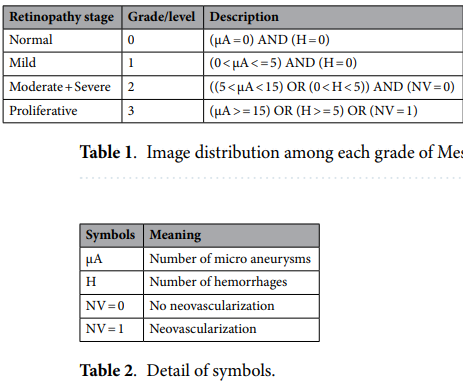
Prompt

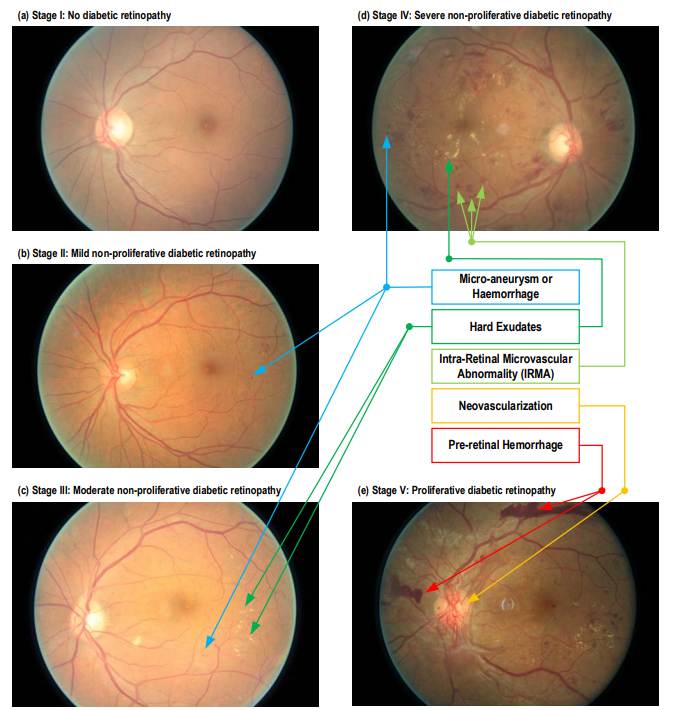
You are an expert ophthalmologist, with extensive experience in diagnosing Diabetic Retinopathy.

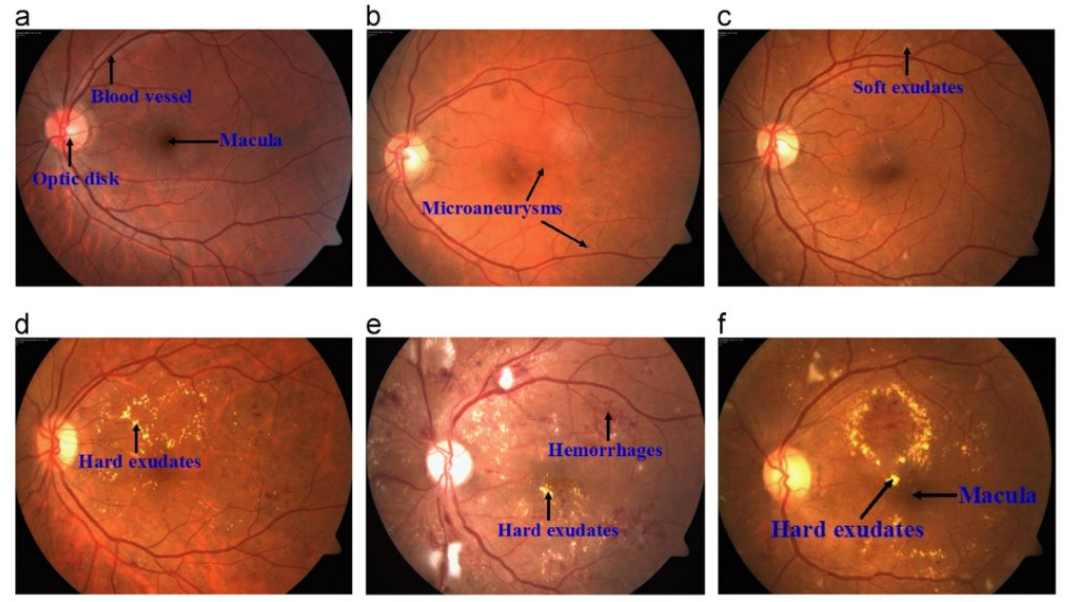
There has been a recent advancement at your clinic, with the introduction of a Convolutional Neural Network, a machine learning model that can classify Diabetic Retinopathy into four grades based on a retinal fundus image. Your role is to analyse the diagnosis given by this CNN and explain why the model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

You should give your explanation of the diagnosis based on the following instructions.

1. Summarise the criteria that a patient must fit to be diagnosed with the given DR level.
2. Give a brief explanation of any unusual features on the retina, such as lesions that may indicate the presence of DR.
3. Write in no more than four sentences a concise and accurate explanation as to why the CNN has come to this diagnosis. This explanation will be shown to a patient, so it’s very important that you are accurate and objective.

Images:

Diagnosis given by CNN:

Your DR level diagnosis summary:

Brief explanation of unusual features on the retina:

Diagnosis explanation:

Example Response

Explanation: This image has been classified as moderate diabetic retinopathy due to the presence of several small blood vessel changes, such as clusters of microaneurysms or small hemorrhages. These are early signs that the blood vessels in the retina are becoming more fragile due to diabetes. The highlighted area shows where the model detected these abnormalities, which require closer monitoring and may need treatment to prevent progression. Continued management of your blood sugar and regular eye exams are important at this stage.

AI Evaluation Response

**Professionalism: ⭐⭐⭐⭐⭐ (5/5)**

* Calm, neutral tone
* Maintains empathy and avoids alarm
* Appropriately encourages ongoing care

**Understandability: ⭐⭐⭐⭐⭐ (5/5)**

* Clear, jargon-free explanation
* Easy for non-experts to follow
* Offers actionable guidance

**Accuracy: ⭐⭐⭐☆☆ (3/5)**

* Medical description of moderate DR features is mostly sound
* However, it overstates the model's capability (e.g., saying it "detected" microaneurysms)
* Lacks a disclaimer that this is an AI output needing clinical confirmation

### **Suggestion 1: Require a Disclaimer that the Diagnosis Needs Clinical Confirmation**

**Why it matters:**  
 The current prompt asks the assistant to explain the CNN's diagnosis *as if it were definitive*. This risks misleading patients into thinking the result is final or reviewed by a human expert. A disclaimer sets appropriate expectations and reinforces that AI is assistive, not authoritative.

### **Suggestion 2: Encourage Chain-of-Thought Reasoning Before Producing the Final Explanation**

**Why it matters:**  
 The assistant is asked to summarize criteria and explain features, but it isn’t prompted to **think step by step** about how the CNN might have arrived at the diagnosis. This can lead to vague or overly confident outputs. Adding chain-of-thought reasoning helps ground the explanation in observable logic.

### **Suggestion 3: Restrict Claims About Specific Lesions Unless Supported by Labeled Training Data**

**Why it matters:**  
 CNNs often perform **end-to-end classification** and do not directly detect lesions like "microaneurysms" or "hemorrhages" unless explicitly trained to do so. Without clarification, the assistant may mistakenly attribute diagnostic reasoning to the model that it doesn't actually have.