

Comparative Study On Optical Coherence Tomography (OCT) Retinal Disease Segmentation

¹Dhiraj Pyakurel, ¹Yokisha Poudel, ¹Sushiksha Prasai, ¹Shailee Acharya, ¹Tisa Shrestha, ¹Amit Kumar Rauniyar

¹Department of Electronics and Computer Engineering, Advanced College of Engineering and Management, Tribhuvan University, Nepal

Email: dhiraj@acem.edu.np, amitkumar.rauniyar@acem.edu.np

Abstract

Optical Coherence Tomography (OCT) is a non-invasive imaging technique that generates high-resolution cross-sectional images of the retina, and has become a critical modality for the diagnosis of retinal diseases. Diabetic Macular Edema (DME) represents one of the most clinically prevalent such conditions, causing significant vision loss and requiring early and precise diagnosis, particularly in resource-constrained settings such as Nepal, where no AI-assisted OCT analysis tool currently exists. This study presents a systematic comparative evaluation of three deep learning-based segmentation architectures — U-Net, U-Net++, and Y-Net — for automated identification of DME and Intraretinal Fluid (IRF) regions in OCT scans. A dataset of retinal OCT images was manually annotated using the Computer Vision Annotation Tool (CVAT) and augmented with geometric transformations to improve model generalization. Models were trained using a hybrid Focal Tversky and Focal Cross-Entropy loss and evaluated on a held-out validation set using standard segmentation metrics: Dice Similarity Coefficient (DSC), Intersection over Union (IoU), precision, and recall. Among the evaluated models, Y-Net achieved the best overall performance, attaining an F1 score of 0.8651, IoU of 0.8142, precision of 0.9210, and recall of 0.8424, outperforming both U-Net++ (F1: 0.8432, IoU: 0.7925) and the baseline U-Net (F1: 0.8246, IoU: 0.7599). These results demonstrate the feasibility of deep learning-based OCT segmentation as a diagnostic support tool in resource-constrained clinical environments.

Keywords: Diabetic Macular Edema, Intraretinal Fluid, Optical Coherence Tomography, Retinal Disease Segmentation, Deep Learning

I. INTRODUCTION

Retinal diseases are among the leading causes of irreversible vision loss worldwide, with Diabetic Macular Edema (DME) representing one of the most clinically prevalent conditions, particularly in aging populations. Optical Coherence Tomography (OCT) has emerged as the gold-standard non-invasive imaging modality in ophthalmology, offering high-resolution, cross-sectional visualization of retinal layers that is critical for early detection and longitudinal monitoring of such diseases [1, 2]. Despite its diagnostic value, the interpretation of OCT images in clinical practice continues to rely predominantly on manual evaluation by trained specialists

— a process that is time-intensive, subject to inter-observer variability, and increasingly unsustainable in the face of growing patient volumes.

The integration of deep learning into medical image analysis has demonstrated substantial promise in automating and augmenting diagnostic workflows. Encoder-decoder architectures, most notably U-Net, have become foundational to biomedical image segmentation owing to their ability to preserve spatial information through skip connections while learning hierarchical feature representations [3–6]. Building upon this foundation, U-Net++ introduced densely connected skip pathways and redesigned feature aggregation to improve segmentation of irregularly shaped structures across multiple scales [7, 8]. Inspired by dual-branch and multi-task encoder-decoder designs explored in the literature [9, 10], we adopt a task-specific dual-decoder Y-Net variant in which a shared encoder feeds two separate decoder branches — one dedicated to IRF segmentation and one to DME segmentation — enabling richer class-specific feature extraction than single-pathway models. These architectures collectively represent the state of the art in retinal image segmentation; however, a rigorous and systematic comparative evaluation of all three on OCT-based DME segmentation remains absent from the literature.

This gap is particularly significant in the Nepali clinical context. Consultations with ophthalmologists at regional hospitals confirmed that DME is among the most frequently encountered retinal conditions in daily practice, and that no AI-assisted tool currently exists to support OCT image segmentation in these settings. The absence of such tools amplifies the diagnostic burden on clinicians and limits the standardization and scalability of retinal disease management [11].

To address this gap, this paper presents a comparative study of U-Net, U-Net++, and Y-Net for the automated segmentation of DME and Intraretinal Fluid (IRF) in OCT images, evaluated using standard metrics including Dice Similarity Coefficient (DSC), Intersection over Union (IoU), precision, and recall. The primary contributions of this work are as follows:

- A systematic comparative evaluation of three encoder-decoder architectures — U-Net, U-Net++, and Y-Net — on OCT-based DME segmentation under identical experimental conditions.
- An analysis of each model’s segmentation quality, architectural complexity, and training efficiency to provide clinically and computationally actionable insights.
- A demonstration of the feasibility of AI-assisted OCT image segmentation as a diagnostic support tool in resource-constrained clinical environments.

The remainder of this paper is organized as follows. Section 2 reviews related work in deep learning-based retinal image segmentation. Section 3 describes the dataset, preprocessing pipeline, and model architectures. Section 4 presents the experimental results, comparative analysis, and discussion of findings, followed by the conclusion in Section 5.

II. LITERATURE REVIEW

Medical image segmentation using deep learning has evolved rapidly over the past decade, with encoder-decoder architectures emerging as the dominant paradigm for pixel-wise classification tasks. Recent uncertainty-aware segmentation frameworks have further improved the robustness of medical image analysis by integrating uncertainty estimation into deep learning pipelines (Bhatt and Shakya 2025). This section synthesizes existing research across three thematic clusters relevant to the present study: foundational segmentation architectures, OCT-specific retinal fluid segmentation methods, and multi-task or frequency-aware extensions of standard models.

A. Foundational Encoder-Decoder Architectures

The encoder-decoder architecture was established as a cornerstone of biomedical image segmentation by Ronneberger et al., whose U-Net model demonstrated that a symmetric contracting-expansive path with skip connections could achieve precise localization even from limited annotated data (Ronneberger et al. 2015). By concatenating high-resolution feature maps from the encoder to corresponding decoder layers, U-Net effectively preserves spatial detail while capturing semantic context — a property especially valuable when segmenting small and structurally complex regions such as retinal lesions. Its simplicity, parameter efficiency, and strong generalization have made it the de facto baseline for virtually all subsequent medical segmentation architectures.

Building directly on this foundation, Zhou et al. proposed U-Net++, which addresses a fundamental limitation of U-Net: the semantic gap between encoder and decoder feature maps of corresponding resolution (Zhou et al. 2018). U-Net++ resolves this through nested and densely connected skip pathways, inserting intermediate convolutional layers between encoder and decoder nodes to progressively reconcile low-level and high-level representations. Combined with deep supervision at multiple decoder depths to stabilize gradient flow, U-Net++ demonstrated statistically significant improvements over U-Net across diverse medical imaging modalities — including lung nodule segmentation in CT, liver segmentation, nuclei segmentation in microscopy, and colon polyp detection — with average IoU gains of 3.4–3.9 points (Zhou et al. 2018). Despite these gains, U-Net++ has primarily been validated on non-retinal datasets, and its performance on OCT-based fluid segmentation, where lesion boundaries are subtle and diffuse, has not been systematically studied.

Multi-scale context fusion has been identified as another productive direction. A U-Net-based method employing a scale-aware feature fusion module assigned differential weights to features extracted at multiple encoding scales, improving segmentation of histopathological images without proportional increases in computational cost (Yuan and Cheng 2024; Lamichhane et al. 2026; Baral et al. 2025). While this approach demonstrated the value of scale-aware feature integration, it was designed for tissue-level histopathology rather than fluid-layer segmentation in OCT,

limiting its direct applicability.

B. OCT Retinal Fluid and DME Segmentation

Segmentation of retinal fluid in OCT images presents distinct challenges relative to general medical image segmentation: fluid pockets are often thin, irregularly shaped, and distributed across multiple retinal layers with low-contrast boundaries. Several dedicated methods have attempted to address these characteristics. An order-constrained regression approach demonstrated that predicting retinal layer positions as a regression problem — while enforcing anatomical ordering constraints — yielded sub-pixel accuracy in layer boundary delineation and strong correlation with expert-annotated fluid volumes (Morelle et al. 2023). This regression-based paradigm offers scalability to large OCT populations, though it is primarily oriented toward layer boundary detection rather than fluid region segmentation per se.

A U-Net-based model for retinal sublayer segmentation using a ResNet34 encoder reported high Dice Similarity Coefficients across retinal layers, confirming the encoder-decoder architecture's suitability for this domain (Yojana and Rani 2023). However, validation was limited to a single dataset, raising concerns about generalizability across scanner types and disease severity levels.

To improve fluid segmentation accuracy for both DME and Age-related Macular Degeneration (AMD), a precision segmentation model incorporating residual blocks, inception modules, and multiscale attention mechanisms was proposed, taking both enhanced OCT images and edge maps as dual inputs (Karn and Abdulla 2024). The use of structural edge information as a secondary input stream is a methodologically notable choice that partially compensates for the indistinct boundaries characteristic of retinal fluid. Similarly, the MCFAR-Net architecture introduced a two-stage Context Feature Enhance Module (CFEM) pipeline, where contrastive loss was used in the first stage to prioritize multi-scale feature maps for thick fluid regions, with the second stage refining segmentation of fine-grained fluid pockets (Geetha Pavani et al. 2024). A Feature Synthesizer module then fused probabilistic outputs from both stages to reduce misclassification errors. While MCFAR-Net reports strong performance in early-stage DME detection, the architecture's complexity introduces additional training overhead and its evaluation is limited to a single domain.

C. Y-Net Variants and Frequency-Aware Segmentation

The Y-Net architecture has been instantiated in several distinct forms across the literature, reflecting two different conceptual lineages. In one lineage, Y-Net is defined as a dual-branch structure designed for simultaneous segmentation and classification. Mehta et al. applied this formulation to breast biopsy whole-slide images, demonstrating that joint optimization of pixel-level tissue segmentation and image-level diagnostic classification yielded a 7% accuracy improvement over state-of-the-art classification baselines (Mehta et al. 2018). The sum-skip-concatenate

operation used in this variant offers memory efficiency over standard concatenation-based skip connections, and compatibility with pre-trained backbone networks adds transferability. However, this formulation is task-specific to classification-coupled segmentation and has not been applied to OCT fluid segmentation.

In a second lineage, Farshad et al. proposed a spatio-spectral Y-Net for retinal OCT segmentation employing parallel spatial and spectral encoding pathways: one branch uses standard convolutions to learn spatial features while the other applies Fourier transforms to capture global frequency-domain representations, with both streams merged in a shared decoder (Farshad et al. 2022). This design explicitly exploits the observation that retinal fluid regions exhibit characteristic spectral signatures, achieving a 13% relative improvement in fluid segmentation Dice score over U-Net with fewer parameters. Ablation studies confirmed that spectral-domain features are the primary driver of Y-Net's gains over the spatial-only baseline.

A further extension, Denoising Y-Net (DeY-Net), appended an auxiliary denoising decoder to the Y-Net backbone combined with Denoising Test-Time Adaptation (DeTTA) to address domain generalization across heterogeneous acquisition environments (Wen et al. 2023). This work identifies domain shift as a persistent and underexplored challenge in medical image segmentation, particularly for datasets acquired across different OCT scanner manufacturers. A more recent multi-scale analysis-driven reformulation of Y-Net incorporated wavelet-based frequency decomposition to explicitly model high- and low-frequency image components, outperforming U-Net, U-Net++, and other baselines across multiple benchmark datasets (Meng et al. 2023). Despite these results, this variant has not been evaluated on retinal OCT data.

The Y-Net variant adopted in the present study is distinct from both lineages described above. Rather than employing dual encoding pathways (as in Farshad et al. (Farshad et al. 2022)) or coupling segmentation with image-level classification (as in Mehta et al. (Mehta et al. 2018)), our architecture uses a single shared encoder feeding two separate, task-specific decoder branches: one dedicated to IRF segmentation and one to DME segmentation. This design is motivated by the observation that IRF and DME, while spatially co-occurring, exhibit sufficiently distinct morphological characteristics to benefit from independent decoding pathways. We adopt the Y-Net naming convention to reflect the branching topology of this shared-encoder, dual-decoder structure, and distinguish it from the spatio-spectral and classification-coupled variants in the literature.

D. Research Gap

The body of literature reviewed above reveals two converging limitations. First, while U-Net (Ronneberger et al. 2015), U-Net++ (Zhou et al. 2018), and various Y-Net formulations (Farshad et al. 2022; Meng et al. 2023; Mehta et al. 2018) have each

demonstrated strong segmentation performance across a range of medical imaging tasks, no study has conducted a controlled, head-to-head comparative evaluation of all three architectures on the same OCT dataset under identical experimental conditions. Existing comparisons are either pairwise, conducted across different datasets, or embedded within studies that primarily propose a novel architecture rather than benchmark existing ones (Geetha Pavani et al. 2024; Karn and Abdulla 2024; Yuan and Cheng 2024). This makes it difficult to draw reliable conclusions about the relative strengths, computational trade-offs, and failure modes of each model for OCT-based DME segmentation specifically. Second, the clinical deployment context — particularly in low-resource settings such as Nepal, where no AI-assisted OCT analysis tool currently exists — has received no attention in the prior literature. The present study addresses both gaps by providing a systematic, reproducible comparative evaluation of U-Net, U-Net++, and Y-Net for DME segmentation in OCT images, offering insights directly relevant to resource-constrained clinical adoption.

III. METHODOLOGY

A. System Model

The system is designed to detect Diabetic Macular Edema (DME) and Intraretinal Fluid (IRF) from OCT scans using deep learning.

1. Data Sources and Tools

- **Data Source:** The *Retinal OCT Image Classification – C8* dataset published by Obuli Sai Naren on Kaggle in 2021 was used as the primary data source.
- **Annotation Tools:** The Computer Vision Annotation Tool (CVAT) was used to manually annotate regions of interest in the images.
- **Software Stack:** The system was developed using Python, PyTorch, NumPy, and OpenCV, and was implemented and tested using VS Code and Google Colab.
- **Hardware Requirements:** An NVIDIA GPU with a minimum of 8 GB of VRAM is required to efficiently train and run the models.

2. System Assumptions

- **Expert Accuracy:** Expert annotations are assumed to be correct and are used as ground truth.
- **Intensity Distribution:** Preprocessing steps are assumed not to remove or distort important structural details of retinal layers.
- **Label Overlap:** IRF and DME regions may spatially co-occur within the same pixel in OCT scans. The segmentation task is therefore treated as multi-label rather than mutually exclusive multi-class, which motivates the use of independent sigmoid activations per output channel rather than a shared softmax. Each channel is thresholded independently at inference time.

3. Operational Process for Reproducibility

To ensure reproducibility, the following steps are followed:

- **Standardization:** All images are resized to a fixed resolution of 256×256 pixels using Lanczos interpolation for OCT images and nearest-neighbor interpolation for segmentation masks to preserve label integrity. Aspect ratio is preserved during resizing using the scaling factor:

$$s = \min\left(\frac{256}{W}, \frac{256}{H}\right)$$

where W and H are the original width and height, and s is the scaling factor used to derive the resized dimensions $W' = sW$ and $H' = sH$. Zero-padding is then applied symmetrically to reach exactly 256×256 pixels where the scaled dimensions fall short. Following padding, pixel intensities are normalized per image to the range $[0, 1]$ by subtracting the per-image minimum and dividing by the per-image intensity range. Augmentation is applied only after normalization, exclusively to the training split.

- **Augmentation:** The training split was expanded from 280 to approximately 2,500 images using a combination of geometric and photometric transformations, including random rotation, horizontal flipping, translation, shearing, and brightness and contrast perturbations. These transformations were applied jointly to each image and its corresponding mask, with geometric operations using bilinear interpolation for OCT images and nearest-neighbor interpolation for masks to preserve label integrity. This can be represented as:

$$(I_a, S_a) = A(I_p, S)$$

where I_p is the resized and normalized image, S is its corresponding mask, and $A(\cdot)$ represents the augmentation operation. The validation and test splits were not augmented.

- **Architectural Deployment:** Three segmentation models are evaluated: standard U-Net, Y-Net (task-specific dual-decoder variant, as described in Section 2), and U-Net++ with CBAM attention mechanisms. CBAM was incorporated into U-Net++ to evaluate whether attention mechanisms complement its dense skip pathway design; U-Net and Y-Net were evaluated in their standard forms to reflect their original architectural intent.
- **Optimization:** The system minimizes a hybrid loss function combining Focal Tversky Loss and Focal Cross-Entropy Loss.

4. Objective

The total loss function is defined as:

$$L = L_{FT} + L_{FCE}$$

The Focal Tversky Loss is defined as:

$$TI = \frac{TP}{TP + \alpha FN + \beta FP}$$

$$L_{FT} = (1 - TI)^\gamma$$

where α and β control the relative penalty on false negatives and false positives respectively, and γ is the focusing parameter. The Focal Cross-Entropy Loss is defined as:

$$L_{FCE} = - (1 - p_t)^\gamma \log(p_t)$$

The final segmentation output is obtained by applying independent sigmoid activations to each output channel:

$$\hat{y}_c = \sigma(f_c(x; \theta)), \quad c \in \{background, IRF, DME\}$$

where $f_c(x; \theta)$ is the network's logit output for channel c and $\sigma(\cdot)$ denotes the sigmoid function. Because IRF and DME regions can spatially co-occur in OCT images, the three output channels are treated as independent binary predictions rather than a mutually exclusive multi-class problem; sigmoid is therefore the appropriate activation. At inference time, each channel is thresholded at 0.5 to produce the final binary mask per class. The model parameters are optimized as:

$$\theta^* = \arg \min_{\theta} L$$

The model is trained until convergence using an early stopping mechanism. Using this system, Diabetic Macular Edema and Intraretinal Fluid can be effectively detected from OCT scans.

B. Dataset

The dataset used in this study is the *Retinal OCT Image Classification – C8* dataset published by Obuli Sai Naren on Kaggle in 2021, comprising OCT images of patients with Diabetic Macular Edema. A total of 400 DME-positive images were selected and manually annotated at the pixel level using CVAT. All 400 images contain both IRF and DME regions simultaneously, confirming the multi-label nature of the task. Each annotated image corresponds to a distinct patient, ensuring that no patient appears across multiple splits.

The 400 annotated images were randomly shuffled and partitioned into three non-overlapping, patient-independent splits prior to any augmentation: 280 images (70%) for training, 60 images (15%) for validation, and 60 images (15%) for testing. Because each image originates from a distinct patient, this image-level split is equivalent to a patient-level split, and no risk of data leakage across splits exists. To improve model generalization, the training split was subsequently expanded from 280 to approximately 2,500 images through data augmentation as described above. The validation and test splits were held fixed at 60 images each and were not augmented.

Prior to training, all images were resized and zero-padded to 256×256 pixels and normalized per image to $[0, 1]$ as described above. A representative sample is shown in Fig. 1.

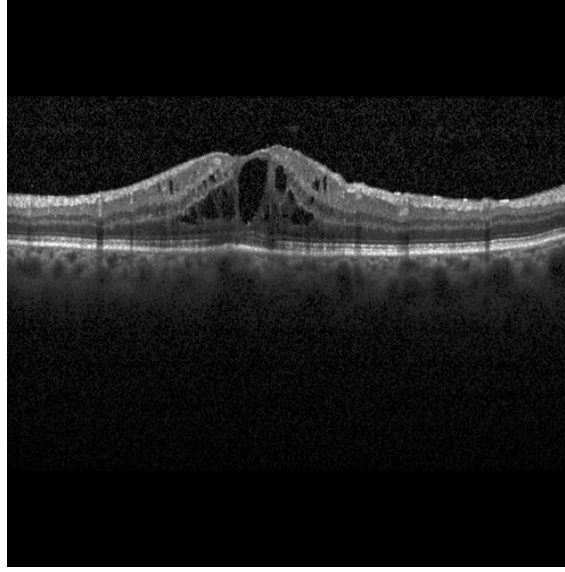


Fig. 1: Sample OCT retinal image with DME from the dataset.

C. Algorithm

Algorithms 1–4 describe the forward passes of each segmentation architecture and the shared training loop.

Algorithm 1: Forward Pass of Standard U-Net for OCT Segmentation

Input: x : 2D OCT scan tensor of shape $(B, 1, H, W)$
Output: $\hat{y}$: Predicted segmentation mask of shape $(B, 3, H, W)$

1 **Notation:** $\mathcal{F}_{DC}(\cdot)$: DoubleConv block (Conv $\rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ Conv $\rightarrow$ BN $\rightarrow$ ReLU); $\downarrow$: 2×2 Max Pooling; $\uparrow$: 2×2 Transposed Convolution; $[\cdot, \cdot]$: Channel-wise concatenation;
// Encoder (Downsampling Path)
2 $d_1 \leftarrow \mathcal{F}_{DC}(x)$;
3 $d_2 \leftarrow \mathcal{F}_{DC}(\downarrow d_1)$;
4 $d_3 \leftarrow \mathcal{F}_{DC}(\downarrow d_2)$;
5 $d_4 \leftarrow \mathcal{F}_{DC}(\downarrow d_3)$;
// Bottleneck
6 $b \leftarrow \mathcal{F}_{DC}(\downarrow d_4)$;
// Decoder (Upsampling Path with Skip Connections)
7 $u_1 \leftarrow \mathcal{F}_{DC}([\uparrow b, d_4])$;
8 $u_2 \leftarrow \mathcal{F}_{DC}([\uparrow u_1, d_3])$;
9 $u_3 \leftarrow \mathcal{F}_{DC}([\uparrow u_2, d_2])$;
10 $u_4 \leftarrow \mathcal{F}_{DC}([\uparrow u_3, d_1])$;
// Classification Head
11 $\hat{y} \leftarrow \text{Conv}_{1 \times 1}(u_4)$;
12 **return** $\hat{y}$

Algorithm 2: Forward Pass of Y-Net (Task-Specific Dual-Decoder)

Input: x : 2D OCT scan tensor of shape $(B, 1, H, W)$
Output: $\hat{y}$: Combined multi-label prediction (Background, IRF, DME)

1 **Notation:** Definitions for $\mathcal{F}_{DC}(\cdot)$, $\downarrow$, $\uparrow$, and $[\cdot, \cdot]$ remain as outlined in Algorithm 1;
 // Shared Encoder
 2 $s_1 \leftarrow \mathcal{F}_{DC}(x)$;
 3 $s_2 \leftarrow \mathcal{F}_{DC}(\downarrow s_1)$;
 4 $s_3 \leftarrow \mathcal{F}_{DC}(\downarrow s_2)$;
 5 $s_4 \leftarrow \mathcal{F}_{DC}(\downarrow s_3)$;
 6 $b \leftarrow \mathcal{F}_{DC}(\downarrow s_4)$; *// Shared Bottleneck*
 // Branch 1: Intraretinal Fluid (IRF) Decoder
 7 $d4_i \leftarrow \mathcal{F}_{DC}([s_4, \uparrow b])$;
 8 $d3_i \leftarrow \mathcal{F}_{DC}([s_3, \uparrow d4_i])$;
 9 $d2_i \leftarrow \mathcal{F}_{DC}([s_2, \uparrow d3_i])$;
 10 $d1_i \leftarrow \mathcal{F}_{DC}([s_1, \uparrow d2_i])$;
 11 $out_{irf} \leftarrow \text{Conv}_{1 \times 1}(d1_i)$; *// Single-channel IRF logit*
 // Branch 2: Diabetic Macular Edema (DME) Decoder
 12 $d4_d \leftarrow \mathcal{F}_{DC}([s_4, \uparrow b])$;
 13 $d3_d \leftarrow \mathcal{F}_{DC}([s_3, \uparrow d4_d])$;
 14 $d2_d \leftarrow \mathcal{F}_{DC}([s_2, \uparrow d3_d])$;
 15 $d1_d \leftarrow \mathcal{F}_{DC}([s_1, \uparrow d2_d])$;
 16 $out_{dme} \leftarrow \text{Conv}_{1 \times 1}(d1_d)$; *// Single-channel DME logit*
 // Multi-label Fusion & Background Channel
 17 $out_{bg} \leftarrow \text{Conv}_{1 \times 1}([d1_i, d1_d])$; *// Background logit from fused features*
 18 $\hat{y} \leftarrow \sigma([out_{bg}, out_{irf}, out_{dme}])$; *// Per-channel sigmoid; shape $(B, 3, H, W)$*
 19 **return** $\hat{y}$

Algorithm 3: Forward Pass of U-Net++ with CBAM Attention

Input: x : 2D OCT scan tensor of shape $(B, 1, H, W)$
Output: $\hat{y}$: Predicted segmentation mask of shape $(B, 3, H, W)$

1 **Notation:** $\mathcal{F}_V(\cdot)$: VGG Block with Convolutional Block Attention Module (CBAM); $\downarrow$: 2×2 Max Pooling; $\uparrow$: Bilinear Upsampling; $[\cdot, \dots, \cdot]$: Dense multi-scale channel-wise concatenation;
 // Depth 0 & 1 (Initial Encoding & First Skip)
 2 $x_{0,0} \leftarrow \mathcal{F}_V(x)$;
 3 $x_{1,0} \leftarrow \mathcal{F}_V(\downarrow x_{0,0})$;
 4 $x_{0,1} \leftarrow \mathcal{F}_V([x_{0,0}, \uparrow x_{1,0}])$;
 // Depth 2 (Nested Features)
 5 $x_{2,0} \leftarrow \mathcal{F}_V(\downarrow x_{1,0})$;
 6 $x_{1,1} \leftarrow \mathcal{F}_V([x_{1,0}, \uparrow x_{2,0}])$;
 7 $x_{0,2} \leftarrow \mathcal{F}_V([x_{0,0}, x_{0,1}, \uparrow x_{1,1}])$;
 // Depth 3 (Nested Features)
 8 $x_{3,0} \leftarrow \mathcal{F}_V(\downarrow x_{2,0})$;
 9 $x_{2,1} \leftarrow \mathcal{F}_V([x_{2,0}, \uparrow x_{3,0}])$;
 10 $x_{1,2} \leftarrow \mathcal{F}_V([x_{1,0}, x_{1,1}, \uparrow x_{2,1}])$;
 11 $x_{0,3} \leftarrow \mathcal{F}_V([x_{0,0}, x_{0,1}, x_{0,2}, \uparrow x_{1,2}])$;
 // Depth 4 (Bottleneck & Final Dense Aggregation)
 12 $x_{4,0} \leftarrow \mathcal{F}_V(\downarrow x_{3,0})$;
 13 $x_{3,1} \leftarrow \mathcal{F}_V([x_{3,0}, \uparrow x_{4,0}])$;
 14 $x_{2,2} \leftarrow \mathcal{F}_V([x_{2,0}, x_{2,1}, \uparrow x_{3,1}])$;
 15 $x_{1,3} \leftarrow \mathcal{F}_V([x_{1,0}, x_{1,1}, x_{1,2}, \uparrow x_{2,2}])$;
 16 $x_{0,4} \leftarrow \mathcal{F}_V([x_{0,0}, x_{0,1}, x_{0,2}, x_{0,3}, \uparrow x_{1,3}])$;
 // Classification Head
 17 $\hat{y} \leftarrow \text{Conv}_{1 \times 1}(x_{0,4})$;
 18 **return** $\hat{y}$

IV. RESULTS AND DISCUSSION

Figures 2–16 present the training and validation curves for all three models across 60 epochs.

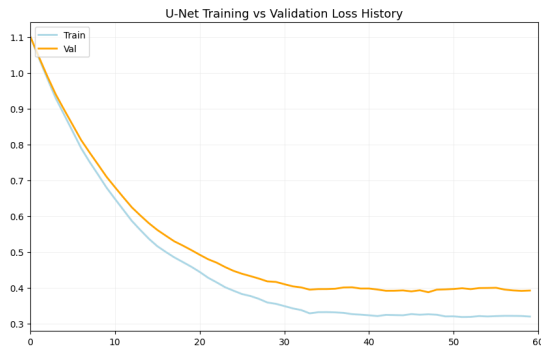


Fig. 2: U-Net: Training vs. Validation Loss

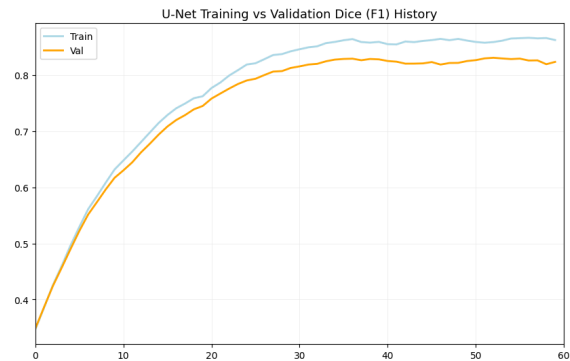


Fig. 3: U-Net: Training vs. Validation Dice/F1

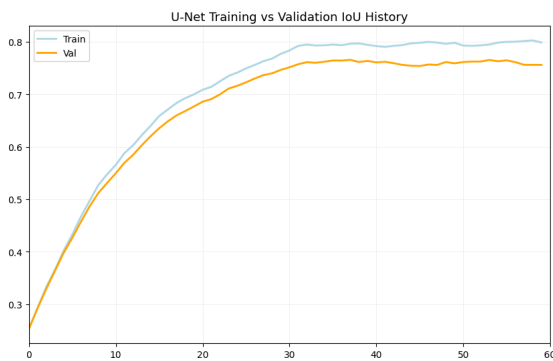


Fig. 4: U-Net: Training vs. Validation IoU

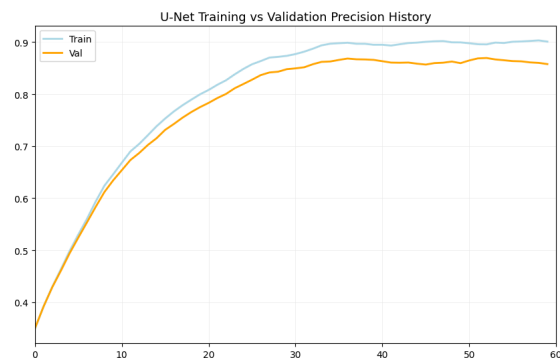


Fig. 5: U-Net: Training vs. Validation Precision

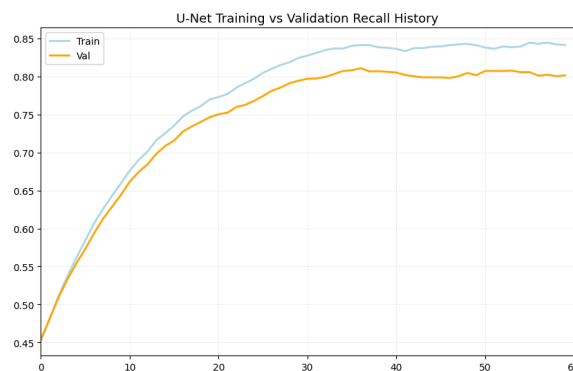
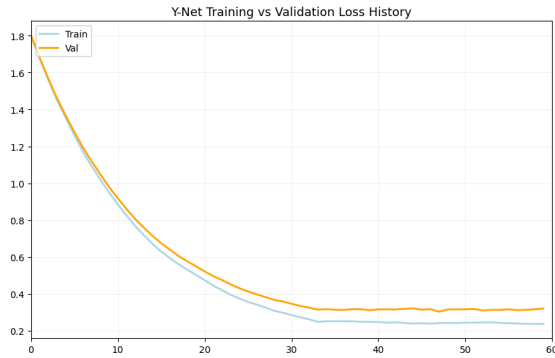
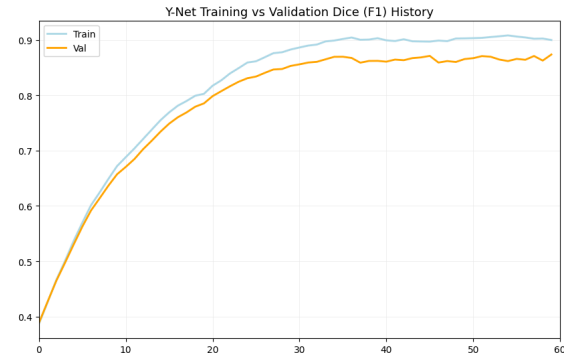
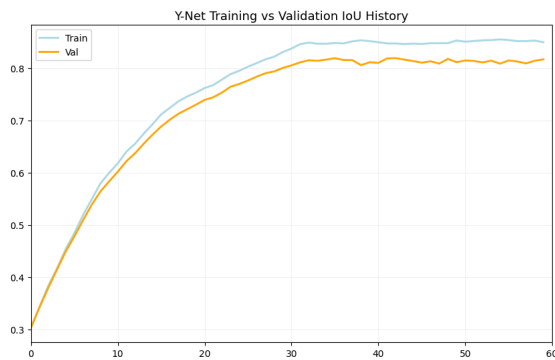
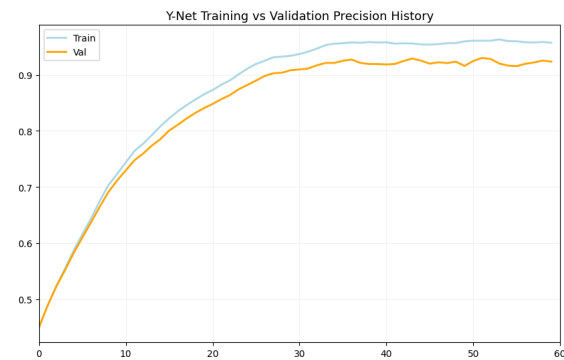
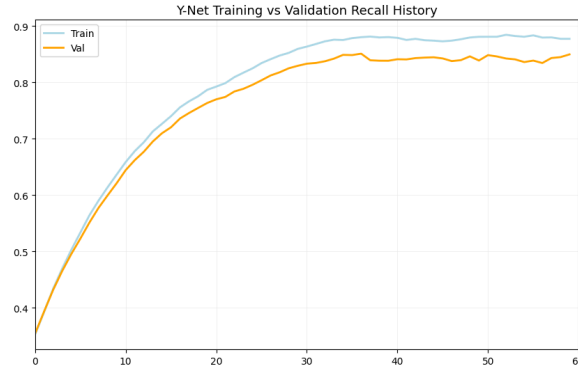
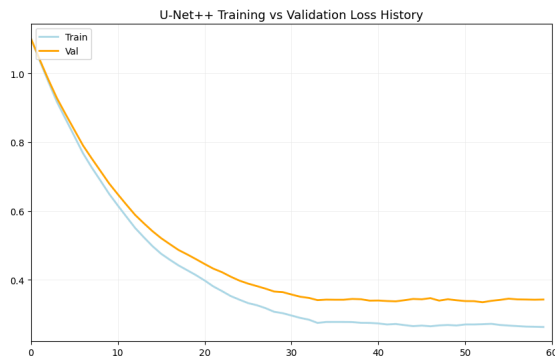
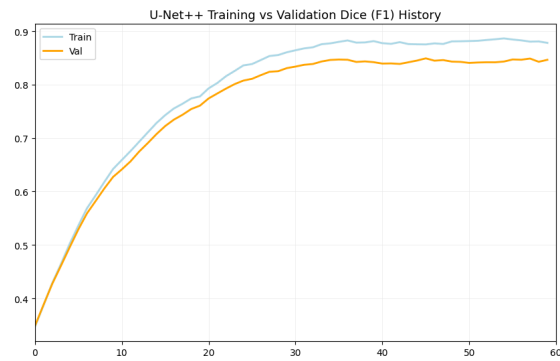
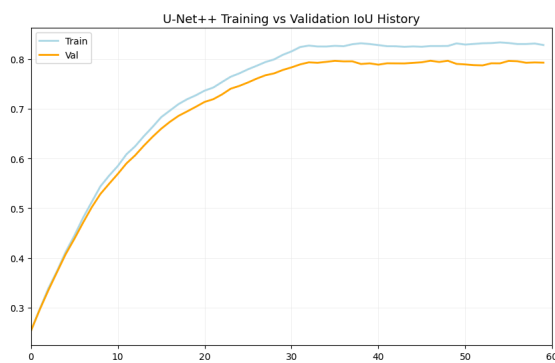
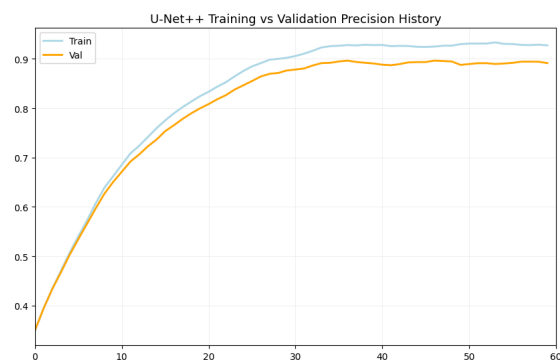
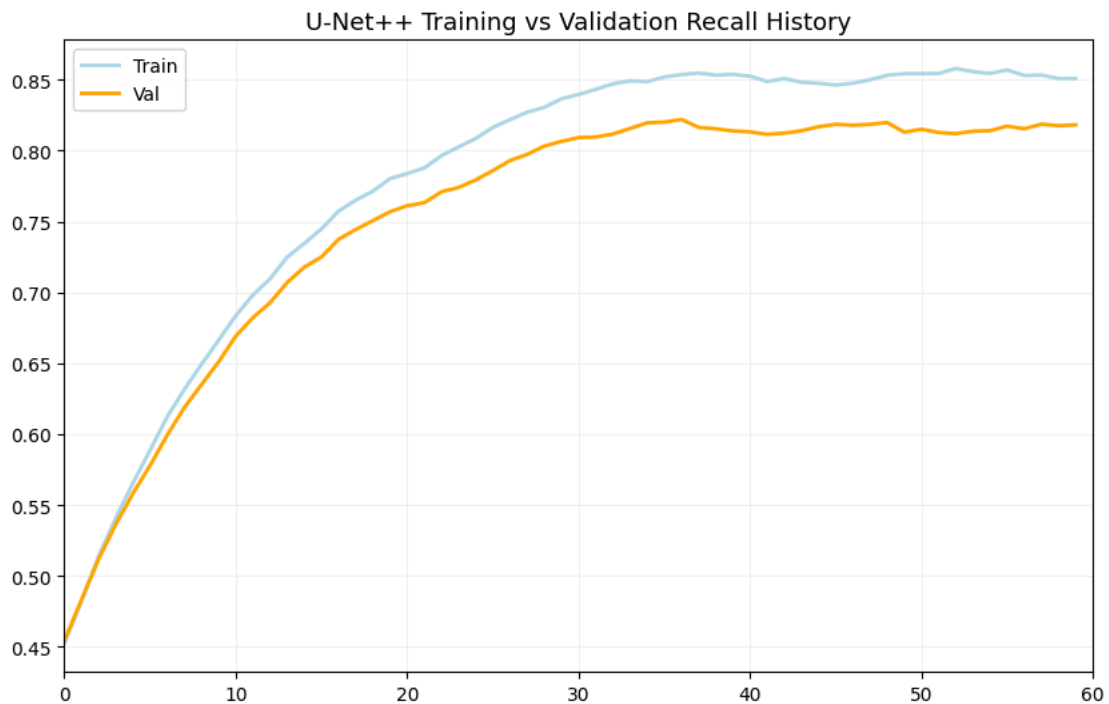


Fig. 6: U-Net: Training vs. Validation Recall

U-Net converged steadily across all metrics, with training and validation curves remaining closely aligned, indicating stable learning without significant overfitting. The validation loss settled around 0.3950 by epoch 60.

**Fig. 7:** Y-Net: Training vs. Validation Loss**Fig. 8:** Y-Net: Training vs. Validation Dice/F1**Fig. 9:** Y-Net: Training vs. Validation IoU**Fig. 10:** Y-Net: Training vs. Validation Precision**Fig. 11:** Y-Net: Training vs. Validation Recall

Y-Net exhibited the fastest initial convergence among the three models, with the validation loss dropping sharply in early epochs before stabilizing at 0.3145 — the lowest among all evaluated architectures. Precision and recall curves showed consistent improvement and minimal divergence between training and validation, suggesting strong generalization.

**Fig. 12:** U-Net++: Training vs. Validation Loss**Fig. 13:** U-Net++: Training vs. Validation Dice/F1**Fig. 14:** U-Net++: Training vs. Validation IoU**Fig. 15:** U-Net++: Training vs. Validation Precision**Fig. 16:** U-Net++: Training vs. Validation Recall

U-Net++ showed competitive convergence behavior, with validation loss stabilizing at 0.3412. A modest gap between training and validation loss in later epochs suggests mild overfitting, likely attributable to the increased model complexity introduced by

the dense skip pathways and CBAM attention modules.

A. Experimental Results

Table 1: Comparison of Model Performance Metrics on the Validation Set (single training run; aggregate scores over the full validation split).

Model	F1	IoU	Prec.	Rec.	Val Loss
U-Net	0.8246	0.7599	0.8620	0.8029	0.3950
Y-Net	0.8651	0.8142	0.9210	0.8424	0.3145
U-Net++	0.8432	0.7925	0.8912	0.8156	0.3412

The performance of the three segmentation models was evaluated on the validation dataset using standard segmentation metrics including F1 Score, Intersection over Union (IoU), Precision, Recall, and Validation Loss. As these models were tuned using the validation set, these metrics reflect validation performance. All reported values are aggregate scores computed over the full validation split from a single training run per model.

- U-Net achieved a validation F1 score of 0.8246 and IoU of 0.7599, with a precision of 0.8620 and recall of 0.8029, along with a validation loss of 0.3950, establishing a strong baseline performance.
- U-Net++ produced improved results over U-Net, achieving a validation F1 score of 0.8432 and IoU of 0.7925, with a precision of 0.8912 and recall of 0.8156, and a validation loss of 0.3412.
- Y-Net demonstrated the best overall performance among the three models, achieving the highest validation F1 score of 0.8651 and IoU of 0.8142, the highest precision of 0.9210 and recall of 0.8424, along with the lowest validation loss of 0.3145.

Recall is a particularly critical metric in medical image segmentation, as it reflects the model's ability to correctly identify positive regions such as lesions or abnormalities. Higher recall reduces the risk of false negatives, which is essential in clinical applications. Among the evaluated models, Y-Net achieved the highest recall of 0.8424, indicating its superior ability to capture relevant regions while maintaining strong precision and overall segmentation quality.

Statistical considerations. The metrics reported in Table 1 are aggregate scores from a single training run per model, computed over the full validation split rather than per-image distributions. As a result, per-image standard deviations, confidence intervals, and formal statistical significance tests (such as paired bootstrap testing on image-level Dice or IoU) could not be derived from the available logged outputs. We acknowledge this as a limitation of the current evaluation. The observed differences between models — Y-Net improving F1 by 2.05 percentage points and IoU by 2.17 percentage points over U-Net++, and by 4.05 and 5.43 percentage points respectively over U-Net — are consistent and directionally aligned across all five reported metrics, suggesting the ranking is unlikely to be a result of random variation. Nevertheless,

future work should log per-image metric scores during evaluation and apply formal statistical testing, such as the Wilcoxon signed-rank test on image-level Dice scores, to confirm the significance of these differences.

B. Discussion

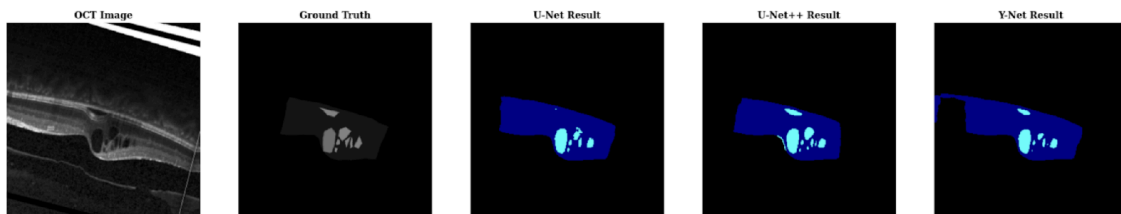


Fig. 17: Visualization of Model Predictions. From left to right: original OCT input, ground truth mask, U-Net output, U-Net++ output, and Y-Net output. DME regions are shown in dark blue; IRF regions are shown in cyan.

To further contextualize the quantitative metrics, a qualitative assessment was performed by visually comparing the model predictions against the ground truth. Figure 11 presents a comparative panel displaying the original input scan, the ground truth mask, and the resulting segmentation outputs from the evaluated architectures. DME regions are represented in dark blue and Intraretinal Fluid (IRF) regions in cyan, corresponding to two distinct segmentation classes.

Both Y-Net and U-Net++ produce strong and visually coherent segmentations, accurately capturing the overall structure of the affected regions. Among them, Y-Net provides the most accurate visual prediction, with better preservation of fine details, smoother boundaries, and closer alignment to the ground truth mask. U-Net++ also demonstrates high-quality segmentation, particularly in maintaining structural consistency across regions, but shows slight limitations in precisely delineating smaller or fragmented fluid areas compared to Y-Net. In contrast, the baseline U-Net, while stable, exhibits comparatively less precise boundary delineation and reduced sensitivity to finer details.

V. CONCLUSIONS, SUGGESTIONS AND RECOMMENDATIONS

A. Discussion

This study investigated the segmentation of OCT images using U-Net, U-Net++, and a task-specific dual-decoder Y-Net, in which a shared encoder feeds two separate decoder branches dedicated to IRF and DME segmentation respectively. Among the three models, Y-Net achieved the best overall segmentation performance on the validation set, obtaining the highest F1 Score (0.8651), IoU (0.8142), Precision (0.9210), Recall (0.8424), and the lowest validation loss (0.3145). The results indicate that Y-Net provided more accurate and reliable segmentation of Intraretinal Fluid (IRF) and Diabetic Macular Edema (DME) regions in OCT images compared to U-Net and U-Net++. Its architecture effectively captured both local and contextual

features through task-specific decoding pathways, leading to improved segmentation consistency and reduced prediction errors. These findings suggest that Y-Net can support ophthalmologists in making faster and more dependable clinical diagnoses through accurate automated OCT image segmentation. It should be noted that all reported metrics are aggregate scores from a single training run per model; future evaluations should include per-image statistical analysis to formally confirm the significance of the observed performance differences.

B. Suggestions and Recommendations

- Since Y-Net achieved the best segmentation performance, future work should focus on optimizing the model to reduce computational complexity and memory consumption for efficient real-time deployment.
- Although U-Net and U-Net++ produced competitive results, their performance can be further improved through architectural enhancements, hyperparameter tuning, and advanced training strategies.
- Additional preprocessing and image enhancement techniques should be explored to improve the segmentation of small or low-contrast retinal abnormalities in OCT images.
- A larger and more diverse OCT dataset should be used to improve the robustness and generalization capability of the segmentation models across different clinical scenarios.
- Future datasets should include explicit patient and volume identifiers to enable fully patient-independent data splitting, eliminating any residual risk of data leakage across training, validation, and test sets.
- Advanced denoising and artifact removal techniques can be incorporated to reduce the effect of image noise on segmentation accuracy.
- Future studies may investigate incremental learning and transfer learning approaches to enable continuous adaptation of the models to new clinical data.
- Future evaluations should log per-image metric scores and apply formal statistical significance tests — such as the Wilcoxon signed-rank test on image-level Dice and IoU scores — to rigorously confirm performance differences across architectures.
- Lightweight and real-time OCT image segmentation systems suitable for clinical applications and portable healthcare devices should be developed to enhance practical usability.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to Er. Prem Chandra Roy for his continuous guidance, valuable suggestions, and constant support throughout this research work. We are also thankful to the Department of Electronics and Computer Engineering at Advanced College of Engineering and Management, affiliated with the Institute of Engineering,

Tribhuvan University, for providing the necessary academic environment and institutional support required for this study. We are equally grateful to the respective doctors and hospitals for their guidance and support during this research. Finally, we would like to thank our family members, friends, and well-wishers for their encouragement and moral support throughout the completion of this work.

AUTHOR CONTRIBUTIONS

All authors actively contributed to this research work and to the preparation of the manuscript. The study concept and research methodology were developed collaboratively. Data collection, system development, experimentation, analysis, and interpretation of results were jointly carried out by the authors. The manuscript was written, reviewed, and revised with the active involvement of all authors, and all authors approved the final version of the manuscript for publication.

DATA AVAILABILITY

The dataset used in this study was obtained from Kaggle (<https://www.kaggle.com/datasets/obulisainaren/retinal-oct-c8>), published by Obuli Sai Naren in 2021. Relevant information supporting the findings of this research has been included within the article. Additional data related to this study are available from the corresponding author upon reasonable request.

CONFLICT OF INTEREST

The authors confirm that there are no conflicts of interest related to this publication. The research was conducted independently, and no personal or financial relationships influenced the outcomes of the study.

REFERENCES

- [1] O. Morelle, M. W. Wintergerst, R. P. Finger, and T. Schultz, "Accurate drusen segmentation in optical coherence tomography via order-constrained regression of retinal layer heights," *Scientific Reports*, vol. 13, no. 1, p. 8162, 2023.
- [2] K. Yojana and L. T. Rani, "Oct layer segmentation using u-net semantic segmentation and resnet34 encoder-decoder," *Measurement: Sensors*, vol. 29, p. 100817, 2023.
- [3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: <http://arxiv.org/abs/1505.04597>
- [4] L. P. Bhatt and S. Shakya, "Deep learning-driven uncertainty estimation for brain tissue segmentation by integrating expectation maximization algorithms," in *2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL)*, 2025, pp. 914–920.
- [5] S. Lamichhane, L. P. Bhatt, and S. Shakya, "Deeprust: A hybrid transformer-cnn model for deepfake detection with zero-knowledge-based blockchain authentication," *Journal of*

- Advanced College of Engineering and Management, vol. 12, no. 01, p. 57–71, May 2026. [Online]. Available: <https://nepjol.info/index.php/JACEM/article/view/93907>
- [6] S. Poudel, L. P. Bhatt, P. C. Prasad, and A. C. Paudel, “Prediction of chronic kidney disease using random forest, xgboost and ann model,” *Journal of Advanced College of Engineering and Management*, vol. 10, no. 1, p. 121–133, Mar. 2025. [Online]. Available: <https://nepjol.info/index.php/JACEM/article/view/76323>
- [7] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: A nested u-net architecture for medical image segmentation,” in *Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4*. Springer, 2018, pp. 3–11.
- [8] R. Baral, S. Satyal, and A. Pokhrel, “Cnn-transformer based speech emotion detection,” *Journal of Advanced College of Engineering and Management*, vol. 10, no. 1, p. 135–145, Mar. 2025. [Online]. Available: <https://nepjol.info/index.php/JACEM/article/view/76324>
- [9] A. Farshad, Y. Yeganeh, P. Gehlbach, and N. Navab, “Y-net: A spatio-spectral dual-encoder network for medical image segmentation,” in *International conference on medical image computing and computer-assisted intervention*. Springer, 2022, pp. 582–592.
- [10] S. Mehta, E. Mercau, J. Bartlett, D. Weaver, J. G. Elmore, and L. Shapiro, “ay-net: joint segmentation and classification for diagnosis of breast biopsy images,” in *International conference on medical image computing and computer-assisted intervention*. Springer, 2018, pp. 893–901.
- [11] S. Sitoula, L. P. Bhatt, A. Pokhrel, and A. Neupane, “Generative ai for mental well-being and insights from high and low resource languages,” in *Enhancing Edge Devices With Generative Intelligence: Architecture, Challenges, and Applications*, S. Shakyia, T. Shahi, E. Ceh-Varela, and M. Cowling, Eds. IGI Global Scientific Publishing, 2026, pp. 391–424. [Online]. Available: <https://doi.org/10.4018/979-8-3373-7340-9.ch012>
- [12] Y. Yuan and Y. Cheng, “Medical image segmentation with unit-based multi-scale context fusion,” *Scientific Reports*, vol. 14, no. 1, p. 15687, 2024.
- [13] P. K. Karn and W. H. Abdulla, “Precision segmentation of subretinal fluids in oct using multiscale attention-based u-net architecture,” *Bioengineering*, vol. 11, no. 10, 2024. [Online]. Available: <https://www.mdpi.com/2306-5354/11/10/1032>
- [14] P. Geetha Pavani, B. Biswal, S. R. Kandula, P. Biswal, G. Siddhartha, T. Niranjana, and B. S. N., “A robust segmentation of retinal fluids from oct images using mcfar-net,” *Neurocomputing*, vol. 599, p. 128059, 2024.
- [15] R. Wen, H. Yuan, D. Ni, W. Xiao, and Y. Wu, “From denoising training to test-time adaptation: Enhancing domain generalization for medical image segmentation,” 2023. [Online]. Available: <https://arxiv.org/abs/2310.20271>

- [16] W. Meng, X. Zhu, H. Li, Y. Li, D. Ma, and S. Xu, "Ynet: Research on deep learning model driven by multi-scale analysis and its application to image segmentation," 2023.