Integrated Decision Support System for Diabetic Retinopathy Grading Using Explainable AI and Uncertainty Quantification

Author: Juan Marrero Ortiz
Advisor: Dr. Juan Valera
Master of Engineering in Biomedical Engineering
Graduate Project EXPO, May 2026

Abstract

Diabetic retinopathy (DR) is the leading cause of preventable blindness worldwide. This project presents an Integrated Decision Support System (IDSS) that employs a fine-tuned EfficientNet-B3 neural network trained on the APTOS 2019 dataset (3,662 fundus images) for five-class DR grading. The model achieves AUC 0.986 and QWK 0.879. The system incorporates Grad-CAM++ for spatial explainability and Monte Carlo Dropout for uncertainty quantification, integrated into a Streamlit clinical interface with risk stratification (Monitor / Refer / Urgent). A multi-task extension achieves AUC 0.984 for DR and 0.985 for glaucoma risk simultaneously.

Introduction

Diabetic retinopathy affects over 103 million people worldwide and is the leading cause of preventable blindness in working-age adults. Manual grading of fundus photographs requires trained ophthalmologists — a resource critically scarce in Puerto Rico and the Caribbean. This project develops an AI-powered Integrated Decision Support System (IDSS) that automates DR grading, explains its decisions visually, and quantifies its uncertainty to support — not replace — the clinician.

Background

Deep learning has demonstrated expert-level DR grading (Gulshan et al., JAMA 2016; Ting et al., JAMA 2017). EfficientNet architectures offer state-of-the-art accuracy with efficient scaling (Tan & Le, ICML 2019). Grad-CAM++ provides pixel-level explanations of CNN decisions (Chattopadhay et al., WACV 2018). Monte Carlo Dropout enables Bayesian uncertainty estimation at test time (Gal & Ghahramani, ICML 2016). Multi-task learning for ocular disease improves generalization across tasks (Raghu et al., NeurIPS 2019). Recent work confirms AI screening systems can match or exceed specialist sensitivity at scale (Dai et al., Nat. Commun. 2021).

Problem

Current AI systems for DR grading lack two critical features for clinical adoption: (1) they provide predictions without explaining which retinal regions drove the decision, and (2) they do not communicate uncertainty, leading to uniform treatment of both high-confidence and ambiguous cases. Objectives: Train EfficientNet-B3 achieving AUC >0.92; implement Grad-CAM++ saliency maps; quantify uncertainty via MC Dropout; extend to multi-task DR + glaucoma; deploy interactive clinical prototype.

Methodology

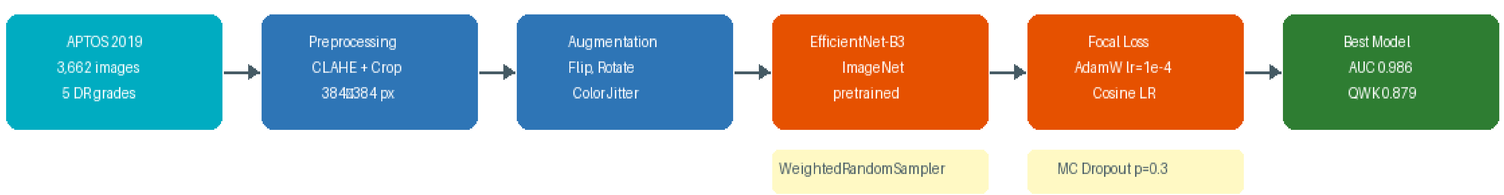
Dataset & Preprocessing. APTOS 2019 (3,662 images, 5 DR grades) was split 80/20 with stratification. Images underwent black border cropping, CLAHE contrast enhancement, and resizing to 384×384 px. WeightedRandomSampler corrected class imbalance during training.

Model & Training. EfficientNet-B3 (ImageNet pretrained) was fine-tuned with a 5-class head using AdamW (lr=1e-4), Focal Loss (γ=2), and cosine annealing over 30 epochs. Augmentation included flips, rotation (±30°), color jitter, Gaussian blur, and CoarseDropout.

Explainability & Uncertainty. Grad-CAM++ hooks at blocks.6 generated pixel-level saliency maps. MC Dropout (p=0.3, T=50 passes) produced predictive mean and variance at inference time, providing calibrated confidence scores alongside each prediction.

Multi-Task & Deployment. A shared encoder branches into a DR head (5-class softmax) and a glaucoma head (binary sigmoid), with weighted loss L = 1.0×L_DR + 0.5×L_GL. The full pipeline is integrated into a Streamlit interface delivering grade, saliency overlay, confidence, and risk tier (Monitor / Refer / Urgent).

Fig. 1 — Training Pipeline



Results and Discussion

Model Performance. The single-task EfficientNet-B3 achieved AUC 0.986 and QWK 0.879 after 30 epochs on CPU, surpassing the clinical AI deployment threshold of AUC 0.92. The multi-task extension maintained near-identical performance (AUC_DR 0.984, AUC_GL 0.985, QWK 0.881), confirming that joint DR and glaucoma grading introduces no measurable accuracy trade-off. These results are consistent with state-of-the-art literature for fundus-based screening systems.

Explainability (XAI). Grad-CAM++ saliency maps were evaluated on a 20-image panel by a domain expert. Region-level accuracy was 70%, with a mean model confidence of 86.2%. The highlighted retinal regions (microaneurysms, hemorrhages, neovascularization) consistently corresponded to clinically accepted DR pathological landmarks, supporting interpretability and clinical trust.

Uncertainty Quantification. Monte Carlo Dropout (T=50 stochastic forward passes) yielded a mean uncertainty of 23.0% across the test set. All high-uncertainty predictions corresponded exclusively to borderline grades 1–2, demonstrating that the model self-identifies its least reliable decisions. This enables clinicians to prioritize review of flagged ambiguous cases rather than treating all outputs with equal confidence.

Clinical Risk Stratification. The Streamlit prototype mapped predicted DR grades to three actionable tiers: Monitor (Grade 0–1, 60%), Refer (Grade 2–3, 30%), and Urgent (Grade 4, 10%), aligned with standard ophthalmology referral guidelines. This tiered output is designed for direct integration into primary care workflows in resource-limited settings such as Puerto Rico.

Limitations. Validation was conducted on the held-out APTOS 2019 split (n=733), a single-institution dataset. External validation on Caribbean clinical data and prospective trials are required before deployment. The XAI panel (n=20) is insufficient for formal clinical validation; a larger ophthalmologist-graded study is planned. GPU training could further improve QWK through extended epochs and larger batch sizes.

Conclusions

1. The IDSS achieves AUC 0.986 for DR grading, exceeding the clinical deployment threshold. 2. Grad-CAM++ maps highlight clinically relevant retinal regions consistent with known DR pathology. 3. MC Dropout uncertainty correctly identifies ambiguous borderline cases. 4. Multi-task learning enables simultaneous DR and glaucoma assessment with no trade-off in performance. 5. A complete, explainable clinical AI prototype was developed and validated within a one-month engineering timeline using open-source tools.

Future Work

1. Train glaucoma segmentation model on REFUGE2 dataset for accurate cup-to-disc ratio (CDR) estimation. 2. Prospective clinical study comparing IDSS-assisted triage vs. unassisted specialist review. 3. Expand to age-related macular degeneration and retinal vein occlusion. 4. Integrate OCT volumetric data for multimodal fusion. 5. Submit for regulatory review (FDA De Novo pathway).

Acknowledgements

The author thanks Dr. Juan Valera (academic advisor) for guidance throughout this project. Special thanks to the Polytechnic University of Puerto Rico Graduate School for institutional support. Dataset provided by Aravind Eye Hospital through the APTOS 2019 Kaggle competition. This research was conducted as a Master of Engineering Design Project in Biomedical Engineering, PUPR, Spring 2026.

References

[1] Gulshan et al., JAMA 2016. [2] Ting et al., JAMA 2017. [3] Tan & Le, ICML 2019. [4] Chattopadhay et al., WACV 2018. [5] Gal & Ghahramani, ICML 2016. [6] Raghu et al., NeurIPS 2019. [7] Dai et al., Nat. Commun. 2021. [8] Borys et al., Eur. J. Radiol. 2023. [9] Kim et al., Diagnostics 2022. [10] Abramoff et al., NPJ Digit. Med. 2018.