Getting Started with AI in Neuropathology

Path-iQ Getting Started Series · No. 5 of 20 · A practical primer for the practising pathologist

July 2026

1. The subspecialty where machine learning already rewrote the classification

Neuropathology occupies a unique position: it is the only pathology subspecialty in which a machine learning classifier has been formally incorporated into the World Health Organization's tumour classification and into routine diagnostic practice worldwide. That classifier is not an imaging model. It is a random forest applied to DNA methylation data.

The lineage begins with Capper and colleagues in *Nature* in 2018, who trained a methylation-based classifier on a reference set of 2,801 samples across 91 CNS tumour classes and released it as a free online tool. The consequences were immediate: diagnoses that had rested on morphology and a handful of immunostains acquired an objective molecular anchor, and entities that morphology could not separate were shown to be distinct.

The current version tells you how far this has travelled. The Heidelberg CNS Tumour Methylation Classifier v12.8, published in *Cancer Cell* in December 2025, was trained on 7,495 methylation profiles and expanded recognised entities from 91 classes to 184 subclasses. Validated by five-fold nested cross-validation, it achieved 95% subclass-level accuracy with a Brier score of 0.028 — indicating well-calibrated probability estimates, which matters more for clinical use than raw accuracy. By the October 2024 data freeze, over 160,000 profiles had been analysed on the platform worldwide.

For a pathologist starting out, this is the single most important lesson available anywhere in the discipline: the highest-impact AI in pathology to date operates on molecular data, not on pixels.

2. The intraoperative revolution

The second major thread is speed. Methylation array profiling takes days. Nanopore sequencing produces sparse methylation data in minutes, and machine learning can classify from sparse data.

Sturgeon, published in *Nature* in 2023, is a patient-agnostic transfer-learned neural network designed for exactly this. It delivered an accurate diagnosis within 40 minutes of starting sequencing in 45 of 50 retrospectively sequenced samples, appropriately abstaining where confidence was insufficient. MethyLYZR, in *Nature Medicine*, reduced this further, returning clinically relevant classification within 15 minutes. A prospective multicentre validation in *Nature Medicine* in 2025 described a workflow delivering real-time methylation classification and copy-number information within a 30-minute intraoperative window, followed by comprehensive molecular profiling within 24 hours.

The clinical implication is substantial. A neurosurgeon's decision about extent of resection currently rests on frozen section and smear — informative, but frequently non-specific. Intraoperative molecular classification changes the information available at the moment the decision is made.

3. Where imaging AI fits

Image-based models in neuropathology have a narrower but real role.

Frozen section and smear support. Rapid intraoperative distinction between tumour, gliosis and normal brain; identification of infiltrating margin.

Grading and mitotic quantification. Ki-67 indices and mitotic counts in meningioma and glioma carry grading weight and are poorly reproducible manually.

Integrated risk models. Meningioma is the best example of morphology and methylation being combined formally — integrated molecular-morphologic classification has been validated retrospectively and prospectively and implemented clinically.

Neurodegenerative disease. Automated quantification of tau, amyloid, alpha-synuclein and TDP-43 burden across large brain-bank cohorts. This is largely a research application, but it is where digital neuropathology delivers most reliably today.

A domain-specific foundation model, NeuroFM, has been developed and benchmarked against general pathology encoders including UNI, UNI2, Virchow, Virchow2 and Prov-GigaPath on brain-specific tasks — a useful signal that generic pathology models under-serve neural tissue.

4. Your first 90 days

Weeks 1–4: establish access to methylation classification, if you do not have it. For most neuropathologists this is a higher-value step than any imaging pilot. Determine your referral pathway, turnaround time, and how classifier output is documented in your reports. If you are already sending cases, audit how often the classifier changed the diagnosis — that is your local evidence base.

Weeks 5–8: audit your intraoperative accuracy. Compare frozen section and smear diagnoses against final integrated diagnoses over the past two years. This baseline determines whether rapid molecular classification would change your practice, and by how much.

Weeks 9–12: if pursuing imaging AI, start with quantification. Ki-67 and mitotic counting in meningioma provide a measurable, low-risk target with a known reproducibility problem.

5. Validation specifics

Calibration matters more than accuracy in this subspecialty. A methylation classifier that returns a low-confidence score is giving you clinically useful information — the case does not match the reference classes well. Understand your platform’s confidence thresholds and never treat a sub-threshold call as a diagnosis. The Brier score reported for v12.8 exists precisely because calibration was treated as a primary endpoint.

Tumour purity is a hard constraint. Low-purity specimens, extensively necrotic tumours and small stereotactic biopsies dilute the methylation signal. Work is under way on classifiers robust to reduced feature sets, but the practical rule stands: sample adequacy determines classifier reliability.

Reference set coverage defines the boundary. A classifier can only assign to classes it was trained on. Rare, novel or paediatric-specific entities may not be represented, and the classifier will assign the closest available class rather than declare novelty — unless calibration is respected.

6. Failure modes to anticipate

Over-reading a low-confidence score. The commonest error. The score is a probability, not a diagnosis.

Classifier–morphology discordance. When these conflict, the answer is not automatically molecular. Discordance should trigger review of specimen identity, purity and adequacy before either result is discarded.

Version drift. A case classified under v11 may classify differently under v12.8, because the class structure changed. Record the classifier version in the report. This is a genuine medicolegal exposure and is frequently overlooked.

Paediatric under-representation. Paediatric CNS tumour diversity is high and reference sets have historically been adult-weighted, although the MNP 2.0 prospective cohort of over 1,200 newly diagnosed paediatric patients has improved this substantially.

7. What to watch next

Cerebrospinal fluid cell-free DNA methylation classification for tumours not amenable to biopsy is progressing rapidly and would extend molecular diagnosis to patients who currently receive none. Expect also further compression of the intraoperative window, and extension of methylation-based classification into sarcoma, sinonasal and other domains — the method has already proven generalisable well beyond the CNS.

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

Which classifier version is in use, and how is the version recorded in the report?

What is the calibrated confidence threshold, and what is the abstention behaviour below it?

How are paediatric and rare entities represented in the reference set?

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

Sill, Sturm, Capper, Sahm et al., *Cancer Cell*, 2025

Heidelberg classifier v12.8; 7,495 profiles; 184 subclasses; 95% accuracy

2

Capper et al., *Nature*, 2018;555

Original methylation classifier, 91 classes

3

Vermeulen et al., *Nature*, 2023;622:842–849

Sturgeon; diagnosis in 40 min in 45/50

4

MethyLYZR, *Nature Medicine*, 2025

Classification within 15 min

5

Prospective multicentre nanopore validation, *Nature Medicine*, 2025

30-min intraoperative window

6

Sturm et al., *Nature Medicine*, 2023;29:917–926

Multiomic paediatric neuropathology

7

Integrated molecular-morphologic meningioma classification, *J Clin Oncol*, 2021;39:3839–3852

Combined risk model

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.