Highlights

Comparative Analysis of Retinal Vessel Segmentation Utilising Convolutional and Transformer-Based Architectures

Morteza Tavakol Sadrabadi,Hamed Agahi

- we provide a Comparison of the performance of Convolution-based and transformer-based architectures in retinal vessel Segmentation
- four different data-sets are considered from high resolution to low resolution
- Transformer-based architectures outperform convolution-based models in high-resolution data-sets.

Comparative Analysis of Retinal Vessel Segmentation Utilising Convolutional and Transformer-Based Architectures

Morteza Tavakol Sadrabadi[a], Hamed Agahi[a,1]

[a]Department of Mechatronics, Islamic Azad University, Shiraz branch, Shiraz, Iran,

ARTICLE INFO

Keywords:
Retinal Vessel
Deep Learning
Segmentation techniques
Vision Transformer

ABSTRACT

Retinal vessel segmentation plays a major role in identifying and screening eye-related diseases such as glaucoma, diabetic retinopathy, and microaneurysm. Utilising the automatic feature extraction capabilities of deep learning models, there is a potential to reduce the reliance on human expertise for disease diagnosis, enabling physicians to make faster and more efficient diagnoses. This study focuses on evaluating the effect of uncomplicated data augmentation techniques on retinal vessel segmentation, offering a comparative analysis across eight diverse Convolutional-based and transformer-based architectures, including UNet, Attention-UNet, Res-UNet, UNet++, UCTransNet, TransUNet, UNeXt and SwinUNet. Each model is trained on four publicly available datasets with different resolutions, from low-res to high-res images, including $DRIVE$, $STARE$,$CHASE-DB1$, and HRF. Subsequently, the study provides comparative results, offering insight into the efficiency of these models in retinal vessel segmentation. Obtained results indicate that even though the overall difference between models' performance is small, convolutional models perform better than transformers when trained on lower-resolution datasets (i.e. *DRIVE* dataset), while transformers are best on Higher-resolution datasets (i.e. *HRF*).

1. Introduction

Retinal vessel segmentation is a crucial component of computer-assisted screening, diagnosis, and treatment for a variety of cardiovascular and eye-related disorders, such as stroke, diabetes, hypertension, and diabetic retinopathy. Furthermore, it can be used for biometric identification by utilising the unique properties of an individual's retinal veins. Given fundus cameras are specialised, and blood vessels are small, retinal vascular segmentation presents a unique set of challenges. The camera and outside environment substantially impact image quality, which in turn affects segmentation performance. Additional difficulties arise from the features of retinal blood vessels, which include their abundance, small size, curvature, and variability in shape. Problems include parallel or closely positioned vessels during segmentation, missing detection of small vessels, and trouble with broken vessels at bifurcations [1].

Recognising the time-consuming and error-prone nature of manual segmentation performed by professionals, there has been a significant effort in automating retinal blood vessel segmentation. The majority of early automatic techniques for segmenting retinal vessels were concentrated on segmenting images of the fundus, examining connections between backdrop and vessel features, and extracting features from the images through the utilisation of matched filters, vessel tracking, and morphological reconstruction-based techniques[1]. Owing to the advancements in Deep Learning and the increased availability of computing power, recent years have witnessed a substantial increase in the number of studies and enhancements in the performance of retinal vessel segmentation algorithms when applied to fundus images.

The field of biomedical image segmentation experienced notable advancements since the introduction of the UNet network [2]. Many medical image segmentation studies rely on UNet and its extension architectures, and the number of research works published using UNet as the baseline model has significantly increased over the past few years. Besides, to address the complexity of medical image analysis and the drawbacks of the naive UNet architecture, a variety of different extensions of this architecture have been proposed over the years [3]. For example, [4] introduced a new architecture based on the UNet by modifying skip connections and sharing weights between layers, which reduced the model's susceptibility to overfitting. [5] proposed a new network named "Octave UNet" for vessel segmentation in colour fundus images using encoder-decoder-based octave convolution networks. In [6], the IterNet, a new network, was introduced based on a deeper mini-UNet that uses weight sharing and skip connections in retinal vessel segmentation. In another study, [7] proposed the Channel Attention Residual UNet to accurately segment retinal vascular and non-vascular pixels, achieving state-of-the-art performance on three datasets, including DRIVE, CHASE DB1 and STARE. In another interesting paper, [8] introduced the SuperVessel network. The method includes a super-resolution auxiliary branch for possible high-resolution detail features, which is used in training and excluded in testing. Furthermore, two modules are included to improve segmentation features: a Feature Interaction Module (FIM) with a restricting loss that focuses on the region of interest for better segmentation and an Upsampling with Feature Decomposition (UFD) module. [9] utilised the UNet++ architecture, which includes a

*Corresponding author
ORCID(s):
1

UNet structure with intermediate nodes that perform multi-resolution aggregation of features combined with data augmentation techniques and achieved comparable results with state-of-the-art retinal image segmentation models.

Following the introduction of the attention mechanism [10] and vision transformers [11], a variety of architectures were developed based on utilising various transformers. For example, [12] proposed a lightweight Spatial Attention UNet (SA-UNet) that employs a spatial attention module for inferring spatial dimensions and structured convolution dropout blocks to prevent overfitting. [13] conducted an empirical study on data processing, architecture design, attention mechanisms, and regularization strategies. Based on the insights gained from this review, they proposed a well-designed deep neural network that incorporated best practices and achieved high performance with AUCs of 0.98279, 0.98862, and 0.98724 on the DRIVE, STARE, and CHASE-DB1 datasets, respectively. However, the obtained performance is still slightly below the state-of-the-art performance obtained by [14] with $AUC - ROC$ equal to 0.9887, 0.9914, and 0.9887 for DRIVE, CHASE-DB1, and STARE datasets, respectively. [15] developed the Connection Sensitive Attention UNet to enhance the accuracy of retinal vessel segmentation. [16] suggested SDAUNet, a retinal vascular segmentation technique that improves the UNet structure by substituting a series deformable convolution module for the convolution module. Integrating lightweight and dual attention modules into the decoder enhances extracting features from retinopathy images, particularly for small vessels with intricate morphology. In another study, [17] proposed a modified UNet with a supervised attention mechanism that uses a Context Squeeze and Excitation module in the decoding module to amplify contextual information for small blood vessels and introduces a Decoder Fusion Module for thorough feature extraction in the encoding part. A Supervised Fusion Mechanism produces the final result and enhances segmentation performance by merging multi-scale characteristics and gathering data from different levels.

The Effect of proper data augmentation techniques on improving retinal image segmentation accuracy is investigated and highlighted in a series of studies, including [18] and [19]. Obtained results of [18] indicated that while the performance gap between different architectures and models are usually small, these minor differences could be obtained or even surpassed through proper data augmentation. In [19], authors used affine transformations like rotate, flip, zoom Out, random crop, shift and shearing and elastic transformations like elastic deformations, grid distortion, optical distortion and Pixel-level transformations like adding white noise, gamma correction, histogram equalization, pixel Dropout, sharpening, blurring and contrast adjustment for data augmentation. In another study, [20] increased the accuracy of retinal vessel segmentation by introducing two new data augmentation modules based on image gamma correction and replacing the ResNet-50 as the encoder section of the UNet network architecture. In [21], authors proposed an SGL scheme to improve the robustness of the model trained on noisy labels to deal with retinal vessel segmentation task's limitations like small-scale datasets and incorrect or incomplete vessel annotations. They utilise concatenated Unet architecture consisting of the enhancement and segmentation modules to learn the enhancement and the segmentation map. They also utilised random cropped patches of 256×256 and applied data augmentations such as horizontal and vertical flips, rotations, transpose and random elastic wrapping.

So far, various studies have focused on developing different architectures and obtaining higher accuracies through utilising deeper or more advanced models and preprocessing and augmentation strategies. Yet, there is space for a comprehensive analysis to compare how different developed architectures and models respond to the image quality and preprocessing and augmentation techniques. Hence, this study evaluates the accuracy of different convolutional and transformer-based architectures in segmenting the retinal vascular tree in the fundus images of retinal vessels, utilising image preprocessing and patch extraction data augmentation strategy. Consequently this paper evaluates the accuracy of nine architectures including naive UNet [2], AttUNet [22], ResUNet [23], UCTransNet [24], TransUNet [25] with ViT-B16 and ResNet50-ViT-B16 backbones, UNet++ [26], UNeXt [27], and SwinUNet [28] over four publicly available datasets, with the different image resolutions including DRIVE [29], STARE [30], CHASE-DB1 [31], and HRF [32] from low-resolution to high-resolution, respectively. All architectures are trained over raw and preprocessed images, and the results are then compared.

The remainder of this paper is hence structured as follows: Section 2 presents an overview of the utilised architectures, datasets, evaluation metrics, and implementation details; Section 3 presents the results obtained through model training experiments; Section 4 provides a discussion on the obtained results and comparison with other studies; and finally Section 5 presents the conclusions of the study.

2. Methodology

2.1. Datasets

This study performs a comparative analysis of the performance of different model architectures on four publicly available datasets with different image resolutions from low-res to high-res images, including DRIVE [29], STARE [30], CHASE-DB1 [31], and HRF [32]. An overview of their specifications is provided below.

DRIVE dataset comprises 40 images, each with dimensions of 584×564 pixels, of which 20 images are considered for training and 20 for testing. These images were acquired as part of a diabetic retinopathy diagnosis program in the Netherlands utilising a Canon CR5 non-mydriatic 3CCD camera with a 45° field of view (FOV).

STARE dataset comprises 20 images, each with dimensions of 605×700 pixels, captured with a 35-degree field of view,

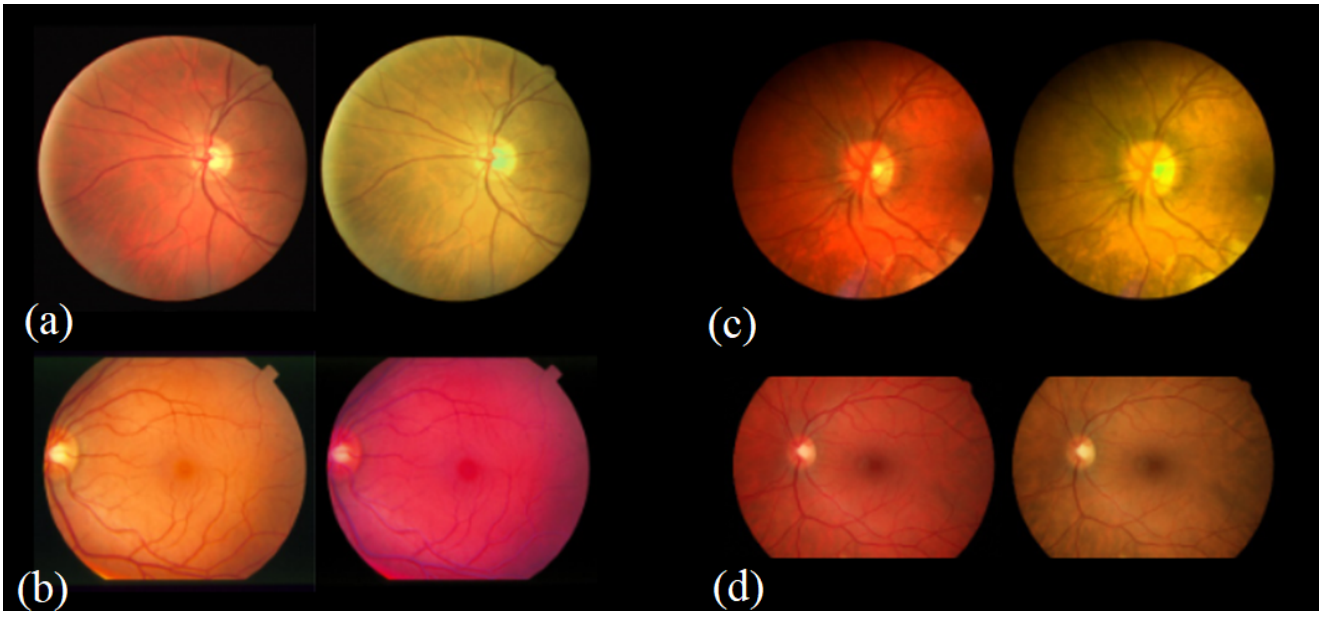

Figure 1: Comparison of raw (left side) and Gamma Corrected and Hue-Saturation-Value adjusted (right side) images from (a) Drive, (b) STARE, (c) CHASE-DB1, and (d) HRF datasets

of which 5 images are allocated for testing and 15 images for training of networks.

CHASE-DB1 dataset contains 28 retinal images with the resolution of 960 × 999 pixels captured using an NM-200-D handheld fundus camera with a 30° field of view, of which 20 images are allocated for training and 8 images for testing.

HRF dataset comprises 45 images and is organised as 15 subsets. Each subset contains one healthy fundus image, one image of a patient with diabetic retinopathy and one glaucoma image, each with dimensions of 3304 × 2336 pixels. In this study, we allocated 30 images for training and 15 for test.

2.2. Image Preprocessing and Augmentation

Image enhancement methods like gamma correction, CLAHE, and spatial transformations like Rotate and Flip have proven useful for retinal image segmentation tasks in different studies (e.g., see [19, 20]). Through a trial and error method, it is found that the random adjustment in Hue-Saturation-Value colour space can be helpful in low-quality datasets like DRIVE. Consequently, in this study, we consider random Hue-Saturation-Value (HSV) Adjustment combined with random Gamma correction as image practice to improve the quality of images and increase the contrast of vessels with the background. Figure 1 compares raw and preprocessed images for random samples from different datasets.

The augmentation methods utilized in this study include rotation and horizontal and vertical Flip, which is carried out utilising the Albumentations [33] library with default configs. Resize augmentation is not utilised in this study as trial and error indicated that it leads to the loss of thin vessels. Besides, it should be mentioned that the data augmentation process is entirely separated from network training to ensure all networks are trained with the same data.

2.3. Utilised Architectures

As previously discussed, this study intends to assess the efficacy of different UNet-based and transformer-based segmentation methods on four different retinal datasets. Thus, the performance of various segmentation models, including UNet [2], AttUNet [22], ResUNet [23], UCTransNet [24], TransUNet [25] with ViT-B16 and ResNet50-ViT-B16 backbones, UNet++ [26], UNeXt [27], and SwinUNet [28] are considered and compared. Following, we present an overview of the employed network architectures.

UNet [2] is a convolutional neural network architecture designed for semantic segmentation tasks in medical image analysis. Architecture is characterised by a U-shaped network structure, with a contracting path to capture context and a symmetric expanding path to enable precise localisation. This design facilitates incorporating both high-level and low-level features, making UNet particularly effective for image segmentation tasks where detailed spatial information is crucial.

AttUNet [22] is an extension of the UNet architecture that incorporates attention mechanisms to enhance its segmentation performance. The network incorporates attention gates at multiple scales to selectively focus on informative regions while suppressing irrelevant backgrounds. This attention mechanism allows the model to dynamically adapt its focus during the segmentation process, making it particularly effective for tasks where the target structures may vary in size, shape, or context.

Table 1
Number of trainable parameters for Utilised architectures

	UNet	AttUNet	ResUNet	UCTransNet	UNet++	TransUNet	TransUNet	UNeXt	SwinUNet
Number of Params ($\times 10^6$)	34.53	34.87	13.04	65.6	9.16	91.67	105.27	1.47	41.38

Figure 2: Selection of 224 × 224 pixel extracted patches and their corresponding mask from DRIVE dataset utilising the primary approach

ResUNet [23] builds upon the UNet architecture and incorporates residual connections inspired by the success of residual networks in image classification tasks. By addressing the vanishing gradient problem, residual connections help to overcome the challenge of training deep networks. This approach was developed for extracting roads from satellite or aerial photos, where precisely outlining road networks is critical for various applications such as urban planning and infrastructure management.

UCTransNet [24] argues that the current UNet framework with a simple skip connection scheme is still challenging to model the global multi-scale context of complex data. Consequently, this architecture proposes a Channel Transformer (CTrans) to replace the skip connections in UNet, which consists of two modules: (*i*) Channel-wise Cross Fusion Transformer (CCT) for the multi-scale encoder feature fusion and (*ii*) Channel-wise Cross Attention (CCA) for the fusion of the decoder features and the enhanced CCT features.

UNet++ [26], a nested UNet architecture, represents an advancement in medical image segmentation designed to address the challenges of capturing intricate details and hierarchical features in medical images. The architecture adds stacked skip connections to the classic UNet structure, resulting in a multi-level feature hierarchy. This stacked architecture allows the model to gather information at several scales, from fine-grained details to wider contextual characteristics. Stacked skip connections improve the flow of information across network tiers, boosting the model's capacity to identify structures in medical images correctly.

TransUNet [25] is a novel architecture for medical image segmentation. Leveraging transformer models, traditionally successful in natural language processing, the authors employ them as robust encoders for capturing intricate features in medical images. This fusion of transformers with UNet architecture enhances the model's ability to understand contextual information and spatial relationships, proving effective for tasks in medical image analysis.

It is worth mentioning that in this paper, we use two Backbones for the encoder with pre-trained weights in TransUNet, namely "$ViT - Base - 16$" and "$ResNet50 - ViT - Base - 16$", which have smaller trainable parameters compared with other backbones.

UNeXt [27] proposes a new convolutional multilayer perceptron (MLP) based network for image segmentation. The paper suggests that the current state-of-the-art medical image segmentation methods, such as UNet and its latest extensions like TransUNet, are unsuitable for rapid image segmentation in point-of-care applications due to their parameter-heavy, computationally complex, and slow-to-use

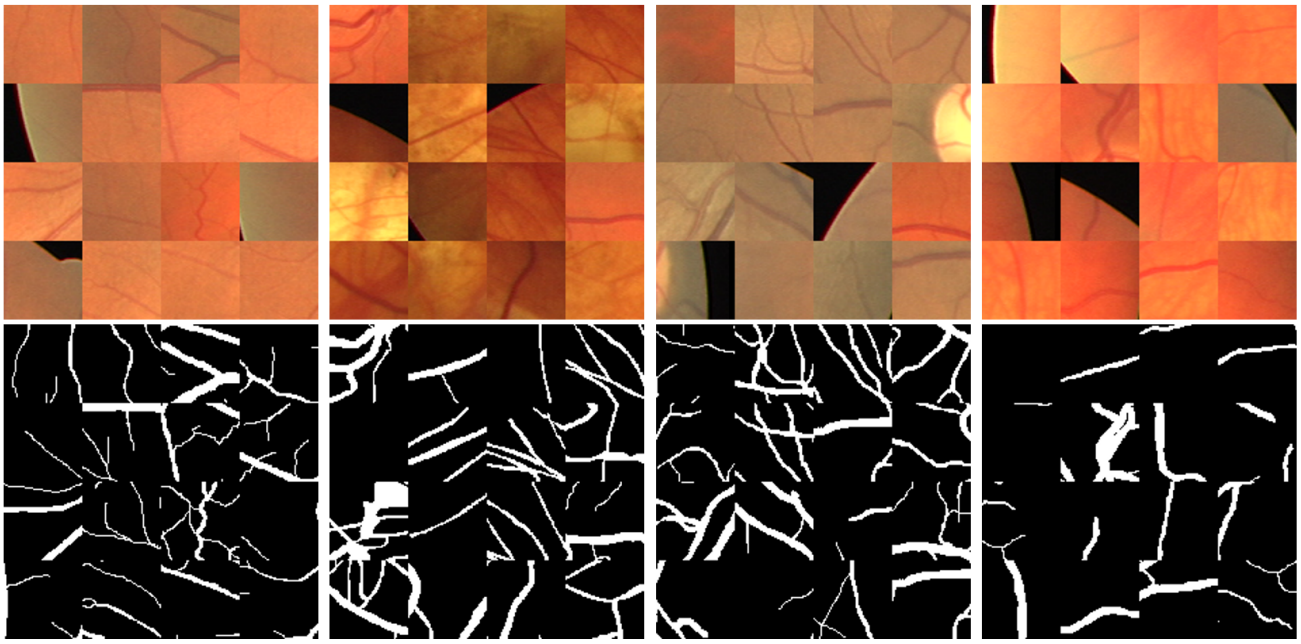

Figure 3: Selection of 224 × 224 pixel patches and their corresponding mask from DRIVE dataset formed utilising the second approach

nature. To address this issue, the paper proposes UNeXt, designed effectively with an early convolutional stage and an MLP stage in the latent stage. The paper proposes a tokenised MLP block that efficiently tokenises and projects the convolutional features and uses MLPs to model the representation. To further boost the performance, the paper proposes shifting the channels of the inputs while feeding into MLPs to focus on learning local dependencies. Using tokenised MLPs in the latent space reduces the number of parameters and computational complexity while resulting in a better representation to help segmentation. The network also consists of skip connections between various levels of encoder and decoder.

SwinUNet [28] is a pure transformer-based U-shaped architecture for medical image segmentation. The encoder, bottleneck, and decoder are all built based on the Swin Transformer block. The input medical images are split into non-overlapping image patches, tokenised and fed into the Transformer-based U-shaped Encoder-Decoder architecture with skip connections for local-global semantic feature learning. The paper shows that Swin-UNet outperforms other fully convolutional or combined transformers and convolution in multi-organ and cardiac segmentation tasks.

Table 1 presents an overview of the number of parameters for the utilised architectures to provide insight into the complexity of different models.

2.4. Evaluation Metrics

Multiple metrics must be utilised to evaluate a model's segmentation performance. The accuracy measures used in this study include the area under the receiver operating characteristics curve (AUC-ROC), the area under the precision-recall curve (AUC-PR), F1-Score, accuracy (Acc), Sensitivity (Sen), Specificity (Spe), which could be calculated as follows:

$$Pre = \frac{tp}{tp + fp} \tag{1}$$

$$Acc = \frac{tp + tn}{tp + tn + fp + fn} \tag{2}$$

$$F1 - Score = \frac{2 \times tp}{2 \times tp + fp + fn} \tag{3}$$

$$Spe = \frac{tn}{fp + tn} \tag{4}$$

$$Sen = \frac{tp}{tp + fn} \tag{5}$$

It should be noted that all metrics in this study are calculated using Scikit-learn library [34].

2.4.1. Patch Extraction and Implementation details

Given the implementation considerations of some models, all models are trained with patches with the size of 224 × 224 extracted from original images for consistency. This paper, however, has investigated two approaches for patch extraction.

The primary approach for extracting the patches from images in this study includes extracting overlapped random patches with the size of 224 × 224 provided that each patch should also be overlapped with the field of view (FOV). Performing a trial and error procedure, it was found that the efficient number of extracted patches for training models equals 1000 patches for the DRIVE and STARE datasets and 3000 patches for the CHASE-DB1. Further augmentation of data did not lead to an increase in the models' accuracy. Specific details on the HRF dataset augmentation are presented in section 3.4. Figure 2 presents samples of extracted 224 × 224 patches and their corresponding masks.

The second approach which is examined exclusively on the *DRIVE* dataset, includes extracting random patches with the size of 56 × 56 with the attitude of reducing the number of pixels outside of the field of view (FOV) as well as saving the border of FOV in random patches. Extracted 56 × 56 patches are then randomly attached in 4 × 4 settings to form 224 × 224 pixel images for the training phase of models. However, it is generally acknowledged that this method introduces artificial edges to the input images, which could force the model to avoid edges and potentially reduce model performance by forming misleading images with a large variety of HSV and Gamma values in different forming patches. Consequently, image enhancements are performed subsequent to the patch extraction to overcome this drawback. Figure 3 presents samples of formed 224 × 224 patches and their corresponding masks utilising the second approach.

For all experiments, a floating learning rate that decreases by validation loss is utilised in order to reach the highest accuracy and global minima. Initial learning rates are set to 0.01 for all networks with a scheduling factor of 0.1, decreasing the learning rate every 5 epochs if validation loss is not decreased. Performing a trial and error procedure, this study utilises the $Adam$ optimiser with weight decay equal to 0.0001 for Unet, AttUNet, UNeXt, UCTransNet, ResUnet, and Unet++. TransUnet and SwinUnet are, on the other hand, trained by utilising the SGD optimiser with weight decay equal to 0.0001 and momentum coefficient equal to 0.9. The loss function is formulated as a composite of Cross Entropy Loss and Dice Loss with Logits, both assigned equal weights. Finally, the early stop is configured for all models conditioned on the validation loss, not showing improvement for 20 consecutive epochs. The number of epochs that each model is trained with is presented under the $Stop$ heading in the training results of the models considered as an indicator of the model computational demand and efficiency.

To train and implement models, an Nvidia 3090 GPU with 24GB of RAM and the PyTorch framework [35] are utilised. A batch size of 10 was employed for the $DRIVE$ dataset, and a batch size of 16 was considered for the $CHASE-DB1$, $STARE$, and HRF datasets.

Table 2
Accuracy measures obtained by different models trained on unprocessed $DRIVE$ dataset

Architecture	AUC -ROC	AUC -PR	F1 -Score	Acc	Sen	Spe	Stop
Unet	0.9769	**0.9060**	**0.8073**	**0.9548**	0.7436	0.9856	92
AttUNet	**0.9771**	0.9051	0.8059	0.9545	0.7413	**0.9856**	78
ResUnet	0.9745	0.8999	0.8035	0.9538	0.7419	0.9847	101
UCTransNet	0.9757	0.9012	0.8017	0.9536	0.7360	0.9854	65
Unet++	0.9742	0.8986	0.8054	0.9537	**0.7526**	0.9830	90
TransUnet -ViT-B16	0.9310	0.7891	0.6943	0.9331	0.5963	0.9823	107
TransUnet -R50-ViT-B16	0.9658	0.8861	0.7989	0.9523	0.7448	0.9825	138
Unext	0.9710	0.8883	0.7903	0.9511	0.7245	0.9841	73
SwinUnet	0.9706	0.8888	0.7945	0.9519	0.7308	0.9841	98

3. Results

This section presents the performance of all models over different datasets. For this purpose, each dataset is presented and discussed separately. An overall discussion of the obtained results and comparisons with other studies is then presented in the next section.

3.1. DRIVE Dataset

As previously discussed in section 2.4.1, all models are trained on the $DRIVE$ dataset utilising two distinct data augmentation and patch extraction methods. All models are initially trained on the 224 × 224 patches extracted randomly from raw images, utilising the first patching approach to compare and evaluate the effectiveness of the preprocessing and data augmentation approaches. The results obtained from this configuration are presented in Table 2. In the next step, models are trained on the patches extracted before the image preprocessing, the results of which are presented in Table 3. The bold values in the tables delineate the top-performing architecture. Comparing the results, it could be concluded that the performing image preprocessing has increased the performance of all models except the UNeXt. From the model performance point of view, it could be observed that the UNet has performed better than other models in three out of six utilised accuracy metrics when trained on patches extracted from both raw and preprocessed images with $AUC-Roc$=0.9769, $AUC-PR$=0.9060, $F1-Score$=0.8073, and Acc=0.9548 for training on raw images and $AUC-Roc$=0.9788, $AUC-PR$=0.9105, and Acc=0.9559 when trained on preprocessed images. However, it should be highlighted that the AttUNet has obtained the highest $AUC-ROC$ and Spe values equal to 0.9771 and 0.9856 for training on unprocessed images. Figure 4 presents a qualitative comparison of the retinal vessel segmentation results obtained from different models trained on the *DRIVE* dataset. It could be observed that generally, UNet and AttUNet have performed better representations of the vessels, especially for smaller and thinner vessels.

In the second step, to compare the effect of different patch extraction approaches described before, all models are trained on both raw and preprocessed images utilising the second patch extraction approach described in Section 2.4.1. The results obtained through this step are presented in

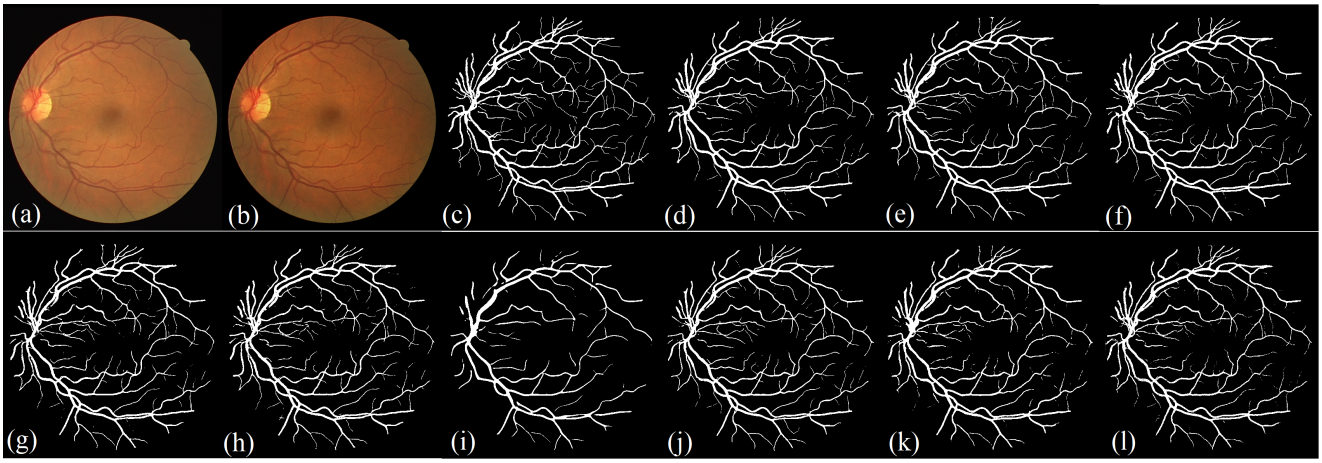

Figure 4: Qualitative comparison of the retinal vessel segmentation results on DRIVE dataset obtained utilising (a) original image, (b) preprocessed image, (c) GroundTruth, (d) UNet, (e) AttUNet, (f) ResUNet, (g)UCTransNet,(h) Unet++, (i) TransUNet-Vit, (j)TransUNet-R50, (k) UNeXt, and (l) SwinTransformer

Table 3
Accuracy measures obtained by different models trained on preprocessed $DRIVE$ dataset

Architecture	AUC -ROC	AUC -PR	F1 -Score	Acc	Sen	Spe	Stop
Unet	**0.9788**	**0.9105**	0.8124	**0.9559**	0.7501	0.9859	72
AttUNet	0.9772	0.9062	0.8087	0.9551	0.7464	0.9855	75
ResUnet	0.9779	0.9081	0.8080	0.9551	0.7416	**0.9863**	93
UCTransNet	0.9775	0.9069	**0.8138**	0.9555	0.7633	0.9836	88
Unet++	0.9787	0.9103	0.8134	**0.9559**	0.7558	0.9850	93
TransUnet -ViT-B16	0.9332	0.7939	0.6884	0.9331	0.5802	0.9846	93
TransUnet -R50-ViT-B16	0.9729	0.8967	0.7969	0.9528	0.7273	0.9857	60
Unext	0.9691	0.8852	0.8013	0.9501	**0.7901**	0.9734	103
SwinUnet	0.9722	0.8944	0.8008	0.9531	0.7407	0.9840	118

Table 4
Accuracy measures obtained by different models trained on unprocessed $DRIVE$ dataset through 56×56 pixel patch extraction technique

Architecture	AUC -ROC	AUC -PR	F1 -Score	Acc	Sen	Spe	Stop
Unet	0.9755	0.8998	**0.8083**	**0.9540**	0.7621	0.9819	90
AttUNet	**0.9756**	**0.9029**	0.8026	**0.9540**	0.7338	**0.9862**	78
ResUnet	0.9642	0.8785	0.7914	0.9504	0.7390	0.9812	180
UCTransNet	0.9747	0.8989	0.8009	0.9534	0.7363	0.9850	75
Unet++	0.9745	0.9000	0.8095	0.9542	**0.7639**	0.9820	118
TransUnet -ViT-B16	0.9404	0.8056	0.7063	0.9349	0.6143	0.9817	76
TransUnet -R50-ViT-B16	0.9746	0.8973	0.8047	0.9534	0.7536	0.9825	62
Unext	0.9734	0.8949	0.7923	0.9520	0.7186	0.9861	82
SwinUnet	0.9713	0.8925	0.7956	0.9523	0.7297	0.9847	158

Table 5
Accuracy measures obtained by different models trained on preprocessed $DRIVE$ dataset through 56×56 pixel patch extraction technique

Architecture	AUC -ROC	AUC -PR	F1 -Score	Acc	Sen	Spe	Stop
Unet	0.9780	0.9087	0.8006	0.9553	0.7415	0.9865	92
AttUNet	0.9776	0.9076	0.8058	0.9549	0.7347	**0.9870**	76
ResUnet	0.9767	0.9050	0.8040	0.9544	0.7348	0.9864	78
UCTransNet	0.9786	0.9095	0.8132	0.9558	0.7556	0.9850	69
Unet++	**0.9789**	**0.9109**	**0.8170**	**0.9564**	**0.7655**	0.9842	63
TransUnet -ViT-B16	0.9259	0.7804	0.6635	0.9301	0.5413	0.9868	87
TransUnet -R50-ViT-B16	0.9721	0.8934	0.8014	0.9528	0.7483	0.9826	49
Unext	0.9699	0.8847	0.7924	0.9507	0.7393	0.9815	66
SwinUnet	0.9714	0.8922	0.7925	0.9520	0.7195	0.9859	118

Table 4 for training on raw images and Table 5 for training on preprocessed images. Results indicate that the AttUNet has performed better while trained on raw images obtaining best values in three out of six metrics with $AUC-ROC$=0.9756, $AUC-PR$=0.9029, and Acc=0.9540, which is still slightly lower than those obtained through training models on randomly extracted 224 × 224 patches (first approach). On the other hand, UNet++ has performed best in five out of six utilised accuracy metrics when trained on patches extracted from both preprocessed images augmented through the second patch extraction approach with $AUC-ROC$=0.9789, $AUC-PR$=0.9109, $F1-Score$=0.8170, and Acc=0.9564 which indicates a slight improvement compared to the first augmentation approach.

3.2. STARE Dataset

Unlike the DRIVE dataset, STARE and other datasets considered are trained through the utilisation of the first patch extraction approach with overlapping randomly extracted 224 × 224 patches. The testing results obtained from different models trained on $STARE$ dataset are presented in Table 6 for training on raw images and Table 7 for training on preprocessed images. Results indicate that the AttUNet has performed best while trained on raw images, obtaining best values in five out of six metrics with $AUC-ROC$=0.9901, $AUC-PR$=0.9330,$F1-Score$=0.8448, Acc=0.924, and Spe=0.9894, which is slightly higher than those obtained through training models on preprocessed images with $AUC-ROC$=0.9878, $AUC-PR$=0.9237, indicating the ineffectiveness of the preprocessing performed. Additionally, the naive UNet has performed comparably with the AttUNet when trained on patches extracted from

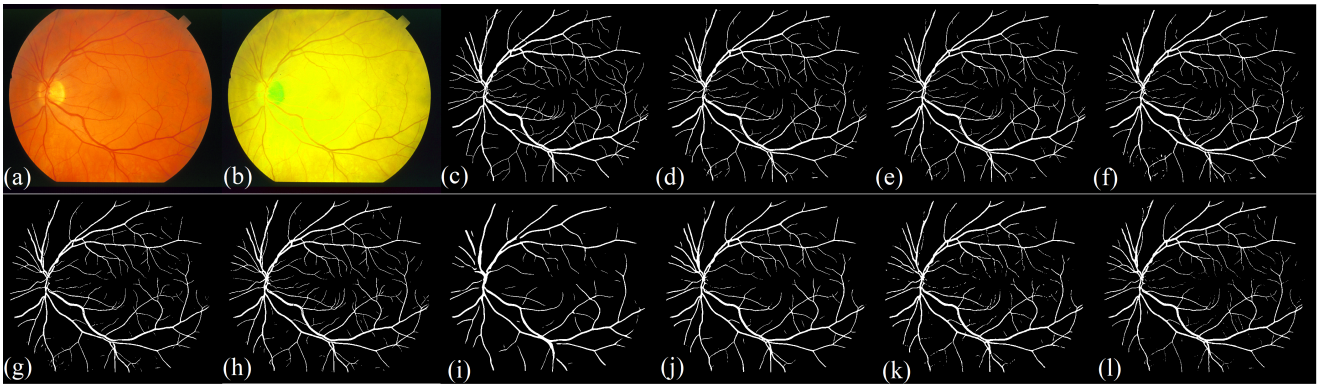

Figure 5: Qualitative comparison of the retinal vessel segmentation results on STARE dataset obtained utilising (a) original image, (b) preprocessed image, (c) GroundTruth, (d) UNet, (e) AttUNet, (f) ResUNet, (g)UCTransNet,(h) Unet++, (i) TransUNet-Vit, (j)TransUNet-R50, (k) UNeXt, and (l) SwinTransformer

Table 6
Accuracy measures obtained by different models trained on unprocessed $STARE$ dataset

Architecture	AUC -ROC	AUC -PR	F1 -Score	Acc	Sen	Spe	Stop
Unet	0.9885	0.9271	0.8427	0.9716	0.8189	0.9872	110
AttUNet	**0.9901**	**0.9330**	**0.8448**	**0.9724**	0.8068	**0.9894**	93
ResUnet	0.9867	0.9178	0.8322	0.9697	0.8069	0.9864	107
UCTransNet	0.9887	0.9277	0.8359	0.9713	0.7867	0.9902	81
Unet++	0.9863	0.9235	0.8412	0.9712	**0.8198**	0.9867	130
TransUnet -ViT-B16	0.9579	0.8199	0.7295	0.9551	0.6501	0.9864	110
TransUnet -R50-ViT-B16	0.9849	0.9222	0.8360	0.9707	0.8016	0.9881	85
Unext	0.9853	0.9143	0.8236	0.9687	0.7867	0.9873	75
SwinUnet	0.9895	0.9276	0.8389	0.9713	0.8018	0.9887	144

Table 7
Accuracy measures obtained by different models trained on preprocessed $STARE$ dataset

Architecture	AUC -ROC	AUC -PR	F1 -Score	Acc	Sen	Spe	Stop
Unet	0.9877	0.9230	**0.8347**	**0.9706**	**0.7973**	0.9884	84
AttUNet	**0.9878**	**0.9237**	0.8305	0.9705	0.7766	**0.9904**	79
ResUnet	0.9818	0.9010	.0.8120	0.9671	0.7640	0.9879	82
UCTransNet	0.9856	0.9161	0.8257	0.9695	0.7749	0.9895	81
Unet++	0.9870	0.9216	0.8365	0.9703	0.8176	0.9859	97
TransUnet -ViT-B16	0.9566	0.8219	0.7304	0.9552	0.6520	0.9864	101
TransUnet -R50-ViT-B16	0.9839	0.9080	0.8229	0.9684	0.7878	0.9870	60
Unext	0.9798	0.8921	0.7997	0.9653	0.7444	0.9879	76
SwinUnet	0.9842	0.9118	0.8249	0.9689	0.7857	0.9878	85

Table 8
Accuracy measures obtained by different models trained on unprocessed $CHASE-DB1$ dataset

Architecture	AUC -ROC	AUC -PR	F1 -Score	Acc	Sen	Spe	Stop
Unet	0.9828	0.8838	0.7972	0.9642	0.7713	0.9835	108
AttUNet	0.9825	0.8826	0.7974	0.9643	0.7694	0.9839	160
ResUnet	0.9761	0.8628	0.7834	0.9617	0.7575	0.9823	182
UCTransNet	**0.9832**	**0.8863**	0.7997	0.9647	0.7722	**0.9840**	130
Unet++	0.9815	0.8797	0.7930	0.9636	0.7634	0.9837	83
TransUnet -ViT-B16	0.9747	0.8514	0.7809	0.9614	0.7530	0.9824	115
TransUnet -R50-ViT-B16	0.9781	0.8750	**0.8258**	**0.9649**	**0.7979**	0.9817	98
Unext	0.9814	0.8769	0.7941	0.9633	0.7755	0.9821	129
SwinUnet	**0.9832**	0.8822	0.8023	0.9644	0.7910	0.9818	142

Table 9
Accuracy measures obtained by different models trained on preprocessed $CHASE-DB1$ dataset

Architecture	AUC -ROC	AUC -PR	F1 -Score	Acc	Sen	Spe	Stop
Unet	0.9806	**0.8799**	0.7927	0.9635	0.7634	0.9836	73
AttUNet	0.9805	0.8797	0.7917	0.9634	0.7603	0.9838	60
ResUnet	0.9746	0.8581	0.7639	0.9602	0.7044	**0.9859**	154
UCTransNet	0.9799	0.8769	0.7902	0.9630	0.7622	0.9832	57
Unet++	0.9791	0.8759	0.7884	0.9629	0.7571	0.9836	83
TransUnet -ViT-B16	0.9738	0.8476	0.7671	0.9590	0.7385	0.9812	58
TransUnet -R50-ViT-B16	**0.9808**	0.8763	0.7925	**0.9637**	0.7596	0.9842	40
Unext	0.9767	0.8629	0.7734	0.9608	0.7323	0.9838	84
SwinUnet	0.9807	0.8720	**0.7935**	0.9629	**0.7811**	0.9811	80

preprocessed images. Figure 5 presents a qualitative comparison of the retinal vessel segmentation results obtained from different models trained on the *STARE* dataset.

3.3. CHASE-DB1 Dataset

The testing results obtained from different models trained on $CHASE-DB1$ dataset are presented in Table 8 for training on raw images and Table 9 for training on preprocessed images. Results indicate that the *UCTransNet* and *TRansUnet-R50* have performed comparably better than other models while trained on raw images, each obtaining best values in three out of six metrics with $AUC-ROC$=0.9832, $AUC-PR$=0.8863, and Acc=0.9840. However, similar to the previous dataset, it is still slightly higher than those obtained through training models on preprocessed images where the highest accuracies obtained by the *TRansUnet-R50* equal to $AUC-ROC$=0.9808, $AUC-PR$=0.8763, and Acc=0.9637. Figure 6 presents a qualitative comparison of the retinal vessel segmentation results obtained from different models trained on the *CHASE-DB1* dataset. It is clear that almost all models have performed equally well in segmenting main vessels yet show poor performance when segmenting capillary arteries.

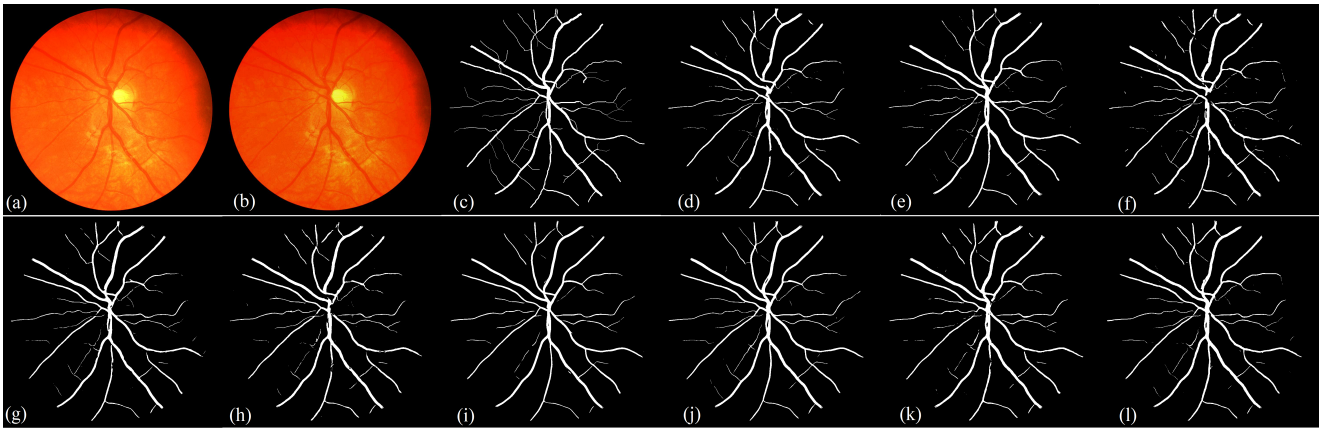

Figure 6: Qualitative comparison of the retinal vessel segmentation results on $CHASE-DB1$ dataset obtained utilising (a) original image, (b) preprocessed image, (c) GroundTruth, (d) UNet, (e) AttUNet, (f) ResUNet, (g)UCTransNet,(h) Unet++, (i) TransUNet-Vit, (j)TransUNet-R50, (k) UNeXt, and (l) SwinTransformer

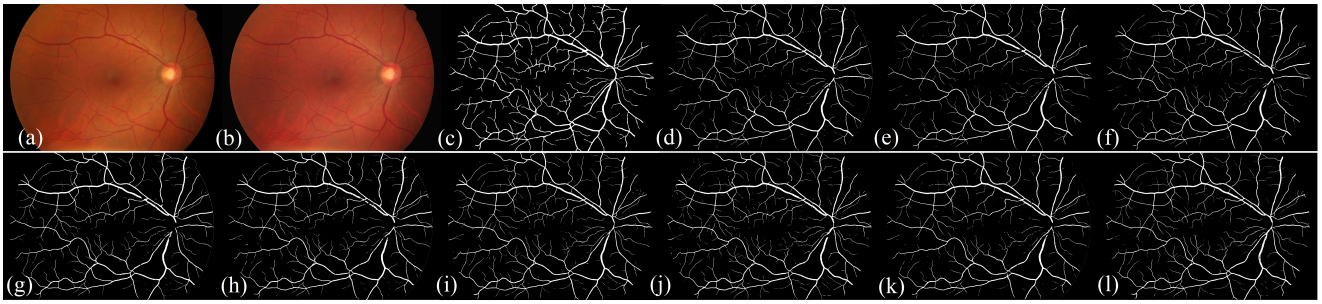

Figure 7: Qualitative comparison of the retinal vessel segmentation results on HRF dataset obtained utilising (a) original image, (b) preprocessed image, (c) GroundTruth, (d) UNet, (e) AttUNet, (f) ResUNet, (g)UCTransNet,(h) Unet++, (i) TransUNet-Vit, (j)TransUNet-R50, (k) UNeXt, and (l) SwinTransformer

3.4. HRF Dataset

Following the patch extraction ratio used for earlier datasets would require extracting an unfeasible number of 35,000 patches for the HRF dataset, which has noticeably higher image resolution than prior datasets in this study. Given the computational and memory constraints and the objectives of our investigation, we chose a more effective strategy. Early attempts to train models from scratch on this dataset resulted in poor performance (results not included). Consequently, Using pre-learned weights from models trained on raw CHASE-DB1 images, we fine-tuned the models on a subset of 3,000 random patches extracted from the HRF dataset. The results were greatly enhanced by utilising this concept of transfer learning, which decreased computational costs and improved all metrics.

The testing results obtained from different models trained on the HRF dataset are presented in Table 10 for training on raw images and Table 11 for training on preprocessed images. Results indicate that the AttUNet has performed best while trained on raw images, obtaining best values in four out of six metrics with $AUC-ROC$=0.9825, $AUC-PR$=0.8937, Acc=0.9540, and Spe=0.9885 which is generally higher than those obtained through training models preprocessed images where $TransUnet-R50$ has outputted better results compared to other models with

Table 10
Accuracy measures obtained by different models trained on unprocessed HRF dataset

Architecture	AUC -ROC	AUC -PR	F1 -Score	Acc	Sen	Spe	Stop
Unet	**0.9825**	0.8902	**0.7966**	0.9642	0.7513	0.9861	40
AttUNet	**0.9825**	**0.8937**	0.7939	**0.9646**	0.7317	**0.9885**	41
ResUnet	0.9794	0.8846	0.7862	0.9634	0.7204	0.9884	50
UCTransNet	0.9820	0.8882	0.7944	0.9639	0.7464	0.9863	55
Unet++	0.9820	0.8909	0.7965	0.9644	0.7468	0.9868	58
TransUnet -ViT-B16	0.9727	0.8515	0.7824	0.9614	0.7439	0.9838	35
TransUnet -R50-ViT-B16	0.9805	0.8828	0.7914	0.9629	**0.7539**	0.9844	14
Unext	0.9812	0.8879	0.7879	0.9634	0.7279	0.9877	58
SwinUnet	0.9821	0.8881	0.7951	0.9639	0.7499	0.9860	30

$AUC-ROC$=0.9682, and Sen=0.72. This indicates a considerable drop in the accuracy of models trained on preprocessed images through gamma correction and random HSV adjustment. Figure 7 presents a qualitative comparison of the retinal vessel segmentation results obtained from different models trained on the *HRF* dataset.

4. Discussion

This section provides a discussion of the results obtained through this study. Regarding the efficiency of the

Table 11
Accuracy measures obtained by different models trained on preprocessed HRF dataset

Architecture	AUC -ROC	AUC -PR	F1 -Score	Acc	Sen	Spe	Stop
Unet	0.9670	0.8544	0.7623	0.9601	0.6877	0.9880	51
AttUNet	0.9640	0.8432	0.7555	0.9589	0.6830	0.9872	42
ResUnet	0.9570	0.8246	0.7268	0.9558	0.6328	**0.9889**	73
UCTransNet	0.9669	0.8523	0.7648	**0.9602**	0.6940	0.9876	36
Unet++	0.9675	**0.8554**	0.7620	0.9601	0.6874	0.9880	69
TransUnet -ViT-B16	0.9609	0.8334	0.7548	0.9579	0.6949	0.9849	14
TransUnet -R50-ViT-B16	**0.9682**	0.8550	0.7707	0.9600	**0.7200**	0.9847	16
UNeXt	0.9656	0.8490	**0.7577**	0.9592	0.6855	0.9873	43
SwinUNet	0.9297	0.7186	0.6545	0.9440	0.5732	0.9820	88

preprocessing pipeline adopted in this study, which includes Gamma corrections and random HSV adjustment, it could be concluded that its positive impact is limited to the low-resolution images of the *DRIVE* dataset where it has led to an improvement of the segmentation accuracy obtained by all models as well as faster convergence of the training procedure and requirement for less training epoch. However, it had an adverse effect on the performance and negligible to no effect on model convergence of models trained on other datasets with higher resolutions.

From the model performance point of view, it could be summarised that the convolutional architectures, such as naive UNet and UNet++, have performed best while trained on low-resolution images of the *DRIVE* dataset. For all other datasets, transformers such as AttUNet, TransUNet and SwinUNet have performed better than convolutional-based models. Overall, the TransUNet-ViT-B16 performed worst among all other models, while AttUNet and TransUNet-R50 performed better than the rest while trained on all datasets except the *DRIVE*.

Table 12 compares the results of this study with those obtained by different researchers. It could be concluded that while we acknowledge that our study did not outperform the state-of-the-art studies in the remaining three datasets, it did achieve slightly better results in one of them, the *STARE* dataset. However, the results we were able to acquire for the rest of the datasets are still comparable, demonstrating competitive performance in line with existing benchmarks in the field.

5. Conclusions

In conclusion, our study delves into the complicated task of retinal vessel segmentation, employing nine different Convolutional-based and transformer-based architectures. Transformers, such as AttUNet, TransUNet, and SwinUNet, perform exceptionally well on datasets with greater resolutions, while convolutional models, like naive UNet and UNet++, perform better on datasets with lower resolutions. Preprocessing methods have differing effects; for example, random HSV adjustment and gamma correction work well for low-resolution photos but are ineffective when used with higher-resolution images. Our findings indicate comparable performance with the state-of-the-art benchmarking of the field and minor improvements in the STARE dataset compared to other research. The conclusions highlight the importance and effectiveness of image-enhancing methods on improving model accuracy and reducing training time, especially for low-resolution datasets.

CRediT authorship contribution statement

Morteza Tavakol Sadrabadi: <Credit authorship details>. **Hamed Agahi:** .

References

[1] Qing Qin and Yuanyuan Chen. A review of retinal vessel segmentation for fundus image analysis. *Engineering Applications of Artificial Intelligence*, 128:107454, 2024.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[3] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net, 2022.

[4] Juntang Zhuang. Laddernet: Multi-path networks based on u-net for medical image segmentation, 2019.

[5] Zhun Fan, Jiajie Mo, Benzhang Qiu, Wenji Li, Guijie Zhu, Chong Li, Jianye Hu, Yibiao Rong, and Xinjian Chen. Accurate retinal vessel segmentation via octave convolution neural network, 2020.

[6] Liangzhi Li, Manisha Verma, Yuta Nakashima, Hajime Nagahara, and Ryo Kawasaki. Iternet: Retinal image segmentation utilizing structural redundancy in vessel networks. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3645–3654, 2020.

[7] Changlu Guo, Márton Szemenyei, Yangtao Hu, Wenle Wang, Wei Zhou, and Yugen Yi. Channel attention residual u-net for retinal vessel segmentation. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1185–1189, 2021.

[8] Yan Hu, Zhongxi Qiu, Dan Zeng, Li Jiang, Chen Lin, and Jiang Liu. Supervessel: Segmenting high-resolution vessel from low-resolution retinal image. In Shiqi Yu, Zhaoxiang Zhang, Pong C. Yuen, Junwei Han, Tieniu Tan, Yike Guo, Jianhuang Lai, and Jianguo Zhang, editors, *Pattern Recognition and Computer Vision*, pages 178–190, Cham, 2022. Springer Nature Switzerland.

[9] M.Jasmin Pemeena Priyadarsini, Sowmiya S, A Jabeena, G.K. Rajini, Ganesan Subramanian, and Ernest Bravin Clinton S. Retinal vessel segmentation using unet++. In *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, pages 1–5, 2023.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[12] Changlu Guo, Márton Szemenyei, Yugen Yi, Wenle Wang, Buer Chen, and Changqi Fan. Sa-unet: Spatial attention u-net for retinal vessel segmentation, 2020.

Table 12
Comparison of the accuracy measures obtained in this study with relevant studies for different datasets

Dataset	*Paper*	*Model*	*Year*	*AUC-ROC*	*AUC-PR*	*F1-Score*	*Acc*	*Sen*	*Spe*
DRIVE	[36]	V-GAN	2017	0.9803	0.9149	0.829	-	-	-
	[4]	Ladder-Net	2018	0.9793	-	0.8202	0.9561	0.7856	0.981
	[15]	CSAU	2019	0.9807	**0.9157**	0.8294	0.9563	0.8349	-
	[37]	Unet	2019	0.9794	-	0.8169	0.9559	0.7728	0.9826
	[12]	SA-Unet	2020	0.9864	-	-	0.9698	0.8212	0.984
	[7]	CAR-Unet	2020	0.9852	-	-	0.9699	0.8135	0.9849
	[5]	Octave-Unet	2020	0.9835	-	0.8127	0.9664	0.8374	0.979
	[14]	RV-GAN	2020	**0.9887**	-	**0.869**	**0.979**	0.7927	**0.9969**
	[19]	Unet	2021	0.9855	-	-	0.9712	-	-
	[20]	Unet	2021	0.9788	-	0.8209	0.9545	0.8227	0.9741
	[21]	Unet	2021	0.9886	-	0.8316	0.9705	**0.838**	0.9834
	Our	Unet	2023	0.9788	0.9105	0.8124	0.9559	0.7501	0.9859
STARE	[36]	V-GAN	2017	0.9838	0.9167	0.834	-	-	-
	[37]	Unet	2019	0.9846	-	0.8219	0.9638	0.7739	0.9867
	[15]	CSAU	2019	0.9834	0.9206	0.8435	0.9673	0.8465	-
	[7]	CAR-Unet	2020	**0.9911**	-	-	0.9743	0.8445	0.985
	[5]	Octave-Unet	2020	0.9875	-	0.8191	0.9713	**0.8664**	0.9798
	[14]	RV-GAN	2020	0.9887	-	0.8323	**0.9754**	0.8356	0.9864
	[20]	Unet	2021	0.9893	-	0.8082	0.9683	0.8908	0.9745
	Our	AttUNet	2023	0.9901	**0.933**	**0.8448**	0.9724	0.8068	**0.9894**
CHASE-DB1	[4]	Ladder-Net	2018	0.9839	-	0.8031	0.9656	0.7978	0.9818
	[37]	Unet	2019	0.9784	-	0.7815	0.96	0.724	**0.986**
	[12]	SA-Unet	2020	0.9905	-	-	0.9755	0.8573	0.9835
	[7]	CAR-Unet	2020	0.9898	-	-	0.9751	0.8439	0.9839
	[5]	Octave-Unet	2020	0.9905	-	0.8313	**0.9759**	0.867	0.984
	[14]	RV-GAN	2020	0.9914	-	**0.8957**	0.9697	0.8199	0.9806
	[20]	Unet	2021	0.9838	-	0.7565	0.9612	**0.8818**	0.9673
	[21]	Unet	2021	**0.992**	-	0.8271	0.9771	0.869	0.9843
	Our	UCTransNet	2023	0.9832	**0.8863**	0.7997	0.9647	0.7722	0.984
	Our	SwinUnet	2023	0.9832	0.8822	0.8023	0.9644	0.791	0.9818
HRF	[15]	CSAU	2019	**0.9867**	**0.9047**	**0.8171**	-	**0.8303**	-
	[7]	Octave-Unet	2020	0.9845	-	0.7963	**0.9698**	0.8076	0.9831
	[8]	SuperVessel	2022	0.8506	-	0.7674	0.9654	0.7229	-
	Our	Unet	2023	0.9825	0.8902	0.7966	0.9642	0.7513	0.9861
	Our	AttUNet	2023	0.9825	0.8937	0.7939	0.9646	0.7317	**0.9885**

[13] Yanzhou Su, Jian Cheng, Guiqun Cao, and Haijun Liu. How to design a deep neural network for retinal vessel segmentation: an empirical study. *Biomedical Signal Processing and Control*, 77:103761, 2022.

[14] Sharif Amit Kamran, Khondker Fariha Hossain, Alireza Tavakkoli, Stewart Lee Zuckerbrod, Kenton M. Sanders, and Salah A. Baker. *RV-GAN: Segmenting Retinal Vascular Structure in Fundus Photographs Using a Novel Multi-scale Generative Adversarial Network*, page 34–44. Springer International Publishing, 2021.

[15] Ruirui Li, Mingming Li, Jiacheng Li, and Yating Zhou. Connection sensitive attention u-net for accurate retinal vessel segmentation, 2019.

[16] Kun Sun, Yang Chen, Yi Chao, Jiameng Geng, and Yinsheng Chen. A retinal vessel segmentation method based improved u-net model. *Biomedical Signal Processing and Control*, 82:104574, 2023.

[17] Zhendi Ma and Xiaobo Li. An improved supervised and attention mechanism-based u-net algorithm for retinal vessel segmentation. *Computers in Biology and Medicine*, 168:107770, 2024.

[18] Zhaolei Wang, Junbin Lin, Ruixuan Wang, and Weishi Zheng. Data augmentation is more important than model architectures for retinal vessel segmentation. In *Proceedings of the 2019 International Conference on Intelligent Medicine and Health*, ICIMH 2019, page 48–52, New York, NY, USA, 2019. Association for Computing Machinery.

[19] Enes Sadi Uysal, M. Şafak Bilici, B. Selin Zaza, M. Yiğit Özgenç, and Onur Boyar. Exploring the limits of data augmentation for retinal vessel segmentation, 2021.

[20] Xu Sun, Huihui Fang, Yehui Yang, Dongwei Zhu, Lei Wang, Junwei Liu, and Yanwu Xu. *Robust Retinal Vessel Segmentation from a Data Augmentation Perspective*, page 189–198. Springer International Publishing, 2021.

[21] Yuqian Zhou, Hanchao Yu, and Humphrey Shi. Study group learning: Improving retinal vessel segmentation trained with noisy labels, 2021.

[22] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018.

[23] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.

[24] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R. Zaiane. Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer, 2022.

[25] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation, 2021.

[26] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer Syeda-Mahmood, Anne Martel, Lena Maier-Hein, João Manuel R.S. Tavares, Andrew Bradley, João Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, and Anant Madabhushi, editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11, Cham, 2018. Springer International Publishing.

[27] Jeya Maria Jose Valanarasu and Vishal M. Patel. Unext: Mlp-based rapid medical image segmentation network. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 23–33, Cham, 2022. Springer Nature Switzerland.

[28] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, pages 205–218, Cham, 2023. Springer Nature Switzerland.

[29] J. Staal, M.D. Abramoff, M. Niemeijer, M.A. Viergever, and B. van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509, 2004.

[30] A.D. Hoover, V. Kouznetsova, and M. Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging*, 19(3):203–210, 2000.

[31] Christopher G. Owen, Alicja R. Rudnicka, Robert Mullen, Sarah A. Barman, Dorothy Monekosso, Peter H. Whincup, Jeffrey Ng, and Carl Paterson. Measuring Retinal Vessel Tortuosity in 10-Year-Old Children: Validation of the Computer-Assisted Image Analysis of the Retina (CAIAR) Program. *Investigative Ophthalmology and Visual Science*, 50(5):2004–2010, 05 2009.

[32] A. Budai, R. Bock, A. Maier, J. Hornegger, and G. Michelson. Robust vessel segmentation in fundus images. *Int J Biomed Imaging*, 2013:154860, 2013. 1687-4196.

[33] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020.

[34] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

[36] Jaemin Son, Sang Jun Park, and Kyu-Hwan Jung. Retinal vessel segmentation in fundoscopic images with generative adversarial networks, 2017.

[37] Zhengyuan Liu. Retinal vessel segmentation based on fully convolutional networks, 2019.