

# 112090\_Revision\_Auto\_Edited.docx

---

WORD COUNT

5258

TIME SUBMITTED

08-AUG-2025 07:31PM

PAPER ID

117640788

***Retrospective Study*****Predicting Lymph Node Metastasis in Colorectal Cancer Using Case-Level Multiple Instance Learning**

Zou LF *et al.* Case-Level LNM Prediction in CRC

Ling-Feng Zou, Xuan-Bing Wang, Jing-Wen Li, Xin Ouyang, Yi-Ying Luo, Yan Luo, Cheng-Long Wang

**Abstract****BACKGROUND**

Accurate lymph node metastasis (LNM) prediction is crucial for managing locally advanced (T3/T4) colorectal cancer (CRC). However, both traditional histopathology and standard slide-level deep learning often fail to capture the sparse, diagnostically critical features of metastatic potential.

**AIM**

To develop and validate a case-level Multiple Instance Learning (MIL) framework that mimics a pathologist's comprehensive review, thereby improving LNM prediction in T3/T4 CRC.

**METHODS**

Whole-slide images (WSIs) from 130 T3/T4 CRC patients were retrospectively collected. A case-level MIL framework, utilizing the CONCH v1.5 and UNI2-h deep learning models, was trained on features from all H&E-stained primary tumor slides for each

patient. These pathology features were subsequently integrated with clinical data, and model performance was evaluated using the area under the curve (AUC).

## RESULTS

The case-level framework demonstrated superior LNM prediction over slide-level training, with the CONCH v1.5 model achieving a mean AUC ( $\pm$  standard deviation, SD) of  $0.899 \pm 0.033$  vs  $0.814 \pm 0.083$ , respectively. Integrating pathology features with clinical data further enhanced performance, yielding a top model with a mean AUC of  $0.904 \pm 0.047$ , in sharp contrast to a clinical-only model (mean AUC  $0.584 \pm 0.084$ ).

Crucially, pathologist review confirmed that model-identified high-attention regions corresponded to known high-risk histopathological features.

## CONCLUSION

A case-level MIL framework provides a superior and more clinically relevant approach for predicting LNM in advanced CRC. By analyzing the entire slide set per patient and integrating clinical data, this method holds significant promise for refining risk stratification and guiding adjuvant therapy decisions, warranting further large-scale validation.

**Key Words:** Colorectal cancer; Lymph node metastasis; Deep learning; Multiple instance learning; Histopathology

**Core Tip:** To better predict lymph node metastasis (LNM) in advanced colorectal cancer (CRC), this pilot study developed a case-level deep learning framework. By analyzing all of a patient's pathology slides—emulating a pathologist's workflow—the model achieved a high AUC of 0.899, outperforming traditional methods. Integrating clinical data further increased accuracy to 0.904. This interpretable approach is a promising tool for refining LNM risk assessment and helping guide adjuvant therapy decisions.

## **INTRODUCTION**

Colorectal cancer (CRC) remains a leading contributor to cancer-related mortality worldwide [1]. Among its stages, T3 and T4 CRC represent locally advanced disease, characterized by tumor invasion through the muscularis propria into the perirectal or pericolic fat (T3) or further into adjacent structures (T4)[2]. A defining factor in managing these stages is the presence of lymph node metastasis (LNM), which critically determines prognosis and the necessity of adjuvant chemotherapy following surgical resection[3,4]. Compared with T1 and T2 tumors, confined to the bowel wall and often node-negative, rarely warranting adjuvant therapy, T3 and T4 tumors, with their higher propensity for LNM, frequently do—especially when lymph nodes are involved (Stage III)[5]. However, even in node-negative T3 or T4 cases, the risk of occult metastases or recurrence persists, complicating treatment decisions. Current histopathological assessment of resection samples, while standard, is limited by sampling inconsistencies and the potential to miss micrometastases, leading to inaccurate LNM classification and suboptimal therapeutic strategies[6,7].

Deep learning models have significantly advanced pathology image analysis and classification. Techniques such as convolutional neural networks (CNNs), transformers, and ensemble networks have proven effective in classifying pathology images, enhancing the accuracy, consistency, and efficiency of medical diagnostics, particularly for tumor detection and grading[8-10]. Deep learning has revolutionized computational pathology, extending its applications beyond diagnosis to uncover novel pathological evidence, including predicting biomarkers, molecular alterations, and LNM from histological data, as well as enabling virtual staining[11-14]. However, their effectiveness in these tasks remains limited, underscoring considerable potential for further development.

Recent advances in deep learning have demonstrated potential for predicting LNM in CRC using histopathological images. While CNNs combined with clinical data achieve moderate performance with an area under the curve (AUC) of approximately 0.74, their accuracy is still comparable to, and not significantly better than, traditional

histopathological assessment[15,16]. This limitation may stem from methodological constraints: Current models typically rely on slide-level labels during training, which assume uniform distribution of high-risk features (*e.g.*, lymphovascular invasion, perineural invasion, tumor budding) across all slides[17]. However, these features are often sparsely distributed, requiring pathologists to meticulously review entire slide sets for accurate diagnosis. Slide-level training may miss these critical but infrequent indicators, potentially limiting model generalization. To address this, we propose a case-level multiple instance learning (MIL) framework that mirrors the comprehensive evaluation performed by pathologists. By aggregating information across all slides per patient, this approach aims to improve the detection of high-risk features, enhance LNM prediction robustness, and uncover clinically relevant pathological patterns for T3/T4 CRC.

## **MATERIALS AND METHODS**

### ***Data Collection***

A cohort of 153 colorectal adenocarcinoma cases diagnosed between 2023 and 2024 was retrospectively identified from the archives of Chongqing Traditional Chinese Medicine Hospital. Inclusion criteria stipulated patients with pathological stage T3 or T4 primary tumors, a minimum of 12 Lymph nodes examined in the resection specimen, availability of at least four primary tumor slides per case, and no history of neoadjuvant therapy[18]. Following the application of these criteria, 130 cases met eligibility requirements and were included in the final study cohort (Figure 1). This study has been approved by the Institutional Review Board of Chongqing Traditional Chinese Medicine Hospital. For each included case, all available hematoxylin and eosin (H&E) stained slides derived from the primary tumor resection were utilized; slides containing metastatic lymph node tissue were excluded from this analysis. The ground truth for LNM status (positive or negative) was established based on the definitive histopathological assessment of all resected lymph nodes, as documented in the final surgical pathology report. The selected primary tumor slides were subsequently

digitized using an F.Q. CytoSense 40P scanner (Guangzhou F.Q. PATHOTECH Co., Ltd.) at 20x objective magnification (0.8 Numerical Aperture [NA], 0.1760  $\mu\text{m}/\text{pixel}$  resolution).

### ***Data processing***

Whole-slide images (WSIs) captured at 20x magnification were segmented into non-overlapping tiles using QuPath. Tumor areas within the WSIs were annotated independently by two pathologists (CLW and LFZ) to guide tile selection. Tiles containing more than 50% background were excluded, with background pixels identified using the Otsu thresholding method to determine the optimal brightness threshold. The Reinhard method was applied to normalize tile colors. To ensure robust dataset splitting and prevent data leakage, patient-level 5-fold Monte Carlo cross-validation was employed, randomly partitioning the cases into training and validation sets across five iterations. This approach ensures that all slides and tiles from a single patient belong exclusively to either the training or validation set within any given fold, preventing the model from being tested on data from patients it has already seen during training. All models in this study utilized the same resulting split datasets for consistency in training and validation.

### ***Deep Learning Strategy and Feature Extraction***

In this study, a clustering-constrained-attention MIL method[19] was applied to perform instance-level clustering of histopathological images without manual annotations. The hybrid neural network architecture, illustrated in Figure 2, integrated feature extraction and clustering to analyze whole-slide images effectively. Each histopathological image was preprocessed at a resolution of 0.5  $\mu\text{m}/\text{pixel}$  to meet the input requirements of the feature extraction models. Regions of interest (ROIs) were resized to  $256 \times 256$  pixels for UNI2-h and  $512 \times 512$  pixels for CONCH v1.5, followed by normalization using ImageNet parameters (mean: [0.485, 0.456, 0.406]; standard deviation: [0.229, 0.224, 0.225])[20,21]. Features were extracted using UNI2-h by loading

model weights from a designated repository [<https://huggingface.co/MahmoodLab/UNI2-h>]. The model processed  $256 \times 256$  ROIs in inference mode, generating feature embeddings with a dimensionality of 1536 per ROI. Concurrently, features were extracted using CONCH v1.5 by loading model weights from a designated repository [[https://huggingface.co/MahmoodLab/conchv1\\_5](https://huggingface.co/MahmoodLab/conchv1_5)]. The model processed  $512 \times 512$  ROIs in inference mode, producing embeddings with a dimensionality of 768 per ROI. Feature extraction was performed on an NVIDIA GPU-enabled environment, and the resulting features were stored as NumPy arrays in HDF5 format. These embeddings served as input for the clustering-constrained-attention MIL framework to enable downstream instance-level analysis.

#### ***Combination of Feature Data***

Feature data from multiple HDF5 files, each containing patch coordinates and embeddings from histopathological images, were combined into a single HDF5 file. The x-coordinates of patches from subsequent files were shifted by an offset of 768 units relative to the maximum x-coordinate of the prior data to prevent overlap. Coordinates and features were concatenated across files, and the resulting datasets were saved to a new HDF5 file with attributes copied from the first file.

#### ***Determination of the Optimal Number of Clusters***

To identify the optimal number of clusters (denoted as  $k$ ), the elbow method was employed. The hierarchical clustering dendrogram was iteratively cut using the `cutree` function to obtain cluster assignments for  $k$  ranging from 1 to 30. For each value of  $k$ , the Sum of Squared Errors (SSE) was calculated in the Uniform Manifold Approximation and Projection (UMAP) space as follows: For each cluster, the centroid was determined as the mean of the UMAP coordinates of all points assigned to that cluster, and the SSE was computed by summing the squared Euclidean distances from each point to its respective cluster centroid. This process quantified the within-cluster variability for each  $k$ .

### ***Dimensionality Reduction Using UMAP***

To analyze and visualize the high-dimensional feature data extracted from histopathological images, UMAP, a non-linear dimensionality reduction technique, was employed. UMAP was applied to principal component scores that had been derived from a preceding Principal Component Analysis (PCA). The dimensionality of the data had been reduced by PCA, with its most significant variance preserved. This step was facilitated to explore complex patterns within the data in a manner that was both computationally efficient and interpretable.

### ***Histopathological Interpretation of Model-Selected Patches in Training Data***

To further investigate the histopathological features that drove the model's predictive capacity, a targeted review of high-attention regions was conducted using only the training dataset. True positive cases within the training set were identified – defined as cases with confirmed LNM that had been accurately predicted as positive by the model. From these true positive cases, the slide exhibiting the highest predicted probability of LNM was selected; specifically, slides exceeding a probability threshold of 0.9 were prioritized. Subsequently, these selected slides were re-processed through the trained deep learning model to extract the image patches that corresponded to the highest attention weights within the model's architecture. These high-attention patches, representing the regions most influential in the model's positive predictions during training, were then subjected to detailed histopathological interpretation by two experienced pathologists (CLW and LFZ). The pathologists, blinded to the model's output, independently reviewed the patches to identify the predominant histological patterns. Any discrepancies were resolved by consensus discussion to arrive at a final interpretation for each identified cluster.

### ***Machine Learning Classifiers***

To develop models for LNM prediction, a diverse set of machine learning algorithms was employed, each with a distinct architectural foundation. Ensemble methods such as Random Forest and Extremely Randomized Trees were included, where predictions were aggregated from numerous decision trees to achieve robust results. Gradient boosting frameworks, including XGBoost and LightGBM, were also utilized, where refined models were built sequentially by correcting errors from previous stages. Support Vector Machines, architecturally designed to define optimal separating boundaries in high-dimensional space, were also employed. Furthermore, Multilayer Perceptron, a neural network architecture with the capability to learn complex patterns through interconnected layers of nodes, was incorporated. These varied architectures, implemented using the validated scikit-learn library in Python, provided a comprehensive approach to predictive modeling for this study.

#### **Model Interpretability and Feature Contribution Analysis**

To provide interpretability for the machine learning models, a post-hoc analysis was conducted using SHapley Additive exPlanations (SHAP). A model-agnostic, kernel-based SHAP methodology was employed, which is appropriate for explaining non-tree-based architectures such as SVMs. SHAP values were subsequently calculated for each patient in the corresponding validation set. These values quantified the marginal contribution of each input feature—including the deep learning-derived pathology score and all clinical variables, which had been transformed *via* one-hot encoding—to the model's predicted probability for the lymph node metastasis-positive class. The results were visualized *via* SHAP summary plots to evaluate both overall feature importance, ranked by mean absolute SHAP value, and the directional influence of feature values on model output.

#### **Computational Environment**

The deep learning models were developed in Python (version 3.11.5) with the PyTorch framework (version 2.3.1+cu118), running on an Ubuntu-based workstation featuring

<sup>5</sup> an Intel Core i9-14900KF CPU, 128 GB of RAM, and an NVIDIA GeForce RTX 4090 GPU.

### *Evaluation of the deep learning-based model*

To evaluate the predictive performance of each machine learning model for LNM, a comprehensive assessment was conducted. AUC was utilized as the primary metric to quantify the discriminative ability of each model, reflecting its capacity to distinguish between cases with and without LNM. In addition to AUC, a range of other performance metrics were calculated to provide a multi-faceted evaluation, <sup>1</sup> including accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), precision, recall, and F1-score. These metrics were employed to assess various aspects of model performance, such as overall correctness, <sup>4</sup> the ability to correctly identify positive cases (sensitivity), the ability to correctly identify negative cases (specificity), and the balance between precision and recall (F1-score).

## **RESULTS**

### *Baseline characteristics*

This study analyzed a cohort of 130 patients with CRC (59 LNM-positive, 71 LNM-negative), from whom a total of 1,016 primary tumor H&E slides were digitized for <sup>1</sup> analysis. The baseline demographic and clinical characteristics are detailed in Table 1. To identify independent clinical predictors of LNM, a multivariate logistic regression was subsequently performed. This analysis revealed that only patients in the 70–79 age group were independently associated with lower odds of LNM (OR <sup>2</sup> 0.24, 95% CI [0.09–0.63],  $P = 0.004$ ), whereas other factors like T stage, tumor location, and CEA level were not significant (all  $P > 0.05$ ; Table S1).

### *Deep Learning Model Performance for LNM Prediction*

The performance of our deep learning models for LNM prediction was evaluated across three key factors: The granularity of training labels (case-level vs. slide-level), the model

architecture (CONCH v1.5 vs. UNI2-h), and the use of pathologist-annotated tumor regions (ROIs). All performance metrics are detailed in Tables 2 and S2.

Our primary finding is that the case-level training strategy, which aggregates information from all of a patient's slides, consistently and significantly outperformed the slide-level approach. The CONCH v1.5 architecture achieved the highest predictive accuracy, yielding a mean AUC of 0.899 with case-level labels. This represents a marked improvement over the 0.814 AUC it achieved with slide-level labels. The CONCH v1.5 model also proved superior to the UNI2-h architecture under identical training conditions, outperforming it in both case-level (mean AUC 0.899 vs. 0.754) and slide-level (mean AUC 0.814 vs. 0.764) configurations.

Counterintuitively, explicitly guiding the models with ROI annotations did not improve – and in some cases, hindered – performance. For the top-performing CONCH v1.5 model, adding ROIs reduced the case-level AUC from 0.899 to 0.838 and had a negligible effect at the slide-level (mean AUC 0.806). Interestingly, this effect was model-dependent, as the UNI2-h model showed a modest performance benefit from ROI annotation in both case-level (mean AUC 0.754 to 0.813) and slide-level (mean AUC 0.764 to 0.774) settings.

Beyond predictive accuracy, we also assessed computational efficiency by measuring the mean epoch duration during training (Table S3). The case-level approach demonstrated consistently superior efficiency compared to its slide-level counterpart across all tested configurations. This advantage was particularly pronounced for the CONCH v1.5 model, where case-level training significantly reduced epoch duration both with and without ROI annotations (Figure 3). These results suggest that our case-level MIL framework offers a dual advantage in both predictive power and computational resource utilization.

#### *Integration of Clinical and Pathology Data for LNM Prediction*

Building upon the findings that pathology-based deep learning models, particularly CONCH v1.5 with case-level labels, demonstrate robust performance in LNM

prediction (achieving an average AUC of 0.899 as detailed in Table 2), we next sought to evaluate whether integrating clinical features could further enhance predictive capabilities. Building upon the findings that pathology-based deep learning models demonstrate robust performance, we next sought to evaluate whether integrating clinical features could further enhance predictive capabilities. To quantify this, we compared the performance of eleven machine learning classifiers trained on two feature sets: Clinical data alone *vs* a combination of clinical and pathology features. The averaged results from the 5-fold cross-validation are summarized in Figure 4. As illustrated by the clear separation between the ROC curves in each panel, models trained exclusively on clinical data (red curves) exhibited limited predictive power. Mean AUCs ( $\pm$  SD) ranged from  $0.497 \pm 0.044$  (ExtraTrees) to  $0.584 \pm 0.084$  (Support Vector Machine, SVM). In stark contrast, integrating the deep learning-derived pathology features (blue curves) led to a dramatic and consistent improvement across all classifiers. This combined approach yielded mean AUCs ranging from  $0.743 \pm 0.061$  (K-Nearest Neighbors, KNN) to a robust  $0.904 \pm 0.047$  (SVM). This makes the SVM the top-performing method in this study. Other models also showed strong performance with the combined data, notably Logistic Regression (LR) and ExtraTrees, which achieved mean AUCs of  $0.889 \pm 0.066$  and  $0.875 \pm 0.06$ , respectively.

To elucidate the predictive drivers of our top-performing SVM model (mean AUC 0.904), we employed SHAP to assess feature contributions across the five cross-validation folds (Figure 5). The analysis consistently revealed that the pathology score – generated by the CONCH v1.5 model trained with case-level labels – was by far the most influential predictor of LNM status. As demonstrated in the SHAP summary plots, higher pathology scores were strongly associated with an increased likelihood of a positive LNM prediction. In contrast, the remaining clinical variables, including Age, T-Stage, Sex, CEA level, and tumor location, exerted a comparatively minor influence on model output. This interpretability analysis confirms that the model's robust predictive power is primarily derived from the rich histopathological information captured by the case-level deep learning framework.

### *Histopathological Interpretation of High-Attention Morphological Features*

To interpret the morphological basis of the model's predictions, the high-attention image patches automatically identified by the model were first subjected to dimensionality reduction and unsupervised clustering. This computational analysis aimed to group patches based on their learned features without prior human input. Using the elbow method, we identified an optimal number of six clusters, a finding corroborated by UMAP visualization, which demonstrated a clear separation of the feature embeddings into six distinct groups (Figure 6). This result indicates that our case-level MIL model successfully learned to stratify histopathological patterns into cohesive morphological categories, confirming its ability to identify recurring tissue phenotypes predictive of LNM.

Following this computational grouping, a detailed histopathological review of representative tiles from each of the six machine-generated clusters was performed by two expert pathologists (CLW and LFZ) to assign a clinical interpretation to each group (Figure 7). Their analysis revealed that each cluster corresponded to a well-established, high-risk feature associated with aggressive tumor behavior. The identified phenotypes included: Poorly differentiated adenocarcinoma (Cluster 1), prominent desmoplastic reaction (Cluster 2), complex glandular architecture (Cluster 3), micropapillary adenocarcinoma (Cluster 4), overt lymphovascular and perineural invasion (Cluster 5), and signet-ring cell carcinoma (Cluster 6). The autonomous identification of this spectrum of high-risk features, subsequently validated by pathologists, demonstrates that the model's decision-making process aligns with established pathological principles, functioning not as a "black box," but as an interpretable tool for recognizing significant indicators of metastatic potential.

### **DISCUSSION**

In this study, we developed and validated a case-level MIL framework to predict LNM in locally advanced CRC from whole-slide images. Our results demonstrate that this

case-level approach, which aggregates information from all tumor slides for a given patient, significantly outperforms standard slide-level methods, achieving a mean AUC of 0.899 compared to 0.814 for slide-level training. Integrating these pathology features with clinical data further improved accuracy, with the top-performing model yielding an AUC of 0.904, in sharp contrast to a model using clinical data alone (AUC 0.584). Crucially, the model's clinical relevance was underscored by pathologist validation, which confirmed that high-attention regions corresponded to known high-risk histological features predictive of LNM.

The performance of our case-level model (AUC 0.899) demonstrates a substantial advance in LNM prediction. It not only compares favorably to previous deep learning studies that achieved AUCs around 0.74[15, 16], but it also vastly outperformed a model built exclusively from our own clinical data (AUC 0.584). Notably, this clinical model's poor performance persisted even though our multivariate analysis identified age as a statistically significant, independent predictor of LNM. This powerfully illustrates that isolated clinical or demographic variables are insufficient for robust risk stratification. The superiority of our framework stems from its case-level MIL strategy, which emulates a pathologist's comprehensive review of an entire case. This holistic approach is uniquely capable of capturing the heterogeneous tumor characteristics and sparsely distributed prognostic features that are the true drivers of metastatic potential.

Recognizing the pilot nature of this study and the inherent limitations of working with a relatively small cohort of 130 CRC cases, we strategically employed a clustering-constrained attention MIL framework. This choice was particularly well-suited for our pilot investigation as this MIL approach demonstrates notable advantages when applied to smaller datasets, which are common in exploratory medical imaging research[19]. Clustering-constrained attention MIL excels in such scenarios by enabling effective instance-level learning from bag-level labels (case-level in our study), allowing the model to extract meaningful patterns from individual histopathology tiles despite the limited overall case numbers. The integrated clustering mechanisms enhance the organization of the feature space, potentially improving model robustness and

generalization even with a smaller training set typical of pilot studies. Concurrently, the attention mechanisms ensure the model prioritizes the most diagnostically significant regions within each WSI for accurate LNM prediction.

The observed superiority of case-level labeled datasets compared to slide-level labeled datasets in our study directly reflects the inherent nature of histopathological assessment for LNM risk. Critical diagnostic features indicative of metastasis, such as lymphovascular invasion, perineural invasion, and tumor budding, are often sparsely distributed across whole slide images within a patient case. These high-risk features are not always present in every slide and can be missed if models are trained solely on slide-level labels, which implicitly assume feature homogeneity across all slides of a case. In clinical practice, pathologists undertake a comprehensive review of all available slides for a given patient to identify these potentially infrequent yet highly prognostic features. This holistic, case-centric evaluation is crucial for accurate LNM risk stratification. Our finding that case-level training significantly enhanced model performance (AUC 0.899 vs. 0.814) strongly suggests that this approach more effectively captures the subtle but critical information distributed across the entire patient case, thereby more closely mirroring and potentially augmenting the diagnostic acumen of expert pathologists. While a direct comparison with pathologists was not performed, a model achieving this level of accuracy demonstrates a high level of diagnostic accuracy for this task. This performance significantly surpasses that of models relying on less comprehensive data or clinical factors alone.

Furthermore, our findings revealed a performance advantage for the CONCH v1.5 model over UNI2-h across most training paradigms. This superiority extended beyond mere predictive accuracy to include model stability and robustness. While both models were trained on identical case-level cross-validation splits, CONCH v1.5 delivered highly consistent performance (mean AUC  $0.899 \pm 0.033$ ), whereas UNI2-h exhibited significant instability, with its AUC fluctuating from 0.607 to 0.954 between folds (mean AUC  $0.754 \pm 0.128$ ). This disparity may be attributed to several factors, including differences in model architecture and input resolution. CONCH v1.5, processing larger

512x512 pixel patches compared to UNI2-h's 256x256, likely benefits from a broader contextual view of the histopathological landscape within each tile. In pathology image analysis, context is paramount; larger tiles can encompass more complex tissue architectures, tumor-stroma interactions, and subtle spatial relationships between different cell types, all of which could be crucial for discerning prognostic features related to LNM[22-24]. In contrast, the smaller ROIs processed by UNI2-h might capture finer details but potentially at the cost of losing broader contextual information.

Interestingly, manual tumor annotation did not improve—and in some instances, even reduced—predictive performance. This counterintuitive result strongly suggests that restricting the model's view to only neoplastic cells detrimentally constrains its learning capacity. We hypothesize this is because the model's accuracy depends on its ability to assess the entire tumor microenvironment (TME), a complex landscape containing competing biological signals that collectively determine metastatic potential[25, 26]. On one hand, the TME harbors a pro-tumorigenic stromal response. This is driven by cancer-associated fibroblasts that mediate the desmoplastic reaction—a feature our model correctly identified in high-attention regions (Cluster 2)—and actively remodel the extracellular matrix to create pathways for tumor cell dissemination [27, 28]. Conversely, the TME can also mount a powerful anti-tumorigenic immune response, organized within Tertiary Lymphoid Structures (TLS). These immune hubs are associated with a lower risk of metastasis and improved patient survival, functioning as local sites for adaptive immunity [29, 30]. Both the dense, reactive stroma and the lymphocytic infiltrates of TLS are visually prominent features on standard H&E slides. It is therefore highly plausible that a model with a holistic, unconstrained view learns to assess the net prognostic impact of these competing forces. By forcing the model to focus only on tumor cells, we prevent it from learning this crucial biological balance. Our clustering-constrained-attention MIL framework, which is designed to discover salient regions without relying on precise localization, is thus uniquely suited to learn from this complete biological context, rendering explicit ROI annotations not only redundant but counterproductive.

A key strength of our case-level MIL framework lies in its profound interpretability. The model's attention mechanism effectively mirrors a pathologist's diagnostic process by focusing on high-risk features, a finding quantitatively confirmed through the unsupervised clustering of high-attention regions. This analysis demonstrates that the model did not operate as an uninterpretable "black box" but instead learned to autonomously identify and group a validated spectrum of histopathological risk factors directly associated with metastatic potential in T3/T4 colorectal cancer. Our subsequent pathological review of these machine-generated clusters validated their clinical significance in a sequential and systematic manner.

The model first identified Cluster 1, characterized by poorly differentiated adenocarcinoma. This finding is highly significant, as the loss of glandular differentiation is a fundamental hallmark of aggressive tumor biology and a potent predictor of metastatic disease[31,32]. By prioritizing these regions, the model correctly learned to associate high-grade cytological atypia and disorganized growth patterns with an increased risk of LNM. Next, the model focused on Cluster 2, which featured a prominent desmoplastic reaction. This highlights its capacity to recognize the dynamic tumor-stroma interface, where a dense fibroblastic response signifies an active, aggressive invasion process integral to tumor progression[33,34].

The analysis then progressed to identifying high-risk architectural patterns. Cluster 3 captured regions defined by complex glandular structures, including cribriform and fused glands. While these patterns are named prognostic factors in other cancers, in colorectal adenocarcinoma their significance lies in their contribution to tumor grading. Such complex and disorganized structures are hallmarks of high-grade morphology, signifying a loss of the normal glandular architecture. This high-grade status is a powerful and well-established predictor of aggressive tumor behavior, including increased invasiveness and a higher propensity for lymph node metastasis[35-40]. Therefore, the model's specific attention to these features demonstrates its ability to discern a critical component of the tumor grading system used by pathologists to assess metastatic risk. Following this, the model identified Cluster 4, which consisted of

micropapillary adenocarcinoma. The specific identification of this pattern is particularly noteworthy, as pathologists recognize this variant as a well-established, high-risk subtype with a strong independent association with LNM and adverse outcomes[41-43].

Crucially, the model proved adept at pinpointing the most direct evidence of metastatic capability. Cluster 5 contained unequivocal examples of overt lymphovascular and perineural invasion. The ability to detect these often focal events, which represent the primary conduits for tumor dissemination, is a testament to the model's clinical utility and its potential to augment a pathologist's review by flagging these critical, high-yield regions[44,45]. Finally, Cluster 6 isolated another aggressive subtype, signet-ring cell carcinoma. The model's specific attention to this variant is highly relevant, given that it is notorious for its rapid clinical course and high propensity for lymphatic spread[46-51]. Collectively, the sequential identification of these six distinct high-risk phenotypes validates our model's alignment with established pathological principles and underscores its power as a robust and interpretable predictive tool.

This study has several strengths. Foremost, the introduction of a case-level MIL framework directly addresses a critical methodological gap in deep learning-based histopathological analysis for LNM prediction. By moving beyond slide-level labeling and mirroring the comprehensive, multi-slide evaluation inherent to pathological practice, we have demonstrated a tangible improvement in predictive performance. The robust AUC achieved, particularly with the CONCH v1.5 model and further enhanced by integrating clinical data, underscores the clinical potential of this approach. Moreover, the qualitative validation provided by expert pathologist review of the high-attention tiles lends significant biological plausibility to our model, demonstrating its ability to identify and prioritize features that align with established pathological knowledge of CRC metastasis. However, as a pilot study, this research has inherent limitations that must be acknowledged. First and foremost, the retrospective nature of our study and its reliance on data from a single institution inherently limit the generalizability of our findings to broader patient populations and diverse clinical

settings. The relatively small sample size of 130 patients, while sufficient to establish proof-of-concept for our case-level approach, is modest for deep learning applications. Therefore, validation in larger, multi-center cohorts is a crucial next step to solidify the robustness and clinical utility of our model. Future research should also prioritize the inclusion of greater demographic and geographic diversity to ensure the model performs equitably across different patient groups. Furthermore, our focus on T3/T4 tumors, while clinically relevant for adjuvant therapy decisions, restricts the applicability of our conclusions to these specific stages of CRC. Future research should explore the performance of this case-level MIL framework across the entire spectrum of CRC stages and in datasets encompassing patients who have received neoadjuvant therapies. Finally, while pathologist review provided valuable insights, a more quantitative approach to characterizing and validating the features within the high-attention tiles could further strengthen the interpretability and clinical translatability of our findings.

This study provides a promising avenue for improving LNM risk assessment in locally advanced CRC. Our case-level MIL framework demonstrates a potential improvement in predictive accuracy compared to slide-level methods, which could contribute to more informed adjuvant therapy decisions in the future. The integration of clinical data further enhances the practicality of this approach within existing clinical workflows. Moreover, this work adds to the growing body of evidence supporting the utility of MIL in computational pathology for histopathological analysis. The model's identification of pathologically relevant features, as validated by expert review, provides a basis for further exploration of deep learning methodologies for feature discovery in cancer. Future research can build upon these findings to more quantitatively investigate identified histological features and explore the broader applicability of case-level MIL strategies in other diagnostic areas.

## **CONCLUSION**

Our study demonstrates that a case-level deep learning approach improves the prediction of LNM in advanced CRC. This method, which analyzes the entirety of patient pathology slides, better reflects clinical practice and enhances accuracy compared to traditional slide-based analysis. Combining image-based features with clinical data further strengthens prediction. Importantly, expert pathologist review validates that the model identifies clinically relevant tumor characteristics. These results indicate that case-level deep learning holds significant promise for refining LNM risk assessment in CRC and advancing computational pathology.

## ORIGINALITY REPORT

---

# 2%

SIMILARITY INDEX

---

### PRIMARY SOURCES

---

1	<a href="http://www.frontiersin.org">www.frontiersin.org</a> Internet	29 words — 1%
2	<a href="http://www.mdpi.com">www.mdpi.com</a> Internet	26 words — < 1%
3	<a href="http://f6publishing.blob.core.windows.net">f6publishing.blob.core.windows.net</a> Internet	21 words — < 1%
4	<a href="#">Aldair Darlan Santos-de-Araújo, Daniela Bassi-Dibai, Izadora Moraes Dourado, Renan Shida Marinho et al. "Prognostic value of the duke activity Status Index Questionnaire in predicting mortality in patients with chronic heart failure: 36-month follow-up study", BMC Cardiovascular Disorders, 2024</a> Crossref	16 words — < 1%
5	<a href="http://arxiv.org">arxiv.org</a> Internet	15 words — < 1%

---

EXCLUDE QUOTES ON

EXCLUDE SOURCES < 12 WORDS

EXCLUDE BIBLIOGRAPHY ON

EXCLUDE MATCHES < 12 WORDS