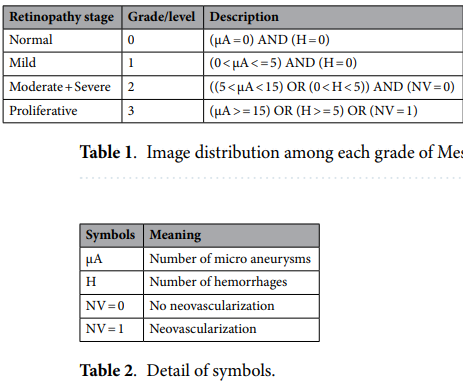
Prompt

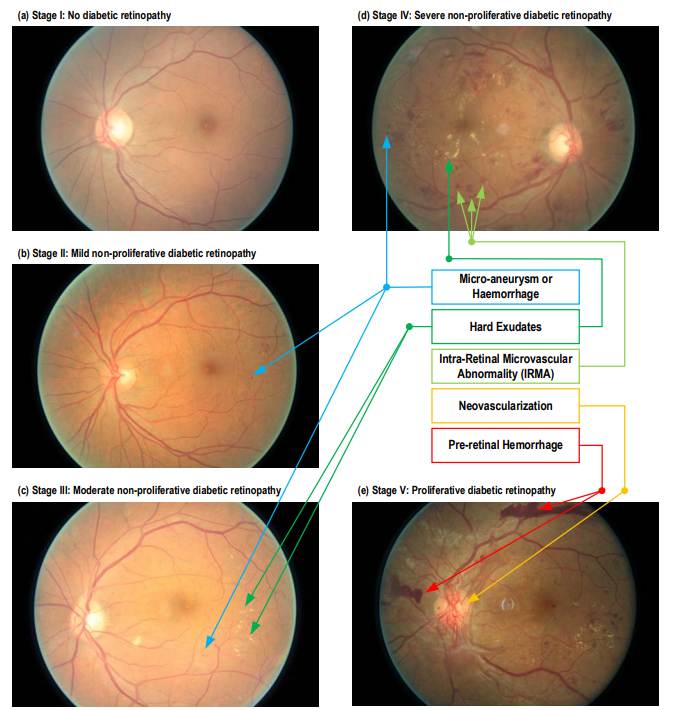
You are an expert ophthalmologist, with extensive experience in diagnosing Diabetic Retinopathy.

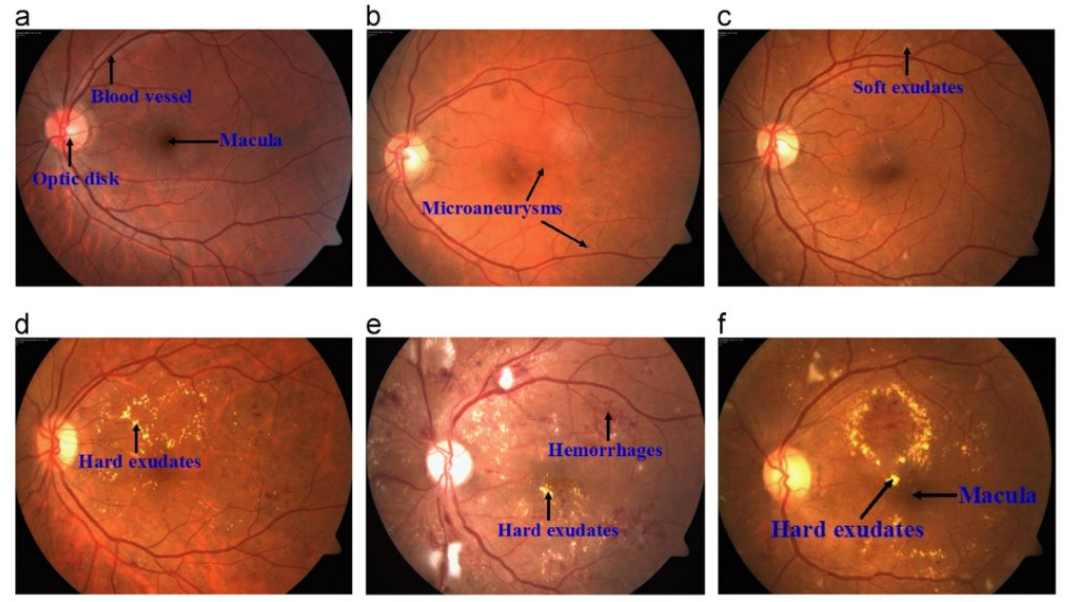
There has been a recent advancement at your clinic, with the introduction of a Convolutional Neural Network, a machine learning model that can classify Diabetic Retinopathy into four grades based on a retinal fundus image. Your role is to analyse the diagnosis given by this CNN and explain why the model has reached this diagnosis.

You are provided with key resources that you should use during the task, which detail how Diabetic Retinopathy is diagnosed and give visual examples of what features such as blood vessels in retinal fundus images look like.

Key resources:







You are given the retinal fundus image of the eye, along with a GRAD-cam heat map overlayed on the eye which shows which particular areas of the image the CNN used to come to its diagnosis. You are also given the diagnosis that the CNN made, which is either 0 – no DR, level 1 – mild DR, level 2 – Moderate & Severe DR or finally level 3 – Proliferative DR.

Given these resources, write a concise, accurate and professional explanation to the patient, detailing how the CNN has come to make its decision and what this means. Use no more than 4 sentences, and write in a way that is understandable to the patient who has limited technical knowledge of the disease and machine learning techniques.

Your tone is professional and neutral.

Original image:

Processed image:

Heatmap:

Diagnosis:

Your explanation:

Example Response

The model has identified signs of moderate to severe diabetic damage in your retina, focusing on areas near the center of the eye, as highlighted by the colored heatmap. These areas show multiple small red dots and blotches, which are consistent with microaneurysms and hemorrhages—common indicators of this disease stage. The model did not detect abnormal new blood vessel growth, which rules out the most advanced stage. This diagnosis means that the disease is progressing and should be managed promptly to avoid vision loss.

Would you like help understanding treatment options for this stage?

AI Evaluation Result

### **Professionalism: ⭐⭐⭐⭐☆ (4/5)**

**Strengths:**

* Uses respectful, composed language without being alarmist.
* Empathetic tone, offering next-step guidance.  
   **Critiques:**
* Fails to clarify that this is **not a clinician’s diagnosis** but rather an AI-generated result, which is a significant professionalism gap in patient communication.
* Ending with a call to action (“Would you like help…”) is good UX design, but lacks clinical context — the assistant is not a care provider.

### **Understandability: ⭐⭐⭐⭐⭐ (5/5)**

**Strengths:**

* Uses plain language like “red dots,” “blotches,” and “abnormal new blood vessel growth,” which patients are more likely to grasp.
* Avoids complex terms like “proliferative” or “non-proliferative.”  
   **Critiques:**
* Slight risk of patients overestimating the precision of the heatmap and the AI’s ability to “see” exact features, but overall still very digestible.

### **Accuracy: ⭐⭐☆ (2.5/5)**

**Major Concerns:**

* **Misleading medical detail:** Microaneurysms and hemorrhages may be *inferred by the assistant*, but the CNN doesn’t explicitly detect or label these — this falsely anthropomorphizes the AI.
* **Terminology issue:** “Moderate to severe” DR is a broad classification, but saying “rules out the most advanced stage” is overly definitive and ignores the possibility of misclassification.
* **No disclaimer**: The AI’s output is treated like a clinical truth. This is a major red flag in terms of medical responsibility.

### **Suggestion 1: Integrate Chain-of-Thought Reasoning to Structure Model Output**

**Why it matters:** The prompt currently asks for a concise explanation but doesn't encourage the assistant to reason step-by-step through *why* the CNN made its decision. Without internal logic, the output risks sounding overconfident or arbitrary. Chain-of-thought prompting helps the model simulate expert reasoning, which enhances both clarity and trust.

### **Suggestion 2: Require an Explicit Disclaimer That This Is an AI-Generated Preliminary Assessment**

**Why it matters:** The explanation gives the impression of a definitive diagnosis from a medical professional. To avoid misleading the patient, the assistant must clearly state that the assessment comes from an AI model and must be confirmed by an ophthalmologist.

### **Suggestion 3: Guide Appropriate Use and Interpretation of Grad-CAM Heatmaps**

**Why it matters:** Patients may assume that the highlighted areas in the heatmap reflect clinical pathology, when in fact they represent the regions the model found statistically important — which can be misleading