Additional file 3. Methods verification appendix

Patient-level multimodal fusion for LSIL-or-above cervical lesion risk stratification

Final analytical specification, sensitivity protocols and evidence boundaries

16 July 2026

1 Purpose and evidence roles

This appendix records the executable analytical specification recovered from the retained source workbooks, modelling scripts, checkpoints, prediction files and run logs. It separates the original diagnostic-context analysis from later sensitivity analyses. No result in this appendix was selected by searching the held-out test set for a better architecture or headline model.

Analysis	Role	Interpretation
Diagnostic-context full model	Primary internally evaluated analysis	Uses contemporaneous clinical variables and a cached image-representation branch; ROC-AUC 0.863.
Fully out-of-fold image representation	Conservative sensitivity analysis	Replaces the cached representation with fold-excluded representations; ROC-AUC 0.814.
Annotation-training comparison	Training-protocol sensitivity	Compares raw-only training with annotation pretraining followed by raw fine-tuning.
Same-split image-backbone benchmark	Exploratory algorithmic benchmark	Compares the full multimodal model with standard image-only architectures on the same locked test patients.
Duplicate-conservative subset	Eligibility sensitivity	Uses prespecified exact- and perceptual-hash criteria; n=38 test patients.
Group OOF, source holdout and deployable variables	Supplementary internal robustness analyses	Uses a separate fixed supplementary configuration; it did not select or replace the primary model.

2 Study setting, cohort and operational endpoint

The study is a single-centre retrospective internal validation analysis using confidential records from the Department of Gynecology, The Second Affiliated Hospital of Guangzhou University of Chinese Medicine / Guangdong Provincial Hospital of Chinese Medicine. Included examination dates span 20 September 2024 to 11 February 2026.

Cohort-flow step	Retained, n	Excluded, n	Reason or role
Clinically labelled authority records	855	–	Valid patient identifier and one of six prespecified groups
Records with linked eligible colposcopy images	377	478	No linked eligible image
Complete clinical-image-ZhengHou cohort	356	21	No linked ZhengHou record among image-linked records
Training partition	254	–	Model fitting and fusion training
Validation partition	51	–	Learner, iteration and threshold selection
Held-out internal test partition	51	–	Internal comparative evaluation

The binary endpoint was derived from the retained routine-clinical case-record group field. Negative groups were control (n=61), same-type persistent infection (n=92) and SPI at the 12-month boundary (n=94). Positive groups were low-grade squamous intraepithelial lesion (n=83), high-grade squamous intraepithelial lesion (n=18) and cervical cancer (n=8). The resulting cohort contained 247 SPI-12-or-below and 109 LSIL-or-above patients. The retained package did not contain a separate pathology-only adjudication table; this is an operational case-record endpoint rather than an independently verified histopathology reference standard.

Patient identifiers were used for modality linkage, quality control and grouped splitting only. They were never prediction features.

3 Patient-level partitioning

Patient leakage groups were assigned to training, validation and test partitions with fixed seed 20260625 and stratification by binary endpoint and colposcopy availability. The partitions contained 78, 15 and 16 positive outcomes, respectively. All records and images belonging to the same leakage group remained in one partition.

Within a prespecified component set, the validation set ranked learners by ROC-AUC and then balanced accuracy at the validation Youden threshold. The held-out test set was not read for learner selection. Several component sets were subsequently compared on the same 51 test patients, so this is an internal comparative test set rather than external confirmation.

4 Clinical predictors and preprocessing

Twenty-two raw clinical predictors were retained:

1. age;
2. number of pregnancies;
3. number of deliveries;
4. number of abortions;
5. menopause status;
6. marital status;
7. contraceptive use;
8. smoking history in years;
9. three or more sexual partners;
10. vaccination status or type;
11. current ThinPrep cytology result;
12. current high-risk human papillomavirus result;
13. koilocytosis;
14. *Trichomonas*;
15. *Candida*;
16. vaginal-discharge cleanliness grade;
17. allergy history;
18. previous high-risk human papillomavirus infection;
19. history of genital warts;
20. history of *Mycoplasma* infection;
21. history of *Chlamydia* infection; and
22. other medical history.

Patient identifiers, patient name, endpoint group, examination date, free-text remarks and same-type infection-duration variables were excluded. Same-type infection duration was excluded because it was structurally related to the negative endpoint definition.

Preprocessing was fitted on the 254-patient training partition and applied unchanged to validation and test data:

- numeric detection required at least 85% of non-missing values to parse as numeric;
- numeric values used training-median imputation and `StandardScaler(with_mean=False)`;
- categorical values used training most-frequent imputation;
- categorical encoding used `OneHotEncoder(handle_unknown="ignore", min_frequency=3)`;
- final encoded clinical dimensionality was 28.

Current cytology, current high-risk HPV and koilocytosis were recorded during the index diagnostic work-up and are diagnosis-proximal. Their inclusion defines the primary result as a contemporaneous diagnostic-context estimate rather than a pre-screening model.

5 Colposcopy image branch

The complete cohort contributed 1,053 images: 347 native, 355 acetic-acid and 351 iodine views. Most patients had three images (338/356); 16 had two, one had one and one had six. Validation and test inference always used raw, unannotated images.

5.1 ConvNeXt-Tiny training specification

Element	Locked specification
Backbone	ConvNeXt-Tiny initialized with ImageNet weights; two-class output
Training input	Resize to 256 pixels; random resized crop to 224 pixels, scale 0.72–1.00
Training augmentation	Horizontal flip probability 0.5; colour jitter: brightness 0.18, contrast 0.18, saturation 0.12, hue 0.03
Validation/test input	Resize to 256 pixels and centre crop to 224 pixels
Normalization	ImageNet channel means and standard deviations
Loss	Class-weighted cross-entropy using patient-level class counts
Optimizer	AdamW
Initial learning rate	0.0001
Weight decay	0.0001
Batch size	24
Data-loader workers	0 in the retained execution configuration
Checkpoint rule	Highest validation patient-level ROC-AUC
Annotation schedule	One annotation-overlay pretraining epoch followed by two raw-image fine-tuning epochs
Fine-tuning learning rate	0.000035
Overlay omission	Probability 0.15 during annotation pretraining
Patient aggregation	Mean image probability across available views

5.2 Scalar out-of-fold image risk

Five-fold stratified out-of-fold image-risk predictions used seed 20260629. Every fusion-training patient's scalar risk was generated by a fold model that had not trained on that patient. Original validation and test patients were excluded from all fold-model training sets. Patient risk was the mean probability across that patient's available raw views.

5.3 Cached diagnostic-context image representation

The original representation branch concatenated patient-level mean and standard-deviation embeddings from an annotation-enhanced checkpoint. A training-fitted standardizer and PCA reduced the representation to 32 components. These cached representation features were not originally generated fully out of fold. They are retained only because they form part of the reported diagnostic-context model; the fully out-of-fold replacement below quantifies the associated sensitivity.

6 Fully out-of-fold image-representation sensitivity

Five retained fold-specific ConvNeXt-Tiny checkpoints were used. Each of the 254 training patients was represented by the single fold checkpoint for which that patient's leakage group was held out. The 51 validation and 51 test

patients were excluded from every fold-model training set; their per-image embeddings were averaged over all five eligible checkpoints.

The 768-dimensional penultimate embedding was extracted after global pooling and head normalization. For each patient, the mean and standard deviation across images were concatenated into 1,536 features. A `StandardScaler` and 32-component PCA were fitted only on the 254 training-patient representations and applied unchanged to validation and test patients. The 32 components explained 89.16% of training-representation variance.

The fixed 77-feature fusion used 28 clinical features, one scalar OOF image risk, 32 fully OOF image components and 16 ZhengHou components. CatBoost parameters were maximum 400 iterations, depth 2, learning rate 0.03, L2 leaf regularization 20, balanced automatic class weights and seed 20260715. The validation set retained 31 trees and selected threshold 0.4238159801. No candidate architecture or parameter was selected on the test set.

The held-out result was ROC-AUC 0.814 (95% CI 0.686–0.923), PR-AUC 0.619 (0.467–0.866), Brier score 0.203, ECE 0.167, balanced accuracy 0.712 and F1 0.612. Stratified confidence intervals and paired differences used 10,000 patient-level bootstrap resamples.

7 ZhengHou semantic branch

ZhengHou text combined tongue colour, tongue-coating colour, pulse and syndrome-pattern/type fields into a plain structured string. The encoder was the locally cached `BAAI/bge-large-zh` model. The retained snapshot identifier was:

`b5d9f5c027e87b6f0b6fa4b614f8f9cdc45ce0e8`

Embeddings were L2-normalized with encoding batch size 32. A training-fitted standardizer and PCA retained 16 components.

LSIL-or-above remained the positive class in all ZhengHou analyses. The ZhengHou-only estimate was reported in that prespecified direction and was not inverted after inspection.

8 Fusion learners and validation-only selection

The diagnostic-context full vector contained 77 inputs: 28 encoded clinical features, one scalar OOF image risk, 32 cached image-representation components and 16 ZhengHou components.

Learner	Candidate specification
Logistic regression	L2 penalty; `C` in {0.03, 0.1, 0.3, 1.0}; balanced class weights; `liblinear`; maximum 3,000 iterations; standardized inputs
Random forest	500 trees; minimum leaf size 3; balanced-subsample class weights
CatBoost	Maximum 400 iterations; learning rate 0.03; depth in {2,3}; L2 leaf regularization in {5,10,20}; log-loss objective; AUC evaluation metric; balanced automatic class weights

The selected diagnostic-context model used CatBoost with maximum 400 iterations, depth 2, learning rate 0.03 and L2 leaf regularization 20. Its validation ROC-AUC was 0.937 and validation balanced accuracy was 0.931. The validation Youden threshold was 0.4036166127 and was applied unchanged to the held-out test set.

For the same-dataset learner benchmark, the highest-ranked candidate within each learner family was fixed from the validation partition before test evaluation. CatBoost, random forest and L2 logistic regression then received the identical 77-feature multimodal matrix and the same 254/51/51 patient-level split. Each learner used its own validation-selected Youden threshold, and no test outcome was used for candidate or threshold selection.

9 Statistical analysis and operating metrics

Primary discrimination metrics were ROC-AUC and precision-recall AUC. Threshold-dependent metrics were balanced accuracy, F1, sensitivity, specificity, positive predictive value and negative predictive value. The primary diagnostic-context confidence intervals used 5,000 stratified patient-level bootstrap resamples that preserved positive and negative counts. Paired model comparisons resampled the same patients and used two-sided empirical bootstrap P values. The same-input learner benchmark used seed 20260629 and 5,000 paired stratified resamples

for CatBoost-minus-comparator differences. Its P values were exploratory and unadjusted for multiplicity. The image-backbone benchmark used 5,000 paired stratified resamples with root seed 20260716 and reported confidence intervals without multiplicity-adjusted hypothesis tests.

Calibration measures were Brier score, expected calibration error with 10 equal-width bins, and a near-unpenalized logistic calibration intercept and slope using `C=1,000,000`, `lbfgs` and maximum 5,000 iterations. Calibration plots used up to five quantile bins. Decision-curve analysis covered thresholds 0.05–0.80 in increments of 0.01 and was exploratory because no clinical action threshold was prespecified.

At threshold 0.4036166127, the diagnostic-context test confusion matrix was TP=14, TN=28, FP=7 and FN=2. Sensitivity was 0.875, specificity 0.800, positive predictive value 0.667 and negative predictive value 0.933.

10 Additional sensitivity and robustness protocols

10.1 Annotation-training ablation

Both protocols used the same five patient-level folds, ConvNeXt-Tiny initialization, optimizer, initial learning rate, class weighting, augmentations and raw-image inference pipeline. Raw-only training used three raw-image epochs per fold. Annotation-informed training used one annotation-overlay pretraining epoch followed by two raw-image fine-tuning epochs. Validation and test inference used raw images only. A protocol-specific Youden threshold was selected on the original validation partition.

On the 51-patient test set, raw-only training achieved ROC-AUC 0.793 and PR-AUC 0.692; annotation-informed training achieved 0.779 and 0.644. The annotation-minus-raw ROC-AUC difference was -0.014 (95% CI -0.148 to 0.109; P=0.846). This did not establish a discrimination advantage from annotation pretraining.

10.2 Same-split image-backbone benchmark

The exploratory architecture benchmark used the same 356 patients, 1,053-image strict manifest, original 254/51/51 split and five retained training-fold manifests as the ConvNeXt image branch. Original validation and test patients were excluded from every fold-model training set. Four additional ImageNet-1K-pretrained image-only backbones were evaluated: ResNet50 (`resnet50.tv_in1k`), EfficientNet-B0 (`efficientnet_b0.ra_in1k`), DenseNet121 (`densenet121.tv_in1k`) and Swin-T (`swin_tiny_patch4_window7_224.ms_in1k`).

All four backbones used the ConvNeXt raw-only preprocessing and optimization specification: resize to 256 pixels, random 224-pixel crop with scale 0.72–1.00, horizontal flipping, colour jitter, ImageNet normalization, class-weighted cross-entropy, AdamW, learning rate 0.0001, weight decay 0.0001, batch size 24 and three raw-image epochs per fold. The best epoch was selected by patient-level ROC-AUC on the fold-held-out training patients. Image probabilities were averaged within patient; original validation and test probabilities were then averaged across the five fold models. A model-specific Youden threshold was fixed on the original validation partition before operating metrics were computed on the locked test partition. The benchmark root seed was 20260716.

The locked-test ROC-AUC/PR-AUC point estimates were 0.721/0.674 for ResNet50, 0.679/0.534 for EfficientNet-B0, 0.752/0.679 for DenseNet121 and 0.743/0.620 for Swin-T. Raw-only and annotation-enhanced ConvNeXt-Tiny yielded 0.793/0.692 and 0.779/0.644, respectively. The full multimodal model yielded 0.863/0.762. Full-minus-image point differences were positive for every comparator; the paired ROC-AUC and PR-AUC intervals against the strongest image-only comparator crossed zero. These comparisons are descriptive and do not establish universal architecture superiority.

10.3 Duplicate-conservative eligibility

Prespecified duplicate components were defined by exact SHA-256 identity, exact perceptual-hash identity, or cross-patient perceptual-hash Hamming distance of four or less. The sensitivity analysis removed patients belonging to these duplicate or leakage-sensitive components. The resulting held-out subset contained 38 patients and 11 positive outcomes. Because this subset differs from the primary n=51 test set, it is not a paired same-sample comparison.

10.4 Separate fixed supplementary configuration

Group OOF, source-batch holdout and deployable-variable analyses used a separate fixed supplementary CatBoost configuration: 300 iterations, depth 5, learning rate 0.025, L2 leaf regularization 10, random strength 1, bagging temperature 0 and balanced automatic class weights. The full diagnostic-context version used 62 features. The

deployable multimodal version used 56 features after current-test and diagnosis-proximal variables were removed. These analyses were not used to select or replace the 77-feature diagnostic-context model.

Five-fold stratified group OOF evaluation of the 62-feature configuration used all 356 patients and yielded ROC-AUC 0.872 and PR-AUC 0.688. Leave-one-source-batch-out ROC-AUC was 0.791 for Batch 1 (n=184; 57 events) and 0.722 for Batch 2 (n=172; 52 events). The 56-feature deployable-variable analysis yielded ROC-AUC 0.686 and PR-AUC 0.441. Source-batch holdout remains internal and is not external validation.

10.5 Complete-case selection audit

The 356 complete-multimodal patients were compared with 499 labelled records outside the modality intersection using variables available before intersection. Mann-Whitney tests and standardized mean differences summarized continuous variables. Chi-square tests and Cramer's V summarized categorical variables, retaining missingness as a category. No unavailable variable was imputed for this comparison. Included patients were younger (median 38 versus 41 years; standardized mean difference -0.238), and source-label and examination-period distributions differed (Cramer's V 0.227 and 0.238). These are descriptive applicability findings, not causal effects.

11 Retained execution environment

Software	Retained version
Python	3.12.10
NumPy	2.3.5
pandas	2.3.0
scikit-learn	1.7.0
PyTorch	2.7.1
torchvision	0.22.1
timm	1.0.15
CatBoost	1.2.8
sentence-transformers	5.4.1

12 Privacy, reproducibility and evidence boundaries

The uploaded supplementary workbooks contain aggregate results only. They exclude patient identifiers, names, filenames, storage paths, raw images, patient-level predictions and embedding arrays. Retained internal reproducibility files include fold checkpoints, image-embedding manifests, patient-level prediction files, preprocessing objects and execution metadata, but these are not included in the direct-submission archive because they may expose confidential linkage information.

The analysis does not claim external validation, prospective validation, independent pathology-only adjudication, pre-screening applicability, established clinical utility or statistically conclusive ZhengHou improvement. The 0.863 result is a diagnostic-context estimate. The fully OOF, duplicate-conservative, source-batch and deployable-variable analyses define the uncertainty and applicability limits that must remain visible in editorial and reviewer correspondence.