Original research article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 11: 1–18

DOI: 10.1177/20552076251393290
journals.sagepub.com/home/dhj
Sage

LLM-BCgrading: Large language model-based Chinese medical long text classification for bladder cancer grade prediction

Xianwei Pan[1,†], Lijie Wen[2,†], Yuhua Li[3], Yijia Zhang[4] and Mingyu Lu[1]

Abstract

Background: Traditional cystoscopic biopsy-based methods for histological grading of bladder cancer (BC) are invasive, subject to sampling errors, and susceptible to interobserver variability among pathologists. To address these challenges, this study explores a large language model (LLM)-based noninvasive approach to BC grade prediction using long Chinese medical texts.

Methods: We retrospectively collected admission records and computed tomography urography (CTU) descriptions from 642 patients pathologically diagnosed with BC. Each paired text was annotated as low grade or high grade according to histopathological results. We developed LLM-BCgrading to leverage HuatuoGPT-7B for Chinese medical long-text representation and integrated a gated multiplicative attention mechanism (GMAM) to selectively emphasize discriminative features. To address class imbalance and clinical risk asymmetry, the model was optimized with a cost-sensitive loss function. Performance was evaluated on a fixed internal test set with additional evaluation on an independent external validation cohort to assess generalizability.

Results: The best-performing configuration combined both admission records and CTU descriptions via an attention-based fusion strategy and GMAM, achieving balanced accuracy of 0.757, macro *F*1 score of 0.749, and macro AUC of 0.740. The ablation results demonstrated that incorporating both texts significantly improved classification performance compared with single-text configurations, and the GMAM consistently outperformed conventional attention mechanisms. Dimensionality experiments identified 256 as the optimal embedding size, balancing computational efficiency and predictive performance.

Conclusion: Our findings demonstrate that LLMs can effectively process Chinese medical long-texts for accurate preoperative prediction of BC grade. Attention-based fusion, cost-sensitive optimization, and interpretability based on Shapley additive explanations further support the robustness and clinical relevance of this LLM-driven framework.

Keywords

Chinese medical long-text classification, largelanguage model, bladder cancer, grade prediction, gated multiplicative attention mechanism

Received: 22 July 2025; accepted: 16 October 2025

[1]College of Artificial Intelligence, Dalian Maritime University, Dalian, China
[2]Department of Urology, The Second Affiliated Hospital of Dalian Medical University, Dalian, China
[3]School of Computer Science and Informatics, Cardiff University, Cardiff, UK
[4]School of Information Science and Technology, Dalian Maritime University, Dalian, China

[†]These authors contributed equally to this work.

Corresponding authors:
Mingyu Lu, College of Artificial Intelligence, Dalian Maritime University, Dalian 116026, China.
Email: lumingyu_professor@126.com

Yuhua Li, School of Computer Science and Informatics, Cardiff University, CF10 3AT Cardiff, UK.
Email: liy180@cardiff.ac.uk

Introduction

The latest statistical data from the International Agency for Research on Cancer (IARC) show that bladder cancer (BC) has become the ninth most commonly diagnosed cancer worldwide in both genders, with an estimated 613,791 new cases and 220,349 deaths in 2022.[1] According to the depth of invasion of the tumor into the bladder, BC can be pathologically classified into two types, that is, non-muscle-invasive bladder cancer (NMIBC) and muscle-invasive bladder cancer (MIBC), which correspond to different treatment decisions. In clinical practice, it remains challenging for urologists to develop prime personalized management strategy for patients with BC, mainly because of the morphological and cytological complexity of tumor progression.[2,3] The histological grade of BC, especially NMIBC, is one of the most significant factors for clinicians to make treatment selection, and predict recurrence and prognosis[4]; therefore, precisely grading BC is of great importance for patients to receive more appropriate treatment and gain more clinical benefits.

According to the World Health Organization grading system for NMIBC introduced in 2004, which still retained the same when updated in 2016 and 2022 (currently known as the WHO 2004/2022 system), the histological grades of NMIBC include papillary urothelial neoplasm of low malignant potential (PUNLMP), noninvasive papillary carcinoma low grade (LG) and high grade (HG), with corresponding PUNLMP, LG and HG rates of 1.5%, 49.8%, and 48.7%, respectively.[5] All MIBC cases should be considered HG BC.[6] Owing to the low proportion of PUNLMP and its low risk of recurrence and progression,[7] we mainly included LG and HG when stratifying patients with BC in this study. To determine the histological grade of BC, cystoscopic examination remains the recommended clinical diagnostic approach, providing tissue evaluation through either cold-cup biopsy or resection, which is regarded as the gold standard for grading.[8] However, biopsy results obtained from sampled tissues may only partially reflect the histopathological characteristics of the tumor, potentially leading to misdiagnosis. The inherent subjectivity of pathological interpretation and the relatively low interobserver reproducibility among pathologists further increase the risk of misclassification in BC grading. Moreover, owing to its invasive nature, cystoscopic biopsy can cause unpredictable adverse effects and considerable patient discomfort. Given these limitations, there is a pressing need to develop a more objective and noninvasive approach for the preoperative prediction of BC grade, thereby supporting optimal clinical management.

With the rapid advancement of artificial intelligence (AI) techniques, numerous medical and clinical problems have been effectively addressed by using computer-aided methods, particularly deep-learning frameworks. In the context of BC, studies have been conducted to improve the quality of the entire treatment process—from diagnosis to prognosis—by applying AI techniques to the analysis of genomics,[9–11] medical images,[12–16] and structured experimental data.[17,18] For AI-based BC grading, several studies have focused on the analysis of histopathological slides to develop deep-learning models capable of automatically predicting BC grade, thereby reducing inconsistencies and interobserver variability among pathologists. Wetteland et al.[19] proposed a two-stage pipeline comprising two primary models: the first segmented whole-slide images into six different types—urothelium, stroma, muscle, blood, damaged tissues, and background; the second extracted tiles from diagnostically relevant urothelial regions at three magnification levels, which were then sequentially processed by a convolutional neural network-based model. This pipeline achieved an average $F1$ score of 0.91 for both LG and HG classes. García et al.[20] presented a clustering-based self-learning framework to specifically grade MIBC from histological images. The model facilitated the classification of histological patches according to varying levels of disease severity and refined the latent feature space prior to the classification stage through the incorporation of a convolutional attention module, reaching a final average accuracy of 0.9034 in a multiclass task. Although previous studies have achieved high performance in BC grading based on histological images, the acquisition of these images is invasive and may cause unpredictable adverse effects on patients. Additionally, the annotations need to be manually provided by experts, which is a time-consuming and expertise-dependent process. To overcome the limitation of being dependent on the annotations of medical images in the field of medical image analysis, Fuster et al. introduced a pipeline in a weakly supervised way to extract urothelium tissue tiles at different magnification levels and employ a nested multiple-instance learning approach with attention to predict the grade.[21] Some researchers have explored both computed tomography (CT)-based deep-learning radiomics nomograms and models that combine CT semantic features with selected clinical variables for accurate prediction of the pathological grade of BC.[22,23] Furthermore, an approach based on multimechanical cellular properties to classify BC cells and early diagnose BC with AI has also been explored, which provided a novel perspective for BC grading.[24] Although accurate grading of BC is crucial for both clinicians and patients, existing AI-based studies on BC grading remain limited in number and predominantly focus on a narrow range of data modalities, such as histopathological images, CT scans, and experimental data, as previously mentioned. Therefore, it is imperative to develop more advanced computer-aided approaches and broaden the scope of data modalities to include sources such as electronic medical records (EMRs), which contain a wealth of valuable information for clinical decision-making.

Despite the extensive use of EMRs in natural language processing (NLP) applications—such as drug

recommendation,[25,26] disease diagnosis,[27,28] and automated international classification of diseases (ICD) coding[29–31]—AI-based prediction of BC grade via EMRs remains an unexplored area. Admission records and CT urography (CTU) descriptions are two textual parts of EMRs and contain abundant critical clinical information about patients. Commonly, admission records mainly consist of basic medical history, initial assessments, and auxiliary examination results, which provide almost the entire medical profile for the patient and can be collected relatively easily from the hospital information system (HIS). CTU descriptions are a main part of CTU reports and contain more details about the organ and neighboring tissues than the radiological diagnosis provided by radiologists. Compared with pathological biopsy procedures for BC grading, extracting critical clinical information from admission records and CTU descriptions via advanced NLP techniques may provide a significantly safer and more patient-acceptable alternative for predicting BC grade.

In recent years, several studies have demonstrated the feasibility of NLP for processing unstructured clinical text in the context of BC. For instance, Narayan et al. developed an NLP model to identify and characterize high-risk NMIBC patients from U.S. EMRs, with a focus on patient cohort identification rather than predictive modeling of pathological grade.[32] Schroeck et al. created a rule-based NLP engine to extract structured pathology variables from narrative pathology reports, primarily targeting post-operative histopathology data.[33] Similarly, Yang et al. employed a context-based NLP approach to determine muscle invasion status from free-text clinical documentation, thereby addressing tumor staging but not grading.[34] While these works underscore the potential of NLP in handling clinical narratives, they are predominantly centered on English-language datasets and largely confined to information extraction or staging tasks after diagnosis. In contrast, our study is the first to leverage domain-specific large language models (LLMs) for preoperative grade prediction of BC using long-form Chinese admission records and CTU descriptions, thus addressing both the linguistic challenges of Chinese medical narratives and the clinical need for noninvasive, pretreatment risk stratification.

While NLP techniques, such as rule-based algorithms and transformer-based models, have been extensively applied in medical text analysis and have undergone great development, it is still challenging to understand very long Chinese medical texts from real clinical health records. The challenges stem from the substantial length of medical texts, the complexity of medical terminology, the inherent ambiguity of the Chinese language, and the variability in narrative styles. Recently, LLMs have been recognized as an effective way to solve the toughest problems in the field of natural language understanding; furthermore, to obtain more accurate results, many domain-specific LLMs have been trained for specific tasks. HuatuoGPT-7B[35] is an LLM trained on a vast Chinese medical corpus, and can be specifically used for Chinese medical note analysis. It has achieved excellent performance in understanding Chinese medical language. To leverage the great capacity of LLMs in understanding long Chinese medical texts, we attempt to embed clinical notes, that is, admission records and CTU descriptions, based on HuatuoGPT-7B to develop an end-to-end deep-learning model for BC grade prediction.

The contributions we made in this paper are as follows:

1. We constructed a Chinese medical text dataset, BCgrading-Text, which comprises admission records and CTU examination descriptions from 642 patients who were pathologically diagnosed with BC in real-world clinical settings. Each case was annotated as either LG or HG BC based on histopathological examination results. To ensure experimental consistency and reproducibility, the dataset is partitioned into fixed training, validation, and test subsets, providing a reliable benchmark for research in Chinese medical long-text classification.
2. We propose an LLM-based model for long-text classification in the Chinese medical domain, specifically designed to process EMRs. Our goal is to explore a more effective strategy for applying LLMs to medical text classification tasks. Notably, we introduce a novel modification to the attention mechanism by replacing the conventional additive residual connection with a multiplicative operation. This design acts as a gating mechanism to enhance the influence of important textual components.
3. We conduct extensive experiments to explore the model architecture and pattern selection. Through a series of ablation studies, we identified an effective and competitive strategy for leveraging LLMs to predict BC grades from lengthy Chinese medical texts, highlighting clear advantages in handling long and complex clinical narratives.

Methods

Dataset construction

Patient inclusion. All procedures performed in our study involving human participants were conducted in accordance with the 1964 Helsinki Declaration and its subsequent amendments or comparable ethical standards. This study was approved by the Institutional Review Board of the Second Affiliated Hospital of Dalian Medical University (Approval No. KY2025-541-01). Owing to the retrospective design and the use of anonymized clinical data, the

requirement for informed consent was waived by the ethics committee.

Patients who were pathologically diagnosed with BC and received treatment at the Second Hospital of Dalian Medical University between 2018 and 2022 were included in our study. The inclusion principles for patients are as follows: (1) CTU scans are performed before cystectomy and (2) bladder biopsy is performed to obtain pathological results. Finally, 642 patients were included in this study. Patient ages ranged from 23 to 98 years, with an average age of 69.04 ± 11.34 years. Among all patients, 189 patients (29.44%) had LG BC, and 453 patients (70.56%) had HG BC, indicating a class imbalance in the dataset with a predominance of HG cases. This imbalance may affect model performance, particularly in recognizing LG BC, and has been considered during model training and optimization.

Text dataset. We extracted two kinds of texts from the EMRs of patients, that is, admission records (T_a) and CTU descriptions (T_c), as processing objects to formulate a clinical profile for patients. After removing patient-private information and performing text cleaning and deduplication, the preprocessed admission records include chief complaint; present illness history; past medical history (general health; history of illnesses and injuries; infectious disease history; surgical and trauma history; blood transfusion history; allergy and vaccination history); personal history; marital and reproductive history; family history; physical examination (vital signs; general condition; skin and mucosa; superficial lymph nodes; head and associated structures; neck; chest; lungs; heart; abdomen; external genitalia; digital rectal examination; spine and limbs; nervous system); specialist findings; auxiliary examination; and preliminary diagnosis. The CTU descriptions are only the detailed narratives of CTU reports provided by radiologists, excluding the diagnostic conclusion. The formats of these two texts were unified among patients, as shown in Figure 1.

Our constructed dataset, known as the BCgrading-Text dataset, comprises 642 pairs of admission records and CTU descriptions from patients with BC. Two urologists —one with 10 years of clinical experience and the other with 15 years of clinical surgical experience—evaluated the included texts and annotated them as LG or HG BC based on histopathological results. In cases of disagreement between their annotations, a senior urologist with 20 years of experience re-evaluated the cases to reconcile discrepancies and minimize interobserver variability. The LG group is labeled 0 (negative), and the HG group is therefore labeled 1 (positive). Accordingly, the task of predicting BC grade is framed as a Chinese medical long-text classification problem.

In our BCgrading-Text dataset, we established fixed splits for the training, validation, and test sets to ensure consistency across experiments. This design aims to provide a benchmark dataset that enables other researchers to reproduce our results. Specifically, the training set comprises 449 cases (132 LG cases and 317 HG cases), the validation set includes 96 cases (28 LG cases and 68 HG cases), and the test set contains the remaining 97 cases (29 LG cases and 68 HG cases).

External validation dataset. To further evaluate the generalizability of the proposed model, we additionally collected an independent dataset from another subspecialty department of our hospital. Although this dataset was obtained from the same institution, the medical records were documented by different surgeons, with notable variations in terminology usage and narrative style, thereby providing a heterogeneous validation cohort that could serve as a proxy for external validation.

This external validation dataset comprised 76 patients with pathologically confirmed BC, including 23 cases of LG and 53 cases of HG. The inclusion and exclusion criteria were consistent with those described in the "Patient inclusion" section. Both admission records and CTU descriptions were extracted and preprocessed following the same pipeline outlined in the "Text dataset" section. To ensure comparability, five representative baseline models and our proposed model were retrained on the internal BCgrading-Text training set and subsequently evaluated on this external validation dataset.

Model architecture

Owing to their powerful capability in language understanding and generating, LLM-based models have shown excellent performance in NLP-related medical tasks, such as automatic ICD coding and drug recommendation. In this study, we propose LLM-BCgrading, an LLM-based neural architecture, to classify BC grades from two types of long Chinese medical texts: (1) admission records and (2) CTU descriptions. We utilize HuatuoGPT-7B, which is a large Chinese-language model trained on real-world data from *doctors* and distilled data from *ChatGPT*, to understand and represent Chinese medical long-texts. As shown in Figure 2, our model comprises two modules: LLM-based text understanding is used for medical text representation via HuatuoGPT-7B, and attention-based feature fusion is adopted to enhance the extracted features for better performance.

LLM-based text understanding. For each patient, two types of texts are provided: the admission record T_a and the CTU description T_c. Each text is fed into HuatuoGPT-7B to obtain the last hidden states from the encoder as its embeddings, with the encoder specifically designed for natural language understanding. According to the HuatuoGPT-7B technical documentation, the

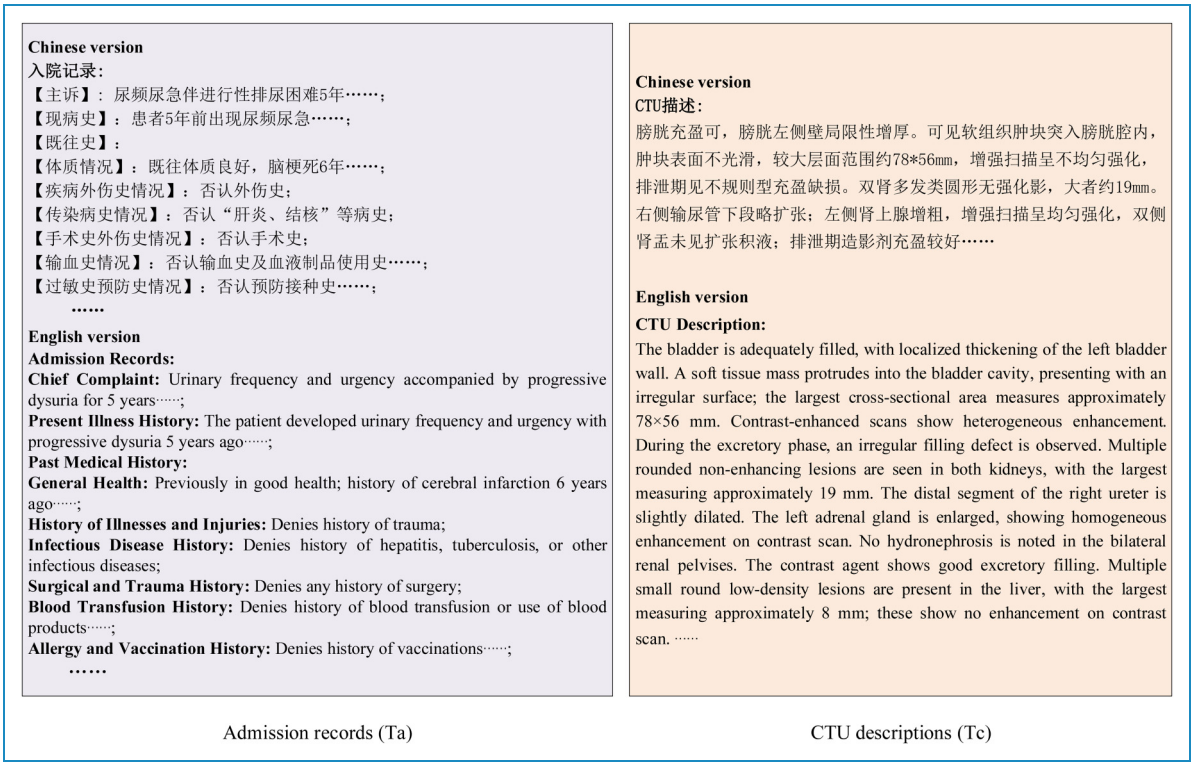

Figure 1. Illustration of the two text formats used in this study. The figure shows the two types of text formats included in our dataset, BCgrading-Text. The left panel shows the admission record structure, and the right panel displays the formats of the CTU descriptions. The Chinese version represents the actual text processed in our study, while the English translation is provided solely to improve the readability of the article. CTU: computed tomography urography.

dimension of the hidden states is set to 3584 which is the embedding dimension of T_a and T_c in this study. The process was formulated as follows:

$$H_a = \text{HuatuoGPT}(T_a) \tag{1}$$

$$H_c = \text{HuatuoGPT}(T_c) \tag{2}$$

where $H_a \in \mathbb{R}^{3584}$ and $H_c \in \mathbb{R}^{3584}$.

Gated multiplicative attention mechanism. To emphasize important semantic features within the text representation, we alternatively employ a variant attention mechanism, named gated multiplicative attention mechanism (GMAM). First, a multihead cross-attention mechanism was applied to compute an intermediate vector A which can be formulated as follows:

$$A = \text{MultiHeadAttn}(H_a, H_c) \tag{3}$$

where $A \in \mathbb{R}^{3584}$, and MultiHeadAttn$(\cdot)$ is the multihead cross-attention mechanism.

Then, we apply a sigmoid activation function to the intermediate vector A to compute a score vector G, which is used as a gate in the next step. Finally, element-wise multiplication was used to form the gated vector H_{gated} which was subsequently normalized via LayerNorm. The process can be denoted as follows:

$$G = \sigma(A) \in \mathbb{R}^{3584} \tag{4}$$

$$H_{\text{gated}} = H_a \odot G \in \mathbb{R}^{3584} \tag{5}$$

$$H_{\text{norm}} = \text{LayerNorm}(H_{\text{gated}}) \tag{6}$$

To reduce the computational cost, mitigate overfitting by avoiding the curse of dimensionality, and filter out redundant or noisy features, dimensionality reduction was performed to project the normalized 3584-dimensional embeddings to 256 dimensions:

$$H_{\text{proj}} = \text{LayerNorm}(\text{ReLU}(W_{\text{proj}} H_{\text{norm}} + b_{\text{proj}})) \tag{7}$$

where $H_{\text{proj}} \in \mathbb{R}^d$, $d = 256$ in our implementation, and W_{proj} and b_{proj} are the learnable parameters.

A feed-forward network (FFN) is applied as follows:

$$\begin{aligned} H_{\text{ffn}} &= \text{FFN}(H_{\text{proj}}) \\ &= \text{Dropout}(W_2 \cdot \text{GELU}(W_1 \cdot H_{\text{proj}} + b_1) + b_2) \end{aligned} \tag{8}$$

where $H_{\text{ffn}} \in \mathbb{R}^{256}$, W_1, W_2, b_1, and b_2 are the learnable parameters.

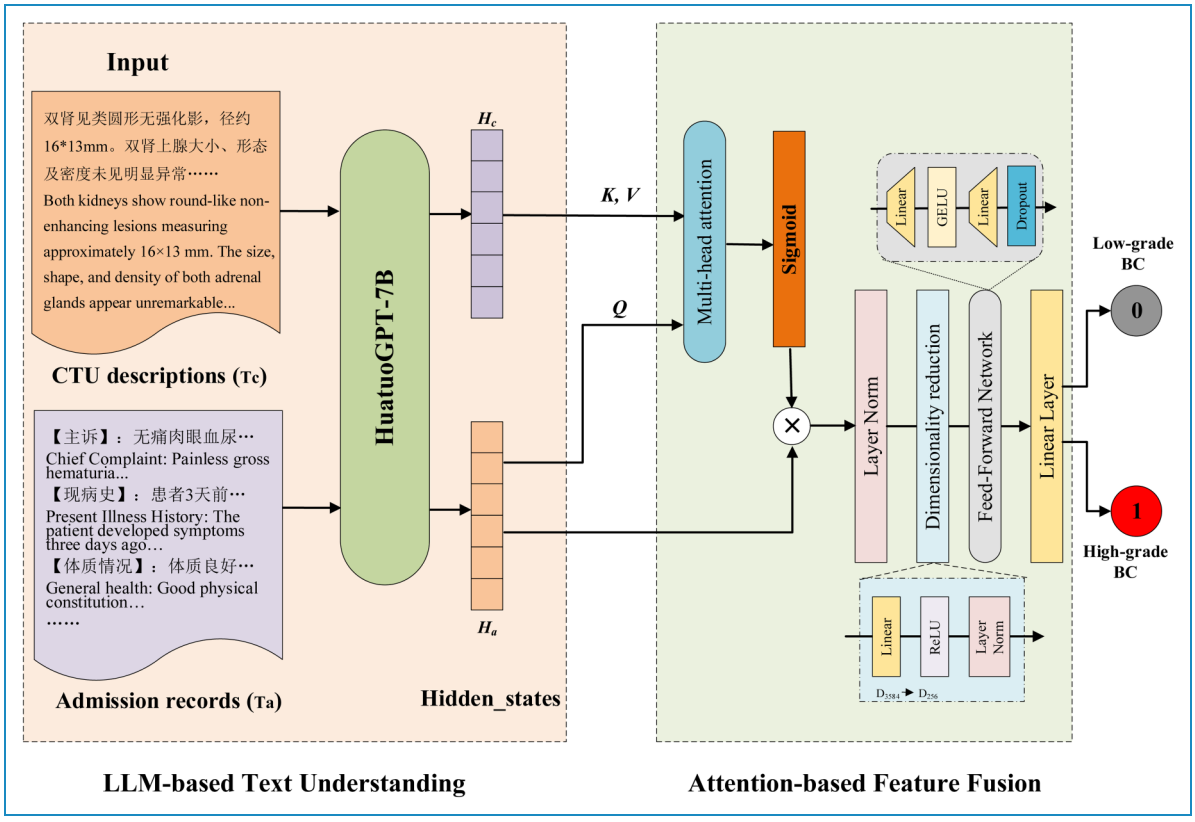

Figure 2. Architecture of the proposed LLM-BCgrading model. Input texts are encoded into embeddings via HuatuoGPT-7B, which is capable of comprehending Chinese medical long-texts and capturing latent information relevant to BC grading. A variant of the attention mechanism is incorporated to strengthen salient features and enhance overall model performance. LLM: large language model; BC: bladder cancer.

The output of the FFN layer H_{ffn} is used for the final prediction and passed through a linear classifier to generate prediction logits over two classes:

$$\hat{y} = \text{Softmax}(W_{\text{cls}} \cdot H_{\text{ffn}} + b_{\text{cls}}) \tag{9}$$

where W_{cls} and b_{cls} are trainable parameters.

Model optimization. The dataset presented a pronounced class imbalance, with HG cases considerably more frequent than LG cases. Such imbalance can bias model predictions toward the majority class, reducing sensitivity to the minority class. In addition, the clinical consequences of misclassification are asymmetric: predicting a HG case as LG (missed diagnosis) is more severe than predicting a LG case as HG (overtreatment).

To address both the statistical imbalance and the clinical risk asymmetry, we employed a cost-sensitive cross-entropy (CS-CE) loss.[36,37] The formulation is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathbb{E}_{\text{p}}[C(y, \hat{y})] \tag{10}$$

where $\mathcal{L}_{\text{CE}}$ is the weighted cross-entropy (WCE) loss and $\mathbb{E}_{\text{p}}[C(y, \hat{y})]$ represents the expected misclassification cost under the predictive distribution:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N} w_{y_i} \log p(y_i \mid \boldsymbol{x}_i) \tag{11}$$

$$\mathbb{E}_{\text{p}}[C(y, \hat{y})] = \frac{1}{N} \sum_{i=1}^{N} \sum_{j} p(y = j \mid \boldsymbol{x}_i) C(y_i, j) \tag{12}$$

The cost matrix C was defined as:

$$C = \begin{bmatrix} 0 & C_{\text{LH}} \\ C_{\text{HL}} & 0 \end{bmatrix}, \quad C_{\text{LH}} = 1.5, \quad C_{\text{HL}} = 3.0 \tag{13}$$

where N is the number of training cases, $\boldsymbol{x}_i$ is the i-th input, $\boldsymbol{y}_i$ is the corresponding ground truth label, w_{y_i} is the class-specific weight, and $p(y_i \mid \boldsymbol{x}_i)$ denotes the predicted probability of the true class, $p(y = j \mid \boldsymbol{x}_i)$ is the predicted probability of class j for case i. C_{LH} penalizes LG cases predicted as HG, and C_{HL} penalizes HG cases predicted as LG. A trade-off coefficient $\lambda = 1.0$ was used in this study.

For comparison, we also evaluated WCE[38] and focal loss with inverse-frequency weighting (Focal-IFW),[39] but the cost-sensitive formulation was ultimately adopted as the

primary loss due to its ability to balance statistical skew with clinical risk.

Evaluation metrics

To comprehensively evaluate model performance, we calculated overall accuracy (Acc), balanced accuracy (Bal-acc), weighted and macro $F1$ scores ($F1$-W and $F1$-M), macro area under the receiver operating characteristic (ROC) curve (AUC-M), and class-specific sensitivity and specificity.

Let TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives for a given class, respectively. The metrics are defined as follows:

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{14}$$

$$\text{Bal-acc} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right) \tag{15}$$

$$\text{Precision}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c} \tag{16}$$

$$\text{Recall}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \tag{17}$$

$$F1_{\text{weighted}} = \sum_{c=1}^{C} \frac{n_c}{N}\left(2\frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}\right) \tag{18}$$

$$F1_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\left(2\frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}\right) \tag{19}$$

$$\text{Sensitivity}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \tag{20}$$

$$\text{Specificity}_c = \frac{\text{TN}_c}{\text{TN}_c + \text{FP}_c} \tag{21}$$

For binary classification, AUC was calculated for each class by treating it as positive against the other class, and then averaged:

$$\text{AUC}_M = \frac{1}{C}\sum_{c=1}^{C}\text{AUC}_c \tag{22}$$

where C is the number of classes; n_c is the number of cases in class c; TP_c, TN_c, FP_c, FN_c represents the number of corresponding metrics for class c; Precision_c and Recall_c refer to the precision and recall of class c, respectively .

Baseline pretrained models

To evaluate the effectiveness of our proposed framework, we compared it with several widely used pretrained language models covering different domains, languages, and scales. These models differ in terms of their pretraining corpora, maximum input length, long-text handling strategies, and embedding dimensionalities. Specifically, we included:

Bio_ClinicalBERT,[40] an English clinical models trained on all notes from MIMIC-III, restricted to a 512-token input limit.

BERT-base-Chinese,[41] a general-domain Chinese model, also limited to 512 tokens, serving as a widely used NLP baseline.

Qwen3-Embedding series (0.6B, 4B, and 8B),[42] three multilingual embedding models that support up to ~32K tokens and provide different embedding dimensionalities (1024, 2560, and 4096, respectively), thereby enabling direct long-sequence representation learning. We simplified these three models of 0.6B, 4B, and 8B as Qwen3-em-0.6B, Qwen3-em-4B, and Qwen3-em-8B, separately.

These diverse baselines allow us to comprehensively assess the contribution of domain adaptation (general vs. clinical vs. medical Chinese), context length, and embedding dimensionality to downstream BC grading.

Model interpretability analysis

To improve the transparency of model predictions, we employed Shapley additive explanations (SHAP)[43] to quantify the contribution of individual textual features to the classification outcome. SHAP values were calculated using the Python *SHAP* package (v0.39.0) within the Google Colab environment. Global interpretability was assessed through summary plots and bar plots of mean SHAP values, which highlighted the relative importance of admission records versus CTU descriptions. Case-level interpretability was further explored by extracting and visualizing the top-contributing sentences from each text for representative patients.

Experiments and results

Implementation details

All the experiments were conducted using PyTorch 2.6.0 + cu124 with CUDA 12.4 support, running on Google Colab Pro with NVIDIA Tesla T4 GPU (16 GB memory), NVIDIA L4 GPU (24 GB memory), and NVIDIA A100 (40 GB memory). The code was implemented in Python 3.11.13 and executed within a Linux-based container environment. The model was trained via the Adam optimizer with a fixed learning rate of $1e-5$. The batch size is set to 32, and the number of epochs is set to 150. We evaluated our model using standard metrics as described in the "Evaluation metrics" section. To train and evaluate the model, we constructed the BCgrading-Text dataset, which consists of 642 paired admission records and CTU descriptions. The dataset was randomly split into training, validation, and test sets at a ratio of 0.70:0.15:0.15, with fixed

case assignments as described in the "Text dataset" section. In addition, an external validation set of 76 cases was collected from another subdepartment to assess model generalizability.

Model configuration pattern selection

We demonstrate the effectiveness of our proposed framework by conducting extensive experiments in this subsection. We compare the results of different data configurations, data fusion strategies, and the efficacy of the GMAM in all model settings to determine the best model configuration for BC grading prediction based on clinical admission records and CTU descriptions.

Data ablation. To evaluate the contribution of different types of clinical notes (i.e. admission records denoted T_a and CTU descriptions denoted T_c in this study) to model performance, we conducted an ablation study by comparing three configurations: using only admission records (T_a-only), using only CTU descriptions (T_c-only), and combining both components ($T_a + T_c$). As shown in Table 1, using CTU descriptions alone (T_c-only) yielded the weakest performance (macro $F1 = 0.567$, balanced accuracy = 0.569), indicating limited discriminatory value. Admission records alone (T_a-only) substantially improved performance (macro $F1 = 0.715$, balanced accuracy = 0.695), highlighting their richer clinical information. When combining both sources ($T_a + T_c$), the model achieved the best balance across metrics, with the highest macro $F1$ (0.749) and balanced accuracy (0.757), driven by a marked improvement in LG sensitivity (0.690), confirming the complementary value of CTU descriptions to admission records.

Data fusion strategies. To investigate the impact of different fusion strategies for the two kinds of data on model performance, we compared three approaches: (1) early fusion by directly concatenating the two text sequences before embedding ($T_a + T_c$ before emb), (2) direct vector concatenation after independent embeddings ($T_a + T_c$ concat), and (3) cross-attention-based fusion after independent embeddings ($T_a + T_c$ attn-based). As summarized in Table 2, fusing T_a and T_c before embedding yielded the highest macro AUC of 0.785 but poor balance between classes, with LG sensitivity only 0.552. $T_a + T_c$ concat fusion improved balanced accuracy to 0.737 and macro $F1$ to 0.739 though LG sensitivity remained limited (0.621). In contrast, the $T_a + T_c$ attn-based fusion achieved the best overall trade-off, with the highest balanced accuracy (0.757) and macro $F1$ (0.749), markedly improving LG sensitivity to 0.690 while maintaining HG sensitivity, and was therefore selected as the final fusion strategy.

Efficacy of GMAM. We investigate the impact of attention mechanism design on model performance by comparing the conventional attention mechanism followed by a residual connection[44] (AttnResidual) with our proposed GMAM. The GMAM replaces the standard additive residual connection with a sigmoid-based multiplicative gating operation, allowing the model to dynamically modulate the importance of the attended features.

Experimental results showed that GMAM consistently outperformed AttnResidual in both T_c-only and T_a-only configurations, yielding higher balanced accuracy and macro $F1$ (as shown in Table 3). This advantage became more evident when combined with an FFN under the $T_a + T_c$-Attnfusion strategy, where GMAM-FFN achieved the most favorable trade-off across metrics, markedly improving LG sensitivity while preserving HG detection. These findings suggest that multiplicative gating provides a more effective way to emphasize salient features and suppress redundant signals than additive residual connections.

The superiority of GMAM over the conventional AttnResidual design is further illustrated in Figure 3. Across all configurations, GMAM consistently achieved higher macro $F1$ scores, with particularly notable gains in the T_c-only (+0.122) and $T_a + T_c$-FFN (+0.083) settings. These improvements indicate that multiplicative gating is especially effective in scenarios where feature interactions are either sparse (T_c-only) or heterogeneous ($T_a + T_c$ fusion). In addition, the incremental advantage of GMAM becomes more pronounced as model complexity increases, such as when combined with FFN layers, suggesting that gating enables deeper architectures to better exploit complementary information without amplifying noise. Together, these findings highlight the robustness and scalability of GMAM as a general attention mechanism for medical text classification.

Comparison of different dimensions

Given the substantial length of the Chinese medical texts involved in this study, we employed HuatuoGPT-7B to generate 3584-dimensional embeddings. To avoid excessive computational cost and overfitting, we applied a linear projection to reduce dimensionality and conducted experiments across multiple target dimensions.

As shown in Table 4, model performance varied with the reduced dimensionality. Very low dimensions ($D = 64$) already provided competitive results (macro $F1 = 0.744$, balanced accuracy = 0.747), while increasing the dimension to $D = 128$ slightly improved overall accuracy (0.794) but at the expense of class balance (balanced accuracy = 0.715). The best trade-off was achieved at $D = 256$, where the model obtained the highest macro $F1$ (0.749) and balanced accuracy (0.757), with substantial improvement in low-grade sensitivity (0.690). Larger dimensions (≥ 512) did not yield further gains; instead, performance gradually declined, likely due to redundancy and overfitting. These results demonstrate that moderate dimensionality reduction not only alleviates computational burden but also enhances

Table 1. Ablation study for different types of clinical notes.

Data types	Acc	Bal-acc	*F*1-W	*F*1-M	AUC-M	Sen-LG/Spe-LG	Sen-HG/Spe-HG
T_c-only	0.701	0.569	0.664	0.567	0.614	0.241/0.897	0.897/0.241
T_a-only	**0.794**	0.695	0.775	0.715	**0.753**	0.448/**0.941**	**0.941**/0.448
$T_a + T_c$	0.784	**0.757**	**0.786**	**0.749**	0.740	**0.690**/0.824	0.824/**0.690**

Acc: accuracy; Bal-acc: balanced accuracy; *F*1-W: weighted *F*1 score; *F*1-M: macro *F*1 scores; AUC-M: macro area under the ROC curve; Sen: sensitivity; Spe: specificity; LG: low grade; HG: high grade; T_a: admission record; T_c: computed tomography urography description.
Bold indicates the best results.

generalization, with $D = 256$ emerging as the most effective embedding size for this task.

The trend is further illustrated in Figure 4, which plots macro *F*1 score against embedding dimensionality. While a reduced dimension of 64 already preserved much of the model's discriminative ability, performance slightly dropped at 128 and then peaked at 256, confirming it as the optimal setting. Beyond this point, larger embedding sizes (512–2048) led to progressively lower macro *F*1 scores, with the steepest decline observed at 1024. This suggests that excessive dimensionality introduced redundancy and overfitting, outweighing the benefits of richer representations. Taken together, both the tabular and graphical analyses indicate that 256 dimensions strike the best balance between information retention and generalization, and were therefore adopted as the final configuration for subsequent experiments.

Comparison of different learning strategy

We compared the three imbalance-handling loss functions and the results are listed in Table 5. The CS-CE achieved the highest overall accuracy (0.784) and weighted *F*1 (0.786), while also yielding the best macro *F*1 (0.749), slightly outperforming Focal-IFW (0.734) and WCE (0.716). Compared with WCE, cost-sensitive training markedly improved HG sensitivity, though with a reduction in LG sensitivity. Focal-IFW produced more balanced class sensitivities (0.724 and 0.779) but did not surpass the macro *F*1 achieved by the CS-CE approach. Although Focal-IFW showed the highest macro AUC, cost-sensitive optimization provided the most favorable balance between class-wise performance and overall accuracy, and was therefore adopted for subsequent experiments.

Comparative evaluation on internal and external datasets

To comprehensively evaluate the effectiveness and robustness of the proposed framework, we compared HuatuoGPT-7B with representative baselines, including Bio_ClinicalBERT, BERT-base-Chinese, and Qwen3-Embedding models (0.6B, 4B, 8B).

On the internal BCgrading-Text test set, HuatuoGPT-7B consistently achieved the highest balanced accuracy (0.757) and macro *F*1 score (0.749), outperforming all baselines (Table 6). Although Qwen3-em-8B obtained the highest macro AUC (0.760), HuatuoGPT-7B demonstrated competitive AUC (0.740) while maintaining superior threshold-dependent performance metrics. Notably, Bio_ClinicalBERT showed relatively poor performance (macro $F1 = 0.613$), reflecting the limited transferability of English-pretrained models to Chinese medical narratives, while BERT-base-Chinese achieved moderate results. The Qwen3 series performed better, particularly the 0.6B model, but still lagged behind HuatuoGPT-7B in overall effectiveness.

The confusion matrices on the internal test set (Figure 5) provide further insight into model behavior. Across all models, predictions for HG BC were relatively stable, whereas LG classification was more variable. Several baselines, including Qwen3-em-8B and Bio_ClinicalBERT, exhibited a higher proportion of false negatives for LG cases. In contrast, HuatuoGPT-7B achieved more balanced sensitivity across both classes, which explains its superior macro *F*1 and balanced accuracy despite not always producing the highest AUC.

To further assess generalizability, we conducted external validation using an independent dataset of 76 patients collected from another subdepartment within the same hospital. All models exhibited performance degradation when transferred to this heterogeneous dataset (Table 6 and Figure 6), yet HuatuoGPT-7B remained the strongest performance with a macro *F*1 score of 0.680 and balanced accuracy of 0.676. Although Qwen3-em-4B achieved a slightly higher macro AUC (0.716), HuatuoGPT-7B showed the most stable macro *F*1 and balanced accuracy, suggesting better calibration under distributional shifts. Importantly, Bio_ClinicalBERT suffered the steepest decline (macro $F1 = 0.412$), highlighting the challenges faced by non-Chinese pretrained models in this task.

Taken together, these results demonstrate that HuatuoGPT-7B delivers superior balanced accuracy and *F*1 compared with widely used transformer-based baselines, while maintaining competitive AUC performance. Moreover, the model exhibited greater robustness to heterogeneous documentation styles in the external validation

Table 2. Ablation experiments for different data fusion strategies.

Data fusion strategy	Acc	Bal-acc	*F*1-W	*F*1-M	AUC-M	Sen-LG/Spe-LG	Sen-HG/Spe-HG
$T_a + T_c$ before emb	0.784	0.717	0.777	0.727	**0.785**	0.552/**0.882**	**0.882**/0.552
$T_a + T_c$ concat	0.784	0.737	0.782	0.739	0.742	0.621/0.853	0.853/0.621
$T_a + T_c$ attn-based	**0.784**	**0.757**	**0.786**	**0.749**	0.740	**0.690**/0.824	0.824/**0.690**

Acc: accuracy; Bal-acc: balanced accuracy; *F*1-W: weighted *F*1 score; *F*1-M: macro *F*1 scores; AUC-M: macro area under the ROC curve; Sen: sensitivity; Spe: specificity; LG: low grade; HG: high grade; T_a: admission record; T_c: computed tomography urography description.
Bold indicates the best results.

Table 3. Performance comparison of models employing AttnResidual and GMAM for BC grade prediction.

Method	Acc	Bal-acc	*F*1-W	*F*1-M	AUC-M	Sen-LG/Spe-LG	Sen-HG/Spe-HG
T_c-AttnResidual	0.742	0.579	0.676	0.564	0.694	0.172/**0.985**	**0.985**/0.172
T_c-GMAM	0.763	0.673	0.749	0.686	0.690	0.448/0.897	0.897/0.448
T_c-AttnResidual-FFN	0.701	0.609	0.688	0.614	0.619	0.379/0.838	0.838/0.379
T_c-GMAM-FFN	0.732	0.611	0.702	0.618	0.613	0.310/0.912	0.912/0.310
T_a-AttnResidual	0.773	0.680	0.757	0.696	0.740	0.448/0.912	0.9128/0.448
T_a-GMAM	0.784	0.687	0.766	0.705	0.740	0.448/0.927	0.927/0.448
T_a-AttnResidual-FFN	0.784	0.678	0.762	0.696	0.718	0.414/0.941	0.941/0.414
T_a-GMAM-FFN	**0.794**	0.695	0.775	0.715	**0.753**	0.448/0.941	0.941/0.448
$T_a + T_c$-Attnfusion-AttnResidual	**0.794**	0.675	0.765	0.696	0.711	0.379/0.971	0.971/0.379
$T_a + T_c$-Attnfusion-GMAM	0.784	0.707	0.774	0.721	0.771	0.517/0.897	0.897/0.517
$T_a + T_c$-Attnfusion-AttnResidual-FFN	0.773	0.650	0.742	0.666	0.723	0.345/0.956	0.956/0.345
$T_a + T_c$-Attnfusion-GMAM-FFN	0.784	**0.757**	**0.786**	**0.749**	0.740	**0.690**/0.824	0.824/**0.690**

Acc: accuracy; Bal-acc: balanced accuracy; *F*1-W: weighted *F*1 score; *F*1-M: macro *F*1 scores; AUC-M: macro area under the ROC curve; Sen: sensitivity; Spe: specificity; LG: low grade; HG: high grade; T_a: admission record; T_c: computed tomography urography description; GMAM: gated multiplicative attention mechanism; BC: bladder cancer; FFN: feed-forward network.
Bold indicates the best results.

dataset, underscoring its potential for practical application in Chinese medical text classification.

SHAP-based model interpretability

To enhance the transparency of model decision-making, we applied SHAP analysis at both the global and case levels. At the global level, the cross-case sentence-level SHAP summary plot (Figure 7(a)) demonstrated distinct contribution patterns between admission records and CTU descriptions. CTU features displayed a wider distribution and more extreme SHAP values, suggesting a stronger case-specific influence on prediction outcomes, whereas admission records contributed more consistently but with moderate magnitudes. The averaged SHAP values (Figure 7(b)) further confirmed that both clinical texts contributed positively to HG prediction and negatively to LG prediction, with CTU showing stronger effects in both directions (0.281 for HG, −0.366 for LG) compared with admission records (0.235 and −0.225, respectively). These findings indicate that while admission texts provide broad contextual information, CTU descriptions capture lesion-focused and discriminative features that play a decisive role in grade classification.

At the case level, we examined the top four sentence-level SHAP contributions from both the admission records

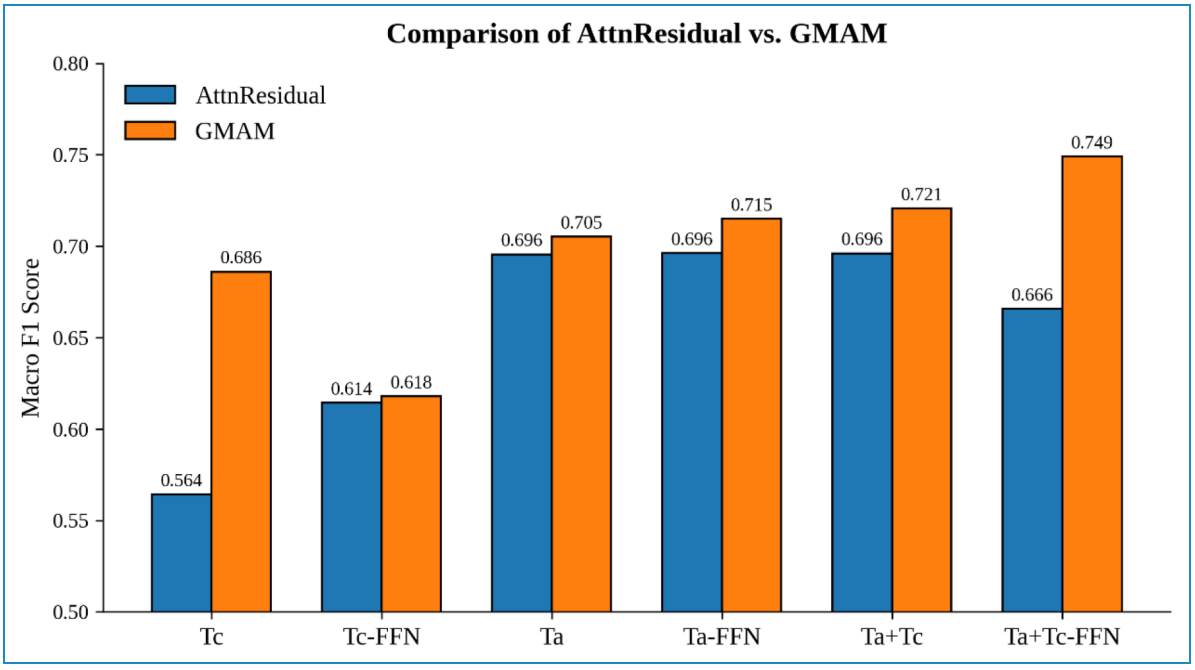

Figure 3. Comparison of macro *F*1 scores between AttnResidual and GMAM across different configurations. The results show that GMAM consistently outperforms AttnResidual in both single-text (T_c, T_a) and multitext ($T_a + T_c$) settings. The GMAM-FFN model achieves the best overall performance, highlighting the benefit of combining gated multiplicative attention with an FFN. GMAM: gated multiplicative attention mechanism; T_c: computed tomography urography description; T_a: admission record; FFN: feed-forward network.

and CTU descriptions for a representative patient (Figure 8). In Figure 8(a), positive contributions mainly arose from the chief complaint of painless gross hematuria and the progression of untreated hematuria with irritative urinary symptoms, whereas negative contributions were associated with normal systemic examination findings. In Figure 8(b), CTU descriptions exerted stronger overall influence, with sentences describing bladder wall thickening with large intraluminal filling defects and multiple enlarged lymph nodes strongly promoting HG prediction, while vertebral nodularity and mild hydronephrosis contributed negatively. Together, these findings highlight the complementary roles of admission narratives and CTU imaging in shaping the model's predictions, providing interpretable insights into how clinical and radiological features jointly drive grade classification.

Discussion

This study proposes an innovative approach for the noninvasive preoperative prediction of BC histological grade by leveraging long-form Chinese medical texts extracted from EMRs, specifically admission records and CTU descriptions. By integrating both text sources through an attention-based fusion strategy, predictive performance was significantly enhanced, achieving a maximum macro *F*1 score of 0.749 and a balanced accuracy of 0.757 with a reduced embedding dimension of 256. These findings demonstrate the feasibility of applying LLMs, exemplified by HuatuoGPT-7B, to capture latent features from Chinese medical narratives for BC grading.

Our results suggest that attention-based fusion offers clear advantages over simply mixed strategies. The superior performance of attention-based fusion arises from its ability to perform selective integration of T_a and T_c. Unlike simple concatenation, which passively merges embeddings and may introduce redundant or noisy features, the attention mechanism allows T_a to query T_c and highlight complementary cues while suppressing irrelevant signals. This context-aware reweighting preserves the richer information in admission records, while amplifying useful details from CTU descriptions only when they add value. As a result, attention-based fusion achieves a more balanced trade-off across classes, particularly improving LG sensitivity without sacrificing HG detection.

The dimensionality analysis suggests that an intermediate embedding size (e.g. 256 dimensions) may provide a favorable trade-off between capturing sufficient semantic information and mitigating overfitting. This finding aligns with broader observations in NLP research, where excessively high-dimensional embeddings often yield diminishing returns.[45,46] Clinically, the relative instability observed in LG predictions highlights the intrinsic difficulty of distinguishing subtle textual cues for these cases. This

Table 4. Comparison of models with different reduced dimensions.

Dimensionality	Acc	Bal-acc	*F*1-W	*F*1-M	AUC-M	Sen-LG/Spe-LG	Sen-HG/Spe-HG
$D = 64$	0.784	0.747	0.785	0.744	**0.762**	0.655/0.838	0.838/0.655
$D = 128$	**0.794**	0.715	0.783	0.731	0.742	0.517/0.912	0.912/0.517
$D = 256$	0.784	**0.757**	**0.786**	**0.749**	0.740	**0.690**/0.824	0.824/**0.690**
$D = 512$	0.784	0.707	0.774	0.721	0.723	0.517/0.897	0.897/0.517
$D = 1024$	0.773	0.680	0.757	0.696	0.707	0.448/0.912	0.912/0.448
$D = 2048$	0.784	0.688	0.766	0.705	0.727	0.448/**0.927**	**0.927**/0.448

Acc: accuracy; Bal-acc: balanced accuracy; *F*1-W: weighted *F*1 score; *F*1-M: macro *F*1 scores; AUC-M: macro area under the ROC curve; Sen: sensitivity; Spe: specificity; LG: low grade; HG: high grade.
Bold indicates the best results.

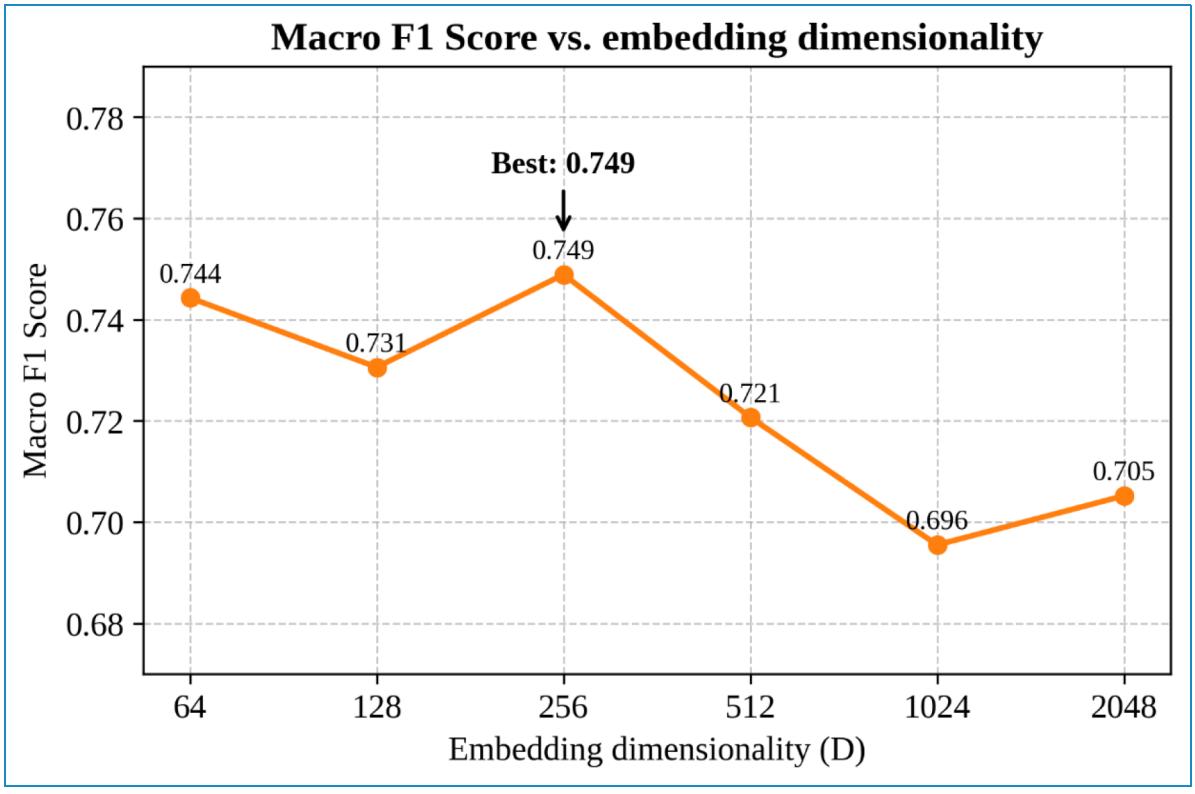

Figure 4. Macro *F*1 score versus embedding dimensionality for the proposed classification model. The plot illustrates the effect of varying embedding dimension *D* on the model's macro *F*1 score performance.

underscores the need for advanced modeling strategies, such as incorporating multimodal information or domain-specific pretraining, to enhance the sensitivity of models in identifying LG BC.

Beyond LLM-based frameworks, numerous studies have demonstrated that optimization strategies can substantially enhance the performance and generalization of classification models in diverse application domains. For instance, metaheuristic algorithms and hybrid optimization techniques have been employed for oral cancer detection,[47] breast cancer recognition,[48] weed detection in precision agriculture,[49] and intrusion detection in IoT systems.[50] Similarly, recent innovations in hybrid feature selection methods[51] have highlighted the importance of integrating optimization into model design. Although these works are applied in different contexts, they share the common methodological goal of

Table 5. Performance comparison of different imbalance-handling loss functions.

Loss	Acc	Bal-acc	*F*1-W	*F*1-M	AUC-M	Sen-LG/Spe-LG	Sen-HG/Spe-HG
WCE	0.732	**0.759**	0.743	0.716	0.747	**0.828**/0.691	0.691/**0.828**
Focal-IFW	0.763	0.752	0.769	0.734	**0.756**	0.724/0.779	0.779/0.724
CS-CE	**0.784**	0.757	**0.786**	**0.749**	0.740	0.690/**0.824**	**0.824**/0.690

Acc: accuracy; Bal-acc: balanced accuracy; *F*1-W: weighted *F*1 score; *F*1-M: macro *F*1 scores; AUC-M: macro area under the ROC curve; Sen: sensitivity; Spe: specificity; LG: low grade; HG: high grade; WCE: weighted cross-entropy; Focal-IFW: focal loss with inverse-frequency weighting; CS-CE: cost-sensitive cross-entropy.
Bold indicates the best results.

Table 6. Comparative performance of baseline models and the proposed HuatuoGPT-7B framework on the BCgrading-Text dataset (internal) and the external validation dataset.

	BCgrading-Text dataset			External validation dataset		
Models	Bal-acc	*F*1-M	AUC-M	Bal-acc	*F*1-M	AUC-M
Bio_ClinicalBERT	0.609	0.613	0.623	0.456	0.412	0.605
Bert-base-Chinese	0.616	0.622	0.656	0.614	0.621	0.682
Qwen3-em-0.6B	0.724	0.737	0.726	0.639	0.649	0.710
Qwen3-em-4B	0.663	0.665	0.665	0.626	0.626	**0.716**
Qwen3-em-8B	0.700	0.727	**0.760**	0.608	0.614	0.660
HuatuoGPT-7B (ours)	**0.757**	**0.749**	0.740	**0.676**	**0.680**	0.700

Bal-acc: balanced accuracy; *F*1-M: macro *F*1 scores; AUC-M: macro area under the ROC curve.
Bold indicates the best results.

improving classification reliability under challenging data conditions. Our adoption of a CS-CE loss function follows the same principle, ensuring that clinically adverse misclassifications are more heavily penalized. Situating our model within this broader optimization-driven research landscape underscores the value of domain-tailored optimization for improving robustness in BC grading tasks.

A notable advantage of our framework lies in its interpretability. While deep-learning models are often criticized as "black boxes," the integration of SHAP analysis provided transparent explanations of how admission records and CTU descriptions influenced predictions. Interestingly, although admission records alone achieved higher overall classification performance than CTU descriptions, SHAP analysis revealed that CTU features exerted more extreme case-specific contributions, whereas admission features contributed more consistently but with moderate magnitudes. This pattern underscores their complementary roles: admission records offer broad contextual information that stabilizes model performance, while CTU descriptions capture lesion-focused cues that can strongly drive decisions in individual cases.

Rather than reiterating performance metrics, interpretability highlights the clinical plausibility of the model—features such as hematuria history, bladder wall thickening, and nodal involvement were consistently identified as key drivers of HG classification, aligning well with established diagnostic reasoning in urology. This alignment between model-derived attributions and clinical knowledge strengthens the credibility of our approach and supports its potential for clinician acceptance. Moreover, interpretability facilitates error analysis by revealing cases where irrelevant or nonspecific findings attenuate predictions, thereby offering concrete opportunities for model refinement. In the broader context of AI in medicine, such transparent interpretability is increasingly recognized as a prerequisite for safe deployment and ethical clinical integration.

Although our study remains at an experimental stage, it is important to consider potential paths for clinical translation. The proposed model could be embedded into HISs as a background decision-support module, where admission records and CTU descriptions are automatically processed and the predicted grade is presented as an advisory flag for clinicians. Deployment would require moderate

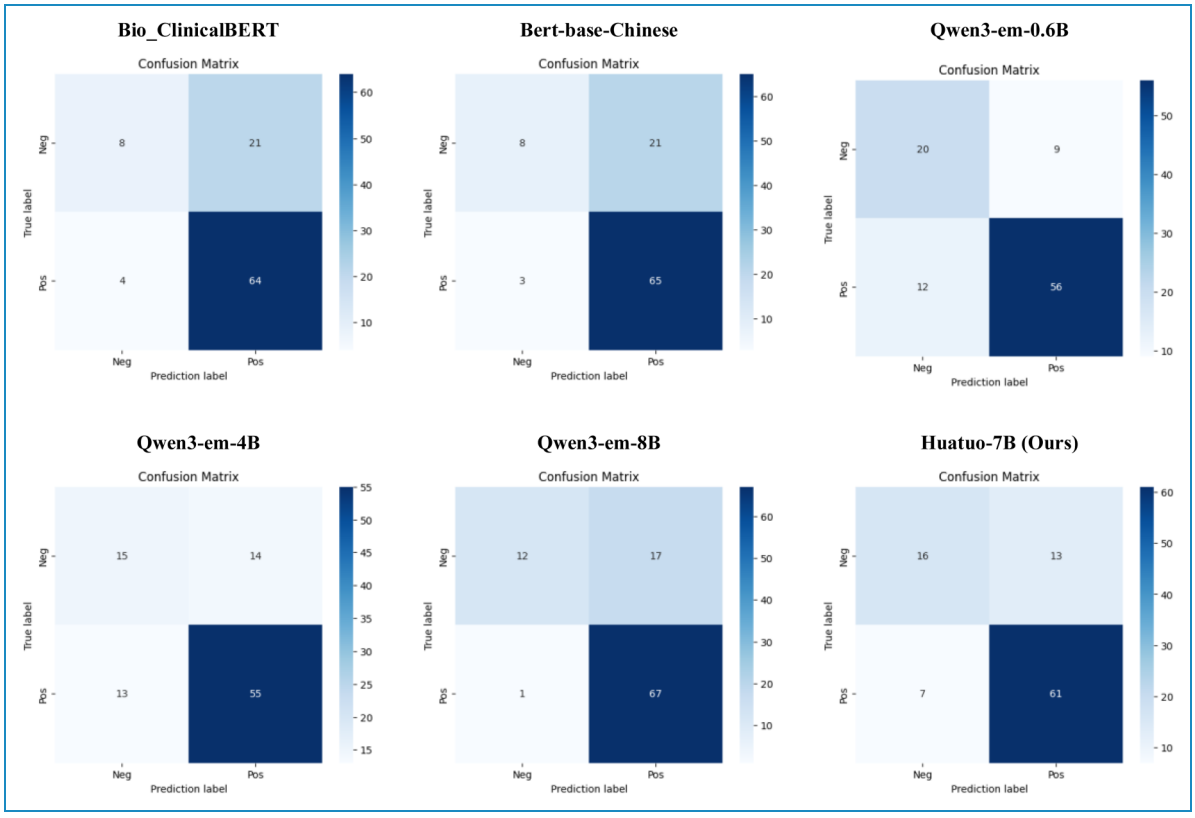

Figure 5. Confusion matrices of baseline models and HuatuoGPT-7B on the BCgrading-Text test set. HuatuoGPT-7B demonstrated more balanced classification between LG and HG cases compared with other models.

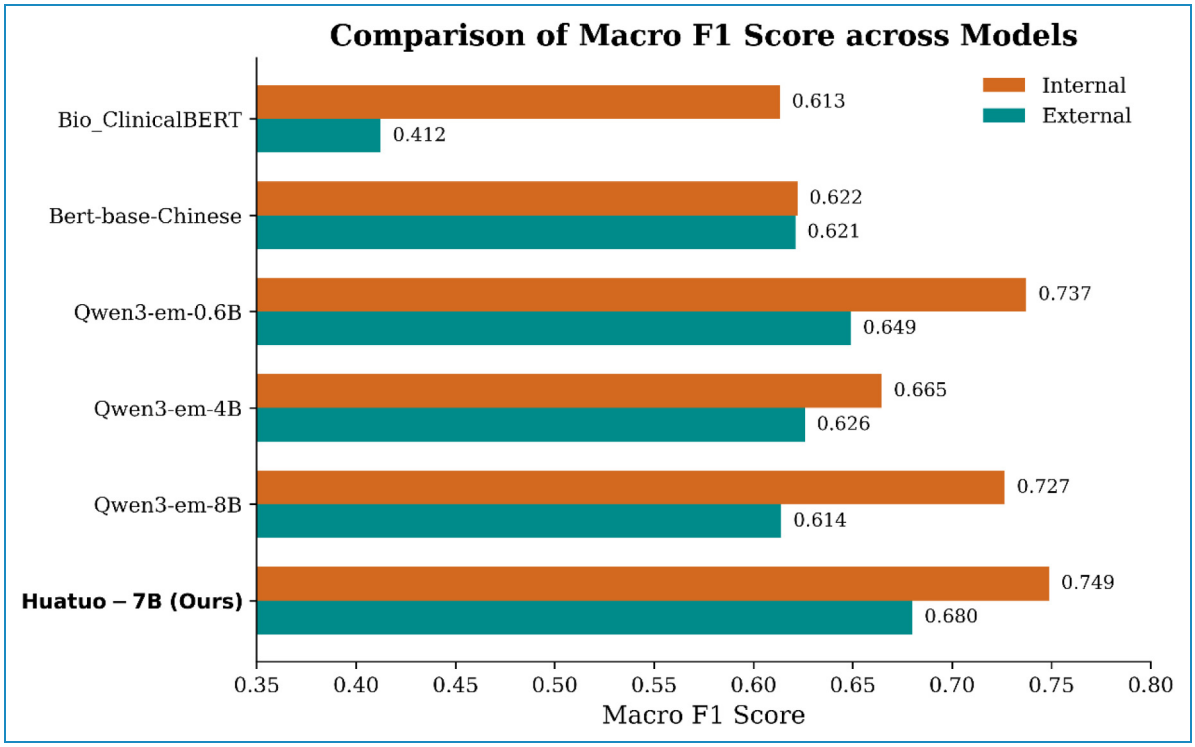

Figure 6. Macro *F*1 scores of different models on the BCgrading-Text dataset and the external validation dataset. All models showed reduced performance under distributional shift, whereas HuatuoGPT-7B maintained the most stable performance across datasets.

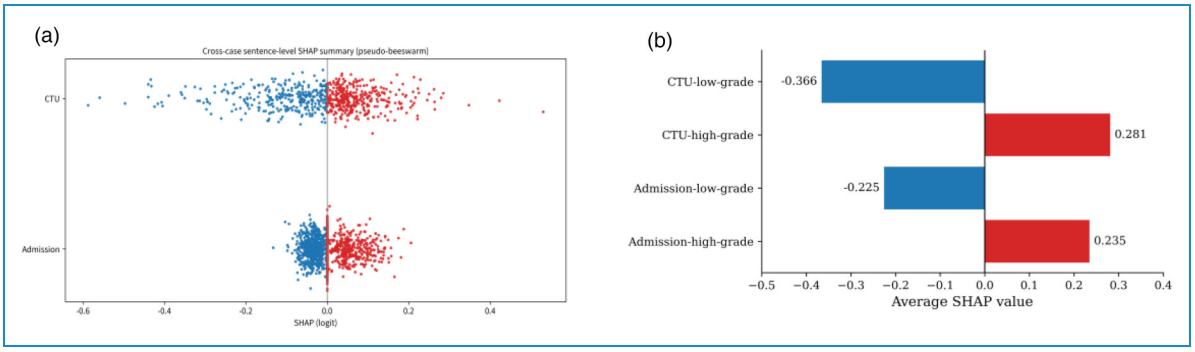

Figure 7. SHAP interpretability of admission records and CTU descriptions in BC grading. (a) Sentence-level SHAP summary plot showing broader and more extreme contributions from CTU compared with admission. (b) Average SHAP values by grade, indicating stronger predictive impact of CTU for both HG and LG cases. SHAP: Shapley additive explanations; CTU: computed tomography urography; BC: bladder cancer.

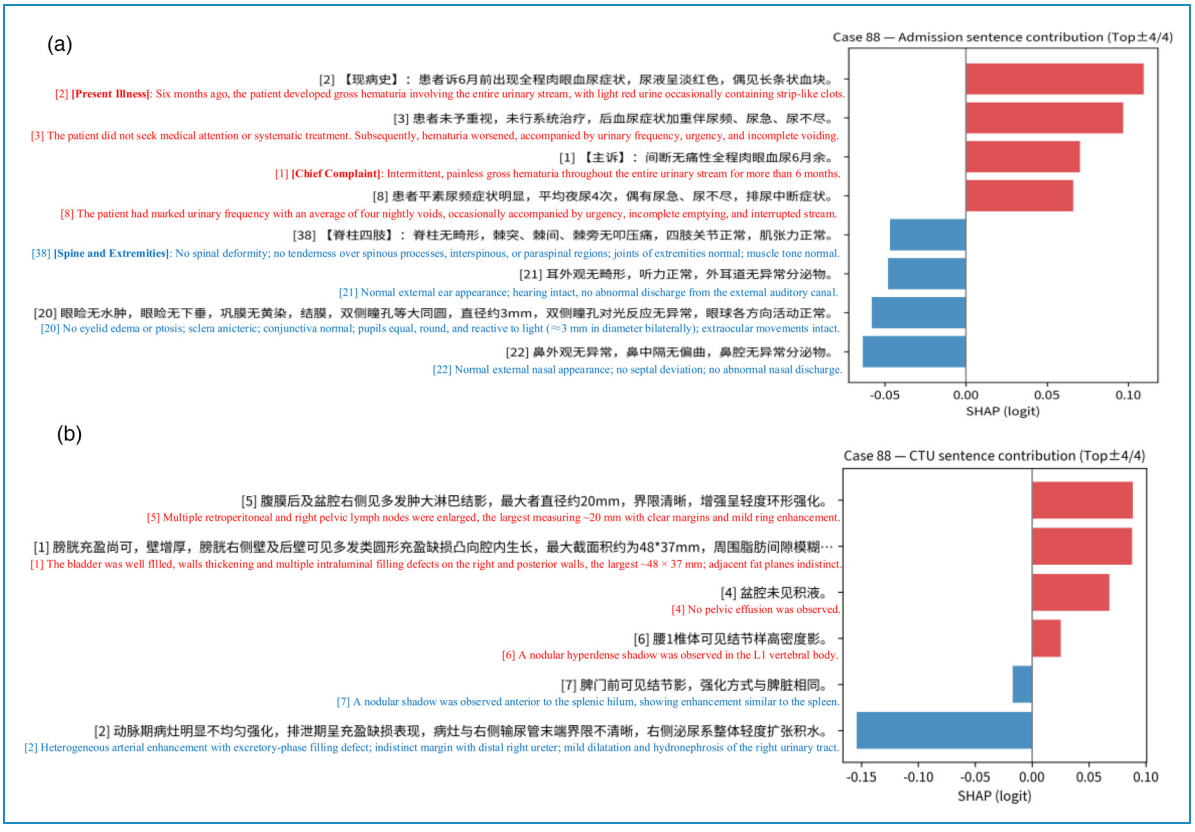

Figure 8. Case-level SHAP interpretability. (a) Admission records: symptom history supported HG prediction, whereas normal systemic findings attenuated it. (b) CTU descriptions: tumor burden and nodal enlargement strongly promoted HG classification, while vertebral and hydronephrotic findings contributed negatively. SHAP: Shapley additive explanations; CTU: computed tomography urography.

computational resources, such as an on-premise server with GPU or CPU acceleration, and the entire pipeline could operate within the hospital network to ensure data privacy. In practice, such a system may reduce the time clinicians spend manually reviewing lengthy clinical narratives and lower the risk of misclassification by providing an additional, objective preoperative reference. Importantly, the tool is designed to support rather than replace pathological confirmation, and further multisite validation is required before real-world use.

Despite these promising results, several limitations remain. First, although an external validation was conducted using an independent dataset from another subspecialty within the same hospital, the study remains a single-center retrospective analysis with limited sample size. The heterogeneity in documentation provided a proxy for real-world variability, but true multicenter datasets are still needed to confirm robustness and generalizability. Second, the framework relies on Chinese-specific LLMs, and the embeddings from HuatuoGPT-7B are not directly transferable to other languages or healthcare systems, which restricts its broader applicability. Although comparative experiments with Bio_ClinicalBERT, BERT-base-Chinese, and the Qwen3-Embedding series demonstrated the superiority of HuatuoGPT-7B in balanced accuracy and macro $F1$, future work will need to explore domain adaptation, multilingual LLMs, and cross-lingual transfer learning to enhance generalizability. Third, performance may be further improved by domain-specific fine-tuning on larger and more diverse BC datasets. Finally, integrating multimodal data such as imaging and laboratory results could enhance model robustness and interpretability, supporting more reliable clinical decision-making.

Conclusions

This study presents an interpretable LLM-based framework for the preoperative prediction of BC grade using Chinese medical long-texts. By fusing admission records and CTU descriptions with an attention-based strategy, optimizing embedding dimensionality, and incorporating a cost-sensitive loss function, the model achieved robust and balanced performance while maintaining clinical plausibility. SHAP analysis further highlighted complementary contributions of the two text sources and enhanced transparency of model decision-making. Although the current work remains limited by its single-center, retrospective nature, the findings underscore the potential of LLM-driven approaches to support clinical decision-making in BC and provide a foundation for future multicenter and multimodal research.

Abbreviations

AI	artificial intelligence
AUC	area under the ROC curve
BC	bladder cancer
CS-CE	cost-sensitive cross-entropy
CTU	computed tomography urography
EMR	electronic medical record
FFN	feed-forward network
FN	false negative
Focal-IFW	focal loss with inverse-frequency weighting
FP	false positive
GMAM	gated multiplicative attention mechanism
HG	high grade
HIS	hospital information system
LG	low grade
LLM	large language model
MIBC	muscle-invasive bladder cancer
NLP	natural language processing
NMIBC	non-muscle-invasive bladder cancer
PUNLMP	papillary urothelial neoplasm of low malignant potential
SHAP	Shapley additive explanations
TN	true negative
TP	true positive
WCE	weighted cross-entropy

Acknowledgements

We would like to thank Dr Bo Yang, Chief Surgeon of the Department of Urology at the Second Affiliated Hospital of Dalian Medical University, and Dr Xiling Zhang, Associate Chief Surgeon of the Department of Urology at the Fourth Affiliated Hospital of China Medical University, for their valuable contributions to the data annotation process and their expert clinical insights.

ORCID iDs

Xianwei Pan https://orcid.org/0000-0002-6397-8079
Lijie Wen https://orcid.org/0000-0002-7016-919X
Yuhua Li https://orcid.org/0000-0003-2913-4478
Yijia Zhang https://orcid.org/0000-0002-5843-4675
Mingyu Lu https://orcid.org/0000-0002-8663-9870

Ethics approval and consent to participate

All procedures performed in our study involving human participants were conducted in accordance with the 1964 Helsinki Declaration and its subsequent amendments or comparable ethical standards. This study was approved by the Institutional Review Board of the Second Hospital of Dalian Medical University (Approval No. KY2025-541-01). Owing to the retrospective design and the use of anonymized clinical data, the requirement for informed consent was waived by the ethics committee.

Author contributions

XWP formulated the framework, conducted the experiments, summarized the results, and wrote the manuscript. LJW collected, interpreted, and annotated the patient data related to bladder cancer, and was a major contributor to manuscript writing and funding acquisition. YHL designed the model framework, analyzed the results, and revised the manuscript. YJZ interpreted the experimental findings, contributed to result discussion, and revised the manuscript. MYL conceptualized the study, supervised the experiments, and secured the funding. All authors read and approved the final version of the manuscript.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Second Hospital of Dalian Medical University, National Natural Science Foundation of China (grant number 2023, No. 62372077).

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Availability of data and materials

The dataset constructed and analyzed during the current study are available from the corresponding author on reasonable request.

Use of AI tools

During the preparation of this manuscript, generative AI tools (ChatGPT) were used solely for language editing and formatting assistance. All scientific content, data analysis, and interpretation were performed entirely by the authors, who take full responsibility for the accuracy and integrity of the work.

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2024; 74: 229–263.
2. Mokkapati S, Manyam G, Steinmetz AR, et al. Molecular profiling of bladder cancer xenografts defines relevant molecular subtypes and provides a resource for biomarker discovery. *Transl Oncol* 2025; 52: 102269.
3. Su Y, Chen L and Yang J. Hesperetin inhibits bladder cancer cell proliferation and promotes apoptosis and cycle arrest by PI3 K/AKT/FoxO3a and ER stress-mitochondria pathways. *Curr Med Chem* 2025; 32: 3879–3904.
4. Sylvester RJ, Van der Meijden AP, Oosterlinck W, et al. Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. *Eur Urol* 2006; 49: 466–477.
5. Van Rhijn BW, Burger M, Lotan Y, et al. Prognostic value of the WHO1973 and WHO2004/2016 classification systems for grade in primary Ta/T1 non-muscle-invasive bladder cancer: a multicenter European association of urology non-muscle-invasive bladder cancer guidelines panel study. *Eur Urol Oncol* 2021; 4: 182.
6. WHO Classification of Tumours Editorial Board. *WHO classification of tumours: urinary and male genital tumours*. Lyon, France, IARC; 2022.
7. Jones TD and Cheng L. Reappraisal of the papillary urothelial neoplasm of low malignant potential (PUNLMP). *Histopathology* 2020; 77: 525–535.
8. Gontero P, Birtle A, Capoun O, et al. European Association of Urology guidelines on non-muscle-invasive bladder cancer (TaT1 and carcinoma in situ)—a summary of the 2024 guidelines update. *Eur Urol* 2024; 86: 531–549.
9. Zhou M, Zhang Z, Bao S, et al. Computational recognition of lncRNA signature of tumor-infiltrating B lymphocytes with potential implications in prognosis and immunotherapy of bladder cancer. *Brief Bioinform* 2021; 22: bbaa047.
10. Meng XY, Zhou XH, Li S, et al. Machine learning-based detection of bladder cancer by urine cfDNA fragmentation hotspots that capture cancer-associated molecular features. *Clin Chem* 2024; 70: 1463–1473.
11. Teng F, Zhang R, Wang Y, et al. Machine learning and Mendelian randomization reveal a tumor immune cell profile for predicting bladder cancer risk and immunotherapy outcomes. *Am J Pathol* 2025; 195: 1141–1157.
12. Woerl AC, Eckstein M, Geiger J, et al. Deep learning predicts molecular subtype of muscle-invasive bladder cancer from conventional histopathological slides. *Eur Urol* 2020; 78: 256–264.
13. Dong Q, Huang D, Xu X, et al. Content and shape attention network for bladder wall and cancer segmentation in MRIs. *Comput Biol Med* 2022; 148: 105809.
14. Freitas NR, Vieira PM, Cordeiro A, et al. Detection of bladder cancer with feature fusion, transfer learning and CapsNets. *Artif Intell Med* 2022; 126: 102275.
15. Sun R, Zhang M, Yang L, et al. Preoperative CT-based deep learning radiomics model to predict lymph node metastasis and patient prognosis in bladder cancer: a two-center study. *Insights Imaging* 2024; 15: 21.
16. Jiang F, Hong G, Zeng H, et al. Deep learning-based model for prediction of early recurrence and therapy response on whole slide images in non-muscle-invasive bladder cancer: a retrospective, multicentre study. *EClinicalMedicine* 2025; 81: 103125.
17. Jobczyk M, Stawiski K, Kaszkowiak M, et al. Deep learning-based recalibration of the CUETO and EORTC prediction tools for recurrence and progression of non-muscle-invasive bladder cancer. *Eur Urol Oncol* 2022; 5: 109–112.
18. Zou XC, Rao XP, Huang JB, et al. Predicting distant metastasis of bladder cancer using multiple machine learning models: a study based on the SEER database with external validation. *Front Oncol* 2024; 14: 1477166.
19. Wetteland R, Giskeødegård GF, Tessem MB, et al. Automatic diagnostic tool for predicting cancer grade in bladder cancer patients using deep learning. *IEEE Access* 2021; 9: 115813–25.
20. García G, Esteve A, Colomer A, et al. A novel self-learning framework for bladder cancer grading using histopathological images. *Comput Biol Med* 2021; 138: 104932.
21. Fuster S, Kiraz U, Eftestøl T, et al. NMGrad: advancing histopathological bladder cancer grading with weakly supervised deep learning. *Bioengineering (Basel)* 2024; 11: 909.
22. Song H, Yang S, Yu B, et al. CT-based deep learning radiomics nomogram for the prediction of pathological grade in bladder cancer: a multicenter study. *Cancer Imaging* 2023; 3: 89.
23. Deng Z, Dong W, Xiong S, et al. Machine learning models combining computed tomography semantic features and selected clinical variables for accurate prediction of the pathological grade of bladder cancer. *Front Oncol* 2023; 13: 1166245.
24. Zhu X, Qin R, Qu K, et al. Atomic force microscopy-based assessment of multimechanical cellular properties for

classification of graded bladder cancer cells and cancer early diagnosis using machine learning analysis. *Acta Biomater* 2023; 158: 358–373.
25. Zhang H, Yang X, Bai L, et al. Enhancing drug recommendations via heterogeneous graph representation learning in EHR networks. *IEEE Trans Knowl Data Eng* 2024; 36: 3024–3035.
26. Li X, Meng F, Wei J, et al. Towards robust drug recommendation based on dual perspective encoder and iterative denoising mechanism. *Expert Syst Appl* 2025; 283: 127784.
27. Ge X, Satpathy A, Williams RD, et al. *DKEC: domain knowledge enhanced multi-label classification for diagnosis prediction*. In: *Proceedings of the 2024 conference on empirical methods in natural language processing*, Miami, Florida, USA, 2024. pp.12798–12813.
28. Cheng H, Li S, Shen T, et al. Enhancing diagnosis prediction with adaptive disease representation learning. *Artif Intell Med* 2025; 163: 103098.
29. Yoo Y and Kim S. How to leverage large language models for automatic ICD coding. *Comput Biol Med* 2025; 189: 109971.
30. Huang CW, Tsai SC and Chen YN. *PLM-ICD: automatic ICD coding with pretrained language models*. In: *Proceedings of the 4th clinical natural language processing workshop*, Seattle, WA, 2022, pp.10–20.
31. Luo J, Wang X, Wang J, et al. *Correlation: boosting automatic ICD coding through contextualized code relation learning*. In: *Proceedings of the 2024 Joint international conference on computational linguistics, language resources and evaluation (LREC-COLING)*, Torino, Italia, 2024, pp.3997–4007.
32. Narayan VM, Siolas D, Meadows ES, et al. Evaluation of a natural language processing model to identify and characterize patients in the United States with high-risk non-muscle-invasive bladder cancer. *JCO Clin Cancer Inform* 2023; 7: e2300096.
33. Schroeck FR, Patterson OV, Alba PR, et al. Development of a natural language processing engine to generate bladder cancer pathology data for health services research. *Urology* 2017; 110: 84–91.
34. Yang R, Zhu D, Howard LE, et al. Context-based identification of muscle invasion status in patients with bladder cancer using natural language processing. *JCO Clin Cancer Inform* 2022; 6: e2100097.
35. Zhang H, Chen J, Jiang F, et al. HuatuoGPT: towards taming language model to be a doctor. In: *Findings of the association for computational linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, 2023, pp.10859–10885.
36. Elkan C. *The foundations of cost-sensitive learning*. In: *Proceedings of the 17th international joint conference on artificial intelligence (IJCAI)*. 2001, pp.973–978.
37. Zhou ZH and Liu XY. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Trans Knowl Data Eng* 2006; 18: 63–77.
38. King G and Zeng L. Logistic regression in rare events data. *Polit Anal* 2001; 9: 137–163.
39. Lin TY, Goyal P, Girshick R, et al. *Focal loss for dense object detection*. In: *Proceedings of the IEEE international conference on computer vision (ICCV)*, Venice, Italy, 2017, pp.2999–3007.
40. Alsentzer E, Murphy J, Boag W, et al. Publicly available clinical BERT embeddings. In: *Proceedings of the 2nd clinical natural language processing workshop (ClinicalNLP)*. Minneapolis, MN. Stroudsburg: Association for Computational Linguistics; 2019, pp.72–78.
41. Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT 2019*. Minneapolis, MN. Stroudsburg: Association for Computational Linguistics; 2019, pp.4171–4186.
42. Bai J, Chen H, Chen S, et al. Qwen technical report. arXiv preprint arXiv: 2309.16609. 2023. https://arxiv.org/abs/2309.16609.
43. Lundberg SM and Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, et al. (eds) *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30. New York, NY: Curran Associates, Inc., 2017, pp.4765–4774.
44. He K, Zhang X, Ren S, et al. *Deep residual learning for image recognition*. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp.770–778.
45. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. arXiv, https://arxiv.org/abs/1301.3781 (2013, accessed 21 June 2025).
46. Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. arXiv. https://arxiv.org/abs/2001.08361 (2020, accessed 21 June 2025).
47. Thippa Reddy G, Kaluri R, Lakshmanna K, et al. Advanced meta-heuristic algorithm based on particle swarm and Al-Biruni earth radius optimization methods for oral cancer detection. *Comput Mater Continua* 2023; 74: 3265–3281.
48. Yildirim M, Hameed IA, Oliva D, et al. Breast cancer detection based on metaheuristic optimization and machine learning. *Comput Mater Continua* 2023; 75: 291–308.
49. Ullo SL, Russo M, Peng H, et al. Weed detection in precision agriculture using deep learning and metaheuristic optimization. *Mathematics* 2022; 10: 4421.
50. Ramalingam S, Elhoseny M, Suresh A, et al. An intrusion detection system using optimized deep learning model for IoT applications. *IEEE Access* 2023; 11: 41671–41684.
51. Zaguia A, Elaziz MA, Dahou A, et al. Innovative feature selection method based on hybrid sine cosine and dipper throated optimization algorithms. *IEEE Access* 2023; 11: 84528–84541.