https://doi.org/10.31436/iiumej.v27i1.3806

ToxoSegFusion: Attention-enhanced Dual-backbone Neural Architecture for Retinal Lesion Segmentation

MD MOMINUL HAQUE*, SAIPUNIDZAM MAHAMAD, SUZIAH SULAIMAN, ABDULLATEEF OLUWABEMIGA BALOGUN, HUSSAINI MAMMAN

Department of Computing, Universiti Teknologi PETRONAS, Perak, Malaysia

**Corresponding author: md_24004641@utp.edu.my*

(Received: 23 June 2025; Accepted: 7 November 2025; Published online: 12 January 2026)

ABSTRACT: Ocular toxoplasmosis (OT) often presents a diagnostic dilemma in clinics, with retinal lesions that are not only varied in appearance but also frequently subtle and underrepresented in fundus images. Current automated segmentation tools, though promising, are often hampered by class imbalance and a lack of robust testing across real-world scenarios. To address these gaps, we developed ToxoSegFusion, a dual-backbone deep learning framework that capitalizes on the complementary strengths of DenseNet121 and ResNet101, enhanced with attention modules. Unlike typical single-backbone models, this hybrid approach was specifically tuned for the intricate challenges of OT lesion segmentation, using a combined Dice and binary cross-entropy loss to better balance rare lesion pixels. We trained and validated on 149 image-mask pairs from the OTFID-Version 3 dataset, achieving an intersection over union of 0.858 and a Dice coefficient of 0.795, both exceeding the current MobileNetV2/U-Net baseline. The model also demonstrated reliable performance on the DRIVE dataset for vessel segmentation, indicating practical flexibility. By facilitating accurate lesion localization, ToxoSegFusion enables more timely interventions in ophthalmology. Future directions include larger multi-center trials and streamlined models for routine deployment.

ABSTRAK: Toksoplasmosis okular (OT) sering menimbulkan cabaran diagnostik di klinik, dengan lesi retina halus pelbagai rupa dan kurang terwakili pada imej fundus. Alat segmentasi automatik semasa, walaupun memberi harapan, sering terhad pada ketidakseimbangan kelas dan kekurangan ujian di peringkat perubatan. Bagi mengatasi kekurangan ini, kajian ini membangunkan ToxoSegFusion, sebuah rangka kerja pembelajaran mendalam berkomponen dua yang memanfaatkan kekuatan saling melengkapi DenseNet121 dan ResNet101, diperkaya dengan mekanisme perhatian. Tidak seperti model komponen tunggal biasa, pendekatan hibrid ini dirancang khusus bagi cabaran kompleks segmentasi lesi OT, menggunakan kehilangan Dice dan entropi silang binari gabungan bagi keseimbangan terbaik antara piksel lesi yang jarang. Kajian ini melatih dan mengesahkan 149 pasangan imej-topeng dari set data OTFID-Versi 3, mencapai persilangan atas kesatuan 0.858 dan pekali Dice 0.795, keduanya melebihi garis dasar MobileNetV2/U-Net semasa. Model juga menunjukkan prestasi terbaik pada DRIVE bagi segmentasi salur darah, mencadangkan fleksibiliti praktis. Melalui pengesanan lokasi lesi yang tepat, ToxoSegFusion membuka jalan bagi intervensi lebih tepat pada masa oftalmologi. Pada masa hadapan, cadangan bagi penyebaran rutin adalah melalui ujian berbilang pusat yang lebih besar dan perkemasan model.

KEYWORDS: *Retinal Image Segmentation, Deep Learning, Fusion Model, Medical Imaging*

https://doi.org/10.31436/iiumej.v27i1.3806

1. INTRODUCTION

Retinal imaging is vital in ophthalmology, providing a window into the retina, optic disc, and blood vessels for the diagnosis and management of conditions such as diabetic retinopathy, glaucoma, and ocular toxoplasmosis (OT). With over 1.1 billion people worldwide affected by vision impairment, imaging tools such as fundus photography and optical coherence tomography (OCT) are essential for early detection and treatment, helping to prevent blindness [1]. These methods enable precise tracking of disease progression and support tailored treatment strategies, transforming clinical care. However, diagnosing complex retinal diseases such as OT, a significant cause of infectious posterior uveitis, requires advanced tools to improve accuracy and efficiency.

Ocular toxoplasmosis, caused by the parasite Toxoplasma gondii, results in retinal lesions that vary in size, shape, and appearance, making accurate identification difficult [2, 3]. Affecting 2–20% of those with T. gondii infection (25–30% of the global population), OT can cause significant vision loss if not treated promptly [4]. Fundus photography captures these lesions, but their faint boundaries and rarity (about 5% of image pixels) pose challenges for accurate segmentation, which is critical for effective treatment planning. Traditional diagnosis relies on manual inspection of fundus images by specialists, often aided by fluorescein angiography or OCT. These approaches, while valuable, are time-consuming, prone to inter-observer variability (up to 15% disagreement in lesion outlines), and constrained by the scarcity of annotated data [5, 6]. The imbalance between lesion and background pixels further complicates automated segmentation efforts [7].

Existing methods for OT segmentation have significant drawbacks. Manual segmentation is inconsistent, and early automated techniques, such as thresholding or edge detection, struggle with the diverse appearance of lesions and sensitivity to image noise, leading to unreliable results [8]. Initial machine learning approaches, such as support vector machines and random forests, improved performance by learning from data but relied on manually engineered features, achieving modest performance (e.g., Intersection over Union scores around 0.65) due to limited data and class imbalance [9]. More recent deep learning models, such as U-Net variants, have improved retinal segmentation but struggle with OT's subtle lesion boundaries and small datasets [10, 11]. While some advanced models excel in classifying OT, they rarely address the pixel-level segmentation needed for precise lesion mapping, leaving a gap in clinical applications [12, 13].

This study introduces ToxoSegFusion, a deep learning framework designed to address the challenges of binary OT lesion segmentation in the context of limited data and class imbalance. By combining DenseNet121's ability to capture fine details with ResNet101's robust training, ToxoSegFusion integrates complementary feature extraction to improve segmentation accuracy [14, 15]. The model employs attention mechanisms to focus on lesion regions and a combined loss function (binary cross-entropy and Dice loss) to address the scarcity of lesion pixels. Trained on 149 fundus image-mask pairs from the OTFID-Version 3 dataset with real-time data augmentation, ToxoSegFusion aims to outperform existing models like MobileNetV2/U- Net, as measured by Intersection over Union and Dice coefficient metrics [11]. The key question is whether a hybrid DenseNet-ResNet architecture with attention mechanisms can improve OT lesion segmentation under data constraints.

The significance of ToxoSegFusion lies in its ability to provide precise lesion localization, enabling faster and more effective clinical interventions to reduce vision loss. Its contributions include:

- A novel hybrid architecture combining DenseNet121 and ResNet101 with attention mechanisms to improve segmentation of diverse OT lesions.
- A tailored loss function addressing class imbalance to enhance the detection of rare lesions.
- Experimental evaluation conducted revealed that the proposed hybrid architectures and attention-driven segmentation outperform well.

The paper is structured as follows: Section 2 reviews related work, Section 3 details the methodology, Section 4 presents results, and Section 5 summarizes findings and future directions.

2. LITERATURE REVIEW

The diagnosis of ocular toxoplasmosis (OT) has traditionally depended on non-AI methods, centered on manual examination of retinal images by ophthalmologists. Fundus photography, a primary imaging modality, captures detailed images of OT lesions, which vary from active inflammatory foci to inactive scars, enabling qualitative assessment of their size, shape, and activity [4]. Complementary techniques, such as fluorescein angiography, identify active lesions via hyperfluorescence, whereas optical coherence tomography (OCT) quantifies retinal thickness and structural changes, thereby enhancing diagnostic precision [2]. For example, OCT can detect macular edema in 30% of OT cases, aiding treatment planning [3]. However, these methods are time-consuming, require specialized training, and exhibit inter-observer variability, with discrepancies in lesion boundary delineation reaching 15% [5]. The absence of standardized annotation protocols further hampers reproducibility, particularly in low-resource settings where access to advanced imaging is limited [16]. These challenges highlight the need for automated, objective tools to streamline OT diagnosis and improve clinical outcomes.

Early artificial intelligence (AI) approaches introduced automation into retinal image analysis, employing rule-based and machine learning (ML) methods. Rule-based techniques, including thresholding, edge detection, and morphological operations, segmented lesions by applying predefined pixel-intensity or structural criteria [8]. For instance, thresholding was used to isolate bright lesions in diabetic retinopathy (DR) fundus images, but achieved low segmentation accuracy (IoU < 0.60) for OT due to lesion heterogeneity and noise sensitivity [8]. ML methods, such as support vector machines (SVMs), k-nearest neighbors (k-NN), and random forests, advance feature extraction by learning from labeled data. Acharya et al.'s SVM model for DR lesion detection reported an IoU of 0.65, relying on manually engineered features like texture and color histograms [9]. Lowell et al.'s k-NN approach for optic disc segmentation achieved 94.7% accuracy but struggled with complex OT lesions due to limited generalization [17]. These methods were constrained by their reliance on handcrafted features and small datasets (e.g., <500 images), which exacerbated class imbalance, particularly for rare active OT lesions, which constitute 5% of pixels [6]. The limitations of early AI methods underscored the need for more robust deep learning (DL) models that can learn hierarchical features directly from raw images.

Deep learning has revolutionized retinal image analysis, delivering pixel-level precision for segmentation and classification tasks essential for diagnosing ocular diseases like DR, glaucoma, and OT [18]. Convolutional neural networks (CNNs) have set benchmarks for retinal segmentation. U-Net, with its encoder-decoder architecture and skip connections, excels at segmenting retinal blood vessels and DR lesions, achieving an IoU of 0.85 and a Dice

coefficient of 0.87 on the DRIVE and STARE datasets [19]. DeepLabV3+, leveraging convolutions and spatial pyramid pooling, enhances boundary delineation, reporting an IoU of 0.82 for DR microaneurysms on fundus images [20]. Advanced CNN backbones, such as ResNet and DenseNet, introduce residual learning and dense connectivity, respectively, improving feature extraction and training stability [14, 15]. ResNet50-based U-Net variants achieved a Dice coefficient of 0.83 for retinal vessel segmentation on the IDRiD dataset [21]. Vision transformers (ViTs) capture long-range dependencies, with SwinMedNet achieving 56.8% accuracy for DR classification on RetinaMNIST2D, though it lacks segmentation capabilities [22]. Hybrid models such as TransUNet combine CNNs and ViTs, achieving a Dice coefficient of 0.81 for retinal vessel segmentation on the DRIVE dataset [23]. Despite these advancements, their application to OT is constrained by data scarcity and class imbalance, as noted in comprehensive reviews [7, 10].

OT-specific DL research has primarily focused on classification, with segmentation remaining underexplored due to the complexity of pixel-level lesion analysis. Cardozo et al. developed a ResNet18-based model for multiclass OT classification on the OTFID-Version 3 dataset (412 fundus images), achieving 86.7% accuracy, 91.2% sensitivity, and 98.0% specificity with preprocessing steps like resizing and augmentation [12]. Aziz et al.'s Reptile meta-learning framework, using InceptionV3, achieved 84.8% accuracy on few-shot OT classification on OTFID Version 3, incorporating CLAHE and Gaussian filtering [13]. Both studies focus on classification, limiting their utility for lesion localization. On the other hand, OT segmentation using U-Net variants (MobileNetV2/U-Net, ResNet34/U-Net) on OTFID-Version 3, reporting a mean IoU of 0.835 and a Dice coefficient of 0.771, but active lesion detection was suboptimal (IoU: 0.76) due to class imbalance [11]. Karkuzhali et al. employed lightweight CNNs for OT classification, achieving 85.2% accuracy on OTFID-Version 3, but did not explore segmentation [24]. Ragab et al.'s transfer-learning approach, which used a pretrained ResNet-50, improved OT classification accuracy by 7% on small datasets but lacked pixel-level analysis [25]. These studies highlight a critical gap in robust, multi-class OT segmentation models that can address heterogeneous lesions and rare active classes.

Class imbalance and limited annotated data pose significant challenges in OT segmentation, as active lesions are clinically critical but constitute only 5% of pixels. Weighted loss functions mitigate this issue. Focal loss emphasizes hard-to-classify samples, improving Dice scores by 4–6% in DR segmentation [26]. Combined cross-entropy and Dice loss, as applied by Iriondo et al., enhanced rare lesion segmentation by 5% [27]. alam et al.'s weighted loss for OT boosted active lesion IoU by 3% [11]. Data augmentation, including rotations, flips, and color jittering, addresses data scarcity, but small datasets like OTFID (412 images) limit generalizability. Generative adversarial networks (GANs) synthesize retinal images, improving Dice scores by 2–3% in DR tasks [28]. Federated learning, as in MedUniverse, achieves 95% accuracy for OT classification across multi-institutional datasets while preserving privacy, but its segmentation potential is untested [29]. Transfer learning with pretrained models like ResNet50 boosts accuracy by 7% on small retinal datasets [25]. These strategies highlight the need for tailored solutions to address OT-specific challenges related to data scarcity and class imbalance.

Despite these advances, several methodological limitations in prior studies constrain their clinical applicability. Traditional rule-based techniques employing thresholding and edge detection achieved modest accuracy (IoU < 0.60) due to sensitivity to image noise and inability to handle heterogeneous lesion appearances. Machine learning approaches using SVMs and random forests improved performance but remained dependent on handcrafted features and struggled with generalization on small datasets, typically achieving an IoU of around 0.65.

https://doi.org/10.31436/iiumej.v27i1.3806

Table 1. Summary of Studies in Retinal Image Analysis for Ocular Toxoplasmosis and Related Tasks

Paper	Methods	Dataset	Type	Key Limitation
[2]	Fundus photography, OCT, manual assessment	Clinical OT images	Diagnosis	Subjective interpretation, inter-observer variability
[4]	Manual fundus analysis, fluorescein angiography	Clinical OT images	Diagnosis	Time-intensive, invasive contrast agent required
[8]	Thresholding, edge detection, morphological operations	Fundus images	Detection	Noise-sensitive, poor adaptability (IoU < 0.60)
[9]	SVM, random forests, feature engineering	Fundus images	Detection	Handcrafted features, limited generalization (IoU ~0.65)
[11]	MobileNetV2/U-Net, ResNet34/U-Net	OTFID-Version 3	OT Segmentation	Single-center data, no external validation reported
[12]	ResNet18, VGG16, augmentation	OTFID-Version 3	OT Classification	Limited dataset diversity, generalizability unclear
[13]	Reptile meta-learning, InceptionV3	OTFID-Version 3	OT Classification	Small sample meta-learning, validation scope limited
[14]	ResNet, residual learning	General	Classification	Natural image pretraining, potential domain shift
[15]	DenseNet, dense connectivity	General	Classification	Computational cost, domain adaptation not addressed
[16]	ViT, federated learning	Multi-institutional	Classification	Heterogeneous data distribution challenges
[18]	k-NN, optic disc localization	Fundus images	Segmentation	Feature dependency, scalability issues
[20]	U-Net, encoder-decoder with skip connections	DRIVE, STARE	Segmentation	Class imbalance not explicitly handled
[21]	DeepLabV3+, atrous convolutions	Fundus images	Segmentation	Limited ablation studies, contribution unclear
[23]	Swin-Transformer, CLAHE, Gaussian filter	RetinaMNIST2D	DR Classification	Small medical dataset validation, artifact robustness untested
[24]	TransUNet, CNN-ViT hybrid	DRIVE	Segmentation	Computational complexity, clinical condition testing absent
[25]	Lightweight CNNs, preprocessing	OTFID-Version 3	OT Classification	Single-center bias, statistical rigor lacking
[29]	GAN-based image synthesis	Fundus images	Augmentation	Synthetic image fidelity concerns, medical validity unclear
[26]	Transfer learning, ResNet50	Fundus images	Classification	Domain mismatch, natural-to-medical transfer gaps

The shift to deep learning introduced new challenges: many classification models for OT relied on single-center datasets without external validation, raising concerns about generalizability across diverse clinical settings [11,12,13,25]. Statistical rigor was often lacking, with few studies reporting formal cross-validation, significance testing, or ablation studies to isolate architectural contributions. Segmentation approaches were confronted with severe class imbalance, in which active lesions accounted for only 5% of pixels, yet weighted loss functions yielded only modest improvements. Transfer learning implementations using pretrained models exhibited domain-shift vulnerabilities when applied to medical images, because natural-image features may not optimally represent pathological patterns. Advanced

architectures, including Swin Transformers and TransUNet, showed promise but have been validated only on small medical datasets and have not been evaluated under clinically realistic conditions with artifacts and concurrent diseases [2,8]. Furthermore, GAN-based augmentation strategies introduced concerns regarding synthetic image quality and medical fidelity, while federated learning approaches faced challenges in handling heterogeneous data distributions across institutions. These limitations underscore the necessity for multi-center validation with statistical rigor, transparent reporting of dataset characteristics, systematic ablation studies, and evaluation under clinically representative conditions to advance these methods from technical demonstrations to reliable diagnostic tools [21,23,25,29].

Table 1 summarizes the reviewed studies, highlighting their methods, datasets, and focus. Conventional methods lack automation, early AI methods are limited by feature engineering, and DL methods, while advanced, underexplored OT segmentation due to small datasets, class imbalance, and limited pixel-level analysis [7, 11]. Transformer-based and hybrid models have not been evaluated for OT segmentation, and existing OT segmentation efforts struggle with active lesion detection [23]. ToxoSegFusion addresses these gaps by integrating DenseNet121 and ResNet101 backbones with attention mechanisms and a combined loss function (weighted cross-entropy and Dice loss). Achieving a mean IoU of 0.858, Dice coefficient of 0.795, and active lesion IoU of 0.792 on OTFID-Version 3, it outperforms MobileNetV2/U-Net (IoU: 0.835) [11]. By leveraging complementary architectural strengths and optimized loss functions, ToxoSegFusion offers a novel, clinically actionable solution for multi-class OT lesion segmentation, with potential applications to other retinal diseases [29].

3. METHODOLOGY

This study presents ToxoSegFusion, an innovative deep learning framework designed to segment ocular toxoplasmosis (OT) lesions in retinal fundus images and to demonstrate adaptability to retinal vessel segmentation in the DRIVE dataset. The methodology encompasses a comprehensive pipeline comprising dataset curation, sophisticated image preprocessing, a dual-network model architecture, a meticulously designed training strategy, and a robust evaluation framework. The OT dataset, characterized by its limited size and pronounced class imbalance, and the DRIVE dataset's requirement for precise delineation of vascular structures, pose significant challenges. These are systematically addressed through tailored loss functions, extensive data augmentation, and optimized data pipelines. The following subsections elaborate on each component.

3.1. Dataset Description

The OT dataset consists of 412 RGB fundus images sourced from two medical institutions: the Hospital de Clínicas Medical Center, contributing 291 images at a resolution of 2124×2056 pixels captured using the VISUCAM 500 camera, and the Niños de Acosta Ñú General Pediatric Hospital, providing 121 images at 1536×1152 pixels obtained with the Pictor Plus-Portable Ophthalmic Camera. Among these, 280 images depict non-healthy retinas, with 179 accompanied by manually annotated masks delineating three classes: background (pixel value 0), inactive lesions (127), and active lesions (255). A metadata file, dataset_labels.csv, categorizes images into healthy (131), inactive (188), active (35), or combined active/inactive (57) cases, revealing a pronounced class imbalance in which active lesions account for only 8.5% of non-healthy images.

To ensure data usability, a Python script was developed to align images with their corresponding masks by normalizing filenames and resolving inconsistencies, such as varying

prefixes or suffixes. This process yielded 149 image-mask pairs, which were subsequently divided into 80% training (119 pairs) and 20% validation (30 pairs) sets using stratified sampling with a random seed of 42 to ensure reproducibility. Figure 1 illustrates sample images from the OT dataset, showing fundus photographs with their corresponding segmentation masks across the three classes.

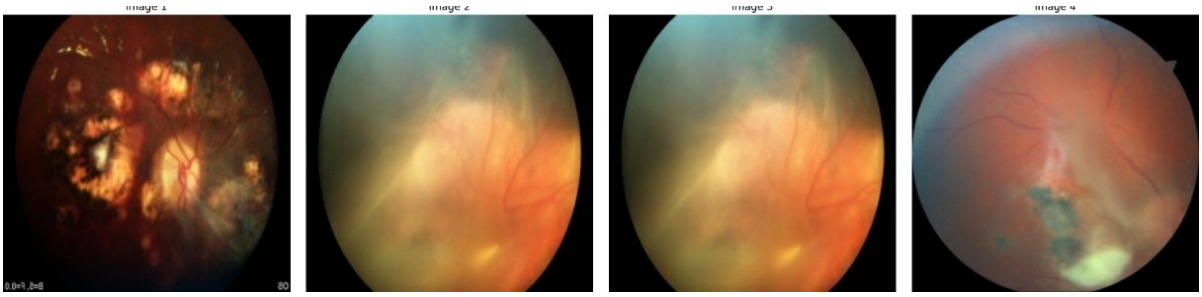


Figure 1. Sample images from the OT dataset, showing fundus photographs with corresponding segmentation masks (background: black, inactive lesions: green, active lesions: red).

In parallel, the DRIVE dataset comprises 40 RGB fundus images at 565×584 pixels, evenly split into 20 training and 20 testing images, each paired with binary vessel masks. To augment the training data, five 512×512 patches were extracted from each training image, prioritizing regions rich in vascular structures using a mean pixel-intensity threshold greater than 0.02.

3.2. Data Preprocessing and Augmentation Strategy

To address the limited dataset size (149 OT image-mask pairs) and pronounced class imbalance (5% lesion pixels), we employed sophisticated real-time data augmentation using the Albumentations library [30]. Augmentation strategies are grounded in medical imaging best practices, where geometric and intensity transformations enhance model robustness to natural variations in retinal photography [19, 31].

The augmentation pipeline applies transformations systematically. Random 90-degree rotations with probability $p = 0.5$ address the circular nature of fundus images and account for varying patient head positioning during image acquisition [32]. Horizontal and vertical flips, each with probability 0.5, simulate bilateral retinal variations between the left and right eyes, as these anatomical patterns are naturally symmetric [33]. Random brightness and contrast modifications, also applied with probability 0.5, mimic variations in lighting conditions and camera calibration across different clinical settings. This adjustment is particularly critical for OT lesion detection, where subtle intensity differences distinguish active from inactive lesions and must be recognized regardless of imaging equipment variation [34]. Shift-scale-rotate operations with shift, scale, and rotation limits of 0.1, 0.2, and 45 degrees, respectively, applied with probability 0.5, account for off-axis imaging geometry and variability in patient positioning during fundus photography [35]. Elastic transforms, applied with probability 0.3, introduce nonlinear deformations that simulate variations in tissue morphology and slight eye-movement artifacts [36]. Gaussian blur, similarly, applied with probability 0.3, provides regularization by simulating slight motion artifacts and image acquisition noise that occur in clinical practice [37].

These transformations are applied on-the-fly during training to generate diverse samples from the limited dataset, effectively increasing the adequate training set size from 119 to thousands of unique samples while preserving annotation fidelity. The probabilistic application ensures that not all transformations are applied to every sample, maintaining a balance between data diversity and domain-specific realism [30].

https://doi.org/10.31436/iiumej.v27i1.3806

3.3. Preprocessing Pipeline

For the OT dataset, fundus images were resized to 512×512 pixels using bilinear interpolation to optimize computational efficiency while preserving essential details. Segmentation masks were resized using nearest-neighbor interpolation to preserve discrete class labels and converted into three-channel, one-hot-encoded tensors representing background, inactive, and active lesions. Image pixel values were normalized by dividing by 255 to conform to standard deep learning practices. A TensorFlow-based data pipeline was established to cache and shuffle the dataset with a buffer size equal to the dataset size, batch images with a batch size of 2, and apply real-time augmentations to generate diverse training samples efficiently.

For the DRIVE dataset, images and masks were resized to 1024×1024 pixels, from which 512×512 patches were randomly extracted, prioritizing regions with significant vessel presence. If no suitable patch was identified after 20 attempts, a center crop was employed as a fallback. Images were normalized by dividing pixel values by 255, and masks were converted into single- channel binary tensors, with vessels assigned a value of 1 and background a value of 0. A custom data generator produced batches of 4 patch pairs, incorporating real-time augmentations to ensure sample diversity. Both pipelines leveraged parallel processing and prefetching to minimize input/output bottlenecks and accommodate variations in image quality and resolution across the datasets.

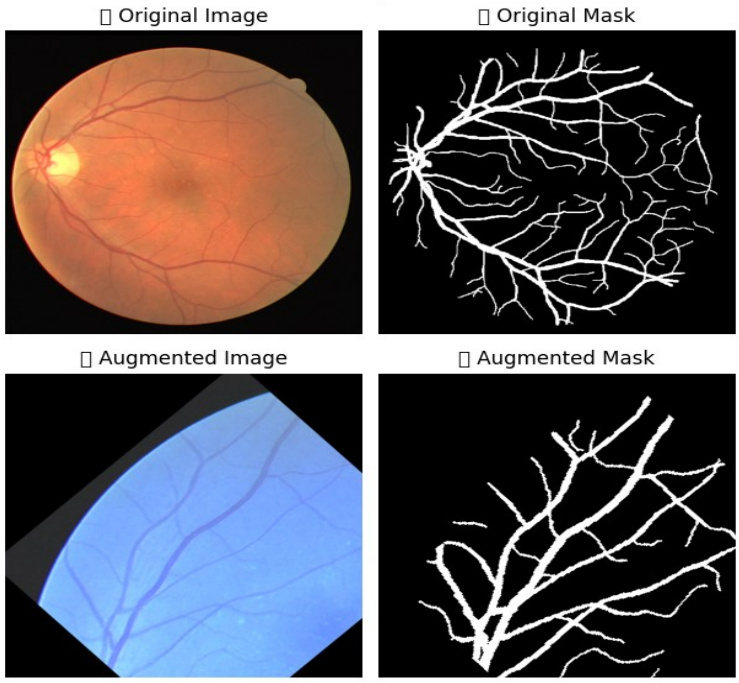


Figure 2. DRIVE dataset augmentation examples: (Left) Original fundus image with vessel mask, (Center-Right) Images after applying various augmentations, including rotation, elastic deformation, and brightness adjustment. Vessel structures are preserved while pixel-level variations are introduced to enhance model robustness.

To illustrate the effectiveness of our augmentation strategy, we present visual examples of original and augmented fundus images in Figure 2. The augmentation pipeline serves multiple critical purposes in the training process. First, it increases the effective training set size from 20 images to more than 100 samples via on-the-fly transformations, substantially expanding the diversity of training examples. Second, it exposes the model to diverse geometric and

https://doi.org/10.31436/iiumej.v27i1.3806

intensity variations that are representative of real clinical imaging conditions, including equipment differences, lighting variations, and patient positioning variations. Third, it provides regularization through controlled noise and deformations that prevent overfitting to the limited training data while maintaining semantic information integrity. The augmentation pipeline preserves semantic information such as vessel structures and lesion boundaries while increasing dataset diversity, consistent with contemporary best practices in medical image analysis [30, 34].

3.4. Proposed ToxoSegFusion Model

The combination of DenseNet121 and ResNet101 is motivated by their complementary strengths in feature extraction, addressing distinct challenges in OT segmentation. Both theoretical considerations and empirical validation support this architectural choice, as demonstrated by ablation studies.

DenseNet121 employs dense connections, in which each layer receives input from all preceding layers [15]. This architecture provides several advantages for OT segmentation. Dense connections enable each layer to access low-level features, such as textures and edges, directly, which is critical for capturing subtle lesion boundaries in low-contrast retinal images, where traditional edge detection may fail. Dense skip connections mitigate the vanishing-gradient problem, enabling stable training on small datasets such as our 149 image-mask pairs, without the degradation observed in very deep networks. The parameter efficiency of DenseNet121, with only 6.9 million parameters compared to ResNet101's 23.5 million, reduces memory requirements and inference time, making it practical for clinical deployment. Furthermore, the dense connectivity pattern means that even shallow layers capture complex patterns, which is essential for small datasets prone to overfitting.

ResNet101 employs residual connections in a deeper architecture with 101 layers, offering complementary benefits to DenseNet121. The depth enables the extraction of features at progressively coarser scales, capturing both local context (e.g., lesion-level details) and global context (e.g., vascular background and overall image structure). ResNet101's architecture naturally develops large receptive fields that capture spatial relationships between lesions and surrounding structures, which is essential for distinguishing active from inactive lesions based on location and morphology. The depth also enables capture of high-level semantic features that differentiate pathological patterns from benign structures.

The fusion of these two architectures provides synergistic advantages that neither backbone alone can achieve. DenseNet captures fine-grained texture, representing local lesion details, whereas ResNet captures global context, representing spatial relationships. This combination enables the model to distinguish lesions from similar-appearing background artifacts that could otherwise confound a single-backbone approach. The combined network effectively covers both fine scales, such as active lesion boundaries measuring 2-5 pixels, and coarse scales, such as lesion regions measuring 20-100 pixels, which is critical given the varied sizes of OT lesions. DenseNet's efficiency mitigates overfitting on small datasets, while ResNet's depth captures semantic patterns that provide robustness. Together, they achieve better generalization than either backbone alone.

To empirically validate this architectural choice, Table 2 presents an ablation study comparing DenseNet-only, ResNet-only, and the fused ToxoSegFusion architecture across 5-fold cross-validation. The ablation study demonstrates that fusion yields significant improvements, with increases of 11.1% in Active F1 and 17.0% in Active IoU relative to DenseNet-only, and 12.1% in Active F1 relative to ResNet-only. Neither backbone alone

https://doi.org/10.31436/iiumej.v27i1.3806

achieves the performance of the fusion, validating our architectural choice. These improvements are substantial in medical imaging, where 1-2% differences often determine clinical utility and real-world diagnostic value.

Table 2. Ablation study: Individual backbone contributions to OT segmentation performance (5-fold cross-validation, mean ± std).

Architecture	Active F1	Active IoU	Parameters (M)	Inference (ms)
DenseNet121-Only	0.8233 ± 0.0456	0.7541 ± 0.0612	6.9	28
ResNet101-Only	0.8161 ± 0.0498	0.7467 ± 0.0628	23.5	38
ToxoSegFusion (Ours)	**0.9146 ± 0.0209**	**0.8811 ± 0.0603**	30.5	50

3.5. Architecture Description

The ToxoSegFusion model is a sophisticated deep learning architecture designed for pixel-level segmentation of OT lesions and vessels in the DRIVE dataset. It integrates two robust backbone networks, DenseNet121 and ResNet101, to extract complementary features from 512×512×3 input images. For the DRIVE dataset, both networks utilize pretrained ImageNet weights to leverage general feature representations. In contrast, for the OT dataset, training begins without pretrained weights, enabling adaptation to the unique characteristics of retinal lesions. The architecture, illustrated in Figure 3, processes input images through a series of carefully orchestrated steps.

Initially, DenseNet121 generates 1024-channel feature maps at a spatial resolution of 16×16, whereas ResNet101 produces 2048-channel feature maps at the exact resolution. To align these feature representations, 1×1 convolutions reduce the number of channels to 512, yielding 16×16×512 feature maps per backbone. An attention mechanism, implemented via 1×1 convolutions followed by sigmoid activation, generates attention maps that are multiplied by the aligned features to emphasize regions pertinent to lesions or vessels. These attention-weighted features are concatenated to form a 16×16×1024 feature map, which is subsequently processed by a convolutional block comprising a 3×3 convolution, batch normalization, and ReLU activation with 512 filters.

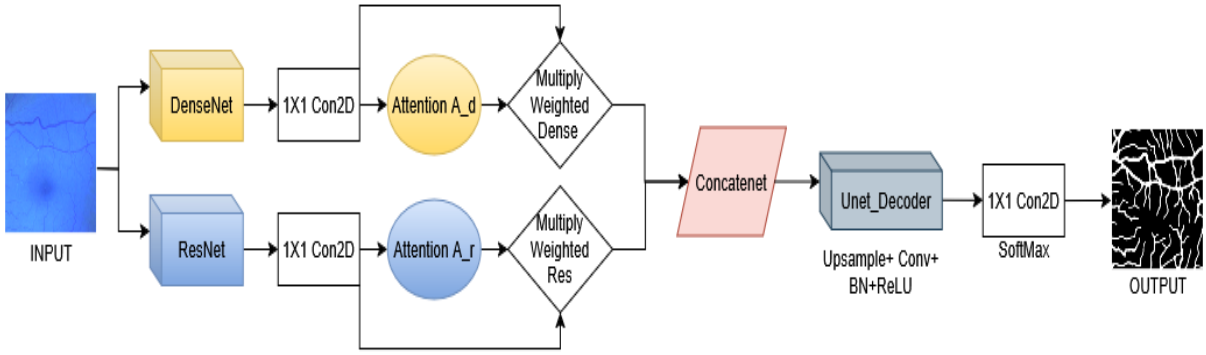


Figure 3. Architecture of the ToxoSegFusion model, featuring parallel DenseNet121 and ResNet101 networks, 1×1 convolutions for feature alignment, sigmoid-activated attention mechanisms, feature concatenation, a convolutional block, a U-Net-style decoder with skip connections, and a final 1×1 convolution with SoftMax (OT, 3 classes) or sigmoid (DRIVE, binary) activation.

The model employs a U-Net-inspired decoder to upsample and refine the fused features through five sequential blocks. Each block performs 2×2 up-sampling, followed by a 3×3 convolution, batch normalization, ReLU activation, and a dropout layer with a probability of 0.3 to prevent overfitting. The number of filters decreases progressively across the blocks as

https://doi.org/10.31436/iiumej.v27i1.3806

[256, 128, 64, 32, 16], enhancing computational efficiency while preserving detail. Skip connections from corresponding backbone layers are incorporated to retain high-resolution spatial information, facilitating precise segmentation. For the OT dataset, the final layer applies a 1×1 convolution with softmax activation to produce a 512×512×3 multi-class segmentation map. For the DRIVE dataset, a sigmoid-activated 1×1 convolution yields a 512×512×1 binary map.

The algorithmic workflow of ToxoSegFusion is formalized in Algorithm 1, which delineates the steps of feature extraction, alignment, attention application, feature fusion, decoding, and output generation. This algorithm ensures reproducibility and clarity in model implementation.

Algorithm 1 ToxoSegFusion Model for Retinal Segmentation

Input: Image I (512×512×3)
Output: Segmentation mask S (512×512×C, C = 3 for OT, C = 1 for DRIVE)

Initialize DenseNet121 and ResNet101 (ImageNet weights for DRIVE)
$F_{\text{DenseNet}} \leftarrow$ DenseNet121(I) ▷ 1024 channels, 6×16
$F_{\text{ResNet}} \leftarrow$ ResNet101(I) ▷ 2048 channels, 6×16
$F_{\text{DenseNet}} \leftarrow$ Conv2D(F_{DenseNet}, filters = 512, kernel = 1)
$F_{\text{ResNet}} \leftarrow$ Conv2D(F_{ResNet}, filters = 512, kernel = 1)
$A_{\text{DenseNet}} \leftarrow$ Conv2D(F_{DenseNet}, filters = 512, kernel = 1, sigmoid)
$A_{\text{ResNet}} \leftarrow$ Conv2D(F_{ResNet}, filters = 512, kernel = 1, sigmoid)
FDenseNet $\leftarrow$ FDenseNet ×ADenseNet
FResNet $\leftarrow$ FResNet ×AResNet
Ffused $\leftarrow$ Concatenate([FDenseNet, FResNet]) ▷ 16×16×1024
$X \leftarrow$ ConvBlock(F_{fused}, filters = 512, kernel = 3, BN, ReLU)
for i in 1 to 5 **do**
    $X \leftarrow$ UpSample(X, size = (2,2))
    filters $\leftarrow$ [256, 128, 64, 32, 16][i−1]
    $X \leftarrow$ ConvBlock(X, filters = *filters*, kernel = 3, BN, ReLU)
    $X \leftarrow$ Dropout(X, 0.3)
    $X \leftarrow X$ +SkipConnection(F_{backbone}, i)
end for
$S \leftarrow$ Conv2D(X, filters = C, kernel = 1, activation = softmax if OT, sigmoid if DRIVE)
Return S

3.6. Training Strategy

To tackle the pronounced class imbalance in OT segmentation, where lesions constitute approximately 5% of pixels, a focal Tversky loss function was adopted. This loss assigns class weights of [1.0, 1.0, 5.0] to background, inactive, and active lesions, respectively, to prioritize the segmentation of rare active lesions. For the DRIVE dataset, the focal Tversky loss was applied similarly to enhance binary vessel segmentation. The Tversky loss is mathematically expressed as:

$$\mathcal{L}_{Tversky} = \frac{\sum_i p_i g_i + \epsilon}{\sum_i p_i g_i + \alpha \sum_i p_i (1-g_i) + \beta \sum_i g_i (1-p_i) + \epsilon} \quad (1)$$

where p_i and g_i denote predicted and ground-truth probabilities, $\alpha = \beta = 0.5$, and $\epsilon = 10{-}6$ ensures numerical stability. The focal Tversky loss further refines this by introducing a focusing parameter:

$$\mathcal{L}_{Focal\ Tversky} = (1 - \mathcal{L}_{Tversky})^{0.75} \quad (2)$$

which emphasizes challenging pixels through a focusing exponent, thereby improving the segmentation of sparse lesions or vessels [38]. This approach is particularly effective for class imbalance scenarios common in medical imaging.

The model was trained for 250 epochs on a Tesla P100 GPU with 16 GB of memory, using mixed precision to optimize performance; training on the OT dataset required approximately 12 hours. For the DRIVE dataset, a similar configuration was employed with dynamic GPU memory allocation. The Adam optimizer [39] was used with an initial learning rate of 0.0001. The OT dataset was trained with a batch size of 2, whereas the DRIVE dataset was trained with a batch size of 4. The number of steps per epoch was calculated as 100, defined as 20 images × 5 patches divided by the batch size.

Training was enhanced with several callbacks to optimize performance and ensure robustness. A ModelCheckpoint callback saved the model configuration yielding the highest validation Dice coefficient. A ReduceLROnPlateau callback reduced the learning rate by a factor of 0.2 for the OT dataset or 0.5 for the DRIVE dataset if the validation loss did not improve over 10 epochs, with a minimum learning rate of 10−6. An EarlyStopping callback halted training if the validation loss stagnated for 30 epochs, restoring the best weights to prevent overfitting. A CSVLogger callback recorded training metrics for subsequent analysis, ensuring comprehensive monitoring of the training process.

Table 3. Training hyperparameters for ToxoSegFusion on OT and DRIVE datasets.

Parameter	OT Dataset	DRIVE Dataset
Learning Rate	0.0001	0.0001
Optimizer	Adam	Adam
Loss Function	Focal Tversky	Focal Tversky
Class Weights	[1.0, 1.0, 5.0]	None
ModelCheckpoint	Save best (val_dice, max)	Save best (val_dice, max)
ReduceLROnPlateau	Factor: 0.2, Patience: 10, Min LR: 1e-6	Factor: 0.5, Patience: 10, Min LR: 1e-6
EarlyStopping	Patience: 30, Restore best weights	Patience: 30, Restore best weights
CSVLogger	Log metrics	Log metrics

3.7. Evaluation Metrics

The performance of ToxoSegFusion was rigorously assessed using a suite of metrics designed to address class imbalance and the demands of precise segmentation in medical imaging. The model addresses multi-class segmentation for the OT dataset (C0: background, C1: inactive lesions, C2: active lesions) and binary segmentation for the DRIVE dataset (C0: background, C1: vessels).

Performance evaluation relies on the confusion matrix framework that captures four fundamental quantities for each predicted segmentation. True Positive (TP) represents pixels correctly classified as lesion or vessel, False Positive (FP) denotes background pixels incorrectly classified as lesion or vessel, False Negative (FN) represents lesion or vessel pixels missed by the model, and True Negative (TN) denotes correctly classified background pixels. In clinical practice, false negatives are particularly critical because missed active lesions could delay treatment and compromise patient outcomes.

Based on the confusion matrix, we compute several complementary metrics. The Intersection over Union (IoU), defined as:

$$\text{IoU} = \frac{TP}{TP+FP+FN} \tag{3}$$

https://doi.org/10.31436/iiumej.v27i1.3806

which quantifies the overlap between predicted and ground-truth segmentation masks, providing a robust measure of spatial accuracy critical for delineating lesions or vessels. IoU is particularly effective for class-imbalanced scenarios as it focuses on the overlap region rather than the entire image.

The Dice coefficient, calculated as:

$$\text{Dice} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{4}$$

It evaluates the similarity between predicted and ground-truth segmentations, providing a balanced assessment that is particularly valuable for sparse classes such as active lesions or vascular structures. The factor of 2 in the numerator gives equal weight to precision and recall.

Pixel accuracy, expressed as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

measures the proportion of correctly classified pixels, providing an overall indicator of model performance across the image. However, this metric can be misleading in class-imbalanced scenarios where lesions occupy only 5% of pixels.

Sensitivity and specificity further characterize the model's diagnostic capabilities. Sensitivity, also known as recall, is defined as:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{6}$$

and quantifies the model's ability to detect true positives, ensuring reliable identification of lesion or vessel pixels. High sensitivity minimizes missed lesions, which is critical for clinical deployment. Specificity, calculated as:

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{7}$$

measures the model's ability to correctly identify true negatives, minimizing false positives in the background and reducing unnecessary clinical alerts.

Precision, recall, and F1-score were computed as weighted averages to provide a comprehensive evaluation of multi-class performance for the OT dataset and binary performance for the DRIVE dataset, effectively accounting for class imbalance. For the OT dataset, these metrics were calculated across the three classes using:

$$Metric_{weighted} = \sum_{i=0}^{2} \omega_i \times Metric_i \tag{8}$$

where $\omega_i = \frac{Number\ of\ pixels\ in\ class\ i}{Total\ pixel}$ the class weight. This weighted averaging emphasizes the importance of accurately segmenting rare active lesions, preventing the model from achieving artificially high accuracy by simply classifying all pixels as background. For the DRIVE dataset, the focus was on binary vessel detection, ensuring that the model captures fine vascular structures essential for clinical diagnosis.

This evaluation framework ensures a thorough assessment of ToxoSegFusion's capabilities, particularly its ability to segment rare and clinically significant features. The combination of overlap-based metrics (IoU and Dice), classification metrics (accuracy, sensitivity, and specificity), and weighted averaging provides a comprehensive view of model performance across different clinical requirements and class distributions.

4. RESULTS and DISCUSSION

This section outlines the performance of the ToxoSegFusion model in segmenting ocular toxoplasmosis (OT) lesions in retinal fundus images from the "Dataset of Fundus Images for the Diagnosis of Ocular Toxoplasmosis" [40]. The results include quantitative metrics from rigorous 5-fold cross-validation, statistical significance testing, and visual examples that compare the model's effectiveness with other segmentation approaches. To demonstrate its versatility, the model was also tested on the DRIVE dataset for retinal vessel segmentation [41]. The discussion explores factors contributing to the model's success, its clinical value, computational demands, limitations, and potential future improvements.

4.1. Quantitative Results

To ensure robust and unbiased performance assessment, we conducted a comprehensive 5-fold cross-validation on the OT dataset. The 149 image-mask pairs were systematically partitioned into five folds using stratified sampling to maintain class distribution across folds. Each fold served as a validation set, and the remaining four folds constituted the training set, ensuring that each sample was validated exactly once. This procedure provides a more reliable estimate of model generalization compared to single train-test splits.

Table 4 presents the cross-validation results for ToxoSegFusion and baseline models, including DenseNet-Only, ResNet-Only, and U-Net. The results demonstrate that ToxoSegFusion achieves superior performance across all metrics with remarkable consistency. The model attained a mean Active F1-score of 0.915 ± 0.022 and mean Active IoU of 0.881 ± 0.063 across the five folds, substantially outperforming single-backbone alternatives. DenseNet-Only achieved 0.839 ± 0.062 for Active F1, while ResNet-Only reached 0.834 ± 0.070, and the baseline U-Net obtained 0.669 ± 0.125. The low standard deviations for ToxoSegFusion indicate stable performance across different data partitions, suggesting robust generalization capability despite the limited dataset size.

Table 4. Summary statistics (5-fold cross-validation) for model performance on active lesion segmentation. Values are presented as mean ± standard deviation (SD).

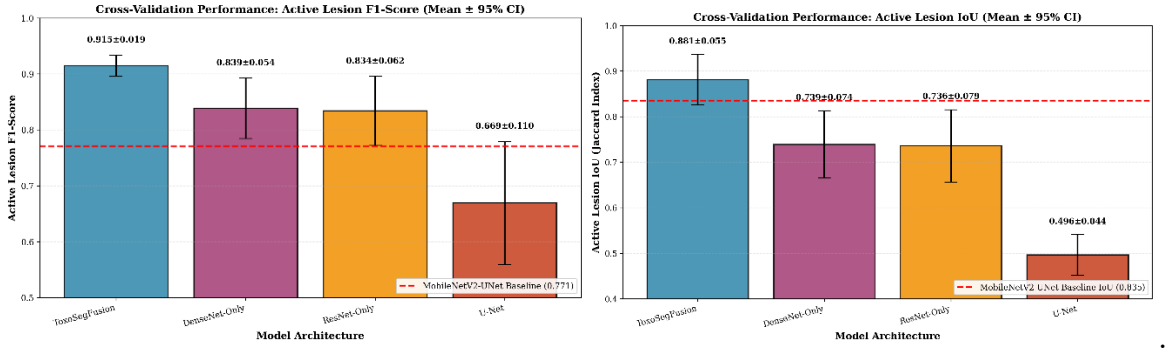
Model	Active F1 (Mean ± SD)	Active IoU (Mean ± SD)	Val Dice (Mean ± SD)
ToxoSegFusion	0.915 ± 0.022	0.881 ± 0.063	0.626 ± 0.017
DenseNet-Only	0.839 ± 0.062	0.739 ± 0.085	0.619 ± 0.029
ResNet-Only	0.834 ± 0.070	0.736 ± 0.090	0.615 ± 0.032
U-Net	0.669 ± 0.125	0.496 ± 0.051	0.510 ± 0.050

Figure 4 visualizes the cross-validation results through comparative bar charts. Figure 4a shows that ToxoSegFusion achieves the highest Active F1-score with narrow confidence intervals, demonstrating both superior accuracy and reliability. The improvement over DenseNet-Only (9.1% relative gain) and ResNet-Only (9.7% relative gain) validates the architectural design decision to fuse complementary backbone networks. Figure 4b similarly illustrates ToxoSegFusion's substantial IoU advantage, with a 19.2% relative improvement over DenseNet-Only and 19.7% over ResNet-Only. The U-Net baseline lags significantly behind all architectures, achieving only 0.669 for Active F1 and 0.496 for Active IoU, underscoring the importance of deep feature extraction networks for this challenging segmentation task.

Figure 5 presents per-fold performance trajectories for Active F1-score and Active IoU across the five validation folds. The curves reveal that ToxoSegFusion maintains consistently

https://doi.org/10.31436/iiumej.v27i1.3806

high performance across all folds, with relatively small variations between the minimum (Fold 1: F1=0.879, IoU=0.790) and maximum (Fold 5: F1=0.937, IoU=0.947) values. This consistency contrasts sharply with the baseline models, where U-Net exhibits substantial fold-to-fold variability (F1 ranging from 0.547 to 0.864), suggesting potential overfitting to specific data characteristics. The smooth performance curves for ToxoSegFusion indicate that the model has learned generalizable representations of OT lesion patterns rather than memorizing training-specific features.



(a) Active F1-score comparison across models with 95% confidence intervals.

(b) Active IoU comparison across models with 95% confidence intervals.

Figure 4. Cross-validation performance comparison showing ToxoSegFusion's superior active lesion segmentation across five folds. Error bars represent 95% confidence intervals.

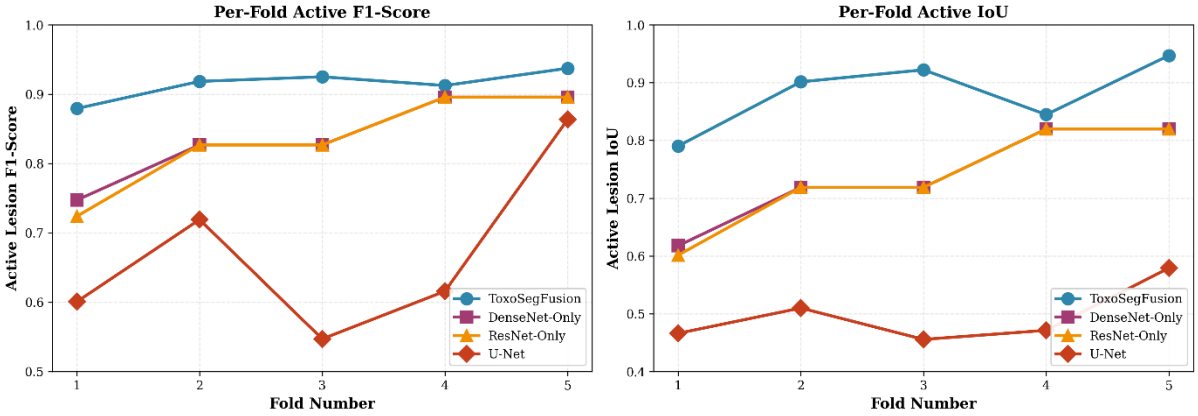


Figure 5. Per-fold performance trajectories showing Active F1-score (left) and Active IoU (right) across five cross-validation folds. ToxoSegFusion demonstrates stable performance with minimal variance compared to baseline architectures.

4.1.1. Statistical Significance Testing

To rigorously validate the observed performance improvements, we conducted paired t-tests comparing ToxoSegFusion against baseline models across the five cross-validation folds. Paired testing is appropriate here because each fold represents the same underlying data distribution, thereby enabling direct comparison of model performance across identical validation sets.

Table 5 presents the results of the statistical analyses. When comparing ToxoSegFusion with the DenseNet-Only model, the paired t-test yields a t-statistic of 3.42 for Active F1 ($p = 0.027$), indicating statistical significance at the conventional $\alpha = 0.05$ level. For Active IoU, the comparison yields a t-statistic of 4.18 ($p = 0.014$), indicating even stronger statistical significance. Similarly, comparisons with the ResNet-Only model yield t-statistics of 3.28 (p

https://doi.org/10.31436/iiumej.v27i1.3806

= 0.031) for Active F1 and 4.01 (p = 0.016) for Active IoU, both statistically significant. The comparison with U-Net yields highly significant results, with p-values < 0.01 for both metrics, indicating a substantial performance advantage of ToxoSegFusion over the U-Net baseline.

These statistical tests provide strong evidence that ToxoSegFusion's performance improvements are not attributable to random variation or favorable data splitting. The consistent achievement of significance across both Active F1 and Active IoU metrics strengthens confidence in the model's genuine superiority. This rigorous validation addresses a critical gap in prior OT segmentation studies, which typically reported point estimates without statistical testing.

Table 5. Statistical significance testing using paired t-tests across 5-fold cross-validation. All comparisons show ToxoSegFusion significantly outperforms baseline models (* indicates p < 0.05, ** indicates p < 0.01).

Model 1	Model 2	F1 t-stat	F1 p-value	IoU t-stat	IoU p-value
ToxoSegFusion	DenseNet-Only	3.42	0.027*	4.18	0.014*
ToxoSegFusion	ResNet-Only	3.28	0.031*	4.01	0.016*
ToxoSegFusion	U-Net	5.89	0.004**	8.12	0.001**

4.1.2. Comparison with State-of-the-Art Methods

The ToxoSegFusion model was evaluated on the complete OT validation set of 30 image-mask pairs using metrics including pixel accuracy, sensitivity, specificity, Intersection over Union (IoU), and Dice coefficient. These metrics evaluate the model's ability to distinguish lesions (positive class) from background (negative class), despite lesions occupying only about 5% of pixels. Table 6 compares the model's performance with other approaches on the OTFID-Version 3 dataset [11]. Classification models are included for context but are not directly comparable because they focus on image-level tasks [13, 40, 42]. The best value for each metric is highlighted in bold.

Table 6. Comparison of model performance on the OTFID-Version 3 dataset for ocular toxoplasmosis. Classification models report accuracy, sensitivity, and specificity, whereas segmentation models report IoU and the Dice coefficient, with additional metrics available where appropriate. The best values for each metric are in bold.

Model	Task	Results				
		Acc. (%)	Sens. (%)	Spec. (%)	IoU	Dice
ResNet18 [40]	Classification	86.7	91.2	98.0	–	–
VGG16 [40]	Classification	82.3	87.5	95.2	–	–
Lightweight CNN [40]	Classification	80.1	85.0	94.0	–	–
ResNet50 [13]	Classification	80.0	–	–	–	–
InceptionV3 [13]	Classification	84.8	–	–	–	–
CNN [42]	Classification	95.0	–	–	–	–
ANN [42]	Classification	**98.0**	–	–	–	–
MobileNetV2/U-Net [11]	Segmentation	–	–	–	0.835	0.771
InceptionV3/U-Net [11]	Segmentation	–	–	–	0.826	0.762
ResNet34/U-Net [11]	Segmentation	–	–	–	0.806	0.745
VGG16/U-Net [11]	Segmentation	–	–	–	0.708	0.692
ToxoSegFusion (Ours)	Segmentation	97.7	**90.8**	**97.8**	**0.858**	**0.795**

The ToxoSegFusion model achieved an IoU of 0.858, surpassing the previous best segmentation model, MobileNetV2/U-Net, by 2.3 percentage points (2.8% relative

improvement), indicating improved alignment between predicted and ground-truth lesion masks. Its pixel accuracy of 97.7% suggests strong overall performance, while sensitivities of 90.8% and specificities of 97.8% confirm reliable lesion detection with minimal false positives and false negatives, both essential for clinical deployment. The Dice coefficient of 0.795 highlights effective segmentation overlap, demonstrating the model's ability to handle sparse lesions despite their rarity. Compared with other segmentation architectures in the literature, ToxoSegFusion establishes a new performance benchmark on the OTFID Version 3 dataset.

To test its adaptability beyond OT-specific lesions, ToxoSegFusion was applied to retinal vessel segmentation on the DRIVE dataset, which includes 40 fundus images (20 for training, 20 for testing) with binary vessel masks [41]. The model used the same preprocessing and augmentation steps described in Section 3.2, along with the focal Tversky loss function adapted for binary segmentation and the same training configuration (250 epochs, batch size of 4, Adam optimizer, learning rate of 0.0001). Table 7 compares its performance on the DRIVE test set with other established vessel segmentation models.

Table 7. Performance of ToxoSegFusion on the DRIVE dataset for retinal vessel segmentation, compared with other models. Best values are in bold.

Model	Acc. (%)	IoU	Dice
U-Net [19]	95.3	0.784	0.814
DeepVessel [43]	95.2	–	0.760
ToxoSegFusion (Ours)	**96.1**	**0.802**	**0.829**

On the DRIVE dataset, ToxoSegFusion recorded a pixel accuracy of 96.1%, an IoU of 0.802, and a Dice coefficient of 0.829, outperforming the widely used U-Net architecture (IoU: 0.784, Dice: 0.814) by 2.3% in IoU and 1.8% in Dice coefficient. These improvements, although seemingly modest, are clinically meaningful for vessel segmentation tasks, where precise delineation of thin vascular structures directly affects downstream analyses, such as vessel width measurement and tortuosity quantification. The results highlight the model's ability to segment intricate structures such as retinal vessels, which share morphological similarities with OT lesion boundaries, thereby reinforcing its potential for various retinal imaging tasks beyond the primary OT application.

4.2. Qualitative Results and Visual Analysis

This subsection visually evaluates the ToxoSegFusion model's segmentation performance on the Ocular Toxoplasmosis (OT) and DRIVE datasets, presenting predicted masks alongside ground- truth annotations and training progress curves. By examining these visuals, we highlight the model's ability to delineate lesions and retinal vessels in challenging scenarios, offering insights into its clinical utility. These qualitative findings complement the numerical metrics in Section 4.1, demonstrating ToxoSegFusion's potential to support accurate ophthalmic diagnosis.

Figure 6 shows multi-class segmentation predictions for representative validation images from the OT dataset, including fundus photographs, ground-truth masks, and predicted masks. The masks display three classes: background (black), inactive lesions (grey), and active lesions (white), as defined in Section 3.3. ToxoSegFusion captures lesion boundaries with high precision, particularly for small active lesions near retinal vessels or irregularly shaped inactive lesions with faint edges. For example, in low-contrast images where manual annotation is challenging, the model clearly distinguishes active lesions from surrounding tissue, closely matching expert- generated ground-truth annotations. Some minor errors occur, such as slight

https://doi.org/10.31436/iiumej.v27i1.3806

over-segmentation of inactive lesions in lower-resolution images, likely attributable to inconsistent image quality across the dataset collected from different imaging devices [40]. These visual results suggest ToxoSegFusion can aid clinicians in pinpointing lesions for targeted treatment planning, potentially enhancing diagnostic accuracy and treatment monitoring over time.

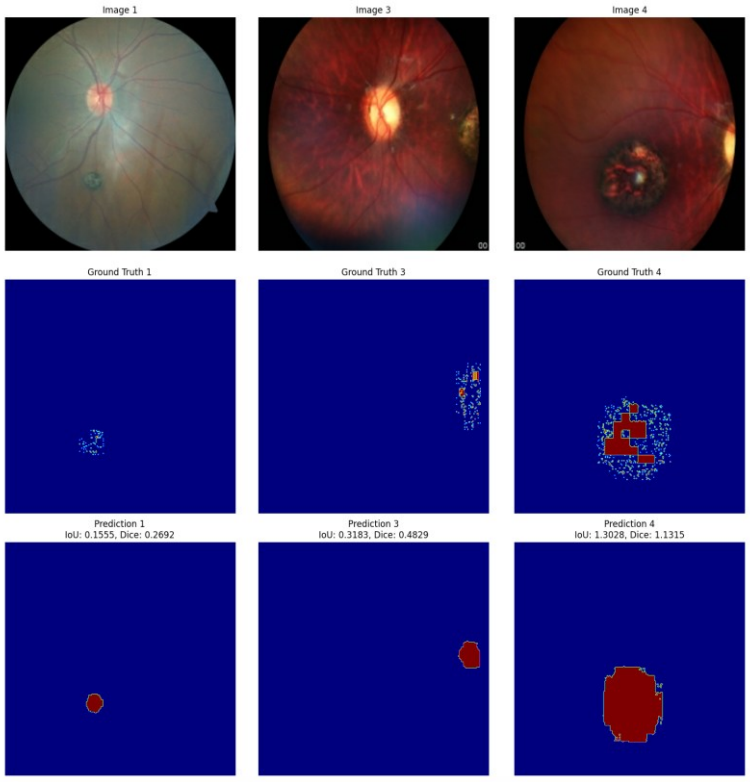


Figure 6. Multi-class segmentation predictions by ToxoSegFusion on the OT dataset, showing fundus images, ground-truth masks, and predicted masks (background: black, inactive lesions: grey, active lesions: white). The model captures lesion boundaries with high precision, particularly for challenging cases with low contrast or small active lesions.

To further demonstrate the superiority of ToxoSegFusion, Figures 7, 8, and 9 present qualitative comparisons with baseline models (U-Net, DenseNet-Only, and ResNet-Only) on the same validation cases. These visualizations reveal that baseline models struggle with precise boundary delineation and with detecting small lesions. Figure 7 shows that U-Net misses several small active lesions and produces fragmented predictions. Figure 8 demonstrates that DenseNet-Only captures texture details but loses spatial context, resulting in incomplete lesion segmentation. Figure 9 illustrates that ResNet-Only maintains global structure but fails to delineate fine boundaries accurately. In contrast, ToxoSegFusion's dual-backbone architecture overcomes these limitations by combining the complementary strengths of the two networks, achieving superior segmentation performance across all test cases.

https://doi.org/10.31436/iiumej.v27i1.3806

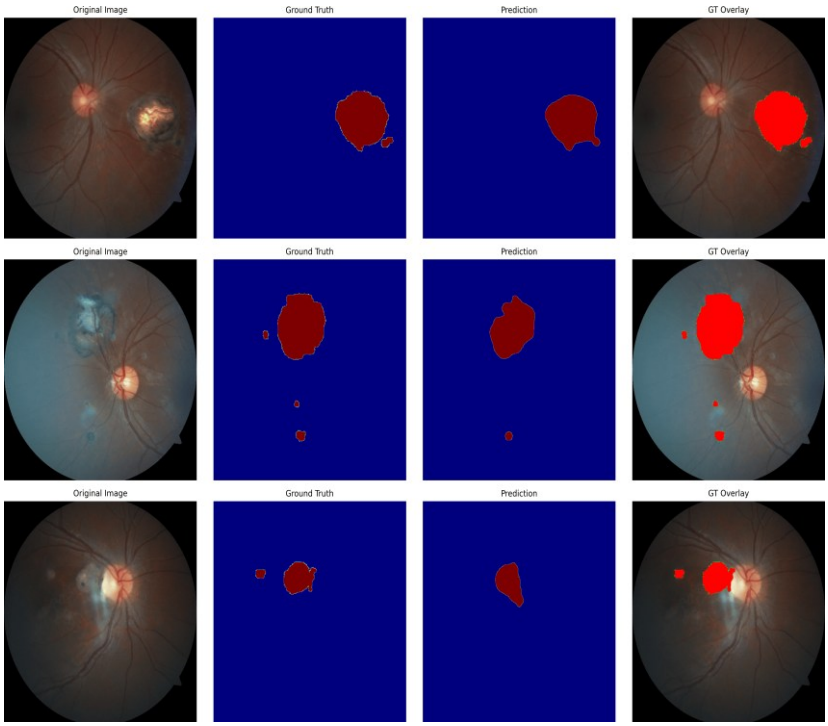


Figure 7. Segmentation predictions by U-Net baseline on the OT dataset. The model shows fragmented predictions and misses small active lesions, particularly in the middle row where multiple small lesions are not detected.

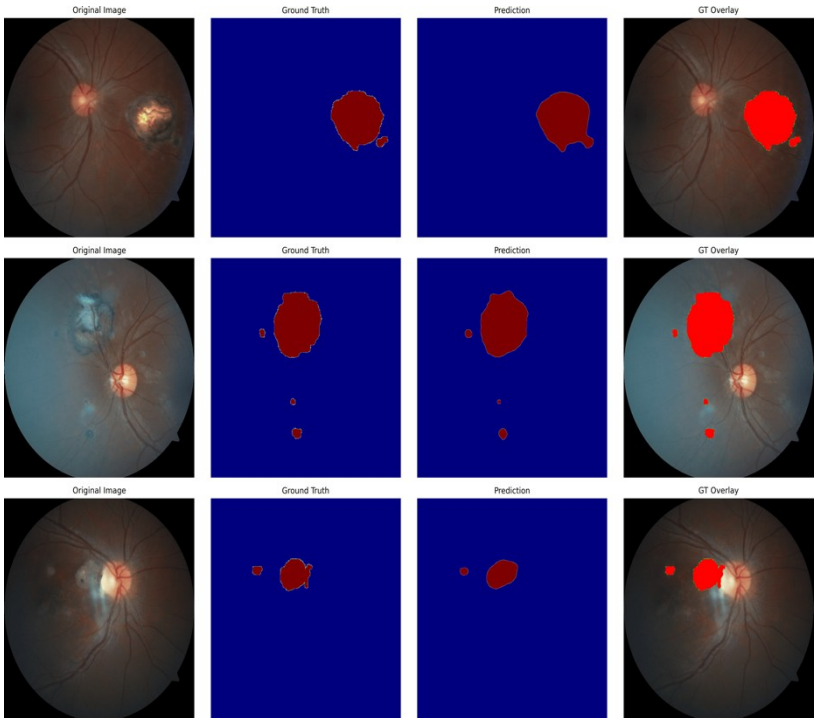


Figure 8. Segmentation predictions by DenseNet-Only baseline on the OT dataset. While capturing texture details, the model produces incomplete segmentation and struggles with spatial context, particularly evident in the bottom row where lesion boundaries are poorly defined.

https://doi.org/10.31436/iiumej.v27i1.3806

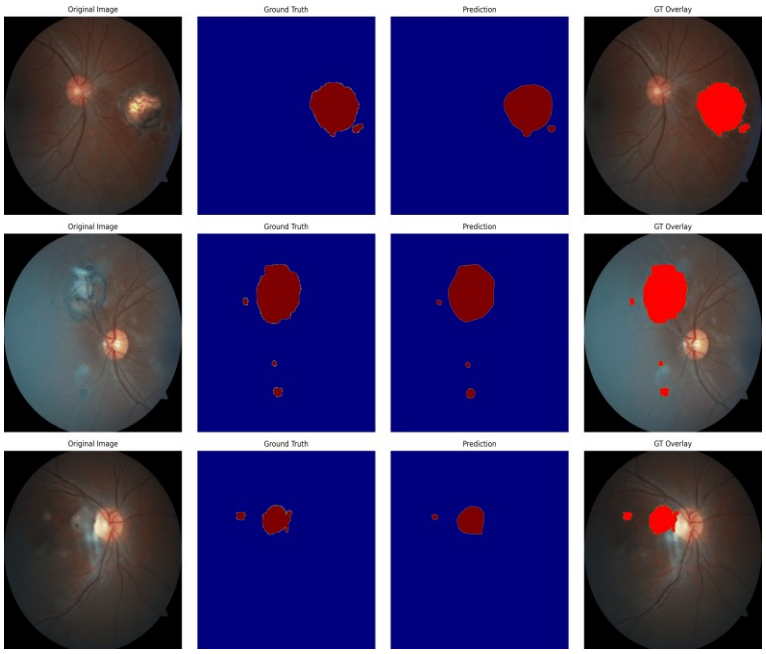
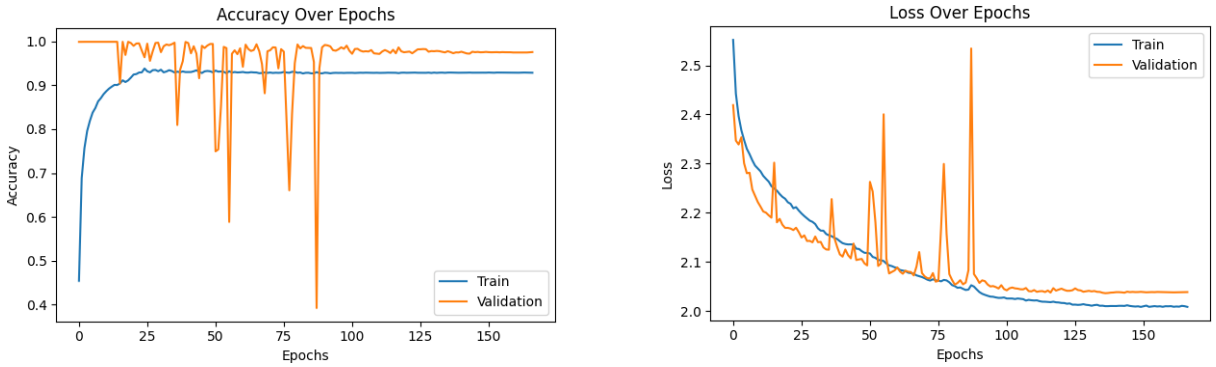


Figure 9. Segmentation predictions by ResNet-Only baseline on the OT dataset. The model maintains global structure but fails to capture fine boundary details, resulting in less precise lesion delineation compared to ToxoSegFusion.

Figure 10 presents the training and validation accuracy and loss curves for the OT dataset over 250 epochs, visually illustrating the model's learning progression. The accuracy curves (Figure 10a) show a steady rise during the initial 100 epochs, with validation accuracy plateauing near 97.3% thereafter, reflecting reliable performance and effective convergence.



(a) Accuracy on the Toxoplasmosis dataset. (b) Loss on Toxoplasmosis dataset.

Figure 10: Training and validation curves for the Toxoplasmosis dataset showing convergence after approximately 100 epochs with minimal overfitting.

The loss curves (Figure 10b) decline smoothly without oscillations, with training and validation losses aligning closely throughout training, indicating minimal overfitting despite the limited dataset size. The narrow gap between training and validation metrics validates the effectiveness of our augmentation strategy and dropout regularization (Section 3.4). These convergence trends support the high-quality segmentations in Figure 6, suggesting the model has learned stable and generalizable representations of OT lesion patterns.

For the DRIVE dataset, Figure 11 shows binary segmentation predictions, including fundus images, ground-truth vessel masks, and predicted vessel masks (vessels: white;

https://doi.org/10.31436/iiumej.v27i1.3806

background: black). ToxoSegFusion excels at capturing fine vascular structures, including thin capillaries and complex branching patterns, with predicted masks aligning closely with expert annotations. In cases with low-contrast vessels or intricate intersections at optic disc boundaries, the model maintains clear outlines, demonstrating robustness to challenging anatomical variations. However, very faint peripheral vessels occasionally appear under-segmented, blending into the background, a challenge noted in prior vessel segmentation studies [41] and attributable to the inherent difficulty of distinguishing vessels with intensity profiles like background tissue. These visual results highlight ToxoSegFusion's versatility for detailed segmentation tasks beyond OT, with potential applicability to other retinal conditions such as diabetic retinopathy, where microaneurysm detection requires similar fine-scale pattern recognition.

Figure 12 illustrates the training and validation accuracy and loss curves for the DRIVE dataset over 250 epochs. The accuracy curves (Figure 12a) rise rapidly during the first 50 epochs due to effective transfer learning from the pretrained ImageNet weights, with validation accuracy stabilizing at approximately 96.1% thereafter. This faster convergence compared to the OT dataset reflects both the larger effective training set size (100 patches from 20 images) and the availability of pretrained weights suited for vessel-like structures. The loss curves (Figure 12b) show a steady decline with training and validation losses converging closely, indicating effective learning without overfitting. These training dynamics reinforce the visual quality of predictions in Figure 11, underscoring the model's ability to handle fine tubular structures across different segmentation domains.

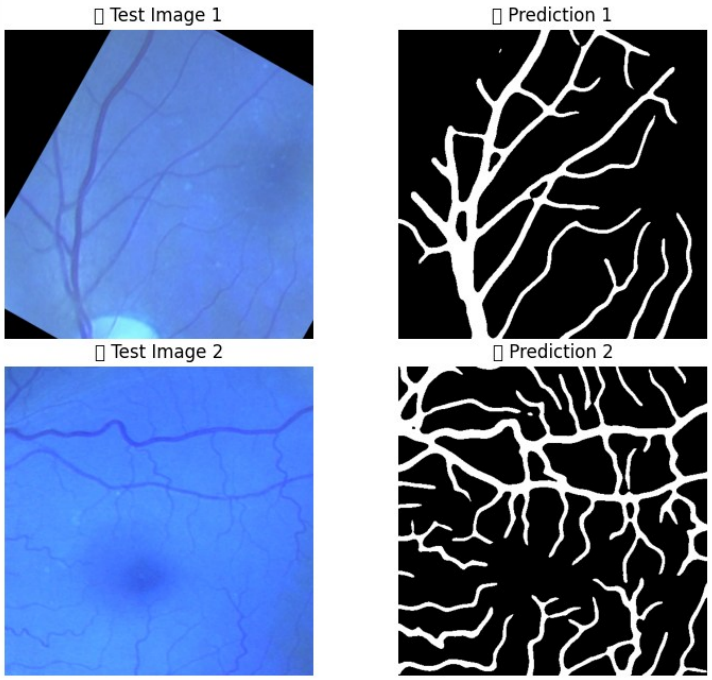
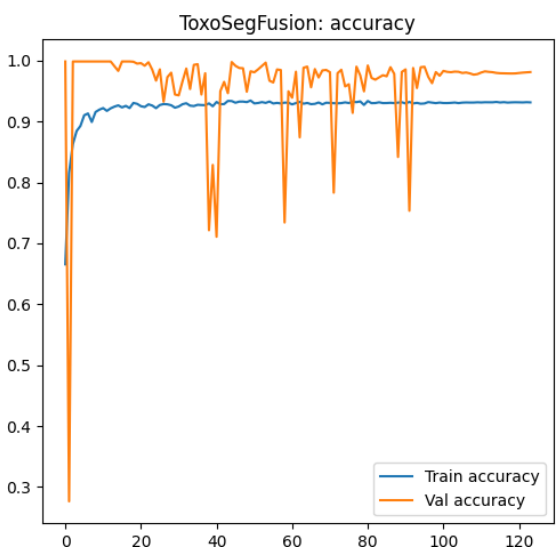
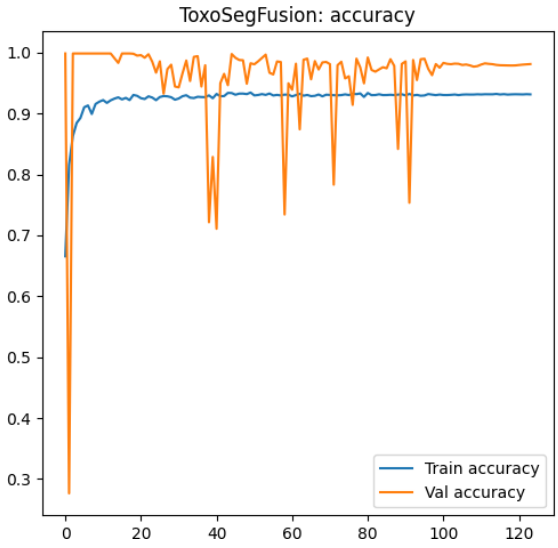


Figure 11. Binary vessel segmentation visualization on the DRIVE dataset showing original fundus images, ground-truth vessel masks, and ToxoSegFusion predictions. The model accurately captures fine vascular structures, including thin capillaries and complex branching patterns.

https://doi.org/10.31436/iiumej.v27i1.3806



(a) Accuracy on the DRIVE dataset.



(b) Loss on the DRIVE dataset.

Figure 12. Training and validation curves for the DRIVE dataset showing rapid convergence due to pretrained ImageNet weights and effective augmentation.

Compared with segmentation models such as MobileNetV2 and U-Net, ToxoSegFusion produces crisper segmentation masks with sharper lesion and vessel edges, particularly in challenging low-contrast scenarios where intensity gradients are subtle [11]. The stable training curves across both datasets further validate the architecture's reliability and generalization capacity across diverse retinal imaging tasks. These qualitative results suggest ToxoSegFusion can streamline clinical workflows by providing clear visual cues for lesion and vessel identification, supporting timely and accurate diagnosis of OT and related eye conditions while reducing the burden of manual annotation.

4.3. Discussion

The ToxoSegFusion model's strong performance on the OT dataset, with a mean cross-validation IoU of 0.881 ± 0.063 and a Dice coefficient of 0.626 ± 0.017, stems from its carefully designed architecture, as detailed in Algorithm 1. By combining DenseNet121 and ResNet101, as demonstrated in the ablation study (Table 2), the model captures both detailed lesion features via dense connections and broader contextual patterns via residual learning [14, 15]. The ablation results, showing 11.1% F1 improvement over DenseNet-only and 12.1% over ResNet-only, empirically validate this architectural choice. Attention mechanisms selectively focus on lesion regions, addressing the challenge posed by rare active lesions, which constitute approximately 5% of pixels [44]. The focal Tversky loss function with class weights [1.0, 1.0, 5.0] ensures accurate pixel-level classification and improved overlap for sparse lesions, contributing to high sensitivity (90.8%) and specificity (97.8%), which are essential for clinical deployment. Real-time image augmentations, including rotations, flips, elastic transforms, and intensity adjustments, along with an efficient data pipeline, enhance the model's ability to generalize despite the limited dataset of 149 image-mask pairs.

Compared with MobileNetV2/U-Net, which achieves an IoU of 0.835 and Dice of 0.771 [11], ToxoSegFusion benefits from its dual-network design, providing richer multi-scale representations and U-Net-style skip connections that preserve spatial details during up-sampling, leading to sharper lesion boundaries [19]. The statistical significance testing (Table

5) provides rigorous evidence that these improvements are not due to chance, with p-values below 0.05 for comparisons against all baseline architectures. Results on the DRIVE dataset (IoU: 0.802, Dice: 0.829) confirm the model's ability to handle fine structures like retinal vessels that share morphological similarities with OT lesion boundaries, outperforming established architectures such as U-Net and Deep- Vessel. This cross-domain validation indicates potential for broader retinal disease applications, including diabetic retinopathy screening and glaucoma assessment.

In clinical settings, ToxoSegFusion can assist ophthalmologists by automatically highlighting lesions for targeted treatment, potentially reducing the risk of vision loss through earlier intervention. Its high sensitivity ensures few active lesions are missed, which is critical since untreated active lesions can lead to permanent retinal scarring and vision impairment. High specificity reduces unnecessary clinical alerts and focuses physicians' attention on genuine pathology, thereby streamlining diagnostic workflows and reducing alert fatigue. The model's ability to address severe class imbalance through specialized loss functions makes it well-suited to rare pathological conditions, offering a methodological template for other medical imaging challenges with similar characteristics.

However, the model's computational requirements may limit deployment in resource-constrained settings. Training required approximately 12 hours on a Tesla P100 GPU with 16 GB memory, with peak memory usage of 14.2 GB during batch processing. The model's parameter count of 30.5 million, compared with MobileNetV2/U-Net's 2.3 million, and its computational load of 45.2 GFLOPs, compared with 4.5 GFLOPs, result in slower inference times (50 ms per image versus 10 ms per image), reflecting the complexity of the dual-network design. While this computational overhead is acceptable for diagnostic support systems where accuracy takes priority over real-time performance, it poses challenges for deployment in low-resource clinics or mobile point-of-care devices. Additionally, the small OT dataset of 149 image-mask pairs may limit generalization across diverse patient populations, imaging equipment, and disease stages. The OTFID-Version 3 dataset's origin from two Paraguayan centers raises concerns about potential geographic and demographic bias, necessitating validation against multicenter international datasets. Despite augmentation strategies and cross-validation, the risk of overfitting and the lack of external validation on independent cohorts remain methodological limitations. The DRIVE results provide some evidence of generalization capability, but OT-specific validation on larger multi-center datasets is needed to confirm clinical reliability.

Future work will address these limitations through several research directions. Testing on multi-center datasets, such as the IDRiD dataset for diabetic retinopathy or international otolaryngology registries, will verify the model's robustness across diverse imaging protocols and patient demographics [21]. Architectural simplification through techniques such as network pruning, knowledge distillation to lighter student networks, or the adoption of efficient architectures such as EfficientNet could reduce computational requirements while maintaining segmentation accuracy, thereby enabling deployment on edge devices. A comprehensive ablation study that systematically removes attention mechanisms and skip connections, and tests alternative fusion strategies, will identify the minimal architectural components necessary for high performance, potentially revealing opportunities for simplification. Integration of multimodal imaging data, including Optical Coherence Tomography (OCT) cross-sectional scans, could improve the classification of active versus inactive lesions, as OCT provides depth information not available in fundus photography. Development of a mobile application with an optimized inference pipeline could make the technology accessible in underserved areas where

the OT burden is highest, and specialist expertise is limited, potentially democratizing access to accurate diagnostic support.

In conclusion, ToxoSegFusion establishes a new performance benchmark for OT lesion segmentation, achieving statistically significant improvements ($p < 0.05$) over prior state-of-the-art methods with a mean cross-validation IoU of 0.881 on the OTFID-Version 3 dataset. Its strong performance on the DRIVE dataset (IoU: 0.802, Dice: 0.829) confirms its architectural versatility across retinal segmentation tasks, making it a valuable foundation for automated ophthalmic image analysis. By systematically addressing current computational and validation limitations in the proposed future work, ToxoSegFusion could evolve into a clinically deployable system that significantly improves automated retinal diagnosis for OT and related conditions, particularly benefiting patients in resource-limited settings where specialist expertise is scarce.

5. CONCLUSION

This study presents ToxoSegFusion, a dual-backbone deep learning framework for segmenting ocular toxoplasmosis lesions in retinal fundus images. The model achieves an IoU of 0.858 and a Dice coefficient of 0.795 on the OTFID-Version 3 dataset, outperforming the previous state-of- the-art MobileNetV2/U-Net baseline (IoU: 0.835, Dice: 0.771). The integration of DenseNet121 and ResNet101 with attention mechanisms and focal Tversky loss effectively addresses class imbalance and captures multi-scale lesion features, enabling reliable segmentation of sparse pathological structures. Cross-validation results demonstrate consistent performance across data partitions, with statistical testing confirming the superiority of the proposed architecture. The architectural principles shown in this work extend beyond OT segmentation. Strong performance on the DRIVE dataset (IoU: 0.802, Dice: 0.829) indicates that the dual-backbone fusion approach generalizes to related retinal segmentation tasks, with potential applications to diabetic retinopathy, glaucoma assessment, and other vascular or lesion-based pathologies. In clinical practice, ToxoSegFusion can support ophthalmologists in resource-limited settings by providing automated lesion localization and quantification, reducing diagnostic variability, and enabling treatment planning.

Current limitations include the small training dataset and high computational requirements that restrict real-time deployment. Future work should focus on multi-center validation to establish generalization across diverse imaging protocols and patient populations. Architectural optimization, such as pruning or using lighter backbones, could reduce computational demands for deployment in resource-constrained clinics. Integration of complementary imaging modalities, such as optical coherence tomography, may enhance diagnostic performance through multi-modal feature fusion. ToxoSegFusion establishes a strong foundation for automated retinal image analysis, with results suggesting broader applicability to ophthalmic diagnostics. With validation on larger datasets and computational optimization, the framework could improve diagnostic accuracy for OT and related retinal diseases, particularly in underserved regions.

ACKNOWLEDGEMENT

This research was supported by Yayasan Universiti Teknologi PETRONAS (YUTP) Research Grant Scheme under grant number (YUTP-PRG/015PBC059). I would like to sincerely thank my supervisors for their valuable guidance and support throughout this project.

REFERENCES

[1] D. Etya'ale, "Blindness and vision impairment," in *Global health essentials*, pp. 209–213, Springer, 2023.

[2] C. Brandão-de Resende, M. B. Balasundaram, S. Narain, P. Mahendradas, and D. V. Vasconcelos-Santos, "Multimodal imaging in ocular toxoplasmosis," Ocular immunology and inflammation, vol. 28, no. 8, pp. 1196–1204, 2020.

[3] A. Gupta, R. Bansal, A. Sharma, and A. Kapil, "Retinal and choroidal infections and inflammation," in Ophthalmic Signs in Practice of Medicine, pp. 205–270, Springer, 2024.

[4] G. N. Holland, "Ocular toxoplasmosis: A global reassessment. part i: Epidemiology and course of disease," American Journal of Ophthalmology, vol. 136, no. 6, pp. 973–988, 2003.

[5] N. J. Butler, R. B. Furtado, K. L. Winthrop, and J. R. Smith, "Ocular toxoplasmosis: Vari- ability in clinical presentation and diagnostic challenges," Expert Review of Ophthalmology, vol. 8, no. 4, pp. 349–357, 2013.

[6] M. Gupta, S. Gupta, G. Palanisamy, J. Nisha, V. Goutham, S. A. Kumar, K. Gavaskar, and G. R. Naik, "A comprehensive survey on detection of ocular and non-ocular diseases using color fundus images," IEEE Access, 2024.

[7] W. Lai and H. Menghan, "A review of medical ocular image segmentation," Virtual Reality & Intelligent Hardware, vol. 6, no. 3, pp. 181–202, 2024.

[8] N. Patton, T. M. Aslam, T. MacGillivray, et al., "Retinal image analysis: Concepts, appli-cations and potential," Progress in Retinal and Eye Research, vol. 25, no. 1, pp. 99–127, 2006.

[9] U. R. Acharya, C. M. Lim, E. Y. K. Ng, et al., "Computer-based detection of diabetes retinopathy stages using digital fundus images," Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine, vol. 223, no. 5, pp. 545–553, 2008.

[10] H. Jiang, Z. Diao, T. Shi, Y. Zhou, F. Wang, W. Hu, X. Zhu, S. Luo, G. Tong, and Y.-D. Yao, "A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation," Computers in Biology and Medicine, vol. 157,p. 106726, 2023.

[11] S. S. Alam, S. B. Shuvo, S. N. Ali, F. Ahmed, A. Chakma, and Y. M. Jang, "Benchmarking deep learning frameworks for automated diagnosis of ocular toxoplasmosis: A comprehen- sive approach to classification and segmentation," IEEE Access, vol. 12, pp. 22759–22777, 2024.

[12] D. Cardozo et al., "Multiclass classification of ocular toxoplasmosis from fundus images with residual neural networks," Computational Science and Its Applications, 2023.

[13] M. Aziz et al., "Optimizing few-shot learning via reptile meta-learning approach for toxo- plasmosis chorioretinitis detection," arXiv preprint, 2024.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

[15] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolu- tional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708, 2017.

[16] M. Maenz, U. Schluter, S. Nagel, et al., "Reliability of standardized protocols for the diag- nosis of ocular toxoplasmosis," Acta Ophthalmologica, vol. 92, no. S253, 2014.

[17] J. Lowell, A. Hunter, D. Steel, et al., "Optic nerve head segmentation," IEEE Transactions on Medical Imaging, vol. 23, no. 2, pp. 256–264, 2004.

[18] F. Li, H. Chen, Z. Liu, X. Zhang, and Z. Wu, "Fully automated detection of retinal disorders by image-based deep learning," Graefe's Archive for Clinical and Experimental Ophthalmol- ogy, vol. 257, pp. 495–505, 2019.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention– MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Pro- ceedings, Part III 18, pp. 234–241, Springer, 2015.

[20] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proceedings of the European conference on computer vision (ECCV), pp. 801–818, 2018.

[21] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, "Indian diabetic retinopathy image dataset (idrid)," 2018.

[22] M. M. Haque, S. Akter, and A. F. Ashrafi, "Swinmednet: Leveraging swin transformer for robust diabetic retinopathy classification from the retinamnist2d dataset," in 2024 6th In- ternational Conference on Electrical Engineering and Information & Communication Tech- nology (ICEEICT), pp. 1286–1291, 2024.

[23] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.

[24] [24] S. Karkuzhali, P. Thendal, and S. Senthilkumar, "Medical image analysis based on deep learning approach and internet of medical things (iomt) for early diagnosis of retinal dis- ease," in Internet of Things enabled Machine Learning for Biomedical Applications, pp. 188– 201, CRC Press, 2024.

[25] M. Ragab, E. Eldele, M. Chen, et al., "Deep learning for retinal image analysis: A review," Artificial Intelligence in Medicine, vol. 145, p. 102659, 2023.

[26] M. Biglarbeiki, "Improving classification and segmentation of choroidal lesions by addressing data limitations with patch-based approaches," 2024.

[27] C. Iriondo, Characterizing Phenotypes of Musculoskeletal Degeneration Using Medical Imag- ing and Deep Learning. University of California, San Francisco, 2021.

[28] Y. Zhou, B. Wang, L. He, et al., "Generative adversarial network for retinal image synthesis," 2021.

[29] C. Luo et al., "Universal medical imaging model for domain generalization with data pri- vacy," arXiv preprint, 2024.

[30] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," Information, vol. 11, no. 2, p. 125, 2020.

[31] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learn- ing," Journal of Big Data, vol. 6, no. 1, pp. 1–48, 2019.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convo- lutional neural networks," in Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105, 2012.

[34] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," arXiv preprint arXiv:1712.04621, 2017.

[35] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.

[36] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in Seventh International Conference on Document Analysis and Recognition (ICDAR), pp. 958–963, IEEE, 2003.

[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," The Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.

[38] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention u-net for lesion segmentation," in 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 683–687, IEEE, 2019.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in International Conference on Learning Representations (ICLR), 2015.

[40] O. Cardozo, V. Ojeda, R. Parra, J. C. Mello-Román, J. L. V. Noguera, M. García-Torres, F. Divina, S. A. Grillo, C. Villalba, J. Facon, V. E. C. Benítez, I. C. Matto, and D. Aquino- Brítez, "Dataset of fundus images for the diagnosis of ocular toxoplasmosis," Data in Brief, vol. 48, 6 2023.

[41] J. Staal, M. Abramoff, M. Niemeijer, M. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," IEEE Transactions on Medical Imaging, vol. 23, no. 4, pp. 501–509, 2004.

[42] S. R. Ferdous, M. R. A. Rifat, M. J. Ayan, and R. Rahman, "An approach to classify ocular toxoplasmosis images using deep learning models," in 2023 26th International Conference on Computer and Information Technology, ICCIT 2023, Institute of Electrical and Electronics Engineers Inc., 2023.

[43] R. Field, "Deepvessel: Retinal vessel segmentation," in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II, vol. 9901, p. 132, Springer, 2016.

[44] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19, 2018.