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.

Integrating Deep Learning Radiomics and Node-RADS for Interpretable Prediction of Colorectal Cancer Lymph Node Metastasis

Ouyang Yongxi
The First People's Hospital of Foshan (Foshan Hospital Affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology

Cuiling Zhu
Foshan Hospital of Traditional Chinese Medicine

Zhenyu Xu
Foshan Fosun Chancheng Hospital

Kaiting Han
The First People's Hospital of Foshan (Foshan Hospital Affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology

Xuanming He
The First People's Hospital of Foshan (Foshan Hospital Affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology

Diliang He
The First People's Hospital of Foshan (Foshan Hospital Affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology

Xin Diao
The First People's Hospital of Foshan (Foshan Hospital Affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology

Chengming Wang
The First People's Hospital of Foshan (Foshan Hospital Affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology

Tingsong Fang
Foshan Hospital of Traditional Chinese Medicine

Yunjun Yang
The First People's Hospital of Foshan (Foshan Hospital Affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology

Zhifeng Xu
xuzf83@126.com

The First People's Hospital of Foshan (Foshan Hospital Affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology

Research Article

Keywords:

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11444576/v1

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

Additional Declarations: No competing interests reported.

Integrating Deep Learning Radiomics and Node-RADS for Interpretable Prediction of Colorectal Cancer Lymph Node Metastasis

Yongxi Ouyang[1], Cuiling Zhu[2,3,] Zhenyu Xu[4], Kaiting Han[1], Xuanming He[1], Diliang He[1], Xin Diao[5], Chengming Wang[1], Tingsong Fang[2,3], Yunjun Yang[1*], Zhifeng Xu[1*]

[1]Department of Radiology, The First People's Hospital of Foshan (Foshan Hospital Affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology, Foshan 528000, China
[2]Department of Radiology, Foshan Hospital of Traditional Chinese Medicine, Foshan 528099, China
[3]The Eighth Clinical Medical College of Guangzhou University of Chinese Medicine, Foshan 528099, China
[4]Department of Radiology, Foshan Fosun Chancheng Hospital, 528031, China
[5]Department of Nuclear Medicine, The First People's Hospital of Foshan (Foshan Hospital Affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology, Foshan 528000, China

Yongxi Ouyang and Cuiling Zhu contributed equally to this work and should be considered co-first author.

Zhifeng Xu and Yunjun Yang contributed equally to this work.

***Corresponding authors:**

E-mail address： xuzf83@126.com(Z. Xu); yangyunjun111@163.com(Y. Yang).
Address: The First People's Hospital of Foshan (Foshan Hospital Affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology, Foshan 528000, China.

ABSTRACT

Background and purpose: To develop and validate an interpretable multimodal fusion model integrating deep learning radiomics (DLR) and Node-RADS score for preoperative prediction of lymph node metastasis (LNM) in colorectal cancer (CRC).

Methods: This retrospective multicenter study included 892 CRC patients (training: Center 1, n=684; external validation: Centers 2/3, n=208). Venous-phase CT tumor regions were segmented. 1394 radiomic and 320 ResNet-50 deep learning features

were extracted; LASSO was used for feature selection. An XGBoost DLR model was integrated with Node-RADS score and clinical variables to construct a multimodal nomogram. Performance was assessed by AUC, calibration, decision curve analysis, and the DeLong test. SHAP analysis provided model interpretability.

Results: LASSO regression identified 24 deep learning features and 9 radiomic features with non-zero coefficients, wherein deep learning features demonstrated higher predictive weights. The XGBoost-based DLR model achieved AUCs of 0.876 and 0.798 in training and external validation cohorts, respectively. The multimodal fusion model, incorporating Node-RADS score and selected imaging features, yielded significantly improved AUCs of 0.900 (95%CI:0.874–0.926) and 0.870 (95%CI:0.821–0.919), outperforming both the DLR model and the clinical model (AUC 0.751/0.641) (all DeLong $P< 0.05$). Calibration curves demonstrated excellent agreement (Hosmer–Lemeshow $P > 0.05$), and decision curve analysis confirmed superior net clinical benefit across wide threshold ranges. SHAP analysis identified DL_27 and lbp_3D_m2_glszm_ZoneEntropy as the dominant predictive features.

Conclusion: This externally validated, interpretable CT-based multimodal model integrating DLR signatures and Node-RADS score enables accurate preoperative prediction of LNM in CRC, offering a noninvasive decision-support tool to guide personalized treatment.

Keywords: Radiomics; Deep learning; Colorectal cancer; lymph node metastasis; Node-RADS score

INTRODUCTION

Colorectal cancer (CRC) remains a leading gastrointestinal malignancy globally, with China accounting for approximately 28% of incident cases and cancer-related deaths—the highest proportion worldwide [1]. The profound survival disparity between early- and late-stage disease (5-year survival exceeding 80% for stage I versus merely 10% for stage IV) underscores the critical importance of accurate preoperative staging [2]. Among all prognostic determinants, lymph node metastasis

(LNM) status represents the single most influential factor guiding therapeutic decisions, particularly regarding neoadjuvant therapy and adjuvant chemotherapy [3,4]. The emergence of total neoadjuvant therapy (TNT) strategies has further amplified the clinical imperative for precise nodal staging, as node-positive patients derive significant survival benefits from preoperative intensification regimens [5].

Despite this unequivocal clinical need, preoperative LNM assessment in CRC confronts a persistent and well-documented diagnostic dilemma. Conventional contrast-enhanced CT—the standard imaging modality for preoperative staging—demonstrates suboptimal accuracy, with recent nationwide data revealing radiological-pathological agreement of only 57% for N staging [6]. This diagnostic inadequacy is particularly consequential in the context of imaging-negative occult lymph node metastasis (OLNM), which occurs in up to 5.8% of node-negative patients following adjuvant chemotherapy and contributes substantially to under-staging and suboptimal treatment allocation [7]. The fundamental limitation lies in traditional morphology-based criteria (size, shape, border characteristics), which exhibit considerable overlap between benign and malignant nodes and yield both false-positive and false-negative rates exceeding 40% in contemporary series [6,8].

In response to these challenges, the Node Reporting and Data System (Node-RADS) was introduced in 2021 to standardize lymph node assessment through a 5-point probability scale integrating size and configurational features [9]. Preliminary validation studies have demonstrated its utility across multiple malignancies, including gastric, lung, bladder, and cervical cancers [10-13]. For colorectal cancer specifically, MRI-based Node-RADS assessment has shown superior performance compared to conventional ESGAR criteria [14]. However, CT-based Node-RADS validation in colon cancer remains limited to small single-center cohorts, and its standalone diagnostic performance—while superior to size criteria alone—remains insufficient for definitive clinical decision-making [6,8].

Concurrent advancements in artificial intelligence have catalyzed radiomics and deep learning approaches capable of extracting high-dimensional, subvisual quantitative features from medical images [15]. Radiomics enables comprehensive tumor phenotyping through high-throughput feature extraction, whereas deep learning architectures can autonomously learn hierarchical representations beyond human perceptual capacity. These techniques have demonstrated considerable promise in CRC applications, including tumor grading, microsatellite instability prediction, and prognostic stratification [16,17]. Nevertheless, dedicated investigations targeting CRC lymph node metastasis remain comparatively scarce, and several critical gaps persist. First, most existing studies are constrained by single-center designs and modest sample sizes, limiting generalizability [18,19]. Second, the predominant focus on either radiomics or deep learning features in isolation fails to capture their complementary information; emerging evidence suggests that integrated approaches substantially outperform individual methodologies [20]. Third, the potential synergy between quantitative imaging biomarkers and structured Node-RADS assessment remains largely unexplored, representing a missed opportunity for diagnostic optimization.Furthermore, the clinical translation of machine learning models has been impeded by their intrinsic "black-box" nature, which engenders hesitancy among clinicians and regulatory bodies. Shapley Additive Explanations (SHAP)—a game theory-based interpretability framework—offers a rigorous solution by quantifying each feature's marginal contribution to individual predictions, thereby enhancing model transparency and fostering clinical trust [21,22]. Recent CRC radiomics studies have successfully employed SHAP analysis to identify dominant predictive features and validate biological plausibility, establishing a methodological precedent for interpretable model development [16,17].

Therefore, this multicenter study was designed to develop and externally validate an interpretable, CT-driven multimodal fusion model integrating complementary information sources: (1) conventional radiomics features capturing hand-crafted texture and phenotypic descriptors; (2) deep learning features encoding

high-dimensional abstract representations; and (3) Node-RADS scores providing structured radiologist-derived probability assessment. We hypothesize that this synergistic integration will substantially outperform individual approaches, while SHAP-based interpretability analysis will elucidate the relative contributions of each feature domain, thereby advancing both predictive accuracy and clinical acceptance. The resultant noninvasive tool holds potential to enhance preoperative precision, optimize patient selection for neoadjuvant therapy, and ultimately improve survival outcomes in colorectal cancer.

METHODS

Study Population

Ethical approval for this retrospective study was derived from the institutional review board of the First People's Hospital of Foshan (Foshan Hospital Affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology (Approval No. 93, 2025). Written informed consent was waived by the Institutional Review Board due to the retrospective nature of this study. Consecutive patients with pathologically confirmed colorectal adenocarcinoma who underwent curative-intent surgical resection and lymphadenectomy between May 2014 and March 2025 at three tertiary referral centers were retrospectively identified. Inclusion criteria were: (1) histopathologically confirmed CRC with documented postoperative lymph node status; (2) preoperative contrast-enhanced abdominal CT performed within 30 days prior to surgery at the enrolling institution; and (3) no prior neoadjuvant chemotherapy or radiotherapy. Exclusion criteria comprised: (1) synchronous or metachronous malignancies; (2) incomplete clinicopathological or imaging data; and (3) CT image quality inadequate for segmentation or Node-RADS assessment.

Following application of eligibility criteria, 892 patients constituted the study cohort. Patients from Center 1 (n=684; May 2014–March 2025) were allocated to the training cohort, while patients from Centers 2 (n=102; September 2022–September 2024) and 3 (n=106; September 2018–March 2025) were pooled to form an

independent external validation cohort. The patient selection flowchart is presented in Fig. 1. Clinicopathological variables included age, sex, preoperative carcinoembryonic antigen (CEA) level, neutrophil-to-lymphocyte ratio (NLR), pathological T stage (American Joint Committee on Cancer, 8th edition), and postoperative lymph node metastasis status. Node-RADS scores were prospectively recorded or retrospectively assigned by standardized chart review as detailed below.

CT Image Acquisition and Tumor Segmentation

All patients underwent preoperative contrast-enhanced abdominopelvic CT following institutional protocols. Non-ionic iodinated contrast (80–100 mL) was administered via antecubital vein at 3–3.5 mL/s, with venous phase images acquired at 60–70 seconds post-injection. Detailed acquisition parameters across institutions are provided in Supplementary Data Table S1.

Regions of interest (ROIs) encompassing the entire primary tumor were manually delineated on venous phase images using ITK-SNAP (version 3.8.0; www.itksnap.org). Segmentation was performed by an abdominal radiologist (5 years of experience) blinded to histopathological outcomes. To assess reproducibility, 50 randomly selected cases were re-segmented by the same observer after a 4-week interval and independently by a second radiologist (8 years of experience). Intra- and inter-observer agreement were quantified using intraclass correlation coefficients (ICC), with ICC $\geq$ 0.75 considered indicative of satisfactory reliability. All segmentations subsequently underwent image preprocessing, including resampling to isotropic voxel dimensions (1 × 1 × 1 mm^3) and grayscale intensity normalization (z-score transformation). Fig.2 summarizes the entire study workflow.

Node-RADS Assessment

Lymph node evaluation was performed independently by two abdominal radiologists (with 5 years of experience and 10 years of experience, respectively), both blinded to histopathological outcomes and clinical data. Peritumoral lymph nodes within the colorectal mesenteric drainage basin were assessed for size,

short-axis diameter, margin characteristics, internal texture, and morphology using the Node-RADS 1.0 scoring system (Supplementary Data Fig. S1). Lymph nodes were classified as enlarged when short-axis diameter exceeded 5 mm for rectal or 10 mm for colonic primary tumors, in accordance with Node-RADS guidelines. For each patient, both size-dominant and morphology-dominant lymph nodes were scored, and the RADS_max—defined as the higher of the two Node-RADS scores—was recorded as the final nodal probability estimate.

Interobserver agreement for Node-RADS scoring was assessed using intraclass correlation coefficient (ICC) with a two-way random-effects model for absolute agreement. An ICC $\geq$ 0.75 was considered indicative of good reliability. In cases of discrepant scoring, a consensus reading was reached through discussion between the two radiologists, and the consensus score was used for subsequent analyses.

Radiomic Feature Extraction

Radiomic feature extraction was conducted using PyRadiomics (version 3.0.1; pyradiomics.readthedocs.io), an open-source platform compliant with Imaging Biomarker Standardization Initiative (IBSI) guidelines. From segmented 3D tumor volumes, 1394 quantitative radiomic features were extracted across three predefined feature classes: (1) shape-based descriptors characterizing three-dimensional tumor morphology; (2) first-order intensity statistics quantifying voxel intensity distribution; and (3) higher-order texture features derived from gray-level co-occurrence (GLCM), gray-level run length (GLRLM), gray-level size zone (GLSZM), neighboring gray-tone difference (NGTDM), and gray-level dependence matrices (GLDM). Complete feature definitions and extraction parameters are provided in Supplementary Materials (Supplementary Data Table S2).

Deep Learning Feature Extraction

Deep learning features were extracted using a ResNet-50 convolutional neural network architecture pretrained on ImageNet. For each tumor, the axial slice demonstrating the maximal cross-sectional area (max-ROI) was identified and served

as input to the network. Transfer learning was implemented via fine-tuning on the training cohort images. Feature embeddings were subsequently derived from the activations of the penultimate fully connected layer (layer fc2), yielding a 320-dimensional deep learning feature vector per patient. This deep transfer learning (DTL) feature set was concatenated with radiomic features for downstream analysis.

Feature Selection and Dimensionality Reduction

All radiomic and deep learning features underwent z-score normalization prior to analysis. Initial feature screening employed Mann–Whitney U tests to identify features significantly associated with lymph node status ($P < 0.05$). To mitigate multicollinearity, pairwise Pearson correlation coefficients were computed; for feature pairs exhibiting $|r| > 0.90$, the feature with lower mean absolute correlation was retained. Subsequently, Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression with five-fold cross-validation was applied to select the most parsimonious and discriminative feature subset. The regularization parameter λ was optimized via minimum deviance criterion, and features with non-zero coefficients at optimal λ were retained for model construction.

Development of Deep Learning Radiomics Signature Model

Selected radiomic and deep learning features were integrated via feature-level fusion and input into five supervised machine learning classifiers: Logistic Regression, Random Forest, eXtreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), and Multilayer Perceptron. Models were developed using five-fold cross-validation on the training cohort, with hyperparameter tuning conducted via grid search. The Synthetic Minority Oversampling Technique (SMOTE) was applied to the training data to address class imbalance. Model discriminative performance was assessed using area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. The optimal classifier was selected based on cross-validated AUC and subsequently designated as the deep learning radiomics (DLR) signature model.

Construction of Clinical Model and Multimodal Fusion Nomogram

A clinical model was constructed using logistic regression incorporating Node-RADS score and other clinicopathological variables. Only variables demonstrating statistical significance in multivariate analysis ($P < 0.05$) were retained. To ensure methodological consistency with the DLR model, the identical XGBoost algorithm was applied following identical five-fold cross-validation framework. A multimodal fusion nomogram was subsequently developed by integrating three complementary feature domains: (1) selected radiomic features, (2) deep learning features, and (3) significant clinical variables including Node-RADS score. Model calibration was assessed using calibration curves and the Hosmer–Lemeshow goodness-of-fit test. Clinical utility was evaluated via decision curve analysis (DCA), quantifying net benefit across clinically relevant threshold probability ranges.

Model Interpretability

To elucidate the internal decision-making mechanisms of the optimal model, Shapley Additive Explanations (SHAP) analysis was implemented. SHAP, a game theory-based unified framework for interpretable machine learning, assigns each feature an importance value representing its marginal contribution to individual predictions. Both global feature importance rankings and local explanation plots were generated to enhance model transparency and facilitate clinical interpretability.

Statistical Analysis

Statistical analyses were performed using SPSS (version 25.0; IBM Corp) and Python (version 3.9.16; Python Software Foundation). Continuous variables were expressed as mean ± standard deviation or median with interquartile range, with between-group comparisons conducted using Student's t-test or Mann–Whitney U test as appropriate. Categorical variables were presented as frequencies and percentages, with comparisons performed using chi-square or Fisher's exact tests. Inter-model AUC comparisons were conducted using the DeLong test. All statistical tests were two-sided, with $P < 0.05$ considered statistically significant.

RESULTS

Patient Characteristics

A total of 892 patients from three independent institutions were included in this multicenter study. Table 1 summarizes the clinicopathological and radiological characteristics of the training (n=684, Center 1) and external validation (n=208, Centers 2 and 3) cohorts. Univariate logistic regression analysis identified Node-RADS score (including maximum LN short diameter) as a statistically significant predictor of lymph node metastasis ($P < 0.05$), warranting its inclusion in subsequent clinical model construction. Conversely, age, sex, carcinoembryonic antigen level, and neutrophil-to-lymphocyte ratio demonstrated no significant association with nodal status (all $P > 0.05$). Clinicopathological and semantic feature distributions were well-balanced between training and validation cohorts, with no statistically significant inter-cohort differences observed (Supplementary Data Table S3).

Feature Selection and Deep Learning Radiomics Model Development

From segmented tumor volumes, 1394 radiomic features were extracted via the PyRadiomics platform, alongside 320 deep learning features derived from a ResNet-50 convolutional neural network pretrained on maximal cross-sectional tumor slices. Least Absolute Shrinkage and Selection Operator (LASSO) regression with five-fold cross-validation identified 24 deep learning features and 9 radiomic features exhibiting non-zero coefficients, which were retained for subsequent modeling (Supplementary Data Fig.2A-2C). Deep learning features demonstrated substantially higher coefficient magnitudes relative to radiomic features, indicating their dominant contribution within the fused feature set.

Five machine learning algorithms—Logistic Regression, Random Forest, eXtreme Gradient Boosting (XGBoost), Adaptive Boosting, and Multilayer Perceptron—were systematically evaluated for deep learning radiomics (DLR) model construction. XGBoost outperformed all competing algorithms, achieving the highest

discriminative performance (AUC 0.900; Table 2). The final DLR model yielded AUCs of 0.876 (95%CI: 0.8505 - 0.9013) in the training cohort and 0.798 (95% CI: 0.7369 - 0.8589) in the external validation cohort (Supplementary Data Fig.2D). Decision curve analysis demonstrated superior net clinical benefit of the DLR model compared to the radiomics-only approach across a wide range of threshold probabilities (Supplementary Data Fig.2E).

Construction of Clinical Model and Multimodal Nomogram

The clinical model was developed using logistic regression incorporating Node-RADS score and other clinically relevant variables, with model selection based on statistical significance in multivariate analysis. Methodological consistency was maintained by applying the identical XGBoost algorithm employed in DLR model development. Subsequently, a multimodal fusion nomogram was constructed by integrating three complementary information domains: (1) selected radiomic features, (2) deep learning features, and (3) clinical variables including Node-RADS score.

Comparative Performance of Predictive Models

Table 3 presents the comparative diagnostic performance of the clinical, DLR, and multimodal fusion models. In the training cohort, the fusion model demonstrated the highest discriminative accuracy (AUC 0.900; 95% CI: 0.8773 - 0.9228), outperforming both the DLR model (AUC 0.876; 95% CI: 0.8505 - 0.9013) and the clinical model (AUC 0.751; 95% CI: 0.7146 - 0.7881) (Fig. 3A). This performance hierarchy was robustly reproduced in the external validation cohort, wherein the fusion model achieved an AUC of 0.870 (95% CI: 0.8225 - 0.9178), substantially exceeding those of the DLR model (AUC 0.798; 95% CI: 0.7369 - 0.8589) and clinical model (AUC 0.641; 95% CI: 0.5614 - 0.7201) (Fig. 3D).

Decision curve analysis revealed that the fusion model provided superior net clinical benefit relative to both comparator models across the majority of clinically relevant threshold probability ranges in both training and validation cohorts (Fig. 3B, 3E). Calibration curves demonstrated excellent agreement between predicted

probabilities and observed lymph node metastasis status for the fusion model, with Hosmer–Lemeshow tests yielding non-significant P values in both training and validation cohorts, indicating adequate goodness-of-fit (Fig. 3C, 3F). Pairwise DeLong tests confirmed that the fusion model's AUC was statistically significantly superior to both the clinical model and the DLR model in training and validation cohorts (all $P < 0.05$).

Based on the constructed multimodal fusion nomogram, we selected two cases with similar lymph node morphology but different metastatic status for lymph node status prediction (Fig. 4). This model integrates multiple variables, including NLR, age, gender, Node-RADS score and radiomics features, providing an intuitive and quantitative assessment of lymph node status for each patient. The nomogram not only achieves satisfactory predictive performance but can also be conveniently applied to individualized risk prediction and clinical management of patients with colorectal cancer.

Model Interpretability via SHAP Analysis

To elucidate the decision-making process of the optimal XGBoost-based fusion model, Shapley Additive Explanations (SHAP) analysis was performed. SHAP summary plots visualized both global feature importance and individual prediction contributions (Fig. 5). Among all integrated features, the deep learning feature DL_27—derived from the penultimate layer of the ResNet-50 architecture—emerged as the most influential contributor to lymph node metastasis prediction. The radiomic feature lbp_3D_m2_glszm_ZoneEntropy, a second-order texture descriptor quantifying heterogeneity within the gray-level size zone matrix, constituted the second most dominant predictive feature. These findings remained consistent across sensitivity analyses, underscoring the robustness of the identified feature hierarchy.

DISCUSSION

Accurate preoperative prediction of lymph node metastasis in colorectal cancer remains an unmet clinical imperative, with profound implications for neoadjuvant therapy selection, surgical planning, and prognostic stratification. This multicenter study demonstrates that a multimodal fusion model integrating deep learning radiomics signatures with Node-RADS scores achieves superior and externally validated predictive performance (AUC 0.900 in training, 0.870 in external validation), substantially outperforming both the DLR model alone (AUC 0.876/0.798) and the Node-RADS-based clinical model (AUC 0.751/0.641). To our knowledge, this investigation represents the first systematic integration of quantitative imaging biomarkers with structured Node-RADS assessment for CRC lymph node prediction, leveraging the largest multicenter cohort reported to date. These findings carry immediate translational relevance and address several critical gaps in the current literature.

The suboptimal performance of the standalone clinical model (AUC 0.641 in external validation) warrants careful examination. Recent investigations have reported widely discrepant estimates of CT-based Node-RADS diagnostic accuracy in colorectal cancer. While Niu et al. [14] demonstrated superior performance of Node-RADS over ESGAR criteria in rectal cancer (AUC 0.862), Leonhardi et al. [6] reported only limited diagnostic accuracy (AUC 0.68) with substantial inter-reader variability ($\kappa = 0.31$) in colon cancer patients, further noting false-negative rates exceeding 60% when applying conventional size thresholds, and corroborated these findings in a nationwide cohort, demonstrating radiological-pathological N-stage concordance of only 57%. Our findings align with the latter, corroborating that Node-RADS alone—despite its standardized framework—remains diagnostically insufficient for definitive clinical deployment. This apparent paradox is biologically plausible: Node-RADS, however systematically applied, remains tethered to macroscopic morphological features accessible to human visual perception (size, border characteristics, internal texture). Yet the histopathological reality of lymph node metastasis encompasses micrometastases (<2 mm) and isolated tumor cells that

manifest no macroscopic correlates, as well as reactive hyperplasia that faithfully mimics malignant morphology. The Node-RADS system, constrained by the fundamental limits of human perceptual capacity, cannot transcend this biological barrier. Our fusion model, by incorporating subvisual radiomic and deep learning features, effectively compensates for this inherent limitation—demonstrating that Node-RADS retains clinical utility not as a standalone arbiter but as a structured semantic scaffold that synergizes with quantitative imaging biomarkers.

The marked performance differential between our DLR model (AUC 0.798) and fusion model (AUC 0.870) in external validation substantiates the principle of diagnostic synergy. This observation resonates with emerging evidence from adjacent domains. Sun et al. [23], investigating MRI-based lateral lymph node prediction in rectal cancer, reported that fusion models integrating deep transfer learning and radiomics features outperformed individual approaches, though their models demonstrated performance degradation on external testing (AUC 0.713–0.741). Similarly, Yang et al.[20] developed a multiparametric MRI-based DLR model achieving AUC 0.878 in training but 0.752 upon external validation, with performance augmentation when deployed as a radiologist decision-support tool. Our study advances beyond these foundational contributions in three distinct dimensions. First, we demonstrate that the synergistic principle extends beyond combining radiomics with deep learning to encompass structured semantic assessment (Node-RADS)—a clinically deployable, guideline-endorsed scoring system. Second, our external validation cohort (n=208) substantially exceeds those reported in comparable studies (n=29–66) [23], enhancing generalizability estimates. Third, we identify the dominant predictive features (DL_27, lbp_3D_m2_glszm_ZoneEntropy) through rigorous SHAP-based interpretability analysis, transforming an opaque predictive algorithm into a transparent decision-support instrument.

The clinical translation of machine learning models has been persistently impeded by their intrinsic "black-box" nature, a barrier we explicitly addressed through systematic SHAP analysis. Zone Entropy—derived from gray-level size zone

matrix texture analysis—quantifies the heterogeneity of homogeneous intensity regions within tumors. Elevated Zone Entropy indicates greater variability in tumor texture patterns, a radiomic phenotype increasingly recognized as a surrogate marker for intratumoral genetic heterogeneity and aggressive biological behavior [17]. That this feature emerged as the dominant radiomic contributor to LNM prediction confers biological plausibility: tumors capable of establishing nodal metastases exhibit greater clonal diversity and architectural disorganization detectable at the subvisual imaging level. The deep learning feature DL_27, extracted from ResNet-50's penultimate layer, represents an abstract, high-dimensional representation imperceptible to human interpretation yet statistically robust across cross-validation folds. This complementary relationship—hand-crafted radiomic features encoding mathematically defined heterogeneity alongside autonomously learned deep features capturing emergent phenotypes—epitomizes the methodological rationale for multimodal integration. By rendering these contributions explicit at both global and individual prediction levels, SHAP analysis transforms model opacity into clinical transparency, addressing a documented regulatory and adoption barrier [17,24]. Yan et al. [25] recently demonstrated similar interpretability benefits in predicting colorectal liver metastasis, identifying texture features (LargeDependenceLowGrayLevelEmphasis) and clinical biomarkers as dominant contributors through SHAP analysis.

This study possesses several distinctive strengths. The three-center design with geographically independent external validation mitigates overfitting concerns that pervade single-institution radiomics investigations [20]. Our sample size (n=892) substantially exceeds recent comparable studies (n=108–310) [6,14,20,23], affording statistical power for robust feature selection and model calibration. The methodological decision to fuse features prior to machine learning classification, rather than ensembling model outputs, enabled the XGBoost algorithm to exploit cross-modal feature interactions—a technical nuance that likely contributed to our superior external validation performance. Furthermore, our exclusive focus on

venous-phase CT enhances clinical translatability, as this sequence is universally acquired in standard CRC staging protocols without requirement for specialized sequences or additional radiation exposure.

Several limitations warrant acknowledgment. First, despite multicenter enrollment, the retrospective design confers inherent selection bias; prospective validation is essential prior to clinical deployment. Second, our Node-RADS assessment was performed by a single radiologist with three years of experience; although intra- and inter-observer reproducibility was formally assessed (ICC ≥0.75), the substantial inter-reader variability documented in larger Node-RADS series (κ = 0.31–0.74) [14,26] suggests that generalizability across radiologists with varying expertise requires dedicated investigation. Third, our segmentation strategy employed single-slice maximal tumor area rather than volumetric whole-tumor annotation. While this approach reduces annotation burden and enhances reproducibility, emerging evidence suggests that volumetric radiomics may capture additional phenotypic information [20]. Fourth, our cohort exclusively comprised treatment-naïve patients; the model's performance in the neoadjuvant setting—increasingly relevant for locally advanced rectal cancer—remains undefined.

Several investigative avenues merit pursuit. External validation in prospective, multi-reader cohorts with diverse experience levels is necessary to establish Node-RADS reproducibility boundaries and define optimal integration strategies with quantitative biomarkers. Comparative effectiveness research directly comparing CT-based and MRI-based multimodal fusion models within the same patient cohort would inform modality selection guidelines. Radiogenomic correlation studies linking dominant predictive features (DL_27, Zone Entropy) to underlying transcriptomic programs—particularly those governing epithelial-mesenchymal transition, lymphangiogenesis, and immune evasion—could transform empirical imaging biomarkers into mechanistically grounded surrogate endpoints. Finally, deployment-ready clinical decision-support interfaces incorporating real-time SHAP

explanation visualization should be developed and subjected to human-factors evaluation.

Conclusion

This multicenter study demonstrates that an interpretable, CT-based multimodal fusion model integrating deep learning radiomics signatures with Node-RADS scores achieves robust, externally validated accuracy for preoperative lymph node metastasis prediction in colorectal cancer. By synergistically combining complementary information domains—subvisual quantitative phenotypes and structured semantic assessment—and operationalizing interpretability through SHAP analysis, this noninvasive tool addresses a critical translational gap and holds promise for enhancing precision in neoadjuvant therapy selection and surgical decision-making.

KEY POINTS

Question: Its multicenter external validation, large sample, and feature fusion mitigate overfitting, boosting generalizability; exclusive venous-phase CT focus ensures clinical translatability.

Findings: This multicenter study demonstrates that an interpretable, CT-based multimodal fusion model integrating deep learning radiomics signatures with Node-RADS scores achieves robust, externally validated accuracy for preoperative lymph node metastasis prediction in colorectal cancer.

Clinical Relevance Statement: Combining subvisual quantitative phenotypes and structured semantic assessments with SHAP-based interpretability, this noninvasive tool bridges a translational gap and may improve precision in neoadjuvant therapy selection and surgical decisions.

Abbreviations

LNM	lymph node metastasis
CRC	colorectal cancer
DLR	deep learning radiomics
ROC	convolutional neural network
AUC	area under the curve

LNM	lymph node metastasis
DCA	decision curve analysis
SHAP	shapley additive explanations
OLNM	ccult lymph node metastasis
TNT	total neoadjuvant therapy
ESGAR	European Society of Gastrointestinal and Abdominal Radiology
ROI	regions of interest
LNs	lymph nodes
SMOTE	Synthetic Minority Oversampling Technique
AJCC	american joint committee on cancer
CEA	carcinoembryonic antigen
NLR	Neutrophil-to-lymphocyte ratio
ICC	intraclass correlation coefficient
CNN	convolutional neural network
DTL	deep transfer learning
LASSO	least absolute shrinkage and selection operator
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
RQS	Radiomics Quality Score
LR	logistic regression
MLP	Multilayer Perceptron
RF	RandomForest
XGBoost	eXtreme Gradient Boosting
AdaBoost	adaptive boosting

DECLARATIONS

Ethics approval and consent to participate

Ethical approval for this retrospective study was derived from the institutional review board of the First People's Hospital of Foshan (Foshan Hospital Affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology (Approval No. 93, 2025).

Consent for publication

All authors have read and approved the final manuscript for publication.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available because of patient privacy restrictions and multicenter institutional agreements, but are available from the corresponding author on reasonable request,

subject to approval by the institutional review boards of the participating centers.

Conflict of Interest

The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Funding

This study was funded by the Medical Scientific Research Foundation of Guangdong Province (B2025808), Science and technology innovation project of Foshan (2420001004459) and Foshan 14th Five-Year Plan Key Discipline Foundation (No. FSGSP145036). The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author contributions

Conceptualization: ZF Xu, YJ Yang, TS Fang. Original draft: YX Ouyang, CL Zhu. Data curation: YX Ouyang, ZY Xu, KT Han, XM He, DL He, X Diao, CM Wang, TS Fang. Methodology: YX Ouyang, CL Zhu, ZY Xu. Formal analysis: CL Zhu. Validation: YX Ouyang, CL Zhu, ZY Xu, KT Han, XM He. Supervision: KT Han. Resources: DL He. Software: X Diao, CM Wang. review & editing: YJ Yang, ZF Xu. Funding acquisition: YJ Yang, ZF Xu. All authors have read and approved the manuscript.

Acknowledgements

Not applicable.

References

Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-249.

[2] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin. 2022;72(1):7-33.

[3] Benson AB, Venook AP, Al-Hawary MM, et al. Colon Cancer, Version 2.2021, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw. 2021;19(3):329-359.

[4] Argilés G, Tabernero J, Labianca R, et al. Localised colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2020;31(10):1291-1305.

[5] BBahadoer RR, Dijkstra EA, van Etten B, et al. Short-course radiotherapy followed by chemotherapy before total mesorectal excision (TME) versus preoperative chemoradiotherapy, TME, and optional adjuvant chemotherapy in locally advanced rectal cancer (RAPIDO): a randomised, open-label, phase 3 trial. Lancet Oncol. 2021;22(1):29-42.

[6] Leonhardi J, Mehdorn M, Stelzner S, et al. Diagnostic accuracy and reliability of CT-based Node-RADS for colon cancer. Abdom Radiol (NY). 2025;50(1):1-7.

[7] DDykstra MA, Gimon TI, Ronksley PE, Buie WD, MacLean AR. Classic and Novel Histopathologic Risk Factors for Lymph Node Metastasis in T1 Colorectal Cancer: A Systematic Review and Meta-analysis. Dis Colon Rectum. 2021;64(9):1139-1150.

[8] Maggialetti N, Greco CN, Lucarelli NM, et al. Applications of new radiological scores: the Node-rads in colon cancer staging. Radiol Med. 2023;128(11):1287-1295.

[9] Elsholtz FHJ, Asbach P, Haas M, et al. Introducing the Node Reporting and Data System 1.0 (Node-RADS): a concept for standardized assessment of lymph nodes in cancer. Eur Radiol. 2021;31(8):6116-6124.

[10] Leonardo C, Flammia RS, Lucciola S, et al. Performance of Node-RADS Scoring System for a Standardized Assessment of Regional Lymph Nodes in Bladder Cancer Patients. Cancers (Basel). 2023;15(3):580.

[11] Meyer HJ, Schnarkowski B, Pappisch J, et al. CT texture analysis and node-RADS CT score of mediastinal lymph nodes - diagnostic performance in lung cancer patients. Cancer Imaging. 2022;22(1):75.

[12] Wu Q, Lou J, Liu J, et al. Performance of node reporting and data system (node-RADS): a preliminary study in cervical cancer. BMC Med Imaging. 2024;24(1):28.

[13] Xu H, Zhao W, Guo W, et al. Prediction Model Combining Clinical and MR Data for Diagnosis of Lymph Node Metastasis in Patients With Rectal Cancer. J Magn Reson Imaging. 2021;53(3):874-883.

[14] Niu Y, Wen L, Yang Y, et al. Diagnostic performance of Node Reporting and Data System (Node-RADS) for assessing mesorectal lymph node in rectal cancer by CT. BMC Cancer. 2024;24(1):716.

[15] Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14(12):749-762.

[16] Chen W, Jia Z, Wan L, et al. Interpretable radiomics model based on dual-layer spectral CT iodine maps for predicting microsatellite instability in colorectal cancer: A two-center study. Eur J Radiol. 2025;192:112357.

[17] Wu Z, Gong L, Luo J, et al. An interpretable CT-based machine learning model for predicting recurrence risk in stage II colorectal cancer. Insights Imaging. 2025;16(1):162.

[18] Cheng Y, Yu Q, Meng W, Jiang W. Clinico-Radiologic Nomogram Using Multiphase CT to Predict Lymph Node Metastasis in Colon Cancer. Mol Imaging Biol. 2022;24(5):798-806.

[19] Eresen A, Li Y, Yang J, et al. Preoperative assessment of lymph node metastasis in Colon Cancer patients using machine learning: a pilot study. Cancer Imaging. 2020;20(1):30.

[20] Yang Y, Han K, Xu Z, et al. Development and Validation of Multiparametric MRI-based Interpretable Deep Learning Radiomics Fusion Model for Predicting Lymph Node Metastasis and Prognosis in Rectal Cancer: A Two-center Study. Acad Radiol. 2025;32(5):2642-2654.

[21] Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Comput Methods Programs Biomed. 2022;214:106584.

[22] Zhang H, Wang K, Liu S, et al. Development and validation of machine learning models for predicting no. 253 lymph node metastasis in left-sided colorectal cancer using clinical and CT-based radiomic features. Cancer Imaging. 2025;25(1):57.

[23] Sun Y, Lu Z, Yang H, et al. Prediction of lateral lymph node metastasis in rectal cancer patients based on MRI using clinical, deep transfer learning, radiomic, and fusion models. Front Oncol. 2024;14:1433190.

[24] Lundberg S.M. & Lee S.I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NIPS 2017) 30, 2017;4768-4777.

[25] Yan X, Duan F, Chen L, et al. A Multimodal MRI-Based Model for Colorectal Liver Metastasis Prediction: Integrating Radiomics, Deep Learning, and Clinical Features with SHAP Interpretation. Curr Oncol. 2025;32(8):431.

[26] Jiang L, Zhuang Z, Tang X, et al. Diagnostic performance of node-RADS classification for primary lymph node assessment in rectal cancer: a modality benchmarking study. J Cancer Res Clin Oncol. 2025;151(4):145.

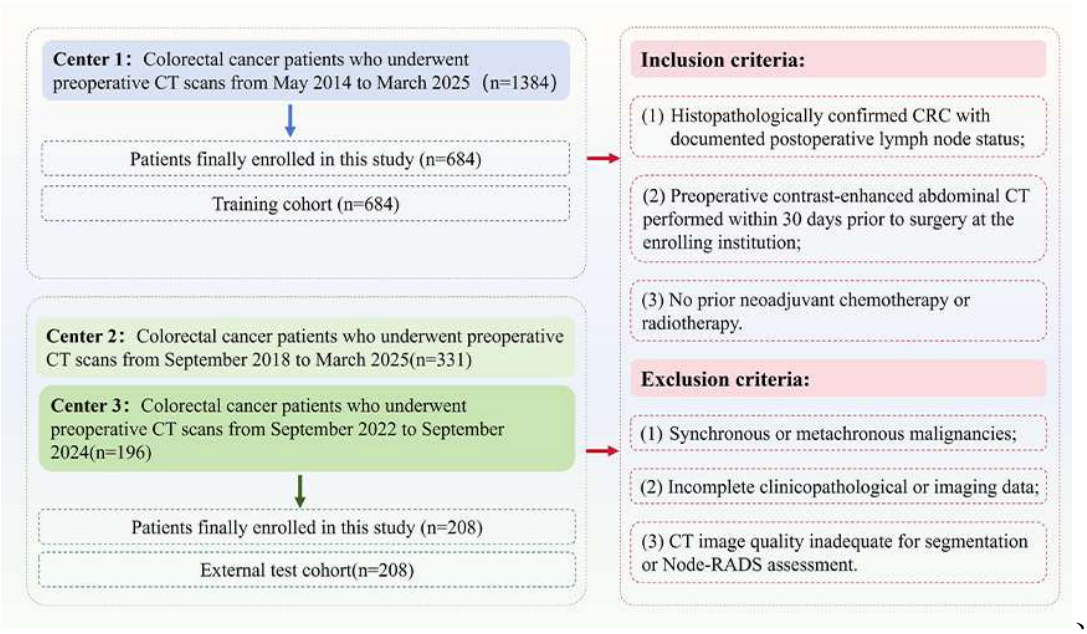


Fig.1: Flowchart of the study subjects based on inclusion and exclusion criteria.

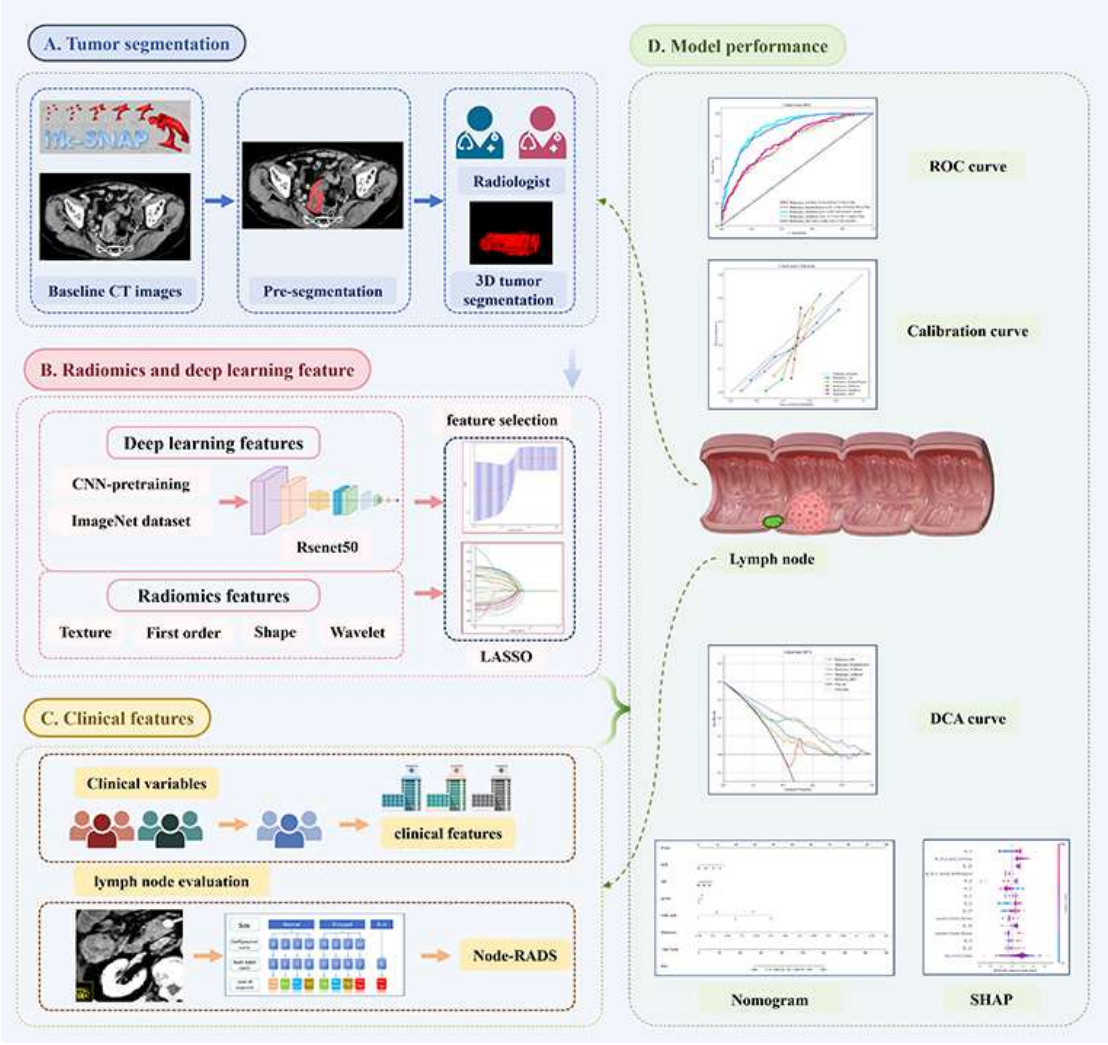


Fig.2: Workflow of radiomic analysis in this study.

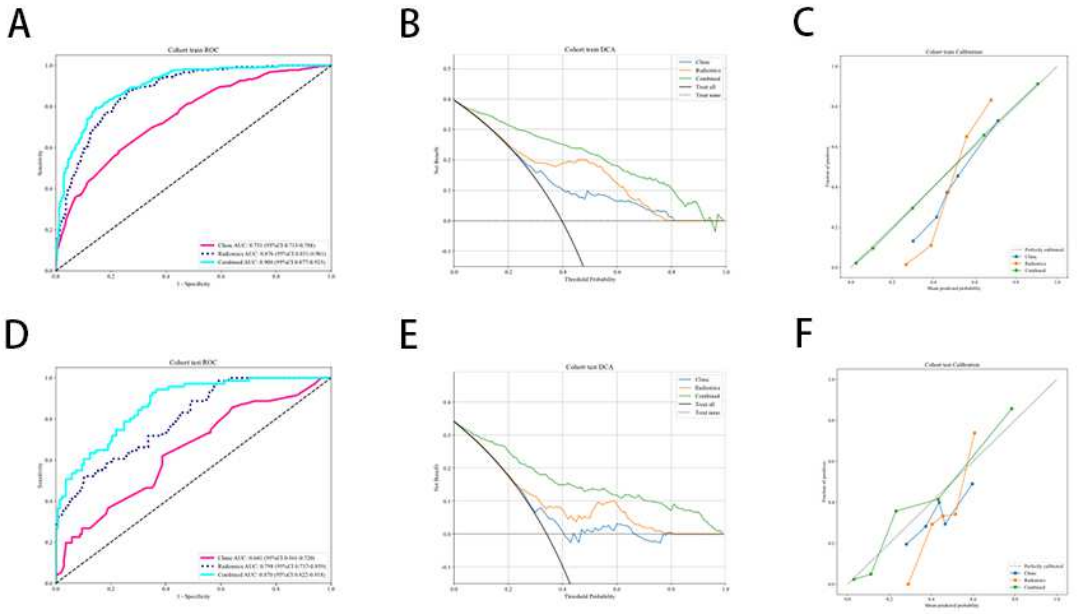


Fig.3: AUC, DCA and Calibration curves of clinical, DTL-radiomics, and nomogram models in the training cohort (A, B, C) and the testing cohort (D, E, F).

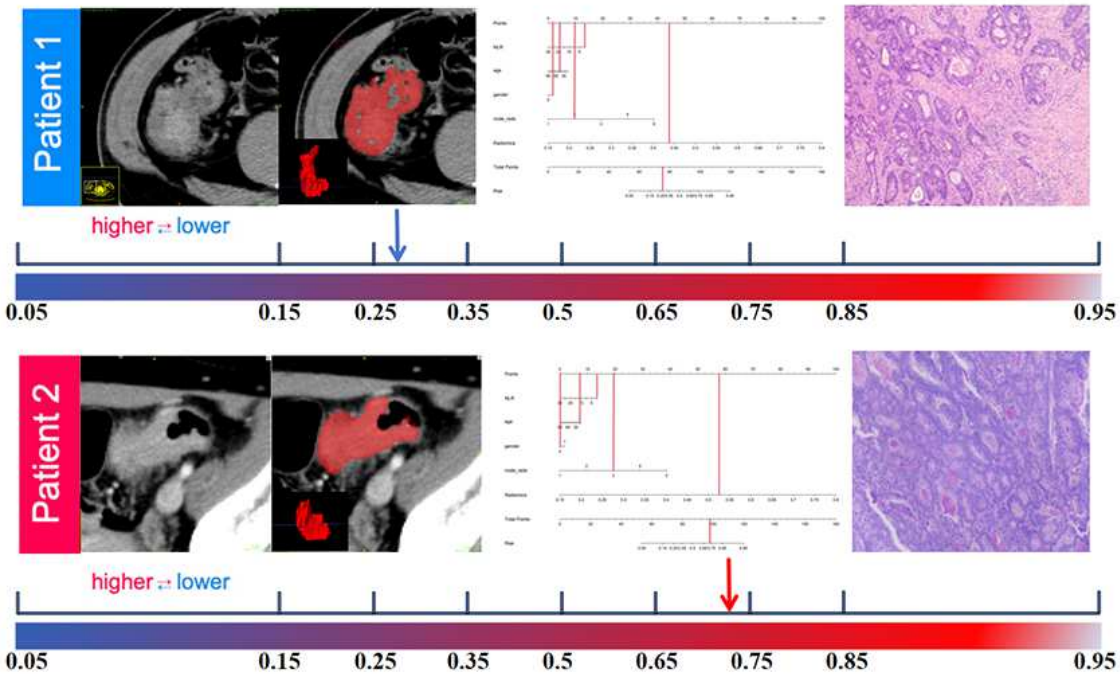


Fig.4: For Patient 1 (with benign lymph nodes) and Patient 2 (with malignant lymph nodes), the corresponding CT images, multimodal fusion nomogram and pathological images are shown. On the prediction maps, colors indicate the risk probability predicted by the nomogram, with blue representing low risk and red representing high risk. The radiomics features and Node-RADS score are the two variables with the greatest contribution (i.e., higher weight) to the final prediction.

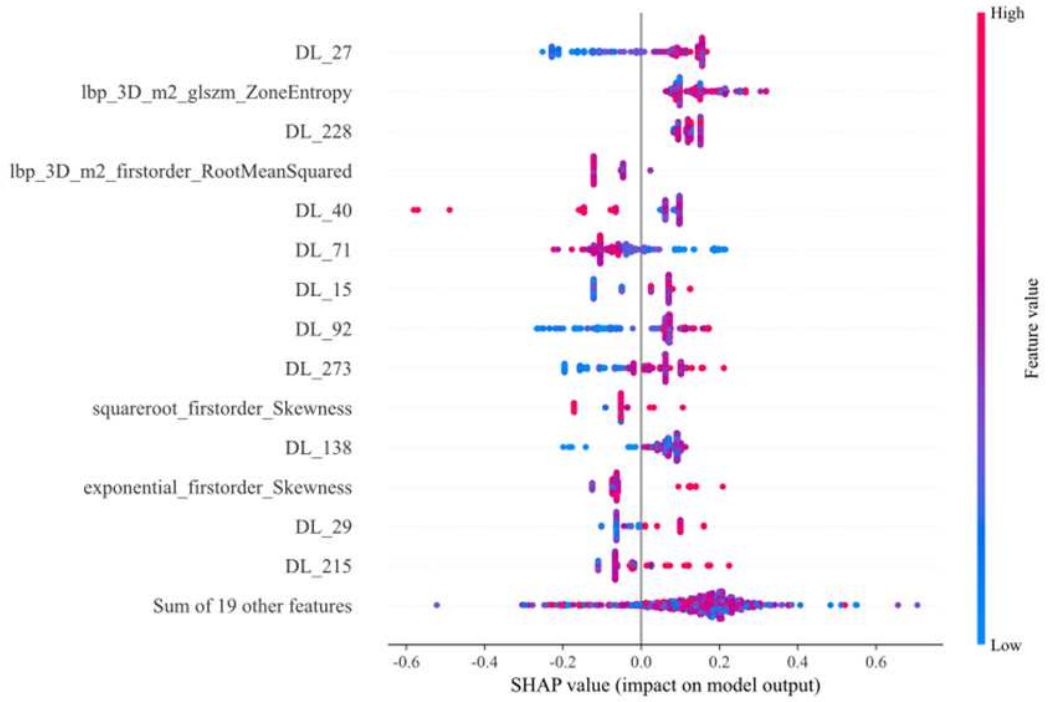


Fig.5: The SHAP plot visually analyzes the decision-making process of the optimal model XGBoost in predicting the lymph node metastasis status of colorectal cancer.

TABLE 1. Clinicopathological Characteristics of Patients in the Training cohort and External test cohort.

Variables	Training cohort(N = 684)			External validation cohort(N = 208)		
	pN- (n =413)	pN+ (n =271)	P value	pN- (n =137)	pN+ (n =71)	P value
NLR	2.28±3.64	1.73±3.35	0.005	5.91±8.59	4.55±3.65	0.477
Age	62.91±12.87	62.16±12.17	0.312	64.67±12.74	62.77±13.39	0.399
LNmax size	5.07±2.21	7.66±3.81	<0.001*	5.04±2.57	7.61±4.81	<0.001*
CEA			0.032			0.112
⩽5ng/ml	271(65.62)	155(57.20)		81(59.12)	33(46.48)	
> 5ng/ml	142(34.38)	116(42.80)		56(40.88)	38(53.52)	
Gender			1.0			1.0
female	182(44.07)	120(44.28)		67(48.91)	35(49.30)	
male	231(55.93)	151(55.72)		70(51.09)	36(50.70)	
Node rads			<0.001*			<0.001*
1	119(28.81)	30(11.07)		38(27.74)	7(9.86)	
2	213(51.57)	117(43.17)		77(56.20)	27(38.03)	
3	68(16.46)	76(28.04)		17(12.41)	21(29.58)	
4	13(3.15)	38(14.02)		4(2.92)	8(11.27)	
5	null	10(3.69)		1(0.73)	8(11.27)	

Training cohort：Colorectal cancer patients who underwent preoperative CT scans at Center 1；External validation cohort：Colorectal cancer patients who underwent preoperative CT scans at Center 2 and Center 3； pN-, negative lymph nodes; pN+, positive lymph nodes; LN, lymph node; LNmax size, maximum LN short diameter; NLR, Neutrophil count/lymphocyte count; CEA, carcinoembryonic antigen; Node-RADS, Lymph Node Reporting and Data System; *P < 0.05.

TABLE 2. The performance comparison of different machine learning models with deep learning and radiomics features.

Model		AUC (95% CI)	ACC	SEN	SPEC	PPV	NPV
LR	Training cohort	0.756(0.7201 - 0.7920)	0.696	0.679	0.707	0.603	0.770
	External validation cohort	0.637(0.5574 - 0.7159)	0.587	0.761	0.496	0.439	0.800
RF	Training cohort	0.741(0.7039 - 0.7784)	0.662	0.745	0.608	0.555	0.784
	External validation cohort	0.644(0.5625 - 0.7256)	0.663	0.577	0.708	0.506	0.764
XGBoost	Training cohort	0.876(0.8505 - 0.9013)	0.792	0.875	0.738	0.687	0.900
	External validation cohort	0.798(0.7369 - 0.8589)	0.769	0.521	0.898	0.725	0.783
AdaBoost	Training cohort	0.732(0.6939 - 0.7693)	0.677	0.613	0.719	0.589	0.739
	External validation cohort	0.653(0.5714 - 0.7346)	0.654	0.592	0.686	0.494	0.764
MLP	Training cohort	0.864(0.8380 - 0.8908)	0.785	0.793	0.780	0.703	0.852
	External validation cohort	0.627(0.5470 - 0.7077)	0.567	0.775	0.460	0.426	0.797

AUC, the area under the receiver operating characteristics curve; ACC, accuracy; SEN, sensitivity; SPEC, specificity; PPV, positive predictive value; NPV, negative predictive value; CI, confidence interval; LR, logistic regression; MLP, Multilayer Perceptron; RF RandomForest; XGBoost eXtreme Gradient Boosting; AdaBoost adaptive boosting.

TABLE 3. The performance comparison of different machine learning models with DLR and clinical features.

Model		AUC (95% CI)	ACC	SEN	SPEC	PPV	NPV
Clinic	Training cohort	0.751(0.7146 - 0.7881)	0.697	0.583	0.772	0.627	0.738
	External validation cohort	0.641(0.5614 - 0.7201)	0.615	0.620	0.613	0.454	0.757
DLR	Training cohort	0.876(0.8505 - 0.9013)	0.792	0.875	0.738	0.687	0.900
	External validation cohort	0.798(0.7369 - 0.8589)	0.769	0.521	0.898	0.725	0.783
Combined	Training cohort	0.900(0.8773 - 0.9228)	0.832	0.793	0.857	0.785	0.863
	External validation cohort	0.870(0.8225 - 0.9178)	0.745	0.930	0.650	0.579	0.947

AUC, the area under the receiver operating characteristics curve; ACC, accuracy; SEN, sensitivity; SPEC, specificity; PPV, positive predictive value; NPV, negative predictive value. CI, confidence interval.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx