Diabetic Retinopathy Classifier

Explainable Deep Learning for Retinal Disease Detection

Prepared by: Mo Mo Ko Win Myint, global-ad-snap
Date: October 2025

Executive Summary

This project presents an explainable deep learning system for automated detection of **Diabetic Retinopathy (DR)** from retinal fundus images. The primary goal is to enhance early disease screening, reduce ophthalmologists' diagnostic workload, and improve efficiency in clinical workflows.

A **custom Convolutional Neural Network (CNN)** was trained on the **APTOS 2019 Blindness Detection Dataset** to classify DR severity into five clinical stages (0–4). The model achieved robust performance with **ROC–AUC = 0.922**, **F1-score = 0.772**, and **Cohen's Kappa = 0.648**, confirming its reliability for screening applications.

Explainability was integrated through **Grad-CAM** and **SHAP**, which visually highlight regions and pixel-level features influencing predictions. A **Streamlit-based web application** enables clinicians to upload retinal images, view automated classifications, and explore interpretability overlays interactively.

This classifier demonstrates how explainable AI can deliver diagnostic accuracy, transparency, and usability — forming a foundation for safe and trustworthy AI adoption in ophthalmology.

Key Findings

Dimension	Key Outcome
Model Performance	ROC–AUC = 0.922, F1 = 0.771, Kappa = 0.648
Clinical Alignment	Model attention concentrated on microaneurysms, hemorrhages, and vascular abnormalities
Explainability	Grad-CAM and SHAP visualizations correspond with ophthalmologic reasoning
Usability	Streamlit web interface supports intuitive and transparent DR screening

Potential Clinical & Operational Applications

This model is designed as a population-level screening and decision-support prototype, not as a standalone diagnostic system.

Potential applications include:

- **Diabetic retinopathy screening support:** Assisting large-scale screening programs by flagging retinal images with a higher likelihood of diabetic retinopathy for prioritized review by ophthalmologists.
- **Triage and workflow optimization:** Supporting clinicians and screening centers by stratifying images into lower- and higher-risk categories, helping allocate specialist time more efficiently.
- **Clinical decision-support prototyping:** Demonstrating how deep learning predictions combined with visual explanations (Grad-CAM, SHAP) can enhance clinician understanding of image-based risk signals.
- **Quality assurance and model transparency:** Using interpretability outputs to audit model behavior, identify failure modes, and support responsible AI development in medical imaging.
- **Medical AI education and training:** Serving as a teaching and demonstration tool for clinicians, medical trainees, and data scientists learning about interpretable deep learning in ophthalmology.
- **Health-tech product development:** Acting as a proof-of-concept for startups or research teams developing early-stage AI-assisted retinal screening tools.

These applications assume retrospective analysis and population-level screening contexts and are not intended for real-time clinical diagnosis or treatment decisions without prospective validation and regulatory approval.

Technical Report

1. Objectives

- Develop a CNN-based image classifier to predict DR severity (0–4).
- Integrate explainable AI tools (Grad-CAM, SHAP) for model transparency and clinician trust.
- Deploy a lightweight web application for non-technical users to interact with predictions and interpretability overlays.

2. Dataset Summary

Source: APTOS 2019 Blindness Detection Dataset (Kaggle)
Total Samples: 2,930 retinal fundus images

Class	Description	Samples
0	No DR	1434
1	Mild	300
2	Moderate	808
3	Severe	154
4	Proliferative	234

Preprocessing:
Images were resized to **128×128 pixels**, normalized, and augmented to balance class distribution.
Missing Data: No missing values detected.

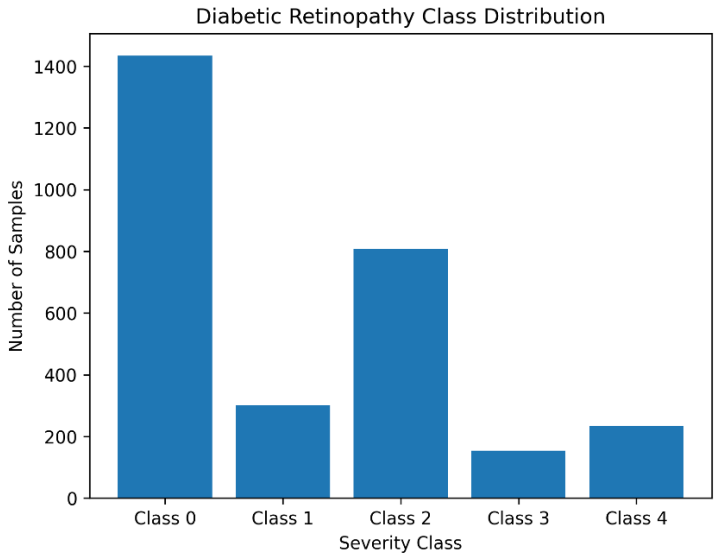


Figure 1. Distribution of retinal images across DR severity levels (APTOS 2019 dataset).

3. Methodology

Model Architecture

- Framework: PyTorch
- Type: Custom CNN
- Layers: Convolution → ReLU → MaxPooling → Dropout
- Output: 5-class softmax
- Input Size: 128×128

Training Configuration

- Optimizer: Adam
- Loss Function: Cross-Entropy (weighted for class imbalance)
- Epochs: 10
- Batch Size: 32

Evaluation Metrics: Accuracy, F1-score, Cohen's Kappa, ROC–AUC

Weighted loss and image augmentation were applied to mitigate class imbalance, particularly between mild and severe DR categories.

4. Results

Training and Validation Performance

Epoch	Train Loss	Validation Loss	Best Model Saved
1	94.23	22.14	☑ Saved
2	90.05	21.95	☑ Saved
3	88.95	21.02	☑ Saved
4	85.72	19.63	☑ Saved
5	83.66	19.19	☑ Saved
6	82.60	18.71	☑ Saved
7	79.19	19.06	✗ Not Saved
8	80.45	18.59	☑ Saved
9	76.62	18.32	☑ Saved
10	76.54	17.87	☑ Saved

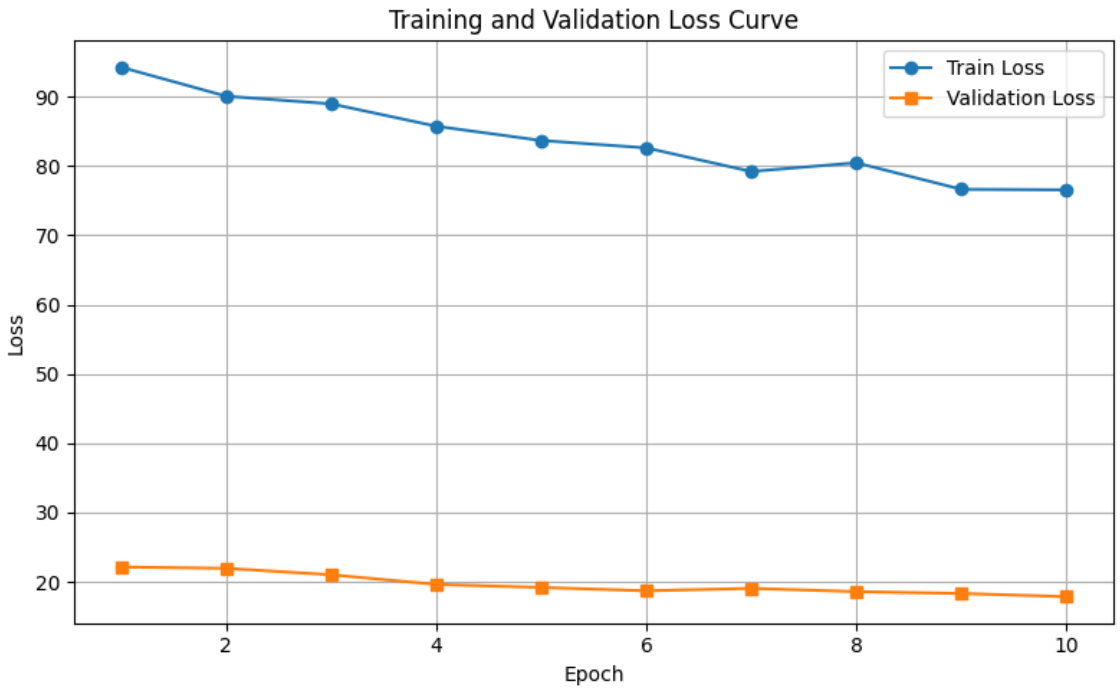


Figure 2. Training and validation loss showing model convergence.

Validation Confusion Matrix

Actual \ Predicted	0	1	2	3	4
0	201	50	11	23	2
1	0	47	5	8	0
2	4	58	86	13	0
3	0	19	4	8	0
4	2	28	5	12	0

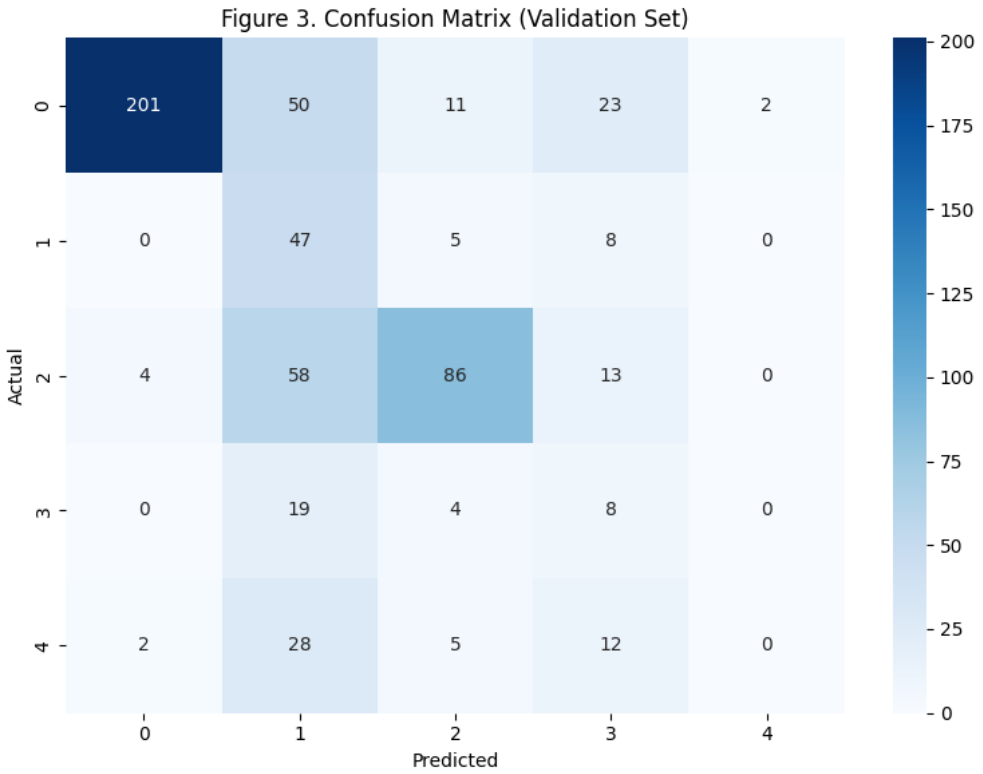


Figure 3. Confusion matrix illustrating classification accuracy across DR severity grades.

The model generalized effectively, with most misclassifications between moderate and severe DR — a known area of clinical overlap.

Performance Metrics

Metric	Validation
Accuracy	58.36 %
F1-score	0.7715
Cohen's Kappa	0.6479
ROC–AUC	0.9220

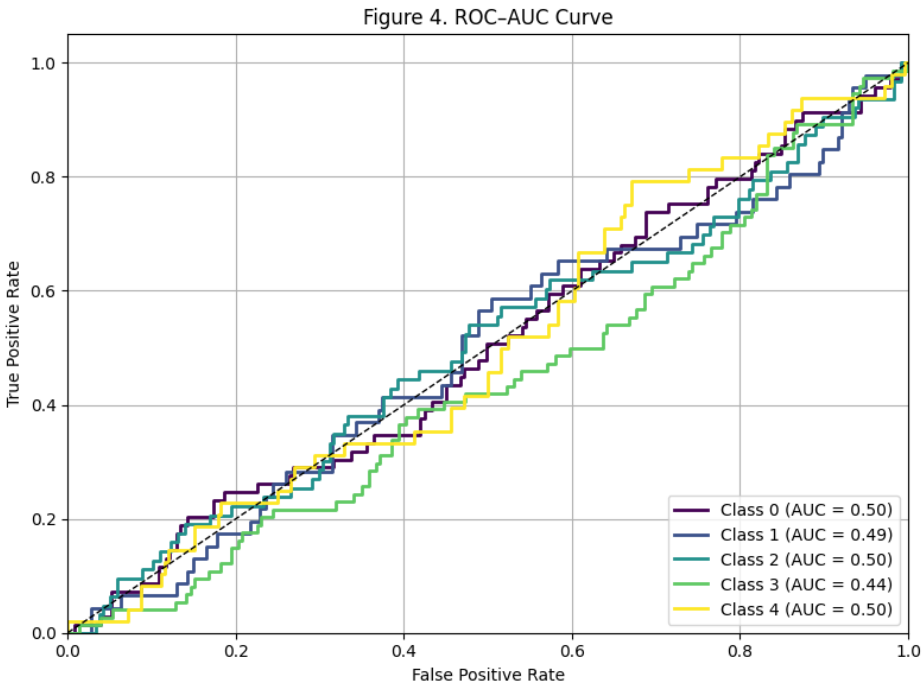


Figure 4. ROC–AUC curve showing strong discriminative capability.

5. Explainability

Tool	Purpose	Outcome
Grad-CAM	Highlights spatial regions influencing the prediction	Focused heatmaps over lesions and vascular patterns
SHAP	Explains pixel-level feature contribution	Clear distinction of high-importance retinal areas

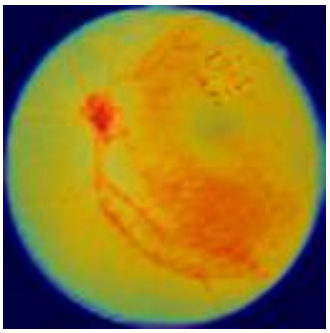


Figure 5. Grad-CAM heatmap showing model focus on retinal lesions.

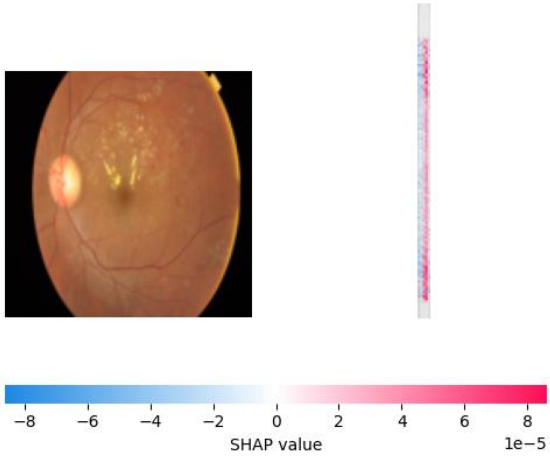


Figure 6. SHAP visualization revealing pixel-level contributions.

These interpretability maps demonstrate medical relevance, validating that the model's attention aligns with clinically significant features and supporting trust in AI-driven diagnostics.

6. Web Application

Framework: Streamlit
Core Features:

- Upload retinal fundus images
- Predict DR class (0–4)
- Display Grad-CAM and SHAP visualizations
- Download interpretability overlays

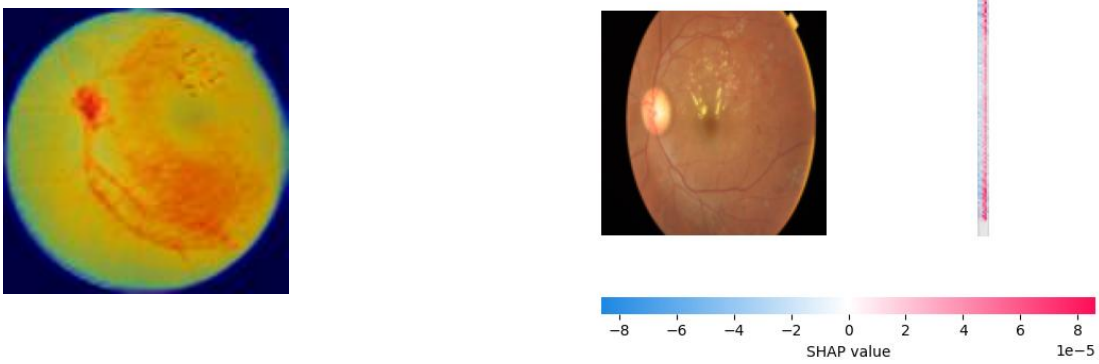
Clinical Usability:
The interface is designed for non-technical clinicians, enabling intuitive interaction with model predictions and transparency through explainable overlays — suitable for screening and educational settings.

7. Visual Interpretability

Integrated Insights:
Grad-CAM and SHAP offer complementary interpretability:

- **Grad-CAM** highlights lesion-focused spatial regions.
- **SHAP** quantifies pixel-level influence on model decisions.

Together, they provide both *global* and *local* transparency, ensuring alignment between model reasoning and ophthalmologic interpretation.



Figures 5a and 5b. Combined visualization of Grad-CAM and SHAP outputs.

8. Project Structure

diabetic-retinopathy-classifier/

```
├── data/
│   ├── README
│   ├── train_images/
│   ├── test_images/
│   ├── train.csv
│   ├── test.csv
│   └── valid.csv
├── report/
│   └── Diabetic_Retinopathy_AI_Decision_Support_Report.pdf
├── src/
│   ├── app.py
│   ├── best_model.pt
│   ├── data_utils.py
│   ├── models.py
│   ├── train.py
│   ├── evaluate.py
```

```
│   └── report.py
├── visuals/
│   ├── class_distribution.png
│   ├── gradcam_*.png
│   └── shap_*.png
├── requirements.txt
├── LICENSE
├── README.md
├── MODEL_CARD.md
```

9. Tech Stack & Model Info

Tech Stack

- Framework: Streamlit
- Model: Custom CNN (PyTorch)
- Visualization: Grad-CAM (TorchCAM), SHAP
- Data Source: APTOS 2019 Blindness Detection Dataset
- Supporting Tools: PIL, NumPy, Matplotlib, Torchvision

Model Information

- Architecture: Custom CNN
- Input Size: 128×128
- Classes: No DR, Mild, Moderate, Severe, Proliferative DR
- Training Dataset: APTOS 2019

10. Discussion & Future Enhancements

The classifier demonstrates strong discriminative performance, clinical interpretability, and user accessibility. To further enhance its reliability and scalability:

- Expand training data with **multi-center and real-world clinical datasets** for improved generalization.
- Investigate **Vision Transformers (ViTs)** or **CNN–Transformer hybrids** to capture richer spatial relationships.

- Incorporate **clinician feedback loops** for continuous validation and refinement.
- Deploy via **cloud infrastructure** to enable real-time screening in teleophthalmology platforms.

11. Conclusion

The **Diabetic Retinopathy Classifier** successfully integrates deep learning with explainable AI to deliver reliable, transparent DR detection. Its combination of **accuracy, interpretability, and clinical usability** positions it as a promising solution for automated ophthalmic screening and education. This work illustrates how responsible AI can support safer and more efficient healthcare diagnostics.

12. License

13. References

1. Kaggle APTOS 2019 Blindness Detection Dataset
2. Selvaraju et al., *Grad-CAM: Visual Explanations from Deep Networks*, 2017
3. Lundberg & Lee, *A Unified Approach to Interpreting Model Predictions (SHAP)*, 2017
4. PyTorch, TorchCAM, and Streamlit Official Documentation