Guideline: Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al.

Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. Radiol Artif Intell 2024; 6: e240300.

CLAIM 2024 explicitly includes pathology among its intended applications. Items that do not apply to a retrospective, fully public, patch-classification benchmark are marked N/A with the reason stated rather than left blank. Section references are to the revised manuscript.

#	Item	Status	Where / note
1	Identify as a study of AI methodology, specifying the category of technology used	Yes	Title ("resource-aware … computational pathology"); Abstract, sentence 1
2	Structured summary: study design, methods, results, conclusions	Yes	Abstract

Introduction

#	Item	Status	Where / note
3	Scientific and clinical background, including the intended use and role of the AI approach	Yes	Introduction, paragraphs 1–4. Intended use is stated as a reproducible benchmark, explicitly **not** clinical deployment
4	Study objectives and hypotheses	Yes	Introduction, final paragraph (six enumerated contributions)

Methods — study design

#	Item	Status	Where / note
5	Prospective or retrospective study	Yes	Methods, *Study design and reporting boundary*: retrospective, public data only
6	Study goal, such as model creation, exploratory study, feasibility study, non-inferiority trial	Yes	Exploratory reproducibility benchmark; stated in Methods and Discussion
7	Reporting guideline adherence and deviations	Yes	Methods; CLAIM primary, STARD-AI consulted, TRIPOD+AI and DECIDE-AI cited with applicability qualifications

Methods — data

#	Item	Status	Where / note
8	Data sources	Yes	Methods, *Datasets and provenance*; exact Kaggle mirrors, access dates, Zenodo DOI of source family
9	Eligibility criteria: how, why and by whom were data included or excluded	Yes	All distributed patches used; no exclusion applied. Stated in Methods
10	Data pre-processing steps	Yes	Methods, *Preprocessing*; 91 descriptors enumerated, no additional stain normalisation

#	Item	Status	Where / note
11	Selection of data subsets, if applicable	Yes	45,000-image stratified subset defined; matched-budget comparison added at revision
12	Definitions of data elements, with references to Common Data Elements	Yes	Nine tissue classes defined with full names at first use
13	De-identification methods	Yes	Methods; public de-identified benchmark, no identifiers present in the arrays or files
14	How missing data were handled	Yes	No missing data; every distributed patch carries a label

Methods — ground truth

#	Item	Status	Where / note
15	Definition of ground truth reference standard	Yes	Class labels as distributed by the source dataset (Kather et al.), read from the archive and filename prefix
16	Rationale for choosing the reference standard	Yes	Methods; the benchmark's own labels are used so results are comparable with published work on the same splits
17	Source of ground truth annotations; qualifications of annotators	Yes	Annotations originate from the source dataset creators; **the present authors performed no annotation.** Stated in Methods and Limitations
18	Annotation tools	**N/A**	No annotation was performed in this study
19	Measurement of inter- and intra-rater variability	**N/A**	No annotation performed. Limitation 6 records that no pathologist adjudicated error cases

Methods — data partitions

#	Item	Status	Where / note
20	Intended sample size and how it was determined	Yes	Fixed by the public benchmark: 89,996 / 10,004 / 7,180. No power calculation applies
21	How data were assigned to partitions; level at which partitions were assigned	Yes	PathMNIST splits used as distributed. **Partition level is image level**, stated explicitly. For the 224 × 224 subset the mirror's folders were used unchanged and the inability to verify slide- or patient-level separation is stated
22	Level at which partitions are disjoint	Yes	Methods; the CRC-VAL-derived test split originates from different patients than the training source, but the 224 × 224 subset is **not** independent of PathMNIST and this is stated

Methods — model

#	Item	Status	Where / note
23	Detailed description of model	Yes	Methods, *Model architectures and optimisation*; layer-by-layer detail in Supplementary File S3
24	Software libraries, frameworks and packages	Yes	Methods; PyTorch 2.8.0, scikit-learn 1.6.1, NumPy 2.0.2, Python 3.9.6, with versions
25	Initialisation of model parameters	Yes	Default PyTorch initialisation; stated in Methods

Methods — training

#	Item	Status	Where / note
26	Details of training approach	Yes	Methods; loss, optimiser, learning rate, weight decay, schedule, epochs, batch size, augmentation
27	Method of selecting the final model	Yes	Checkpoint with best validation macro-F1; no early stopping. Stated in Methods
28	Ensembling techniques, if applicable	**N/A**	No ensembling used

Methods — evaluation

#	Item	Status	Where / note
29	Metrics of model performance	Yes	Accuracy, macro-F1, balanced accuracy, macro one-vs-rest AUROC, log loss, ECE, MCE, multiclass Brier
30	Statistical measures of significance and uncertainty	Yes	2,000-replicate class-stratified bootstrap; **paired** bootstrap for between-model differences; both explicitly labelled conditional patch-level
31	Robustness or sensitivity analysis	Yes	Fifteen image perturbations; feature-block ablation; six-seed repetition; controlled resolution experiment; cross-platform re-execution (S5)
32	Methods for explainability or interpretability	Yes	Transparent colour-texture baselines with random-forest feature importance, described as low-level image descriptors with an explicit non-causality statement
33	Validation or testing on external data	Partial	The test split derives from CRC-VAL-HE-7K, a different patient source, but is a public split used in prior work and was not accessed only once. Reported as “CRC-VAL-derived MedMNIST test split”; the term “held-out” is deliberately avoided

Results

#	Item	Status	Where / note
34	Flow of participants or cases	Yes	Figure 1; Supplementary Table S1 gives per-class counts per split
35	Demographic and clinical characteristics of cases	**N/A**	No patient-level demographic or clinical metadata is distributed with these datasets. Recorded as limitation 5
36	Performance metrics for optimal model on all data partitions	Yes	Tables I–IV; validation and test reported separately for every model
37	Estimates of diagnostic accuracy and their precision	Yes	Point estimates with 2,000-replicate 95% intervals; paired intervals for model differences
38	Failure analysis of incorrectly classified cases	Yes	Confusion matrices and per-class metrics for all models (S4); representative cases (S6). Described as **observed confusion patterns**, not adjudicated failure modes

Discussion

#	Item	Status	Where / note
39	Study limitations, including potential bias, statistical uncertainty and generalisability	Yes	Discussion, eight enumerated limitations; dataset colour bias is treated as the study's central finding rather than a footnote
40	Implications for practice, including the intended use and clinical role	Yes	Conclusions; support for the reporting protocol is explicitly separated from support for any classifier, and clinical deployment is excluded

Other information

#	Item	Status	Where / note
41	Registration number and registry name	**N/A**	Not a clinical trial; no registry applies to a retrospective public-benchmark analysis
42	Where the full study protocol can be accessed	Yes	Public repository; the analysis scripts constitute the executable protocol
43	Sources of funding and other support; role of funders	Yes	Declarations: no specific external funding
44	Conflicts of interest	Yes	Declarations: none

Summary

Status	Count
Fully addressed	38
Partially addressed, with the limitation stated	1 (item 33)
Not applicable, with reason stated	5 (items 18, 19, 28, 35, 41)
Unaddressed	0

Item 33 is the one substantive partial. We judged it more accurate to state that strict one-time test-set access cannot be documented than to claim a stronger form of independence than the study design supports.