A Teleophthalmology Screening Platform for Diabetic Retinopathy with Lesion-Based Evidence and Ordinal Grading

Laura Bernardes[1], Artur Heckler[1], Alejandro Pereira[2], Marcelo Dias[2], Marilton Aguiar[2], Daniel Welfer[3], Carlos Santos[1]

[1]Federal Institute Farroupilha (IFFar)
Alegrete – RS – Brazil

[2]Federal University of Pelotas (UFPel)
Pelotas – RS – Brazil

[3]Federal University of Santa Maria (UFSM)
Santa Maria – RS – Brazil

carlos.santos@iffarroupilha.edu.br

Abstract. *Diabetic retinopathy screening in public health settings demands scalable workflows, traceability, and clinically meaningful visual evidence. This paper presents an open teleophthalmology platform that couples multi-label lesion segmentation with ordinal DR grading (0–4) and exposes the inference trail through an interactive interface. Using the public IDRiD dataset, we evaluated segmentation performance across U-Net++, Attention U-Net, and Swin-Unet, and grading performance with EfficientNetV2-S and ConvNeXt-Tiny, including QWK, Macro-F1, Weighted-F1, AUC (OvR), per-class reports, and confusion matrices. Beyond model metrics, the contribution is a healthcare-oriented system design that operationalizes visual evidence, class-wise threshold calibration, and experiment governance, enabling remote triage scenarios and supporting future external and clinical validation in partner ophthalmology clinics.*

1. Introduction

Diabetic retinopathy (DR) represents one of the main preventable causes of vision loss in populations with diabetes. For this reason, screening programs play a strategic role in early detection and timely referral. In the context of the Brazilian Unified Health System (SUS), care demand, the geographic distribution of patients, and the limited availability of specialists make teleophthalmology models attractive, in which image acquisition takes place at remote units and centralized reading supports clinical workflows. In this setting, a robust computational solution must not only classify images but also provide traceable visual evidence, allow task-specific calibration, and record artifacts for auditing and validation.

AI systems for DR already achieve high performance on public benchmarks and in clinical studies; however, moving from the laboratory to real-world screening requires integrating inference, visual evidence, operational feasibility, and result governance. In addition, the ordinal nature of DR grading and the strong class imbalance in datasets influence both evaluation and system calibration.

In this work, we propose a teleophthalmology screening platform that integrates two complementary modules: (i) multi-label segmentation of retinal lesions, namely microaneurysms (MA), hemorrhages (HE), hard exudates (EX), and soft exudates (SE), to generate local evidence and quantitative measures; and (ii) ordinal grading to support case prioritization. The platform provides an interactive interface that displays masks, overlays, heatmaps, and class probabilities, along with explicit lesion-specific threshold controls. Rather than proposing a new neural architecture, the main contribution lies in integrating established models into a reproducible and auditable healthcare-oriented platform, designed to support teleophthalmology workflows and future field validation.

As an additional contribution, we describe the teleophthalmology workflow, define traceability requirements, report model-level experimental results, and discuss the limitations that must be addressed before clinical use, including external validation, locked test-set evaluation, and cross-site robustness assessment.

2. Related Work

Related work on DR in fundus images has evolved along two complementary directions: predictive models for grading and lesion detection, and software-mediated screening workflows that require traceability, system integration, and patient safety. Deep learning-based systems have shown high performance for identifying referable DR in clinical data and diverse cohorts [Gulshan et al. 2016, Ting et al. 2017]. Trials of autonomous systems in primary care also highlight the importance of operational fit, quality control, and referral criteria [Abràmoff et al. 2018]. More recent studies reinforce that teleophthalmology solutions must go beyond accuracy, incorporating usability, auditability, prospective validation, and robustness to real-world acquisition variability [Arenas-Cavalli et al. 2022, Nakayama et al. 2023, Chokshi et al. 2024, Farahat et al. 2024].

Public datasets play an essential role in reproducible research. IDRiD, released in the context of the ISBI 2018 challenge, provides both dense lesion annotations (MA, HE, EX, SE) and severity labels (0–4) [Porwal et al. 2020]. This combination enables pipelines that connect local evidence to global severity outcomes. However, public benchmarks do not replace external validation on local clinical data, particularly when the target scenario involves public health workflows and heterogeneous acquisition conditions.

For lesion segmentation, encoder–decoder architectures remain strong baselines. U-Net++ improves multiscale fusion through nested skip connections [Zhou et al. 2018], Attention U-Net emphasizes relevant regions through attention mechanisms [Oktay et al. 2018], and Swin-Unet incorporates hierarchical Transformer-based attention [Cao et al. 2021]. For grading, EfficientNetV2 and ConvNeXt are efficient and widely used backbones for transfer learning [Tan and Le 2021, Liu et al. 2022]. Nevertheless, DR grading is ordinal and imbalanced, which motivates metrics such as QWK, Macro-F1, Weighted-F1, AUC (OvR), confusion matrices, and per-class reports.

Although the literature offers strong models and promising clinical evidence, many approaches still treat grading and lesion analysis separately or do not specify how results and artifacts are governed during experimental and teleophthalmology workflows. Our contribution is therefore not a new neural architecture, but an integrated platform that combines ordinal grading, multi-label lesion segmentation, threshold calibration, visual evidence, and systematic artifact logging to support traceable teleophthalmology screen-

ing.

3. Materials and Methods

3.1. Datasets

We conducted experiments using the IDRiD dataset [Porwal et al. 2020], which contains fundus images and annotations for both lesion segmentation (MA, HE, EX, SE) and DR grading across five levels (0–4), corresponding to 0 (No DR), 1 (Mild), 2 (Moderate), 3 (Severe), and 4 (Proliferative DR). For grading, we applied an 80/20 stratified split to the IDRiD training set, yielding a validation set with $n = 83$ images and the following class distribution: 0 (27), 1 (4), 2 (27), 3 (15), and 4 (10). This distribution reveals class imbalance, with very low representation of class 1, which affects per-class metrics and motivates the use of class-balancing techniques.

The reported quantitative results are based on a development-stage validation protocol. Although this protocol supports model comparison during prototype construction, it does not replace locked test-set evaluation, cross-validation, or external validation on independent clinical data. These additional evaluations are planned as part of the next validation stage.

3.2. Computational Environment and Reproducibility

We implemented the pipeline in PyTorch and ran it on an NVIDIA Tesla T4 GPU in a Google Colab environment, with artifact persistence on Google Drive. The system records: (i) model checkpoints; (ii) CSV logs for training and validation; (iii) per-class metrics and reports; (iv) raw and normalized confusion matrices; and (v) prediction figures and dashboards. This design supports auditing, reproducibility, and the direct generation of tables and figures for scientific publications.

The same environment was sufficient for training, evaluation, and interactive demonstration of the prototype. However, systematic latency and throughput analysis was not part of the present study. Future evaluations will include inference time, memory requirements, and deployment scenarios involving cloud-based execution and lower-cost computational infrastructure.

3.3. Pipeline Overview

The platform implements an end-to-end pipeline in which two complementary modules share a consistent preprocessing stage, including resizing, normalization, and basic quality control. The first module performs multi-label segmentation, producing probability maps that are binarized using lesion-specific thresholds. This provides visual evidence, such as masks, overlays, and heatmaps, as well as objective quantification, such as relative lesion areas.

The second module performs ordinal grading (levels 0–4), providing the final classification and probability vectors for case prioritization and uncertainty analysis. An interactive interface (Section 5) operationalizes this integration, allowing users to select models and adjust thresholds. By combining local lesion-level evidence with global grading synthesis, the system enables a comprehensive screening and case review workflow. Figure 1 summarizes the proposed pipeline.

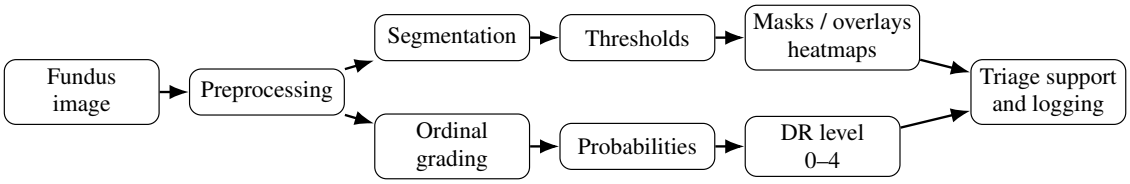


Figure 1. Overview of the proposed teleophthalmology screening pipeline, integrating lesion-level evidence, ordinal grading, threshold calibration, and artifact logging.

3.4. Lesion Segmentation

To handle high image resolution and the strong imbalance between lesions and background, especially in MA and SE, we adopt a patch-based training strategy using 384×384 patches, which we resize to 512×512 at the network input. To maximize the model's exposure to informative regions, we apply oversampling of positive patches (probability 0.9) defined by the union of available lesion masks. Images follow ImageNet-based normalization.

To mitigate false positives in highly reflective regions, we remove the optic disc area from the training targets via controlled dilation (radius 14). We evaluate U-Net++, Attention U-Net, and Swin-Unet, trained using BCEWithLogits and Soft Dice loss functions. To ensure fair comparability, segmentation training follows a single set of hyperparameters and preprocessing steps, applied identically to all evaluated models. We apply data augmentation with flips, rotations, and photometric perturbations. Additional settings appear in Table 1.

During post-processing, we binarize the per-lesion probability maps in the range $[0, 1]$ using class-specific thresholds: $MA = 0.07$, $HE = 0.12$, $EX = 0.18$, and $SE = 0.18$. These values are persisted in a JSON file and exposed via sliders for real-time adjustment, allowing calibration of the sensitivity-specificity trade-off without changing model weights.

Table 1. Training settings for lesion segmentation.

Setting	Value
Strategy	patch-based (with oversampling of positive patches)
Patch size	384×384 patches; 512×512 network input
Positive sampling	$\texttt{pos\_patch\_prob} = 0.90$
Evaluated models	U-Net++, Attention U-Net, Swin-Unet
Optimizer / LR	AdamW / 3×10^{-4}
Epochs / batch	18 / 2
Loss	BCEWithLogits + Soft Dice (multi-label)
Augmentations (train)	geometric + photometric perturbations
Acceleration	AMP when GPU available

We treat segmentation of the four lesion types (MA, HE, EX, and SE) as a multi-label problem, producing one binary mask per class after applying class-specific thresholds. We evaluate performance using the pixel-wise metrics $Dice_c$ and IoU_c for each lesion class c:

$$\mathrm{Dice}_c = \frac{2\,\mathrm{TP}_c}{2\,\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}, \qquad \mathrm{IoU}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}. \tag{1}$$

As aggregate measures, we report macro-averaged scores ($mDice$ and $mIoU$) to mitigate bias toward more frequent classes:

$$\mathrm{mDice} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{Dice}_c, \qquad \mathrm{mIoU} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{IoU}_c. \tag{2}$$

We also include qualitative analyses through overlays and heatmaps to inspect error patterns and support clinical interpretation.

3.5. DR Grading

For ordinal classification (levels 0–4), we adapt EfficientNetV2-S and ConvNeXt-Tiny by replacing their final classification layers with five-class outputs and fine-tuning the models initialized with pretrained weights. Both grading backbones were implemented using the `timm` library and initialized with pretrained weights: `tf_efficientnetv2_s.in21k_ft_in1k` for EfficientNetV2-S and `convnext_tiny.fb_in22k_ft_in1k` for ConvNeXt-Tiny.

We conduct the experiments using an 80/20 stratified split of IDRiD, yielding a validation set of 83 images. Images follow a preprocessing workflow consisting of resizing and padding to 512×512 to preserve global context, followed by cropping to 448×448.

To mitigate class imbalance, we combine class weights in the loss function (weighted CrossEntropy) with a WeightedRandomSampler in the DataLoader. We train the models using AdamW (learning rate of 2×10^{-4} and weight decay of 1×10^{-4}) for 10 epochs with a batch size of 8. During training, we apply moderate data augmentation, including flips and photometric perturbations via RandomBrightnessContrast and HueSaturationValue. Detailed configurations appear in Table 2.

Table 2. Training configuration for DR grading on IDRiD.

Setting	Value
Evaluated backbones	EfficientNetV2-S; ConvNeXt-Tiny
Pretrained weights	`timm`: IN-21k pretraining, IN-1k fine-tuning variants
Split	Stratified 80/20 (validation: $n = 83$)
Input / crop	448×448 (training: random; validation: center)
Optimizer / LR	AdamW / 2×10^{-4}
Weight decay	1×10^{-4}
Epochs / batch size	10 / 8
Normalization	ImageNet (mean/std)
Class balancing	Class weights (CE) + WeightedRandomSampler
Augmentations	Flips + moderate photometric perturbations

For the grading task, we report Macro-F1, Weighted-F1, $\mathrm{AUC}_{\mathrm{OVR}}$, and Quadratic Weighted Kappa (QWK), as well as raw and row-normalized confusion matrices and per-class reports. Macro-F1 emphasizes class-balanced behavior, Weighted-F1 summarizes performance according to class support, AUC (OvR) evaluates probabilistic separability, and QWK explicitly accounts for the ordinal nature of DR severity.

4. System Architecture and Teleophthalmology Workflow

The platform architecture was designed to support teleophthalmology workflows in which screening decisions must remain explainable, traceable, and operationally adjustable. Accordingly, beyond running inference, the system organizes evidence and artifacts to support auditing, clinical discussion, and field validation.

4.1. Healthcare-Oriented Requirements

In applied healthcare computing scenarios, system utility depends on requirements that go beyond accuracy: (i) traceability and auditability, including logging of weights, versions, metrics, confusion matrices, and per-run reports; (ii) transparency through visual evidence, including per-lesion masks, overlays, and heatmaps; (iii) operational calibration through adjustable class-specific thresholds; (iv) reproducibility, supported by standardized execution and artifact persistence; and (v) modularity, allowing model selection and replacement without interface changes.

4.2. Functional Layers and Result Governance

The architecture comprises four layers, with emphasis on governance of the inference and evaluation cycle:

1. **Data and integrity:** IDRiD indexing, consistency checks, stratified splits, and loading of masks and labels for lesion segmentation and DR grading.
2. **Models and inference:** independent execution of lesion segmentation (MA/HE/EX/SE) and ordinal grading (0–4), with versioned checkpoints and configurable loading.
3. **Evaluation and governance:** automatic and persistent generation of metrics and artifacts, including training history, per-class reports, confusion matrices, figures, and segmentation threshold files.
4. **Interface and interaction:** a Gradio-based application for model selection, end-to-end execution, and integrated visualization of masks, overlays, heatmaps, and grading probabilities.

The directory structure and deterministic logging reduce dependence on the notebook state and support auditing. This design aligns with the transition to field validation, where tracking which model and threshold produced each analyzed output becomes essential.

5. Interactive Interface and Teleophthalmology Workflow

The interface supports remote screening with interpretable evidence. The user uploads a fundus image, selects the segmentation and grading models, and runs end-to-end inference. The system returns: (i) an image overlay with lesion regions; (ii) scrollable

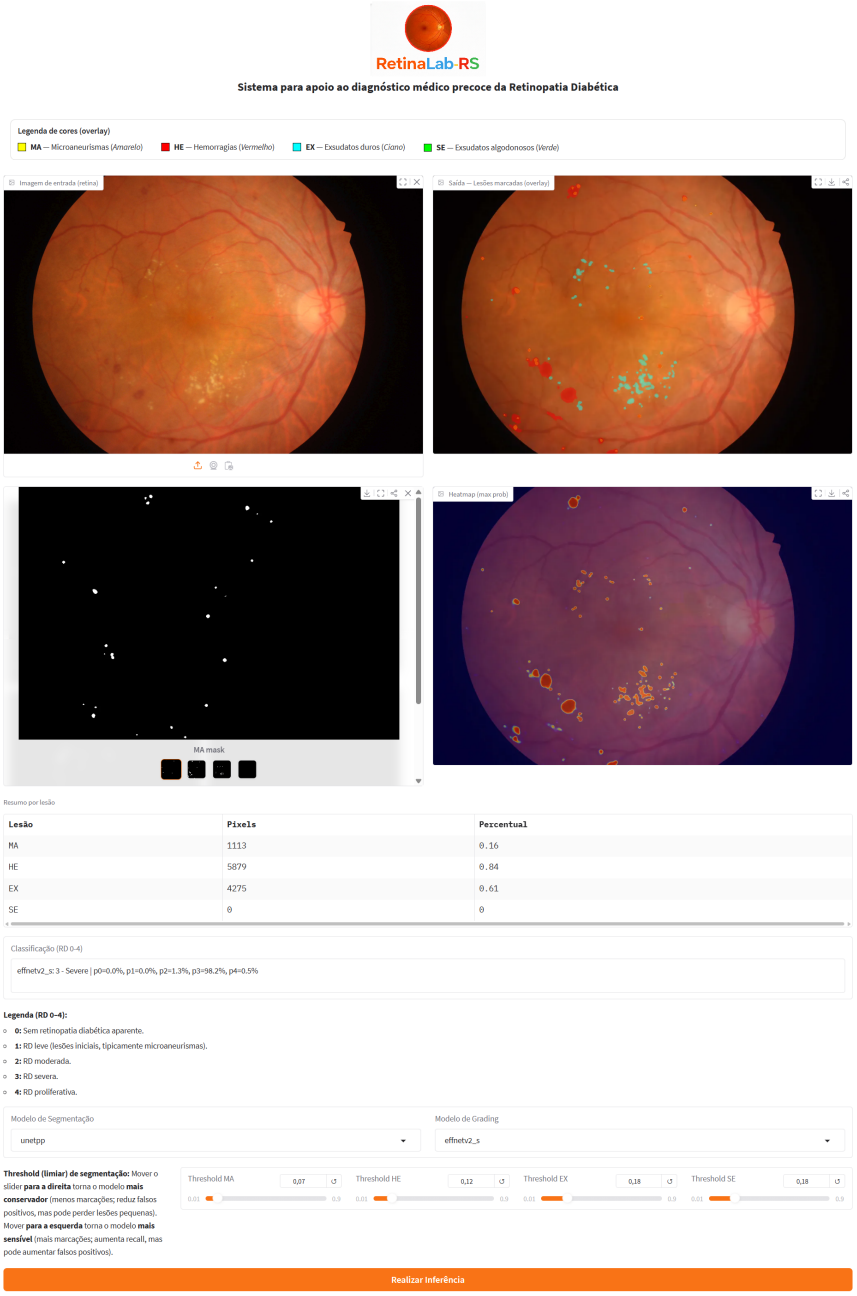


Figure 2. Functional prototype interface of the platform (Gradio), integrating lesion segmentation, class-specific threshold controls, and DR grading with probability outputs.

per-lesion masks; (iii) relative area estimates per class; (iv) the final grading class and per-level probabilities; and (v) lesion-specific threshold controls for real-time adjustment. Figure 2 illustrates the functional prototype.

Beyond individual inspection, the prototype displays the four binary masks corresponding to the lesion types (MA, HE, EX, and SE), generated after applying class-specific thresholds. Figure 3 illustrates this panel, which allows comparison of the spatial distribution of different lesions within the same image and verification of consistency with the reported grading outcome.

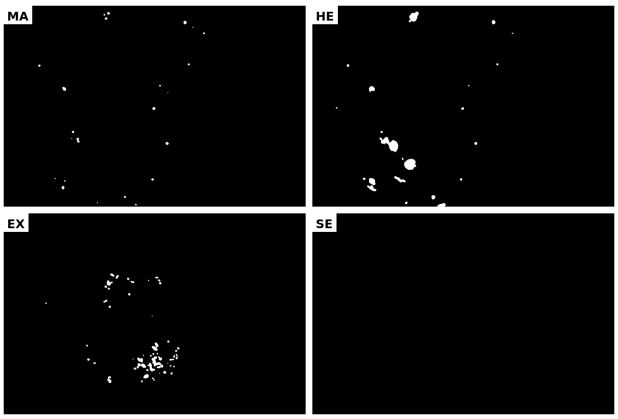


Figure 3. Binary masks generated by the prototype for the four lesion types (MA, HE, EX, and SE) after applying class-specific thresholds.

To support interpretation and screening, the system also returns an annotated image (overlay) and an evidence heatmap. The overlay highlights the segmented regions on the original image, whereas the heatmap summarizes the segmenter response on a continuous scale, emphasizing areas with higher lesion probability. Figure 4 presents these complementary outputs.

Figure 4. Example of the visual outputs produced by the prototype: (left) annotated image with lesion overlay; (right) evidence heatmap highlighting regions with higher segmenter response.

6. Experimental Results

6.1. Segmentation Results

Table 3 reports, for each architecture, the best validation performance selected according to the mean Dice score, averaged over the four lesion classes. U-Net++ achieved the best overall trade-off (Dice_m = 0.221), with more pronounced gains for HE and EX, which typically occupy larger regions and exhibit more stable contrast in fundus images.

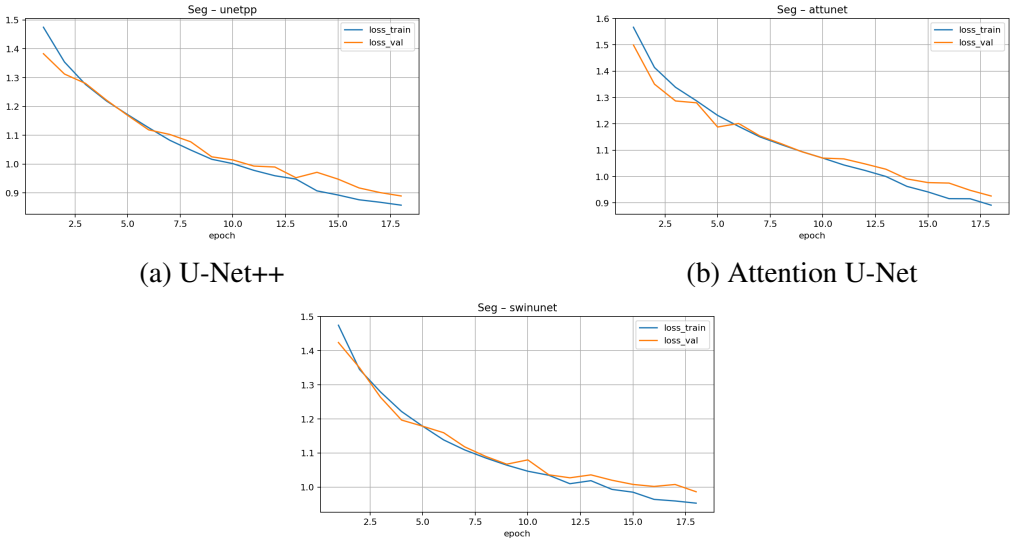
Table 3. Segmentation comparison on IDRiD (best epoch selected by mean Dice).

Model	Dice_m	Dice_{MA}	Dice_{HE}	Dice_{EX}	Dice_{SE}	mIoU
U-Net++	0.221	0.049	0.377	0.458	0.000	0.146
Att U-Net	0.187	0.014	0.329	0.403	0.000	0.123
Swin-Unet	0.080	0.000	0.094	0.188	0.037	0.046

MA and SE remain the most challenging classes. MA corresponds to a very small and thin lesion, making it sensitive to patch scale, noise, and illumination variation. SE

is rare in the dataset and can be confused with bright patterns or acquisition artifacts, which explains the near-zero Dice scores for U-Net++ and Attention U-Net. Even with positive-patch oversampling, low prevalence and subtle morphology can lead to conservative predictions.

Swin-Unet performed worse in the evaluated setting (Dice_m = 0.080), although it showed a small positive signal for SE (Dice_{SE} = 0.037). This result suggests that Transformer-based architectures may require more data and/or more careful tuning to outperform convolutional networks on moderately sized and imbalanced lesion datasets. Figure 5 displays the loss curves. Although the curves indicate stable learning behavior, they also suggest that additional epochs, early stopping, or hyperparameter optimization could further improve segmentation performance.



(a) U-Net++

(b) Attention U-Net

(c) Swin-Unet

Figure 5. Training and validation loss curves during segmentation model training.

6.2. Grading Results

Table 4 summarizes the metrics for the evaluated grading models on the validation set ($n = 83$). Because DR severity is ordinal (0–4), QWK is particularly informative, as it penalizes distant errors more strongly than confusions between adjacent classes. EfficientNetV2-S reached QWK = 0.889, indicating high ordinal agreement and suggesting that most errors occur between neighboring grades.

Table 4. DR grading comparison on IDRiD (validation set).

Model	QWK	Macro-F1	Weighted-F1	AUC (OVR)	n
EfficientNetV2-S	0.889	0.679	0.745	0.904	83
ConvNeXt-Tiny	0.718	0.179	0.118	0.797	83

Macro-F1 weights all classes equally and is sensitive to failures in minority classes, whereas Weighted-F1 summarizes performance according to the empirical class distribution. For EfficientNetV2-S, the values (Macro-F1 = 0.679; Weighted-F1 = 0.745;

AUC = 0.904) indicate consistent performance in both final decision quality and probabilistic separability. By contrast, ConvNeXt-Tiny showed low Macro-F1 and Weighted-F1 despite a moderate AUC, suggesting that its probabilities contain some discriminative signal, but the final argmax decision does not translate into well-calibrated per-class predictions.

Beyond the aggregated metrics, Table 5 reports the per-class performance of EfficientNetV2-S. Class 1 was the most unstable due to its very low support (4 samples), whereas classes 0, 3, and 4 showed more consistent performance. Class 2 presented partial confusion with neighboring grades, which is expected in ordinal grading tasks.

Table 5. Per-class metrics for EfficientNetV2-S (validation set).

Class	Precision	Recall	F1	Support
0	0.844	1.000	0.915	27
1	0.500	0.250	0.333	4
2	0.750	0.556	0.638	27
3	0.632	0.800	0.706	15
4	0.800	0.800	0.800	10

7. Discussion and Healthcare Implications

The quantitative and qualitative results suggest that integrating ordinal grading and multi-label lesion segmentation provides a pragmatic strategy for screening: grading summarizes disease severity and supports case prioritization, while masks and overlays enable rapid verification of whether the global decision is supported by local evidence. This is important because a triage-oriented system should not provide only a final class label; it should also expose evidence that can be inspected and audited by healthcare professionals.

From the perspective of applied healthcare computing, two aspects stand out. First, class-specific threshold parameterization makes the sensitivity-specificity trade-off explicit, allowing adaptation to different operational objectives without changing model weights. Second, automatic logging of metrics and artifacts, including training curves, per-class reports, confusion matrices, threshold files, and output images, provides a governance layer for reproducibility, auditing, and version comparison.

The results should be interpreted as evidence of prototype feasibility rather than definitive clinical validation. The grading module achieved strong ordinal agreement with EfficientNetV2-S, but the validation set contains only 83 images and is imbalanced, with particularly low support for class 1. The segmentation module also revealed limitations for small or rare lesions, especially MA and SE. These results indicate that the platform can expose useful visual evidence, but model outputs must still be interpreted with caution and should not be used as autonomous diagnostic decisions.

The evaluated neural networks are established architectures from the literature. Therefore, the contribution of this work is not a new segmentation or grading architecture, nor a claim of state-of-the-art performance. Instead, the contribution is the integration of lesion segmentation, ordinal grading, threshold calibration, visual evidence, and experiment governance into an interactive teleophthalmology platform. Direct comparison with the state of the art would require harmonized protocols, including the same datasets, train/test partitions, preprocessing, metrics, and evaluation conditions.

Several limitations must be acknowledged. First, the current quantitative evaluation is based on a single holdout validation split; future work will include k-fold cross-validation and locked test-set evaluation. Second, the study uses a single public dataset, which limits robustness assessment across camera devices, acquisition protocols, and population characteristics. Third, segmentation models were trained for 18 epochs and grading models for 10 epochs under a fixed protocol for comparability and computational feasibility; longer training, early stopping, and hyperparameter tuning may improve performance. Fourth, systematic inference-time and hardware-requirement evaluation was not included and will be considered in the operational validation stage.

In terms of applicability, the platform aligns with teleophthalmology use cases and has the potential to support remote screening and risk-based case prioritization. As next steps, we plan field validation in two partner ophthalmology clinics, combining performance evaluation on real-world data with usability analysis, including reading time, interpretability of evidence, and adequacy to clinical workflow. To support demonstrative use, the platform is available in a public Hugging Face Space: `https://bit.ly/4e9qyvt`.

8. Conclusion

This work presents an integrated system for supporting diabetic retinopathy screening from fundus images, combining lesion segmentation and ordinal grading within a reproducible workflow and an interactive interface. In experiments on IDRiD, U-Net++ achieved the best overall segmentation performance (mean Dice of 0.221), whereas EfficientNetV2-S attained strong ordinal agreement for grading (QWK of 0.889 on the validation set, $n = 83$). The grading evaluation also included Macro-F1, Weighted-F1, AUC (OvR), and per-class reports, providing a broader view of model behavior under class imbalance.

Beyond quantitative metrics, the main contribution lies in an architecture designed for auditability, interpretability, and interaction, integrating lesion-level evidence, ordinal severity estimation, threshold calibration, and systematic artifact logging. Future work will include validation in two partner ophthalmology clinics, locked test-set evaluation, k-fold cross-validation, systematic inference-time assessment, direct comparison with state-of-the-art approaches under harmonized protocols, and robustness analysis under different acquisition conditions.

Acknowledgements

This study was partially funded by the National Council for Scientific and Technological Development (CNPq), Brazil.

Ethical approval

This study used exclusively the public IDRiD dataset, which consists of anonymized fundus images with no identifiable information. The study did not collect participant data or access personal data; therefore, it did not require additional informed consent or approval from an ethics committee.

Use of generative AI tools

We used generative AI tools only to assist with minor language revisions during manuscript review. All study design, data collection, data analysis, and scientific interpretation were conceived and carried out by the authors, who assume full responsibility for the content of this work.

References

Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., and Folk, J. C. (2018). Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine*.

Arenas-Cavalli, J. T. et al. (2022). Clinical validation of an artificial intelligence-based diabetic retinopathy screening tool for a national health system. *Eye*.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M. (2021). Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*.

Chokshi, T., Cruz, M. J., Ross, J., and Yiu, G. (2024). Advances in teleophthalmology and artificial intelligence for diabetic retinopathy screening: a narrative review. *Annals of Eye Science*, 9.

Farahat, Z. et al. (2024). Diabetic retinopathy screening through artificial intelligence algorithms: a systematic review. *Survey of Ophthalmology*, 69(5):707–721.

Gulshan, V., Peng, L., Coram, M., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nakayama, L. F., Ribeiro, L. Z., Silva, P. S., et al. (2023). Artificial intelligence for telemedicine diabetic retinopathy screening: a review. *Annals of Medicine*, 55(2):2258149.

Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N., Kainz, B., Glocker, B., and Rueckert, D. (2018). Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., and Meriaudeau, F. (2020). Idrid: Diabetic retinopathy – segmentation and grading challenge. *Medical Image Analysis*, 59:101561.

Tan, M. and Le, Q. V. (2021). Efficientnetv2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Ting, D. S. W., Cheung, C. Y.-L., Lim, G., et al. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2018). UNet++: A nested U-Net architecture for medical image segmentation. *arXiv preprint arXiv:1807.10165*.