npj Precision Oncology

https://doi.org/10.1038/s41698-026-01415-z

Article in Press

Development and validation of a multimodal AI-agent system for prognosis analysis of bladder urothelial carcinoma

Received: 20 August 2025

Accepted: 31 March 2026

Cite this article as: He, Q., Tan, H., Xiao, B. *et al.* Development and validation of a multimodal AI-agent system for prognosis analysis of bladder urothelial carcinoma. *npj Precis. Onc.* (2026). https://doi.org/10.1038/s41698-026-01415-z

Quanhao He, Hao Tan, Bangxin Xiao, Xiang Peng, Canjie Peng, Yiwen Tan, YingJia Liu, Youde Cao, Fa Jin Lv, Wenlong Zhao, Xiaofeng Yue, Weiyang He & Mingzhao Xiao

Title: Development and validation of a multimodal AI-agent system for prognosis analysis of bladder urothelial carcinoma

Quanhao He[1], Hao Tan[1], Bangxin Xiao[1], Xiang Peng[1], Canjie Peng[1], Yiwen Tan[2], YingJia Liu[3], Youde Cao[4,5,6], Fa Jin Lv[7], Wenlong Zhao[8,10], Xiaofeng Yue[9,10], Weiyang He[1,10], Mingzhao Xiao[1,10]

[1]Department of Urology, The First Affiliated Hospital of Chongqing Medical University, Chongqing 400016, peoples R China. [2]Department of Pathology, The Second Affiliated Hospital of Chongqing Medical University, Chongqing 400016, peoples R China. [3] Department of Pathology, Yongchuan Hospital of Chongqing Medical University, Chongqing 402177, China. [4] Department of Pathology, College of Basic Medicine, Chongqing Medical University, 1 Yixueyuan Road, Yuzhong Distinct, Chongqing, 400016, P.R. China. [5]Department of Clinical Pathololgy Laboratory of Pathology Diagnostic Center, Chongqing Medical University, 1 Yixueyuan Road, Yuzhong Distinct, Chongqing 400016, P.R. China. [6]Molecular Medicine Diagnostic and Testing Center, Chongqing Medical University,1 Yixueyuan Road, Yuzhong Distinct, Chongqing, 400016, P.R. China. [7]Department of Radiology, The First Affiliated Hospital of Chongqing Medical University, Chongqing 400016, peoples R China. [8]College of Artificial Intelligence Medicine, Chongqing Medical University, Chongqing, 400016, China. [9] Department of Urology, The Third Affiliated Hospital of Chongqing Medical University, Chongqing, 401120, China.[10]Correspondence: Wenlong Zhao(cqzhaowl@cqmu.edu.cn), Xiaofeng Yue(652282@hospital.cqmu.edu.cn), Weiyang He(weiyang361@stu.163.com), Mingzhao Xiao(mingzhaoxiao@hospital.cqmu.edu.cn).

Abstract

Precise survival risk stratification for bladder urothelial carcinoma (BUC) remains a clinical challenge. We developed and validated a multimodal AI agent that integrates textual, radiographic, and pathological data from 1185 patients across four medical centers to predict survival risk. The agent employs LLMs to standardize pathology reports, interactive deep learning networks for precise CT image segmentation, and extracts features from CT scans and whole slide images using CTVisionNet and MacroVisionNet. The multimodal fusion framework, MATCH-Net, integrates these features with microscopic pathology information and clinical text embeddings using a multi-head attention mechanism to generate a comprehensive prognostic score. In multi-center validation, MATCH-Net demonstrated robust performance (C-index ranging from 0.836 to 0.874) and effectively stratified patients into high-and low-risk groups, identifying potential candidates responsive to adjuvant chemotherapy. Furthermore, the framework enabled the quantification of novel, interpretable prognostic biomarkers and provides a reliable and clinically applicable solution for personalized BUC prognosis.

Keywords: bladder urothelial carcinoma, multimodal deep learning, prognosis prediction, biomarker exploration, AI agent

Introduction

Bladder urothelial carcinoma (BUC), a malignant neoplasm arising from urothelial cells, stands as one of the most prevalent cancers of the urinary tract[1,2]. Despite advances in therapeutic strategies, including intravesical chemotherapy, immunotherapy, and radical surgery, the five-year survival rate for BUC remains suboptimal and exhibits substantial inter-patient variability[1,3]. This clinical challenge underscores the urgent need for precise survival risk stratification to guide personalized diagnosis and treatment[4-6].

While preoperative computed tomography (CT) scans can provide critical anatomical details of tumor size and morphology, their interpretation relies heavily on the subjective assessment of radiologists[7]. Similarly, postoperative Whole Slide Images (WSIs) offer crucial microscopic insights, yet the manual evaluation of their complex features introduces inherent subjectivity. Conventional prognostic methods like TNM staging fail to fully capture the tumor's complex biological heterogeneity, leading to generalized patient management and suboptimal outcomes[8].

Recent advancements in computational radiology and quantitative pathology have presented promising alternatives to traditional approaches[9,10]. Deep learning algorithms have proven particularly valuable in enhancing diagnostic accuracy and prognostic prediction. In radiology, automated segmentation algorithms like the widely used nnU-Net and 3D U-Net often lack mechanisms for incorporating expert-guided constraints, which is critical for improving segmentation accuracy[11-13]. In pathology, WSI analysis is commonly framed as a Multiple Instance Learning (MIL) paradigm, exemplified by attention-based methods like AttMIL, Patch-GCN, and TransMIL[14-16]. More recently, foundation models such as CONCH and UNI have further enhanced feature extraction capabilities[17,18].

Despite these successes, existing approaches face significant limitations. Standard MIL models often lack integration with the prior knowledge embedded in textual pathology reports, complicating decision interpretability[19,20]. More crucially, conventional analytical workflows frequently suffer from dimensionality reduction. Traditional radiomics methods condense 3D volumes into global statistical descriptors, while MIL approaches aggregate local patch features into a 1D bag-level vector. Although advanced architectures like Patch-GCN and TransMIL effectively model inter-patch spatial correlations, their final aggregation step inherently compresses these features. Consequently, direct and visualizable information regarding the explicit macroscopic distribution of diverse tissue components is often lost[21,22]. This limitation not only complicates model interpretability but also hinders the discovery of quantifiable prognostic biomarkers linked to the tumor’s structural architecture. In contrast, emerging techniques like interactive segmentation

have demonstrated the effectiveness of incorporating user guidance to enhance accuracy[23,24]. Meanwhile, previous studies have indicated that macroscopic tissue distribution information can influence prognosis, and deep learning algorithms can assist in identifying measurable prognostic biomarkers[25,26].

Beyond these modality-specific challenges, a broader limitation in current prognostic modeling is the fragmentation of clinical data, failing to incorporate essential multimodal information. Consequently, there is a scarcity of AI-driven prognostic systems capable of executing a comprehensive data processing pipeline that fuses textual, pathological, and radiographic data for clinical applications[27,28].Addressing this gap, our research introduced a multimodal AI agent designed to integrate diverse prognostic information for a more thorough and accurate assessment of survival risk in BUC patients.

The agent follows a multi-stage process to perform its analysis. It begins by using Large Language Models (LLMs) to read and organize unstructured text from pathology reports into a consistent, structured layout. In the radiology component, the agent deploys an interactive Swin-UNETR network to locate BUC areas and employs a CTVisionNet to learn radiographic prognostic features. In the pathology component, it utilizes prior knowledge to decouple and sparsify WSIs into local and global slide representations. Specifically, a fully supervised patch classification network (BUCSegNet) generates a sparse tissue distribution heatmap, while a vision-language foundation model (CONCH) aligns pathological knowledge with tumor patch images to create a tumor phenotype heatmap[17]. These representations are then processed by MacroVisionNet to identify macro-level prognostic predictors.

Finally, the agent culminates its analysis in MATCH-Net, a multimodal deep learning framework. This framework integrates textual features from a pretrained ClinicalBERT, radiographic features from CTVisionNet, macroscopic features from MacroVisionNet, and microscopic features from the UNI self-supervised histopathology network[18]. The

fusion is guided by a Pre-Gated Multi-Head Contextual Gated Attention (PG-MHCA) mechanism, facilitating the generation of a final, unified prognostic prediction.

Results

Baseline characteristics

A comprehensive dataset consisting of preoperative CT scans, postoperative pathology images, and clinical data from 1185 individuals was aggregated from the First Affiliated Hospital of Chongqing Medical University (CQMFH), the Second Affiliated Hospital of Chongqing Medical University (CQMSH), the Third Affiliated Hospital of Chongqing Medical University (CQMTH), and Yongchuan Hospital of Chongqing Medical University (YCH). From the CQMFH cohort (December 1, 2012, to June 1, 2025), 564 patients were randomly allocated as the training cohort (CQMFH-T), and 243 patients were assigned as the internal validation cohort (CQMFH-V) based on a 7:3 ratio. 127 patients from CQMSH (May 1, 2013, to June 1, 2025), 101 patients from CQMTH (November 1, 2017, to June 1, 2025), and 150 patients from YCH (January 2, 2016, to June 1, 2025) were enrolled in the external validation cohorts (**Figure 1, Supplementary Figure 1**). The median follow-up periods with interquartile range (IQR) were 29.40 (11.28-60.42) months for the CQMFH-T cohort, 25.80 (8.20-62.10) months for the CQMFH-V cohort, 14.30 (6.09-27.39) months for the CQMSH cohort, 12.72 (7.85-52.79) months for the CQMTH cohort, and 23.10 (9.72-43.46) months for the YCH cohort (**Table 1**).

Interactive segmentation and CTVisionNet construction

The proposed framework employs an interactive segmentation algorithm based on geodesic distance transformation, which encodes user prior knowledge to generate an initial BUC lesion mask (**Figure 2, Supplementary Figure 2**). Among the evaluated architectures, Swin-UNETR demonstrated enhanced performance compared to 3D U-Net and UNETR, achieving initial Dice similarity coefficients (DSC) of 80.03%–81.75% (**Supplementary Table 1**). Crucially,

to maximize precision, we incorporated an interactive refinement procedure (**Supplementary Figure 5**); this optimization step improved the segmentation quality, elevating the final DSC scores to a range of 83.01%–84.32%. **Tables 2-3** present detailed segmentation metrics. Subsequently, we trained and validated the CTVisionNet prognostic model utilizing the processed BUC area via the 3D-cropbox. The C-index for CTVisionNet was 0.764 (0.711-0.816) in the CQMFH-T cohort, 0.707 (0.614-0.786) in the CQMFH-V cohort, 0.746 (0.601-0.877) in the CQMSH cohort, 0.791 (0.666-0.895) in the CQMTH cohort, and 0.701 (0.557-0.837) in the YCH cohort. After completing the CTVisionNet construction, we froze the trained weights to extract the radiographic feature vector $K_{radio} \in \mathbb{R}^{2048}$, which was then utilized to construct the multimodal prognostic system.

Pathological knowledge-guided representation and WSI decoupling

To decouple WSIs into discrete tissue distribution patterns, we employed BUCSegNet to classify image patches into eight distinct categories: tumor region, connective tissue, muscularis tissue, lymphovascular region, non-relevant region (non-ROI), adipose tissue region, empty region, and lymphocyte region. The Receiver Operating Characteristic (ROC) curve demonstrated that the AUC of the BUCSegNet patch classification network ranged from 0.9900 (95% CI: 0.9893–0.9908) to 0.9937 (95% CI: 0.9931–0.9942) across enrolled cohorts **(Supplementary Figure 11).** Following the construction of the BUCSegNet network, we employed it to infer all tile images, integrating their corresponding coordinates to generate a tissue probability heatmap, which functions as local knowledge-guided slide representation. Subsequently, we employed LLMs to summarize and extract key prognostic information relevant to BUC diagnosis as described in the pathology reports. Utilizing CONCH, a CLIP-based vision-language model, we designed prognostic information as text prompts and calculated the probability of each patch aligning with corresponding prompts to generate a tumor phenotype heatmap, which functioned as a global knowledge-guided slide representation[17].

Generation and evaluation of structured pathology reports from LLMs

To generate standardized, structured pathology reports from unstructured narrative medical texts, we evaluated the performance of several deployable LLMs using a dual-phase evaluation strategy. Our evaluation was two steps: first, we assessed information extraction accuracy by calculating the proportion of correctly identified parameters in a direct field-by-field comparison against the ground truth. Second, we evaluated text generation quality by comparing the complete AI-generated report to a golden standard report, which was programmatically generated from the ground truth labels using a deterministic, rule-based system. This textual comparison utilized ROUGE-L to measure lexical similarity and BERTScore to assess semantic equivalence, providing a comprehensive, multi-faceted assessment of the agent's ability to produce both accurate and coherent structured reports. Among the models tested, Gemma-27B and DeepSeek-70B demonstrated superior performance (Gemma-27B: BERTScore F1 0.998–0.999, ROUGE-L F1 0.961–0.983; DeepSeek-70B: BERTScore F1 consistent at 0.999, ROUGE-L F1 0.966–0.978). Detailed performance metrics for each model are presented in **Figure 3**. To numerically represent the clinical information contained within the structured pathology reports, we leveraged a pre-trained ClinicalBERT model to derive contextually rich semantic embeddings $K_{TextVision}$ for downstream prognostic tasks[29].

MacroVisionNet and MATCH-Net construction

By employing the concatenated knowledge-guided slide representation, we developed and verified MacroVisionNet to identify macroscopic prognostic information. The C-index value for MacroVisionNet was 0.813 (0.768-0.855) in the CQMFH-T cohort, 0.765 (0.655-0.864) in the CQMFH-V cohort, 0.779 (0.635-0.898) in the CQMSH cohort, 0.771 (0.646-0.876) in the CQMTH cohort, and 0.790 (0.664-0.905) in the YCH cohort. Following the establishment of MacroVisionNet, the parameters were fixed to retrieve macro feature vector $K_{MacroVision} \in \mathbb{R}^{2048}$, which was then incorporated with K_{radio} and $K_{TextVision}$ to form the final multimodal prognostic network. To build MATCH-Net, a self-supervised pretrained network (UNI) was applied to extract patch-level microscopic features ($K_{MicroVision}$) from

the corresponding WSIs. The architecture of MATCH-Net was intended to integrate K_{radio}, $K_{TextVision}$, $K_{MacroVision}$, and $K_{MicroVision}$ for generating final prognostic predictions. **Figure 2** illustrates the overall framework of MATCH-Net. The C-index value for MATCH-Net was 0.854 (0.819-0.889) in the CQMFH-T cohort, 0.846 (0.784-0.903) in the CQMFH-V cohort, 0.849 (0.750-0.930) in the CQMSH cohort, 0.874 (0.788-0.946) in the CQMTH cohort, and 0.836 (0.706-0.934) in the YCH cohort. MATCH-Net showed enhanced prognostic performance relative to CTVisionNet, MacroVisionNet, and various state-of-the-art models, such as TransMIL, AttMIL, and Patch-GCN[14-16]. **Figure 4** displays the time-dependent AUCs for CTVisionNet, MacroVisionNet, and MATCH-Net. Detailed ablation analysis results are provided in **Supplementary Tables 2-4 and Figure 5**.

Assessment of model risk stratification, prognosis prediction, and clinical utility.

Patients underwent stratification into high- and low-risk groups based on specific risk thresholds established from the CQMFH training cohort (CTVisionNet: -2.941; MacroVisionNet: -2.424; MATCH-Net: -3.258). In univariable analyses, all three frameworks (CTVisionNet, MacroVisionNet, and MATCH-Net) demonstrated robust and notable risk stratification capabilities across both training and validation cohorts. The high-risk groups consistently showed increased survival risks, with Hazard Ratios (HRs) ranging from 3.50 (95% CI: 1.85–6.65; $p < 0.001$) to 21.63 (95% CI: 7.11–65.86; $p < 0.001$) (**Figure 4**). To additionally evaluate the stratification performance of these models after accounting for clinical covariates, multivariable Cox regression analyses were conducted. After controlling for covariates, the HR values for CTVisionNet, MacroVisionNet, and MATCH-Net ranged from 0.60 (95% CI: 0.18–2.02; P=0.411) to 18.39 (95% CI: 4.54–74.50; $P < 0.001$) across all cohorts (**Tables 2-3**). Meanwhile, we evaluated the correlation between chemotherapy status and survival outcomes in both high- and low-risk groups (**Supplementary Figure 9**). For patients categorized as high-risk by MATCH-Net and MacroVisionNet, those who received ACT showed improved overall survival, with HRs value ranging from 0.12 (95% CI: 0.04–0.39; $p < 0.001$) to 0.46 (95% CI: 0.22–0.96; $p = 0.038$). In contrast, no

advantage was observed for patients in the low-risk group, with HRs value ranging from 0.27 (95% CI: 0.03–2.20; p = 0.223) to 0.58 (95% CI: 0.12–2.85; p = 0.504). These findings indicate that our model effectively identifies a subset of patients with aggressive tumor biology where ACT is correlated with improved outcomes.

To evaluate clinical utility and measure the added prognostic value beyond conventional clinical parameters, we contrasted a comprehensive clinical baseline model (incorporating age, sex, pT stage, pN stage, M stage, grade, and LVI status) with an integrated model supplemented by AI-derived risk scores. The addition of our prognostic risk scores enhanced the baseline model's predictive performance, resulting in a statistically robust increase in the C-index ranging from 0.034 (95% CI 0.004-0.064; p = 0.030) to 0.084 (95% CI 0.045-0.124; P < 0.001) across enrolled cohorts. Full performance metrics, including the C-index with 95% CIs, Decision Curve Analysis (DCA), calibration curves, and Net Reclassification Improvement (NRI) with 95% CIs, are provided in **Supplementary Tables 5-8** and **Supplementary Figures. 13-14**. Furthermore, subgroup analyses were performed for all models, with detailed HR values reported in **Supplementary Tables 9-11**.

BUC prognostic biomarker qualification

Based on our previous findings, we qualified several potential quantitative BUC prognostic biomarkers[26]. These biomarkers include Integrated Muscle Tumor Score (IMTS), Tumor Muscle Infiltration Fraction (TIM), Tumor-infiltrating Lymphocytes (TILs), Tumor Fraction Score (TFS), Inflammation Fraction Score (IFS), and Tumor Muscle Co-localization Score (Coloc_M). Comprehensive definitions of these prognostic biomarkers are presented in the **Supplementary Note 1**. By performing Cox regression and Kaplan-Meier analysis on the enrolled cohorts, we confirmed the prognostic reliability of these potential biomarkers. Both IMTS and Coloc_M exhibited consistent prognostic efficacy. The HR values in the Coloc_M and IMTS high-risk groups ranged from 3.12 (95% CI 1.64–5.96; p < 0.001) to 12.35 (5.02–30.38; p < 0.001) across the enrolled cohorts. Comprehensive Kaplan-Meier curves and associated HR values for

potential biomarkers are presented in **Supplementary Figure 12.**

Attention visualization and multimodal interpretability

In addition to achieving enhanced C-index and refined risk stratification performance, MATCH-Net delivers strong interpretability, offering insights into the synergy and individual contributions of different modalities to the terminal prognostic score. Specifically, we employed a Grad-CAM heatmap for radiographic images and a saliency attribution heatmap for macroscopic images [30]. We highlighted the contributions of formatted pathology reports via IG heatmaps. Additionally, the visualization of attention scores facilitated the demonstration of interactions between text, radiology, and pathology modalities. **Figures. 6-7** present the model interpretability analysis for the CQMFH-T and CQMFH-V cohorts, while the remaining analyses for other cohorts are presented in the **Supplementary Figures. 15-17.**

AI agent construction and clinical application

We developed an integrated clinical support system structured as a multi-agent framework. This system manages a set of specialized tools to perform tasks such as standardizing pathology reports, enabling interactive medical image segmentation, and conducting a multi-stage analysis for prognosis. User queries, submitted through a web interface, are processed by a dynamic routing engine that first applies keyword-based rules for efficiency and then leverages LLMs for more complex requests. Meanwhile, the framework incorporates essential safety features, including input guardrails and a human-in-the-loop validation mechanism, to ensure the reliability of its outputs in a clinical context. An example of the system's workflow and agent architecture is available in the **Supplementary Movie 1 and Supplementary Figure 10.**

Discussion

In the present research, we introduce an AI agent-based multimodal deep learning framework engineered to facilitate interactive lesion segmentation, standardize pathology reporting, quantify biomarkers, and predict prognosis for BUC

patients. The credibility and practical utility of this study are demonstrated through several key factors: (1) A robust cohort of 807 BUC patients from CQMFH supported the multimodal model construction and internal validation; (2) The transferability and clinical relevance of this prognostic framework were validated through its consistent performance across multiple large-scale medical centers; (3) By integrating textual, radiographic, macroscopic, and microscopic data via MATCH-Net, we achieved optimized and highly accurate risk stratification for BUC patients; (4) A novel interactive BUC segmentation agent was developed to enhance segmentation accuracy and efficiency; (5) We qualified several prospective BUC prognostic biomarkers and developed a pipeline to generate standardized pathology reports; (6) An autonomous end-to-end multimodal AI agent system was implemented based on this framework to improve its applicability in clinical practice.

Despite the swift evolution of medical image segmentation techniques, spanning from the classic U-net architecture to the SAM model, comparatively few studies have incorporated user interaction within the network to iteratively enhance segmentation results for optimal efficacy. Previous studies have indicated that geodesic-based U-Net architectures exhibit excellent segmentation results in both 2D and 3D medical imaging tasks. In the current study, we demonstrate that Swin-UNETR, leveraging the vision transformer framework, is more appropriate for interactive BUC segmentation tasks, surpassing the performance of 3D-Unet. The interactive Swin-UNETR emerges as an efficient tool for the precise annotation and refinement of BUC segmentation masks. The practical utility of this method is evidenced by its efficiency, requiring minimal user intervention (about 9–11 clicks for initialization and 3–6 for refinement). Furthermore, CTVisionNet exhibited predictive stability across varying levels of inputs. The performance remained consistent when transitioning from coarse 3D ROIs to refined radiologist-verified masks, which indicates that the model can effectively mitigate the impact of minor segmentation discrepancies frequently encountered in practical applications.

Although single-modality-trained models like CTVisionNet and MacroVisionNet have demonstrated robust risk differentiation performance across enrolled cohorts, the multi-modality integrated MATCH-Net exhibits even superior capabilities. In comparison to models solely incorporating text modality, radiology modality, and pathology modality, the fully integrated MATCH-Net consistently outperformed all other model versions. The results demonstrated the key significance and considerable advantages of utilizing multimodality data in enhancing predictive accuracy and model interpretability. We also observed that in the high-risk groups defined by both MacroVisionNet and MATCH-Net, patients who received adjuvant chemotherapy showed a valuable survival advantage. This association underscores the potential of our AI-agent system to identify patients with aggressive tumor biology who might derive the greatest benefit from intensified therapy, thereby informing postoperative treatment counseling. Nevertheless, as a retrospective analysis, the observed survival trends should be interpreted as evidence of the model's stratification utility rather than a definitive causal claim for treatment efficacy.

To further deconstruct the contributions of individual modalities to this superior performance, we performed a comprehensive set of ablation studies on MATCH-Net. MATCH-Net attains both enhanced robustness and greater interpretability by combining input across multiple modalities. The visualization of co-attention heatmaps revealed that textual, macroscopic, and radiographic features guide the model's attention to specific locations, emphasizing the complementary properties of the integrated modalities. At present, BUC risk stratification recommendations primarily focus on conventional methods, including pathological grade and staging. However, they do not incorporate WSI's tissue distribution information, such as tumor infiltration depth, width, or the proportion of immune cells. Building on our previous research, we further assessed the prognostic significance of biomarkers (e.g., Coloc_M, IMTS) across multi-institutional cohorts[26]. In contrast, performance variations observed in other candidate biomarkers highlight their

sensitivity to inter-site spatial heterogeneity, suggesting that these indices require further refinement in larger-scale studies.

Several shortcomings of this study require further consideration. First, the retrospective design of this study may include inherent selection bias. The positive performance of the multimodal prognostic approach in external validation cohorts indicates minimal bias. Nonetheless, additional

validation via prospective investigations in extensive, multicenter clinical trials is still required. The observed survival benefit of ACT treatment in high-risk patients may be affected by unmeasured confounders, including patient frailty or physician judgment, and the lack of adjustment for additional covariates. Consequently, our findings are hypothesis-generating and should not inform clinical ACT decisions. Second, practical hurdles in infrastructure and workflow currently limit immediate clinical integration. In current stage, the model necessitates considerable computational capacity and complicated data integration, which may not quickly accessible in conventional clinical settings. Third, prospective validation in real-world settings is an essential subsequent step. While we employed standardized simulations to mitigate the intrinsic uncertainty of interactive segmentation, it is essential to observe the system inside real clinical workflows to accurately evaluate its impact on physician autonomy and diagnostic efficiency. Finally, the proposed AI agent is conceived as a dynamic framework. Future versions will integrate genomic and spatial transcriptomic data to improve biological accuracy, while the integration of Retrieval-Augmented Generation (RAG) will allow the system to adapt dynamically to the most recent clinical guidelines, thereby maintaining continuous clinical relevance[31,32].

Methods

Ethics statement and consent to participate

The study was conducted in accordance with the Declaration of Helsinki. This retrospective study involved data from four centers: The First Affiliated Hospital of Chongqing Medical University (CQMFH), The Second Affiliated Hospital of Chongqing Medical University (CQMSH), The Third Affiliated Hospital of Chongqing Medical University (CQMTH), and Yongchuan Hospital of Chongqing Medical University (YCH). Ethical approval was granted by the Ethics Committee of The First Affiliated Hospital of Chongqing Medical University (Approval No. K2024-187-07), which was recognized by the institutional review boards of CQMSH, CQMTH, and YCH. The requirement for informed consent was waived by the ethics committees due to the retrospective nature of the study and the fact that all data were anonymized and de-identified prior to analysis.

Patient cohorts

This study comprised patients from four cohorts (CQMFH, CQMSH, CQMTH, and YCH) who possessed complete clinical information, CT scans, and pathology slides. The CQMFH data functioned as the development cohort for model training and validation, and the other cohorts (CQMSH, CQMTH, and YCH) were utilized for external validation to evaluate the model's robustness. Data for this retrospective analysis were obtained from medical records covering the period from December 1, 2012, to the end follow-up on June 1, 2025. Following ethical approval, baseline clinical, CT, and follow-up data were retrieved and verified from January to July 2025. Detailed data distribution and patient selection criteria can be found in **Supplementary Figure 1.**

CT image acquisition and ROI sketching procedure

Contrast-enhanced CT scans were acquired using diverse scanning equipment from multiple manufacturers, including Philips, Siemens, GE Healthcare, and Toshiba. Acquisition parameters were optimized as follows: tube voltage (90–120 kV), tube current (125–360 mAs with automatic dose modulation), reconstruction matrix (512×512 matrix), and slice thickness (1.0–5.0 mm). For prognostic analysis and segmentation, we employed images from the nephrographic phase,

obtained after intravenous injection of iohexol (300 mgI/mL) at a flow rate of 3.0 mL/s followed by a saline flush. To ensure the biological relevance of our baseline assessment, we selected the final scan performed prior to surgical resection for each patient, ensuring that no intervening therapy was administered during the interval between imaging and surgery. Three-dimensional regions of interest (ROIs) for model training were created by a strict multi-reader consensus approach. Initial tumor segmentations were manually delineated by two independent abdominal radiologists using ITK-SNAP (v.3.8.0), referencing synchronized axial, sagittal, and coronal views to ensure precise tumor boundary definition. Discrepancies among observers in the segmented volumes were resolved by a senior radiologist with over 20 years of expertise in multi-phase urological CT assessment, hence assuring a reliable ground truth for the CTVisionNet and interactive segmentation networks.

Interactive segmentation framework construction

The manual labeling of BUC lesions is a time-consuming and labor-intensive process. Although segmentation algorithms offer a solution, current solutions lack the interactivity needed for user-guided refinement. To overcome these limitations, we developed a deep learning-based interactive framework that facilitates rapid ROI delineation and allows for user-guided refinement. Our BUC interactive segmentation framework uses Swin-UNETR as its backbone due to its advanced performance in medical image segmentation, as demonstrated by comparisons in **Supplementary Table 1.** The U-shaped design of Swin-UNETR combines a CNN-based decoder with a Swin Transformer encoder, allowing for multi-resolution feature fusion through the use of skip connections[33]. This design, particularly the transformer-based encoder, is highly effective for complex 3D medical imaging tasks because of its ability to capture long-range dependencies (**Supplementary Figure 4).**

The segmentation procedure initiates with user interactions on CT images, including clicks or scribbles, to delineate the 3D borders of the BUC region. To reduce the necessity for additional user input, these interactions are encoded into

geodesic distance transform maps[23]. This method guarantees a balance between automation and precision, offering a more efficient and interactive means for users to enhance the segmentation results (**Supplementary Figure 2**). The geodesic distance transform steps are detailed below:

Let S_f and S_b denote the sets of pixels corresponding to foreground and background scribbles. For a given pixel i in the image I, the unsigned geodesic distance from i to a scribble set $S\left(S \in \left\{S_f, S_b\right\}\right)$ is given by:

$$G(i,S,I) = \min_{j \in S} D_{geo}(i,j,I) \tag{1}$$

$$D_{geo}(i,j,I) = \min_{p \in P_{i,j}} \int_0^1 \left| \nabla I\left(p(s)\right) \cdot u(s) \right| d_s \tag{2}$$

Let $P_{i,j}$ be the set of all paths between pixels i and j, with p representing a potential path parameterized by $s \in [0,1]$. $u(s)$ represents a unit vector that is tangent to the direction of the path. In the absence of user scribbles, a geodesic distance map is automatically generated with random integers. The interactive segmentation framework receives a three-channel input (the original image, user foreground interactions, and user background interactions) to produce an initial segmentation mask. Subsequently, the user can provide additional guidance signals, prompting the framework to iteratively refine the segmentation probability maps based on the new inputs.

To effectively encode these new interactions, we employ the proposed Exponential Geodesic Distance (EGD) transform to generate interaction-based hint maps. Let C_i^f denote the refined EGD foreground hint map, which is derived from the foreground term ($e^{-D_i^f}$) and the background term ($e^{-D_i^b}$) computed via the geodesic distance transform of user clicks. The value of C_i^f ranges from 0 to 1, representing the affinity between pixel i and the foreground interaction points. Additionally, let I_i^f denote the initial foreground probability map predicted by the Swin-UNETR.

Furthermore, we propose an information fusion strategy to refine the initial segmentation prediction using the interaction encoding map C^f. The objective is to design a fusion algorithm that rectifies the initial prediction I^f based on the geodesic proximity of pixel i to the refinement clicks. Intuitively, regions next to interaction points should be more

affected by user inputs than distant regions. Accordingly, we define the refined foreground probability R_i^f for pixel i as follows:

$$C_i^f = \frac{e^{-D_i^f}}{e^{-D_i^f} + e^{-D_i^b}} \tag{3}$$

$$\alpha_i = e^{-\min\left(D_i^f, D_i^b\right)} \tag{4}$$

$$R_i^f = \left(1-\alpha_i\right)*I_i^f + \alpha_i*C_i^f \tag{5}$$

where $\alpha_i \in [0,1]$ serves as an adaptive weighting factor. As pixel i approaches any refinement click, α_i tends toward 1.0, causing the refined probability R_i^f to be dominated by the interaction hint C_i^f. The entire interactive segmentation procedures are presented in **Supplementary Figure 5.** Total click count serves as a quantitative measure of model usability. As illustrated in **Supplementary Figure 7**, the majority of subjects required 9 to 11 clicks for initial segmentation and 3 to 6 clicks for refinement.

WSI file acquisition

For every individual in the training and validation cohorts, a representative H&E-stained tumor slide was selected. These slides, which contained comprehensive tumor and surrounding tissue information, were digitized at 40x magnification using four scanners: KF-PRO-020, KF-PRO-005-EX, and KF-PRO-040-HI (Jiangfeng BioInformation Technology, Ningbo, China), alongside the SQS-600P (Shengqiang Technology, Shenzhen, China). The corresponding pixel dimensions at the specimen level for these scanners were 0.246 μm, 0.246 μm, 0.252 μm, and 0.206 μm, respectively.

Tissue probability and segmentation heatmap generation

We adopted the BUCSegNet tile classification network to enable local pathological knowledge-guided slide representation (**Supplementary Figure 6**). This network generates tissue probability and segmentation heatmaps. Patches for training were annotated into eight regions: tumor, connective tissue, muscularis tissue, lymphovascular tissue,

non-ROI, adipose tissue, empty, and lymphocytes. To improve annotation efficiency and precision, we used the Segment Anything Model (SAM) in QuPath[34]. Two proficient pathologists independently labeled each tile, with a senior pathologist adjudicating any inconsistencies. The BUCSegNet feature extractor, $f_{BUCSegNet\_conv}$, is an adapted ResNeXt50 network, where the final fully connected layers $f_{BUCSegNet\_fc}$ are configured with an output dimension of 8[35]. For model inference, WSIs were first partitioned into tiles with registered coordinates. The OTSU algorithm was implemented to exclude non-tissue regions from tissue thumbnails. Patches of 256×256 pixels were then extracted at 20x magnification from the tissue-containing areas of these thumbnails. Let $p_l(x, y)$ be the local knowledge probability and $s_l(x, y)$ the local knowledge classification result for each input patch and its coordinates $i(x, y)$. These are defined as follows:

$$p_l(x, y) = softmax(f_{BUCSegNet\_fc}(f_{BUCSegNet\_conv}(i(x, y)))) \quad (6)$$

$$s_l(x, y) = argmax(p_l(x, y)) \quad (7)$$

The outputs $p_l(x, y)$ and $s_l(x, y)$ from continuous patch-level inference are aggregated to construct the corresponding local knowledge probability heatmaps P_l and local knowledge segmentation heatmaps S_l.

Tumor phenotype probability and segmentation heatmap generation

After generating the local slide representation, we obtained the coordinates and distribution of tumor tissue. While BUCSegNet differentiates tissue components, global pathological information (Tumor phenotype), such as tumor classification, grading, and invasion status, is often subject to the biases of subjective pathologist evaluations. Additionally, achieving accurate classification with fully supervised learning is challenging due to the difficulty of obtaining extensive and precise annotations. To address this, we used CONCH, a vision-language model leveraging contrastive learning on large-scale image-text pairs. CONCH's ability to recognize visual representations via textual prompts and its state-of-the-art performance in visual-language tasks make it suitable for our purpose.

To construct the global knowledge-driven representation, tumor-localized coordinates derived from the initial local analysis were first targeted. Utilizing the DeepZoomGenerator from OpenSlide, image patches were harvested from these locations and subsequently normalized to a 448×448 pixel format to meet CONCH's input specifications. The framework generated high-dimensional image and text embeddings for each tile. The computed patch-level probability distributions were derived from the cosine similarity with the logit scale application (**Supplementary Figure 6**).

Let v_i represent the image feature embeddings, v_t the text feature embeddings, and e^{logit_scale} a learnable scaling factor. For a patch and its specific coordinates $i(x, y)$, the global knowledge patch probability, $p_g(x, y)$, and the global knowledge classification result, $s_g(x, y)$, are defined as follows:

$$p_g(x, y) = softmax(v_i \times v_t^T \times e^{logit\_scale}) \quad (8)$$

$$s_g(x, y) = argmax(p_g(x, y)) \quad (9)$$

The continuous patch inference process combines $p_g(x, y)$ and $s_g(x, y)$ to create the corresponding tumor phenotype probability and segmentation heatmaps.

Generation of structured pathology reports

To address the challenge of transforming unstructured narrative pathology reports into a standardized, structured format suitable for clinical research and data aggregation, we developed and deployed a novel automated agent powered by LLMs. Our methodology is centered on a sophisticated two-stage prompting strategy, specifically designed to maximize both the accuracy of data extraction and the logical integrity of the final structured output. The first stage initiates the process by guiding the LLMs to perform a deep semantic analysis of the raw narrative text. In this phase, it is tasked not only with extracting explicit data points such as patient demographics and histological grade but also with the more complex inferential task of determining the correct pTNM stage from descriptive text, all based on a comprehensive set of predefined rules. This initial step is crucial for converting nuanced, non-standardized language into a preliminary

structured representation. To enhance the reliability of this output, the second stage of our strategy introduces a critical self-correction and refinement loop. Here, the LLM is prompted to meticulously review its own initial output, acting as an automated quality control agent. The primary purpose of this stage is to verify the internal medical logic and to autonomously correct any detected inconsistencies or errors, such as making sure the assigned T stage is entirely consistent with the textual description of tumor invasion. This two-step approach, combining initial interpretation with subsequent logical validation, was chosen to systematically improve the fidelity and accuracy of the structured data. The final, validated output is then used to programmatically generate a coherent, standardized report, thereby achieving our primary objective of creating a reliable, automated tool for medical data standardization.

Evaluation of structured pathology report accuracy

To rigorously evaluate our LLM-based agent, we benchmarked its performance against a manually curated ground truth dataset containing key clinical parameters, including patient age, gender, pathological TNM stage, histological grade, and lymphovascular invasion (LVI) status. The evaluation framework was designed to assess two critical aspects of performance: the accuracy of discrete information extraction and the quality of the final generated text. Our evaluation was two steps: First, we assessed information extraction accuracy by calculating the proportion of correctly identified parameters in a direct field-by-field comparison against the ground truth. Second, we evaluated text generation quality by comparing the complete AI-generated report to a golden standard report, which was programmatically generated from the ground truth labels using a deterministic, rule-based system. This textual comparison utilized ROUGE-L to measure lexical similarity and BERTScore to assess semantic equivalence, providing a comprehensive, multi-faceted assessment of the agent's ability to produce both accurate and coherent structured reports[36,37].

CTVisionNet construction procedure

The construction of CTVisionNet began with a 3D-cropbox to extract and fill ROIs. As depicted in **Supplementary**

Figure 3, this workflow comprises a three steps: ROI cropping, background filling, and area saving. To facilitate survival prediction, a 3D SEResNext50 network was deployed to derive prognostic risk scores from the processed BUC images. The CTVisionNet $f_{CTVision}$ integrates three distinct functional units: a CT-based image encoding module $f_{CTVision\_enco}$, a feature compression and stabilization module $f_{CTVision\_stab}$, and a prognostic prediction module f_{pred}. A 3D SEResNext50 feature extractor $f_{CTVision\_enco}$ maps the 3D-cropbox information R into a 2048-dimensional feature vector ($K_{radio} \in \mathbb{R}^{2048}$). To down-sample this vector to $S_{radio} \in \mathbb{R}^{32}$ while augmenting model resilience, the feature compression and stabilization module $f_{CTVision\_stab}$ was employed. This specific unit incorporates a fully connected layer integrated with batch normalization and a rectified linear unit. Finally, the prognosis risk score (RS_{radio}) was determined based on $S_{radio} \in \mathbb{R}^{32}$ through a fully connected layer (f_{pred}) optimized by a survival-specific loss function. The model equations are formulated as follows:

$$K_{radio} = f_{CTVision\_enco}(R) \tag{10}$$

$$S_{radio} = f_{CTVision\_stab}(K_{radio}) = ReLU(BN(FC(K_{radio}))) \tag{11}$$

$$RS_{radio} = f_{pred}(S_{radio}) \tag{12}$$

MacroVisionNet construction procedure

MacroVisionNet $f_{MacroVision}$ shares the three-part structure of CTVisionNet, including an encoding module ($f_{MacroVision\_enco}$), a feature compression and stabilization module ($f_{MacroVision\_stab}$), and a prediction module (f_{pred}). For survival prediction from WSI macroscopic tissue information, we used a ResNeXt50 network to encode the concatenated local and global slide representations (P) into a 2048-dimensional feature vector. To accommodate the channel number of P, we adjusted the input channel count of ResNeXt50 to 16. The $f_{MacroVision\_stab}$ module further compressed the encoded macro feature vector $K_{MacroVision} \in \mathbb{R}^{2048}$ to $S_{MacroVision} \in \mathbb{R}^{32}$. The final patient-level risk score $RS_{MacroVision}$ was then derived from $S_{MacroVision}$ using f_{pred}. Detailed model equations are as follows:

$$K_{MacroVision} = f_{MacroVision\_enco}(P) \tag{13}$$

$$S_{MacroVision} = f_{MacroVision\_stab}(K_{MacroVision}) = ReLU(BN(FC(K_{MacroVision}))) \tag{14}$$

$$RS_{MacroVision} = f_{pred}(S_{MacroVision}) \tag{15}$$

MATCH-Net construction procedure

At present, most methodologies for multimodal integration utilize late-fusion strategies. However, substantial divergence exists among various data modalities, such as pathology images, radiographic features, and unstructured clinical text, which presents a substantial obstacle to their effective integration. Late-fusion paradigms restrict the capacity for profound, cross-modal interactions, thereby diminishing the interpretability of prognostic models. To uncover explainable correlations between microscopic, macroscopic, radiographic, and textual elements, we conceptualized a weakly supervised, multimodal deep learning framework: the Multimodal Pre-Gated Attention Transformer with Contextual Hierarchies (MATCH-Net). This architecture leverages a Pre-Gated Multi-Head Contextual Gated Attention (PG-MHCA) mechanism to effectively integrate these diverse data sources.

Informed by co-attention transformer architectures that capture dependencies between heterogeneous data types, our model learns the interactions between high-dimensional microscopic histological features and the more structured radiographic, macroscopic, and semantic textual features to estimate prognostic outcomes [22]. We first employ a self-supervised learning model (UNI) to extract patch-level microscopic features $K_{MicroVision}$ from WSIs[18]. Concurrently, we process narrative clinical reports using a pre-trained ClinicalBERT model to generate powerful semantic embeddings $K_{TextVision}$. These features, along with radiographic K_{radio} and macroscopic $K_{MacroVision}$ features, are subsequently passed to their respective encoder layers.

Subsequently, the encoded radiographic, macroscopic, and textual features are used as queries to guide the extraction of relevant information from the vast feature space of the WSIs. To effectively suppress noise and emphasize task-relevant

regions within the gigapixel WSI prior to complex interaction, we introduce a Pre-Gated Attention (PGA) mechanism. The PGA computes a gating vector based on the relevance between the modality query and WSI patches, which is then used to re-weight the WSI features.

$$\mathrm{G} = \sigma\left((K_{Mod} W_Q^p)(K_{MicroVision} W_K^p)^T\right) \tag{16}$$

$$K_{MicroVision} = K_{MicroVision} \odot \mathrm{G}^T \tag{17}$$

$K_{mod} \in \{K_{radio}, K_{MacroVision}, K_{TextVision}\}$ represents the features of a specific modality. The PGA generates a gating mask G and refines the microscopic features $K_{MicroVision}$. σ represents the sigmoid function. W_Q^p and W_K^p are learnable projection matrices for the pre-gating mechanism.

These re-weighted features $K_{MicroVision}$ then serve as the keys and values for the PG-MHCA mechanism which is further enhanced with a Contextual Attention Gate (CAG) to dynamically adjust the attention output based on contextual information[38,39]. The formulation of the proposed mechanism is defined as follows:

$$Q = K_{Mod} W_q, K = K_{MicroVision} W_k, V = K_{MicroVision} W_v \tag{18}$$

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{19}$$

$$MultiHead(Q,K,V) = Concat(h_1,\ldots,h_H)W_O \tag{20}$$

$$c = CAG\left(Q, MultiHead(Q,K,V)\right) \tag{21}$$

$$Output = MultiHead(Q,K,V) + c \tag{22}$$

Following the co-attention step, the features from each modality are processed by their respective transformer encoders. These encoders comprise several layers of multi-head self-attention, allowing the model to capture complex intra-modal dependencies. After this encoding, the transformed features are passed through gated attention heads to produce pooled representations that summarize the most salient information from each modality. Finally, the pooled embeddings from the micro-pathology, macro-pathology, radiology, and text modalities are fused into a unified feature vector, which is

passed to a linear classifier to produce the final prediction. To assess the relative weight of each modality, we calculated the L2 norm of the respective feature vectors during the forward inference pass. The variations in prognostic contribution scores across all study cohorts are depicted in **Supplementary Figure 8.** Thorough analyses model dimension optimization, the impact of various fusion strategies, and ablation studies are provided in **Supplementary Table 2.**

Interpretability framework and visual explanations.

To enhance the transparency of the inference process within our deep learning models, we leveraged Grad-CAM, saliency maps, and attention score maps to elucidate and visualize the underlying mechanisms of CTVisionNet, MacroVisionNet, and MATCH-Net. Specifically, Grad-CAM was utilized to clarify the relevance of feature maps within CTVisionNet by analyzing gradients in the final convolutional layer. Saliency attribution heatmaps were derived by computing the MacroVisionNet risk gradient relative to the input pixels. To probe the interplay between radiographic, textual, macroscopic, and microscopic attributes, we extracted PG-MHCA co-attention scores alongside their spatial WSI coordinates. These scores were projected onto corresponding pathological thumbnail images to generate co-attention score heatmaps. Concurrently, we utilized integrated gradients (IG) to assess the significance of textual feature properties[40]. We illustrated the weighted contributions of structured pathology reports using IG heatmaps. In this context, positive attributions within the structured reports drive an increase in the output value (indicating high risk), while negative attributions serve to decrease the output value (indicating low risk).

Loss function definition

The interactive segmentation model was trained using the Dice loss function, while the BUCSegNet patch classifier was trained with cross-entropy loss. We chose not to employ the standard negative Cox partial log-likelihood loss for our survival tasks. This decision was made because its mini-batch dependency poses a notable challenge when working with heterogeneous pathology WSIs, where microscopic patch features vary greatly between slides. Instead, for the

CTVisionNet, MacroVisionNet, and MATCH-Net survival models, we adopted a category-based negative log-likelihood loss function as detailed in reference[41]. We divided the continuous overall survival time T_s into four distinct intervals: $[T_0, T_1]$, $[T_1, T_2]$, $[T_2, T_3]$, and $[T_3, T_4]$, according to the uncensored patient's survival time quartiles in the training set. Subsequently, we discretized the survival time T_i for the $i-th$ patient into one of these intervals.

$$T_i = n \, if \, T_s \in \left[T_n, T_{n+1}\right) \tag{23}$$

Leveraging the patient's ultimate integrated representation M_{final}, the prediction layer computed the hazard function M_h and the survival function M_s, which are defined as:

$$M_h\left(n \middle| M_{final}\right) = P\left(T_i = n \middle| T_i \geq n, M_{final}\right) \tag{24}$$

$$\begin{aligned} M_s\left(n \middle| M_{final}\right) &= P\left(T_i > n \middle| M_{final}\right) \\ &= \prod_{s=1}^{n}\left(1 - M_h\left(s \middle| M_{final}\right)\right) \end{aligned} \tag{25}$$

The discrete survival log-likelihood objective can consequently be expressed as:

$$\begin{aligned} L_i = &-c_i \, log\left(M_s\left(Y_i \middle| M_{final}\right)\right) \\ &-\left(1-c_i\right) log\left(M_s\left(Y_i - 1 \middle| M_{final}\right)\right) \\ &-\left(1-c_i\right) log\left(M_h\left(Y_i \middle| M_{final}\right)\right) \end{aligned} \tag{26}$$

In this context, Y_i denotes the ground-truth label for the i-th individual, whereas ci represents the binary censoring indicator (ci =1 if the patient was censored at the end of the follow-up period, and ci =0, otherwise). To ensure the impactful contribution of non-censored patient data, we employed a weighted sum to derive the final training loss.

Statistical analysis

Overall survival (OS) served as the primary clinical endpoint, measured as the time interval from surgical resection to

all-cause mortality or the final follow-up census. We rigorously evaluated the model's discriminative performance using the concordance index (C-index) and time-dependent area under the receiver operating characteristic curves (AUC) at 1, 3, and 5-year intervals. To avoid arbitrary thresholds, we employed maximally selected rank statistics on the training cohort to identify the optimal cut-off for stratifying patients into high- and low-risk groups. Survival distributions were estimated via the Kaplan-Meier method and compared using the log-rank test, while hazard ratios (HRs) were derived from Cox proportional hazards regression to quantify risk status. Beyond discrimination, we assessed calibration accuracy through calibration plots and evaluated clinical utility via Decision Curve Analysis (DCA). Crucially, to quantify the incremental value of our multimodal agent over a clinical-only baseline, we computed the Net Reclassification Improvement (NRI) and C-index improvement. The statistical robustness of these comparative metrics was validated using stratified bootstrap resampling (1,000 iterations). All analyses were two-sided, with statistical significance set at $P < 0.05$.

Software and Development Environment

The technical framework for this study was implemented using a suite of specialized Python libraries. The AI agent's backend architecture was developed utilizing LangGraph (v0.4.10), with the corresponding frontend interface constructed via FastAPI (v0.115.13). We leveraged Ollama for the local management and deployment of the large language models integral to our agent. In the domain of medical image analysis, the MONAI framework (v1.3) was employed for image processing tasks, while the OpenSlide library (v1.1.2) was specifically used for pathology slide manipulation. All model training and validation were conducted within a Python (v3.12.6) environment built on PyTorch (v2.0.0). All computational tasks were carried out on a high-performance workstation featuring a 48 GB NVIDIA A6000 GPU.

References

1 Nadal, R., Valderrama, B. P. & Bellmunt, J. Progress in systemic therapy for advanced-stage urothelial carcinoma.

Nature Reviews Clinical Oncology **21**, 8-27, doi:10.1038/s41571-023-00826-2 (2024).

2 Hemenway, G. *et al.* Advancements in Urothelial Cancer Care: Optimizing Treatment for Your Patient. *American Society of Clinical Oncology Educational Book* **44**, e432054, doi:10.1200/EDBK_432054 (2024).

3 Lopez-Beltran, A., Cookson, M. S., Guercio, B. J. & Cheng, L. Advances in diagnosis and treatment of bladder cancer. *BMJ (Clinical research ed.)* **384**, e076743, doi:10.1136/bmj-2023-076743 (2024).

4 Soualhi, A. *et al.* The incidence and prevalence of upper tract urothelial carcinoma: a systematic review. *BMC Urology* **21**, 110, doi:10.1186/s12894-021-00876-7 (2021).

5 Compérat, E. *et al.* Current best practice for bladder cancer: a narrative review of diagnostics and treatments. *Lancet (London, England)* **400**, 1712-1721, doi:10.1016/s0140-6736(22)01188-6 (2022).

6 Shen, J., Li, Z., Wang, R., Ding, G. & Zhang, Y. Bladder cancer diagnostic and prognostic models from DNA methylation by multi algorithm machine learning. *NPJ precision oncology*, doi:10.1038/s41698-025-01195-y (2025).

7 Raman, S. P. & Fishman, E. K. Bladder Malignancies on CT: The Underrated Role of CT in Diagnosis. **203**, 347-354, doi:10.2214/ajr.13.12021 (2014).

8 Compérat, E. *et al.* Updated pathology reporting standards for bladder cancer: biopsies, transurethral resections and radical cystectomies. *World journal of urology* **40**, 915-927, doi:10.1007/s00345-021-03831-1 (2022).

9 Perez-Lopez, R., Ghaffari Laleh, N., Mahmood, F. & Kather, J. N. A guide to artificial intelligence for cancer researchers. *Nature Reviews Cancer*, doi:10.1038/s41568-024-00694-7 (2024).

10 McGenity, C. *et al.* Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. *npj Digital Medicine* **7**, 114, doi:10.1038/s41746-024-01106-8 (2024).

11 He, J. *et al.* Development of a deep learning model for T1N0 gastric cancer diagnosis using 2.5D radiomic data in preoperative CT images. *NPJ precision oncology* **9**, 249, doi:10.1038/s41698-025-01055-9 (2025).

12 Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **18**, 203-211, doi:10.1038/s41592-020-01008-z (2021).

13 Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016.* (eds Sebastien Ourselin *et al.*) 424-432 (Springer International Publishing).

14 Shao, Z. *et al.* Transmil: Transformer based correlated multiple instance learning for whole slide image classification. **34**, 2136-2147 (2021).

15 Chen, R. J. *et al.* in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24.* 339-349 (Springer).

16 Saldanha, O. L. *et al.* Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology. *NPJ precision oncology* **7**, 35, doi:10.1038/s41698-023-00365-0 (2023).

17 Lu, M. Y. *et al.* A visual-language foundation model for computational pathology. *Nature Medicine* **30**, 863-874, doi:10.1038/s41591-024-02856-4 (2024).

18 Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nature Medicine* **30**, 850-862, doi:10.1038/s41591-024-02857-3 (2024).

19 Yuan, Y. *et al.* Cell graph analysis in hepatocellular carcinoma: predicting local recurrence and identifying spatial relationship biomarkers. *NPJ precision oncology* **9**, 261, doi:10.1038/s41698-025-01042-0 (2025).

20 Pisula, J. I. *et al.* Explainable, federated deep learning model predicts disease progression risk of cutaneous squamous cell carcinoma. *NPJ precision oncology* **9**, 205, doi:10.1038/s41698-025-00997-4 (2025).

21 Lu, M. Y. *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering* **5**, 555-570, doi:10.1038/s41551-020-00682-w (2021).

22 Chen, R. J. *et al.* Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* **40**,

865-878.e866, doi:https://doi.org/10.1016/j.ccell.2022.07.004 (2022).

23 Wang, G. *et al.* DeepIGeoS: a deep interactive geodesic framework for medical image segmentation. **41**, 1559-1572 (2018).

24 Luo, X. *et al.* MIDeepSeg: Minimally interactive segmentation of unseen objects from medical images using deep learning. **72**, 102102 (2021).

25 Liang, J. *et al.* Deep learning supported discovery of biomarkers for clinical prognosis of liver cancer. *Nature Machine Intelligence* **5**, 408-420, doi:10.1038/s42256-023-00635-3 (2023).

26 He, Q. *et al.* Integrated multicenter deep learning system for prognostic prediction in bladder cancer. *NPJ precision oncology* **8**, 233, doi:10.1038/s41698-024-00731-6 (2024).

27 Lee, Y., Ferber, D., Rood, J. E., Regev, A. & Kather, J. N. How AI agents will change cancer research and oncology. *Nature Cancer* **5**, 1765-1767, doi:10.1038/s43018-024-00861-7 (2024).

28 Ferber, D. *et al.* Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nature Cancer*, doi:10.1038/s43018-025-00991-6 (2025).

29 Huang, K., Altosaar, J. & Ranganath, R. J. a. p. a. Clinicalbert: Modeling clinical notes and predicting hospital readmission. (2019).

30 Selvaraju, R. R. *et al.* Grad-CAM: visual explanations from deep networks via gradient-based localization. **128**, 336-359 (2020).

31 Suzuki, H. *et al.* Exploring drug resistance via intercellular crosstalk using spatial transcriptomics in high-grade serous ovarian carcinoma. *NPJ precision oncology* **9**, 345, doi:10.1038/s41698-025-01122-1 (2025).

32 Li, T. *et al.* Computational pathology annotation enhances the resolution and interpretation of breast cancer spatial transcriptomics data. *NPJ precision oncology* **9**, 310, doi:10.1038/s41698-025-01104-3 (2025).

33 Hatamizadeh, A. *et al.* in *International MICCAI brainlesion workshop.* 272-284 (Springer).

34 Kirillov, A. *et al.* in *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 4015-4026.

35 Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 5987-5995.

36 Ganesan, K. J. a. p. a. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. (2018).

37 Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. J. a. p. a. Bertscore: Evaluating text generation with bert. (2019).

38 Chen, R. J. *et al.* in *2021 IEEE/CVF International Conference on Computer Vision (ICCV).* 3995-4005.

39 Chen, R. J. *et al.* Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. *IEEE transactions on medical imaging* **41**, 757-770, doi:10.1109/tmi.2020.3021387 (2022).

40 Kokhlikyan, N. *et al.* Captum: A unified and generic model interpretability library for pytorch. (2020).

41 Zadeh, S. G., Schmid, M. J. I. t. o. p. a. & intelligence, m. Bias in cross-entropy-based training of deep survival networks. **43**, 3126-3137 (2020).

Data availability

The WSIs, nephrographic CT scans, and annotation data used for both the training and validation sets are subject to institutional restrictions. Due to patient privacy obligations and Institutional Review Board (IRB) approvals, these data

are not publicly available. However, they can be accessed upon reasonable request from the corresponding author, pending approval from the IRBs and legal departments of all participating centers.

Code availability

The source code is available online (https://github.com/hqh1997/MMS_AI_agent).

Acknowledgements

We acknowledge the support from the Medical Health Care Ecosystem Innovation Team of the First Affiliated Hospital of Chongqing Medical University (CYYY-DSTDXM-202409), the Postgraduate Education Reform Project of the First Affiliated Hospital of Chongqing Medical University (jgxm-202501), and the Chongqing Municipal Education Commission's 14th Five-Year Key Discipline Support Project (No. 20240101). We thank all pathologists, radiologists, and related staff at the participating institutions for their assistance in data collection. Computing work was partly supported by the Supercomputing Center of Chongqing Medical University.

Author contributions

QHH, HT, BXX, YWT, XP, WYH, and MZX conceived and designed the study; HT, CJP, XFY, XP, and XZ collected the data. QHH, HT, and CJP evaluated images. YJL, YWT, and DYC labeled the pathological slide images. FJL supervised and annotated the radiographic images. QHH, WLZ, and XP trained and developed the AI system. QHH, BXX, and YWT analyzed and interpreted the data and wrote the original draft of the manuscript. QHH and XFY were responsible for revising the manuscript and performing supplementary experiments. WLZ, XFY, WYH, and MZX supervised and directed the study.

Competing interests

The authors declare that no competing interests exist.

Figure 1. A schematic overview of the multimodal prognostic AI agent for bladder urothelial carcinoma (BUC).

Our proposed AI agent system is built by integrating several trained modules to achieve four primary capabilities: interactive lesion segmentation, automated generation of standardized pathology reports, WSI decoupling for biomarker qualification, and survival risk assessment based on multimodal data.

Figure 2. Detailed structure of the multimodal prognostic model. MATCH-Net integrates features from multiple modalities for prognosis. The model extracts image features from CTVisionNet, macroscopic pathological features from MacroVisionNet, and textual features from ClinicalBERT. These features are fused with microscopic pathological features extracted from a self-supervised pre-trained network via a Pre-Gated Multi-Head Contextual Gated Attention (PG-MHCA) mechanism. Following this, a fusion module processes the combined features to ultimately provide a patient prognosis assessment. PGA, Pre-gated Attention; CAG, Contextual Attention Gate; SSL, self-supervised learning.

Figure 3. Comparative evaluation of LLM-generated structured pathology reports across multiple cohorts. The performance of various Large Language Models (LLMs) in producing standardized, structured pathology reports is assessed using both classification-based and NLP metrics. (A–C) Results for the CQMFH cohort. (D–F) Results for the CQMSH cohort. (G–I) Results for the CQMTH cohort. (J–L) Results for the YCH cohort. Within each row, the heatmaps (left columns) display classification performance across different clinical categories, with numerical values and color intensity representing the scores for each model. The violin plots (middle and right columns) illustrate the distribution of ROUGE-L F1 and BERTScore F1 scores, respectively. CQMFH, The First Affiliated Hospital of Chongqing Medical University; CQMSH, The Second Affiliated Hospital of Chongqing Medical University; CQMTH, The Third Hospital of Chongqing Medical University; YCH, Yongchuan Hospital of Chongqing Medical University; LLM, Large Language Model; NLP, Natural Language Processing.

Figure 4. Performance comparison and ablation experiment results of multimodal survival prediction models across the enrolled cohorts. (A–C) Time-dependent AUCs for (A) CTVisionNet, (B) MacroVisionNet, and (C) MATCH-Net across all enrolled cohorts. (D) Radar chart of ablation results for MATCH-Net. (E) Radar chart comparing MATCH-Net against established

state-of-the-art (SOTA) deep learning models. PGA, Pre-gated Attention; CAG, Contextual Attention Gate; WSI only, model trained solely on pathology image features; Without CAG, MATCH-Net with CAG module removed; Without PGA, MATCH-Net with PGA module removed; Without macroscopic feature, MATCH-Net with macro modality removed; Without radiographic feature, MATCH-Net with radio modality removed; Without textual feature, MATCH-Net with text modality removed. Concat and Gate-concat describe the fusion strategy used to combine features from different modalities. Small, medium, and large define the internal embedding dimension for all feature encoders and fusion layers within the model, corresponding to feature sizes of 128, 256, and 512, respectively. CQMFH-T, The First Affiliated Hospital of Chongqing Medical University (Training cohort); CQMFH-V, CQMFH internal validation cohort; CQMSH, The Second Affiliated Hospital of Chongqing Medical University; CQMTH, The Third Hospital of Chongqing Medical University; YCH, Yongchuan Hospital of Chongqing Medical University.

Figure 5. Kaplan–Meier survival analysis across the enrolled cohorts. Kaplan-Meier survival analysis was performed for CTVisionNet (A-E), MacroVisionNet (F-J), and MATCH-Net (K-O) across the enrolled cohorts. Survival differences between groups were assessed using the two-sided log-rank test, and HRs with 95% CIs were derived from univariable Cox proportional hazards regression. HR, Hazard Ratio; CI, Confidence Interval; CQMFH-T, The First Affiliated Hospital of Chongqing Medical University (Training cohort); CQMFH-V, CQMFH internal validation cohort; CQMSH, The Second Affiliated Hospital of Chongqing Medical University; CQMTH, The Third Affiliated Hospital of Chongqing Medical University; YCH, Yongchuan Hospital of Chongqing Medical University.

Figure 6. Interpretability analyses of the multimodal AI agent system in the CQMFH training cohort. The Grad-CAM and saliency attribution heatmaps highlight the regions of focus for CTVisionNet and MacroVisionNet. In the Co-Attention heatmaps, the highlighted areas represent regions of interest guided by their respective modalities. The highlighted text in the structured pathology reports indicates the specific textual descriptions that the multimodal model focused on. (A) Low-risk prediction example. (B) High-risk prediction example. CQMFH, The First Affiliated Hospital of Chongqing Medical University.

Figure 7. Interpretability analyses of the multimodal AI agent system in the CQMFH validation cohort. The Grad-CAM and saliency attribution heatmaps highlight the regions of focus for CTVisionNet and MacroVisionNet. In the Co-Attention heatmaps, the highlighted areas represent regions of interest guided by their respective modalities. The highlighted text in the structured pathology reports indicates the specific textual descriptions that the multimodal model focused on. (A) Low-risk prediction example. (B) High-risk prediction example. CQMFH, The First Affiliated Hospital of Chongqing Medical University.

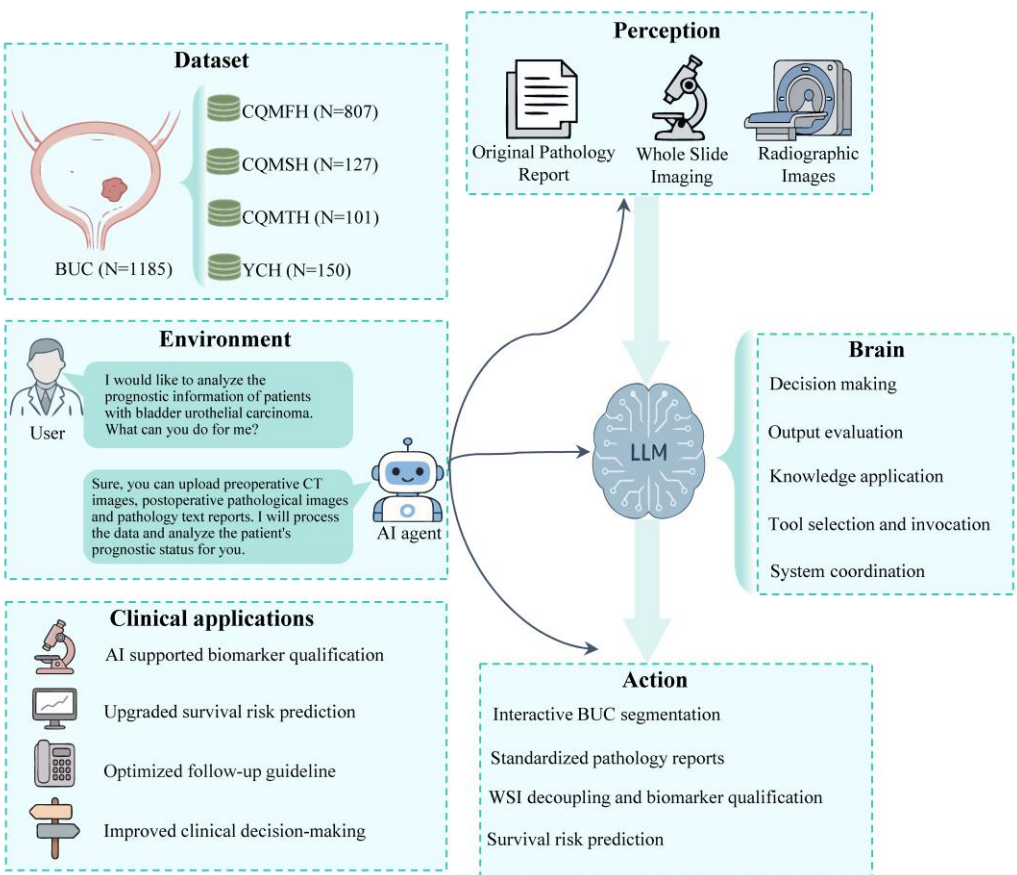

Figure1 A schematic overview of the multimodal prognostic AI agent for bladder urothelial carcinoma (BUC). Our proposed AI agent system is built by integrating several trained modules to achieve four primary capabilities: interactive lesion segmentation, automated generation of standardized pathology reports, WSI decoupling for biomarker qualification, and survival risk assessment based on multimodal data.

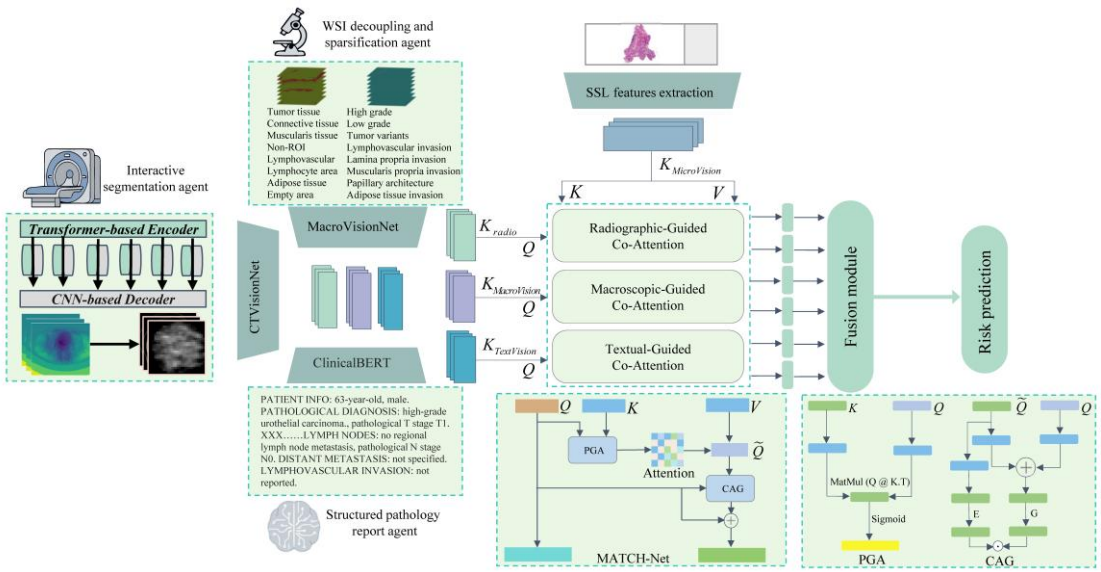

Figure 2. Detailed structure of the multimodal prognostic model. MATCH-Net integrates features from multiple modalities for prognosis. The model extracts image features from CTVisionNet, macroscopic pathological features from MacroVisionNet, and textual features from ClinicalBERT. These features are fused with microscopic pathological features extracted from a self-supervised pre-trained network via a Pre-Gated Multi-Head Contextual Gated Attention (PG-MHCA) mechanism. Following this, a fusion module processes the combined features to ultimately provide a patient prognosis assessment. PGA, Pre-gated Attention; CAG, Contextual Attention Gate; SSL, self-supervised learning.

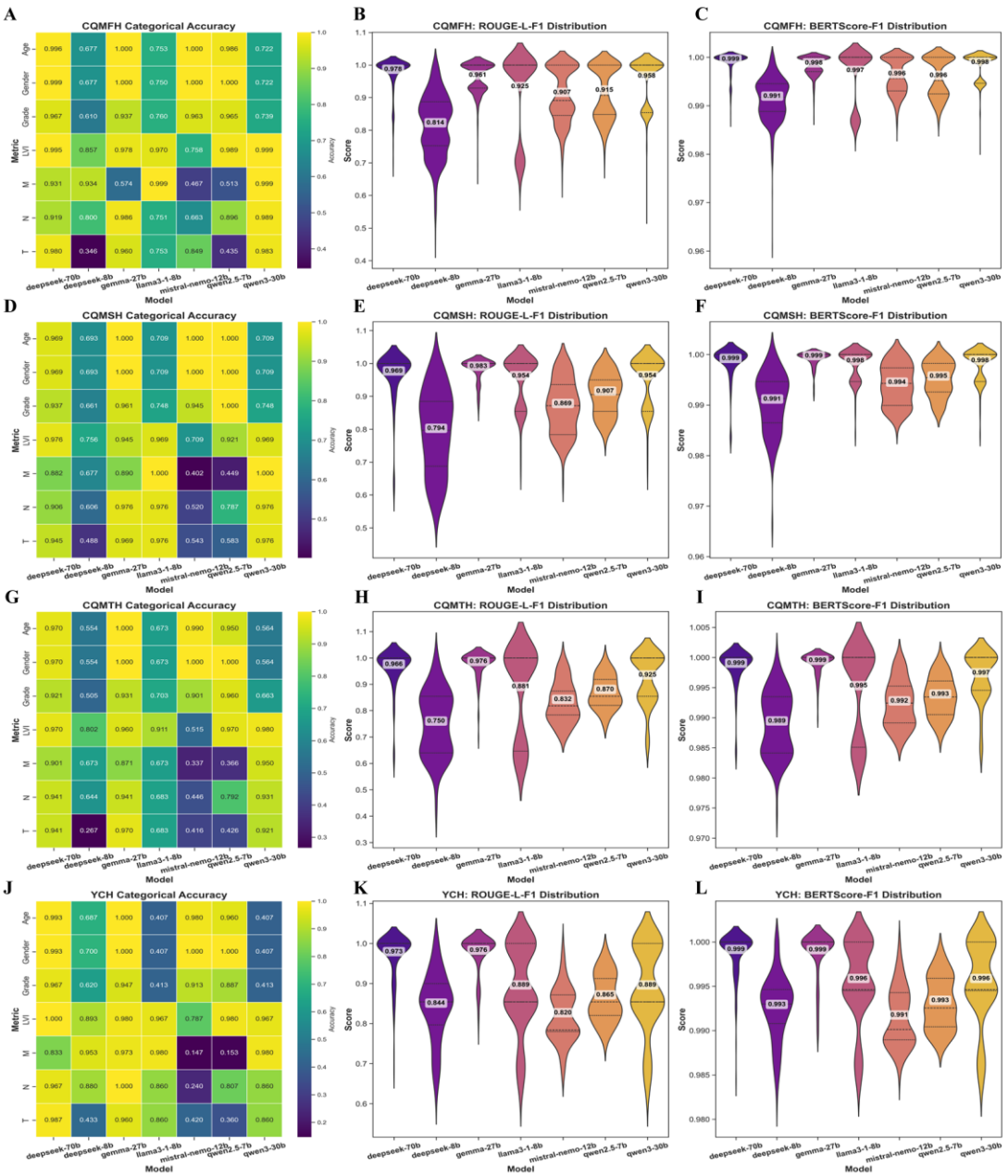

Figure 3. Comparison of the generated standardized, structured pathology reports across different LLMs. The performance of various Large Language Models (LLMs) in producing standardized, structured pathology reports is assessed using both classification-based and NLP metrics. (A–C) Results for the CQMFH cohort. (D–F) Results for the CQMSH cohort. (G–I) Results for the CQMTH cohort. (J–L) Results for the YCH cohort. Within each row, the heatmaps (left columns) display classification performance across different clinical categories, with numerical values and color intensity representing the scores for each model. The violin plots (middle and right columns) illustrate the distribution of ROUGE-L F1 and BERTScore F1 scores, respectively. CQMFH, The First Affiliated Hospital of Chongqing Medical University; CQMSH, The Second Affiliated Hospital of Chongqing Medical University; CQMTH, The Third Hospital of Chongqing Medical University; YCH, Yongchuan Hospital of Chongqing Medical University; LLM, Large Language Model; NLP, Natural Language Processing.

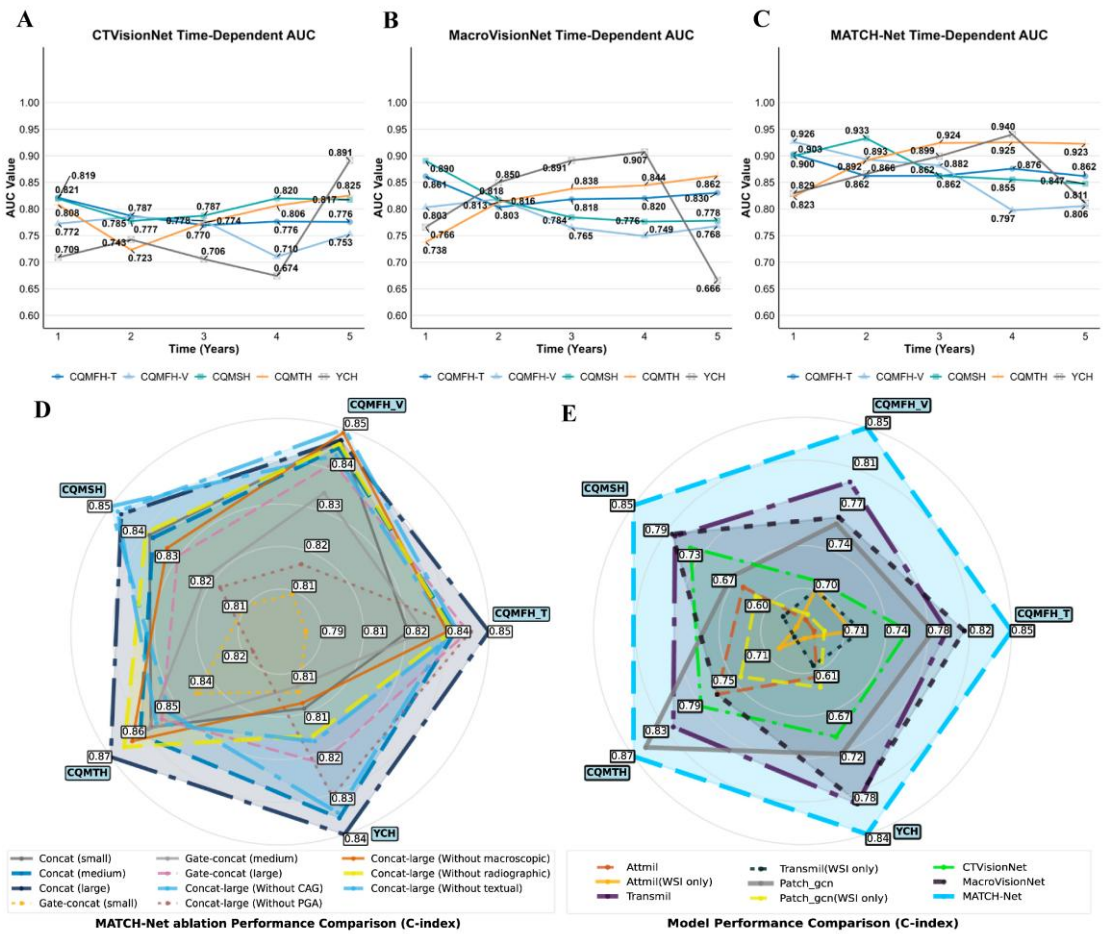

Figure 4. Performance comparison and ablation experiment results of multimodal survival prediction models across the enrolled cohorts. (A–C) Time-dependent AUCs for (A) CTVisionNet, (B) MacroVisionNet, and (C) MATCH-Net across all enrolled cohorts. (D) Radar chart of ablation results for MATCH-Net. (E) Radar chart comparing MATCH-Net against established state-of-the-art (SOTA) deep learning models. PGA, Pre-gated Attention; CAG, Contextual Attention Gate; WSI only, model trained solely on pathology image features; Without CAG, MATCH-Net with CAG module removed; Without PGA, MATCH-Net with PGA module removed; Without macroscopic feature, MATCH-Net with macro modality removed; Without radiographic feature, MATCH-Net with radio modality removed; Without textual feature, MATCH-Net with text modality removed. Concat and Gate-concat describe the fusion strategy used to combine features from different modalities. Small, medium, and large define the internal embedding dimension for all feature encoders and fusion layers within the model, corresponding to feature sizes of 128, 256, and 512, respectively. CQMFH-T, The First Affiliated Hospital of Chongqing Medical University (Training cohort); CQMFH-V, CQMFH internal validation cohort; CQMSH, The Second Affiliated Hospital of Chongqing Medical University; CQMTH, The Third Hospital of Chongqing Medical University; YCH, Yongchuan Hospital of Chongqing Medical University.

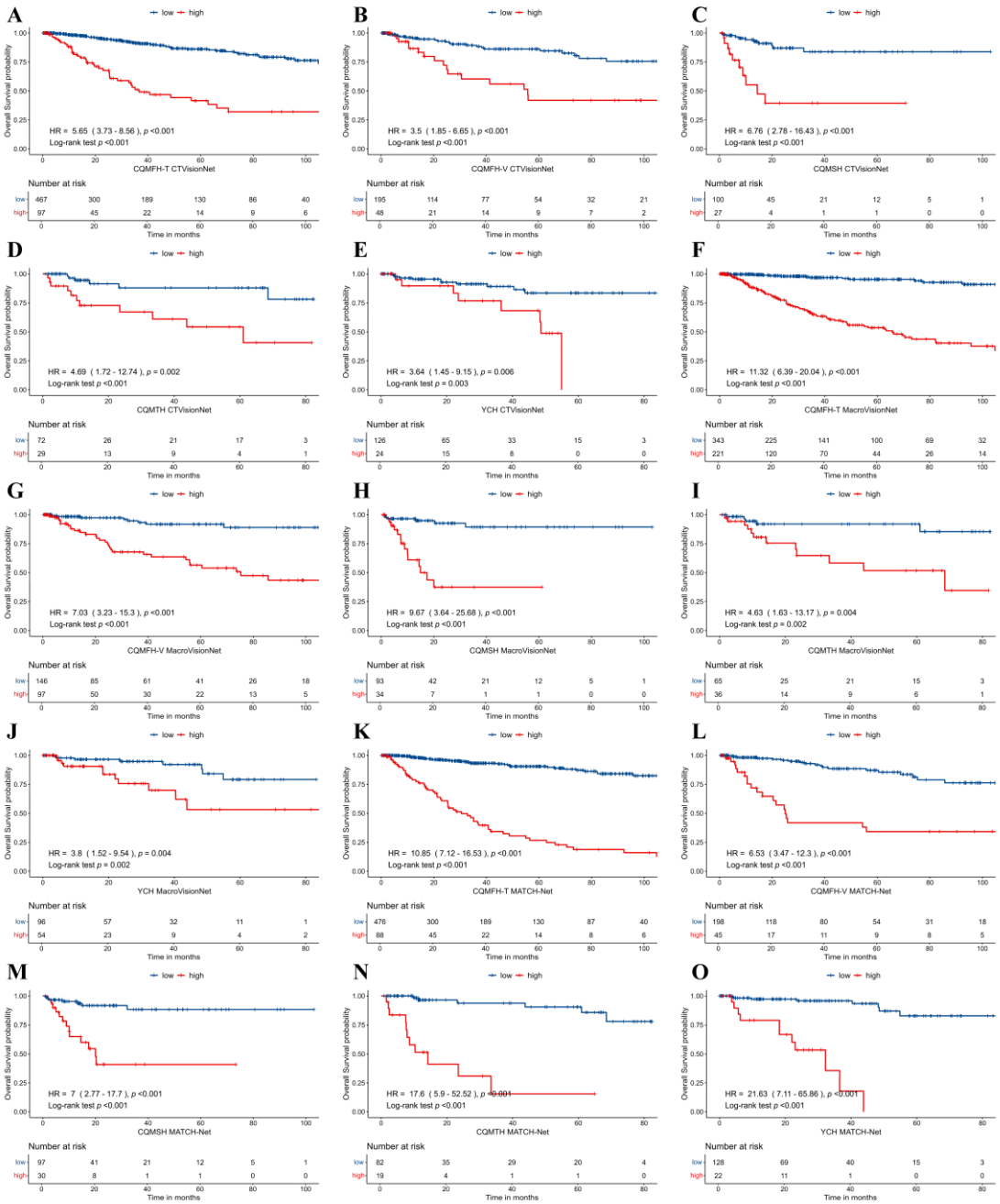

Figure 5. Kaplan–Meier survival analysis across the enrolled cohorts. Kaplan-Meier survival analysis was performed for CTVisionNet (A-E), MacroVisionNet (F-J), and MATCH-Net (K-O) across the enrolled cohorts. Survival differences between groups were assessed using the two-sided log-rank test, and HRs with 95% CIs were derived from univariable Cox proportional hazards regression. HR, Hazard Ratio; CI, Confidence Interval; CQMFH-T, The First Affiliated Hospital of Chongqing Medical University (Training cohort); CQMFH-V, CQMFH internal validation cohort; CQMSH, The Second Affiliated Hospital of Chongqing Medical University; CQMTH, The Third Affiliated Hospital of Chongqing Medical University; YCH, Yongchuan Hospital of Chongqing Medical University.

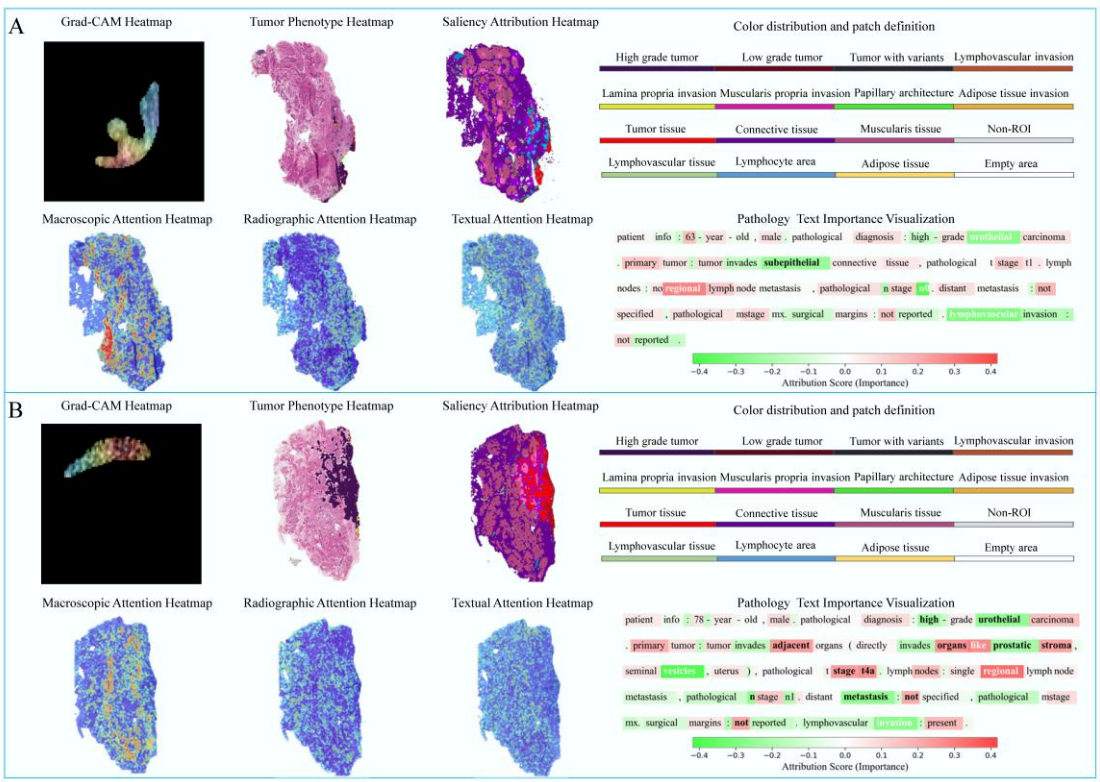

Figure 6. Interpretability analyses of the multimodal AI agent system in the CQMFH training cohort. The Grad-CAM and saliency attribution heatmaps highlight the regions of focus for CTVisionNet and MacroVisionNet. In the Co-Attention heatmaps, the highlighted areas represent regions of interest guided by their respective modalities. The highlighted text in the structured pathology reports indicates the specific textual descriptions that the multimodal model focused on. (A) Low-risk prediction example. (B) High-risk prediction example. CQMFH, The First Affiliated Hospital of Chongqing Medical University.

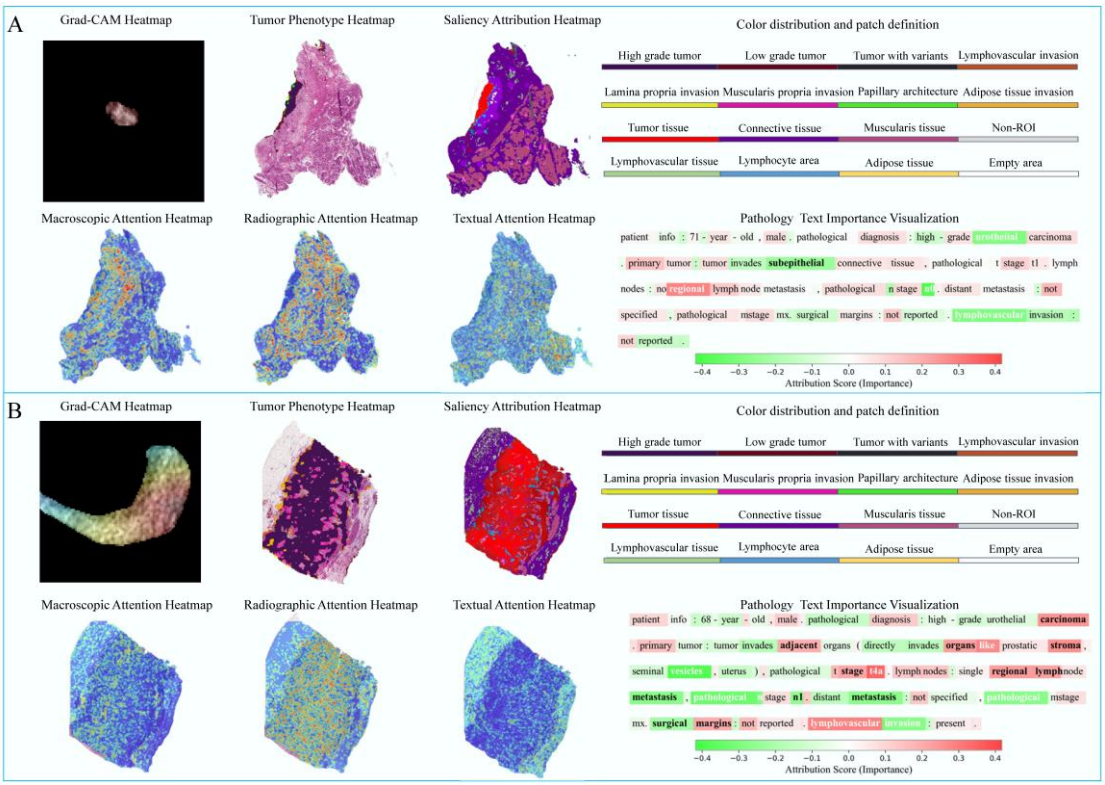

Figure 7. Interpretability analyses of the multimodal AI agent system in the CQMFH validation cohort. The Grad-CAM and saliency attribution heatmaps highlight the regions of focus for CTVisionNet and MacroVisionNet. In the Co-Attention heatmaps, the highlighted areas represent regions of interest guided by their respective modalities. The highlighted text in the structured pathology reports indicates the specific textual descriptions that the multimodal model focused on. (A) Low-risk prediction example. (B) High-risk prediction example. CQMFH, The First Affiliated Hospital of Chongqing Medical University.

	CQMFH (n= 807)		CQMSH	CQMTH	YCH	P value
	training (n=564)	validation (n=243)	(n=127)	(n=101)	(n=150)	
Age						p<0.05
⩽60 years	165(29.3%)	78(32.1%)	21(16.5%)	27(26.7%)	34(22.7%)	
61-70 years	213(37.8%)	80(32.9%)	49(38.6%)	28(27.7%)	54(36.0%)	
71-80 years	145(25.7%)	70(28.8%)	43(33.9%)	37(36.6%)	43(28.7%)	
>80 years	41(7.3%)	15(6.2%)	14(11.0%)	9(8.9%)	19(12.7%)	
Gender						0.52
Female	93(16.5%)	35(14.4%)	26(20.5%)	15(14.9%)	29(19.3%)	
Male	471(83.5%)	208(85.6%)	101(79.5%)	86(85.1%)	121(80.7%)	
Grade						p<0.05
High	369(65.4%)	147(60.5%)	78(61.4%)	59(58.4%)	91(60.7%)	
Low	186(33.0%)	92(37.9%)	44(34.6%)	32(31.7%)	57(38.0%)	
Not reported	9(1.6%)	4(1.6%)	5(3.9%)	10(9.9%)	2(1.3%)	
M stage						0.46
M0	563(99.8%)	243(100.0%)	126(99.2%)	101(100.0%)	150(100.0%)	
M1	1(0.2%)	--	1(0.8%)	--	--	
Pathological T stage						p<0.05
(Ta, T0,Tis)	75(13.3%)	44(18.1%)	25(19.7%)	17(16.8%)	21(14.0%)	
T1	163(28.9%)	62(25.5%)	28(22.0%)	23(22.8%)	33(22.0%)	
T2	108(19.1%)	37(15.2%)	23(18.1%)	16(15.8%)	17(11.3%)	
T3	52(9.2%)	34(14.0%)	12(9.4%)	8(7.9%)	6(4.0%)	
T4	15(2.7%)	6(2.5%)	5(3.9%)	1(1.0%)	4(2.7%)	
Tx	151(26.8%)	60(24.7%)	34(26.8%)	36(35.6%)	69(46.0%)	
Pathological N stage						p<0.05
N0	129(22.9%)	48(19.8%)	54(42.5%)	35(34.7%)	18(12.0%)	
N+	12(2.1%)	16(6.6%)	6(4.7%)	7(6.9%)	3(2.0%)	
Nx	423(75.0%)	179(73.7%)	67(52.8%)	59(58.4%)	129(86.0%)	
Lymphovascular invasion						p<0.05
Absent	298(52.8%)	137(56.4%)	60(47.2%)	35(34.7%)	58(38.7%)	
Present	33(5.9%)	13(5.3%)	13(10.2%)	11(10.9%)	9(6.0%)	
Not reported	233(41.3%)	93(38.3%)	54(42.5%)	55(54.5%)	83(55.3%)	
Follow-up time	29.40	25.80	14.30	12.72	23.10	
(median [IQR])	[11.28,60.42]	[8.20,62.10]	[6.09,27.39]	[7.85,52.79]	[9.72,43.46]	

Table 1: Baseline patient characteristics. Data are presented by median (interquartile range) or n (%).Variables are compared using the Chi-squared test or Fisher's exact test when the expected frequency in any count was less than 5. IQR, interquartile range; CQMFH, The First Affiliated Hospital of Chongqing Medical University; CQMSH, The Second Affiliated Hospital of Chongqing Medical University; CQMTH, The Third Hospital of Chongqing Medical University; YCH, Yongchuan Hospital of Chongqing Medical University.

Cohort	CQMFH-T			CQMFH-V		
Dice(%)	83.01±5.76	83.01±5.76	83.01±5.76	83.27±7.69	83.27±7.69	83.27±7.69
IOU(%)	71.34±7.89	71.34±7.89	71.34±7.89	72.01±10.37	72.01±10.37	72.01±10.37
SEN(%)	88.78±8.28	88.78±8.28	88.78±8.28	87.89±8.35	87.89±8.35	87.89±8.35
SPE(%)	99.99±0.01	99.99±0.01	99.99±0.01	99.99±0.01	99.99±0.01	99.99±0.01
Model	CTVisionNet	MacroVisionNet	MATCH-Net	CTVisionNet	MacroVisionNet	MATCH-Net
Low risk (n)	467	343	478	195	146	197
High risk (n)	97	221	86	48	97	46
Overall HR (95% CI)	5.65(3.73~8.56)	11.32(6.39~20.04)	10.85(7.12~16.53)	3.50(1.85~6.65)	7.03(3.23~15.30)	6.53(3.47~12.30)
multivariable Cox regression [a]	5.57(3.66~8.47)	11.21(6.28~20.01)	12.32(7.86~19.3)	2.85(1.43~5.66)	8.65(3.76~19.92)	7.21(3.69~14.08)
multivariable Cox regression [b]	2.84(1.79~4.51)	5.87(3.11~11.07)	6.91(3.98~11.97)	1.7(0.8~3.61)	3.19(1.23~8.26)	3.63(1.52~8.64)
C-index	0.764 (0.711-0.816)	0.813 (0.768-0.855)	0.854 (0.819-0.889)	0.707 (0.614-0.786)	0.765 (0.655-0.864)	0.846 (0.784-0.903)
C-index (1 years)	0.821 (0.689-0.928)	0.860 (0.756-0.950)	0.903 (0.830-0.965)	0.724 (0.443-0.920)	0.803 (0.562-0.961)	0.932 (0.835-1.000)
C-index (3 years)	0.768 (0.672-0.857)	0.819 (0.740-0.894)	0.862 (0.798-0.927)	0.781 (0.611-0.926)	0.770 (0.585-0.940)	0.881 (0.772-0.969)
C-index (5 years)	0.773 (0.683-0.867)	0.832 (0.753-0.907)	0.863 (0.784-0.927)	0.754 (0.591-0.899)	0.769 (0.610-0.916)	0.803 (0.664-0.936)
C-index P value						
AUC (1 years)	0.821 (0.733-0.909)	0.861 (0.794-0.929)	0.903 (0.857-0.949)	0.772 (0.621-0.924)	0.803 (0.648-0.958)	0.926 (0.871-0.982)
AUC (3 years)	0.770 (0.698-0.842)	0.818 (0.762-0.875)	0.862 (0.813-0.911)	0.778 (0.669-0.887)	0.765 (0.636-0.895)	0.882 (0.816-0.948)
AUC (5 years)	0.776 (0.706-0.845)	0.83 (0.773-0.888)	0.862 (0.81-0.913)	0.753 (0.637-0.868)	0.768 (0.658-0.878)	0.806 (0.708-0.904)

Table 2. Comparative evaluation of segmentation fidelity and prognostic efficacy for the CTVisionNet, MacroVisionNet, and MATCH-Net architectures across training and validation cohorts. Dice, dice similarity; IoU, Intersection over union; SEN, sensitivity; SPE, specificity; C-index, Concordance Index; AUC, Area Under the Curve; CI, Confidence Interval; HR, Hazard Ratio; CQMFH-T, The First Affiliated Hospital of Chongqing Medical University (Training cohort); CQMFH-V, CQMFH internal validation cohort. Dice, SEN, SPE, and IoU are reported as means with standard deviations. C-index, AUC, and HR are presented with 95% confidence intervals (CIs). C-index comparison is performed utilizing stratified bootstrap resampling with 1,000 iterations. [a] Multivariable Cox regression controlled for age, gender, [b] Multivariable Cox regression controlled for age, gender, T stage and tumor grade.

Cohort	CQMSH			CQMTH			YCH		
Dice(%)	84.32±7.09	84.32±7.09	84.32±7.09	84.27±5.02	84.27±5.02	84.27±5.02	83.79±6.73	83.79±6.73	83.79±6.73
IOU(%)	73.51±10.23	73.51±10.23	73.51±10.23	73.12±7.19	73.12±7.19	73.12±7.19	72.65±9.52	72.65±9.52	72.65±9.52
SEN(%)	87.66±8.68	87.66±8.68	87.66±8.68	84.65±6.94	84.65±6.94	84.65±6.94	88.64±7.06	88.64±7.06	88.64±7.06
SPE(%)	99.99±0.01	99.99±0.01	99.99±0.01	99.99±0.01	99.99±0.01	99.99±0.01	99.99±0.01	99.99±0.01	99.99±0.01
Model	CTVisionNet	MacroVisionNet	MATCH-Net	CTVisionNet	MacroVisionNet	MATCH-Net	CTVisionNet	MacroVisionNet	MATCH-Net
Low risk (n)	100	93	102	72	65	74	126	96	135
High risk (n)	27	34	25	29	36	27	24	54	15

HR (95% CI)	6.76(2.78~16.43)	9.67(3.64~25.68)	7.00(2.77~17.70)	4.69(1.72~12.74)	4.63(1.63~13.17)	17.60(5.90~52.52)	3.64(1.45~9.15)	3.80(1.52~9.54)	21.63(7.11~65.86)
multivariable Cox regression [a]	7.78(3.05~19.84)	10.78(3.94~29.46)	7.49(2.9~19.37)	4.61(1.6~13.24)	3.32(1.08~10.25)	17.32(5.27~56.87)	2.63(0.89~7.8)	7.81(2.57~23.75)	31.13(8.31~116.61)
multivariable Cox regression [b]	2.77(0.85~8.96)	4.94(1.04~23.44)	2.18(0.68~6.96)	1.93(0.58~6.43)	0.6(0.18~2.02)	10.54(2.99~37.12)	1.95(0.53~7.12)	2.92(0.73~11.63)	18.39(4.54~74.5)
C-index	0.746 (0.601-0.877)	0.779 (0.635-0.898)	0.849 (0.750-0.930)	0.791 (0.666-0.895)	0.771 (0.646-0.876)	0.874 (0.788-0.946)	0.701 (0.557-0.837)	0.790 (0.664-0.905)	0.836 (0.706-0.934)
C-index (1 years)	0.815 (0.590-0.979)	0.890 (0.703-1.000)	0.899 (0.756-1.000)	0.810 (0.549-1.000)	0.736 (0.446-1.000)	0.816 (0.550-0.987)	0.715 (0.307-0.976)	0.764 (0.417-1.000)	0.833 (0.515-1.000)
C-index (3 years)	0.783 (0.556-0.991)	0.784 (0.530-1.000)	0.865 (0.649-1.000)	0.776 (0.545-0.980)	0.835 (0.643-1.000)	0.919 (0.745-1.000)	0.717 (0.480-0.920)	0.887 (0.688-1.000)	0.895 (0.693-1.000)
C-index (5 years)	0.814 (0.540-1.000)	0.779 (0.510-1.000)	0.851 (0.566-1.000)	0.825 (0.584-1.000)	0.863 (0.649-1.000)	0.919 (0.761-1.000)	0.887 (0.649-1.000)	0.676 (0.350-1.000)	0.818 (0.539-1.000)
C-index P value									
AUC (1 years)	0.819 (0.685-0.954)	0.890 (0.776-1.000)	0.900 (0.817-0.982)	0.808 (0.658-0.958)	0.738 (0.584-0.891)	0.823 (0.687-0.960)	0.709 (0.507-0.911)	0.766 (0.582-0.950)	0.829 (0.666-0.992)
AUC (3 years)	0.787 (0.644-0.931)	0.784 (0.613-0.955)	0.862 (0.721-1.000)	0.774 (0.634-0.914)	0.838 (0.724-0.952)	0.924 (0.835-1.000)	0.706 (0.559-0.853)	0.891 (0.795-0.987)	0.899 (0.800-0.998)
AUC (5 years)	0.817 (0.659-0.976)	0.778 (0.603-0.954)	0.847 (0.675-1.000)	0.825 (0.689-0.961)	0.862 (0.743-0.982)	0.923 (0.838-1.000)	0.891 (0.777-1.000)	0.666 (0.434-0.898)	0.811 (0.653-0.969)

Table 3. Comparative evaluation of segmentation fidelity and prognostic efficacy for the CTVisionNet, MacroVisionNet, and MATCH-Net architectures in the external validation cohorts. Dice, dice similarity; IoU, Intersection over union; SEN, sensitivity; SPE, specificity; C-index, Concordance Index; AUC, Area Under the Curve; CI, Confidence Interval; HR, Hazard Ratio; CQMSH, The Second Affiliated Hospital of Chongqing Medical University. CQMTH, The Third Affiliated Hospital of Chongqing Medical University. YCH, Yongchuan Hospital of Chongqing Medical University. Dice, SEN, SPE, and IoU are reported as means with standard deviations. C-index, AUC, and HR are presented with 95% confidence intervals (CIs). C-index comparison is performed utilizing stratified bootstrap resampling with 1,000 iterations. [a] Multivariable Cox regression controlled for age, gender, [b] Multivariable Cox regression controlled for age, gender, T stage and tumor grade.