pixel intensities are scaled to a specific range such as [0, 1] or [-1, 1]. This can be done by applying a linear transformation to the normalized pixel intensities as represented by equation 7.

$$I_{NS} = \frac{I_{Normalize}-a}{b-a} \tag{7}$$

Where $a = min\ (I_{Normalize()})$ is minimum pixel intensity and $b = max\ (I_{Normalize()})$ is maximum pixel intensity. The resulting image I_{NS} has pixel intensities that are scaled to the range [0, 1] based on the minimum and maximum pixel intensities in the normalized image.

3.5.3 Contrast Limited Adaptive Histogram Equalization (CLAHE)

The CLAHE method divides the input image into contextual "tiles", each of which has a histogram used to match the output to a target histogram distribution (Reza, A.M., 2004). For an image with dimensions $W \times H$ and tiles of size $w \times h$, the total number of tiles can be calculated using equation 8.

$$T = \frac{W \times H}{w \times h} \tag{8}$$

The histograms of the obtained tiles are generated using the clip limit C_L of the image from the equation 9.

$$C_L = N_{CL} \times N_{avg} \tag{9}$$

Here N_{CL} is Normalized Contrast Limit and N_{avg} is average count of pixels. N_{avg} is obtained using the equation 10.

$$N_{avg} = \frac{N_x \times N_y}{N_g} \tag{10}$$

Here N_x, N_y, and N_g are the number of gray scales, x and y dimensions of the pixels, respectively. The relationship produces a mean of clip pixel values from equation 11.

$$N_{CP} = \frac{N\sum C_L}{N_g} \tag{11}$$

Here $N\Sigma C_L$ is overall number of C_L. The remaining pixels are redistributed using the equation 12.

$$R = \frac{N_g}{N_r} \tag{12}$$

The bi-linear interpolation is implemented to minimize false borders in the image and to combine neighboring tiles. The CLAHE methodology focuses on enhancing local contrast to overcome the limitations of global approaches. Important hyper-parameters for this approach include the tile size and clip limit. The use of hyper-parameters improperly may result in an inappropriate impact on image quality. The best options for clip limit, clip size (10, 10), and other parameters are selected after a thorough analysis. A picture histogram created from the raw photos and the CLEHE shows the intensity levels of the image pixels. Using photographs with an intensity range of pixels 0-255, the study uses a statistical representation of the data. When compared to the original raw image, a CLAHE image contrast is found to be noticeably better.

3.6 Instrumentation Requirement

Experiments with the UNet combined with a Vision Transformer model were conducted using Kaggle and Google

Colab, with TensorFlow for deep learning computations. Data preprocessing was performed on an 8th-generation i5 device with Jupyter Notebook, while an 8-core NVIDIA P100 GPU was utilized from Kaggle and Google Colab. The learning rate for the UNet-Vision Transformer model started at 10^{-4} and was adjusted as needed to optimize the training process. The batch size was set to 10, with 12 layers in the network. Input images were sized at 256×256×3, with a patch size of 16×16 for the Vision Transformer backbone.

4. Results

Medical images are more complex than typical images. To effectively generalize and classify these images, the models have been initially trained on a training dataset. Throughout the training and testing phases, hyper-parameter tuning and visualization of results were conducted. The collected dataset was split into three sections: 80% of the images were allocated for training, 10% for validation, and the remaining 10% for testing. The model underwent training for 35 epochs, during which it was updated based on the training set, which was further divided into batches. For each batch, the model computed the loss (the discrepancy between predicted and actual outputs) and the prediction accuracy. After each epoch, the model's performance was assessed using the validation set, and the model weights were saved if there was an improvement in validation loss, utilizing the "Early Stopping" technique.

4.1. Training Results

The training performance of the model during process has been evaluated using dice coefficient, validation dice coefficient, training accuracy, training loss, validation accuracy and validation loss. Each of the results has been generate and analyzed to evaluate model performance.

4.2 Dice Coefficient and Validation Dice Coefficient Results

The figure 5 and 6 shows the dice coefficient and validation dice coefficient of the model at each epoch during the training process. The Dice Coefficient values for the training set start at 0.6247 and gradually improve, reaching 0.8514 by the final epoch. This steady rise reflects the model's learning and adaptation to the data. The Validation Dice Coefficient follows a similar trend, starting at 0.7541 and increasing to 0.8525.

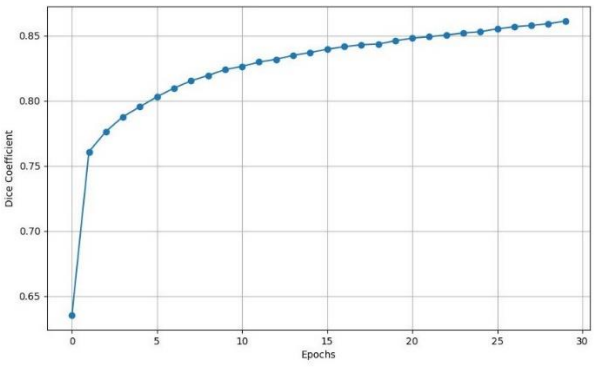

Figure 4. Epoch versus dice coefficient plot

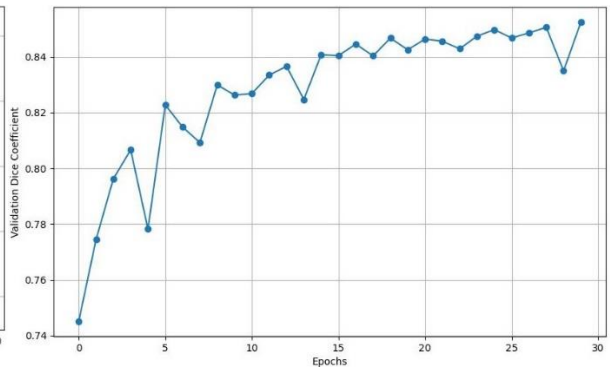

Figure 5. Epoch versus validation dice coefficient plot

4.3 Training Accuracy and Training Loss Results

The figure 7 and 8 represents the relationship between epoch, training accuracy, and training loss of a model during the training phase. The training accuracy increased from 0.7966 to 0.9341, indicating that the model became progressively better at correctly classifying the images. Concurrently, the training loss decreased from 0.3753 to 0.1486, showing that the model predictions became more accurate with reduced errors over time.

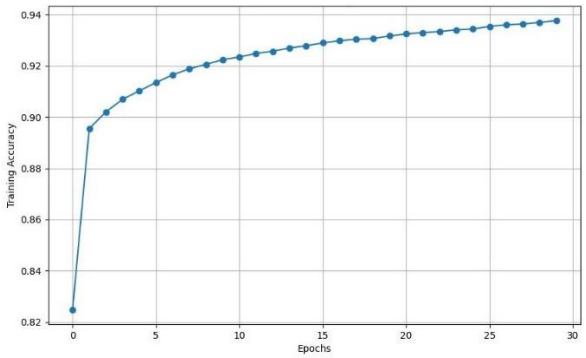

Figure 6. Epoch versus training accuracy plot

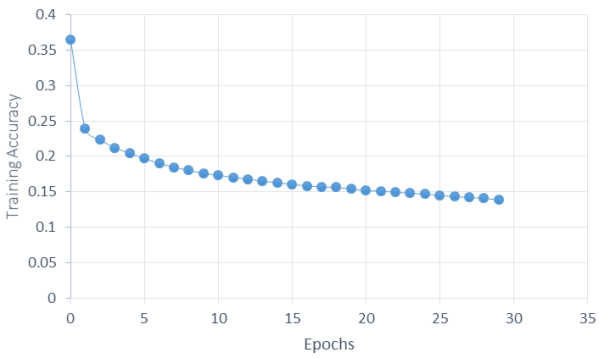

Figure 7. Epoch versus training loss plot

4.4 Validation Accuracy and Validation Loss Results

The figure 9 and 10 represents the validation accuracy and validation loss and its relationship with the epoch. The validation accuracy started at 0.885647178 and consistently improved, reaching a peak of 0.933357298 by the final epoch. This indicates a steady enhancement in the model's ability to correctly identify and segment the images. Concurrently, the validation loss showed a decreasing trend, beginning at 0.247276068 and reducing to 0.148534998.

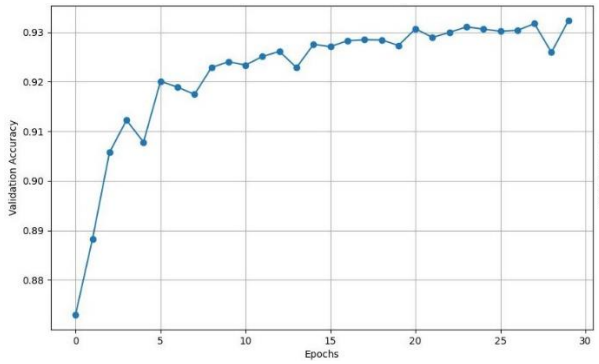

Figure 8. Epoch versus validation accuracy plot

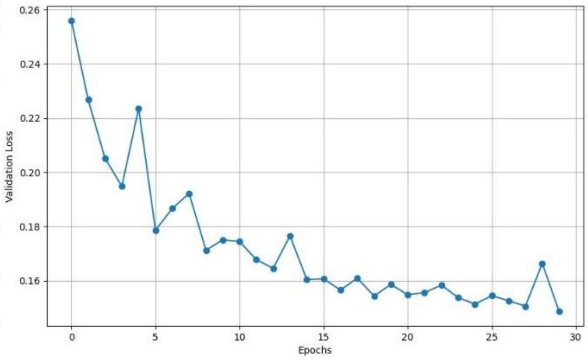

Figure 9. Epoch versus validation loss plot

4.5 Testing Results

The image 11 shows the segmentation performance of the UNet-ViT combined model for the ISIC dataset. The leftmost image is the input image, the middle image is the ground truth mask of the input image and the rightmost image is the predicted segmentation.

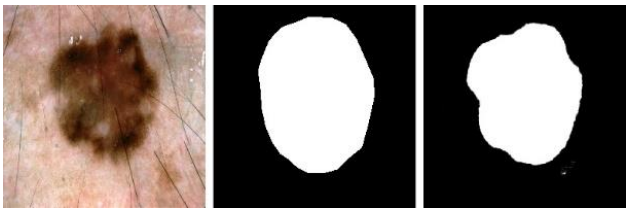

Figure 10. Leftmost image: Original Input Image. Middle image: Ground Truth of Segmented Mask and Rightmost Image: Predicted Mask

4.5.1 Confusion Matrix

As shown in the figure 12, the model performance on prediction with the ISIC dataset has been evaluated using a confusion matrix. The normalized *TP, TN, FP* and *FN* are found to be 0.97, 0.83, 0.034 and 0.17 respectively.

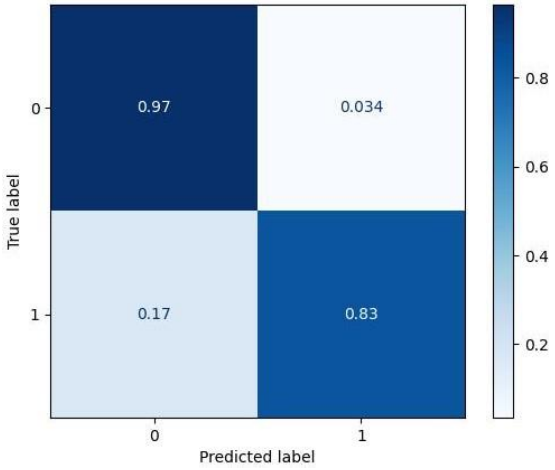

Figure 11. Confusion Matrix

4.5.2 ROC and PR Curves

The *ROC* and *PR* Curve of the UNet-ViT Model for the ISIC testing data have been generated as shown in figure 13 and 14 with *AUC* of 0.96 and 0.91 respectively.

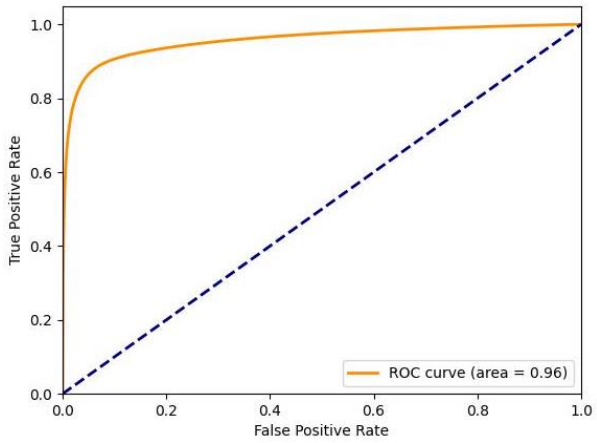

Figure 12. ROC Curve

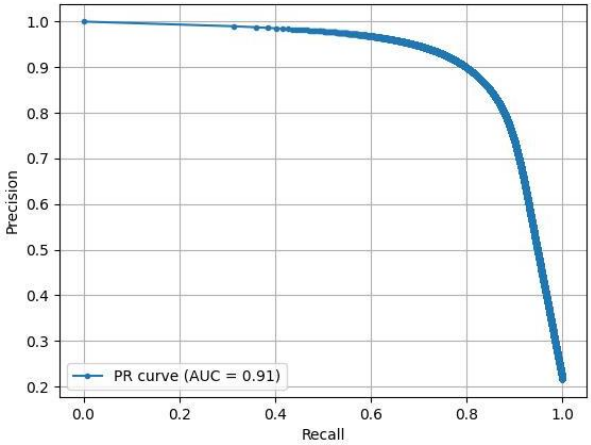

Figure 13. PR Curve

5. Discussion and Analysis

On analyzing these results for the ISIC dataset, the *accuracy, precision, recall,* and *F1-score* of the model during the testing process using the ISIC image dataset are found to be as shown in the table 3. The *learning rate* has been updated automatically as 10^{-2} from epoch 10 up to epoch 19, 10^{-3} from epoch 11 up to epoch 29 and 10^{-7} from epoch 30 up to epoch 35.

Table 2. Performance Metrics on ISIC dataset

Metrics	Results
Accuracy	0.9378
Precision	0.8713
Recall	0.8345
F1-Score	0.8525

6. Conclusion and Future Works

The UNet model, combined with the Vision Transformer as its backbone, achieved an *accuracy* of 0.9378, a *precision* of 0.8713, a *sensitivity* of 0.8345, and an *F1-Score* of 0.8525 for the ISIC dataset. Vision Transformers have proven to be highly scalable and efficient in managing large datasets and processing images simultaneously. These research findings can enhance the diagnostic capabilities of medical professionals by providing valuable insights and supporting decision-making. In the future, existing deep learning models could be integrated with other hybrid deep learning approaches to improve analysis and prediction accuracy. Such hybrid models could play a significant role in deepening the understanding of data and its characteristics.

Medical image segmentation can progress toward instance and panoptic segmentation by adapting existing techniques to meet specific challenges in healthcare imaging. Instance segmentation focuses on identifying and delineating individual objects within an image, enabling the differentiation between multiple instances of the same class, such as overlapping tumors or organs (Yang, R., Yu, J., Yin, J., Liu, K. and Xu, S., 2022). By employing models like UNet and Mask R-CNN and creating comprehensive datasets with detailed annotations, medical professionals can enhance the precise localization of structures, which is vital for treatment planning and surgical navigation. On the other hand, panoptic segmentation combines instance and semantic segmentation, providing a holistic understanding of an image by classifying all pixels while distinguishing between different instances (Zhang, D., Song, Y., Liu, D., Jia, H., Liu, S., Xia, Y., Huang, H. and Cai, W., 2018). This approach allows for a comprehensive view of anatomy, helping clinicians grasp the relationships between various structures, especially in complex cases with overlapping organs. The future of medical analysis and research can greatly benefit from these advancements, as improved accuracy in detecting and delineating structures will enhance diagnostics and treatment strategies. Furthermore, automated analysis tools can streamline workflows for radiologists, while detailed insights into anatomical variations support personalized medicine.

Acknowledgements

The authors would like to express sincere gratitude to former ***Asst. Prof. Dinesh Baniya Kshatri***, MSISE (M.Sc. in Informatics and Intelligent Systems Engineering) program coordinator, for providing invaluable guidance, insightful comments, meticulous suggestions, and encouragement for the conduct of this research work.

References

Hesamian, M.H., Jia, W., He, X. and Kennedy, P., 2019. Deep learning techniques for medical image segmentation: achievements and challenges. Journal of digital imaging, 32, pp.582-596.

Strudel, R., Garcia, R., Laptev, I. and Schmid, C., 2021. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 7262-7272).

Gao, Y., Zhou, M. and Metaxas, D.N., 2021. UTNet: a hybrid transformer architecture for medical image segmentation. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24 (pp. 61-71). Springer International Publishing.

Shen, Q., Li, Y., Jin, J. and Liu, B., 2023, September. Q-net: Query-informed few-shot medical image segmentation. In Proceedings of SAI Intelligent Systems Conference (pp. 610-628). Cham: Springer Nature Switzerland.

Huang, X., Deng, Z., Li, D. and Yuan, X., 2021. Missformer: An effective medical image segmentation transformer. arXiv preprint arXiv:2109.07162.

Gao, Y., Zhou, M., Liu, D., Yan, Z., Zhang, S. and Metaxas, D.N., 2022. A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. arXiv preprint arXiv:2203.00131.

Li, S., Sui, X., Luo, X., Xu, X., Liu, Y. and Goh, R., Medical image segmentation using squeeze-and-expansion transformers. arXiv 2021. arXiv preprint arXiv:2105.09511.

Jha, D., Tomar, N.K., Sharma, V. and Bagci, U., 2024, January. TransNetR: transformer-based residual network for polyp segmentation with multi-center out-of-distribution testing. In Medical Imaging with Deep Learning (pp. 1372-1384). PMLR.

Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P. and Johansen, H.D., 2019, December. Resunet++: An advanced architecture for medical image segmentation. In 2019 IEEE international symposium on multimedia (ISM) (pp. 225-2255). IEEE.

Qiu, Z., Wang, Z., Zhang, M., Xu, Z., Fan, J. and Xu, L., 2022, April. BDG-Net: boundary distribution guided network for accurate polyp segmentation. In Medical Imaging 2022: Image Processing (Vol. 12032, pp. 792-799). SPIE.

Pogorelov, K., Ostroukhova, O., Jeppsson, M., Espeland, H., Griwodz, C., de Lange, T., Johansen, D., Riegler, M. and Halvorsen, P., 2018, June. Deep learning and hand-crafted feature-based approaches for polyp detection in medical videos. In 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS) (pp. 381-386). IEEE.

Alexey, D., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929.

Liu, X., Song, L., Liu, S. and Zhang, Y., 2021. A review of deep-learning-based medical image segmentation methods. Sustainability, 13(3), p.1224.

Siddique, N., Paheding, S., Elkin, C.P. and Devabhaktuni, V., 2021. U-net and its variants for medical image segmentation: A review of theory and applications. IEEE access, 9, pp.82031-82057.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N. and Terzopoulos, D., 2021. Image segmentation using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence, 44(7), pp.3523-3542.

Shaker, A.M., Maaz, M., Rasheed, H., Khan, S., Yang, M.H. and Khan, F.S., 2024. UNETR++: delving into efficient and accurate 3D medical image segmentation. IEEE Transactions on Medical Imaging.

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W. and Wu, J., 2020, May. Unet 3+: A full-scale connected unet for medical image segmentation. In ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 1055-1059). IEEE.

Reza, A.M., 2004. Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement. Journal of VLSI signal processing systems for signal, image and video technology, 38, pp.35-44.

Yang, R., Yu, J., Yin, J., Liu, K. and Xu, S., 2022. A dense R-CNN multi-target instance segmentation model and its application in medical image processing. IET Image Processing, 16(9), pp.2495-2505.

Zhang, D., Song, Y., Liu, D., Jia, H., Liu, S., Xia, Y., Huang, H. and Cai, W., 2018. Panoptic segmentation with an end-to-end cell r-cnn for pathology image analysis. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11 (pp. 237-244). Springer International Publishing.