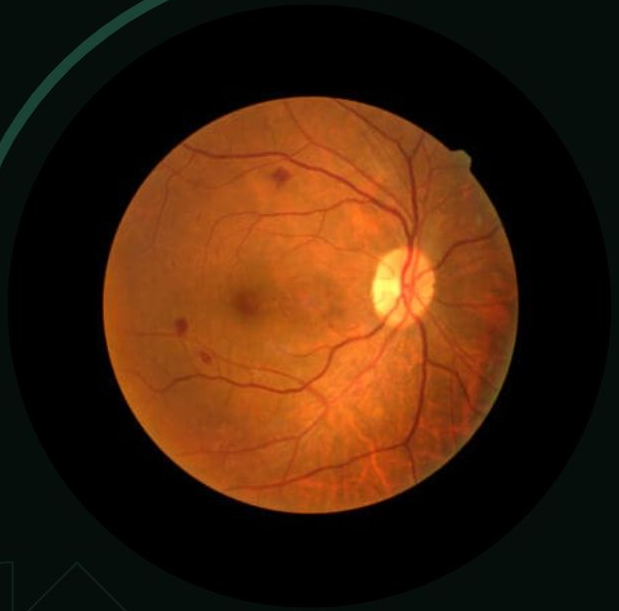
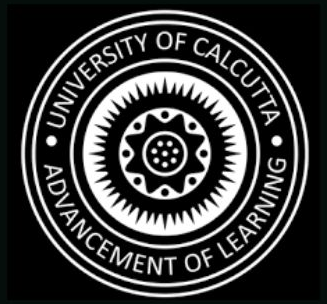


Integrating Explainable AI in Diabetic Retinopathy Detection

MCA-P41, MCA 4th Sem, 2025



Submitted By : -

Arkaprabha Ray	544-1112-1220-20	C91/MCA/232003
Prantik Bhattacharya	544-1111-0667-20	C91/MCA/232012
Rahul Das	544-1112-0306-20	C91/MCA/232016

Under the supervision of:-
Prof. Arpan Murmu
Assistant Professor,AKCSIT,University of Calcutta

Table of Contents

Understanding Diabetic Retinopathy: Causes and Impact

Retinal Damage

Diabetic retinopathy is caused by prolonged high blood sugar, which damages the retina's blood vessels.

This damage can lead to vision impairment.

Stages of DR

Diabetic retinopathy has four stages: mild, moderate, severe, and proliferative.

Each stage indicates increasing severity, from microaneurysms to complete retinal damage.

Global Prevalence

Millions of diabetic patients worldwide are at risk of developing diabetic retinopathy.

It is a leading cause of preventable blindness globally.

The Role of AI in Transforming Diabetic Retinopathy Diagnosis

1

Data Collection

High-quality retinal images are collected from diverse sources.

These images are reviewed and labeled by trained professionals based on standardized severity scales.

2

Stage Detection

AI models employ deep learning techniques to identify and classify stages of diabetic retinopathy from retinal images.

This automation streamlines the diagnostic process, making it more accessible and less reliant on specialized expertise.

3

Interpretablility

Visualization techniques transforms DL models from "black boxes" into more transparent and understandable systems.

This interpretability is vital for debugging models, identifying and mitigating biases, and ultimately, developing more reliable and robust AI applications in diverse fields.

Notable research works:-

Optimized Deep CNN for Detection and Classification of Diabetic Retinopathy and Diabetic Macular Edema

Organization: Published in BMC Medical Imaging
Year: 2024
Description: This study presents an automated approach for detecting and grading DR and Diabetic Macular Edema (DME) using retinal fundus images. The methodology involves preprocessing with DWT, segmentation using ANN, feature extraction with AGF, feature selection via RF, and classification using a Deep CNN optimized with the Chicken Swarm Algorithm (CSA). The proposed system achieved an accuracy of 97.91%.

Lesion Detection and Grading of Diabetic Retinopathy via Two-stages Deep Convolutional Neural Networks

Organization: Tencent AI Lab, China
Year: 2017
Description: This study proposes a two-stage DCNN approach for DR analysis. The first stage focuses on lesion detection, while the second stage grades DR severity. By integrating local and global networks and introducing an imbalanced weighting map, the method enhances lesion localization and grading performance, achieving results comparable to trained human observers.

Automatic Detection of Diabetic Retinopathy Using Custom CNN and Grad-CAM

Organization: LaROSERI Laboratory, Chouaib Doukkali University, Morocco
Year: 2020
Description: This study proposes a lightweight custom CNN architecture for DR diagnosis using Optical Coherence Tomography (OCT) images. Leveraging transfer learning with MobileNet and employing Grad-CAM for visual explanations, the model achieved an accuracy of 80%, precision of 85%, and recall of 80.5%.

Ensemble of Convolutional Neural Networks for Automatic Grading of Diabetic Retinopathy and Macular Edema

Organization: Indian Institute of Technology (IIT) Hyderabad, India
Year: 2018
Description: This research employs an ensemble of CNNs for automated grading of DR and macular edema. Utilizing transfer learning and a max-voting approach, the ensemble achieved an accuracy of 83.9% for DR grading and 95.45% for macular edema grading on test datasets.

Dataset Preparation and Preprocessing

1

Data Collection

The dataset was sourced from the APTOS 2019 Blindness Detection dataset.

It consists of 3,662 retinal images labeled by trained doctors.

2

Data Cleaning

Images were reviewed to ensure quality and accuracy.

This step is crucial for reliable model training.

3

Data Augmentation

Techniques like rotation and flipping were applied to enhance dataset diversity.

These methods help mitigate the risk of overfitting.

4

Image Resizing

All images were resized to 224x224 pixels for consistency.

This standardization is essential for compatibility with CNNs.

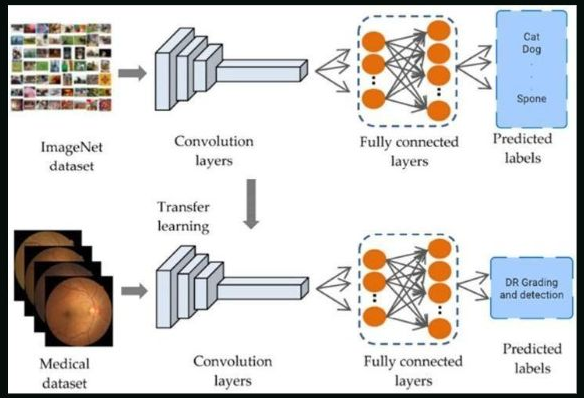
5

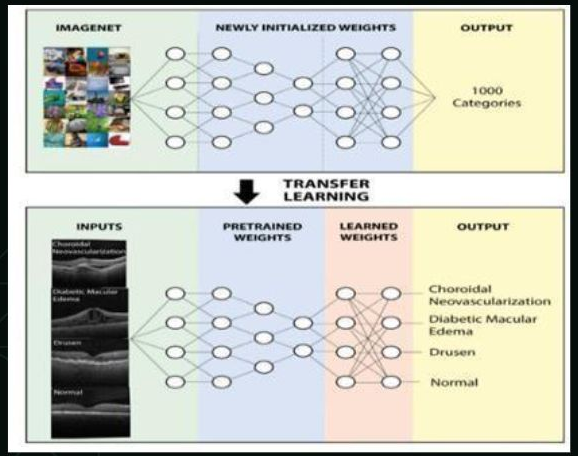
Normalization Techniques

Pixel values were normalized to a standard range to stabilize training.

Normalization improves the overall performance of the model.

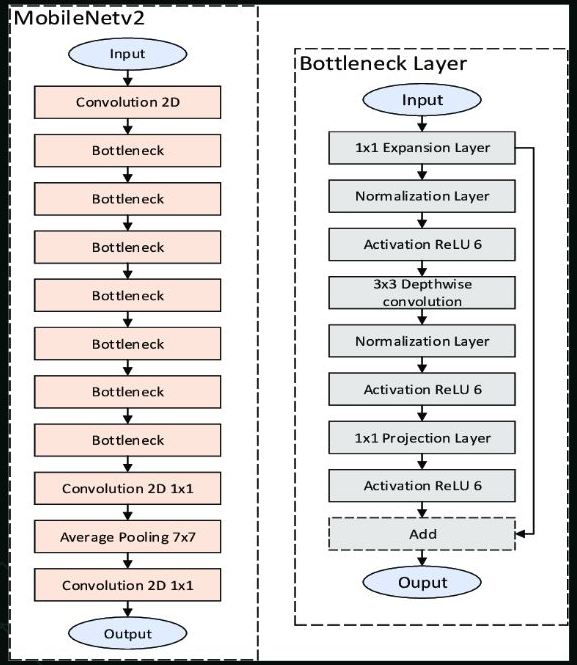
Fine tuning various Pretrained models and evaluating their performance

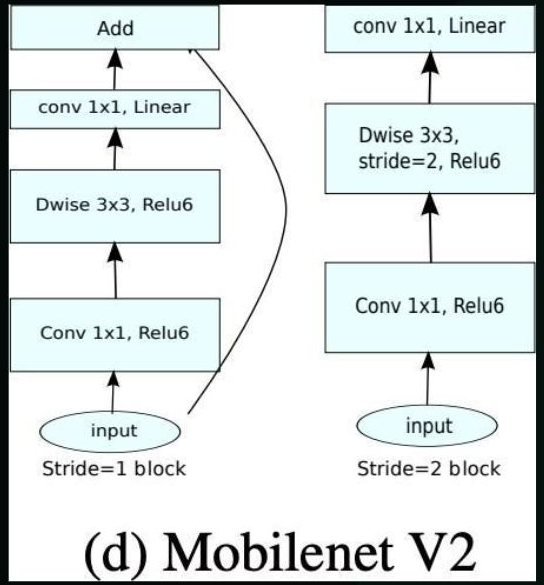




	model	train_accuracy	val_accuracy	Training time (sec)
0	MobileNetV2	0.7242	0.7720	16.75
1	DenseNet169	0.7229	0.7660	25.34
2	ResNet152V2	0.7205	0.7599	25.97
3	InceptionV3	0.7057	0.7538	19.17
4	DenseNet201	0.7326	0.7538	27.69
5	NASNetMobile	0.7040	0.7508	29.25
6	MobileNet	0.7316	0.7508	14.66
7	InceptionResNetV2	0.6949	0.7477	28.43
8	DenseNet121	0.7161	0.7447	41.24
9	ResNet50V2	0.7245	0.7416	17.16
10	Xception	0.7097	0.7386	18.19
11	ResNet101V2	0.7222	0.7204	21.22
12	VGG16	0.6719	0.7112	15.75
13	VGG19	0.6521	0.7021	15.04
14	ResNet50	0.6524	0.6687	17.24
15	ResNet152	0.6372	0.6626	28.66
16	ResNet101	0.6291	0.6353	22.20
17	MobileNetV3Large	0.5509	0.6049	18.08
18	MobileNetV3Small	0.4727	0.4833	17.25
19	EfficientNetB1	0.4784	0.4802	22.41
20	EfficientNetB6	0.4798	0.4802	27.34
21	EfficientNetB0	0.4835	0.4802	20.60
22	EfficientNetB2	0.4710	0.4802	23.08
23	EfficientNetB3	0.4791	0.4802	25.13
24	EfficientNetB4	0.4804	0.4802	27.15
25	EfficientNetB7	0.4788	0.4802	28.17
26	EfficientNetB5	0.4761	0.4802	27.78

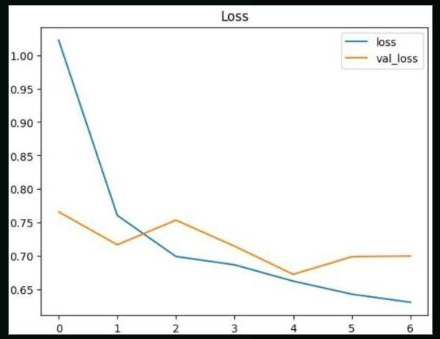
MobileNetV2 architecture

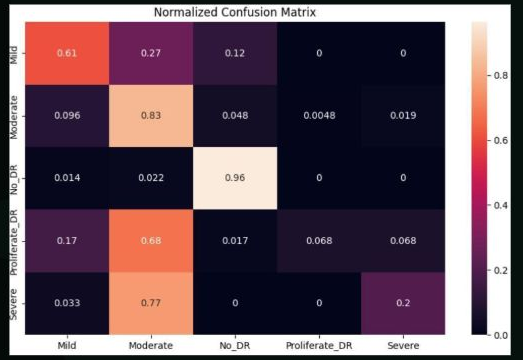


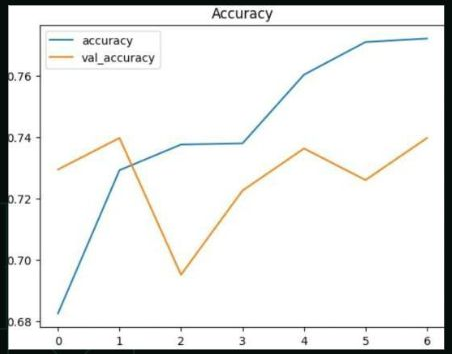


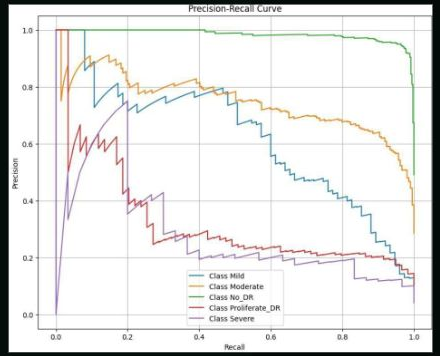
Performance evaluation using MobileNetV2 pretrained model

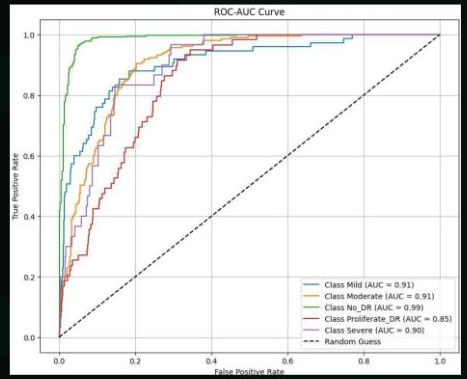
	precision	recall	f1-score	support
Mild	0.56	0.61	0.59	75
Moderate	0.66	0.83	0.73	209
No_DR	0.95	0.96	0.95	360
Proliferate_DR	0.80	0.07	0.12	59
Severe	0.43	0.20	0.27	30
accuracy			0.79	733
macro avg	0.68	0.54	0.53	733
weighted avg	0.79	0.79	0.76	733











Interpretability Methods

GRAD-CAM

Gradient-weighted Class Activation Mapping

Visualizes important regions in the image influencing the model's decision.

LIME

Local Interpretable Model-agnostic Explanations

Builds local surrogate models to explain individual predictions.

SHAP

SHapley Additive exPlanations

Uses Shapley values from game theory to assign feature importance.

Integrated Gradients

Measures feature importance by accumulating gradients over input interpolations.

Why GRAD-CAM?

1. Produces clear visual heatmaps for CNN models
2. Works without modifying or retraining the original model.
3. Highly effective for image-based deep learning tasks
4. Intuitive for both technical and non-technical audiences.

Why not others?

1. LIME -> Less reliable for image data; superpixel segmentation may miss deep features.
2. SHAP -> High computational cost; explanations are harder to visualize for images.
3. Integrated Gradients -> Needs a suitable baseline image; results are less intuitive compared to Grad-CAM heatmaps.

Working & Output of GRAD-CAM

1. **Forward Pass**
   Pass the input image through the CNN to get the prediction and feature maps from the last convolutional layer

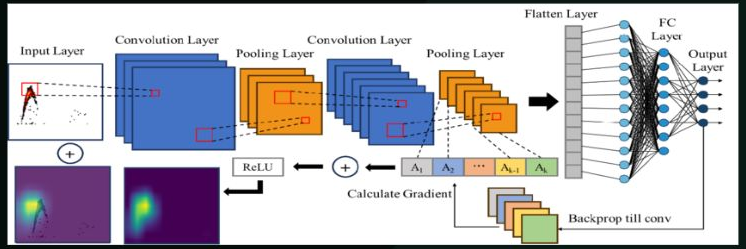
2. **Gradient Calculation**
   Compute the gradients of the target class score (e.g., predicted class) with respect to the feature maps.

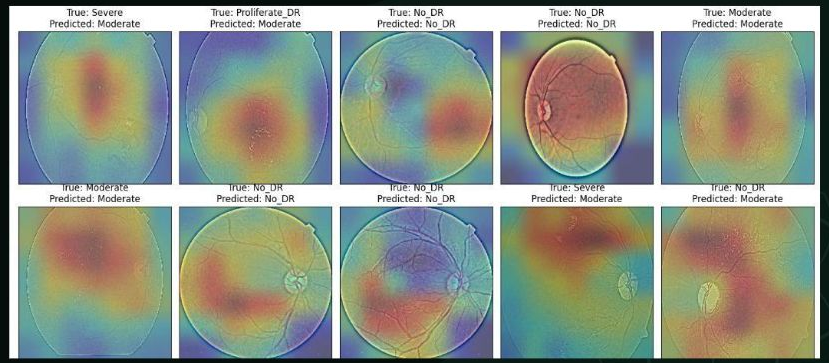
3. **Global Average Pooling**
   Calculate the average of the gradients for each feature map channel to get importance weights.

4. **Weighted Combination**
   Multiply each feature map by its corresponding weight and sum them up.

5. **ReLU Activation**
   Apply ReLU to the combined map to retain only the positive influences (focus on class-specific features).

6. **Upsampling**
   Resize the resulting heatmap to match the input image dimensions for visualization.





Our model:-

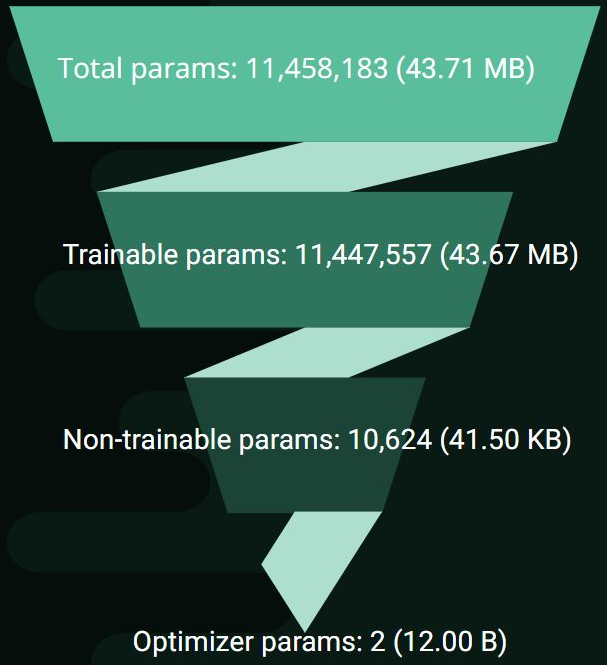
Based on the experiments we have conducted, we created a convolutional neural network (CNN) with residual connections, which appears to be inspired by the ResNet architecture but with some modifications.

Architecture Overview:-

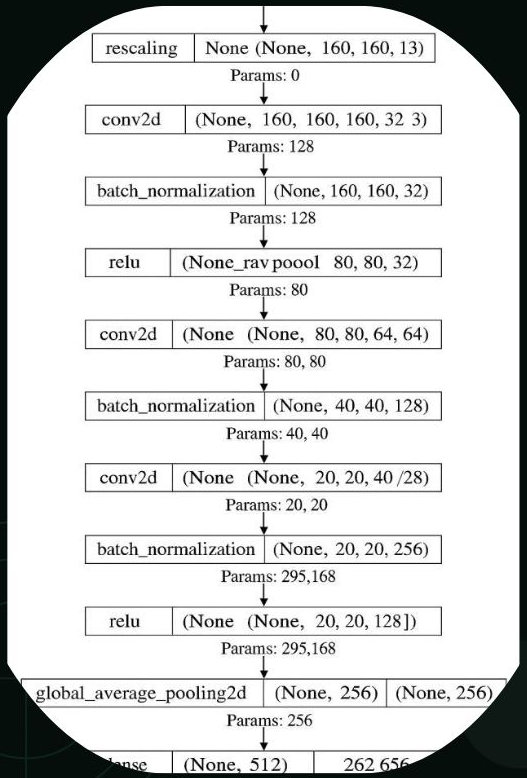
01 **An initial convolutional block**

02 **Multiple residual blocks with increasing filter sizes**

03 **A final classification head with dense layers**



Key Features of our model:-



Residual Connections

The model uses skip connections (Add layers) that allow gradients to flow directly through the network, helping with training deeper architectures.

Global Average Pooling

Used before the final dense layers to reduce spatial dimensions while preserving channel information.

Progressive Downsampling

The spatial dimensions are reduced from 224×224 to 7×7 through successive convolutional and pooling operations while increasing the number of filters (64 → 128 → 256 → 512).

Regularization

Includes dropout in the final layers to prevent overfitting.

Batch Normalization

Applied after every convolutional layer to stabilize and accelerate training.

Model Performance

Accuracy Achieved

87.7%

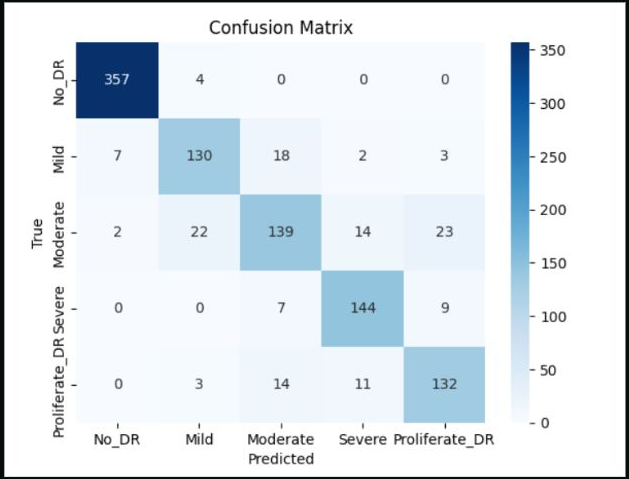
Precision Rate

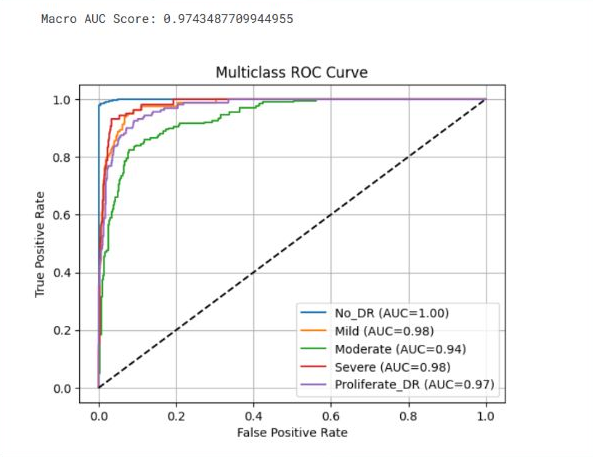
84%

Recall Performance

84.2%

Model Evaluation:-





	precision	recall	f1-score	support
No_DR	0.98	0.99	0.98	361
Mild	0.82	0.81	0.82	160
Moderate	0.78	0.69	0.74	200
Severe	0.84	0.90	0.87	160
Proliferate_DR	0.79	0.82	0.81	160
accuracy			0.87	1041
macro avg	0.84	0.84	0.84	1041
weighted avg	0.86	0.87	0.86	1041

Future Enhancements and Deployment Strategies

1

Model Enhancements

Incorporate additional features such as patient demographics and clinical history to improve predictive performance.

Explore ensemble techniques to combine multiple models for better accuracy.

2

Real-Time Integration

Develop a mobile or web-based application for real-time diabetic retinopathy detection in clinical settings.

Optimize the model for deployment on edge devices like smartphones and tablets.

3

Scalable Deployment

Ensure seamless integration into existing healthcare workflows.

Adapt the system for efficient use in resource-constrained environments.

Official Clinical Implementations of CNN and Grad-CAM in Diabetic Retinopathy Diagnosis

RetCAD v1.3.1 by Thirona (Netherlands)

Clinical Use: Implemented in a tertiary hospital screening program.

Performance: Achieved an Area Under the ROC Curve (AUC) of 0.988 for detecting referable DR.
Impact: Enabled a 96% reduction in ophthalmologist workload with minimal false negatives.

VeriSee DR by Acer (Taiwan)

Clinical Use: Deployed in hospitals across Taiwan for DR screening.

Performance: Reported a sensitivity of 92.3% and specificity of 93.7% in real-world settings.
Regulatory Approval: Received clearance from Taiwan's Food and Drug Administration.

IDx-DR (Digital Diagnostics, USA)

Clinical Use: First autonomous AI system approved by the U.S. FDA for DR screening.

Performance: Demonstrated high sensitivity and specificity in detecting more-than-mild DR.
Regulatory Approval: FDA-approved for use in primary care settings without the need for an eye specialist.

Teleophthalmology Programs in Canada

Clinical Use: Implemented across various provinces, including Alberta and Ontario, to screen for DR in remote communities.

Impact: Reduced average wait times for specialist consultations and increased screening coverage.

Sankara Nethralaya's Teleophthalmology Initiative (India)

Clinical Use: Since 2003, has screened over 450,000 patients in rural India using mobile units equipped with fundus cameras.

Impact: Provided early detection and treatment of DR, preventing vision loss in underserved populations.

Thank You