Ensembled Foundation-Model Pipeline for Reasoning-Guided Pathology Report Generation: A REG^2 Challenge Submission

Mikołaj Olesiński

Wrocław University of Science and Technology, Wrocław, Poland

Abstract. We describe our submission to the REG^2 (Pathologist Reasoning-Guided Report Generation) Challenge, which requires a step-by-step diagnostic chain-of-thought (CoT) and structured pathology report from a gigapixel whole-slide image (WSI, Interface 1, Metric A), together with free-text answers about a region of interest that correctly reject non-informative background (Interface 0, Metric B). Our pipeline encodes stratified, artifact-filtered patches with a frozen foundation model, aggregates them with a five-fold ensemble of attention-based multiple-instance-learning heads, selects a diagnostic-template graph with a second ensembled classifier, and generates answers and the final report with a LoRA-adapted medical vision-language model, optionally conditioned on retrieved similar training cases. Every trainable component is validated with five-fold cross-validation rather than trained on the full corpus, given a single, non-revocable final submission. We report our platform-measured Test Phase 1 result (Overall Score 0.7504, obtained through a one-time reopened submission that does not count toward the official leaderboard or ranking, granted after the phase's regular window had closed), structural reasoning metrics from clean cross-validation, and a fine-tuning/retrieval ablation on a held-out set, explicit about which numbers come from the platform itself, which are local proxies, and where known evaluation leakage inflates a number. Our official, final Test Phase 2 result (Overall Score 0.6775) has since been released and is reported alongside these figures.

Keywords: Computational Pathology · Whole-Slide Images · Chain-of-Thought Reasoning · Report Generation · Multiple Instance Learning · Retrieval-Augmented Generation · Foundation Models

1 Introduction

Pathology examination is the diagnostic gold standard in oncology, yet requires a pathologist to visually synthesize a gigapixel whole-slide image while identifying clinically significant, often small, morphological features – a slow, cognitively demanding process. Automated report generation from WSIs has recently become feasible with vision-language foundation models, but prior benchmarks,

including the previous edition of this challenge, scored only the final report text with lexical and shallow semantic metrics that capture neither *how* a diagnosis was reached nor whether intermediate reasoning steps are individually correct.

The REG^2 challenge addresses this by pairing each WSI with a structured chain-of-thought (CoT) reasoning trace – question/answer/next-question triples emulating how pathologists narrow down a diagnosis – in addition to the final CAP-style report. Submissions are graded on two interfaces: *Workflow Reasoning* (Interface 1, WSI $\rightarrow$ full CoT trace ending in a report, Metric A, 70%) and *Visual Grounding* (Interface 0, an ROI crop and question $\rightarrow$ a free-text answer that must correctly reject non-informative background, Metric B, 30%). The corpus spans $\sim$11,220 training cases across seven organs, five countries, and four scanner models; Test Phase 2 (70 cases) is the final, single-submission evaluation.

Two constraints of this task setting shaped nearly every design decision below. **The platform allows exactly one non-revocable final submission**, with no opportunity to observe a failure on the true test set and correct it; and **the deployment container has no internet access and is invoked once per case with every model loaded from scratch**, ruling out any component that depends on a live download or persistent warm state. Both pushed us toward architectures that trade a little peak performance for (a) some form of held-out local validation even for the final deployed model, and (b) predictable, bounded resource use.

Section 2 describes the pipeline outlined above (Fig. 1) and each component, including the patch-selection/artifact-filtering procedure (Fig. 2), which required substantially more iteration than expected. Section 3 reports results, distinguishing platform-measured outcomes, local offline validation, and local proxy measurements. Section 4 discusses remaining weaknesses and what we would do differently with more time.

2 Methods

2.1 Data Acquisition and Preprocessing

Training WSIs ($\sim$11,220 cases) and the paired CoT/report corpus were obtained from the organizers' CDN. A multi-connection downloader cut average per-file download time by roughly an order of magnitude over single-connection, making encoding GPU- rather than I/O-bound. WSIs are deleted immediately after encoding, since raw pixels are not needed once patch embeddings are computed.

Patch selection. A WSI is several gigapixels and cannot be encoded in full. Patch selection depends on whether the TIFF exposes a low-resolution pyramid overview level: **41/42 (97.6%) of a random 42-WSI sample did not** (Fig. 2c), so blind random (y, x) sampling is the dominant path in practice, not a rare fallback. When an overview level is available, we instead use a stratified approach: tissue regions are located from the pyramid overview, a coarse 8×8 grid (in downsampled overview-pixel coordinates) is imposed over the tissue mask,

and candidate 224×224 px full-resolution patches are drawn evenly from every non-empty grid cell. Under either path, roughly $3\times$ the target count (up to 512 patches/WSI) is read in parallel and ranked by pixel-intensity variance, an inexpensive structural-complexity proxy, keeping the top-scoring patches.

On the worst observed case, blind random sampling recovered only 30/512 tissue patches; widening the candidate pool 1,024→8,192 and adding a cheap 64×64-px pre-filter before full-resolution reads recovered 236/512 on the same case ($8\times$), at ∼12–16s/WSI extra cost.

Artifact filtering. Ranking candidates by pixel-intensity variance alone systematically favored scanner artifacts (ink, tissue folds, blur) over genuine tissue, since both produce high local contrast. On a 56-WSI baseline (8/organ), **29.9%** of top-ranked candidates per WSI were artifacts, varying by organ (breast worst at 46.3%, prostate best at 14.5%; Fig. 2a), with 34% of WSIs exceeding 30% contamination. We addressed this with a three-signal rejection rule (saturation, Laplacian-variance texture, hue) reached after seven iterations, each closing one false positive/negative found by visual inspection of held-out crops (Fig. 2b): saturation alone either rejects pale genuine tissue or lets smooth non-tissue gradients through, depending on threshold; texture and hue each cover a gap the other misses. On four known-hard validation WSIs, the final rule keeps 92–97% of candidates as tissue typically, and recovers a case that previously kept only 1/30 (3.3%) before the pyramid-fallback fix above. This filter was re-evaluated on a fresh 56-WSI/8-per-organ sample (a seeded resample, since the original EXP-019 WSIs were not retained locally): the rejection rule discards a mean 2.9% of raw candidates (organ range 0.2–6.2%), and a visual spot-check of kept top-ranked patches across all 7 organs found 18/21 crops genuine tissue, with residual contamination concentrated in one WSI showing a tissue-fold pattern the filter's saturation/texture/hue signals did not catch – consistent with folds being a morphological rather than color/texture artifact, a concrete direction for a future iteration.

Stain normalization. Data originates from five countries and four scanner models, introducing H&E color variation unrelated to tissue morphology. Stain normalization is applied only to patches retained after artifact filtering (above), not before – the filter's thresholds are tuned on raw, un-normalized color. We apply GPU-based Macenko stain normalization [5], fitting one reference from the first 5 WSIs processed: candidate patches are pooled across these slides and the single patch whose pixel-intensity variance is closest to the pool's median is used as the fit target, rather than an arbitrary single WSI, so the reference is not sensitive to one unusual slide. The fitted reference is serialized and reused at inference to avoid a train/inference color mismatch.

2.2 Slide-Level Feature Encoder

We use H-Optimus-1 (Bioptimus, ViT-Giant/14) as a frozen patch encoder, run entirely offline from local weights, since the deployment environment has no in-

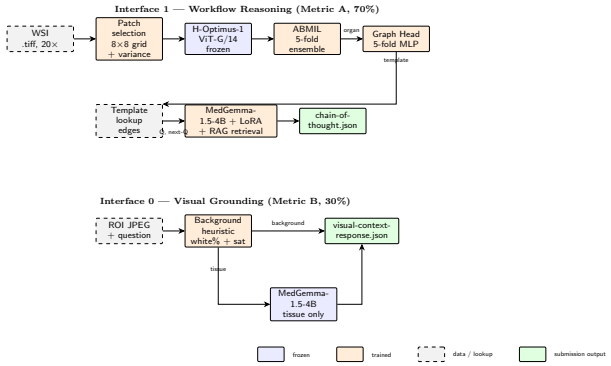

Fig. 1. Both submitted interfaces. Interface 1 (top): WSI → full chain-of-thought trace ending in a structured report (Metric A). Interface 0 (bottom): ROI crop + question → answer, rejecting background before invoking the language model (Metric B). Blue = frozen, orange = trained, dashed grey = static lookup, green = output.

ternet access. The choice was informed by the previous challenge edition, whose top-ranked team benchmarked five foundation models and found H-Optimus-1 strongest and most stable across institutions and scanners – a domain-specific signal weighted above generic leaderboards. We rejected two alternatives: a Virchow2 [12] ensemble, since GPU memory was already near the deployment limit with one encoder; and TITAN [1], co-trained with a different tile encoder (CONCH) and thus incompatible with H-Optimus/Virchow2 embeddings.

2.3 Multiple-Instance Learning Aggregation

Patch embeddings are aggregated into a slide-level representation with a trained attention-based MIL head [4] rather than mean-pooling, so uninformative patches can be down-weighted instead of diluting the diagnostic signal. The aggregator is trained jointly on organ and diagnostic-template classification rather than organ labels alone; ablating this against alternative MIL architectures (DSMIL, TransMIL) trained only on organ labels showed that *what* the embedding is optimized for mattered more than the attention architecture – organ-only training produced embeddings that were harder to separate for downstream template selection, even with equal or better organ accuracy.

Rather than one model on the full dataset, we train five independent ABMIL instances on five cross-validation folds and average softmax organ probabilities at inference – a variance-reduction argument at zero extra inference cost that, given a single non-revocable submission, also remains the only architecture permitting any held-out local validation.

2.4 Diagnostic Template Selection (Graph Head)

Each CoT trace follows one of several dozen organ-specific question/answer graphs ("templates"), pre-mined from the training corpus. We treat template selection as classification rather than generating graph structure token-by-token: a small head predicts, per organ, which pre-mined template the case follows, and the graph's edges are read directly from a lookup table. This trades template novelty for Edge-F1 reliability – edges must match ground truth verbatim after canonicalization, so paraphrased questions would score zero regardless of correctness – and lets the generative model focus entirely on answer text.

We compared a plain MLP head against a nearest-prototype (cosine-similarity) classifier with class re-weighting, motivated by many templates having only a handful of examples. Evaluated on the final multi-task-ABMIL embeddings, the plain MLP consistently outperformed the prototype classifier and was adopted as final; class re-weighting was kept available but unused, giving no consistent benefit once the embedding itself was multi-task-trained. As with the encoder, five folds are trained and averaged.

2.5 Answer and Report Generation

Free-text answers (for questions not covered by a high-confidence lookup of the most common training answer, above a fixed per-(organ, question) consensus threshold) and the final report are generated by MedGemma 1.5 4B [2] with a LoRA adapter [3], rather than the frozen base model or a full fine-tune. LoRA keeps the base model reusable for both interfaces and small enough to package under the platform's size/load-time constraints; a low-rank adapter over the attention projections was judged sufficient since the task is stylistic/domain adaptation rather than new visual capability. An early adapter version trained on text-only sequences while inference conditions generation on the slide thumbnail for several steps; this mismatch was found and fixed by retraining on the full corpus.

2.6 Retrieval-Augmented Generation (RAG)

For the diagnosis step and report-style conditioning, we retrieve the nearest training cases by cosine similarity in the slide-embedding space (the same space the MIL encoder produces) against an index built from the full corpus. Retrieval excludes the query case itself by exact filename match, so a query WSI can never retrieve its own report; this is a no-op at deployment (the test WSI is never in the training index) but is what makes the local ablation (Section 3.2), which reuses training WSIs as queries, measure genuine retrieval rather than trivial self-lookup. Most final-report scoring criteria reward near-literal overlap with reference text (Section 3), favoring a retrieved, genuinely similar case over free generation. A behavior-consistency step constrains style-example retrieval to cases whose diagnosis behavior (benign / malignant / in-situ) agrees with what was already established earlier in the same chain, falling back to plain nearest-neighbor

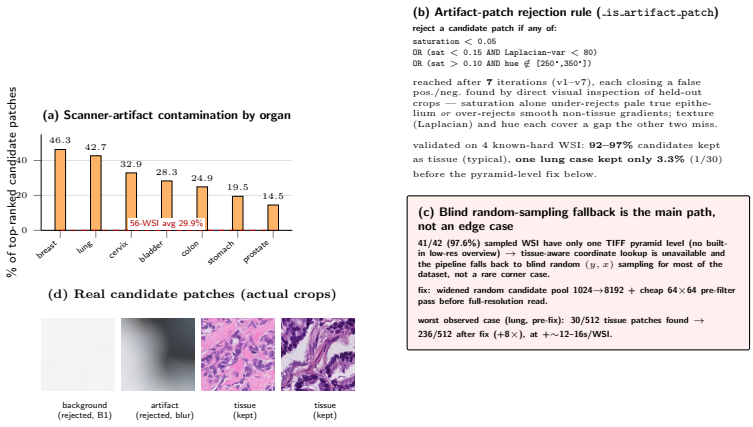

Fig. 2. (a) Scanner-artifact contamination among top-variance-ranked candidates, by organ, 56-WSI baseline, pre-filtering. (b) Final artifact-rejection rule and iterations to reach it. (c) The single-pyramid-level finding driving blind random sampling for most WSIs. (d) Real 224×224 candidate crops, classified by the production rule.

retrieval otherwise – added after free-generation and pure-retrieval components were observed to reach different conclusions for the same case.

2.7 Visual Grounding Interface (Metric B)

Interface 0 receives an ROI thumbnail and a question, and must recognize non-informative background and reject it without a diagnostic claim, otherwise answering from the visible tissue. Background rejection uses a lightweight heuristic (near-white pixel fraction plus mean saturation) rather than a learned classifier: the official background ROIs, generated via unsupervised foreground–background separation [8], are visually unambiguous, and a heuristic is inexpensive, deterministic, with no silent failure mode. When it fires, a fixed non-diagnostic string is returned, which by construction cannot make an incorrect diagnostic claim on background; otherwise the frozen MedGemma 1.5 4B model answers directly from the ROI image and question. All weights – foundation encoder, LoRA adapter, RAG index – are packaged into the model artifact rather than fetched on demand, and every model loads from scratch per container invocation, ruling out any design assuming warm state between cases.

3 Results

A note on what these numbers are, in line with recommended practice for transparent validation reporting [6,10]: our container was not build-verified until after Test Phase 1's regular window had closed, so we had no official score from that window; we subsequently requested and were granted a one-time reopened

Test Phase 1 submission (per organizer forum policy) for paper-reporting purposes only – this score does not count toward the official Test Phase 1 leaderboard or ranking, and it was obtained after our Test Phase 2 submission was already final, so it could not have informed it. Our official Test Phase 2 (final) result was released after Test Phase 1 and is reported in Section 4. Every other number below is therefore either a clean, held-out cross-validation measurement or a local development-set measurement with an explicit leakage caveat – reported explicitly, not omitted.

3.1 Structural Reasoning Metrics (Clean 5-Fold Cross-Validation)

BPV and Edge-F1 (reasoning-path metrics adapted for this challenge from prior chain-of-thought evaluation work [7]) depend only on ABMIL/graph-head template selection, not generated text, so they can be measured with a standard, leakage-free 5-fold cross-validation protocol, compared against our platform-measured Test Phase 1 result (a one-time reopened submission granted for paper reporting only, Section 3): Binary Path Validity (5% of Metric A) reached 74.0% macro-/72.9% pooled-average under clean cross-validation vs. 45.4% platform-measured; Edge-F1 (30%) reached 92.4%/92.2% vs. 82.0% platform-measured; Table 1 gives the corresponding Test Phase 2 values.

3.2 LoRA and RAG Ablation (62-WSI Development Set)

MESS and Final Report (FR) Score depend on generated text and were measured on a 62-WSI local development set with the organizers' official scoring code. **Caveat:** all 62 WSI are part of the LoRA training corpus, so MESS/FR below are an upper bound, not held-out performance – we observed near-verbatim reproduction of a ground-truth report on one case, direct evidence of memorization. The RAG index excludes the query WSI itself by filename (Section 2), so the FR Score increment attributed to RAG (0.557→0.633) reflects retrieval of a genuinely different case, not a self-match; the memorization caveat above is a separate mechanism, LoRA's training-corpus overlap. BPV/Edge-F1 here are further inflated by a separately measured ensemble-leak (4/5 folds saw these cases in training; BPV +13.1pp, Edge-F1 +1.6pp vs. the clean CV above), so should not be compared to it directly. The table's Test Phase 1 row instead gives our single platform-measured, held-out result for direct comparison against these local upper bounds; the final row adds the official Test Phase 2 result (Section 4).

LoRA is the single largest lever among the local conditions (Metric A +0.18 absolute, almost entirely from FR Score and MESS – BPV/Edge-F1 are unaffected by definition). RAG contributes a smaller further gain concentrated in FR Score, consistent with three of its four components (70% of the FR weight) rewarding near-literal reference overlap over paraphrase. The platform-measured Metric A (0.644) lands between the base and LoRA-adapted local conditions – above the no-LoRA base model (0.547) but below both LoRA conditions (0.727, 0.760) – consistent with LoRA still helping on held-out data, just by less than

Table 1. LoRA/RAG ablation, 62-WSI development set (upper-bound numbers, see text), vs. the platform-measured held-out Test Phase 1 result and the official, final Test Phase 2 result.

Condition	BPV	Edge-F1	MESS	FR Score	Metric A
Base MedGemma (no LoRA, no RAG)	0.694	0.883	0.654	0.210	0.547
+ LoRA (no RAG)	0.694	0.883	0.819	0.557	0.727
+ LoRA + RAG (final)	0.661	0.890	0.828	0.633	0.760
Official Test Phase 1 (held-out)	0.454	0.820	0.717	0.489	0.644
Official Test Phase 2 (final)	**0.343**	**0.751**	**0.619**	**0.356**	**0.539**

the leakage-inflated local estimate suggests; Section 4 breaks down where the remaining gap comes from.

To isolate memorization from genuine generalization, we trained a second LoRA adapter on only `fold0_train` (8,968 cases) and evaluated it on 28 held-out, organ-stratified `fold0_val` cases it never saw. Metric A rose 0.639→0.798 (FR Score 0.362→0.769) with zero training/evaluation overlap, directly showing LoRA's local gains are not solely memorization; BPV/Edge-F1 were essentially unchanged, as expected since they depend only on ABMIL/graph-head, not LoRA. This adapter used a faster final-report-only recipe rather than Table 1's full regime (compute-budget constraint), so the two results are not directly comparable step-for-step.

3.3 Visual Grounding (Metric B) Proxy Measurement

We lack local access to the organizers' Metric B judge model (Qwen3-8B [11]), so B2/B3 use a PubMedBERT cosine-similarity proxy and should be read as a lower bound (short clinical sentences sit around 0.4–0.6 cosine similarity even when semantically equivalent). B1 is exact, not proxied: production returns a fixed string on any detected background ROI, so heuristic accuracy on background ROIs directly equals B1. On 18 official ROI pairs: B1 = 1.000 (6/6 background ROIs correctly rejected); B2 proxy trivially 1.0; B3 proxy improved 0.447→0.594 after fixing a prefill bug that stopped the tissue-description model immediately after echoing its own prompt. Officially, Metric B scored a perfect 1.0000 on Test Phase 1, confirming our conservative B3 proxy was indeed a lower bound. Combined with Metric A (0.6435) at 70%/30% weighting, our platform-measured Test Phase 1 Overall Score is $0.70 \times 0.6435 + 0.30 \times 1.0000 = 0.7504$; our official, final Test Phase 2 Overall Score, released after Test Phase 1, is $0.70 \times 0.5393 + 0.30 \times 1.0000 = 0.6775$ (Section 4).

4 Discussion

The platform-measured Test Phase 1 result (Section 3.1, Table 1) checks these local estimates against a real held-out platform score for the first time, and the

check is mixed. The structural half of Metric A transferred only partly despite two local cross-validation measurements a month apart agreeing within noise (Section 3.1): Edge-F1 held up reasonably (92.2% pooled CV vs. 82.0% platform-measured) but BPV dropped substantially (72.9% vs. 45.4%), a larger gap than expected for a metric independent of generated text – with 350 held-out cases in Test Phase 1, this is unlikely to be pure small-sample noise, pointing instead to a genuine template-selection weakness on unseen data rather than a distributional artifact of a small evaluation set. The text-quality half (MESS, FR Score) was already flagged as an upper bound from LoRA-corpus leakage (Table 1); the platform-measured result confirms this, with FR Score falling further than MESS (0.633→0.489 vs. 0.828→0.717), consistent with FR Score's larger literal-overlap component being more memorization-sensitive.

Test Phase 2 (Overall Score 0.6775, Metric A 0.539) is the platform's single, non-revocable final evaluation and the number that determines competition ranking. Every Metric A component fell further than in Test Phase 1 (BPV 0.454→0.343, Edge-F1 0.820→0.751, MESS 0.717→0.619, FR Score 0.489→0.356), while Metric B held at a perfect 1.0000. One plausible contributor: Test Phase 2 has only 70 cases against Test Phase 1's 350, so a single atypical case carries roughly five times the weight on the mean, consistent with (though not proof of) ordinary small-sample variance rather than a new failure mode; we did not identify a specific cause and state this as an open question, not a conclusion.

The single largest process risk was schedule, not architecture: our Docker container was not build-verified until Test Phase 1 had already closed, so Test Phase 2 was our first real build→upload→platform-inference run on unseen data. This also delayed catching a Docker base image missing a C compiler that silently blocked our LoRA adapter from merging at runtime, so early evaluations may have unknowingly run the base model instead of our fine-tuned one – caught only once real platform inference became available, now fixed and verified. Front-loading a minimal submission earlier would have de-risked this independently of model quality, the clearest process change for next time.

Template selection is also capped at BPV ≈74% macro across ∼160 configurations, a real data/feature ceiling – top-3 template scoring instead of argmax is the one unexplored lever.

Responding to a reviewer request, we compared the spatial spread of variance-ranked vs. randomly-retained patches from the same filtered pool (two independent samples, 56 and 105 WSI): variance-ranking touches under a third of the tissue-bearing grid cells that random retention does. Whether this costs anything in downstream structural accuracy remains open – kept patches are mostly genuine tissue, not artifacts (Section 2); full numbers in the Supplementary Material.

Data Use Declaration and Acknowledgments

Data were provided by the REG^2 organizers [9] (CC BY-NC-SA, seven medical centers, ethics-approved). We used publicly released, gated H-Optimus-1 and MedGemma 1.5 4B weights under their licenses, no additional private data.

Code: `https://github.com/mikolaj-olesinski/pathology-report-generation-reg2026`.

References

1. Ding, T., Wagner, S.J., Song, A.H., et al.: Multimodal whole slide foundation model for pathology. arXiv preprint arXiv:2411.19666 (2024)
2. Google DeepMind: Medgemma technical report. `https://github.com/google-health/medgemma` (2025)
3. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
4. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning. pp. 2127–2136. PMLR (2018)
5. Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro pp. 1107–1110 (2009)
6. Maier-Hein, L., et al.: Metrics reloaded: recommendations for image analysis validation. Nature Methods **21**(2), 195–212 (2024)
7. Nguyen, M.V., Luo, L., Shiri, F., Phung, D., Li, Y.F., Vu, T., Haffari, G.: Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 2862–2883 (2024)
8. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9**(1), 62–66 (1979)
9. REG^2 Challenge Organizers: Pathologist reasoning-guided report generation challenge: Structured description of the challenge design. `https://zenodo.org/records/19848983` (2026)
10. Reinke, A., et al.: Understanding metric-related pitfalls in image analysis validation. Nature Methods **21**(2), 182–194 (2024)
11. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
12. Zimmermann, E., Vorontsov, E., Viret, J., et al.: Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv preprint arXiv:2408.00738 (2024)