

Medical Imaging and Computer Vision

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 10 of "Healthcare Analytics and AI: Building Systems That Actually Work"

The essentials: what to know, why it matters at the bedside, and how models fail.

One-sentence thesis. Medical imaging AI rarely fails because the model cannot classify the image; it fails because the clinic, the scanner, the workflow, and the patient do not look like the benchmark dataset.

PART 1: How Vision Models See

Convolution and CNNs

Start with the radiologist's actual problem. A nodule, an edge, a suspicious texture can sit in any corner of the film, and it means the same thing wherever it lands. A network that could only recognize a nodule in the top-left corner would be useless, because the next patient's nodule will be somewhere else. The question a vision model has to answer is therefore not just *what* is in the image but *where a pattern could appear*, anywhere at all. This is the problem convolution solves.

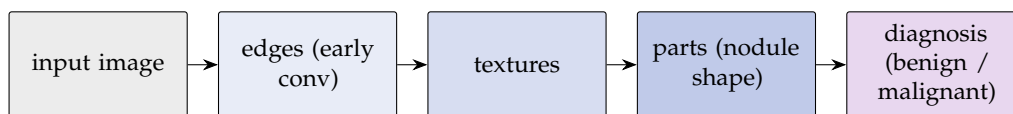
The intuition is a small stamp. Take a little filter a few pixels square that lights up when it meets one specific pattern, say a bright-to-dark edge, and slide it across every position in the image, recording how strongly the pattern matches at each spot. Because the *same* filter is reused at every location, an idea called weight-sharing, the network detects that pattern wherever it appears and does so with a handful of weights instead of one per pixel. Stack these layers and the features compound: early filters catch edges and textures, deeper ones assemble those into nodule shapes

and tissue architecture, and the final layer turns them into a diagnosis. That reuse is also why CNNs are translation-tolerant and efficient enough to train.

For readers who want the operation in symbols, a convolution is exactly that sliding multiply-and-sum of the image I against the filter K :

$$(I * K)(i, j) = \sum_m \sum_n I(i + m, j + n) K(m, n)$$

CNNs erased decades of handcrafted feature engineering, but a medical image is not a JPEG. A chest CT is 300-500 DICOM slices, each carrying scanner meta-data, and before any filter slides across it you must convert to scanner-independent Hounsfield units, apply the correct window (lung, mediastinal, or bone), and resample to uniform voxel spacing. **Watch out.** Skip windowing and the model trains on data no radiologist would recognize. If you do not understand DICOM metadata, you are not doing medical imaging AI, you are doing noise classification.



shallow → deep layers: from local pixels to diagnostic concepts

The CNN feature hierarchy. Early layers detect edges and textures; deeper layers compose them into parts and, finally, a diagnostic decision.

The Architectures That Matter

No single architecture wins every imaging task; the right one follows from the clinical question you are answering. Five define the field. **ResNet** added skip connections that let gradients flow through 50-plus layers

and remains the default classification backbone. **U-Net** is an encoder-decoder with skip connections built for **segmentation**, and its auto-configuring descendant nnU-Net wins most segmentation challenges. **DenseNet-121** powered CheXNet's radiologist-level

chest X-ray reads. **Vision Transformers (ViT)** drop convolution for self-attention, letting every patch of the image attend to every other at once and capturing global relationships a local filter would miss.

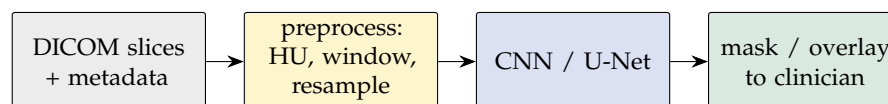
ANALOGY

A CNN builds understanding locally, brick by brick from small receptive fields. A ViT lets every patch of the image talk to every other patch at once, like a tumor board where everyone in the room compares notes simultaneously.

The task decides the tool: classification (“is there a nodule?”) favors ResNet-style backbones, while delineation (“outline the exact tumor”) requires U-Net-style segmentation. That global view carries an audit cost. **Watch out.** A 2026 study found Grad-CAM saliency maps degrade badly on ViTs versus CNNs because attention is non-local, so if your regulatory context demands faithful saliency explanations (Chapter 8), a ViT may explain itself worse than a ResNet would.



A normal PA chest radiograph. DenseNet-121 (CheXNet) reads images like this, but the model never sees an X-ray, only DICOM pixel values and header metadata it must map to findings. Image: Mikael Häggström, CC0.



skip this and you train on noise

The medical-imaging pipeline. The amber preprocessing step — Hounsfield-unit conversion, windowing, and resampling — is what turns raw DICOM into something a model, and a radiologist, would recognize.

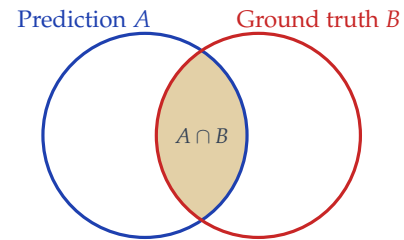
Two Different Tasks, Two Different Metrics

Two clinical questions need two different answers. “Is there a tumor?” is **classification**: assign one label to the whole image, scored with AUC, sensitivity, and specificity. “Where exactly is the tumor, pixel by pixel?” is **segmentation**: assign a label to every pixel and score the result by how well the predicted mask overlaps the truth. A model can ace one and fail the other, so the metric has to match the decision the output feeds.

To grade an outline you need a number for overlap, and two are standard. Dice and Intersection-over-Union (IoU) both compare the predicted mask A with the ground-truth mask B ; Dice weights the shared region more generously, and both run from 0 (no overlap) to 1 (a perfect trace).

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|} \quad \text{IoU} = \frac{|A \cap B|}{|A \cup B|}$$

Segmentation is what drives measurement: tumor volume, organ boundaries for radiation planning, calcium scoring, everywhere a rough bounding box is not enough. Classification tells you a fire exists in the building; segmentation draws the exact outline of the flames on the floor plan, and Dice and IoU grade how faithfully your outline traces the real one.



*Dice weights the shaded overlap $|A \cap B|$;
IoU divides it by the full union $|A \cup B|$.
Both run 0 (no overlap) to 1 (perfect).*

Dice and IoU measure how well a predicted mask A overlaps the ground-truth mask B . The shaded lens is the intersection both metrics reward.

Transfer Learning and Self-Supervised Pretraining

Medical datasets are small and expensively labeled, so almost no one trains a network from scratch. **Transfer learning** starts from weights learned on a large source dataset and fine-tunes them on your smaller medical target. The catch is that the source matters: ImageNet pretraining transfers reasonably to radiology but often fails on pathology, because the diagnostic grammar is different.

ANALOGY

Using ImageNet pretraining for pathology is like training a resident on vacation photos and then handing them biopsy slides. Both are images, but the grammar differs: in ImageNet, texture is secondary to shape, while in histopathology texture *is* the diagnosis.

The fix is to pretrain on medicine itself. **Self-supervised** methods learn from unlabeled medical images directly, and domain-specific foundation models, CheXFound on 1M-plus chest X-rays, EchoCare on 4.5M ultrasound images from 23 countries, Athena on 115M pathology patches, now beat ImageNet initialization on their own modalities. The era of “just use ImageNet weights for everything” is over, and for good reason: ImageNet-pretrained models overfit to confounders, famously learning that a ruler in a dermoscopy photo signals melanoma (because clini-

cians measure suspicious lesions) rather than learning the pathology itself. **Watch out.** Stain variability is a distribution-shift trap unique to pathology: slide color depends on staining protocol, reagent age, section thickness, and scanner, so a model trained at one lab can fail at another without stain normalization or color augmentation.

OOD Detection and Quality Gating

Once deployed, a model meets inputs unlike anything it trained on, out-of-distribution (OOD), and a network will confidently classify pure garbage. Many imaging systems therefore add a quality gate that rejects images below a threshold rather than guessing. But a filter that is too strict is a bouncer who turns away paying guests who belong inside: every rejected image is a patient who receives no result. Your quality threshold is therefore a clinical decision balancing diagnostic precision against population coverage, not a free technical parameter, and it should be set with the clinicians and patients who live with the consequences.

Watch out. Generative multimodal models raise the stakes, because they can hallucinate findings, calling “small bilateral pleural effusions” on a clear film. The MedBLINK benchmark found the best multimodal models hit only 65 percent on basic perception (image orientation, contrast) against 96.4 percent for humans, so their fluent diagnostic reasoning can rest on shaky perception.

The Thailand Diabetic Retinopathy Case

The clearest lesson in this chapter is a model that was excellent in the lab and failed in the field. Google’s diabetic retinopathy model scored **91.4 percent** sensitivity and **95.4 percent** specificity, on par with ophthalmologists. Deployed across eleven rural Thai clinics, its quality filter rejected **21 percent** of images, and patients told to return another day frequently did not.

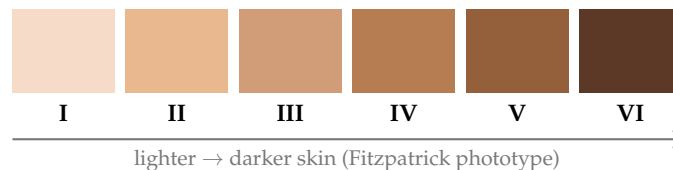
None of the causes were algorithmic. Bright exam rooms produced glare, clinics were not dilating pupils, some fundus cameras needed repair, and cloud processing depended on unreliable internet. The training data had come from dilated patients photographed on well-maintained cameras in controlled suites, none of which matched the field. The lab sensitivity number never saw any of this coming.

Lesson	What it means
Your quality filter is a clinical decision	Maximizing accuracy also maximizes patients who get no diagnosis
Train on the data you will actually see	Distribution shift showed up as a categorical failure to process 1 in 5 inputs
Workflow integration is not optional	The workflow <i>is</i> the deployment; the technology does not exist outside it

Subgroup Fairness: Skin Tone Bias

A headline accuracy number is a photograph taken from far enough away that the disparities blur out, and dermatology is where that blur hides real harm.

Dermatology training data is overwhelmingly light-skinned: one widely used dataset contains exactly one image of Fitzpatrick type VI (0.07 percent), and a 2024 review of 106,000 images found only 11 depicting dark skin. A model learns only what it is shown.



The Fitzpatrick skin-phototype scale (I–VI). The dataset above held a single Type VI image; a model learns only what it is shown. Original figure.

The consequence is measurable. On the Diverse Dermatology Images benchmark, model sensitivity fell from **0.69** on lighter skin to **0.23** on darker skin, a threefold gap. Because melanoma already kills Black patients at higher rates through later-stage diagnosis, a model that confidently calls a dark-skin lesion “benign” amplifies an existing disparity under the

cover of technological authority. **Watch out.** “The model works well on average” is not acceptable when the average hides a threefold gap that maps onto real health inequities. Evaluate every imaging model stratified by skin tone, age, and sex before deployment; if performance drops for a subgroup, fix the model or explicitly restrict its intended-use population.



Acral lentiginous melanoma on the sole, the subtype that predominates in darker skin and is diagnosed late. This is the presentation a model trained on light-skinned data is most likely to miss. Image: Xavier-Júnior et al., CC BY 4.0.

“FDA-Cleared” Is Not “FDA-Approved”

A clearance stamp is easy to over-read. Over 1,356 AI devices are authorized (77 percent radiology), but 97 percent came through the 510(k) pathway, which shows **substantial equivalence to a predicate device**, not independent proof of safety and effectiveness. De Novo and PMA demand more evidence but are rare for imaging AI. **Watch out.** The predicate chain drifts: device C clears against B, which cleared against A, until C barely resembles the device actually validated, and clearance only ever evaluated one dataset at one moment. Treat it as necessary but insufficient, and validate locally on your patients, your equipment, and your workflow.

Should You Build This?

Build it when the task is narrow, the labels are expert-grade, and the deployment environment resembles

the training distribution, and when you have budgeted for the unglamorous parts: domain-appropriate pretraining, quality thresholds tuned to real clinics, and performance stratified by every subgroup the model will actually see. **Think twice when** laboratory metrics are the whole pitch, the way Google’s retinopathy system brought 91.4 percent sensitivity to Thailand and then rejected 21 percent of field images, or when darker skin is nearly absent from the training data and “works well on average” conceals a threefold sensitivity gap. **Who bears the risk:** patients from underrepresented groups, whose lesions the model was never taught to see; frontline clinic staff, whose workflow absorbs every rejected image; and the screening program itself, measured by the patients who never come back.

Quick Self-Test

1. You need to measure a tumor’s volume for radiation planning. Is this classification or segmentation, and which metric (AUC or Dice) do you report?
2. Why does ImageNet pretraining transfer to radiology reasonably but often fail on pathology?
3. Google’s Thailand model had excellent lab sensitivity yet failed in the field. Name two environmental causes and the metric that hid the problem.
4. In one sentence, why is a quality-filter threshold a clinical decision rather than an engineering default?
5. A dermatology model reports 0.86 average AUC. What single analysis must you run before believing it is safe to deploy, and why?
6. What does 510(k) clearance actually establish, and why is the predicate chain a concern?

If You Remember One Thing

Lab performance is not field performance, and headline accuracy is not subgroup accuracy. The technical problem of classifying the image was solved years ago; the unsolved problem is everything around it, the miscalibrated camera, the bright room, the undilated patient, the darker skin absent from training, the workflow no one redesigned. That gap is where most medical imaging AI goes to die, and closing it is a design requirement, not an afterthought.

Richard J. Young, Ph.D. · richard@deepneuro.ai · Senior AI Research Scientist, UnitedHealth Group · UNLV Lee Business School · ORCID 0000-0002-1109-7552