DR Eye Platform

An AI-Powered Diabetic Retinopathy Grading & Clinical Referral System

Prepared by a team of 3:

- **Jana sayedahmad**
- **Mhamad El-khatib**
- **Karim hajj chhade**

Supervised by: Dr. Abbass Rammal

Lebanese University · Faculty of Information & Documentation · Data Science 2025–2026

~20 minutes

28 Slides

0.8888	82%	5	2
QWK Score	Accuracy	Models	Languages

Presentation Roadmap

section 1

Problem & Motivation

7 min

Slides 3–10

- Global DR crisis
- Lebanon access gap
- Why AI?
- Literature & benchmark

section 2

Model & Results

8 min

Slides 11–19

- 5 architectures compared
- QWK 0.8888
- Grad-CAM explainability
- Confidence scoring

Section 3

Platform & Demo

5 min

Slides 20–28

- Full Flask platform
- 4-step patient flow
- WhatsApp delivery
- Limitations & future

01

The Problem

Why diabetic retinopathy demands an AI solution — right now

Diabetic Retinopathy · Global Scale · Lebanon Gap · Why AI

Section 1 — Slides 3–10

Presented by : Mhamad El-khatib

Diabetic Retinopathy: Scale & Silence

537M

adults with diabetes worldwide (IDF 2021)

103M

affected by diabetic retinopathy

28M

at vision-threatening stage

34.6%

of diabetics develop DR (Arora 2024)

Entirely Silent

No symptoms in early or middle stages. By the time vision blurs, irreversible damage has occurred. Screening is the only defense.

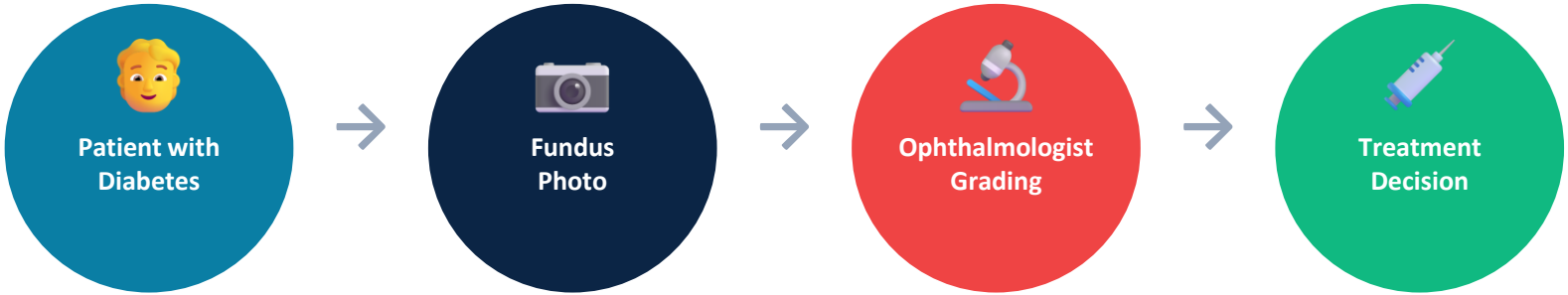
Specialist Shortage

In Lebanon and across the Arab world, the ophthalmologist-to-patient ratio is critically low. Rural patients may never receive a single eye exam.

Treatment Exists

Laser therapy, anti-VEGF injections, and surgery are highly effective — but only when applied early. Detection is the only barrier.

The Bottleneck: Screening Cannot Scale



Waiting Times

Patients wait weeks to months for an ophthalmology appointment in Lebanon.

Cost Barrier

Specialist consultation fees are unaffordable for many diabetic patients, especially elderly and rural.

Geographic Gap

Most ophthalmologists are in Beirut. Patients in Bekaa, North, or South have almost no access.

Workforce Shortage

Lebanon does not have enough trained ophthalmologists to screen its entire diabetic population annually.

DR Eye Platform: Two Gaps, One Solution

We address two distinct but equally important problems simultaneously:

❌ Clinical Gap

Grading retinal photographs requires a trained ophthalmologist. Slow, expensive, and geographically inaccessible for most Lebanese patients.

✅ AI Grading Engine

ConvNeXt-Base deep learning model grades retinal photos in under 1 second. QWK = 0.8888 — matching the published clinical benchmark.

❌ Access Gap

Even patients who get screened struggle to reach a specialist. Phone calls, secretaries, long waiting lists — all create friction that prevents timely treatment.

✅ Clinical Referral Platform

Bilingual Arabic/English Flask web app. Grading → AI report → Doctor directory → WhatsApp delivery in one seamless workflow.

🔑 **Key insight: most published DR systems stop at the model. We built the complete clinical pathway — from upload to specialist contact.**

Literature Foundation: Why AI Works

2016 | Landmark validation

Gulshan et al. · JAMA

Inception-v3 trained on 128,175 images achieves AUC 0.991 — comparable to ophthalmologists. Established AI screening as clinically viable.

2023 | Class imbalance

Vij & Arora · Multimedia Tools

Weighted sampling outperforms oversampling/undersampling for imbalanced DR datasets.

2020 | Our benchmark

Chetoui & Akhloufi · IEEE EMBC

EfficientNet on APTOS 2019 with Grad-CAM achieves QWK 0.89. Demonstrated explainable AI + high accuracy together.

2024 | Generalization

Arora et al. · Scientific Reports

DR affects 34.6% of diabetics. Distribution shift is the key barrier to real-world deployment.

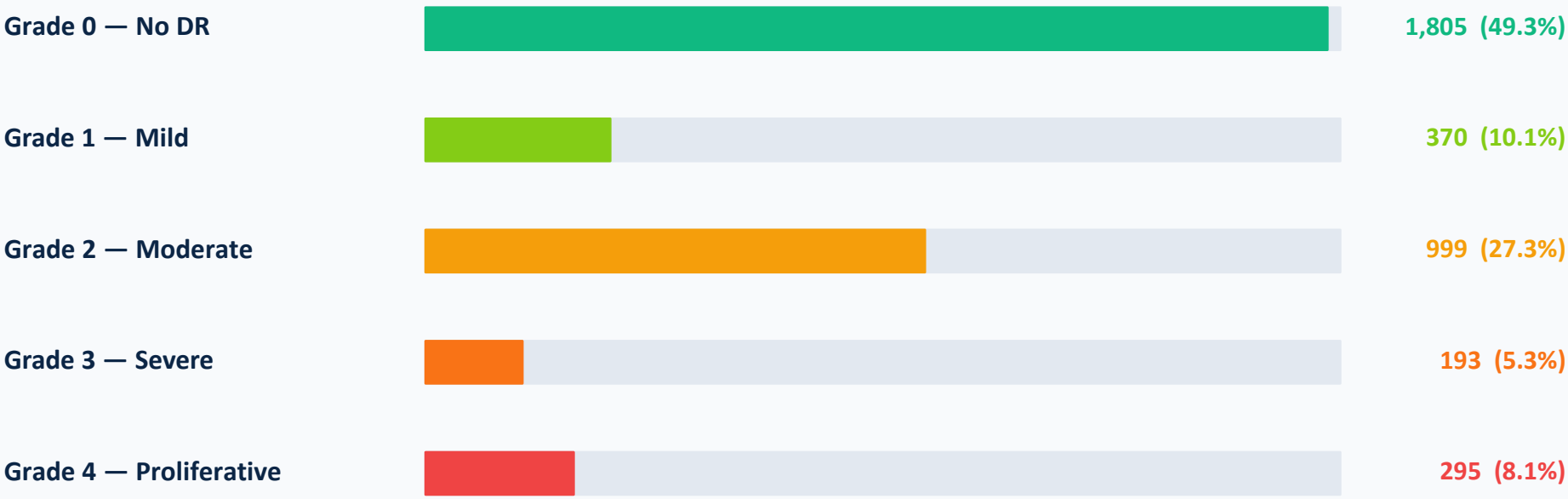
2017 | Grad-CAM method

Selvaraju et al. · ICCV

Introduced Gradient-weighted Class Activation Mapping — visual explanation technique showing WHERE the model looks.

Dataset: APTOS 2019 — 3,662 Retinal Photographs

Asia Pacific Tele-Ophthalmology Society · 5 severity grades · Rural Indian clinics · Single archive (Colab-compatible)



3,662	2,929	733	9.3:1
Total images	Training (80%)	Validation (20%)	Class imbalance ratio ⚠

3-Stage Preprocessing Pipeline

1

Black Border Crop

Binary threshold (tol=7) → largest contour → bounding box crop

Removes non-clinical background that varies by camera model. Prevents model from learning camera-style features instead of retinal features.

2

Resize to 224×224

Bilinear interpolation to fixed square resolution (380×380 for EfficientNetB4)

Neural networks require fixed input dimensions. EfficientNetB4 uses 380×380 — its compound-scaled native resolution.

3

CLAHE Enhancement

L channel of LAB space · clipLimit=2.0 · tileGridSize=8×8

Makes microaneurysms, hemorrhages, and exudates visible. Local enhancement avoids noise amplification of global methods.

WeightedRandomSampler

Weight = 1/class_size. Severe (193 imgs) sampled ~9.4× more than No DR (1,805). Corrects 9.3:1 imbalance without discarding data.

Augmentation (training only)

Horizontal flip (p=0.5) · Rotation ±15° · ImageNet normalization. Physiologically plausible only — no color jitter or extreme rotation.

Evaluation Metric: Quadratic Weighted Kappa

Why standard accuracy fails — and what we use instead:

❌ Standard Accuracy

- A model that predicts No DR for EVERY image achieves 49.3% accuracy — with zero clinical value
- Treats all misclassifications equally: Grade 0→4 error = Grade 2→3 error
- Clinically indefensible — a Grade 4 predicted as Grade 0 is catastrophic

✅ Quadratic Weighted Kappa

- Corrects for chance agreement — a 49% predictor scores ~0.0
- Penalizes large grade gaps QUADRATICALLY: Grade 0→4 = penalty 1.0, Grade 2→3 = penalty 0.0625
- Standard for ordinal medical tasks — used in all 5 reviewed papers and APTOS competition

QWK Interpretation Scale:

< 0.20	0.21–0.40	0.41–0.60	0.61–0.80	0.81–1.00
Slight	Fair	Moderate	Substantial	Almost Perfect

02

Models & Results

Five architectures · QWK 0.8888 · Grad-CAM · Confidence scoring

🧠 SimpleCNN → ResNet50 → DenseNet121 → EfficientNetB4 → ConvNeXt-Base

Section 2 — Slides 11–19

Presented by: Jana sayedahmad

Five Architectures: 2015 → 2022

All trained under identical conditions — differences in results are purely architectural

Baseline

QWK 0.6087

SimpleCNN

3 conv blocks from scratch
No pretrained weights
Establishes lower bound

2015

QWK 0.8272

ResNet50

Residual skip connections
Solves vanishing gradient
50 layers deep

2017

QWK 0.8489

DenseNet121

Dense connections — every
layer sees all prior layers
Strong feature reuse

2019

QWK 0.8415

EfficientNetB4

Compound scaling of depth
width, and resolution
380×380 native input

2022

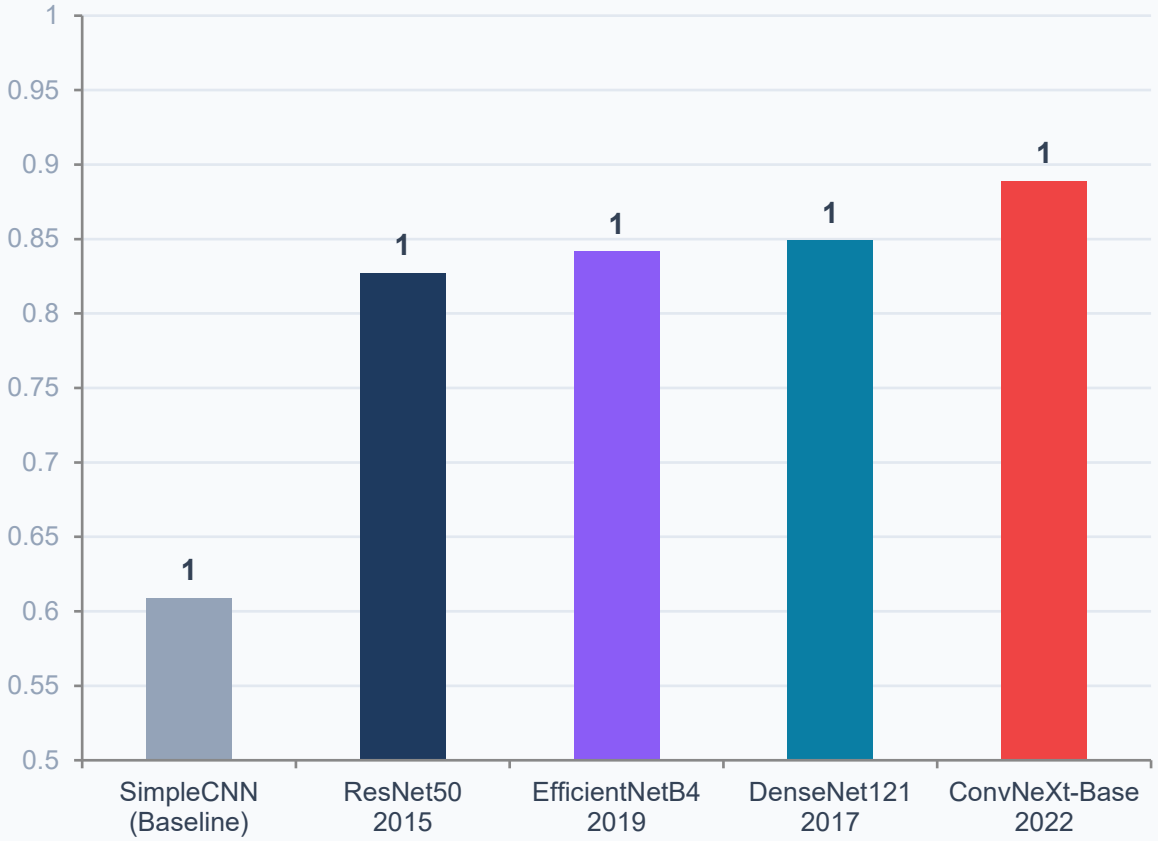
QWK 0.8888

ConvNeXt-Base

CNN redesigned with Transformer principles
7×7 depthwise conv, GELU

Identical protocol for all 4 pretrained models: Phase 1 — freeze backbone, 3 epochs, LR=0.001 → Phase 2 — full fine-tune, 5–8 epochs

Results: Quadratic Weighted Kappa — All Five Models



+0.2185

Largest single jump
SimpleCNN → ResNet50
Transfer learning impact

+0.0616

ResNet50 → ConvNeXt
across 4 architecture
generations (2015–2022)

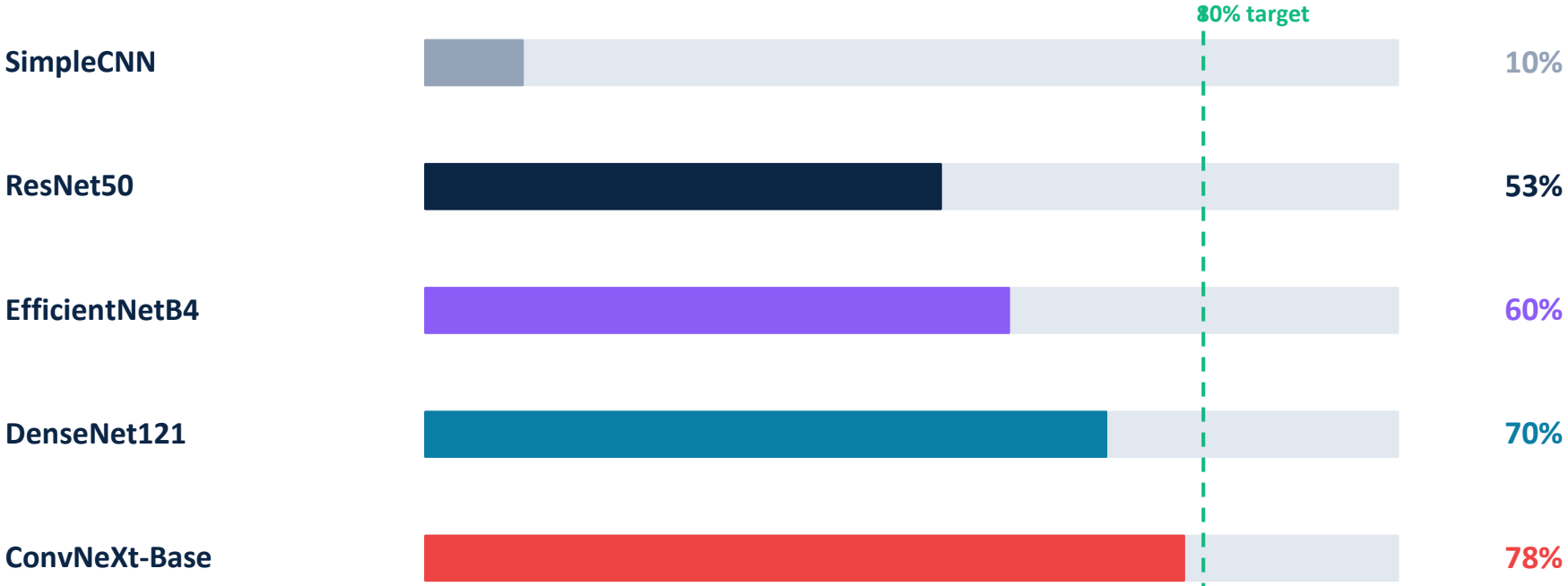
0.8888

vs 0.89

Published benchmark
(Chetoui 2020)
✅ Effectively matched

The Clinical Story: Moderate DR Recall

Grade 2 (Moderate) recall — the inflection point where ophthalmology referral becomes time-sensitive:



Complete Model Comparison

Model	Year	QWK	Accuracy	Moderate Recall	Proliferative F1	vs Benchmark
SimpleCNN	Baseline	0.6087	57%	0.10	0.22	−0.28
ResNet50	2015	0.8272	72%	0.53	0.38	−0.063
EfficientNetB4	2019	0.8415	75%	0.60	0.44	−0.049
DenseNet121	2017	0.8489	79%	0.70	0.51	−0.041
ConvNeXt-Base	**2022**	**0.8888**	**82%**	**0.78**	**0.61**	**−0.001**
Benchmark	**2020**	**~0.89**	**—**	**—**	**—**	**TARGET**

★ *ConvNeXt-Base trained for 8 Phase 2 epochs (vs 5 for other models) to allow full convergence — justified by the superior performance gain.*

★ *EfficientNetB4 slightly below DenseNet121 despite being newer: smaller effective batch size (16 vs 32) due to 380×380 resolution likely limits convergence in 5 epochs.*

Grad-CAM: Making the AI Explain Itself

A clinician cannot trust what they cannot understand. Grad-CAM shows WHERE the model looks when making each prediction.

🔢 How Grad-CAM Works

1. Compute gradient of predicted class score w.r.t. final conv layer activations

2. Global average pool gradients → per-channel importance weights α

3. Weighted sum of feature maps + ReLU → coarse heatmap

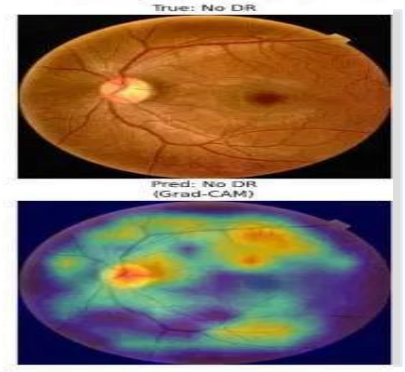
4. Upsample to input resolution → overlay with jet colormap

5. Red = high importance · Blue = low importance

Implemented with pytorch-grad-cam library

Target: conv_head layer of ConvNeXt-Base

✅ Case 1: Correct

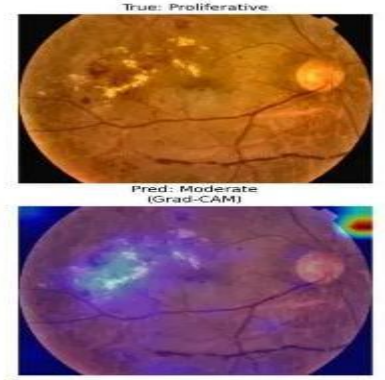


True: Grade 2
Predicted: Grade 2

Confidence: 86.5% ✅ HIGH

Heatmap correctly highlights hard exudates

⚠️ Case 2: Flagged



True: Grade 4
Predicted: Grade 2

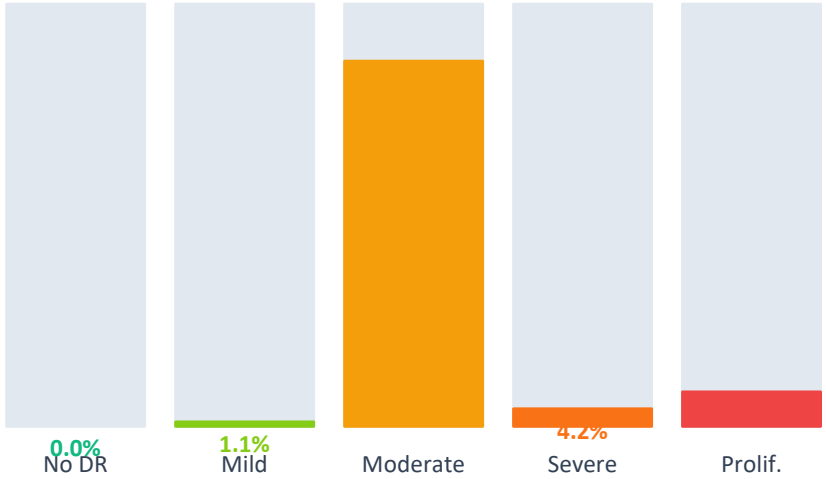
Confidence: 37.4% ⚠️ LOW

Safety flag triggered: refer to specialist

Confidence Scoring: The Safety Mechanism

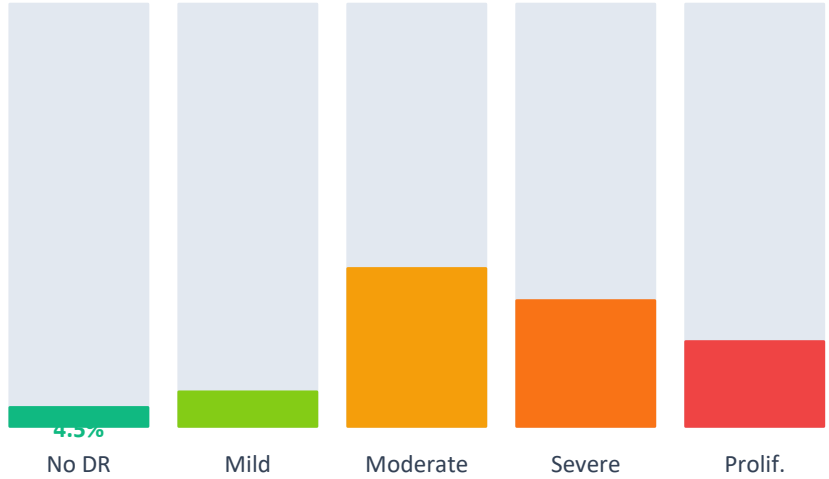
Confidence = max(softmax(logits)) · Threshold: 60% · Below → flagged for specialist review

Case 1 — Correct Prediction (Grade 2 Moderate)



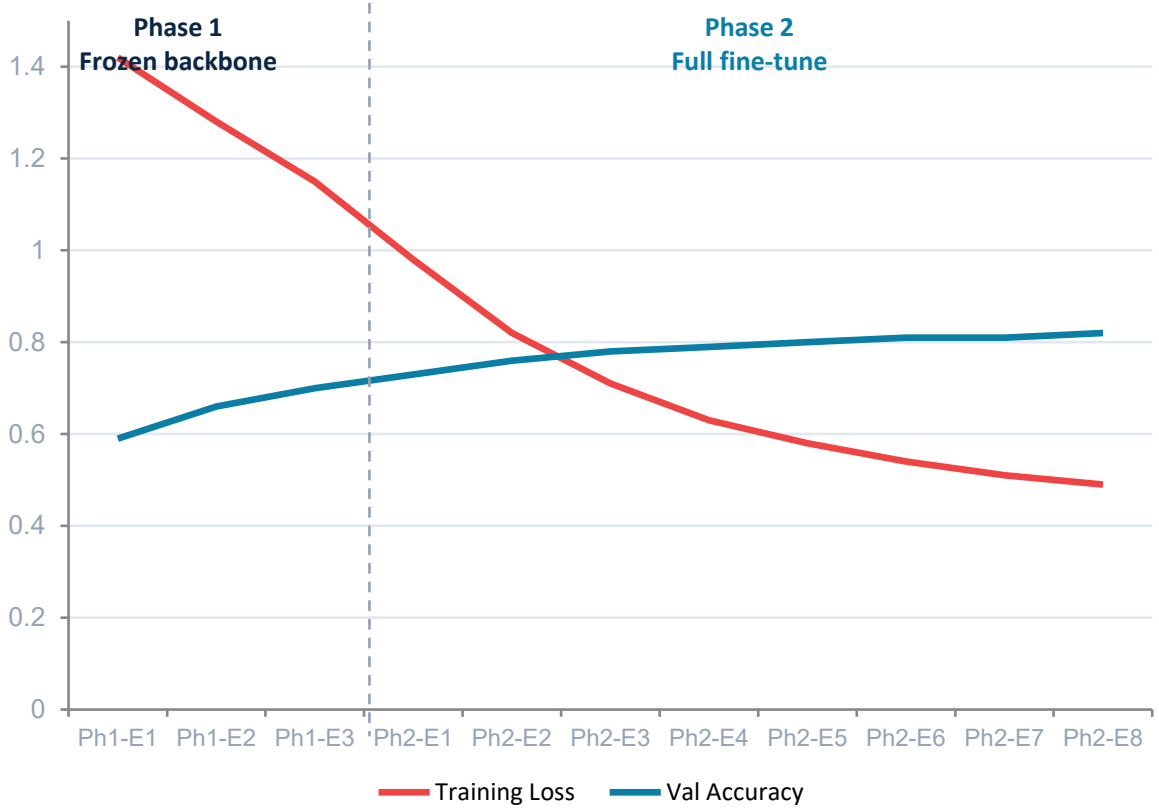
Confidence: 86.5% ✅ HIGH — No flag — Prediction accepted

Case 2 — Misclassified (True: Grade 4 Proliferative)



Confidence: 37.4% ⚠️ LOW — Flagged → specialist review required

ConvNeXt-Base: Training History



Training Config

Architecture	ConvNeXt-Base
Input size	224 × 224
Phase 1 LR	0.001
Phase 2 LR	0.00005
Batch size	32
Phase 1	3 epochs (frozen)
Phase 2	8 epochs (full)
Optimizer	Adam
Loss	CrossEntropyLoss
Final QWK	0.8888 ✅

Technical Contributions: What We Built

C1

Systematic Multi-Architecture Comparison

5 models · identical protocol · 2015–2022 trajectory · QWK 0.6087→0.8888. Every improvement is reproducible and directly comparable.

C2

Matched Published Benchmark

QWK 0.8888 on APTOS 2019 — 0.001 below Chetoui 2020 (0.89) on the same dataset. State-of-the-art level with free-tier GPU.

C3

Integrated Explainability + Safety

Grad-CAM visual explanations + softmax confidence flagging deployed together in a live system. Correctly self-identified the sole misclassification.

C4

Transfer Learning on Limited Data

Demonstrated that ConvNeXt-Base pretrained weights + careful fine-tuning achieves clinical-level performance on just 2,929 training images.

03

The Platform

Flask · Claude/Groq AI · WhatsApp · Arabic + English · 4-step patient workflow

🌐 Screen → Chat → Report → Send to Doctor

Section 3 — Slides 20–28

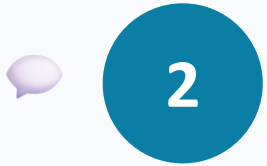
Presented by: Karim hajj chhade

4-Step Patient Workflow

STEP 1

Upload or Skip

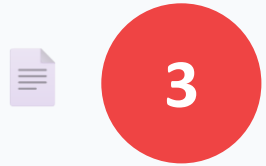
- Drag-and-drop retinal photo
- ConvNeXt grades instantly
- No photo? → Skip to chat
- Image upload is OPTIONAL



STEP 2

AI Symptom Chat

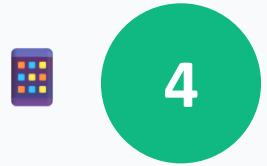
- 4 structured questions
- Numbered options + free text
- Medication photo upload
- Groq LLM — fast & free



STEP 3

Medical Report

- Doctor-facing clinical report
- 3 preprocessing images shown
- Medication photo included
- English + Arabic generated



STEP 4

Send to Doctor

- Filter by city & specialty
- One-click WhatsApp delivery
- Full report + images sent
- PDF download available

Platform Architecture — 6 Modular Components

AI Grading Engine	PyTorch 2.1 + timm + ConvNeXt-Base + pytorch-grad-cam	ml.py
LLM Integration	Groq API (llama-3.1-8b-instant) — free, fast, no quota limits	ai.py
Web Framework	Flask 3.0 — Server-side sessions (flask-session) — Jinja2 templates	app.py
Database Layer	SQLite 3 — patients, doctors, screenings, chat messages	models (inline)
Communication Layer	Twilio WhatsApp Business API + ReportLab PDF generation	services.py
Frontend	Bootstrap 5.3 + Vanilla JS (AJAX chat + drag-drop upload)	static/

Why Groq instead of Claude/Gemini: Both Claude API and Gemini free tier hit quota limits during testing. Groq provides the same LLM

Bilingual Arabic/English + WhatsApp Delivery

🌐 Full Arabic Support

✓ **Chatbot responds in chosen language**

✓ **Reports generated NATIVELY in Arabic — not translated from English**

✓ RTL layout throughout Arabic UI

✓ WhatsApp messages in Arabic

✓ PDF report bilingual side-by-side

✓ Language toggle on every page

📱 WhatsApp Report Delivery

👁️ **DR Eye Platform — New Report**

👤 Patient: [Name] · Age: 45
🔬 DR Grade: 2 — Moderate
📊 AI Confidence: 86.5%
🚨 Urgency: Moderate — Refer within 3 months
📷 Retinal image: Included

📋 Clinical Summary (Doctor):
Moderate NPDR detected. Hard exudate clusters present in temporal quadrant. Anti-VEGF assessment recommended.

⏰ 2025-07-21 14:32
*_Sent by DR Eye Platform_*

Live Demo

The complete patient journey in real time

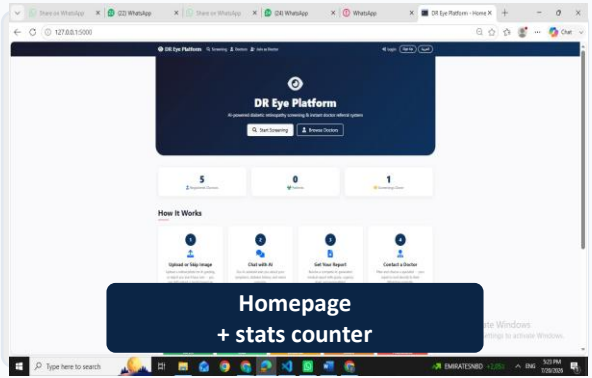
① Upload retinal image → instant AI grade + preprocessing images

② Answer 4 questions → medication photo upload

③ Generate bilingual medical report → PDF download

④ Filter doctors → Send report via WhatsApp → confirmation

Platform Screenshots



Homepage
+ stats counter

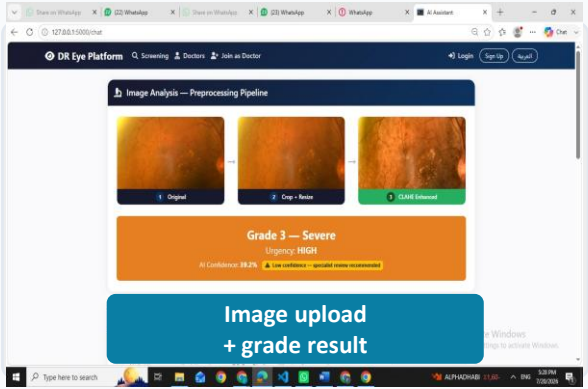
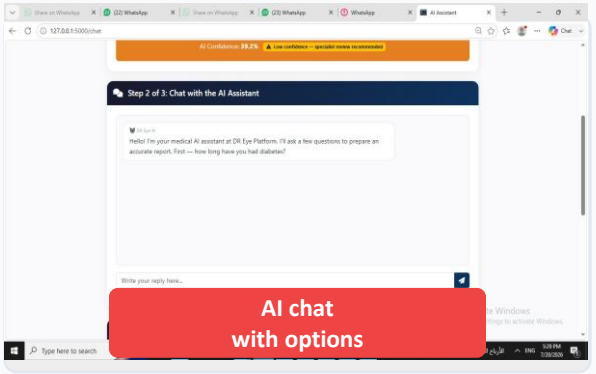
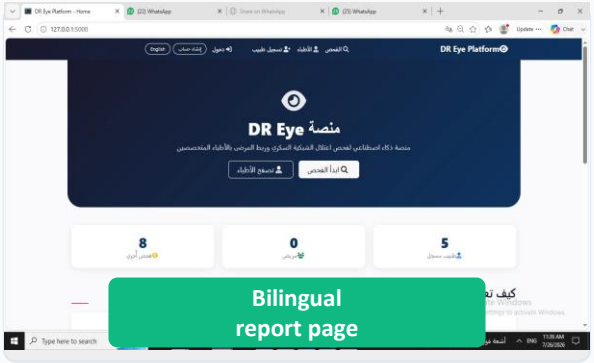


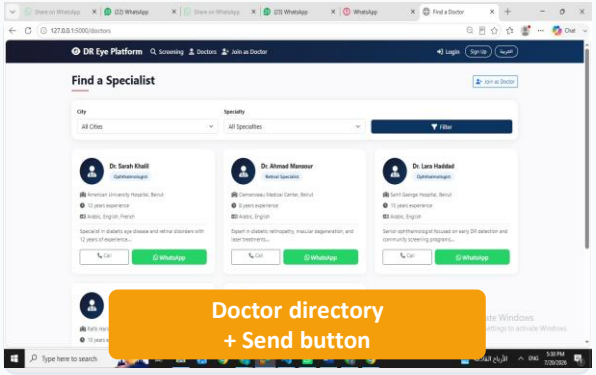
Image upload
+ grade result



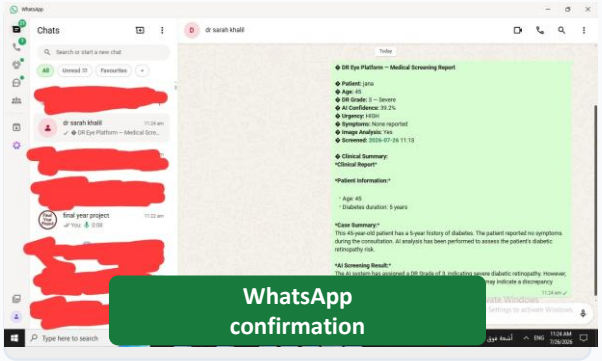
AI chat
with options



Bilingual
report page



Doctor directory
+ Send button



WhatsApp
confirmation

Speaker 3

Honest Limitations — Documented & Traceable

We document 8 limitations with full transparency. Each has a specific cause and a specific resolution path:

L1 **Dataset Size**
APTOS 2019 (3,662) vs planned EyePACS (88k). Colab multi-part archive incompatibility. APTOS enables valid benchmark comparison.

L2 **Distribution Shift**
Model validated on APTOS 2019 only (Indian clinics, 2019 cameras). Untested on Lebanese patients or different camera hardware (Arora 2024).

L3 **MC Dropout Incomplete**
Monte Carlo Dropout planned for uncertainty. Default dropout=0 in timm ConvNeXt; GPU quota exhausted before retraining. Softmax used instead.

L4 **Local Deployment**
Flask dev server + ngrok. No persistent URL, no HTTPS, no production security. Suitable for prototype demonstration only.

L5 **Twilio Sandbox**
WhatsApp delivery requires doctor opt-in to sandbox first. Production needs Meta-approved WhatsApp Business Account.

Future Work — 8 Prioritized Steps

P1	Train on full EyePACS (88k images) — resolve Colab archive via cloud compute	Highest
P2	Complete Monte Carlo Dropout — retrain 2 epochs, full ECE calibration	High
P3	Cross-dataset validation on IDRiD — measure distribution shift before any clinical use	Critical
P4	Deploy on Railway.app — PostgreSQL, HTTPS, 24/7 persistent URL	Medium
P5	Upgrade Twilio to Business Account — unrestricted WhatsApp delivery	Medium
P6	Lebanese clinical user study — ophthalmologist feedback on Grad-CAM + reports	High
P7	Appointment booking integration — complete care pathway to confirmed consultation	Medium
P8	Mobile app (React Native) — fundus camera smartphone adapter for rural access	Long-term

Thank You

DR Eye Platform — Lebanese University · Data Science 2025–2026

0.8888	82%	5	2 Lang	4-Step
QWK Score	Accuracy	Models	AR + EN	Workflow

Supervised by Dr. Abbass Rammal

Questions & Discussion