Online Resource 2: Model card and dataset datasheet

Article: When Does Supervision Beat Reconstruction? Magnification Shift and the Robustness of AI Quality Control in Digital Pathology

Journal: Journal of Medical Systems

Authors: Konstantinos Moutselos, Ilias Maglogiannis

Corresponding author: Konstantinos Moutselos, Department of Digital Systems, University of Piraeus, Piraeus, Greece; kmouts@unipi.gr

Reference numbers in square brackets refer to the reference list of the article. Sections, tables and figures without the prefix S are those of the article.

Part A: Model card

A1 Models

Model	Type	Input	Output	Labels used in training
probe_all	Standardized L2 logistic regression on UNI2-h class tokens (1,536-d)	256-pixel tokens at ≈1.0 µm/px	Token probability of "artifact"; slide mask after smoothing, τ = 0.5 and morphology	Yes (AIRAQc, four types)
probe_aug	As probe_all, trained with an added ×0.5/×2-resampled copy of every token	as above	as above	Yes
K2all	PixCell-1024 + LoRA (rank 16), trained on 2,969 artifact-free patches	1,024-pixel patches at ≈1.0 µm/px	Per-cell denoising error; slide mask by rule R0	No
H2all	K2all recipe + anchor and error-space hinge (λ = 0.1)	as above	as above	Yes (artifact positions)

- Software and weights: PixCell-1024; UNI2-h revision d517a8dd; Phikon-v2 revision 2ae989a9 (second-encoder variant only).
- Model weights are subject to the licences of the base models. The LoRA adapters and the probe parameters are in the Zenodo deposit (https://doi.org/10.5281/zenodo.23448122).

A2 Intended use

- **Purpose:** research tool for pixel-level detection of four artifact types (out-of-focus, tissue fold, pen marking, air bubble) in H&E whole-slide images, as a quality-control step before human review or computational analysis.
- **Operating condition:** slides read at ≈1.0 µm per pixel (10×), with the physical resolution verified (Section 2.4 of the article).
- **Not intended for:**
  - clinical diagnosis or clinical decision-making;

- stains other than H&E;
- artifact types outside the four labelled ones (the reconstruction detectors flag them, but this was not validated);
- use without the monitoring described in A5.

A3 Performance across settings (summary)

Setting	probe_all	Reconstruction (H2all or per-fold H2)
Cross-validation (16 slides)	mean sensitivity 0.787; clean false positives 0 of 3,446,579 cells	0.443; clean FP fraction 0.0086
Development benchmark (24 slides)	F1 0.816; 2,724 clean FP cells	H2 > K2 by +0.06 F1 (both seed pairs)
SlideGuard, blind, non-TCGA (28 slides)	precision 0.291	precision 0.102; artifact Dice not different
SlideInspect clean controls (5× source)	clean FP fraction 0.21	higher out-of-focus sensitivity at matched FP (+0.23 to +0.28)
Resampled benchmark (Δ-B)	clean FP ×227; F1 0.82 → 0.49	(input not resampled)
Fırat, original / resampled (Δ-F)	clean FP 0.037 → 0.114	0.036 → 0.017
Calibration (benchmark)	ECE 0.051 → 0.296 under Δ-B	not defined (scores are not probabilities)

- probe_aug under Δ-B: clean FP 12,596 cells, F1 0.80, ECE 0.033. At the original read: F1 −0.037 and clean FP ×4.6 relative to probe_all.
- **Subgroups:**
  - by organ (Fırat): the resampling increase concentrated in kidney and small intestine;
  - by cohort (SlideGuard): the probe had the highest precision and lowest recall in every cohort;
  - mouse skin: the reconstruction detectors fell below the published range.

A4 Known limitations

- **Training data:** 16 TCGA slides. The benchmark's effective size is about 6 slides for annotated pixels. Two training slides entered training at a wrong read (≈2.0 µm/px).
- **Magnification:** probe performance and calibration degrade sharply when tissue reaches the model at another scale, and the size depends on the encoder. Scale augmentation was validated only against the resampling operator it was trained on.
- **Label scope:** marker ink outside the tissue-only annotations is not detected by the probe unless labelled.
- **Hinge:** H2 loses out-of-focus sensitivity in a few tissues where it reconstructs blur as normal.
- **Combination rules:** rules combining probe and reconstruction inherit the probe's failure under shift.

A5 Monitoring considerations

- **Verify the read:** log the physical resolution of every pyramid level read and reject reads outside 0.9–1.1 µm per pixel unless resampled explicitly.
- **Probe flag rate on clean tissue (M1):**
  - in development, 0.00026 of clean tissue at τ 0.5;

  - after resampling, 0.059;
  - AUROC 0.936 (SlideInspect) and 0.894 (resampled Fırat) against the development slides.

  A sustained rise on slides expected to be clean is a shift alarm. Embedding distance (M3) rose under every new source, including those where the probe worked, and is not a failure signal on its own.

- **Probe–reconstruction disagreement (M2):** a second label-free alarm with similar AUROCs.
- **Probabilities:** do not interpret probe probabilities at face value after a change of scanner, protocol or magnification until calibration has been rechecked.

A6 Ethical and licensing notes

- Only public or controlled-access de-identified data were used (Part B).
- The base models carry non-commercial or restricted licences: UNI2-h under its model licence; Phikon-v2 non-commercial. Derived files are released under CC BY-NC 4.0, except Phikon-derived files (probe parameters and evaluation outputs), which carry the Owkin non-commercial licence as its ShareAlike term requires; embeddings are not released.

Part B: Dataset datasheet

Dataset	Source and terms	Content used	Labels	Magnification route	Role
TCGA + AIRAQc	TCGA diagnostic slides (GDC); AIRAQc annotations [14]	Of 50 released slides (50 patients): 8 excluded (no usable patches or too little tissue), 2 held out; 16 training slides (4 folds) and 24 benchmark slides (11 clean, 13 artifact)	Pixel masks, four types; tissue only	Pyramid level near 1.0 µm/px; four slides re-read at 10× (Section 2.4)	Training, cross-validation, benchmark
SlideGuard [5]	Benchmark of the MICCAI 2026 paper; annotations held by the organizers; scored via Codabench 16457	53 WSIs: 25 TCGA, 10 DHMC lung, 6 lymphatic (Polish institute), 12 mouse skin	Hidden; the server returns per-slide Dice, precision, recall and artifact-free Dice	Guard: 0.91–1.01 µm/px, no upsampling	Blind external evaluation
Dartmouth Lung Cancer Histology Dataset [17]	Research-use agreement; written approval to submit predictions, report scores and show ≤ 3 slides	The 10 DHMC slides of SlideGuard	as SlideGuard	as SlideGuard	as SlideGuard
SlideInspect [16]	Public test set; images reproduced with the authors' permission	289 H&E cases, 55 clean controls	Pixel masks (out-of-focus, fold, mounting)	Native 5×, upsampled ×2 (bilinear)	External, shifted
Fırat [18, 19]	Zenodo, CC BY 4.0	2,127 images from	Binary fold masks	0.49 µm/px,	External, matching

Dataset	Source and terms	Content used	Labels	Magnification route	Role
	(Version 2, corrected slide mapping)	302 slides; 899 clean, 1,228 fold		downsampled ×2 (box)	magnification; resampling experiment

Known issues recorded during the study:

- **AIRAQc annotations:**
  - pooled metrics use every annotated pixel;
  - 0.30% of scored annotated cells are anti-aliased edges without a type colour;
  - 0.07% (one slide) carry a blue no-tissue label;
  - marker ink off tissue is never annotated.
- **SlideGuard:** the cohort of non-TCGA files was assigned from file names, group sizes and scanner metadata. The record describes the DHMC slides as breast; the files are lung adenocarcinoma.
- **Fırat:** two fold images have empty masks and were excluded from fold endpoints. A burned-in scale bar was masked for all detectors.
- **SlideInspect:** the native magnification differs from training, so results combine magnification with stain, scanner and organ differences.