

Dear Editor,

Thank you for the opportunity to revise and resubmit our manuscript. We are very grateful to the Editor and all three reviewers for their time, expertise, and insightful feedback. The comments were highly constructive and have enabled us to significantly improve the clarity, rigor, and overall quality of our paper.

We have carefully considered every comment and have revised the manuscript accordingly.

Below, we provide a detailed, point-by-point response to each of the reviewers' comments, outlining the specific revisions made. All changes in the main manuscript text have been highlighted for your convenience.

Reviewer #1,

Thank you for your detailed review and valuable feedback on our manuscript. In response to your specific comments, we have made the following revisions:

Grammatical Errors: A native English speaker with professional expertise in the research area has thoroughly reviewed and revised the manuscript to correct all grammatical errors, ensuring improved clarity and fluency.

Reference Formatting: We have revised references 3, 13, 23, 31, and 37 to align with the journal's official formatting guidelines, utilizing the recommended reference management tool.

We believe these revisions fully address your comments and enhance the manuscript's quality. Thank you again for your insightful feedback, which has significantly contributed to improving our work.

Reviewer #2,

Comment 1 (Abstract): It would be worthy adding the confidence intervals for the AUC which will make the reporting more robust as most viewers on the paper read the abstract while searching the web. This improves the understanding of the statistical validity of the results while searching.

Response: We thank the reviewer for this valuable suggestion to report a measure of statistical robustness in the abstract. We agree completely that providing a measure of variability is crucial for interpreting the results. Given that our performance metrics are

derived from a 5-fold cross-validation, we have carefully considered the most appropriate way to represent this variability. We have chosen to report the standard deviation (SD) of the mean AUC. The SD directly communicates the consistency and stability of our model's performance across the different validation folds, which is a primary goal of cross-validation reporting. While a confidence interval for the mean could also be calculated, the SD more directly reflects the fold-to-fold variation of our model's performance. For readers interested in the precision of the estimate within each individual fold, the 95% confidence intervals are provided for all five folds in the main text.

Comment 2 (Results): The study flow chart (fig 1) last right-hand box should be changed to "absence of at least 12 LN" to obtain the correct meaning.

Response: We thank the reviewer for identifying this confusing phrasing in our study flowchart. The reviewer is correct that this box represents an exclusion criterion, and the original text was misleading. To clarify the meaning, we have revised the text in the final exclusion box of Figure 1 to read "Absence of at least 12 LN (n=4)". The figure has been updated in the manuscript to reflect this change, and we have refined the figure legend for better clarity.

Comment 3 (Discussion): This section should start with a short summary of the main points in the results (with appropriate statistical metrics without SD/95% CI etc.) after which authors should compare the evidence compiled with that in existing literature. Authors discuss several important points in the discussion but better to stick on to the above suggestion as most readers view on the abstract and discussion of the study.

Response: We thank the reviewer for this excellent recommendation to improve the structure and flow of the Discussion section. We agree that beginning with a direct summary of the main results enhances clarity and impact. Accordingly, we have restructured the opening of the Discussion. It now begins with a concise summary of our study's principal findings, presenting the key statistical metrics without their confidence intervals or standard deviations for readability, as suggested. This summary is immediately followed by a comparison of our findings with those in the existing literature. We believe this new structure better highlights the main contributions of our work and makes the key takeaways more accessible to the reader.

Reviewer #3,

Comment 1. Study Design and Data

1a. Sample Size and Generalizability

Reviewer's Comment: The cohort (n = 130) is small for deep learning. While justified as a pilot, the single-center, retrospective design limits generalizability. Recommendation: Acknowledge this limitation prominently and validate findings in larger, multi-center cohorts. Include demographic diversity in future work.

Response: We thank the reviewer for their insightful comments regarding the study's limitations. We fully agree that the sample size of 130 cases is modest for a deep learning study and that the single-center, retrospective design inherently limits the generalizability of our findings. As this was a pilot study intended to establish proof-of-concept for our case-level MIL framework, these were known constraints.

To address this, we have prominently expanded the limitations section in our Discussion to explicitly state these points. We have also incorporated the reviewer's excellent suggestions, emphasizing the critical need for future validation in larger, multi-center, and demographically diverse cohorts to confirm the robustness and clinical utility of our model.

1b. Temporal Scope and LNM Confirmation

Reviewer's Comment: Cases from 2023–2024 may lack sufficient follow-up for metastasis confirmation. Recommendation: Clarify follow-up duration and LNM confirmation criteria (e.g., histopathology vs. imaging).

Response: We thank the reviewer for highlighting the need for this crucial clarification. We wish to clarify that our study was designed to predict synchronous lymph node metastasis, meaning the LNM status determined by the definitive histopathological examination of the surgical resection specimen itself, not long-term metastatic outcomes or disease recurrence. The "ground truth" label for each patient is therefore derived directly from the gold-standard final pathology report at the time of surgery.

To make this explicit in the manuscript, we have revised the "Data Collection" subsection within the Materials and Methods. The revised text now clearly states that LNM status was based on the histopathological assessment of all resected lymph nodes, which is the definitive standard for nodal staging. This clarifies that long-term follow-up was not a component of this particular study's design.

Comment 2. Methodology

2a. ROI Annotation Paradox

Reviewer's Comment: ROI annotations reduced CONCH v1.5 performance (AUC: 0.90 → 0.84). The hypothesis that stroma contains prognostic signals is intriguing but underdeveloped. Recommendation: Discuss stromal biology's role in CRC metastasis (e.g., tumor microenvironment interactions) and quantify stromal features in high-attention patches.

Response: We thank the reviewer for this insightful observation and excellent recommendation. We agree that this "paradox" is one of the more intriguing findings of our study and that our initial discussion was underdeveloped. The reviewer's hypothesis that prognostically critical information resides in the tumor microenvironment (TME) is well-supported by recent literature.

To address this, we have significantly expanded the Discussion section to elaborate on the potential role of stromal biology in CRC metastasis. The new text discusses tumor-stroma interactions, the desmoplastic reaction, and how excluding these features via strict ROI annotation may inadvertently discard valuable predictive signals, thus explaining the performance drop.

Regarding the excellent suggestion to quantify stromal features in high-attention patches, we agree this is a crucial next step for mechanistic understanding. However, this would require developing a new segmentation model and conducting a substantial new analysis. We believe this is beyond the scope of the current revision but have acknowledged it as a key direction for future work in our revised limitations section.

2b. Model Selection and Comparison

Reviewer's Comment: CONCH v1.5 (512px tiles) outperformed UNI2-h (256px), likely due to contextual information. However, computational costs/resource requirements are unaddressed. Recommendation: Report training time, hardware specs, and inference speed to assess clinical feasibility.

Response: We thank the reviewer for this insightful comment. We agree with their assessment that the larger tile size used by CONCH v1.5 likely contributed to its superior performance by capturing more contextual information. The reviewer is also correct that a discussion of computational costs is essential for assessing clinical feasibility and reproducibility, and this was an important omission in our original manuscript.

To address this thoroughly, we have made the following revisions:

Hardware and Software Specifications: We have added a new subsection to the Materials and Methods titled "Computational Environment." This section now provides a

transparent account of the hardware (CPU, GPU, RAM) and software (Python, PyTorch versions) used for all experiments.

Quantitative Performance Analysis: We have incorporated a new paragraph and a new multi-panel figure (Figure 3) into the Results section. This provides a detailed quantitative comparison of the per-epoch training durations for each of the eight model configurations, allowing for a clear assessment of the computational trade-offs between different models and training strategies.

Comment 3. Results and Analysis

Comment 3a. Baseline Characteristics (Table 1)

Reviewer's Comment: Age stratification shows LNM prevalence drops in 70–79y (35%) vs. <60y (66%), but multivariate analysis (e.g., correcting for stage/location) is absent. Recommendation: Perform multivariate regression to identify independent LNM predictors.

Response: We sincerely thank the reviewer for this excellent recommendation. Performing the multivariate analysis has added significant depth to our study.

To address this, we conducted the recommended multivariate logistic regression analysis on the clinical features. Interestingly, the analysis revealed that patients in the 70–79 age group had significantly lower odds of LNM compared to patients younger than 60 (OR 0.24, $p=0.004$). However, other clinical variables, including T stage, tumor location, and CEA level, were not found to be independent predictors of LNM.

Crucially, despite the statistical significance of this one age group, a predictive model built using only these clinical features still demonstrated poor discriminative ability (mean AUC 0.59), as detailed in our results. This finding powerfully reinforces our central thesis: while a single clinical factor may show an independent association with LNM, the clinical feature set as a whole is insufficient for robust risk stratification. The deep learning-derived pathology features provide the essential and overwhelmingly dominant predictive signal required for high accuracy.

We have presented this new analysis in a supplementary table (Table S1) and have integrated these findings and their implications into both the Results and Discussion sections of the manuscript.

Comment 3b. AUC Variability

Reviewer's Comment: UNI2-c shows high AUC fluctuation (Fold 3: 0.95; Fold 2: 0.61),

suggesting instability (Table 2). Recommendation: Investigate causes (e.g., data imbalance per fold) and report standard deviations for all metrics (beyond AUC).

Response: We thank the reviewer for this insightful observation regarding the performance variability of the UNI2-c model. We agree that this is a significant finding that warrants a thorough discussion. We believe it highlights a key difference in the inherent capabilities and robustness of the model architectures themselves.

Interpretation in the Discussion: We have revised our Discussion section to use this observation as further evidence for the superiority of the CONCH v1.5 framework. We now explicitly discuss how CONCH v1.5 demonstrates not only higher predictive accuracy but also far greater stability, using the standard deviation of the AUCs (0.033 for CONCH-c vs. 0.128 for UNI2-c) as direct quantitative evidence. We then connect this superior stability to CONCH's model design—specifically its use of larger input tiles—which likely provides more stabilizing contextual information for making predictions.

Comprehensive Performance Reporting: To provide the comprehensive data requested, we have updated Table S2 that reports the mean and standard deviation for all key performance metrics (including accuracy, sensitivity, specificity, etc.) for all deep learning models. We have also updated the corresponding table for the machine learning classifiers (Table S4) for consistency and clarity.

Comment 3c. Clinical Data Integration

Reviewer's Comment: SVM outperforms other models (AUC: 0.91), but feature importance is unclear. Do pathology features dominate predictions? Recommendation: Use SHAP/LIME to explain feature contributions (e.g., CEA vs. histology).

Response: We thank the reviewer for this crucial suggestion. We fully agree that understanding the feature contributions is essential for validating and interpreting our top-performing model. As recommended, we have performed a feature contribution analysis using SHapley Additive exPlanations (SHAP) on our best-performing SVM model.

This analysis is now presented as a new Figure 5 and is discussed in a new subsection in the Results. The results unequivocally confirm the reviewer's hypothesis: the deep learning-derived pathology feature (CONCH score) is overwhelmingly the most dominant predictor of LNM status across all cross-validation folds. Other clinical variables had a comparatively minor impact. This finding strongly supports our central conclusion that the model's high accuracy is driven primarily by the rich histopathological information captured by our case-level framework.

Comment 4. Figures and Visualization

Comment 4a. Figure 2 (Workflow)

Reviewer's Comment: Low-resolution TIFF impedes readability of text/diagrams.
Recommendation: Reformat using vector graphics (e.g., SVG) and enlarge labels.

Response: We thank the reviewer for this important feedback and for the excellent suggestion to use vector graphics. We agree that the previously submitted figure was of insufficient quality and apologize for the error. While vector formats like SVG are indeed ideal for diagrams, our Figure 2 is a composite illustration that integrates several raster-based images—specifically, the representative histopathology tiles. It is technically infeasible to convert these photographic elements into a pure vector format without losing their essential visual information. Therefore, to address the core issue of readability while preserving the integrity of all figure components, we have re-exported the entire figure from its source file as a high-resolution TIFF.

Comment 4b. Figures 4 and 5 (ROC Curves)

Reviewer's Comment: Overlapping curves in panels A–E obscure model-specific trends.
Recommendation: Plot mean ROC curves per model across folds with confidence bands.

Response: We thank the reviewer for this excellent recommendation to improve the clarity of our visualizations. We agree completely that the original presentation of individual ROC curves for each fold was cluttered and made direct model comparisons difficult.

To address this, we have fully adopted the reviewer's suggestion. We have removed the original Figures 4 and 5 and replaced them with a single, consolidated figure (now Figure 4). This new figure presents a much clearer, side-by-side comparison for each of the eleven classifiers. For each classifier, we now plot the mean Receiver Operating Characteristic (ROC) curve, averaged across all five cross-validation folds, with the standard deviation represented as a shaded confidence band. This new visualization provides an immediate and unambiguous comparison of model performance when using clinical features alone versus the combined clinical and pathology feature set. We are confident that this revision significantly enhances the clarity and impact of our results.

Comment 4c. Figure 7 (Clusters)

Reviewer's Comment: Cluster 3 ("complex glandular architecture") lacks CRC-specific

prognostic evidence (cited literature focuses on lung cancer). Recommendation: Replace with CRC references (e.g., tumor budding/grading systems).

Response: We are very grateful to the reviewer for identifying this significant error. The reviewer is correct; the original citation was inappropriate for the context of colorectal cancer (CRC), and we sincerely apologize for this oversight.

To ensure our discussion is supported by appropriate, CRC-specific evidence, we have replaced the incorrect citation with a new reference that validates the prognostic significance of "complex glandular architecture." Furthermore, following the reviewer's valuable suggestion, we have also added an additional reference to strengthen the link between these architectural features and established CRC tumor grading systems. These changes ensure that the existing argument in our discussion is now grounded in scientifically accurate and relevant literature.

Comment 5a. Pathologist-Level Performance

Reviewer's Comment: Claims that the model "approaches expert performance" are overstated without direct comparison to pathologists' assessments.

Response: We thank the reviewer for this critical and completely valid point. We agree that our original language was imprecise and could be interpreted as overstating our claims, as this study did not include a formal head-to-head comparison between our model's performance and that of pathologists on the same case set. This was a significant oversight in our wording.

To correct this, we have carefully reviewed and revised the entire manuscript to remove any language that states or implies "pathologist-level" or "expert" performance. Instead, our revised claims are now precisely qualified to reflect what our study actually demonstrates: that the model achieves a high level of diagnostic accuracy for this task and that its decision-making process aligns with established pathological principles. We now frame the model's potential as a valuable tool to support or augment a pathologist's workflow, rather than replace or match it. We believe this revised wording accurately reflects our findings without making any unsubstantiated claims.