Getting Started with AI in Dermatopathology

Path-iQ Getting Started Series · No. 4 of 20 · A practical primer for the practising pathologist

July 2026

1. Why dermatopathology is a natural target

Dermatopathology combines two properties that make it unusually attractive to algorithm developers: enormous specimen volume dominated by a small number of common diagnoses, and a notoriously poor interobserver reproducibility on the cases that matter most.

The melanocytic lesion is the defining problem. Concordance between experienced dermatopathologists on intermediate melanocytic proliferations — dysplastic naevi, Spitzoid lesions, melanoma in situ — is persistently mediocre, and the medicolegal consequences of error run in both directions. A tool that improves consistency here has obvious value, and unlike many subspecialties, dermatopathology has both the case volume and the diagnostic ambiguity to make that value measurable.

2. Where the evidence currently stands

Dermatopathology has moved through three technical generations quickly. First-generation convolutional networks handled binary tasks — basal cell carcinoma detection was demonstrated as far back as 2013. Hybrid CNN–transformer systems followed. The current generation is dominated by large domain-specific foundation models.

The most notable of these is PanDerm, a multimodal dermatology foundation model evaluated across 28 datasets and a broad task range spanning skin cancer diagnosis, phenotyping, risk stratification, inflammatory disease diagnosis, lesion segmentation, metastasis prediction and prognosis. It achieved state-of-the-art performance across tested tasks using less than 10% of the labelled data required by prior models. In real-world evaluation it outperformed dermatopathologists by 10.2% in identifying early-stage melanoma; with a human-in-the-loop configuration, dermatopathologist accuracy improved by 11%.

Read those two numbers together, because the second is the one that matters for practice. The standalone comparison is a research result. The assisted-performance improvement is the deployment case.

Commercially, PathAssist Derm launched in early 2025 as a general dermatopathology triage and classification tool, and Aisencia has been deployed in dermatopathology clinics internationally. Peer-reviewed independent performance data on both remains limited — a gap worth pressing vendors on directly.

3. The task map for skin

Melanocytic lesion triage. Ranking cases by probability of malignancy so that the highest-risk slides reach the dermatopathologist first, and so that referred second-opinion cases are prioritised.

Melanoma subtyping. Superficial spreading, nodular and acral lentiginous subtypes differ in mutation profile, spread pattern and outcome. Deep learning has been applied to automate subtyping, though rare subtype under-representation is a persistent limitation in the published work.

Basal and squamous cell carcinoma detection and margin assessment. High-volume, morphologically stereotyped, and the most immediately practical application. Automated margin assessment on Mohs frozen sections is an active area with a clear workflow benefit.

Breslow thickness and mitotic rate. Quantitative measurements with direct staging implications and demonstrable interobserver variability. Automated mitotic figure detection is among the better-validated quantification tasks in pathology generally.

Inflammatory dermatoses. The largest unmet need and the least developed. Pattern-based diagnosis of interface, spongiotic, psoriasiform and granulomatous reactions requires integrating architecture, cell composition and clinical context. Model coverage here is thin, and modules for inflammatory disease are only now emerging.

4. Your first 90 days

Weeks 1–2: separate the two clinical streams. Dermatology AI and dermatopathology AI are different fields with different regulatory status and different evidence. Clinical dermoscopy algorithms have a far larger literature; do not import their performance claims into your histology conversation.

Weeks 3–6: choose margin assessment or melanocytic triage. Margin assessment offers a fast, measurable workflow win with low diagnostic risk. Melanocytic triage offers higher clinical value but demands a more careful governance conversation. Pick one, not both.

Weeks 7–12: build a discordance corpus. Dermatopathology departments usually hold rich second-opinion and consultation archives. These cases — where experts disagreed — are the most informative validation material you own, and far more discriminating than a consecutive series dominated by straightforward BCCs.

5. Validation specifics for skin

Section orientation and level matter enormously. Shave, punch and excision specimens present entirely different architecture. Confirm which specimen types the model was trained on; a model built on excisions will misbehave on tangentially sectioned shaves.

Special stains and immunohistochemistry are integral to dermatopathology practice — SOX10, Melan-A, p16, PRAME — and most H&E-based models ignore them entirely. If your diagnostic pathway depends on IHC, an H&E-only algorithm is answering a different question than the one you ask.

Pigmentation is a genuine confounder in both directions. Heavily pigmented lesions obscure nuclear detail; models trained on predominantly light-skinned populations may perform differently across skin phototypes. Demand demographic composition data on the training set, and validate on your own population.

6. Failure modes to anticipate

Spitzoid and desmoplastic lesions. The categories where human concordance is worst are also where training labels are least reliable. A model trained on noisy labels cannot exceed the noise.

Regression and naevoid melanoma. Melanomas that lack the architectural cues models rely on are systematically under-detected.

Re-excision and scar specimens. Post-procedure change mimics malignancy and is under-represented in training data.

Over-calling dysplasia. Sensitivity-optimised commercial tools flag liberally by design. In a high-volume practice this can generate a substantial false-positive burden and, if it drives IHC ordering, a real cost increase. Measure this during shadow deployment.

7. What to watch next

Two directions matter. First, genuine multimodal integration — linking the clinical image, the dermoscopic image and the histology of the same lesion. Dermatology is one of the few disciplines where all three exist routinely, and PanDerm's design points this way. Second, inflammatory dermatoses: whichever group solves pattern-based inflammatory diagnosis will address the larger share of daily dermatopathology workload, and will do so in a domain where clinical correlation is unavoidable — a hard problem for any image-only system.

Vendor due-diligence checklist

Use this before any procurement conversation. The first seven questions apply to every AI tool in pathology; the last three are specific to this subspecialty. A vendor unable or unwilling to answer them in writing has told you something important.

Applies to any pathology AI tool

What is the intended use statement, and what is explicitly excluded from it?

What is the regulatory status in my jurisdiction — cleared, CE-marked, research use only?

Can I see performance on an independent external cohort, not only internal validation?

What is the prevalence of the target condition in the validation cohort versus my practice?

What is the failure mode when the model is uncertain — does it abstain, or always answer?

What monitoring does the vendor provide for performance drift after deployment?

Who is liable if an algorithmic output contributes to a diagnostic error?

Specific to this subspecialty

Was this trained on shave, punch or excision specimens, and is performance reported separately?

What is the skin phototype distribution of the training population?

Does it cover inflammatory dermatoses, or melanocytic and keratinocytic lesions only?

Ask for answers in writing, and keep them. When the tool underperforms in your laboratory — and at some point it will — the gap between what was claimed and what was validated is where the conversation starts.

8. Key references

#

Source

Note

1

Yan et al. (PanDerm)

28 datasets; +10.2% vs dermatopathologists on early melanoma; +11% with human-in-loop

2

Deep Learning Image Processing Models in Dermatopathology, *Diagnostics*, 2025;15:2517

CNN to foundation model evolution; PathAssist Derm, Aisencia

3

Translating Features to Findings: DL for Melanoma Subtype Prediction, *Dermatopathology*, 2025;12(4):42

Subtyping; dataset imbalance limitations

4

AI in Dermatopathology: An Update, *Diagnostics*, 2026;16:1702

Current literature review

5

Deep Learning for Skin Melanocytic Tumors in WSI: A Systematic Review

28 studies, four research trends

Glossary of terms used in this primer

Whole-slide image (WSI) — a digitised glass slide, typically gigapixel in scale, produced by a scanner and viewed on screen rather than under a microscope.

Foundation model — a large model pretrained on vast unlabelled image data to learn general-purpose visual features, then adapted to specific tasks with comparatively little labelled data.

Weakly supervised learning — training where only a slide-level label is available (for example “contains cancer”) rather than region-by-region annotation. Most clinical pathology AI is built this way, because exhaustive annotation is unaffordable.

AUROC — area under the receiver operating characteristic curve, a single figure summarising discrimination across all thresholds. 0.5 is chance, 1.0 is perfect. It says nothing about calibration or about performance at the threshold you would actually use.

Calibration — whether a model’s stated probability matches observed frequency. A well-calibrated model that says “80% confident” is right about 80% of the time. Calibration matters more than accuracy when a clinician acts on the number.

Abstention — the model declining to answer when confidence falls below a threshold, rather than returning its best guess. Essential wherever rare entities are in play.

Batch effect — systematic differences arising from scanner, stain lot, laboratory or site rather than from biology. The leading cause of models performing worse elsewhere than where they were built.

Automation bias — the human tendency to defer to a system that is usually correct, and so to stop checking independently. A property of the workflow, not of the algorithm.

Predetermined Change Control Plan (PCCP) — a regulator-agreed specification of changes a manufacturer may make to an authorised device without a new submission. Its existence acknowledges that these systems are not static.

Shadow deployment — running an algorithm in parallel with routine practice, without acting on its output, to characterise its behaviour on your own material before it touches a live workflow.

Path-iQ — Global Pathology AI Review. Prepared for practising pathologists beginning AI adoption.