Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Detection First, Grading Second: A Secondary Analysis of Stage-Specific Errors in Two-Stage Colon CAD

Sulav Dahal

sulav.dahal@mavs.uta.edu

The University of Texas at Arlington

Bipul Bhattarai

University of South Dakota

Research Article

Keywords: colon cancer, computer-aided diagnosis, secondary analysis, two-stage pipeline, class imbalance, histological grading, reporting standards

Posted Date: July 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10336605/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Detection First, Grading Second: A Secondary Analysis of Stage-Specific Errors in Two-Stage Colon CAD

Sulav Dahal
Department of Information Systems and Operations Management
The University of Texas at Arlington, Arlington, TX 76019, USA
sulav.dahal@mavs.uta.edu

Bipul Bhattarai
Department of Biomedical Engineering
University of South Dakota, Vermillion, SD, USA
bipul.bhattarai@usd.edu

***Abstract*—Background: Pathologists confirm malignancy before assigning histological grade, yet many CAD systems collapse both decisions into one multiclass score. A published two-stage colon pipeline reported near-perfect detection and harder grading on public histopathology patches [3], [6].**

Objective: This position-style secondary analysis argues that detection-first CAD should be scored with stage-specific metrics, and quantifies what that means using the published confusion matrices from [6].

Methods: We do not claim a new dataset or a newly trained backbone. We reanalyze reported Stage 1 binary and Stage 2 grading confusion matrices to compute adjacent-grade error shares, a stage-tagged failure taxonomy, and prevalence-adjusted positive predictive value (PPV) for Stage 1 under non-balanced prevalence. Framing follows TRIPOD/DECIDE-AI-style reporting principles applied to a staged architecture [8], [9], [13].

Results: Of 130 SVM grading errors, 107 (82.3%) are adjacent-grade swaps; the ensemble reduces total errors to 107 and Grade 3 false negatives from 42 to 27, but adjacent mistakes remain 86.0% of residual errors. Stage 1 sensitivity is 1.0 with specificity 0.9978 on the balanced set; at 1% cancer prevalence the same specificity yields PPV ≈82%, versus ≈99.8% at 50% prevalence.

Conclusions: Staging mirrors pathology workflow and makes clinical costs inspectable. The new contribution here is error anatomy and prevalence restatement—not a new leaderboard number [2], [9].

***Index Terms*—colon cancer, computer-aided diagnosis, secondary analysis, two-stage pipeline, class imbalance, histological grading, reporting standards**

I. Introduction

Colorectal cancer remains a major global burden [1]. At the microscope, two questions arise in sequence: Is this tissue malignant? If yes, how differentiated is the tumor? Treatment intensity depends heavily on the second answer [12], [15], [18].

Computer-aided diagnosis should respect that sequence [2], [5]. A published two-stage colon CAD system separates binary detection from three-class grading using shared ResNet50 embeddings [6]. **This paper is not a new empirical training study.** It is a secondary analysis and position piece: we reinterpret the operating points and confusion matrices from [6] under a stage-specific cost lens, applying existing reporting-guidance ideas (TRIPOD, DECIDE-AI, imbalance-aware metrics) to that architecture [8], [9], [13].

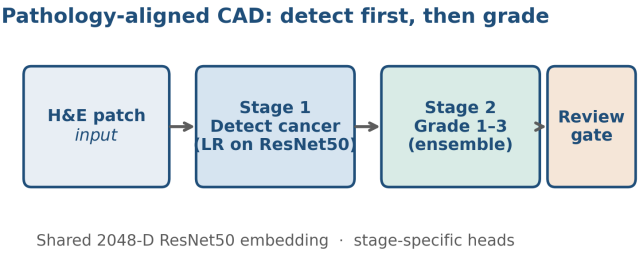


Fig. 1. Detection-first CAD mirrors pathology order: confirm malignancy, then assign grade, then escalate uncertain cases.

Contributions.

1) A concise workflow thesis: detect first, grade second, score each stage separately.
2) A *new* breakdown of published grading errors into adjacent versus non-adjacent swaps, plus a stage-tagged failure taxonomy with counts.
3) Prevalence-adjusted Stage 1 PPV calculations showing why balanced-patch specificity is not a screening PPV.

II. Related Work

Deep learning surveys document strong benchmark performance and persistent domain shift [5]. Colorectal texture and survival histology models established that learned features can encode clinically meaningful structure [4], [15]. Transfer learning from ImageNet remains standard for limited-label medical tasks [7], [14]. Colon detectors with varied backbones report high but protocol-dependent accuracies [10], [11], [17].

TRIPOD and DECIDE-AI already call for transparent reporting, intended use, and human oversight [8], [9]. Saito and Rehmsmeier caution that ROC/AUC can obscure imbalance costs [13]. Our claim is narrower: those principles should be applied *stage-wise* when CAD mirrors pathology's detect-then-grade sequence [6], [12].

TABLE I
PUBLISHED STAGE-WISE OPERATING POINTS [6].

Stage	Task	n	Headline metric
1	Benign vs adenocarcinoma	10,000	Acc. 99.89%; AUC 1.0000
2 (SVM)	Grades 1–3	560	Acc. 76.79%
2 (ensemble)	Grades 1–3	560	Acc. 80.89%

III. Why Staging Is Not Cosmetic

Clinical accounting. Stage 1 can prioritize sensitivity; Stage 2 can prioritize minority-grade recall [13]. A miss-cancer error is not interchangeable with a Grade 2/Grade 3 swap [9].

Dataset honesty. Binary LC25000-style corpora can be large and balanced [3]; graded cancer sets are smaller and skewed (Grade 3 17.7% in [6]). Fused multiclass training mixes those regimes.

Auditability. Stage tags make governance writable: S1-FN demands immediate re-review; S2-G3miss demands therapy-intensity adjudication [2], [12].

IV. Evidence Base and Secondary Analysis Methods

A. Source pipeline (not re-trained here)

We use the published two-stage design in [6]: ResNet50 2048-D embeddings; logistic regression for benign versus adenocarcinoma on n=10,000 balanced patches; RBF-SVM and soft-voting ensemble (weights $[2, 1, 1]$) for Grades 1–3 on n=560 cancer images. Headline anchors: Stage 1 accuracy 99.89%, AUC 1.0000, 0 cancer false negatives, 11 benign false positives; Stage 2 SVM accuracy 76.79%; ensemble 80.89%; Grade 3 recall 57.58%→72.73%.

B. What is new in this paper

From the published confusion matrices in [6] we compute:

- adjacent-grade error counts ($|i - j|$=1) versus non-adjacent errors;
- Grade 3 miss composition (called Grade 1 vs Grade 2);
- a stage-tagged failure inventory;
- Stage 1 PPV at prevalences 50%, 20%, 10%, 5%, and 1%, holding sensitivity/specificity fixed at the reported confusion-matrix values.

We do not claim multi-site validation or a new backbone.

V. Reanalysis Results

A. Published stage-wise anchors

B. Adjacent-grade anatomy of Stage 2 errors

Most grading mistakes are adjacent-grade swaps, not random label noise. The ensemble's clinically material gain is fewer Grade 3 misses (especially fewer Grade 3→Grade 1 jumps), not elimination of adjacency. Accuracy alone (+4.1 points) understates that structure [13].

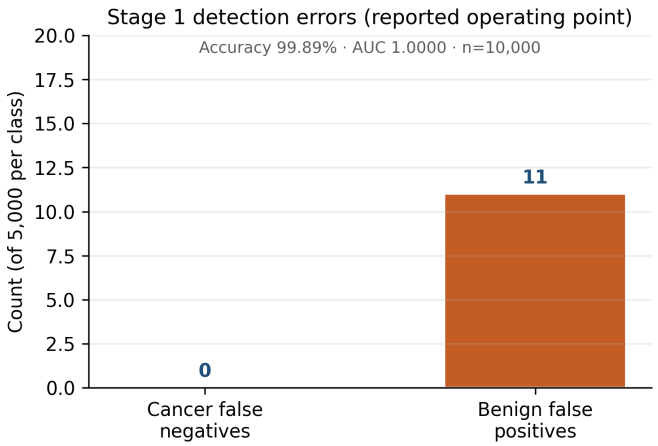


Fig. 2. Stage 1 published errors: 0 cancer false negatives; 11 benign false positives.

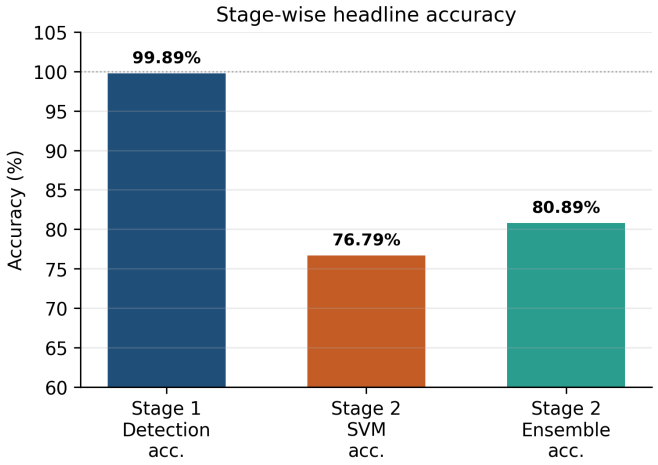


Fig. 3. Published stage-wise accuracy: detection near ceiling; grading remains harder.

TABLE II
SECONDARY ANALYSIS OF PUBLISHED GRADING CONFUSION MATRICES [6].

Model	Errors	Adjacent	Non-adj.	Adj. share
SVM	130	107	23	82.3%
Ensemble	107	92	15	86.0%

C. Stage-tagged failure taxonomy with counts

D. Prevalence-adjusted Stage 1 PPV

On the balanced detection set, specificity is $4989/5000$=0.9978 and sensitivity is 1.0. Holding those fixed: Balanced-patch accuracy therefore cannot be quoted as screening PPV without a prevalence assumption [8], [13].

TABLE III
SELECTED OFF-DIAGONAL CELLS AND GRADE 3 MISSES [6].

Quantity	SVM	Ensemble	Δ
Grade 2→Grade 1	54	39	−15
Grade 3 false negatives	42	27	−15
of which →Grade 2	20	17	−3
of which →Grade 1	22	10	−12
Macro-F1	0.747	0.797	+0.050

TABLE IV
FAILURE TAXONOMY APPLIED TO PUBLISHED COUNTS [6].

Tag	Meaning	Count
S1-FN	Cancer called benign	0
S1-FP	Benign called cancer	11
S2-adj (SVM)	Adjacent-grade swap	107
S2-G3miss (SVM)	Grade 3 under-call	42
S2-adj (ens.)	Adjacent-grade swap	92
S2-G3miss (ens.)	Grade 3 under-call	27

TABLE V
STAGE 1 PPV IF THE SAME SENSITIVITY/SPECIFICITY HELD AT LOWER PREVALENCE.

Prevalence	PPV	Flags / 1000 cases
50% (study mix)	99.78%	501
20%	99.13%	202
10%	98.06%	102
5%	95.99%	52
1%	82.12%	12

VI. DISCUSSION AND RECOMMENDATIONS

Detection-first design is best read as an application of existing clinical-AI reporting principles to a staged pathology workflow—not as a claim of a brand-new evaluation theory [8], [9]. End-to-end multiclass models remain useful for leaderboards, but they hide whether errors are S1-FN or S2-G3miss [5], [10].

For authors of staged colon CAD papers: (i) report S1-FN/S1-FP counts; (ii) report Grade 3 recall, macro-F1, and adjacent-error share; (iii) restate detection PPV at a declared prevalence; (iv) keep pathologist override for residual Grade 3 risk [2], [6], [12].

For readers: treat 99.89% detection and 80.89% grading as answers to different clinical questions. The reanalysis shows grading difficulty is concentrated in adjacent morphology, which is exactly where human review should focus.

VII. LIMITATIONS

All counts derive from a single conference evaluation on public patches [3], [6]. We did not retrain ResNet50, did not access multi-site slides, and did not run prospective concordance. Prevalence-adjusted PPVs assume sensitivity/specificity transfer when prevalence changes—an approximation, not a prospective screening trial [15], [16]. Graded n=560 limits precision of Grade 3 estimates.

VIII. CONCLUSION

Two-stage colon CAD should be judged by whether each stage's errors are clinically priced. Secondary analysis of published confusion matrices shows that grading errors are mostly adjacent-grade swaps, that ensembles mainly buy Grade 3 miss reduction, and that Stage 1 PPV collapses when prevalence leaves the balanced patch setting [6], [9], [13]. Those are the quantities worth reporting next—not another blended accuracy.

REFERENCES

[1] H. Sung et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[3] A. A. Borkowski et al., "Lung and colon cancer histopathological image dataset (LC25000)," arXiv:1912.12142, 2019.

[4] J. N. Kather et al., "Multi-class texture analysis in colorectal cancer histology," *Scientific Reports*, vol. 6, Art. no. 27988, 2016.

[5] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[6] S. Dahal and B. Bhattarai, "Two-Stage Deep Learning Pipeline for Colon Cancer Detection and Histological Grading Using Transfer Learning and Ensemble Methods," in *Proc. 6th Int. Conf. Advanced Research in Computing (ICARC)*, Belihuloya, Sri Lanka, 2026, pp. 1–6.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[8] G. S. Collins et al., "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)," *Annals of Internal Medicine*, vol. 162, no. 1, pp. 55–63, 2015.

[9] B. Vasey et al., "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," *Nature Medicine*, vol. 28, no. 5, pp. 924–933, 2022.

[10] R. Goyal and A. Dudi, "Colon cancer detection using deep learning algorithm," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 13, no. 4, 2025.

[11] T. H. Jaware et al., "Deep learning model for colon cancer classification using InceptionV3," *J. Electr. Syst.*, vol. 20, no. 4s, 2024.

[12] World Health Organization, *WHO Classification of Tumours: Digestive System Tumours*, 5th ed. Lyon, France: IARC, 2019.

[13] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, Art. no. e0118432, 2015.

[14] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[15] J. N. Kather et al., "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLoS Medicine*, vol. 16, no. 1, Art. no. e1002730, 2019.

[16] M. Macenko et al., "A method for normalizing histology slides for quantitative analysis," in *Proc. IEEE Int. Symp. Biomedical Imaging (ISBI)*, Boston, MA, USA, 2009, pp. 1107–1110.

[17] M. A. Talukder et al., "Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning," arXiv:2206.01088, 2022.

[18] Y. Su et al., "Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis," *Computers in Biology and Medicine*, 2022.