Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Clinical-Prompt-Enhanced Deep Learning Model for Precise Preoperative Prediction of Gallbladder Polypoid Lesion Subtypes

**Qingyu Tang**
First Affiliated Hospital of Xi'an Jiaotong University
**Zhibo Wang**
Xi'an Jiaotong University
**Zhenqi Tang**
First Affiliated Hospital of Xi'an Jiaotong University
**Hengchao Liu**
First Affiliated Hospital of Xi'an Jiaotong University
**Hongzhan Wang**
First Affiliated Hospital of Xi'an Jiaotong University
**Yubo Ma**
First Affiliated Hospital of Xi'an Jiaotong University
**Jiashu Song**
First Affiliated Hospital of Xi'an Jiaotong University
**Kangpeng Li**
First Affiliated Hospital of Xi'an Jiaotong University
**Minghui Dou**
First Affiliated Hospital of Xi'an Jiaotong University
**Ziyang Peng**
First Affiliated Hospital of Xi'an Jiaotong University
**Dong Zhang**
First Affiliated Hospital of Xi'an Jiaotong University
**Zhimin Geng**
First Affiliated Hospital of Xi'an Jiaotong University
**Qi Li**
docliqi@163.com

First Affiliated Hospital of Xi'an Jiaotong University

**Research Article**

**Keywords:** Gallbladder polypoid lesions, Cholecystectomy, Multimodal deep learning, Preoperative classification, Ultrasound imaging, Clinical decision support

**Posted Date:** July 9th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10013216/v1

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

**Additional Declarations:** No competing interests reported.

## Abstract

# Background

To develop a clinical-prompt-enhanced deep learning framework that noninvasively predicts gallbladder polypoid lesion (GBPL) pathology across cholesterol polyps, adenomas, and carcinomas using preoperative ultrasound and routinely available clinical text, addressing overtreatment associated with size-based assessment.

# Methods

We retrospectively enrolled 516 cholecystectomy patients with histologically confirmed GBPLs (January 2015–January 2023). Patients were randomly stratified into training set (n = 362) and an independent test set (n = 154). Lesions were segmented using nnU-Net v2 to construct region-of-interest (ROI)-guided inputs, achieving a mean Dice score of 0.797 ± 0.02. We developed GB-DualmodalNet, which converts clinical variables into natural-language prompts via semantic mapping and fuses with ROI-guided VGG16 image features using cross-modal attention strategy. Performance was assessed on the independent test set using macro-averaged area under the receiver operating characteristic curve (AUC), accuracy, and standard diagnostic metrics. A two-reader study on the independent test set was performed to evaluate changes in clinician performance with artificial intelligence (AI) assistance.

# Results

In the test cohort, the ROI-guided image-only model achieved macro-AUC 0.840 (95% CI 0.787–0.890). GB-DualmodalNet achieved the highest macro-AUC of 0.852 (95% CI 0.801–0.902), with accuracy 70.1% and specificity 79.2%. Subtype-specific performance yielded AUC 0.933 for carcinoma, 0.758 for adenoma, and 0.865 for cholesterol polyps. In 10–15 mm gray-zone subgroup (n = 48), GB-DualmodalNet correctly classified 19 of 22 benign lesions as low risk, corresponding to a specificity of 86.4%. Grad-CAM + + highlighted lesion margins and the attachment base, and SHapley Additive exPlanations (SHAP) identified long diameter, short diameter, and age as the most influential clinical features. AI assistance improved pooled clinician macro-AUC from 0.782 to 0.865.

# Conclusion

GB-DualmodalNet provides an automated, interpretable approach for preoperative three-class GBPL prediction and improves performance over image-only and clinical-only baselines. Multimodal deep learning can improve GBPL risk prediction and reduce unnecessary cholecystectomies, though prospective multicenter validation is required before clinical use.

## 1. Introduction

Gallbladder polypoid lesions (GBPLs) are focal mucosal protrusions detected in approximately 5% of adults undergoing abdominal ultrasonography, and the majority are identified incidentally[1]. Histologically, GBPLs span a clinically heterogeneous spectrum: cholesterol polyps and other non-neoplastic lesions, which have negligible malignant potential and are typically managed by surveillance; adenomas, which are premalignant and generally treated by simple cholecystectomy; and gallbladder carcinomas, for which radical or extended resection may be required when invasive disease is suspected[2–4]. Because different pathological subtypes carry markedly different malignant potential and management strategies, accurate preoperative subtyping and risk prediction are clinically essential to align surgical management with the underlying biology of the lesion.

Current international guidelines, including the recent ESGAR/EAES/EFISDS/ESGE update[4], the Society of Radiologists in Ultrasound consensus[5], and the updated Canadian Association of Radiologists recommendations[6], rely primarily on lesion size and a small number of risk factors to decide between surveillance and cholecystectomy, with a long diameter of 10 mm widely used as the surgical threshold[7]. This size-based criterion has well documented limitations: in a prospective evaluation it achieved a sensitivity of 68.1% and a specificity of 70.2% for neoplastic polyps, which translates into a substantial proportion of unnecessary cholecystectomies for cholesterol polyps and, conversely, missed neoplasia in some lesions below the cut-off[7]. The 10–15 mm "gray zone" is particularly problematic because most lesions in this range still prove benign at histopathology, yet most are currently resected[8, 9]. A non-invasive preoperative tool able to refine this decision is therefore of direct surgical interest.

Deep learning applied to transabdominal ultrasound has shown promise in this setting, but most published models address only binary classification (neoplastic versus non-neoplastic) or focus exclusively on gallbladder cancer detection[10–13].Kim et al. trained an ensemble convolutional neural network on 501 cases and reported an area under the receiver operating characteristic curve (AUC) of 0.896 with images alone, rising to 0.908 after adding age and polyp size[12]. Gupta et al. recently reported a prospective evaluation of a deep-learning classifier for gallbladder cancer in 273 patients, achieving an AUC of 0.887 with performance comparable to radiologists[10–13]. These studies confirm the technical feasibility of ultrasound-based AI, but they do not directly inform the surgical decision between surveillance, simple cholecystectomy, and consideration of radical resection.

In addition, when clinical variables are incorporated alongside images, they are typically concatenated as numerical features, which discards the semantic content that clinicians actually use during preoperative assessment[14].

We therefore developed and internally validated a multimodal deep learning model, GB-DualmodalNet, that integrates region-of-interest-guided ultrasound features with clinical variables encoded as natural-language prompts through a cross-modal attention mechanism, with the explicit objective of supporting preoperative three-class assessment of GBPLs within the surgical care pathway. A two-reader study was performed to evaluate whether model outputs could improve clinician diagnostic performance over routine size-based assessment, with particular attention to the 10–15 mm gray-zone subgroup most affected by overtreatment. The study was conducted and reported in accordance with the TRIPOD + AI statement. The source code is available on https://github.com/Qingyut/GB-DualmodalNet.

## 2. Methods

# 2.1 Study design and cohort

The present study was designed as a single-center, retrospective, observational diagnostic accuracy study. Data were collected from the First Affiliated Hospital of Xi'an Jiaotong University. The source population consisted of 557 consecutive adult patients who underwent cholecystectomy at the First Affiliated Hospital of Xi'an Jiaotong University between January 2015 and January 2023 and had a preoperative diagnosis of GBPL on transabdominal ultrasound. The final analytical cohort comprised 516 patients after exclusion applying predefined lesion diameter range (n = 9), image quality (n = 15), pathological confirmation (n = 6), and data-completeness (n = 11) criteria (Fig. 1). Lesions with a long diameter > 25 mm were excluded because such masses often no longer represent the typical spectrum of polypoid lesions for which preoperative subtype discrimination is clinically challenging, and their inclusion could artificially inflate discrimination by introducing morphologically obvious advanced cases. Carcinomas staged T3 or higher were excluded because the study focused on preoperative differentiation within surgically relevant polypoid lesions and early-stage malignant disease, whereas advanced tumors usually present as overt mass-forming or infiltrative lesions with different clinical decision pathways. Patients were randomly partitioned by stratified sampling into a training set (n = 362) and a test set (n = 154). Table 1 presents a detailed comparison of demographic variables, biochemical indicators, and ultrasonographic features between the training and test set. For each enrolled patient, the following clinical variables were extracted from the electronic medical record and ultrasound reports: demographic data (age, sex); routine laboratory indicators measured within one week before surgery (white blood cell count, red blood cell count, platelet count, hemoglobin, alanine aminotransferase, aspartate aminotransferase, alkaline phosphatase, gamma-glutamyl transferase, total bilirubin, direct bilirubin, albumin, and carcinoembryonic antigen); and ultrasound-reported lesion characteristics (long diameter, short diameter, number of polyps, polyp location, base morphology, gallbladder wall thickening, echogenicity, and polyp detection time). Availability of demographic data, pathological diagnosis, and at least the ultrasound-reported lesion characteristics listed above was required for inclusion; cases with missing values in any of these fields were excluded under the "Unavailable clinical data" criterion. Laboratory indicators were used only for descriptive comparison between the training and test sets and were not entered into the model.

The inclusion criteria were as follows:(1) Patients aged over 18 years; (2) Preoperative ultrasound diagnosis of GBPLs; (3) Underwent surgical treatment; (4) Availability of pathological diagnosis confirming GBPLs. The exclusion criteria: (1) Lesion long diameter > 25 mm; (2) Pathologically confirmed gallbladder carcinoma staged T3 or higher; (3) Unavailable clinical data; (4) Low image quality.

# 2.2 Ethics approval

The study protocol strictly adhered to the ethical principles of the Declaration of Helsinki and was approved by the Ethics Committee of the First Affiliated Hospital of Xi'an Jiaotong University (No. 2019G-40). Given the retrospective nature of the study and the exclusive use of de-identified patient data, the requirement for informed consent was waived.

# 2.3 Ultrasound image acquisition

All transabdominal ultrasound examinations were performed at the Department of Ultrasound, the First Affiliated Hospital of Xi'an Jiaotong University. Because the study cohort spanned an 8-year period (January 2015–January 2023), examinations were acquired on multiple commercially available ultrasound systems used in routine clinical practice during the study period, including Mindray Resona R9 (Mindray Biomedical Electronics Co., Ltd., Shenzhen, China), Philips EPIQ 7 (Philips Healthcare, Bothell, WA, USA), and GE LOGIQ E9 (GE Healthcare, Milwaukee, WI, USA)]. Convex array transducers with a frequency range of 1–6 MHz were primarily used for survey scanning of the gallbladder. All patients fasted for at least 8 hours before examination and were scanned in the supine and left lateral decubitus positions according to our institutional gallbladder imaging protocol. For each patient, representative gray-scale images that clearly demonstrated the maximal lesion dimension, the polyp base, the lesion contour, and internal echogenicity were retrospectively retrieved. All examinations were performed by 2 sonographers with more than 3 years of experience in hepatobiliary ultrasonography.

# 2.4 Image quality control and preprocessing modules

We developed an automated quality-control and preprocessing pipeline for ultrasound images. Corrupted or truncated images were excluded and logged. Images were then cropped to retain only the ultrasound fan region. All images passing quality control were resized to 224×224 pixels and normalized channel-wise using ImageNet mean and standard deviation values. To improve visualization of intralesional structures while avoiding

noise over-amplification, contrast-limited adaptive histogram equalization (CLAHE) was applied (clip limit = 2.0). Implementation details of CLAHE and dual-expert validated segmentation reference standard are provided in Supplementary Methods 1 and 2.

## 2.5 Lesion segmentation and ROI mask generation

nnU-Net v2 was used to delineate lesion regions and to provide the imaging input branch for multimodal models because of its high accuracy and strong generalization. All network configurations were trained using a 5-fold cross-validation approach. The nnU-Net v2 evaluated each configuration by calculating the average Dice score across all foreground classes, resulting in a scalar metric per configuration. The segmentation network was trained solely on lesion contours without using or inferring pathological labels. After training, the segmentation parameters were frozen and the model served as a unified preprocessing stage for all classification experiments. Because ROI masks encode only spatial priors and contain no malignancy information, they do not introduce label leakage. Analysis was performed at the patient level. In patients with multiple gallbladder polyps, the lesion with the largest long diameter was prioritized.

## 2.6 clinical prompt engineering and semantic mapping

We proposed GB-DualmodalNet as a novel dual-stream architecture that processes visual and textual information in parallel, followed by deep fusion via a cross-modal attention mechanism (Fig. 2). The detailed architecture of GB-DualmodalNet is presented in Supplementary Methods 3. In conventional medical AI models, structured clinical tabular data are typically processed through one-hot encoding or numerical normalization and subsequently fed into a multilayer perceptron (MLP). Although this approach preserves numerical information, it fails to capture the rich medical semantics underlying individual variables. To better exploit such domain-specific prior knowledge, we developed a descriptive clinical prompted module that transforms structured clinical variables into natural language descriptions. The collected variables included demographic characteristics, laboratory biomarkers, and ultrasound-reported lesion characteristics. Candidate variables were first extracted from the electronic medical record and ultrasound reports, after which ten clinically relevant variables were selected for the multimodal model based on domain knowledge and training-set-only univariable analysis, including age, sex, long diameter, short diameter, number of polyps, base morphology, gallbladder wall thickening, echogenicity, lesion location, and polyp detection time[15]. Each variable was mapped to standardized text descriptors using a dedicated semantic mapping table (Supplementary Table S1). Within the pretrained PubMedBERT embedding space, such descriptors lie closer to clinically related concepts, thereby encoding deeper semantic associations.

## 2.7 Visual explanations

To improve transparency and clinical interpretability, SHapley Additive exPlanations (SHAP) and Grad-CAM + + were applied to visualize the contributions of clinical variables and to highlight the imaging regions that drive model decisions[16, 17]. For each class, representative high-confidence correct and incorrect predictions were exported to support qualitative error analysis. Additionally, SHAP-based explanations were used to identify the features most influential for the model predictions[16].

## 2.8 Reader study

A reader study was conducted to evaluate the impact of AI assistance on clinician diagnostic performance. Two radiologists with > 3 years' experience in hepatobiliary ultrasonography independently reviewed all cases in the independent test set. Each clinician assessed the cases under two conditions: (1) unaided assessment based on routine ultrasonographic features and guideline-informed size criteria, and (2) AI-assisted assessment with access to the subtype probabilities and risk suggestions generated by GB-DualmodalNet. The order of cases was randomized for each reader and each condition, and a washout period was applied between the unaided and AI-assisted reading sessions to minimize recall bias. Clinicians were blinded to pathological results and to each other's assessments. In both conditions, clinicians were instructed to make a single three-class diagnosis for each case, with pathological diagnosis used as the reference standard. Diagnostic performance metrics were computed separately for each reader. Pooled performance was calculated by averaging reader-specific metrics. Macro-AUCs and corresponding 95% CIs were estimated using a bootstrap resampling strategy on the test set.

## 2.9 Statistical analysis

Statistical analysis was performed using Python version 3.10.16 (https://www.python.org). Continuous variables were summarized as Mean ± standard deviation (SD) and as median with interquartile range (IQR); their distribution was assessed using the Shapiro-Wilk test. Between the training and test sets, continuous variables were compared using the independent samples t-test when approximately normally distributed or the Mann–Whitney U test otherwise. Categorical variables were expressed as counts with percentages and compared using the chi-square test or Fisher's exact test. For model evaluation, we calculated accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and macro-averaged AUC. 95% CIs for AUC and other performance metrics were estimated using nonparametric bootstrapping with patient-level resampling. The primary endpoint was macro-averaged AUC for three-class classification of cholesterol polyps, adenomas, and gallbladder carcinoma in the independent test set. Secondary endpoints included accuracy, sensitivity, specificity, PPV, NPV, F1 score, subtype-specific AUCs, diagnostic performance in the 10–15 mm subgroup, and changes in clinician performance with AI assistance. Differences between correlated AUCs were assessed using the DeLong test. All statistical tests were two-tailed, and a P-value < 0.05 was considered statistically significant.

No formal a priori power calculation was performed because this was a retrospective model-development and internal-validation study. The sample size was determined by inclusion of all consecutive eligible patients during the study period. Precision of the primary performance estimate was

assessed using bootstrap-derived 95% confidence intervals.

# 3. Results

### *3.1 Performance of the proposed GB-DualmodalNet*

*3.1.1 Segmentation performance*

Segmentation masks were used to exclude all image pixels outside the ROI and polyp boundaries. An nnU-Net v2 segmentation deep learning model was applied as the preprocessing module of the ROI-guided model for subsequent diagnostic prediction, with 1000 epochs in each cross-validation fold. Performance was stable across folds (Dice 0.740–0.820), with an overall mean of 0.797 ± 0.029 (Supplementary Fig. S1a, S1b).

*3.1.2 Overall diagnostic performance comparison*

In addition to the original ultrasound images and segmented ROI masks, we incorporated clinical text from ultrasound reports as a third modality. Univariable and multivariable analyses revealed that all variables except sex and time since polyp detection were significantly associated with the outcome (Supplementary Table S2). Nevertheless, given that patient sex is an important and readily available clinical characteristic, and time since polyp detection is a relevant and easily accessible element of the medical history based on clinical experience, we retained both variables in the multimodal model.

Among ROI-guided image-only backbones, VGG16 achieved best performance (AUC = 0.840), followed by VGG19 (AUC = 0.829), SwinT (AUC = 0.802), ResNet18 (AUC = 0.801), ViT (AUC = 0.788), ResNet34 (AUC = 0.766), and DenseNet121 (AUC = 0.740) (Supplementary Table S3). We therefore selected VGG16 as the visual backbone because it provided the best discrimination in our dataset and, given the limited sample size, offered a favorable balance between locality bias, optimization stability, and model complexity. The image-only baseline model, which used VGG16 without clinical data, achieved an AUC of 0.608 without the visual preprocessing module and 0.840 with it (Supplementary Fig.S2a, S2b). Imaging features alone provide meaningful discriminatory power, but they remain limited by overlapping sonographic appearances among cholesterol polyps, adenomas, and early carcinomas, especially in small lesions or in the presence of acoustic noise, variable acquisition quality, and incomplete representation of contextual clinical risk factors (Supplementary Fig.S3 a-g).

The proposed GB-DualmodalNet attained the highest macro-AUC of 0.852 (95% CI: 0.801–0.902), with an accuracy of 70.1% and specificity of 79.2% in the test cohort (Fig.3a, 3b, Table 2). In contrast, the VGG16 plus MLP concatenation multimodal model yielded an AUC of 0.815, indicating that early fusion of visual and clinical features provides no additional benefit (Supplementary Fig.S4a).

Together, these findings demonstrate that integrating semantically enriched clinical prompts via cross-modal attention more effectively models the relationship between clinical context and imaging features than simple feature concatenation.

*3.1.3 Subtype-specific analysis*

GB-DualmodalNet demonstrated strong performance across all three lesion subtypes, with particularly effective discrimination of malignant and premalignant lesions of greatest clinical concern. The model achieved an AUC of 0.933 for carcinoma, 0.758 for adenoma, and 0.865 for cholesterol polyp detection (Supplementary Fig. S4b). This performance can largely be attributed to the cross-modal attention mechanism: when high-risk terms appear in the clinical prompt, the model more effectively focuses on imaging cues associated with early invasion. Importantly, the model maintained high specificity for cholesterol polyps, which is clinically valuable for avoiding unnecessary resections and aligns with the goals of precision treatment.

### *3.2 Subgroup analysis*

For clinically ambiguous 10–15 mm “gray-zone” polyps, a focused subgroup analysis was performed on 48 patients from the test cohort (Supplementary Table S4). Under conventional guidelines that recommend surgery strictly for polyps >10 mm, the false-positive rate reached 45.8%. In contrast, GB-DualmodalNet showed markedly better discrimination in this subgroup. Among the 22 benign large polyps, 19 were correctly classified as low risk. Furthermore, the model correctly identified all malignant and premalignant lesions, including 2 early gallbladder carcinomas and 24 adenomas. Statistically, while maintaining 100% sensitivity for neoplastic lesions, GB-DualmodalNet improved specificity from 0% by size criteria alone to 86.4%.

### *3.3 AI-assisted clinician performance*

To assess the influence of AI assistance on clinical decision-making, a two-reader study was performed using the independent test set. Two clinicians experienced in hepatobiliary ultrasonography reviewed all test cases under two conditions: first, an unaided assessment relying on standard ultrasonographic interpretation and established size-based guidelines; and second, an AI-assisted assessment in which they could access the subtype probabilities and risk suggestions provided by GB-DualmodalNet. In both settings, the clinicians made the final diagnostic decisions.

*3.3.1 Individual reader performance*

Table S5 summarizes the diagnostic performance of each reader with and without AI assistance. Without AI support, Reader 1 and Reader 2 achieved macro-AUCs of 0.789 (95% CI: 0.732–0.841) and 0.774 (95% CI: 0.715–0.830), respectively. Both readers demonstrated high sensitivity for neoplastic lesions, but relatively low specificity for cholesterol polyps, indicating a tendency toward false-positive classification of benign lesions. With AI assistance, both readers showed consistent improvements across multiple performance metrics. Macro-AUC increased to 0.872 (95% CI: 0.829–0.914) for Reader 1 and 0.858 (95% CI: 0.812–0.902) for Reader 2. These gains were primarily driven by increased specificity for cholesterol polyps, while sensitivity for neoplastic lesions remained high or improved further. The improvements were consistent for both readers, despite their minor differences in unaided performance.

*3.3.2 Pooled reader performance and confusion matrix analysis*

To provide an overall assessment of clinician performance, pooled metrics were calculated by averaging individual reader results. As presented in Table 3, pooled accuracy improved from 0.643 without AI assistance to 0.747 with AI assistance. The pooled macro-AUC rose from 0.782 (95% CI: 0.724–0.835) to 0.865 (95% CI: 0.821–0.910). Sensitivity for neoplastic lesions increased from 0.919 to 0.968, while specificity for cholesterol polyps improved from 0.500 to 0.685. Analysis of the confusion matrix based on pooled reader results (Supplementary Fig.S5a, S5b) indicated that AI assistance mainly reduced false-positive classifications of benign cholesterol polyps as adenomas, without compromising the accurate identification of gallbladder carcinoma or increasing false-negative cancer predictions. Misclassifications continued to occur most frequently between cholesterol polyps and adenomas.

***3.4 Visualization explainability***

By making decisive image regions and clinical features visually apparent and clinically coherent, the explainable AI modules improve the transparency of GB-DualmodalNet, ensuring that its reasoning aligns with established medical knowledge. SHAP analysis identified polyp long diameter, polyp short diameter, and patient age as the most influential features (Fig.3c). Grad-CAM++ visualization results indicated that in gallbladder carcinoma cases, the heatmaps precisely localize to the base of the polyp and the interface with the liver bed, which were critical areas for assessing invasive growth. In contrast, heatmaps generated by the baseline model occasionally diffuse toward adjacent gallbladder wall regions or imaging artifacts. For cholesterol polyps, the model focuses primarily on intralesional hyperechoic foci, consistent with cholesterol crystal deposition, and displays multi-point attention patterns in cases with multiple lesions (Fig.4).

## 4. Discussion

In this retrospective diagnostic accuracy study, a multimodal deep learning model that combined region-of-interest-guided ultrasound features with clinical information encoded as natural-language prompts achieved a macro-averaged AUC of 0.852 (95% CI 0.801–0.902) for the three-class preoperative classification of cholesterol polyps, adenomas, and carcinomas. Discrimination was highest for gallbladder carcinoma (AUC 0.933), followed by cholesterol polyp (0.865) and adenoma (0.758). In the clinically challenging 10–15 mm gray-zone subgroup (n = 48), the model preserved 100% sensitivity for malignancy while improving specificity from 0% under the size-based guideline to 86.4%. In an AI-assisted reader study, pooled clinician macro-AUC increased from 0.782 to 0.865 and pooled accuracy from 0.643 to 0.747, driven primarily by improved specificity for cholesterol polyps.

Previous artificial intelligence studies of gallbladder ultrasound have largely framed the problem as binary discrimination between neoplastic and non-neoplastic lesions, or as targeted detection of gallbladder carcinoma in isolation[10–13]. Although technically tractable, these formulations do not align with the actual preoperative decision faced by hepatobiliary surgeons, which is essentially three-way: surveillance for cholesterol polyps, simple cholecystectomy for adenomas, and consideration of radical or extended resection for suspected carcinoma. To our knowledge, the present model is the first internally validated multimodal framework that delivers preoperative probabilities for all three of these clinically actionable categories from a routine transabdominal ultrasound and a small set of routinely available clinical variables. Despite addressing a more difficult three-class problem, subtype-specific performance was comparable to or higher than prior binary models in the corresponding categories. For example, a carcinoma AUC of 0.933 in the present study versus 0.887 reported by Gupta et al. for binary cancer detection[13], supporting the feasibility of subtype-level classification in this setting.

A second contribution of this work concerns the way clinical context is incorporated. In prior multimodal models, structured clinical variables are typically one-hot encoded or numerically concatenated with image features, which preserves numerical magnitude but discards the semantic associations clinicians use to interpret findings[14]. By mapping each variable to a standardised textual descriptor and fusing it with image features through a cross-modal attention module that draws on medically pre-trained text embeddings, GB-DualmodalNet leverages domain-specific priors at the representation level rather than at the input level. The incremental gain in macro-AUC over the ROI-guided image-only baseline was modest (0.840 to 0.852, $P < 0.05$), but the practical benefit was concentrated where it matters most clinically in the gray-zone subgroup and in supporting clinician decision-making during the reader study.

The clinical relevance of these results lies primarily in the management of the 10–15 mm gray-zone polyp, which currently drives most of the overtreatment in this disease. Under the size-only guideline, every lesion in this size range is referred for cholecystectomy regardless of underlying pathology, generating a false-positive rate of 45.8% in our test cohort. If a tool with the operating characteristics of the present model were applied to

triage these patients, 19 of 22 patients (86%) with benign lesions in this subgroup would have been correctly classified as low risk while all 26 premalignant or malignant lesions would have been retained for surgical management. Although the absolute numbers are small and require prospective confirmation, the direction and magnitude of this effect are consistent with the broader literature on overtreatment in GBPLs[1, 8, 9] and suggest a plausible role for AI-based triage at the preoperative multidisciplinary assessment stage. The model is best positioned as a decision-support tool that complements, rather than replaces, surgical judgement and existing guideline-based criteria.

The reader study showed consistent improvements with AI assistance across both readers and across multiple performance metrics, with the largest gain in specificity for cholesterol polyps. The two participating clinicians had more than three years of experience in hepatobiliary ultrasonography, which approximates routine practice in many tertiary centres but does not fully capture the spectrum of clinical experience from trainees to subspecialist experts. The unaided pooled macro-AUC of 0.782 in our readers is broadly consistent with previously reported clinician performance for GBPL classification[12, 13] and supports the external validity of the baseline; nevertheless, two readers represent a small sample, and generalization of the AI-assistance effect to broader clinical settings will require larger, more heterogeneous reader studies.

Strengths of this study include a clinically relevant three-class outcome aligned with the surgical decision pathway, a transparent multimodal architecture with both spatial (Grad-CAM++) and feature-level (SHAP) explanations, a dedicated subgroup analysis of the 10–15 mm gray-zone lesions of greatest surgical interest, and an AI-assisted reader study that quantifies practical clinical benefit beyond raw model metrics. The work has several important limitations. First, the design is single-centre and retrospective, and the source cohort consists exclusively of patients who underwent cholecystectomy, which introduces spectrum bias because lesions managed conservatively are absent. Second, no external or temporal validation has been performed. Third, sample sizes for the carcinoma class and for the gray-zone subgroup remain modest, and the corresponding results should be interpreted with appropriate caution. Fourth, only static ultrasound frames were used, and dynamic information potentially relevant to lesion mobility and vascularity was not captured. Prospective multicentre validation, prospective embedding within a surgical decision pathway, and head-to-head comparison with contrast-enhanced and endoscopic ultrasound-based approaches will be needed before clinical deployment can be considered.

In summary, a clinically prompted multimodal deep learning model can perform preoperative three-class classification of gallbladder polypoid lesions on routine transabdominal ultrasound, with subtype-specific discrimination comparable to or higher than prior binary models, and can improve clinician performance when used as a decision-support tool. The most surgically relevant finding is the substantial improvement in specificity for benign lesions in the 10–15 mm gray zone without loss of malignancy detection, suggesting that AI-based triage may help reduce unnecessary cholecystectomy in this subgroup. Prospective external and multicenter validation is required before these findings can be translated into routine surgical practice.

## Declarations

**Ethics approval and consent to participate**

The study protocol strictly adhered to the ethical principles of the Declaration of Helsinki and was approved by the Ethics Committee of the First Affiliated Hospital of Xi'an Jiaotong University (No. 2019G-40). Given the retrospective nature of the study and the exclusive use of de-identified patient data, the requirement for informed consent was waived.

**Consent for publication**

Informed consent was obtained from the individual, who had signed the power of attorney document.

**Availability of data and materials**

De-identified data are not publicly available because of institutional and patient privacy restrictions but may be made available from the corresponding author upon reasonable request and with appropriate institutional approval. The code used for model training and evaluation is available at https://github.com/qingyut/GB-DualmodalNet.

**Competing interests**

The authors declare that they have no competing interests.

**Funding**

This work was supported by the National Natural Science Foundation of China (No. 62576271), Key Research and Development Program of Shaanxi Province (No. 2025SF-YBXM-386 and 2022-SF-606), and Institutional Foundation of The First Affiliated Hospital of Xi'an Jiaotong University (No.2024-QN-15). The funding body had no role in the design of the study, the design of the study and collection, analysis, and interpretation of data, and in writing the manuscript.

**Authors' contributions**

QT and ZW contributed equally to this work as co-first authors. Conceptualization and methodology: QL, ZG, QT. Data collection: QT, JS, YM, HW, HL, KL. Data annotation: MD, QL. Reader study: DZ, ZG. Software and implementation: QT, ZW, ZP. Formal analysis: QT, ZW. Writing: QT. Review and editing: QL, QT, ZT, ZW. Supervision: ZG, QL. Funding acquisition: QL, DZ. All authors read and approved the final manuscript.

**Acknowledgements**

We thank all authors and study participants for their contributions.

## References

1. Szpakowski J-L, Tucker L-Y. Outcomes of Gallbladder Polyps and Their Association With Gallbladder Cancer in a 20-Year Cohort. JAMA Network Open 2020;3:e205143-e205143. doi: 10.1001/jamanetworkopen.2020.5143.
2. Wang K, Xu Q, Xia L, Sun J, Shen K, Liu H*, et al.* Gallbladder polypoid lesions: Current practices and future prospects. Chinese Medical Journal 2024;137:1674-1683. doi: 10.1097/cm9.0000000000003019.
3. Wennmacker SZ, Lamberts MP, Di Martino M, Drenth JPH, Gurusamy KS, van Laarhoven C. Transabdominal ultrasound and endoscopic ultrasound for diagnosis of gallbladder polyps. Cochrane Database of Systematic Reviews 2018. doi: 10.1002/14651858.CD012233.pub2.
4. Foley KG, Lahaye MJ, Thoeni RF, Soltes M, Dewhurst C, Barbu ST*, et al.* Management and follow-up of gallbladder polyps: updated joint guidelines between the ESGAR, EAES, EFISDS and ESGE. European Radiology 2022;32:3358-3368. doi: 10.1007/s00330-021-08384-w.
5. Kamaya A, Fung C, Szpakowski J-L, Fetzer DT, Walsh AJ, Alimi Y*, et al.* Management of Incidentally Detected Gallbladder Polyps: Society of Radiologists in Ultrasound Consensus Conference Recommendations. Radiology 2022;305:277-289. doi: 10.1148/radiol.213079.
6. Fung CI, Kamaya A, Brahm GL, Bird JR, Kirkpatrick IDC. Recommendations for the Management of Incidentally Detected Gallbladder Polyps: Update of the 2020 Canadian Association of Radiologists Recommendations for the Management of Incidental Hepatobiliary Findings in Adults. Can Assoc Radiol J 20258465371251346728. doi: 10.1177/08465371251346728.
7. Wennmacker SZ, van Dijk AH, Raessens JHJ, van Laarhoven CJHM, Drenth JPH, de Reuver PR*, et al.* Polyp size of 1 cm is insufficient to discriminate neoplastic and non-neoplastic gallbladder polyps. Surgical Endoscopy 2019;33:1564-1571. doi: 10.1007/s00464-018-6444-1.
8. Choi TW, Kim JH, Park SJ, Ahn SJ, Joo I, Han JK. Risk stratification of gallbladder polyps larger than 10 mm using high-resolution ultrasonography and texture analysis. European Radiology 2018;28:196-205. doi: 10.1007/s00330-017-4954-1.
9. Wang L-F, Wang Q, Mao F, Xu S-H, Sun L-P, Wu T-F*, et al.* Risk stratification of gallbladder masses by machine learning-based ultrasound radiomics models: a prospective and multi-institutional study. European Radiology 2023;33:8899-8911. doi: 10.1007/s00330-023-09891-8.
10. Basu S, Gupta M, Rana P, Gupta P, Arora C. RadFormer: Transformers with global–local attention for interpretable and accurate Gallbladder Cancer detection. Medical Image Analysis 2023;83:102676. doi: https://doi.org/10.1016/j.media.2022.102676.
11. Hasan MZ, Rony MAH, Chowa SS, Bhuiyan MRI, Moustafa AA. GBCHV an advanced deep learning anatomy aware model for accurate classification of gallbladder cancer utilizing ultrasound images. Sci Rep 2025;15:7120. doi: 10.1038/s41598-025-89232-5.
12. Kim T, Choi YH, Choi JH, Lee SH, Lee S, Lee IS. Gallbladder Polyp Classification in Ultrasound Images Using an Ensemble Convolutional Neural Network Model. J Clin Med 2021;10. doi: 10.3390/jcm10163585.
13. Gupta P, Basu S, Rana P, Dutta U, Soundararajan R, Kalage D*, et al.* Deep-learning enabled ultrasound based detection of gallbladder cancer in northern India: a prospective diagnostic study. The Lancet Regional Health - Southeast Asia 2024;24. doi: 10.1016/j.lansea.2023.100279.
14. Yuan K, Zhang X, Yang Q, Deng X, Deng Z, Liao X*, et al.* Risk prediction and analysis of gallbladder polyps with deep neural network. Computer Assisted Surgery 2024;29:2331774. doi: 10.1080/24699322.2024.2331774.
15. Cho JH, Park JY, Kim YJ, Kim HM, Kim HJ, Hong SP*, et al.* Hypoechoic foci on EUS are simple and strong predictive factors for neoplastic gallbladder polyps. Gastrointestinal Endoscopy 2009;69:1244-1250. doi: 10.1016/j.gie.2008.10.017.
16. Sylvester S, Sagehorn M, Gruber T, Atzmueller M, Schöne B. SHAP value-based ERP analysis (SHERPA): Increasing the sensitivity of EEG signals with explainable AI methods. Behavior Research Methods 2024;56:6067-6081. doi: 10.3758/s13428-023-02335-7.
17. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV); 2018. pp. 839-847.

## Tables

**Table 1.** Baseline patient characteristics.

| Characteristics | Training set | Test set | P-value |
|---|---|---|---|
| | (N=362) | (N=154) | |
| Sex | | | 0.930 |
| Male | 140 (38.7) | 59 (38.3) | |
| Female | 222 (61.3) | 95 (61.7) | |
| Age (year) median (IQR) | 48 (35-58) | 46 (35-57) | 0.528 |
| Polyp detection time (months) | 12 (0.5-48) | 12 (2-48) | 0.564 |
| Number of polyps | | | 0.572 |
| Single | 170 (47.6) | 64 (42.7) | |
| Double | 19 (5.3) | 8 (5.3) | |
| Multiple (≥3) | 168 (47.1) | 78 (52.0) | |
| Polyp size | | | |
| Long diameter, mm, mean (SD) | 12.86 ± 10.24 | 12.13 ± 6.64 | 0.415 |
| Short diameter, mm, mean (SD) | 7.90 ± 4.64 | 7.56 ± 4.15 | 0.428 |
| Polyp location | | | 0.195 |
| Neck | 40 (11.0) | 10 (6.5) | |
| Body | 309 (85.4) | 141 (91.6) | |
| Bottom | 11 (3.0) | 3 (1.9) | |
| Wall thickening | | | 0.227 |
| <4 mm | 302 (85.6) | 135 (90) | |
| ≥4 mm | 51 (14.4) | 15 (10) | |
| Base | | | 0.084 |
| Narrow base | 191 (52.8) | 66 (42.9) | |
| Broad base | 110 (30.4) | 61 (39.6) | |
| Pedicle | 34 (9.4) | 15 (9.7) | |
| Echogenicity | | | 0.356 |
| Low | 38 (10.5) | 17 (11.0) | |
| Medium | 24 (6.6) | 13 (8.4) | |
| Strong | 297 (82.0) | 120 (77.9) | |
| Laboratory Examination | | | |
| WBC ($10^9$/L) | 6.16 ± 2.36 | 5.93 ± 2.14 | 0.299 |
| RBC ($10^{12}$/L) | 4.57 ± 0.62 | 4.58 ± 0.59 | 0.768 |
| PLT ($10^9$/L) | 221.23 ± 64.56 | 208.68 ± 62.97 | 0.038 |
| Hgb (g/L) | 138.59 ± 20.89 | 137.92 ± 20.69 | 0.508 |
| ALT (U/L) | 30.13 ± 56.96 | 24.21 ± 16.49 | 0.213 |
| AST (U/L) | 28.04 ± 47.22 | 22.87 ± 14.77 | 0.193 |
| ALP (U/L) | 78.52 ± 52.20 | 77.29 ± 24.70 | 0.721 |
| GGT (U/L) | 30.86 ± 47.36 | 23.64 ± 25.16 | 0.080 |
| TB (umol/L) | 16.03 ± 20.628 | 14.44 ± 5.873 | 0.187 |
| DB (umol/L) | 5.03 ± 16.63 | 3.81 ± 2.39 | 0.377 |

| | | | |
|---|---|---|---|
| ALB (g/l) | 44.08 ± 21.81 | 42.99 ± 5.66 | 0.548 |
| CEA (u/l) | 2.30 ± 1.98 | 1.79 ± 1.47 | 0.066 |
| Pathology | | | 0.829 |
| Cholesterol polyp | 226 (62.4) | 92 (59.7) | |
| Adenoma | 99 (27.3) | 46 (29.8) | |
| Carcinoma | 37 (10.2) | 16 (10.4) | |

IQR, interquartile range; SD, standard deviation; WBC, white blood cell count; RBC, red blood cell count; PLT, platelet count; Hgb, hemoglobin; ALT, alanine aminotransferase; AST, aspartate aminotransferase; ALP, alkaline phosphatase; GGT, gamma-glutamyl transferase; TB, total bilirubin; DB, direct bilirubin; ALB, albumin; CEA, carcinoembryonic antigen.

**Table 2.** Performance of the GB-DualmodalNet with different multimodal fusion strategies compared to ROI-guided single modal or whole-image baseline.

| Model | AUC (95% CI) | Accuracy | Sensitivity | Specificity | PPV | NPV | F1 score |
|---|---|---|---|---|---|---|---|
| Whole-image baseline | 0.608 (0.524–0.678) | 0.487 | 36.4% (28.9%–44.8%) | 68.2% (63.7%–72.9%) | 0.389 | 0.682 | 0.369 |
| ROI-guided image | 0.840 (0.787–0.890) | 0.662 | 65.7% (55.8%–75.0%) | 81.5% (77.0%–85.8%) | 0.625 | 0.789 | 0.635 |
| Traditional fusion | 0.815 (0.757–0.868) | 0.688 | 65.4% (55.4%–75.1%) | 82.4% (78.1%–86.7%) | 0.640 | 0.814 | 0.645 |
| GB-DualmodalNet (proposed) | 0.852 (0.801–0.902) | 0.701 | 58.0% (48.3%–67.7%) | 79.2% (75.1%–83.4%) | 0.649 | 0.845 | 0.596 |

*AUC, receiver operating characteristic curve; CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value; ROI, region of interest.*

**Table 3.** Pooled diagnostic performance of two readers**.**

| Condition | Accuracy | AUC (95% CI) | Sensitivity | Specificity | F1-score |
|---|---|---|---|---|---|
| Clinicians without AI | 0.643 | 0.781 (0.723–0.834) | 91.9% (85.7%–96.1%) | 50.0% (42.6%–57.4%) | 0.595 |
| Clinicians assisted by AI | 0.747 | 0.866 (0.822–0.911) | 96.8% (91.9%–99.1%) | 68.5% (61.2%–75.1%) | 0.665 |

*AI, artificial intelligence; AUC, receiver operating characteristic curve; CI, confidence interval.*

## Figures

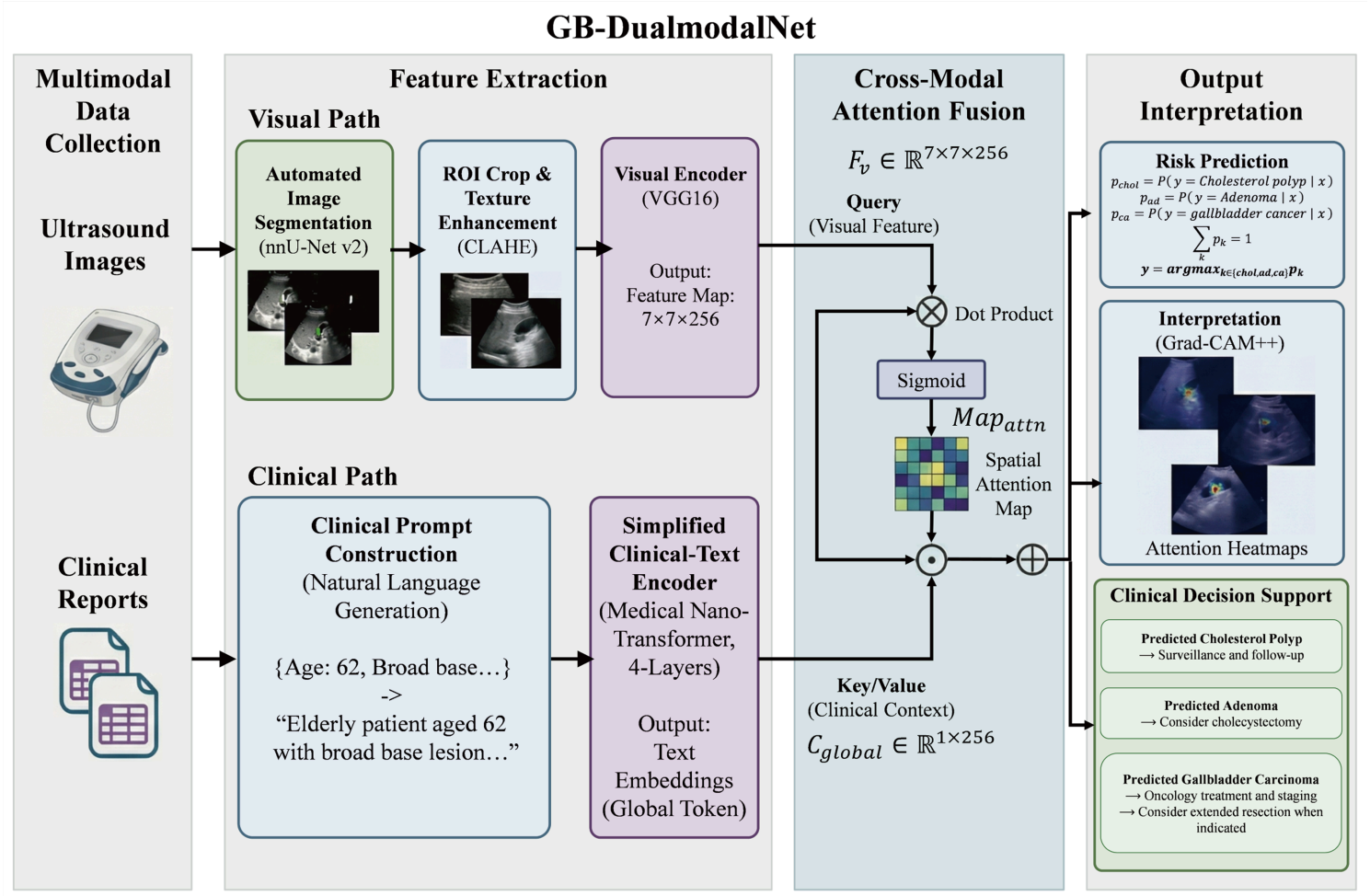



**Figure 1**

The flowchart of patient selection and model evaluation.

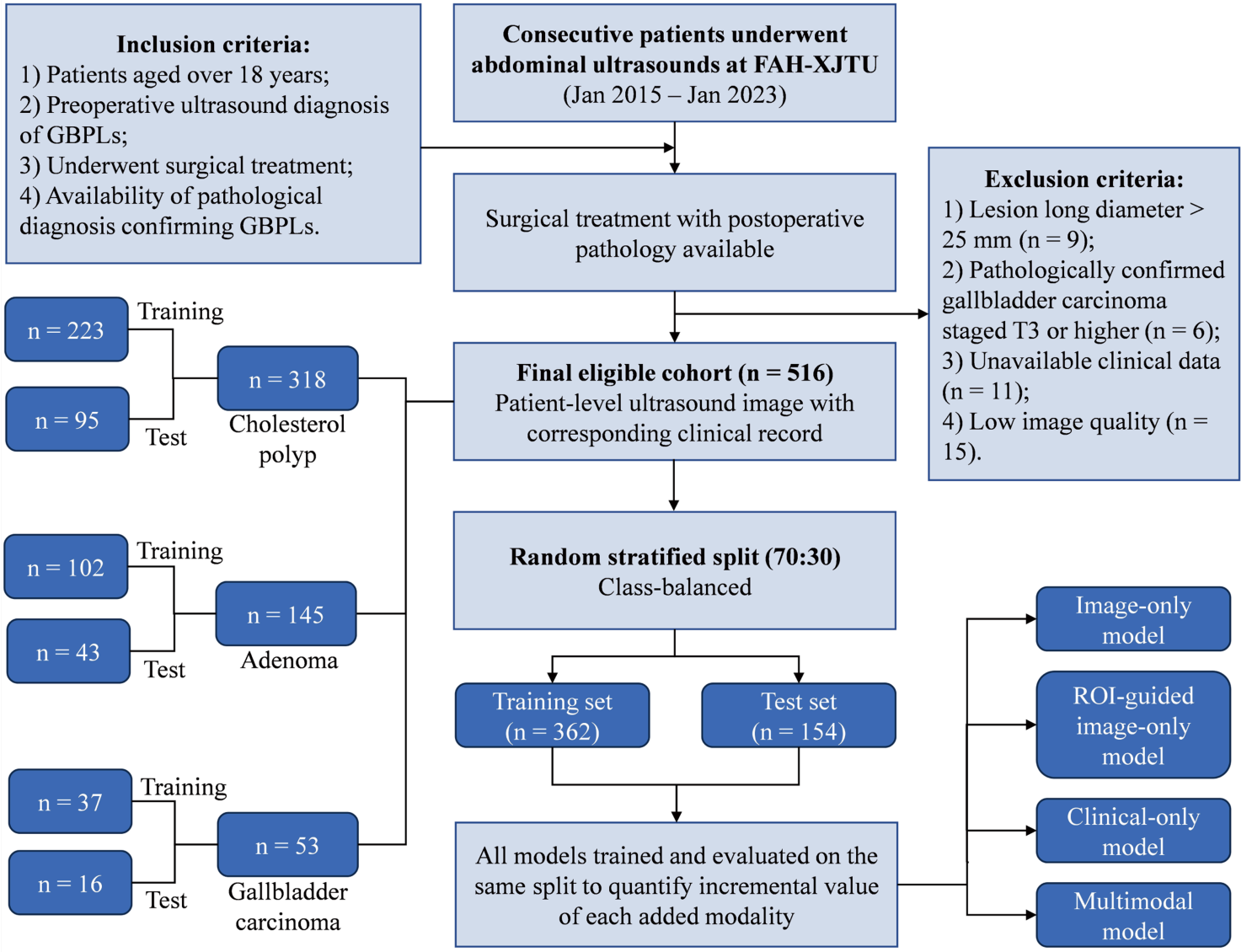



**Figure 2**

Schematic workflow of GB-DualmodalNet and cross-modal fusion strategy.

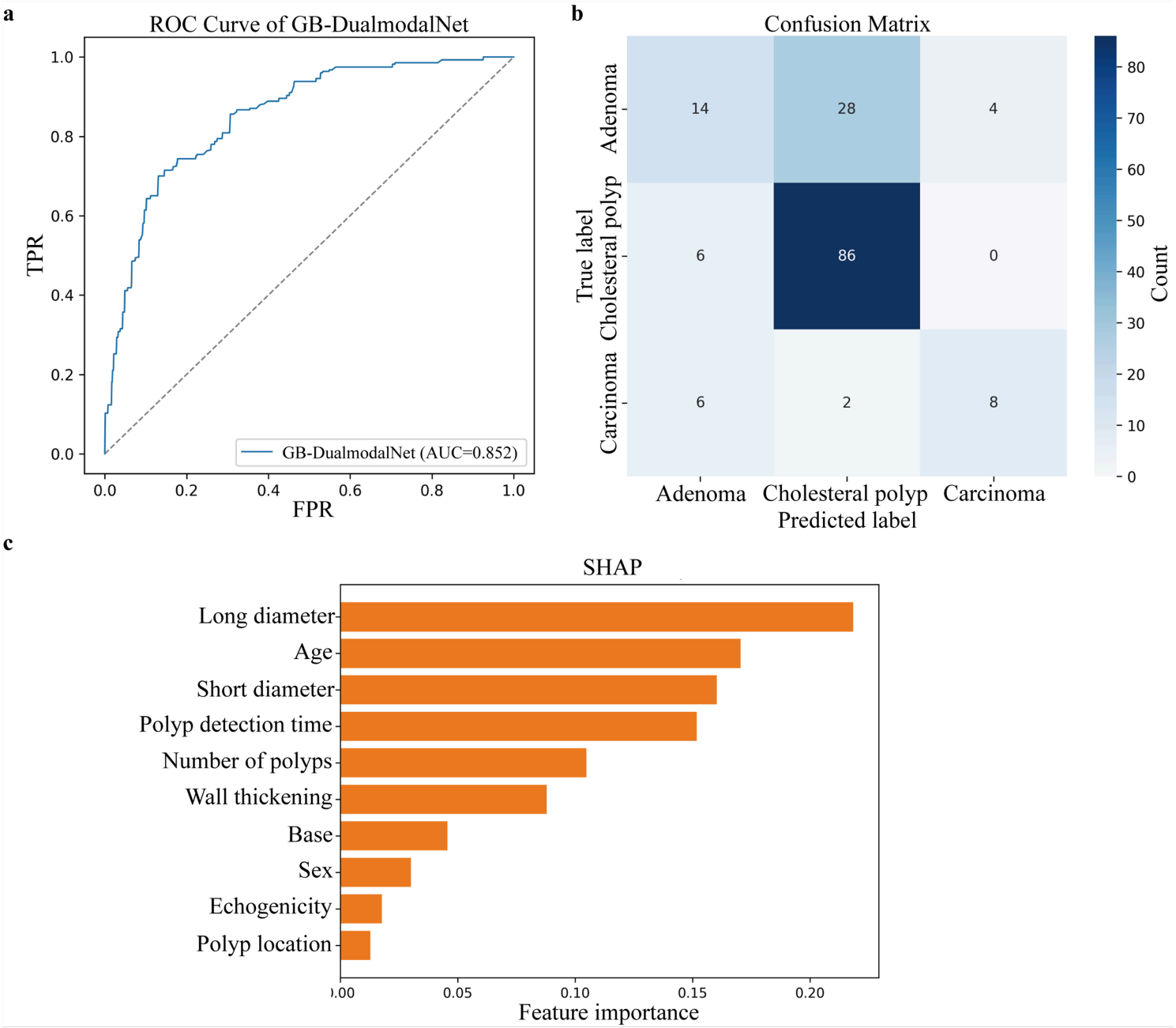



**Figure 3**

Diagnostic performance of GB-DualmodalNet model in the independent test cohort. (a) ROC curve for test-set subtype predictions. (b) Confusion matrix for test-set predictions. (c) SHAP-based feature importance ranking for clinical variables for test-set predictions.

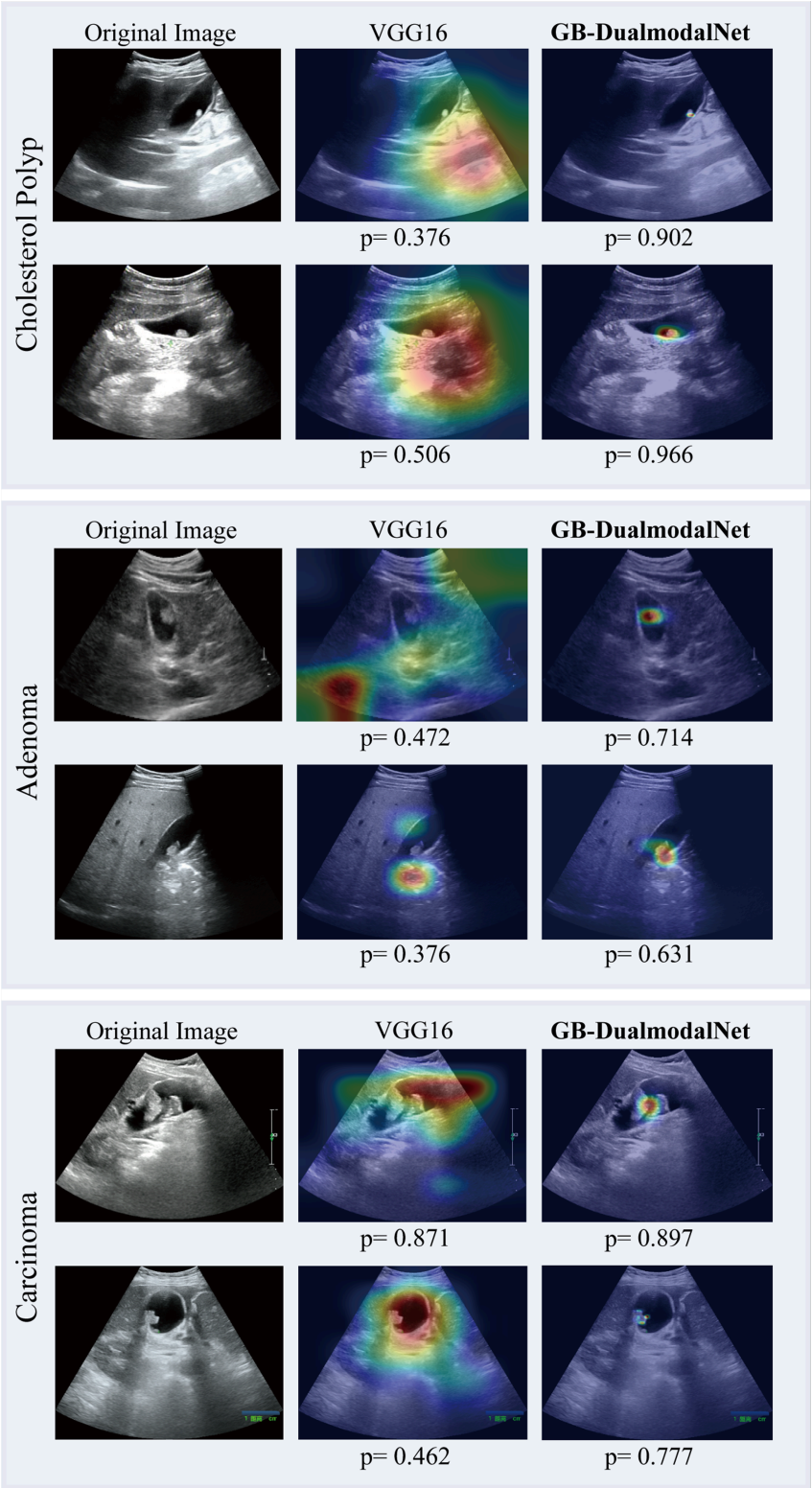



**Figure 4**

Representative Grad-CAM++ visualizations comparing the image-only baseline and GB-DualmodalNet.

## Supplementary Files

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

- SupplementaryMaterial.docx