Experimental Evaluation of Public Retinal Vessel Segmentation Datasets (2020–2025) with Deep Learning: An Empirical Study

Robert Ngabo Mugisha
College of Engineering
Carnegie Mellon University Africa
Kigali, Rwanda
Rngabomu@alumni.cmu.edu

Geoffrey Munyaneza
Kenan Flagler Business School
University of North Carolina
North Carolina, United States
gmunyane@alumni.cmu.edu

Fidele Nsanzumukunzi
The Roux Institute
Northeastern University
Portland ME, United States
fnsanzum@alumni.cmu.edu

Mediatrice Dusenge
African Centre of Excellence in Data Science
University of Rwanda
Kigali, Rwanda
dusengemeddy@gmail.com

Josue Uzigusenga
College of Sciences and Technology
University of Rwanda
Kigali, Rwanda
uzigusengajosue@gmail.com

Theophilla Igihozo
College of Medicine and Health Sciences
University of Rwanda
Kigali, Rwanda
igihozot2@gmail.com

Fabrice Mpozenzi
Department of Computer Science
Abilene Christian University
Texas, Abilene, United States
fabricempozenzi@gmail.com

Emmanuella Nuwayo
College of Sciences and Technology
University of Rwanda
Kigali, Rwanda
emmanuellanuwayo@gmail.com

Prince Shema Musonerwa
College of Engineering
Carnegie Mellon University Africa
Kigali, Rwanda
musonerwaprince@gmail.com

Benny Uhoranishema
College of Engineering
Carnegie Mellon University Africa
Kigali, Rwanda
buhorani@alumni.cmu.edu

Jean De Dieu Niyonteze
Goizueta Business School
Emory University
Atlanta, GA 30322, United States
jniyont@emory.edu

***Abstract*— Retinal vessel segmentation is a cornerstone of computer-aided diagnosis systems for ophthalmic and systemic diseases. However, segmentation performance varies substantially across studies due to differences in dataset quality, annotation consistency, and imaging conditions. This study presents a comprehensive experimental evaluation of eleven publicly available retinal vessel segmentation datasets commonly used in research between 2020 and 2025, with eight of them fully assessed under identical conditions using the U-Net++ architecture and three included for reference and contextual comparison. The results reveal notable variability in segmentation accuracy driven by disparities in image resolution, sample size, and annotation precision. Among the evaluated datasets, the Retinal Vascular Tree Analysis (RETA) Benchmark achieved the best overall performance (Dice = 0.80 ± 0.015; Accuracy = 0.97 ± 0.016), outperforming all other datasets. CHASE-DB1, AV-DRIVE, HRF, and STARE also achieved Dice ≥ 70% and Accuracy > 95%, confirming their robustness and suitability for retinal vessel segmentation benchmarking. In contrast, smaller or older datasets such as Digital Retinal Images for Vessel Extraction (DRIVE), LES-AV from the LES-AV database, and IOSTAR vessel segmentation dataset produced moderate results but still captured major vessel structures effectively. These findings underscore the critical importance of dataset quality and standardization in achieving reliable segmentation outcomes and highlight the need to develop harmonized, multimodal, and large-scale retinal image datasets to advance reproducibility, generalization, and clinical applicability of deep learning–based retinal analysis.**

Keywords—Dataset Benchmarking, Deep Learning, Medical Image Analysis, Retinal Vessel Segmentation, U-Net++

I. INTRODUCTION

Retinal vessel segmentation is a crucial task in medical image analysis that serves as a foundational step in the diagnosis and monitoring of various ocular and systemic diseases. It enables the extraction of clinically significant vascular features, including vessel width branches [1], which function as key biomarkers for early detection and disease progression. The pathological signs, such as hemorrhages and microaneurysms, typically manifest around blood vessels; therefore, accurate segmentation is essential for their reliable detection [2].

Dataset benchmarking significantly influences outcomes, making transparency essential for reproducibility and meaningful comparison. Furthermore, comprehensive metadata not only validates reported results but also enables researchers to replicate analyses, experiment with modifications, and effectively debug models [3]. However, the reliability of these evaluations can be compromised when models are trained on noisy ground truth, which has been shown to significantly degrade performance [4]. To address such challenges in segmentation accuracy, Attention U-Net++ introduces an attention mechanism between nested convolutional blocks, allowing features extracted at different levels to be selectively merged based on task relevance [5]. In addition, U-Net++ is capable of automatically identifying and discarding redundant decoding blocks without loss of precision, thereby improving segmentation accuracy and reducing model complexity considerably [6].

979-8-3315-5723-2/26/$31.00 ©2026 IEEE

This paper benchmarks eleven datasets under identical conditions to recommend the most suitable one for retinal image segmentation. The findings help determine which datasets provide the best support for model development, generalization, and real-world deployment, thereby promoting reproducibility, fairness, and reliability in retinal image segmentation research.

II. Related Work

Over the past few years, publicly accessible datasets containing expert-labeled ground truths have become widely used for developing, validating, and comparing models. Retinal image segmentation models are also evaluated using these datasets. The segmentation of blood vessels is a challenging task due to their small size, structure, and color similarity to the background [7].

In addition, datasets often comprise diverse images that vary in region of interest, resolution, and applicability, introducing challenges for segmentation methods such as variations in illumination, intra-class differences, and complex backgrounds [8]. Therefore, the quality and standardization of images are crucial when selecting a dataset. For instance, models have achieved higher Dice scores on the Digital Retinal Images for Vessel Extraction (DRIVE) dataset, whereas the Structured Analysis of the Retina (STARE) dataset has yielded higher Area Under the Curve (AUC) values. This difference arises because DRIVE includes more finely labeled vessels in its ground truth, whereas STARE contains images with various retinal diseases, increasing its variability and complexity [7]. A benchmark comparison among the DRIVE, High-Resolution Fundus Segmentation (HRF), and CHASE-DB1 datasets led to the selection of the DRIVE dataset, as it achieved an accuracy above 95% [9]. Datasets containing high-quality expert annotations—ranging from image-level diagnostic labels to pixel-level vessel or lesion masks—are essential for training and validating supervised learning methods in retinal image analysis [10]. The size of training images in a dataset plays a key role in obtaining a model with good generalization ability free of overfitting and coarse prediction. It boosts the efficiency of the model and improves its overall performance [11]. Moreover, using a widely adopted dataset allows for meaningful comparisons with results from previously conducted studies.

In prior studies, for instance, Gabor-filtered retinal images from the green channel are fed into U-Net++ to segment blood vessels precisely. After determining the blood vessels in the binary image, Histogram of Oriented Gradients (HOG) and Local Binary Pattern (LBP) features are extracted and used to diagnose the disease in the retinal image, outperforming previous models on the DRIVE and Methods to Evaluate Segmentation and Indexing Techniques in the field of Retinal Ophthalmology (Messidor) datasets. The paper thus builds directly on prior U-Net++ applications in medical image segmentation to improve diagnostic accuracy in retinal disease analysis [12].

Until recently, improved U-Net++ architecture was applied to automatically segment macular edema regions in retinal Optical Coherence Tomography (OCT) images. The improved U-Net++ integrated with an enhanced ResNet backbone in this study achieved superior segmentation accuracy and generalization on the Duke University Diabetic Macular Edema (DME) dataset, outperforming earlier models like Fully Convolutional Network (FCN) and Retinal Layer segmentation network (ReLayNet) in Dice, Intersection over Union (IoU), and AUC metrics [13]. Prior applications of U-Net++ have proven its effectiveness in medical image segmentation by enhancing feature fusion and segmentation across diverse imaging tasks.

The previous reviews presented a detailed comparison of retinal vessel segmentation performance across widely used benchmark datasets. Several public datasets were evaluated based on results reported by multiple studies. Among these, the DRIVE dataset achieved sensitivity between 0.72 and 0.98, specificity from 0.95 to 0.99, accuracy ranging between 0.94 and 0.98, Area Under the Receiver Operating Characteristic Curve (AUC-ROC) values from 0.94 to 0.99, and Harmonic Mean of Precision and Recall (F1-score) between 0.76 and 0.86. The STARE dataset reported sensitivity between 0.65 and 0.99, specificity of 0.96 to 0.99, accuracy of 0.95 to 0.99, AUC-ROC between 0.98 and 0.99, and F1-scores around 0.77 to 0.86. Similarly, the CHASE-DB1 dataset yielded sensitivity ranging from 0.76 to 0.98, specificity between 0.96 and 0.99, accuracy between 0.67 and 0.98, AUC-ROC from 0.96 to 0.99, and F1-scores between 0.79 and 0.82. Meanwhile, the HRF dataset demonstrated sensitivity above 0.75, specificity ranging from 0.97 to 0.98, accuracy above 0.66, AUC-ROC values above 0.95 and F1-scores between 0.78 and 0.80 [14].

Moreover, another review discussed about distribution of diabetic retinopathy (DR) cases and severity levels varies widely across public fundus image datasets, influencing their suitability for different diagnostic and research purposes. Datasets such as DRIVE, Diabetic Retinopathy, Hypertension, Age-related Macular Degeneration (DR HAGIS), Diabetic Retinopathy image DataBase (DriDB), CHASE DB1, Annotated Germs for Automated Recognition (AGAR300), Diabetic Retinopathy DataBase 0 (DiaRetDB0), and Diabetic Retinopathy DataBase 1 (DiaRetDB1) consist entirely of patients with diabetic retinopathy (100% affected), making them highly suitable for lesion segmentation, progression modeling, and feature detection tasks. In contrast, Brazilian Multilabel Ophthalmological Dataset (BRSET) includes approximately 95% of patients with no apparent retinopathy, and Joint Shantou International Eye Centre (JSIEC) has about 85% in the same category, which supports the development of screening models for early detection. The Asia Pacific Tele-Ophthalmology Society (APTOS) dataset contains approximately 48% no apparent retinopathy cases, 28% with moderate non-proliferative DR and smaller proportions (about 5–10%) for mild (mild non-proliferative DR, typically microaneurysms only), severe (severe non-proliferative DR, characterized by extensive retinal abnormalities), and proliferative DR. The Dataset for Diabetic Retinopathy (DDR) dataset includes around 50% (no apparent retinopathy), 35% moderate DR and lower proportions across other levels. Eye Picture Archive Communication System (EyePACS) dataset is more imbalanced, with ~75% no apparent DR, ~10% moderate DR, and 5–7% spread among mild, severe, and proliferative stages. Indian Diabetic Retinopathy Image Dataset (IDRiD) presents a diverse distribution, with roughly 30% each in (no apparent retinopathy) and (moderate non-proliferative DR, involving multiple microaneurysms and possible hemorrhages), and 10–20% [15].

Similarly, JSIEC provides a relatively uniform distribution, with all five levels represented by 15–25% each. Finally, on Messidor, results shows that approximately 45% of patients exhibit no retinopathy, while the remaining three levels (based on microaneurysms, hemorrhages, and neovascularization) are distributed fairly evenly at 15–20% each. These variations in disease stage prevalence underscore the importance of selecting datasets aligned with the specific clinical or computational task [15]. The lack of evaluative studies recommending the most suitable public datasets for retinal image segmentation underscores the need for further research to clarify which datasets are most appropriate for this specific segmentation task.

III. Methodology

A. Datasets and Preprocessing steps

This study evaluated the segmentation performance of the U-Net++ model using eleven publicly available retinal vessel segmentation datasets commonly used between 2020 and 2025.The datasets includes DRIVE [16], STARE [17], CHASE DataBase 1 (CHASE-DB1) [18], HRF [19], REtinal vascular Tree Analysis (RETA) Benchmark [20], Retinal Images vessel Tree Extraction (RITE) [21], Fundus Image Vessel Segmentation (FIVES) [22], Artery/Vein Digital Retinal Images for Vessel Extraction (AV-DRIVE) [23], LES-AV from the LES-AV database [24], Optical coherence tomography angiography (OCTA) dubbed OCTA-500 [25], and IOSTAR vessel segmentation dataset [26]. For the AV-DRIVE dataset, which provides separate artery and vein annotations, the two classes were merged into a single binary vessel mask to maintain consistency with other datasets. Ambiguous or uncertain pixels at artery–vein crossings were treated as background during both training and evaluation to avoid label conflicts.

All input images were resized to 512 × 512 pixels and normalized to a standard intensity range. To enhance local contrast and vessel visibility, Contrast-Limited Adaptive Histogram Equalization (CLAHE) was applied uniformly across datasets. Consistent preprocessing ensured fair comparison across different imaging modalities and resolutions.

B. UNet++ architecture and justification

The U-Net++ model was adopted in this study because previous research has demonstrated that it consistently outperforms the original U-shaped convolutional network (U-Net) in segmentation accuracy, sensitivity, and Dice coefficient. Unlike the plain skip connections in U-Net, U-Net++ employs nested dense skip pathways that bridge the semantic gap between encoder and decoder feature maps, allowing the model to learn fine vessel details more effectively [27]. For implementation, the model was trained using the Adam optimizer with a learning rate of 1×10^{-4} and binary cross-entropy loss. A cyclical learning rate (CLR) strategy and early stopping (patience = 10) were employed to enhance stability and prevent overfitting. Each model was trained for 60 epochs with a batch size of 6 on an NVIDIA GeForce RTX 4070 GPU.

C. Metrics used for evaluation, Training set and Cross-validation Strategies

To quantitatively assess segmentation performance, two widely adopted evaluation metrics, namely accuracy and Dice score, were employed. Accuracy (Eq. 1) quantifies the proportion of correctly classified pixels, encompassing both vessel and background regions, thereby providing an overall estimate of model correctness [28], [29], [30]. Dice score (Eq. 2) measures the spatial overlap between the predicted segmentation and the ground-truth mask, highlighting the model's ability to accurately detect vessel pixels relative to false positives and false negatives [28], [29], [30]. In these equations, TP denotes true positives (correctly identified vessel pixels), TN represents true negatives (correctly identified background pixels), FP refers to false positives (background pixels misclassified as vessels), and FN indicates false negatives (missed vessel pixels). All metrics were computed on a per-image basis and subsequently averaged across the five cross-validation folds. This procedure ensured that the reported mean ± standard deviation values reliably reflected performance variability and provided complementary insights into both overall accuracy and detailed vessel delineation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Dice\ Score = \frac{2TP}{2TP + FP + FN} \tag{2}$$

A five-fold cross-validation strategy was employed for each dataset to rigorously evaluate segmentation performance and assess the model's generalization capability under identical experimental conditions. During each iteration, the model was trained on four folds (combined training data) and validated on the remaining fold, ensuring that every sample contributed to both training and validation across runs. Model performance was subsequently assessed using the Dice coefficient, accuracy, and loss metrics. After completing all folds, the performance measures were averaged to yield the final mean ± standard deviation values, thereby providing a robust estimation of the model's stability and generalization. Collectively, these aggregated outcomes formed the basis for the reported metrics and qualitative visualizations, thereby ensuring fair, consistent, and reproducible comparisons across all evaluated datasets.

IV. Results and Discussion

Table I summarizes the segmentation performance evaluation of the U-Net++ model across eleven publicly available retinal vessel segmentation datasets. Among them, eight datasets were fully evaluated using five-fold cross- validation, while FIVES, OCTA-500, and RITE are included for reference but were not completed under identical experimental conditions due to computational constraints. The reported values represent the mean validation Dice coefficient and accuracy (± standard deviation) across folds, reflecting the model's generalization stability and reproducibility.

Unlike prior studies that report performance on individual datasets in isolation, this work systematically quantifies dataset-dependent variability, offering practical guidance that can directly influence experimental design and fair comparison in future retinal segmentation research.

Among the eight fully evaluated datasets, RETA Benchmark (2022) outperformed all others, achieving the highest validation Dice (0.80 ± 0.015) and Accuracy (0.97 ± 0.016). CHASE-DB1 (2012), AV-DRIVE (2013), STARE (1998-2000), and HRF (2013) also exceeded Dice 0.70 and Accuracy 0.95 with minimal variance, reaffirming their reliability as benchmarks for retinal vessel segmentation. These findings underscore the progressive improvements in dataset quality and annotation precision achieved over the past decades. In contrast, the DRIVE (2004) dataset, although historically the most widely used, did not perform well under the same configuration, achieving a low validation Dice of 0.41 ± 0.05 and a moderate accuracy of 0.88 ± 0.002. This poor Dice score is likely due to the small sample size (40 images) and strong background–vessel class imbalance, which limited the model's ability to capture fine vascular details.

Specifically, the RETA Benchmark (2022) demonstrated superior image quality, precise annotations, and strong generalization, achieving the top performance among all datasets with a validation Dice of 0.80 ± 0.015 and Accuracy of 0.97 ± 0.016. CHASE-DB1 (2012) and AV-DRIVE (2013) followed closely (Dice = 0.74 ± 0.005 and 0.73 ± 0.011, Accuracy = 0.97 for both), highlighting their robustness even with smaller datasets. The STARE (1998-2000) dataset performed consistently (Dice = 0.72 ± 0.002, Accuracy = 0.96 ± 0.002), indicating that reliable annotations can compensate for limited diversity.

Meanwhile, HRF, IOSTAR, and LES-AV datasets showed lower Dice scores ranging between 0.64 and 0.71, mainly due to uneven illumination, inconsistent annotations, and small sample sizes. The FIVES (2021) and OCTA-500 (2024) datasets, though promising in resolution and scale, required significantly longer training times and could not be fully completed under identical configurations, revealing the computational limitations of higher-resolution volumetric data. The RITE (2013) dataset was excluded from direct comparison due to retrieval issues encountered during our study period, although it has been publicly available in prior research studies. Figs. 1–3 depict the training and validation performance curves of the U-Net++ model on the RETA Benchmark (2022) dataset, one of the best performers in this study.

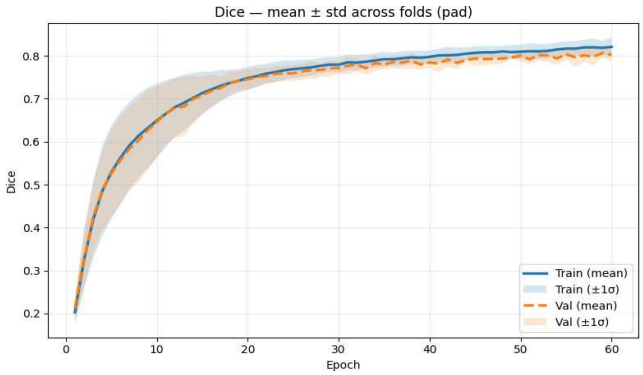

Fig. 1. Training and validation Dice coefficient (mean ± standard deviation) across five folds on the RETA Benchmark dataset using the U-Net++ model

TABLE I. PERFORMANCE COMPARISON OF THE U-NET++ MODEL ACROSS MULTIPLE RETINAL VESSEL SEGMENTATION DATASETS USING FIVE- FOLD CROSS-VALIDATION.

Dataset / Reference / Publication Year	**# Images / Modality / Resolution in Pixels**	**Segmentation results of the K-fold cross-validation of the U-Net++ model (k = 5).** *Average: Mean± Std*		**Comments**
		Dice	Accuracy	
DRIVE / [16] 2004	40 / Color fundus images / Fundus photography / 768×584	0.41 ± 0.05	0.88 ± 0.002	The model did not perform under the same configuration. High validation accuracy but low Dice due to strong background– vessel imbalance; model failed to capture fine vessel structures.
STARE / [17] / 1998-2000	20 / Fundus/ 700×605	0.72 ± 0.002	0.96 ± 0.002	Achieved stable accuracy with moderate Dice score.
FIVES / [22] 2021	800 / Fundus / 2048×2048 masks	-	-	Required long training time (~2 days per fold); training was not completed.
CHASE-DB1 / [18] / 2012	28 / Fundus / 1280×960	0.74 ± 0.005	0.97 ± 0.001	Performed well despite limited sample size.
HRF / [19] / 2013	45 / Color Fundus / 3504×2336	0.71 ± 0.010	0.96 ± 0.001	Limited adaptability due to small dataset size.
IOSTAR / [26] / 2016	30 / Color fundus photographs / 1024×1024 masks	0.66 ± 0.045	0.93 ± 0.005	Lower performance due to small and variable image quality.
RETA Benchmark [20] / 2022	81 / Color fundus photographs / 2896 × 1944	0.80 ± 0.015	0.97 ± 0.016	Recent dataset showing strong generalization.
RITE / [21] / 2013	40 / Fundus photography / 565 × 584	-	-	Excluded from comparison due to retrieval issues during the study period; no results reported.
AV-DRIVE / [23] / 2013	40 / Fundus / 565 × 584	0.73 ± 0.011	0.97 ± 0.001	Reliable artery–vein classification benchmark
LES-AV / [24] / 2018	22 / Fundus / 1620 × 1444 and 1958×2196	0.64 ± 0.072	0.97 ± 0.001	Performance limited by inconsistent annotations.
OCTA-500 / [25] / 2024	500 subjects / OCT & OCTA volumes / 400×400×640 and 304×304×640	-	-	Challenging to train using the same model and configuration.

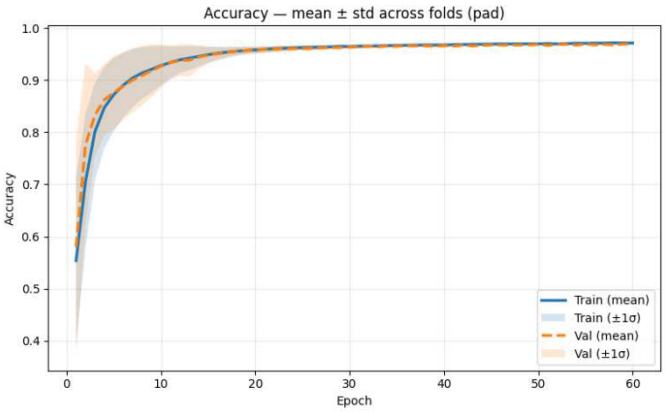

Fig. 2. Training and validation accuracy (mean ± standard deviation) across five folds on the RETA Benchmark dataset using the U-Net++ model

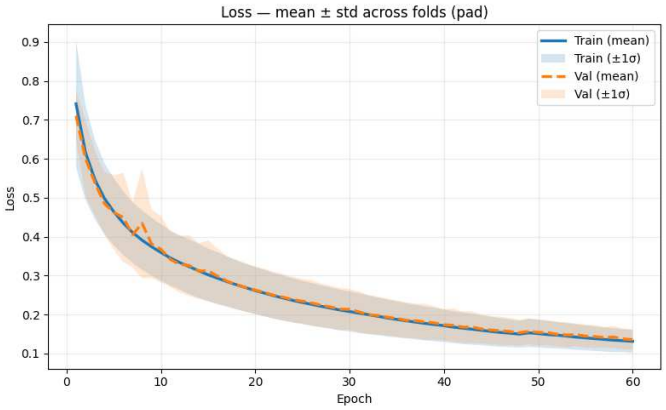

Fig. 3. Training and validation loss (mean ± standard deviation) across five folds on the RETA Benchmark dataset using the U-Net++ model

Overall, even datasets yielding lower quantitative Dice values demonstrated the model's capability to extract vessel morphology reliably. Fig. 4 presents qualitative segmentation results obtained by the U-Net++ model on two representative datasets—(A) LES-AV and (B) CHASE-DB1. In CHASE-DB1, the model accurately segmented fine vascular branches and bifurcations, producing predictions that closely match the ground truth (GT). In LES-AV, although the Dice score was lower, the model still captured the main vessel structures, with minor deviations observed in low-contrast regions due to illumination variability and inconsistent annotations. These results confirm that, despite differences in quantitative performance, U-Net++ effectively preserves vessel topology and demonstrates strong generalization across datasets of varying image quality and complexity.

It should be noted that the Dice coefficient was treated as the primary evaluation metric in this study due to its suitability for highly imbalanced segmentation tasks such as retinal vessel extraction, while accuracy was reported as a complementary measure for consistency. Metrics such as Precision, Sensitivity, IoU, and AUC-ROC were not emphasized to avoid redundancy and to ensure fair comparison across heterogeneous datasets.

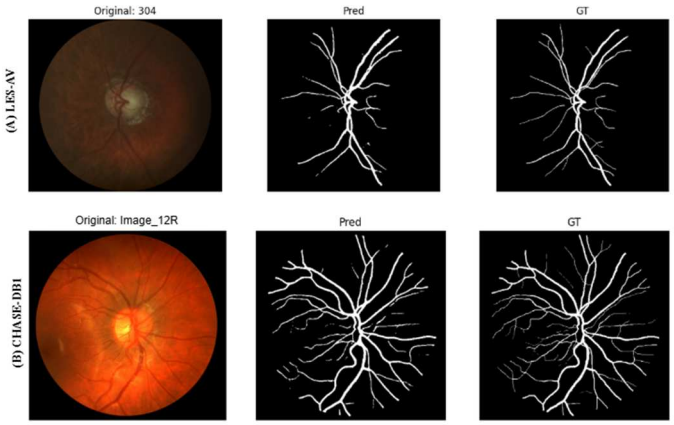

Fig. 4. Qualitative segmentation results obtained by the U-Net++ model. (A) LES-AV dataset and (B) CHASE-DB1 dataset. Columns show the original fundus image, predicted vessel segmentation (Pred), and ground truth (GT).

V. Conclusion and Fuuture Outlook

To the best of our knowledge, this study provides one of the most comprehensive benchmarking analyses of retinal vessel segmentation datasets using the U-Net++ model under identical experimental settings. The evaluation covered eleven publicly available datasets and revealed considerable variability in segmentation performance depending on dataset characteristics such as image resolution, annotation quality, and sample size.

Among all evaluated datasets, the RETA Benchmark (2022) demonstrated the best overall generalization, outperforming others with a Dice of 0.80 ± 0.015 and Accuracy of 0.97 ± 0.016. CHASE-DB1 (2012), AV-DRIVE (2013), HRF (2013), and STARE (1998-2000) also maintained strong results (Dice > 70%, Accuracy > 95%) with low variance. Collectively, these datasets—especially RETA—exhibit high annotation consistency and imaging diversity, making them the most suitable benchmarks for developing and evaluating deep learning models in retinal vessel segmentation.

Although the DRIVE dataset has been widely adopted in past studies, it underperformed in this work under the same configuration—likely due to background–vessel imbalance and its limited number of images. Nonetheless, even datasets with lower Dice scores, such as LES-AV and IOSTAR, were still capable of capturing the main vessel structures, confirming the robustness of U-Net++ in preserving vascular topology qualitatively.

Looking forward, future research should emphasize dataset harmonization, standardized annotation protocols, and multi-center data expansion to improve reproducibility and clinical translation.

In addition, evaluating these datasets under alternative architectures would help assess the extent to which the observed performance rankings are model-dependent, particularly given that U-Net++ has been shown in prior studies to outperform many conventional convolutional architectures, while emerging transformer-based and hybrid models may exhibit different generalization behaviors. Incorporating transformer-based and hybrid attention mechanisms may further enhance sensitivity to fine vessel structures. Moreover, developing large-scale, balanced

datasets that integrate diverse imaging modalities—such as fundus photography, OCT, and OCTA—will be crucial for advancing accurate, generalizable, and explainable AI models for retinal disease diagnosis and screening.

REFERENCES

[1] A. Khandouzi, A. Ariafar, Z. Mashayekhpour, M. Pazira, and Y. Baleghi, "Retinal vessel segmentation: A review of classic and deep methods," Ann. Biomed. Eng., vol. 50, no. 9, pp. 1292–1314, 2022.

[2] T. M. Khan, S. S. Naqvi, A. Robles-Kelly, and I. Razzak, "Retinal vessel segmentation via a multi-resolution contextual network and adversarial learning," Neural Netw., vol. 165, no. 1, pp. 310–320, 2023.

[3] R. Longjohn, M. Kelly, S. Singh, and P. Smyth, "Benchmark data repositories for better benchmarking," Adv. Neural Inf. Process. Syst., vol. 37, pp. 86435–86457, 2024.

[4] J. Li, R. Li, R. Han, and S. Wang, "Self-relabeling for noise-tolerant retina vessel segmentation through label reliability estimation," BMC Med. Imaging, vol. 22, no. 1, Art. no. 198, Dec. 2022.

[5] C. Li, X. Chen, Z. Chen, S. Zhang, and Y. Xia, "Attention UNet++: A nested attention-aware U-Net for liver CT image segmentation," in Proc. IEEE Int. Conf. Image Process. (ICIP), Oct. 2020, pp. 345–349.

[6] Y. Chen, B. Ma, and Y. Xia, "α-UNet++: A data-driven neural network architecture for medical image segmentation," in Proc. MICCAI Workshop Domain Adaptation and Representation Transfer (DART), ser. Lecture Notes Comput. Sci., vol. 12444, pp. 3–12, 2020.

[7] S. Buia, D. Popescu, and L. Ichim, "Blood vessel segmentation in retinal images using neural networks," in Proc. 12th Int. Symp. Adv. Topics Electr. Eng. (ATEE), Bucharest, Romania, Mar. 2021.

[8] H. Mittal, A. C. Pandey, M. Saraswat, S. Kumar, R. Pal, and G. Modwel, "A comprehensive survey of image segmentation: Clustering methods, performance parameters, and benchmark datasets," Multimedia Tools Appl., vol. 81, no. 24, pp. 35001–35026, Oct. 2022.

[9] S. Subramani, A. Deb, and R. R. J. Rachel, "Comparative study of deep learning approaches for retinal blood vessel segmentation," in Proc. Int. Conf. Multi-Agent Syst. Collaborative Intell. (ICMSCI), 2025, pp. 1434–1442.

[10] T. Bahr, T. A. Vu, J. J. Tuttle, and R. Iezzi, "Deep learning and machine learning algorithms for retinal image analysis in neurodegenerative disease: Systematic review of datasets and models," Transl. Vis. Sci. Technol., vol. 13, no. 2, Art. no. 16, 2024.

[11] O. O. Sule, "A survey of deep learning for retinal blood vessel segmentation methods: Taxonomy, trends, challenges and future directions," IEEE Access, vol. 10, pp. 38202–38236, 2022.

[12] M. S. Gargari, M. H. Seyedi, and M. Alilou, "Segmentation of retinal blood vessels using U-Net++ architecture and disease prediction," Electronics, vol. 11, no. 21, Art. no. 3512, Nov. 2022.

[13] Z. Gao, X. Wang, and Y. Li, "Automatic segmentation of macular edema in retinal OCT images using improved U-Net," Appl. Sci., vol. 10, no. 16, Art. no. XXXX, Aug. 2020.

[14] C. Chen, J. H. Chuah, R. Ali, and Y. Wang, "Retinal vessel segmentation using deep learning: A review," IEEE Access, vol. 9, pp. 111985–112004, 2021.

[15] V. Conquer, T. Lambolais, G. Andrade-Miranda, and B. Magnier, "Comprehensive review of open-source fundus image databases for diabetic retinopathy diagnosis," Sensors, vol. 25, no. XX, Art. no. XXXX, 2025.

[16] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," IEEE Trans. Med. Imaging, vol. 23, no. 4, pp. 501–509, Apr. 2004.

[17] A. D. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," IEEE Trans. Med. Imaging, vol. 19, no. 3, pp. 203–210, Mar. 2000.

[18] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman, "An ensemble classification-based approach applied to retinal blood vessel segmentation," IEEE Trans. Biomed. Eng., vol. 59, no. 9, pp. 2538–2548, Sep. 2012.

[19] A. Budai, R. Bock, A. Maier, J. Hornegger, and G. Michelson, "Robust vessel segmentation in fundus images," Int. J. Biomed. Imaging, vol. 2013, Art. no. 154860, 2013.

[20] X. Lyu, L. Cheng, and S. Zhang, "The RETA benchmark for retinal vascular tree analysis," Sci. Data, vol. 9, no. 1, Art. no. 775, Dec. 2022.

[21] Q. Hu, M. D. Abramoff, and M. K. Garvin, "Automated separation of binary overlapping trees in low-contrast color retinal images," Pattern Recognit., vol. 46, no. 8, pp. 2117–2128, Aug. 2013.

[22] K. Jin et al., "FIVES: A fundus image dataset for artificial-intelligence-based vessel segmentation," Sci. Data, vol. 9, no. 1, Art. no. 435, Dec. 2022.

[23] T. A. Qureshi, M. Habib, A. Hunter, and B. Al-Diri, "A manually-labeled, artery/vein classified benchmark for the DRIVE dataset," in Proc. 26th IEEE Int. Symp. Computer-Based Medical Systems (CBMS), Jun. 2013, pp. 485–488.

[24] J. I. Orlando, J. B. Breda, K. Van Keer, M. B. Blaschko, P. J. Blanco, and C. A. Bulant, "Towards a glaucoma risk index based on simulated hemodynamics from fundus images," in Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI), ser. Lecture Notes Comput. Sci., vol. 11071, pp. 65–73, 2018.

[25] M. Li, K. Huang, Q. Xu, J. Yang, Y. Zhang, Z. Ji, K. Xie, S. Yuan, Q. Liu, and Q. Chen, "OCTA-500: A retinal dataset for optical coherence tomography angiography study," Med. Image Anal., vol. 93, Art. no. 103092, 2024.

[26] Idiap Research Institute, "IOSTAR retinal vessel segmentation dataset," Martigny, Switzerland, n.d. [Online]. Available: https://www.idiap.ch/software/bob/docs/bob/bob.db.iostar/stable/ (accessed Jan. 2026).

[27] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in Proc. Int. Workshop Deep Learn. Med. Image Anal., ser. Lecture Notes Comput. Sci., vol. 11045, pp. 3–11, 2018.

[28] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Stanford, CA, USA, Oct. 2016, pp. 565–571.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI), ser. Lecture Notes Comput. Sci., vol. 9351, pp. 234–241, 2015.

[30] T. Schlosser, M. Friedrich, T. Meyer, and D. Kowerko, "A consolidated overview of evaluation and performance metrics for machine learning and computer vision," Junior Professorship of Media Computing, Chemnitz Univ. Technol., Chemnitz, Germany, Tech. Rep. 9107, 2024.