European Radiology
EUROPEAN SOCIETY OF RADIOLOGY

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Risk stratification of gallbladder masses by machine learning-based ultrasound radiomics models: a prospective and multi-institutional study

Li-Fan Wang[1] · Qiao Wang[2,3] · Feng Mao[4] · Shi-Hao Xu[5] · Li-Ping Sun[2,3] · Ting-Fan Wu[6] · Bo-Yang Zhou[1] · Hao-Hao Yin[1] · Hui Shi[2,3] · Ya-Qin Zhang[2,3] · Xiao-Long Li[1] · Yi-Kang Sun[1] · Dan Lu[1] · Cong-Yu Tang[1] · Hai-Xia Yuan[1,7] · Chong-Ke Zhao[1] · Hui-Xiong Xu[1]

Abstract

Objective This study aimed to evaluate the diagnostic performance of machine learning (ML)–based ultrasound (US) radiomics models for risk stratification of gallbladder (GB) masses.

Methods We prospectively examined 640 pathologically confirmed GB masses obtained from 640 patients between August 2019 and October 2022 at four institutions. Radiomics features were extracted from grayscale US images and germane features were selected. Subsequently, 11 ML algorithms were separately used with the selected features to construct optimum US radiomics models for risk stratification of the GB masses. Furthermore, we compared the diagnostic performance of these models with the conventional US and contrast-enhanced US (CEUS) models.

Results The optimal XGBoost-based US radiomics model for discriminating neoplastic from non-neoplastic GB lesions showed higher diagnostic performance in terms of areas under the curves (AUCs) than the conventional US model (0.822–0.853 vs. 0.642–0.706, $p < 0.05$) and potentially decreased unnecessary cholecystectomy rate in a speculative comparison with performing cholecystectomy for lesions sized over 10 mm (2.7–13.8% vs. 53.6–64.9%, $p < 0.05$) in the validation and test sets. The AUCs of the XGBoost-based US radiomics model for discriminating carcinomas from benign GB lesions were higher than the conventional US model (0.904–0.979 vs. 0.706–0.766, $p < 0.05$). The XGBoost-US radiomics model performed better than the CEUS model in discriminating GB carcinomas (AUC: 0.995 vs. 0.902, $p = 0.011$).

Conclusions The proposed ML-based US radiomics models possess the potential capacity for risk stratification of GB masses and may reduce the unnecessary cholecystectomy rate and use of CEUS.

Clinical relevance statement The machine learning-based ultrasound radiomics models have potential for risk stratification of gallbladder masses and may potentially reduce unnecessary cholecystectomies.

Key Points

● *The XGBoost-based US radiomics models are useful for the risk stratification of GB masses.*

● *The XGBoost-based US radiomics model is superior to the conventional US model for discriminating neoplastic from non-neoplastic GB lesions and may potentially decrease unnecessary cholecystectomy rate for lesions sized over 10 mm in comparison with the current consensus guideline.*

Li-Fan Wang and Qiao Wang contributed equally to this work.

✉ Hai-Xia Yuan
    haixiayuan75@163.com

✉ Chong-Ke Zhao
    zhaochongke123@163.com

✉ Hui-Xiong Xu
    xu.huixiong@zs-hospital.sh.cn

1   Department of Ultrasound, Institute of Ultrasound in Medicine and Engineering, Zhongshan Hospital, Fudan University, Shanghai, China

2   Department of Medical Ultrasound, Center of Minimally Invasive Treatment for Tumor, Shanghai Tenth People's Hospital, Ultrasound Education and Research Institute, School of Medicine, Tongji University, Shanghai, China

3   Shanghai Engineering Research Center of Ultrasound Diagnosis and Treatment, Shanghai, China

4   Department of Medical Ultrasound, First Hospital of Ningbo University, Ningbo, Zhejiang, China

5   Department of Ultrasonography, the First Affiliated Hospital of Wenzhou Medical University, Wenzhou, Zhejiang, China

6   Bayer Healthcare, Radiology, Shanghai, China

7   Department of Ultrasound, Zhongshan Hospital of Fudan University (Qingpu Branch), Shanghai, China

- *The XGBoost-based US radiomics model could overmatch CEUS model in discriminating GB carcinomas from benign GB lesions.*

Keywords Ultrasound · Gallbladder masses · Machine learning · Radiomics · Risk stratification

Abbreviations

AUC	Area under the curve
CEA	Carcinoembryonic antigen
CEUS	Contrast-enhanced ultrasound
GB	Gallbladder
ICC	Intraclass correlation coefficient
ML	Machine learning
ROI	Region of interest
US	Ultrasound
XGBoost	Extreme gradient boosting

Introduction

With the improvement and wide use of imaging examinations, gallbladder (GB) masses have been detected in approximately 5% of the global population, presenting as polypoidal intraluminal masses or space-occupying masses in GB imaging [1, 2]. Non-neoplastic and neoplastic lesions (GB carcinomas and adenomas) constitute the disease spectrum and differ in risk and prognosis [1, 3]. GB carcinomas are associated with high mortality rates with high aggression, early metastasis, and poor prognosis without specific symptoms at an early stage [4, 5]. Early identification and adequate surgical intervention for neoplastic lesions are vital for preventing the occurrence or progression of GB carcinomas [6, 7].

Currently, common guidelines recommend cholecystectomy for GB lesions sized over 10 mm [8, 9]. However, numerous non-neoplastic lesions are sized over 10 mm, and follow-up may be an alternative to unnecessary surgical intervention [10]. The size of some GB neoplastic lesions is also below this threshold [9]. Moreover, in terms of management, laparoscopic or simple cholecystectomy is appropriate for benign GB lesions. For GB carcinomas, comprehensive assessment is required to determine a specific treatment strategy, such as radical cholecystectomy or extensive resection [2, 11]. Therefore, accurate preoperative risk discrimination of GB neoplastic from non-neoplastic lesions or GB malignant from benign lesions is a prerequisite for decision-making.

Conventional ultrasound (US) is the first-line imaging modality used for the diagnosis of GB diseases [12, 13]. Several previous studies have revealed some US features were associated with GB neoplastic or malignant lesions; however, the universally accepted risk factor was the size of the lesion [14–16]. Nevertheless, the risk stratification of

GB lesions using only the visual interpretation of US images is challenging [17]. Contrast-enhanced ultrasound (CEUS) is an additional imaging technology for differentiating GB diseases by providing vascularity and perfusion information [18–20]. Some studies have reported that perfusion features and GB wall integrity in CEUS are useful in discriminating malignant from benign GB lesions [21–23]. However, high empirical dependence, increased cost, low availability, and potential contraindications to US contrast agents limit the clinical application of CEUS [24].

Recently, radiomics has drawn widespread attention for medical image recognition tasks [25]. Radiomics provides quantitative and comprehensible features extracted through computerised algorithms, which can serve as imaging biomarkers beyond the visual interpretation of human beings [26–28]. A few studies have applied radiomics approaches to identify true and pseudo-GB polyps or neoplastic GB polyps [29, 30]. By taking advantage of its ability to process quantitative image information, machine learning (ML)–based computational methods have been introduced to improve the diagnostic accuracy for disease classification [31].

Combining the extracted quantitative features and ML-based computational methods may yield a promising effect in the risk stratification of GB masses. Thus, the aim of our study was to investigate the diagnostic performance of ML-based US radiomics models in the risk discrimination of GB masses from two clinical perspectives: (1) discrimination of neoplastic from non-neoplastic lesions, and (2) discrimination of malignant from benign lesions (both neoplastic and non-neoplastic lesions). Furthermore, we investigated whether ML-based US radiomics models could overmatch CEUS in diagnostic performance.

Materials and methods

This multi-institutional study was approved by the ethics committee of the institution (No: SHSYIEC–4.1/21–263/01; 2022-187R), and informed consent was obtained. The prospective protocol is registered at www.chictr.org.cn (ChiCTR2200056165).

Patients

Between August 2019 and October 2022, 609 consecutive patients were enrolled from Institution 1 (Shanghai Tenth People's Hospital) and Institution 2 (Zhongshan Hospital,

Fudan University) as a model development dataset, following the inclusion criteria: (a) patients with non-mobile GB masses as observed during conventional US examination within 1 month before surgery and (b) those who underwent surgical resection and had pathologically confirmed GB masses. The exclusion criteria for patients were as follows: (a) surgically diagnosed with GB stones ($n = 11$); (b) incomplete clinical data ($n = 25$); (c) poor quality of US images ($n = 30$); (d) complete loss of GB structure and inability to distinguish the lesion from adjacent liver tissue, which could not delineate the lesion outline manually as a region of interest (ROI) ($n = 4$); and (e) other carcinomas of the digestive system simultaneously, to avoid the influence on carbohydrate antigen 19-9 (CA19-9) and carcinoembryonic antigen (CEA) levels ($n = 3$). The largest lesion was selected as the target in cases with multiple lesions. Eventually, 536 patients with 536 GB masses were prospectively enrolled and randomly divided into training and validation sets in a 7:3 ratio.

Following the same criteria used for the model development dataset, 56 consecutive patients with 56 GB masses from Institution 3 (First Affiliated Hospital of Wenzhou Medical University) and 48 consecutive patients with 48 GB masses from Institution 4 (First Hospital of Ningbo University) were prospectively enrolled as two independent external test sets (A and B, respectively) between June 2020 and October 2022.

Among the enrolled 640 patients, 95 underwent additional CEUS examination. Excluding five patients with poor image quality of CEUS, 90 patients with 90 GB masses were enrolled as a "comparison" set to compare the diagnostic performance between the ML-based US radiomics model and the CEUS model. A detailed flowchart of the patient selection process is presented in Fig. 1. The basic demographic and clinicopathological characteristics of the patients were collected from the medical record systems (Table 1).

Conventional US and CEUS data acquisition

The details of conventional US and CEUS examinations are provided in Supplementary Material S1. All US and CEUS data in the Digital Imaging and Communications in Medicine format were stored in the respective imaging system platform.

US imaging segmentation and radiomics features extraction

The greyscale US imaging segmentation was processed using the ITK-SNAP 3.6.0 programme (http://www.itksnap.org) [32]. The ROI was manually segmented by two
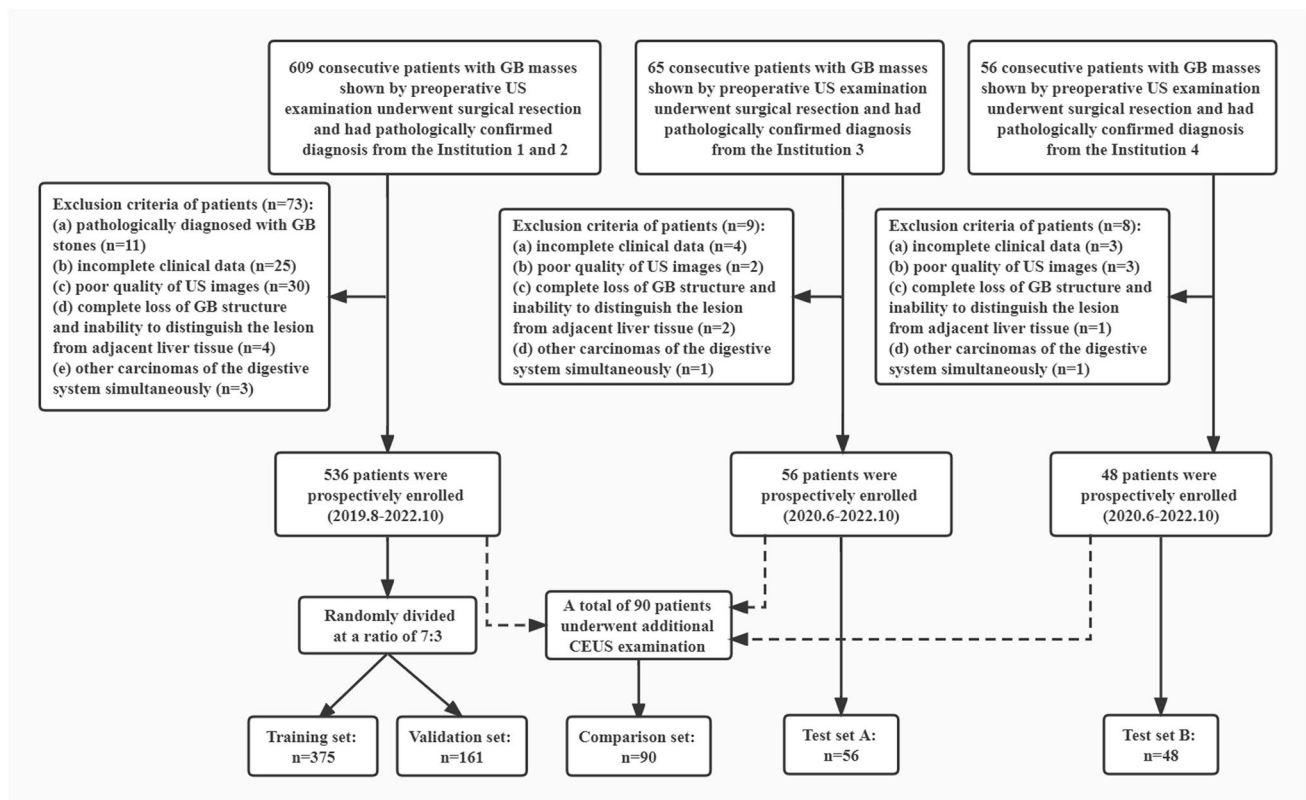

Fig. 1 Flowchart of patient selection in this study. US, ultrasound; CEUS, contrast-enhanced ultrasound

Table 1 The demographic data of 640 patients from four institutions

Characteristics	Training set	Validation set	Test set A	Test set B	Comparison set
No. of patients	375	161	56	48	90
Sex					
Female	219 (58.4)	95 (59.0)	35 (62.5)	23 (47.9)	47 (52.2)
Male	156 (41.6)	66 (41.0)	21 (37.5)	25 (52.1)	43 (47.8)
Mean age, years (range)*	54.9 ± 13.4 (22–84)	53.4 ± 13.5 (18–77)	51.4 ± 16.5 (23–81)	57.2 ± 13.0 (28–78)	57.1 ± 13.2 (27–84)
Presence of gallstone	71 (18.9)	33 (20.5)	6 (10.7)	2 (4.2)	15 (16.7)
CA199					
< 27 µ/mL	336 (89.6)	139 (86.3)	51 (91.1)	45 (93.8)	76 (84.4)
≥ 27 µ/mL	39 (10.4)	22 (13.7)	5 (8.9)	3 (6.2)	14 (15.6)
CEA					
< 5 ng/mL	355 (94.7)	154 (95.7)	55 (98.2)	46 (95.8)	86 (95.6)
≥ 5 ng/mL	20 (5.3)	7 (4.3)	1 (1.8)	2 (4.2)	4 (4.4)
No. of tumors	375	161	56	48	90
Mean tumor size, mm (range)*	18.9 ± 12.2 (6.0–88.0)	14.7 ± 7.9 (6.0–50.0)	16.2 ± 8.6 (9.4–44.7)	18.2 ± 11.7 (8.0–57.0)	21.8 ± 15.0 (7.0–83.9)
GB carcinomas	56 (14.9)	27 (16.8)	11 (19.6)	11 (22.9)	17 (18.9)
Tis	4 (7.1)	1 (3.7)	0 (0.0)	1 (9.1)	1 (5.9)
T1a	3 (5.4)	2 (7.4)	1 (9.1)	1 (9.1)	2 (11.8)
T1b	9 (16.1)	2 (7.4)	1 (9.1)	2 (18.2)	4 (23.5)
T2	32 (57.1)	18 (66.7)	8 (72.7)	2 (18.2)	8 (47.1)
T3	8 (14.3)	4 (14.8)	1 (9.1)	5 (45.5)	2 (11.8)
Benign GB lesions	319 (85.1)	134 (83.2)	45 (80.4)	37 (77.1)	73 (81.1)
Cholesterol polyp	144 (38.4)	88 (54.7)	26 (46.4)	20 (41.7)	28 (38.4)
Adenoma	78 (20.8)	37 (23.0)	8 (14.3)	8 (16.7)	14 (19.2)
Adenomyomatosis	71 (18.9)	7 (4.3)	6 (10.7)	6 (12.5)	26 (35.6)
Inflammatory polyp	26 (6.9)	2 (1.2)	5 (8.9)	3 (6.3)	5 (6.8)
Neoplastic GB lesions	134 (35.7)	64 (39.8)	19 (33.9)	19 (39.6)	31 (34.4)
Non-neoplastic GB lesions	241 (64.3)	97 (60.2)	37 (66.1)	29 (60.4)	59 (65.6)

Data in parentheses are percentages except for special indications

GB gallbladder, *CEA* carcinoembryonic antigen

*Data are presented as the mean ± standard deviation, with ranges in parentheses

radiologists (with 7 and 8 years of experience in abdominal US, respectively), detailed in Supplementary Material S2.

For each ROI, 1070 radiomics features (Supplementary Material S3) were extracted using the IFoundry software (Intelligence Foundry 1.2, GE Healthcare).

Key US radiomics features selection and construction of the ML-based US radiomics models

The intraclass correlation coefficient (ICC > 0.8), t-test, Spearman's correlation ($r < 0.8$), and least absolute shrinkage and selection operator were used to select the key US radiomics features. Eleven ML classifiers (Supplementary Material S4) were used with the selected radiomics features to construct 11 ML-based US radiomics models. The fivefold

cross-validation was used to adjust the structural parameters [33]. The best-performing ML classifier with the highest area under the curve (AUC) in the validation set was regarded as the optimal ML-based US radiomics model (Fig. 2).

Construction of the ML-based US radiomics incorporated clinical characteristics models

Logistic regression analysis was used to identify independent predictors of clinical characteristics, including sex and age of the patients, presence of gallstone, and levels of CA19-9 and CEA. The independent clinical predictors were then combined with the optimal US radiomics models to construct the ML-based US radiomics-incorporated clinical characteristics models (Fig. 2).
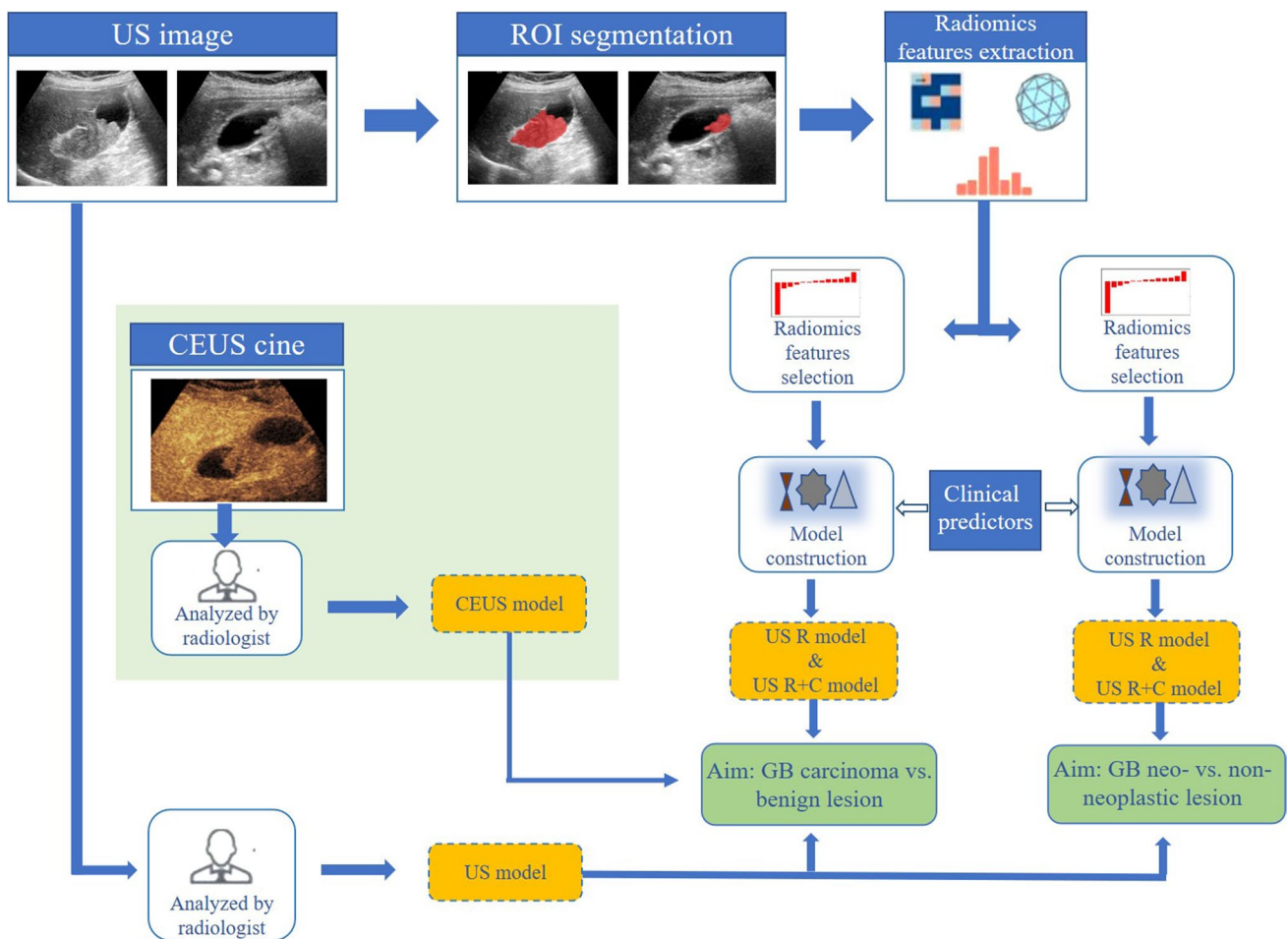
Fig. 2 Study workflow of the US, CEUS, US R, and US R + C models for risk stratification of gallbladder masses. US, ultrasound; CEUS, contrast-enhanced ultrasound; US R, ultrasound radiomics model; US R + C, ultrasound radiomics incorporated clinical characteristics model; GB, gallbladder; ROI, region of interest

Conventional US and CEUS data analysis and construction of the conventional US and CEUS models

A radiologist with more than 10 years of experience in abdominal US and CEUS analysed the conventional US images of 640 patients. After a buffer period of 1 week, the CEUS cines of 90 patients were analysed by the same radiologist. The details of data analyses are provided in Supplementary Material S5.

Conventional US and CEUS models for GB mass prediction were constructed using logistic regression analysis with a fivefold cross-validation.

Statistical analysis

Data analyses were performed using Python (version 3.8.8) and SPSS software (version 22.0; IBM Corporation). Normality was tested using the Kolmogorov–Smirnov test. Continuous variables were compared using the t-test and

described as mean and standard deviation. Categorical variables were compared using the chi-square or McNemar's test. The diagnostic performance was evaluated using AUC. DeLong's test was used to assess differences between AUCs. Statistical significance was set at a p value of less than 0.05.

Results

Demographic and clinicopathological characteristics of patients

Eventually, there were 375, 161, 56, and 48 patients in the training set, validation set, and test sets A and B, respectively (Table 1). In the training set, elevated CA19-9 levels, elevated CEA levels, mean age, sex, and presence of gallstone were significantly different between patients with GB carcinomas and those with benign lesions. Elevated CA19-9 levels, elevated CEA levels, mean age, and presence of gallstone were statistically different between patients with

Table 2 The diagnostic performance of models in discrimination of the GB neoplastic lesions from non-neoplastic ones

Diagnostic models	AUC	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)	p value
Validation set							
Conventional US model	0.706 (0.620–0.793)	51.6 [33/64] (38.8–64.1)	89.7 [87/97] (81.4–94.7)	76.7 [33/43] (61.0–87.7)	73.7 [87/118] (64.7–81.2)	74.5 [120/161] (67.1–81.1)	
US radiomics model	0.837 (0.775–0.900)	62.5 [40/64] (49.5–74.0)	96.9 [94/97] (90.6–99.2)	93.0 [40/43] (79.9–98.2)	79.7 [94/118] (71.1–86.3)	83.2 [134/161] (76.6–88.7)	0.006#
US radiomics + clinical predictors model	0.832 (0.767–0.897)	73.4 [47/64] (60.7–83.3)	84.5 [82/97] (75.5–90.8)	75.8 [47/62] (63.0–85.4)	82.8 [82/99] (73.6–89.4)	80.1 [129/161] (73.1–86.0)	0.818*
Test set A							
Conventional US model	0.642 (0.484–0.799)	52.6 [10/19] (29.5–74.8)	81.1 [30/37] (64.3–91.4)	58.8 [10/17] (33.5–80.6)	76.9 [30/39] (60.3–88.3)	71.4 [40/56] (57.8–82.7)	
US radiomics model	0.822 (0.696–0.948)	63.2 [12/19] (38.6–82.8)	97.3 [36/37] (84.2–99.9)	92.3 [12/13] (62.1–99.6)	83.7 [36/43] (68.7–92.7)	85.7 [48/56] (73.8–93.6)	0.045#
US radiomics + clinical predictors model	0.855 (0.742–0.968)	73.7 [14/19] (48.6–89.9)	94.6 [35/37] (80.5–99.1)	87.5 [14/16] (60.4–97.8)	87.5 [35/40] (72.4–95.3)	87.5 [49/56] (75.9–94.8)	0.425*
Test set B							
Conventional US model	0.660 (0.497–0.823)	52.6 [10/19] (29.5–74.8)	79.3 [23/29] (59.7–91.3)	62.5 [10/16] (35.9–83.7)	71.9 [23/32] (53.0–85.6)	68.8 [33/48] (53.8–81.3)	
US radiomics model	0.853 (0.724–0.982)	84.2 [16/19] (59.5–95.8)	86.2 [25/29] (67.4–95.5)	80.0 [16/20] (55.7–93.4)	89.3 [25/28] (70.6–97.2)	85.4 [41/48] (72.2–93.9)	0.008#
US radiomics + clinical predictors model	0.877 (0.761–0.993)	68.4 [13/19] (43.5–86.4)	96.6 [28/29] (80.4–99.8)	92.9 [13/14] (64.2–99.6)	82.4 [28/34] (64.8–92.6)	85.4 [41/48] (72.2–93.9)	0.288*

Data in parentheses are 95% confidence interval, and data in brackets are used to calculate percentages

GB gallbladder, *US* ultrasound, *AUC* area under the curve, *PPV* positive predictive value, *NPV* negative predictive value

#p value of AUCs differ between the conventional US and US radiomics models

*p value of AUCs differ between the US radiomics model and US radiomics + clinical predictors model
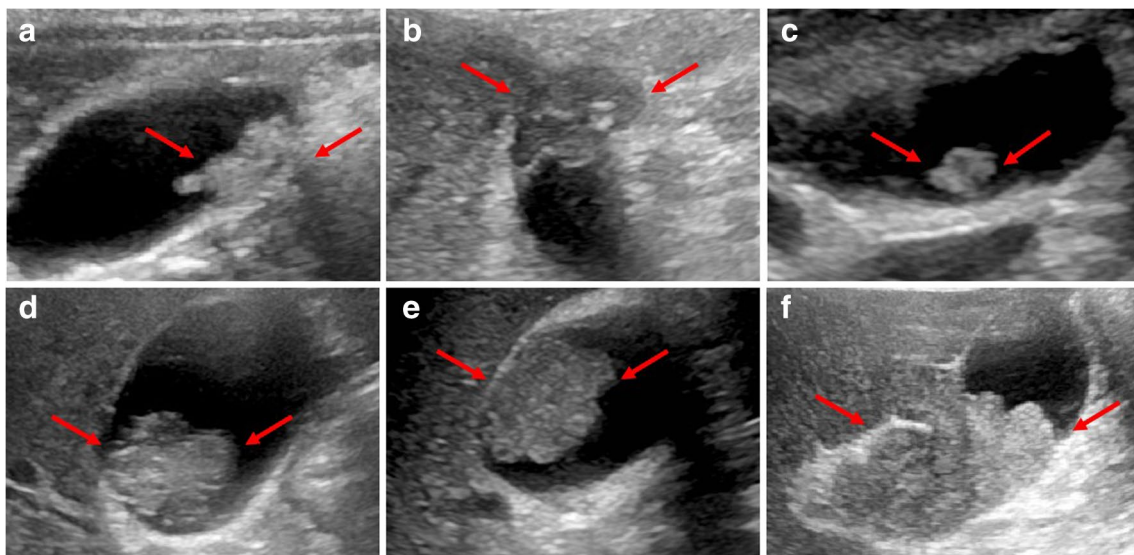

Fig. 3 Examples correctly diagnosed by ultrasound radiomics models. **a** A case of a 48-year-old woman with a cholesterol polyp. **b** A case of a 76-year-old man with an adenomyomatosis. **c** A case of a 61-year-old woman with an adenoma. **d** A case of a 72-year-old woman with a tubular adenoma. **e** A case of a 63-year-old woman with a gallbladder carcinoma. **f** A case of a 59-year-old woman with a gallbladder carcinoma. The lesions are indicated by red arrows

neoplastic lesions and those with non-neoplastic GB lesions and those with non-neoplastic GB lesions (all $p < 0.05$) (Supplementary Material S6).

Construction and validation of the ML-based US radiomics models

To discriminate between neoplastic and non-neoplastic GB lesions, 61 radiomics features in the training set were selected as key features after dimension reduction (Supplementary Material S7). The extreme gradient boosting (XGBoost)–based US radiomics model showed the highest AUC of 0.837 compared with the other 10 ML algorithms in the validation set (Supplementary Material S8). In the two external test sets, the model showed AUCs of 0.822 and 0.853 (Table 2). Examples of cases correctly diagnosed by the XGBoost-based US radiomics model are shown in Fig. 3a–c. There were 67, 32, and 24 patients with non-neoplastic GB lesions sized over 10 mm in the validation set and test sets A and B, respectively. Among them, the XGBoost-based US radiomics model distinguished 65, 31, and 21 non-neoplastic GB lesions and potentially reduced unnecessary cholecystectomy rate significantly in a speculative comparison with the current guidelines (3.1%, 3/97 vs. 53.6%, 52/97 for validation set, $p < 0.001$; 2.7%, 1/37

vs. 64.9%, 24/37 for test set A, $p < 0.001$; and 13.8%, 4/29 vs. 62.1%, 18/29 for test set B, $p = 0.001$).

To discriminate GB carcinomas from benign lesions, 33 critical US radiomics features were reserved using the same feature reduction and selection methods (Supplementary Material S7). The XGBoost-based US radiomics model showed the highest AUC of 0.904 in comparison to the other 10 ML algorithms in the validation set (Supplementary Material S8). In the two external test sets, the model showed AUCs of 0.909 and 0.979 in test sets A and B, respectively (Table 3). Example cases diagnosed correctly using the XGBoost-based US radiomics model are shown in Fig. 3d–f.

Comparison between the conventional US models and the ML-based US radiomics models

The largest diameter (≥ 16.1 mm) was identified as an independent predictor of neoplastic GB lesions in the training set (Supplementary Material S9). The diagnostic performance of the XGBoost-based US radiomics model was superior to that of the conventional US model based on the largest diameter in terms of AUCs (0.822–0.853 vs. 0.642–0.706, all $p < 0.05$) (Table 2 and Fig. 4a–c).

Table 3 The diagnostic performance of models in discrimination of GB carcinomas from benign lesions

Diagnostic models	AUC	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)	p value
Validation set							
Conventional US model	0.766 (0.649–0.884)	59.3 [16/27] (39.0–77.0)	94.0 [126/134] (88.2–97.2)	66.7 [16/24] (44.7–83.6)	92.0 [126/137] (85.8–95.7)	88.2 [142/161] (82.2–92.7)	
US radiomics model	0.904 (0.839–0.969)	70.4 [19/27] (49.7–85.5)	98.5 [132/134] (94.2–99.7)	90.5 [19/21] (68.2–98.3)	94.3 [132/140] (88.7–97.3)	93.8 [151/161] (88.9–97.0)	0.021#
US radiomics +clinical predictors model	0.908 (0.857–0.960)	70.4 [19/27] (49.7–85.5)	99.3 [133/134] (95.3–99.9)	95.0 [19/20] (73.1–99.7)	94.3 [133/141] (88.8–97.3)	94.4 [152/161] (89.7–97.4)	0.864*
Test set A							
Conventional US model	0.706 (0.515–0.897)	54.5 [6/11] (24.6–81.9)	88.9 [40/45] (75.2–95.8)	54.5 [6/11] (24.6–81.9)	88.9 [40/45] (75.2–95.8)	82.1 [46/56] (69.6–91.1)	
US radiomics model	0.909 (0.794–1.000)	63.6 [7/11] (31.6–87.6)	97.8 [44/45] (86.8–99.9)	87.5 [7/8] (46.7–99.3)	91.7 [44/48] (79.1–97.3)	91.1 [51/56] (80.4–97.0)	0.046#
US radiomics +clinical predictors model	0.930 (0.822–1.000)	72.7 [8/11] (39.3–92.7)	100.0 [45/45] (90.2–100.0)	100.0 [8/8] (59.8–100.0)	93.8 [45/48] (81.8–98.4)	94.6 [53/56] (85.1–98.9)	0.196*
Test set B							
Conventional US model	0.764 (0.582–0.947)	63.6 [7/11] (31.6–87.6)	89.2 [33/37] (73.6–96.5)	63.6 [7/11] (31.6–87.6)	89.2 [33/37] (73.6–96.5)	83.3 [40/48] (69.8–92.5)	
US radiomics model	0.979 (0.942–1.000)	100.0 [11/11] (67.9–100.0)	94.6 [35/37] (80.5–99.1)	84.6 [11/13] (53.7–97.3)	100.0 [35/35] (87.7–100.0)	95.8 [46/48] (85.8–99.5)	0.011#
US radiomics +clinical predictors model	0.967 (0.918–1.000)	81.8 [9/11] (47.8–96.8)	97.3 [36/37] (84.2–99.9)	90.0 [9/10] (54.1–99.5)	94.7 [36/38] (80.9–99.1)	93.8 [45/48] (82.8–98.7)	0.577*

Data in parentheses are 95% confidence interval, and data in brackets are used to calculate percentages

GB gallbladder, *US* ultrasound, *AUC* area under the curve, *PPV* positive predictive value, *NPV* negative predictive value

#p value of AUCs differ between the conventional US and US radiomics models

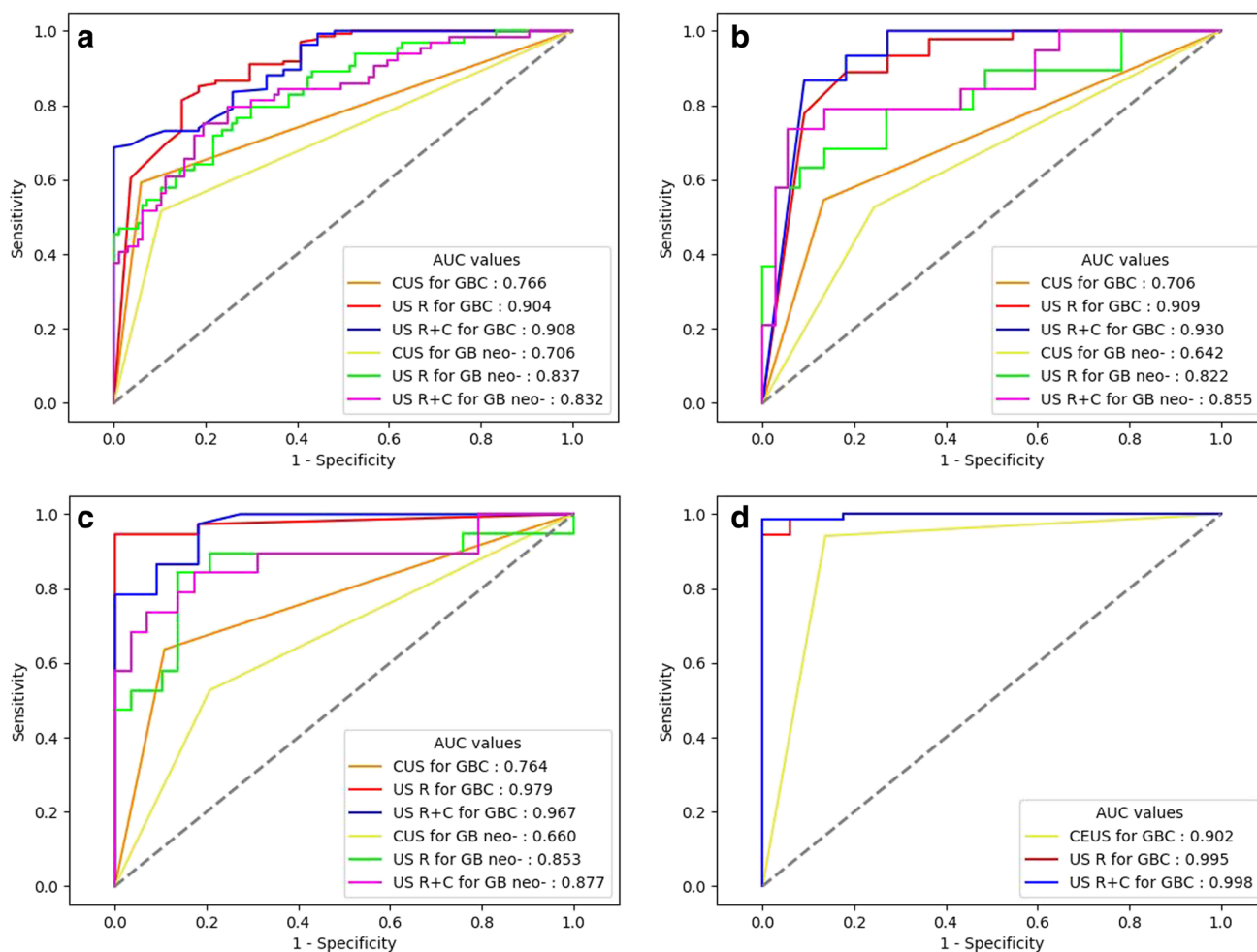*p value of AUCs differ between the US radiomics model and US radiomics + clinical predictors model

Fig. 4 The receiver operating characteristic curves of different methods in the validation (**a**), test A (**b**), test B (**c**), and comparison (**d**) sets. CUS, conventional ultrasound; CEUS, contrast-enhanced ultrasound; US R, ultrasound radiomics model; US R + C, ultrasound radiomics incorporated clinical characteristics model; GBC, gallbladder carcinoma; GB neo-, neoplastic gallbladder lesion

The largest diameter (≥ 21.0 mm) and discontinuity of the GB wall were independent predictors of GB carcinomas (Supplementary Material S9). The diagnostic performance of the XGBoost-based US radiomics model was superior to that of the conventional US model based on the largest diameter and discontinuity of the GB wall in terms of AUCs (0.904–0.979 vs. 0.706–0.766, all $p < 0.05$) (Table 3 and Fig. 4a–c).

Comparison between ML-based US radiomics-incorporated clinical characteristics models and ML-based US radiomics models

Elevated CA19-9 levels (≥ 27 U/mL) and age (> 59 years) were independent clinical predictors of neoplastic GB lesions in the training set. Likewise, elevated CA19-9 levels (≥ 27 U/mL), presence of gallstone, female, and age (> 59 years) were predictors of GB carcinomas (Supplementary Material S10). For the discrimination of neoplastic GB lesions and GB carcinomas, after adding independent clinical predictors, no significant difference in AUCs (all $p > 0.05$ in the validation set, and test sets A and B) was observed between the US radiomics-incorporated clinical characteristics models and US radiomics models (Tables 2 and 3; Fig. 4a–c).

Comparison between the CEUS models and ML-based US radiomics models

Compared with benign lesions, GB carcinomas often displayed hyper-enhancement at the early phase (94.1% vs. 49.3%), washed out earlier (33.1 ± 7.6 s vs. 47.2 ± 14.0 s), and discontinuity of the GB wall (64.7% vs. 4.1%) on CEUS (all $p < 0.05$) (Supplementary Material S11). The optimal cutoff value for washout time was 38 s. These three features were identified as independent characteristics of GB

Table 4 The diagnostic performance of CEUS models in discrimination of GB carcinomas from benign lesions in the "comparison" set

CEUS models	AUC	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)
CEUS characteristics						
Hyper-enhancement extent at early phase	0.724 (0.609–0.839)	94.1 [16/17] (69.2–99.7)	50.7 [37/73] (38.8–62.5)	30.8 [16/52] (19.1–45.3)	97.4 [37/38] (84.6–99.9)	58.9 [53/90] (48.6–68.5)
Washout time within 38 s	0.802 (0.683–0.921)	82.4 [14/17] (55.8–95.3)	78.1 [57/73] (66.6–86.6)	46.7 [14/30] (28.8–65.4)	95.0 [57/60] (85.2–98.7)	78.9 [71/90] (69.3–86.1)
Discontinuity of GB wall	0.803 (0.660–0.946)	64.7 [11/17] (38.6–84.7)	95.9 [70/73] (87.7–98.9)	78.6 [11/14] (48.8–94.3)	92.1 [70/76] (83.0–96.7)	90.0 [81/90] (81.9–94.9)
Diagnostic criterion 1	0.847 (0.756–0.939)	94.1 [16/17] (69.2–99.7)	75.3 [55/73] (63.6–84.4)	47.1 [16/34] (30.2–64.6)	98.2 [55/56] (89.2–99.9)	78.9 [71/90] (69.3–86.1)
Diagnostic criterion 2	0.902 (0.822–0.983)	94.1 [16/17] (69.2–99.7)	86.3 [63/73] (75.8–92.9)	61.5 [16/26] (40.7–79.1)	98.4 [63/64] (90.5–99.9)	87.8 [79/90] (79.3–93.2)

Data in parentheses are 95% confidence interval, and data in brackets are used to calculate percentages

Diagnostic criterion 1: Discontinuity of GB wall or washout time within 38 s

Diagnostic criterion 2: Any two of the following three independent characteristics appearing on CEUS: discontinuity of the GB wall, washout time within 38 s, and hyperenhancement extent in the early phase

CEUS contrast-enhanced ultrasound, *GB* gallbladder, *AUC* area under the curve, *PPV* positive predictive value, *NPV* negative predictive value

carcinomas (Supplementary Material S12). When discontinuity of the GB wall or washout time within 38 s was considered (Diagnostic criterion 1), AUC was 0.847. When any of the two characteristics appeared (Diagnostic criterion 2), AUC increased to 0.902 (Table 4).

In comparison with the Diagnostic criterion 2 on CEUS, the AUC of the XGBoost-based US radiomics model was higher for GB carcinomas prediction (0.995 vs. 0.902, $p = 0.011$) in the "comparison" set. No significant difference in AUCs was observed between the XGBoost-based US radiomics-incorporated clinical characteristics model and XGBoost-based US radiomics model (AUC: 0.998 vs. 0.995, $p = 0.384$) (Table 5 and Fig. 4d). Example cases that had atypical manifestations on CEUS but were diagnosed

correctly by the XGBoost-based US radiomics model are shown in Fig. 5.

Discussion

Our multi-institutional and prospective study indicated that the diagnostic performance of the XGBoost-based US radiomics models was significantly superior to that of conventional US models in terms of AUC for the diagnosis of neoplastic GB lesions and GB carcinomas. Meanwhile, the XGBoost-based US radiomics model could potentially reduce the unnecessary cholecystectomy rate in a speculative comparison with the current consensus guidelines

Table 5 The diagnostic performance of models in discrimination of GB carcinomas from benign lesions in the "comparison" set

Diagnostic models	AUC	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)	*p* value
CEUS model[†]	0.902 (0.822–0.983)	94.1 [16/17] (69.2–99.7)	86.3 [63/73] (75.8–92.9)	61.5 [16/26] (40.7–79.1)	98.4 [63/64] (90.5–99.9)	87.8 [79/90] (79.3–93.2)	
US radiomics model	0.995 (0.986–1.000)	94.1 [16/17] (69.2–99.7)	98.6 [72/73] (91.6–99.9)	94.1 [16/17] (69.2–99.7)	98.6 [72/73] (91.6–99.9)	97.8 [88/90] (92.2–99.7)	0.011*
US radiomics + clinical predictors model	0.998 (0.992–1.000)	100.0 [17/17] (77.1–100.0)	98.6 [72/73] (91.6–99.9)	94.4 [17/18] (70.6–99.7)	100.0 [72/72] (93.7–100.0)	98.9 [89/90] (94.0–100.0)	0.384[#]

Data in parentheses are 95% confidence interval, and data in brackets are used to calculate percentages.

US ultrasound, *CEUS* contrast-enhanced ultrasound, *AUC* area under the curve, *PPV* positive predictive value, *NPV* negative predictive value

[†]CEUS model adopted Diagnostic criterion 2

*p value of AUCs differ between the US radiomics model and CEUS model

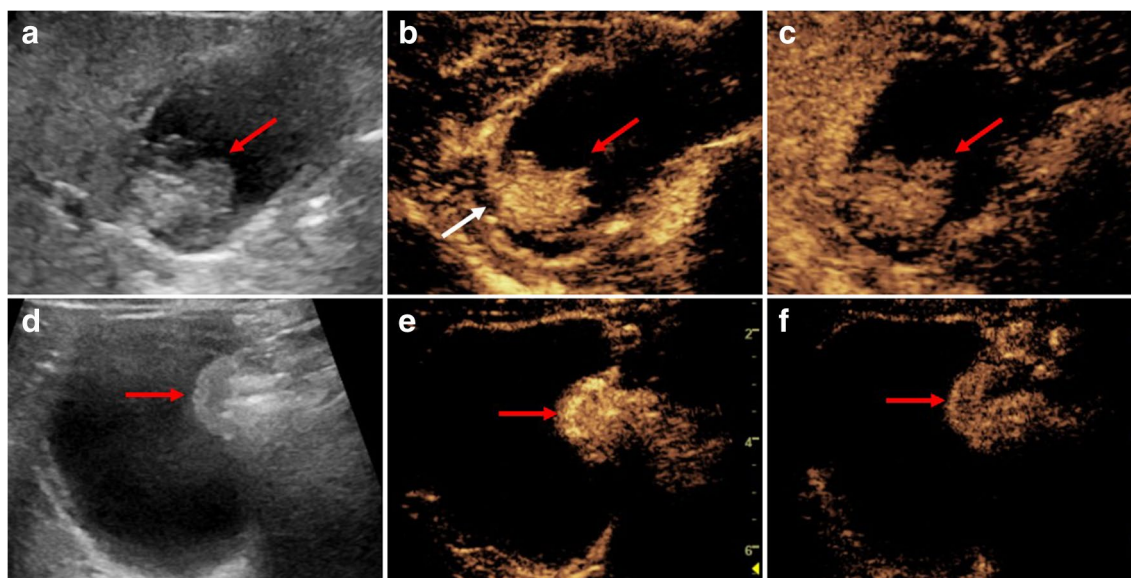[#]p value of AUCs differ between the US radiomics model and US radiomics + clinical predictors model

Fig. 5 Examples of cases with atypical manifestation on CEUS diagnosed correctly by the ultrasound radiomics method. **a–c** A case of a 72-year-old woman with a tubular adenoma. **a** US exhibits a hyperechoic lesion of 31.0 mm (red arrow). **b** CEUS exhibits hyper-enhancement (red arrow) at 19 s after contrast agent injection and local discontinuity of the gallbladder wall (white arrow). **C** The lesion turns into hypo-enhancement (red arrow) at 36 s after contrast agent injection. **d–f** A case of a 70-year-old woman with a gallbladder carcinoma. **d** US exhibits a hyperechoic lesion of 15.0 mm (red arrow). **e** CEUS exhibits hyper-enhancement (red arrow) at 17 s after contrast agent injection. **f** The lesion turns into hypo-enhancement (red arrow) at 68 s after contrast agent injection and is misdiagnosed as adenoma by CEUS

(cholecystectomy for lesions sized over 10 mm) for non-neoplastic GB lesions sized over 10 mm. Furthermore, the diagnostic performance of the XGBoost-based US radiomics model could overmatch that of the CEUS model in discriminating GB carcinomas from benign lesions.

GB masses are common clinically, and their diagnosis can be challenging because of conflicting and unclear imaging manifestations [34]. The 2017 European Joint Guidelines recommend cholecystectomy for GB polypoid lesions sized over 10 mm [35]. The threshold setting of the guideline is based on the limitation of risk stratification among these GB lesions using routine imaging modalities to minimise the omission of neoplastic lesions. Nevertheless, this strategy also leads to unnecessary cholecystectomy, which results in potential post-cholecystectomy syndrome and increases the financial burden on patients [10, 36–38]. On the other hand, precise and timely diagnosis of malignant and benign GB lesions is a topic of great interest for evaluating the best treatment options and prognosis assessment [11, 39, 40]. Therefore, devising methods to minimise the unnecessary cholecystectomy rate and reduce the missed diagnosis of GB carcinoma as much as possible promises to be of great interest in clinical practice. To the best of our knowledge, our multi-institutional study is the first to develop ML-based US radiomics models for the risk stratification of GB masses from two clinical perspectives: diagnosis of neoplastic lesions and GB carcinomas.

Constantly evolving and improving radiomics approaches are a series of processes that utilise quantitative image analysis to extract features in conjunction with computerised algorithms for feature selection, and finally used for diagnosis or prediction [41–43]. We selected interpretable ML classifiers that can effectively exploit the obtainable data. In comparison with the other 10 ML classifiers, our study showed that the XGBoost classifier had the highest diagnostic performance for discriminating both neoplastic GB lesions and GB carcinomas. XGBoost is an ML algorithm that improves the gradient boosting machine framework with system optimisations and algorithm enhancements and is outstanding in parallel computing, missing value processing, overfitting control, and prediction generalisation [44].

Previous studies have reported that size of the lesion (14–16 mm) is a main risk factor for neoplastic GB lesions [15, 17, 45]. The suggested risk factors of GB carcinomas include size of the lesion (20–22 mm), sessile shape, and discontinuity of the GB wall [9, 17, 46]. Similarly, our study found that the largest diameter (>16.1 mm) was a significant independent predictor of neoplastic GB lesions, and the largest diameter (>21.0 mm) and discontinuity of the GB wall were independent predictors of GB carcinomas. However, with respect to GB carcinoma discrimination, the accuracy of US is moderate (approximately 70%) [20]. In our study, the accuracy was increased to 91.1–95.8% when the XGBoost-based US radiomics model was used. Choi

et al reported that the accuracy for predicting a neoplastic polyp was 66.9–72.1% in high-resolution US [17]. Our XGBoost-based US radiomics model was able to improve the accuracy to 83.2–85.7%.

As a complement to conventional US, CEUS can significantly improve the AUC (approximately 0.90) in discriminating GB carcinomas [20, 47]. Generally, destruction of the GB wall, irregular vascularity, heterogeneous enhancement, and earlier contrast agent wash-out time (28–40 s) are characteristics of GB carcinoma on CEUS [21, 22, 48, 49]. However, certain problems can affect the clinical application of CEUS. It requires the setup of an effective venous channel and some patients may be contraindicated or allergic to contrast agents. Moreover, high empirical dependence and the requirement for high-end US machines usually cause CEUS to be performed in hospitals with appropriate expertise and resources [24]. For ML-based radiomics, digital US images fed to the proposed models have already been routinely acquired without extra effort. This cost-effective approach would help radiologists differentiate benign GB lesions from malignant ones, leading to better utilisation of healthcare resources.

Yuan et al analysed spatial and morphological features and constructed an SVM-based US radiomics model to distinguish true and pseudo-GB polyps with an AUC of 0.898 [30]. Our study extracted more varieties of US radiomics features and utilised different ML algorithms to establish optimal ML-based US radiomics models for more effective risk stratification of GB masses. Several studies have suggested that female, older age, presence of gallstone, and elevated CA19-9 and CEA levels are potential risk factors for GB carcinomas [40, 50–52]. In this study, after adding independent clinical predictors, the incorporated US radiomics models did not show significantly superior diagnostic performance compared with US radiomics model alone. A possible reason for this is that these clinical characteristics might be of limited value in the risk prediction of GB masses, in contrast to US radiomics features.

This study had several limitations. First, the sample size was relatively limited. However, the study was produced from a multi-centre dataset and was tested prospectively in two external test sets involving different US scanners, which indicated the reliability of the results. Second, since surgery-confirmed pathology was used as the gold standard, GB lesions sized less than 6 mm were not included, which are commonly not recommended for cholecystectomy [34]. Third, GB lesions which could not be demarcated from the adjacent liver tissue were excluded. These cases lost the GB structure, and the ROI could not be delineated manually. Meanwhile, the study focused on GB masses; other GB diseases presenting only as thickening of the GB wall were not included. Lastly, our proposed models were not compared

with current treatment guidelines in the real clinical practice, which need to be conducted further in a randomised clinical trial to validate our results.

In conclusion, our proposed ML-based US radiomics models are capable of preoperatively predicting the risk of GB masses and have the potential to decrease unnecessary cholecystectomy rate, and substitute for the use of CEUS.

Declarations

Guarantor The scientific guarantor of this publication is Hui-Xiong Xu.

Conflict of interest These authors declare that they have no conflict of interest.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was obtained from patients in this study.

Ethical approval Institutional Review Board approval was obtained.

Methodology
● prospective
● diagnostic or prognostic study
● performed at four institutions

References

1. McCain RS, Diamond A, Jones C, Coleman HG (2018) Current practices and future prospects for the management of gallbladder polyps: a topical review. World J Gastroenterol 24:2844–2852
2. Ganeshan D, Kambadakone A, Nikolaidis P et al (2021) Current update on gallbladder carcinoma. Abdom Radiol (NY) 46:2474–2489
3. Mellnick VM, Menias CO, Sandrasegaran K et al (2015) Polypoid lesions of the gallbladder: disease spectrum with pathologic correlation. Radiographics 35:387–399
4. Ramachandran A, Srivastava DN, Madhusudhan KS (2021) Gallbladder cancer revisited: the evolving role of a radiologist. Br J Radiol 94:20200726
5. Randi G, Franceschi S, La Vecchia C (2006) Gallbladder cancer worldwide: geographical distribution and risk factors. Int J Cancer 118:1591–1602
6. Jang JY, Kim SW, Lee SE et al (2009) Differential diagnostic and staging accuracies of high resolution ultrasonography, endoscopic

ultrasonography, and multidetector computed tomography for gallbladder polypoid lesions and gallbladder cancer. Ann Surg 250:943–949

7. Kai K, Aishima S, Miyazaki K (2014) Gallbladder cancer: clinical and pathological approach. World J Clin Cases 2:515–521

8. Babu BI, Dennison AR, Garcea G (2015) Management and diagnosis of gallbladder polyps: a systematic review. Langenbecks Arch Surg 400:455–462

9. Bhatt NR, Gillis A, Smoothey CO, Awan FN, Ridgway PF (2016) Evidence based management of polyps of the gall bladder: a systematic review of the risk factors of malignancy. Surgeon 14:278–286

10. Park HY, Oh SH, Lee KH, Lee JK, Lee KT (2015) Is cholecystectomy a reasonable treatment option for simple gallbladder polyps larger than 10 mm? World J Gastroenterol 21:4248–4254

11. Coburn NG, Cleary SP, Tan JC, Law CH (2008) Surgery for gallbladder cancer: a population-based analysis. J Am Coll Surg 207:371–382

12. Yu MH, Kim YJ, Park HS, Jung SI (2020) Benign gallbladder diseases: imaging techniques and tips for differentiating with malignant gallbladder diseases. World J Gastroenterol 26:2967–2986

13. Elmasry M, Lindop D, Dunne DF et al (2016) The risk of malignancy in ultrasound detected gallbladder polyps: a systematic review. Int J Surg 33:28–35

14. Kim JH, Lee JY, Baek JH et al (2015) High-resolution sonography for distinguishing neoplastic gallbladder polyps and staging gallbladder cancer. AJR Am J Roentgenol 204:W150-159

15. Kim JS, Lee JK, Kim Y, Lee SM (2016) US characteristics for the prediction of neoplasm in gallbladder polyps 10 mm or larger. Eur Radiol 26:1134–1140

16. Lee JS, Kim JH, Kim YJ et al (2017) Diagnostic accuracy of transabdominal high-resolution US for staging gallbladder cancer and differential diagnosis of neoplastic polyps compared with EUS. Eur Radiol 27:3097–3103

17. Choi TW, Kim JH, Park SJ et al (2018) Risk stratification of gallbladder polyps larger than 10 mm using high-resolution ultrasonography and texture analysis. Eur Radiol 28:196–205

18. Cheng Y, Wang M, Ma B, Ma X (2018) Potential role of contrast-enhanced ultrasound for the differentiation of malignant and benign gallbladder lesions in east Asia: a meta-analysis and systematic review. Medicine (Baltimore) 97:e11808

19. Negrão de Figueiredo G, Mueller-Peltzer K, Armbruster M, Rübenthaler J, Clevert DA (2019) Contrast-enhanced ultrasound (CEUS) for the evaluation of gallbladder diseases in comparison to cross-sectional imaging modalities and histopathological results. Clin Hemorheol Microcirc 71:141–149

20. Liang X, Jing X (2020) Meta-analysis of contrast-enhanced ultrasound and contrast-enhanced harmonic endoscopic ultrasound for the diagnosis of gallbladder malignancy. BMC Med Inform Decis Mak 20:235

21. Xie XH, Xu HX, Xie XY et al (2010) Differential diagnosis between benign and malignant gallbladder diseases with real-time contrast-enhanced ultrasound. Eur Radiol 20:239–248

22. Liu LN, Xu HX, Lu MD et al (2012) Contrast-enhanced ultrasound in the diagnosis of gallbladder diseases: a multi-center experience. PLoS One 7:e48371

23. Yuan Z, Liu X, Li Q et al (2021) Is contrast-enhanced ultrasound superior to computed tomography for differential diagnosis of gallbladder polyps? a cross-sectional study. Front Oncol 11:657223

24. Liu Z, Zhang XY, Shi YJ et al (2017) Radiomics analysis for evaluation of pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer. Clin Cancer Res 23:7253–7262

25. Huang YQ, Liang CH, He L (2016) Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. J Clin Oncol 34:2157–2164

26. Lambin P, Leijenaar RTH, Deist TM et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 14:749–762

27. Lambin P, Rios-Velazquez E, Leijenaar R et al (2012) Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 48:441–446

28. Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: images are more than pictures, they are data. Radiology 278:563–577

29. Jeong Y, Kim JH, Chae HD et al (2020) Deep learning-based decision support system for the diagnosis of neoplastic gallbladder polyps on ultrasonography: preliminary results. Sci Rep 10:7700

30. Yuan HX, Yu QH, Zhang YQ et al (2020) Ultrasound radiomics effective for preoperative identification of true and pseudo gallbladder polyps based on spatial and morphological features. Front Oncol 10:1719

31. Zhao CK, Ren TT, Yin YF et al (2021) A comparative analysis of two machine learning-based diagnostic patterns with thyroid imaging reporting and data system for thyroid nodules: diagnostic performance and unnecessary biopsy rate. Thyroid 31:470–481

32. Yushkevich PA, Piven J, Hazlett HC et al (2006) User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31:1116–1128

33. Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7:91

34. Foley KG, Lahaye MJ, Thoeni RF (2022) Management and follow-up of gallbladder polyps: updated joint guidelines between the ESGAR, EAES, EFISDS and ESGE. Eur Radiol 32:3358–3368

35. Wiles R, Thoeni R, Barbu S et al (2017) Management and follow-up of gallbladder polyps: joint guidelines between the European Society of Gastrointestinal and Abdominal Radiology (ESGAR), European Association for Endoscopic Surgery and other Interventional Techniques (EAES), International Society of Digestive Surgery - European Federation (EFISDS) and European Society of Gastrointestinal Endoscopy (ESGE). Eur Radiol 27:3856–3866

36. Isherwood J, Oakland K, Khanna A (2019) A systematic review of the ae- tiology and management of post cholecystectomy syndrome. Surgeon 17:33–42

37. Womack NA, Crider RL (1947) The persistence of symptoms following cholecystectomy. Ann Surg 126:31–55

38. Tarnasky PR (2016) Post-cholecystectomy syndrome and sphincter of Oddi dysfunction: past, present and future. Expert Rev Gastroenterol Hepatol 10:1359–1372

39. Misra S, Chaturvedi A, Misra NC, Sharma ID (2003) Carcinoma of the gallbladder. Lancet Oncol 4:167–176

40. Tian YH, Ji X, Liu B et al (2015) Surgical treatment of incidental gallbladder cancer discovered during or following laparoscopic cholecystectomy. World J Surg 39:746–752

41. Zhou W, Yang Y, Yu C et al (2021) Ensembled deep learning model outperforms human experts in diagnosing biliary atresia from sonographic gallbladder images. Nat Commun 12:1259

42. Ji GW, Zhang YD, Zhang H et al (2019) Biliary tract cancer at CT: a radiomics-based model to predict lymph node metastasis and survival outcomes. Radiology 290:90–98

43. Xiang F, Liang X, Yang L, Liu X, Yan S (2022) Contrast-enhanced CT radiomics for prediction of recurrence-free survival in gallbladder carcinoma after surgical resection. Eur Radiol 32:7087–7097

44. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining-KDD 2016, San Francisco, CA, USA, pp 785–794

45. Cha BH, Hwang JH, Lee SH et al (2011) Pre-operative factors that can predict neoplastic polypoid lesions of the gallbladder. World J Gastroenterol 17:2216–2222

46. French DG, Allen PD, Ellsmere JC (2013) The diagnostic accuracy of transabdominal ultrasonography needs to be considered when managing gallbladder polyps. Surg Endosc 27:4021–4025

47. Zhang HP, Bai M, Gu JY, He YQ, Qiao XH, Du LF (2018) Value of contrast-enhanced ultrasound in the differential diagnosis of gallbladder lesion. World J Gastroenterol 24:744–751

48. Chen LD, Huang Y, Xie XH et al (2017) Diagnostic nomogram for gallbladder wall thickening mimicking malignancy: using contrast-enhanced ultrasonography or multi-detector computed tomography? Abdom Radiol (NY) 42:2436–2446

49. Sun LP, Guo LH, Xu HX et al (2015) Value of contrast-enhanced ultrasound in the differential diagnosis between gallbladder adenoma and gallbladder adenoma canceration. Int J Clin Exp Med 8:1115–1121

50. Ganeshan D, Kambadakone A, Nikolaidis P, Subbiah V, Subbiah IM, Devine C (2021) Current update on gallbladder carcinoma. Abdom Radiol (NY) 46:2474–2489

51. Wang YF, Feng FL, Zhao XH et al (2014) Combined detection tumor markers for diagnosis and prognosis of gallbladder cancer. World J Gastroenterol 20:4085–4092

52. Shukla VK, Gurubachan, Sharma D, Dixit VK, Usha (2006) Diagnostic value of serum CA242, CA 19-9, CA 15-3 and CA 125 in patients with carcinoma of the gallbladder. Trop Gastroenterol 27:160-165