Faculty of Computing
Sabaragamuwa University of Sri Lanka

ICARC

ICARC 2026

Sabaragamuwa University of Sri Lanka

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

Robert Ngabo Mugisha[1], Geoffrey Munyaneza[2], Fidele Nsanzumukunzi[3], Mediatrice Dusenge[4], Josue Uzigusenga[5], Theophilla Igihozo[6], Fabrice Mpozenzi[7], Emmanuella Nuwayo[5], Prince Shema Musonerwa[1], Benny Uhoranishema[1] and Jean De Dieu Niyonteze[8]

[1]College of Engineering, Carnegie Mellon University Africa, Kigali, Rwanda
[2]Kenan Flagler Business School, University of North Carolina, North Carolina, United States
[3]The Roux Institute, Northeastern University, Portland ME, United States
[4]African Centre of Excellence in Data Science, University of Rwanda, Kigali, Rwanda
[5]College of Sciences and Technology, University of Rwanda, Kigali, Rwanda
[6]College of Medicine and Health Sciences, University of Rwanda, Kigali, Rwanda
[7]Department of Computer Science, Abilene Christian University, Texas, Abilene, United States
[8]Goizueta Business School, Emory University, Atlanta, GA 30322, United States

Presenting Author : Jean De Dieu Niyonteze

Introduction

- Retinal vessel segmentation is essential in medical image analysis:
- It enables accurate identification of blood vessels.
- This is critical for detecting and monitoring eye and systemic diseases.
- Transparent datasets benchmark is critical for reproducible and reliable model comparisons.
- Detailed metadata strengthens result validation and supports replication, modification, and debugging.
- Model performance can be significantly reduced when trained on noisy ground-truth data

Introduction cont’

- This study conducts a comprehensive evaluation of the U-Net++ model for retinal vessel segmentation.
- Eleven widely recognized, publicly available datasets are used, including DRIVE, STARE, CHASE-DB1, HRF, the RETA Benchmark, RITE, FIVES, AV-DRIVE, LES-AV, OCTA-500, and IOSTAR
- The selected datasets have been extensively applied in research from 2020 to 2025.
- The evaluation accounts for diverse imaging conditions and vascular characteristics.
- Datasets include DRIVE, STARE, CHASE-DB1, HRF, RETA Benchmark, RITE, FIVES, AV-DRIVE, LES-AV, OCTA-500, and IOSTAR

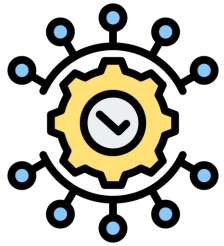
Limitation of the Existing literature

- Datasets often comprise diverse vary in region of interest, image resolution, and applicability, making benchmark inconsistent.
- Variations in illumination, intra-class differences, and complex backgrounds challenge segmentation methods.
- Lack of image quality standardization affects the reliability of benchmark evaluations
- Dataset selection is often not well aligned with specific clinical or computational tasks
- There is a lack of evaluative studies that recommend the most suitable public datasets for retinal image segmentation

Our study contribution

- Provide guidance for determining and recommending the most appropriate datasets for specific segmentation tasks.
- Enables reproducibility of segmentation results through transparent benchmarking practices
- Facilitates replication of experiments and further model improvement or modification

Methodology



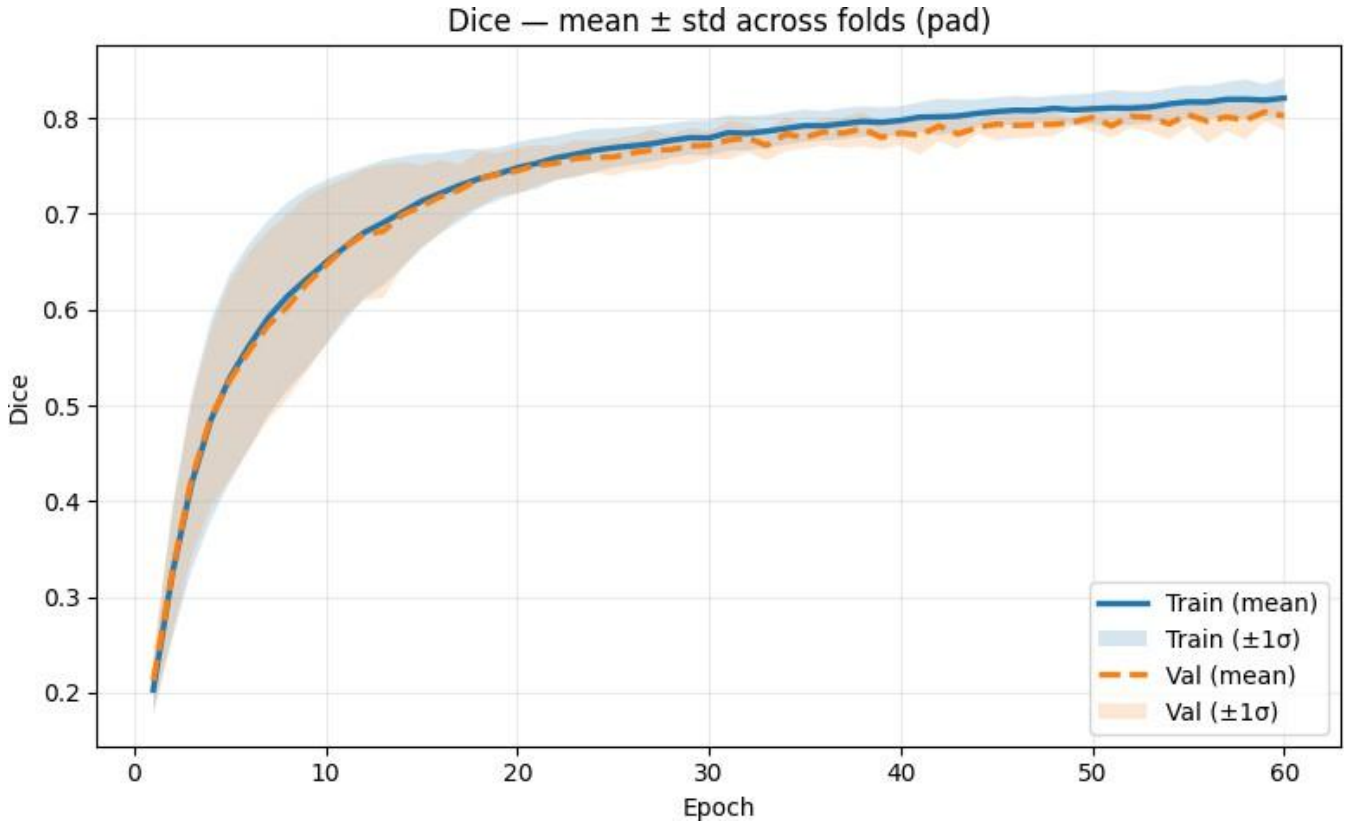
- A five-fold cross-validation strategy was employed for each dataset to rigorously:
  - Evaluate segmentation performance and
  - Assess the model's generalization capability under identical experimental conditions.
- All images were resized to 512 × 512 pixels.
- Image intensities were normalized to a standard range.
- CLAHE was applied to all datasets.
- Model trained with Adam optimizer (LR = 1×10^{-4}) and binary cross-entropy loss.
- Cyclical learning rate and early stopping (patience = 10).
- Training for 60 epochs with batch size 6 on an NVIDIA RTX 4070 GPU.

RESULTS AND DISCUSSION

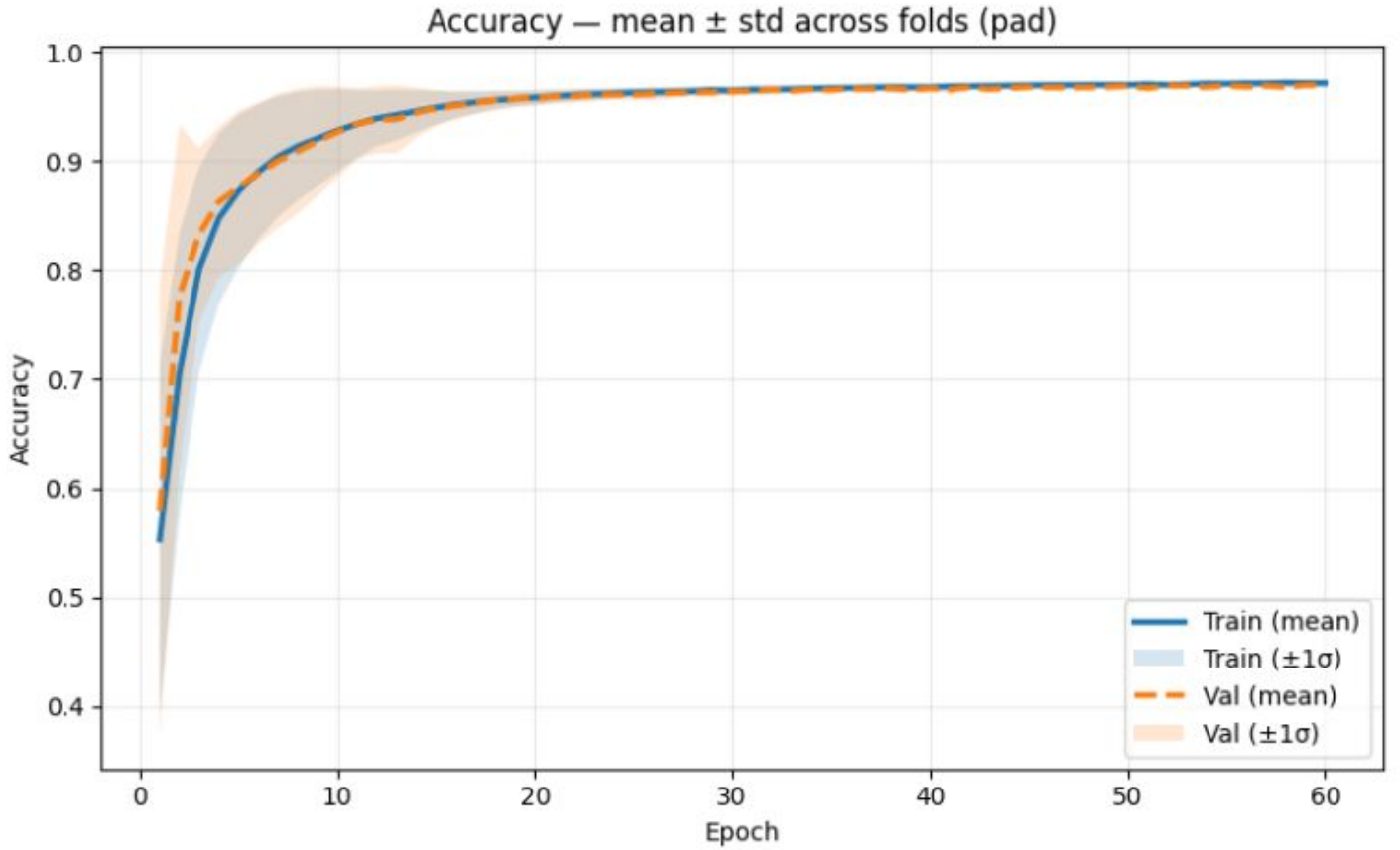
Results of the U-Net++ model across multiple retinal vessel segmentation datasets using five- fold cross-validation

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 / 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 / 1998-2000	20 / Fundus/ 700×605	0.72 ± 0.002	0.96 ± 0.002	Achieved stable accuracy with moderate Dice score.
FIVES / 2021	800 / Fundus / 2048×2048 masks	-	-	Required long training time (~2 days per fold); training was not completed.
CHASE-DB1 / 2012	28 / Fundus / 1280×960	0.74 ± 0.005	0.97 ± 0.001	Performed well despite limited sample size.
HRF / 2013	45 / Color Fundus / 3504×2336	0.71 ± 0.010	0.96 ± 0.001	Limited adaptability due to small dataset size.
IOSTAR / 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 2022	81 / Color fundus photographs / 2896 × 1944	0.80 ± 0.015	0.97 ± 0.016	Recent dataset showing strong generalization.
RITE / 2013	40 / Fundus photography / 565 × 584	-	-	Excluded from comparison due to retrieval issues during the study period; no results reported.
AV-DRIVE / 2013	40 / Fundus / 565 × 584	0.73 ± 0.011	0.97 ± 0.001	Reliable artery–vein classification benchmark
LES-AV / 2018	22 / Fundus / 1620 × 1444 and 1958×2196	0.64 ± 0.072	0.97 ± 0.001	Performance limited by inconsistent annotations.
OCTA-500 /2024	500 subjects / OCT & OCTA volumes / 400×400×640 and 304×304×640	-	-	Challenging to train using the same model and configuration.

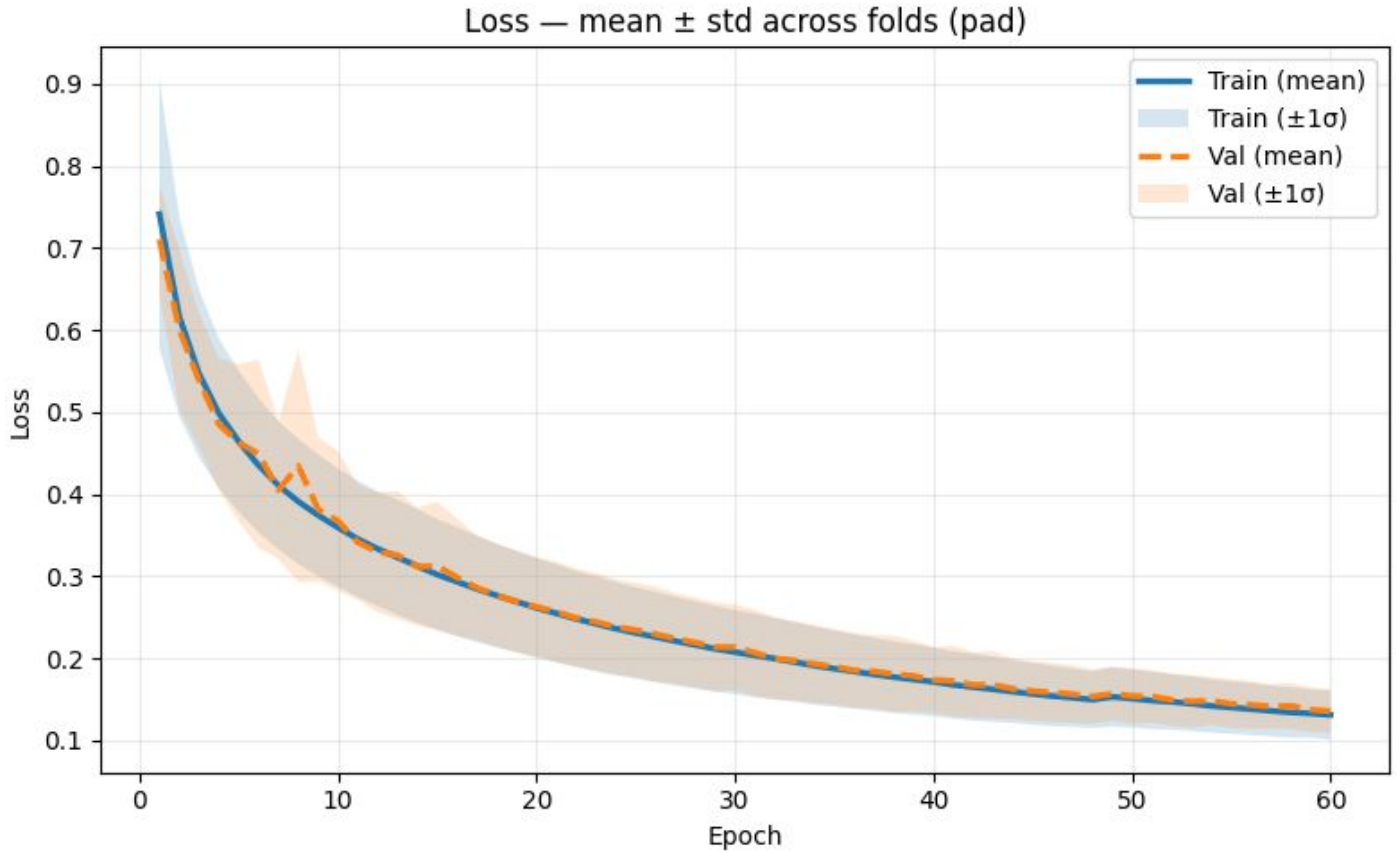
Training and validation Dice coefficient (mean ± standard deviation) across five folds on the RETA Benchmark dataset using the U-Net++ model



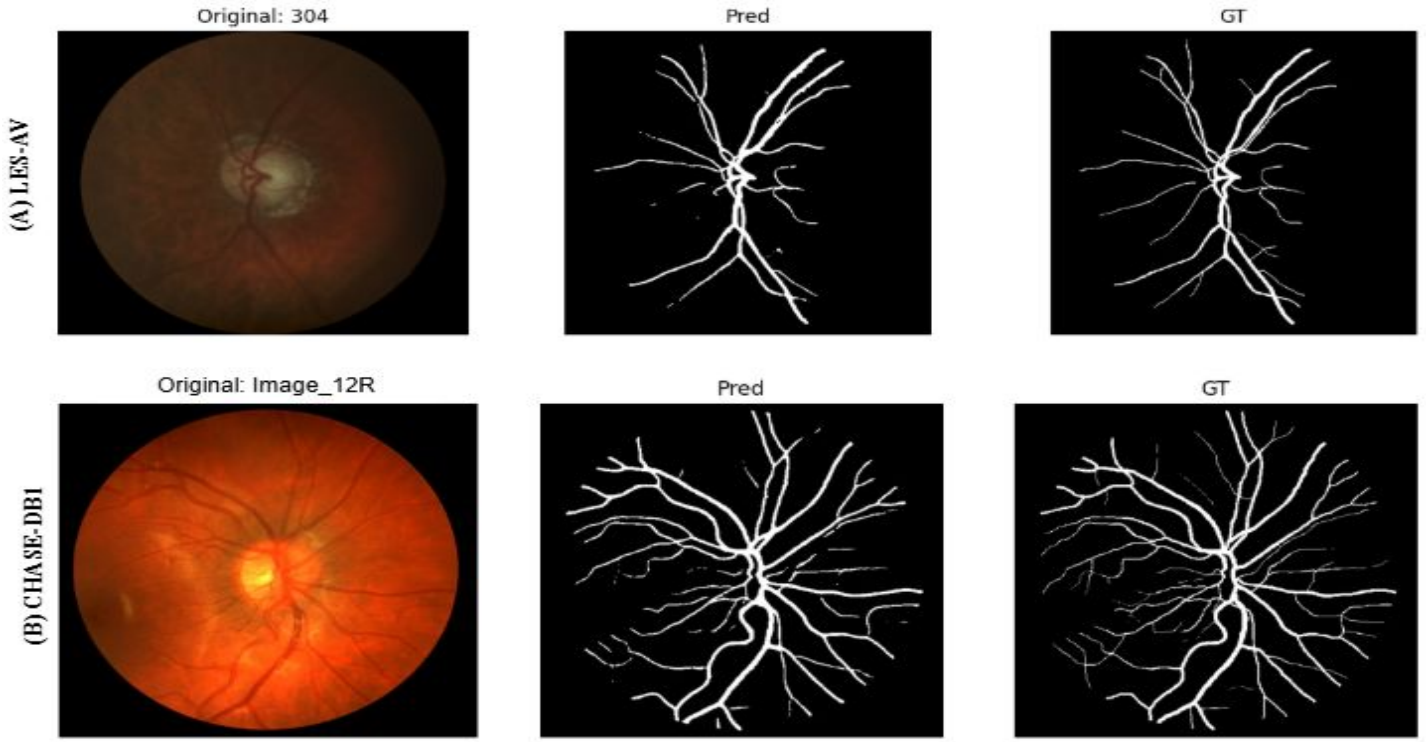
Training and validation accuracy (mean ± standard deviation) across five folds on the RETA Benchmark dataset using the U-Net++ model



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



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).



The best-performing dataset 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

outperformed all other datasets. CHASE-DB1, AV-DRIVE, HRF, and STARE also achieved Dice ≥ 70% and Accuracy > 95%

Conclusion

- The study provides a comprehensive benchmark of U-Net++ across eleven public retinal vessel datasets under identical conditions.
- Significant performance differences were observed due to dataset's characteristics (resolution, annotation quality, sample size).
- The RETA Benchmark (2022) showed the best overall generalization performance.
- Despite its popularity, DRIVE under-performed, likely due to background–vessel imbalance and limited data.
- Datasets with lower Dice scores (e.g., LES-AV and IOSTAR) still captured major vessel structures, highlighting the robustness of U-Net++ in preserving vascular topology

Future works

- Emphasize on datasets harmonization, standardized annotation protocols, and multi- center data expansion to improve reproducibility and clinical translation.
- Evaluate existing datasets using alternative model architectures to determine if performance rankings are model-dependent.
- Investigate the use of transformer-based and hybrid attention mechanisms to improve sensitivity to fine vessel structures.
- Develop large-scale, balanced datasets that include diverse imaging modalities (fundus photography, OCT, OCTA).
- Focus on creating datasets that support accurate, generalization, and explainable AI models for retinal disease diagnosis and screening


Q & A

👋


THANK
YOU!