npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01991-7

Machine learning model for differentiating xanthogranulomatous cholecystitis and gallbladder cancer in multicenter largescale study

Check for updates

Ke Zhang[1,7], Jiajia He[2,7], Weiyue Ji[3], Qunyan Pan[2], Weilv Xiong[4], Liping Wang[5], Weiqi Sun[6], Liting Xie[1] ✉ & Tianan Jiang[1] ✉

Preoperative differentiation between xanthogranulomatous cholecystitis (XGC) and gallbladder cancer (GBC) remains challenging due to overlapping clinical and imaging features. This multicenter retrospective study developed a machine learning (ML) model, LIDGAX, using preoperative clinical, imaging, and laboratory data from 1246 patients (554 XGC, 692 GBC). Twelve variables were identified as independent predictors via multivariate logistic regression and least absolute shrinkage and selection operator analyses. LIDGAX achieved area under the curve (AUC) values of 0.94 (internal validation) and 0.88 (external testing), outperforming the other five ML models. Calibration and decision curve analyses demonstrated its superior clinical utility. Compared to six radiologists, LIDGAX improved sensitivity (1.2–8.5%), specificity (0.0–4.6%), and balanced accuracy (1.8–6.6%), while reducing average diagnostic time per patient by 30.44–35.76 s. LIDGAX was deployed on an open-source online platform, maintaining high performance (AUC 0.95, accuracy 0.92). This non-invasive tool shows strong potential for clinical translation in preoperative differentiation of XGC and GBC.

In real-world clinical scenario, gallbladder diseases are primarily categorized as benign cholecystitis or malignant gallbladder cancer. Xanthogranulomatous cholecystitis (XGC) is a rare type of chronic inflammatory gallbladder diseases characterized histologically by focal or diffuse inflammatory infiltration of foamy cells, multinucleated giant cells, lymphocytes, and fibroblasts[1]. In contrast, gallbladder cancer (GBC) is the most aggressive malignancy of the biliary tract[2], with most cases having a poor prognosis due to its aggressive nature and limited therapeutic options[2,3]. Accurate differentiation between XGC and GBC has critical implications for treatment decisions. XGC, being a chronic inflammatory condition, is typically managed with laparoscopic cholecystectomy[4]. Conversely, GBC often necessitates more aggressive interventions depending on the stage, including liver resection, bile duct excision, lymph node dissection, and possibly neoadjuvant chemoradiotherapy before surgery[5]. Misdiagnosing GBC as XGC could result in undertreatment, such as lack of preoperative surgical planning, incomplete resection, or inadequate follow-up, potentially accelerating disease progression. Conversely, misidentifying XGC as GBC could lead to unnecessary surgical procedures like liver resection or extensive lymph node dissection, as well as increased complications and resource usage. Therefore, distinguishing XGC from GBC preoperatively is crucial in clinical practice, which could reduce intraoperative frozen section misdiagnosis risks and guide postoperative surveillance strategies. However, this differentiation is highly challenging due to the overlapping clinical characteristics and imaging features of the two diseases, such as abdominal pain, jaundice, gallbladder wall thickening, and invasion of adjacent organs[6,7]. Additionally, misdiagnosis is common, with reported rates ranging from 10% to 30%[6,8–10]. Thus, non-invasive preoperative biomarkers are needed to improve differentiation between XGC and GBC and reduce overtreatment.

[1]Department of Ultrasound Medicine, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China. [2]Department of Ultrasound Medicine, Beilun District People's Hospital, Ningbo, Zhejiang, China. [3]Department of Radiology, Beilun District People's Hospital, Ningbo, Zhejiang, China. [4]Department of Ultrasound Medicine, Huzhou Central Hospital, Huzhou, Zhejiang, China. [5]Department of Ultrasound Medicine, Qingchun Hospital of Zhejiang Province, Hangzhou, Zhejiang, China. [6]Department of Ultrasound Medicine, The Second Affiliated Hospital, Jiaxing University, Jiaxing, Zhejiang, China. [7]These authors contributed equally: Ke Zhang, Jiajia He. ✉e-mail: 18258204889@163.com; tiananjiang@zju.edu.cn

Preoperative diagnostic imaging for gallbladder diseases commonly includes ultrasound (US), contrast-enhanced computed tomography (CECT), and magnetic resonance imaging (MRI). Previous studies have explored imaging features on US, CT, and MRI to differentiate XGC from GBC. For instance, gallbladder mucosal line continuity, a low-density border surrounding the lesion, diffuse gallbladder wall thickening, hypoattenuated or hypoechoic nodules in the thickened walls, and the presence of calculi were found to be strongly associated with XGC[11–14]. Lee et al.[15] compared the diagnostic performance of these imaging modalities and demonstrated that MRI had the highest accuracy, followed by US and CT. Each imaging technique has distinct advantages and limitations. US is frequently used for initial screening due to its high temporal resolution, ability to observe blood flow, convenience, absence of radiation, and low cost, despite its lower spatial resolution. CT is effective for visualizing liver and lymph node involvement but involves radiation exposure and lower soft tissue resolution. MRI provides superior soft tissue resolution and effectively shows gallbladder wall invasion but is time-consuming and susceptible to respiratory motion artifacts. Therefore, integrating the features from these imaging modalities may enhance the differential diagnosis of XGC and GBC.

Machine learning (ML) algorithms, as novel non-invasive approaches, can flexibly and efficiently analyze high-throughput data, enabling the discovery of complex relationships between variables[16]. Due to their advanced capabilities, various ML techniques are widely used to identify disease risk factors, predict treatment outcomes in patients with tumors, and support clinicians in real-world practice[17–20]. However, a previous study has indicated that different ML methods can produce varying performance results[21]. Therefore, identifying the most effective ML techniques is essential for ensuring accurate and reliable predictions and classifications in clinical applications. To our knowledge, no current research has employed multiple ML methods on large-scale, multi-center datasets to differentiate XGC from GBC in real-world settings.

In this study, we developed distinct ML-based models to differentiate between XGC and GBC using preoperative clinical characteristics, imaging features, and laboratory tests. We compared the performance of these models and validated them on an independent, external, multi-center testing cohort to assess generalizability. We then evaluated the optimal model against results from a reader study involving six radiologists with varying levels of experience. Finally, we explored the application of the most effective model in real-world clinical settings, including outpatient, inpatient, and physical examination settings. Figure 1 illustrates the framework of the proposed ML-based model.

Results

Baseline characteristics

The baseline characteristics of patients in the four cohorts are summarized in Table 1. Between January 2023 and February 2024, a total of 1246 patients were included in the analysis, comprising 554 patients diagnosed with XGC and 692 with GBC (Fig. 2). The median age across the overall dataset was 63.0 years, with 574 males (46.1%) and 672 females (53.9%). The training cohort consisted of 674 patients (326 with XGC and 348 with GBC), while the internal validation cohort included 169 patients (82 with XGC and 87 with GBC). The external testing cohort contained 279 patients, distributed as 90 XGC and 189 GBC cases. Detailed definitions of clinical data, laboratory tests, and imaging features are available in Supplementary Tables 1, 2. Further details on the distribution of XGC and GBC across all cohorts can be found in Supplementary Table 3.

Construction of ML-based models

A total of 79 variables were collected, including clinical characteristics ($n = 9$), imaging features ($n = 19$), and laboratory tests ($n = 51$). To identify and retain only the most relevant indicators, univariate and multivariate logistic regression analyses were performed on all variables (Supplementary Table 4). The multivariate analysis identified 20 variables as independently associated with either XGC or GBC. Specifically, the variables independently associated with XGC included male, epigastric pain, hyperechoic findings on US, presence of gallbladder stones, regular gallbladder morphology, reduced gallbladder size, presence of intramural nodules,

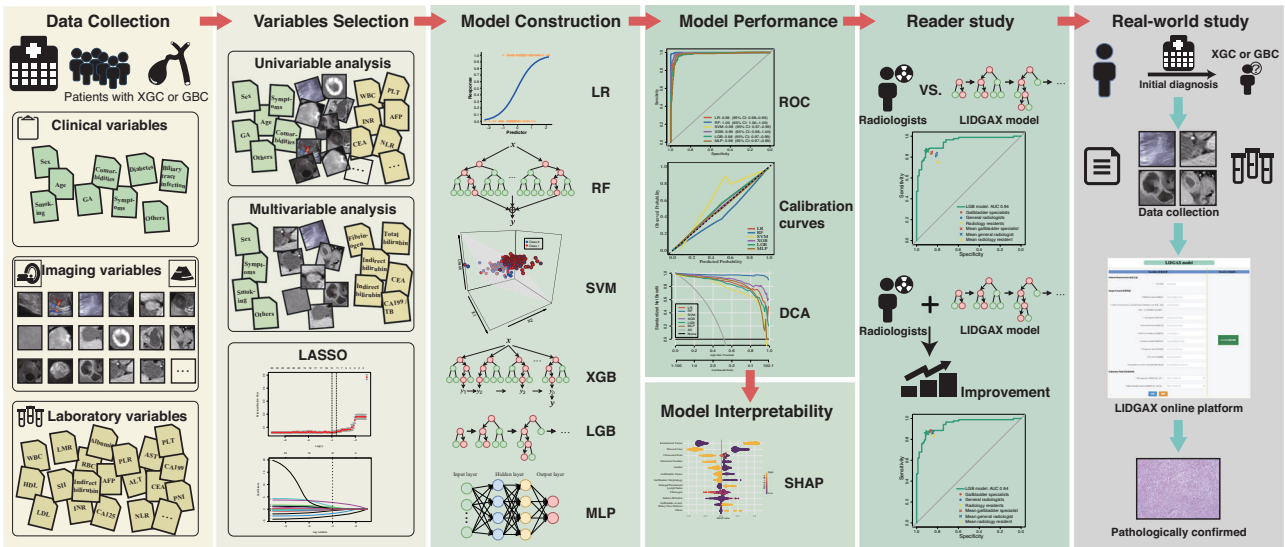


Fig. 1 | Overall design of the framework. (1) Data collection. Clinical, imaging, and laboratory variables were collected for patients with XGC and GBC. (2) Variable selection. Univariate and multivariate logistic regression analyses, followed by LASSO regression, were conducted to select relevant variables, resulting in 12 robust features. (3) Model construction. Six ML models (LR, RF, SVM, XGB, LGB, and MLP) were constructed using the selected features. (4) Model performance. The six models were evaluated using AUC, calibration curves, and DCA across training, internal validation, and external testing cohorts. (5) Model interpretability. SHAP was used to interpret and visualize the feature importance of the optimal model (LIDGAX). (6) Reader study. LIDGAX's diagnostic performance was compared to radiologists, both unassisted and LIDGAX-assisted, in distinguishing XGC from GBC. (7) Real-world study. An online platform for LIDGAX was developed to facilitate its application in clinical practice, enabling translation to real-world settings. XGC Xanthogranulomatous cholecystitis, GBC gallbladder cancer, LASSO least absolute shrinkage and selection operator, LR logistic regression, RF random forest, SVM support vector machine, XGB eXtreme gradient boosting, LGB light gradient boosting, MLP multilayer perceptron, ROC receiver operating characteristic, AUC area under the curve, DCA decision curve analysis, SHAP SHapley Additive exPlanations, LIDGAX LightGBM Intelligent Differentiator for XGC and GBC.

Table 1 | The baseline characteristics of the training, internal validation, external testing, and real-world cohorts

Variables	Overall (*n* = 1246)	The training cohort (*n* = 674)	The internal validation cohort (*n* = 169)	The external testing cohort (*n* = 279)	Real-world cohort (*n* = 124)
Category					
XGC	554 (44.5)	326 (48.4)	82 (48.5)	90 (32.3)	56 (45.2)
GBC	692 (55.5)	348 (51.6)	87 (51.5)	189 (67.7)	68 (54.8)
Clinical variables					
Sex					
Female	672 (53.9)	361 (53.6)	84 (49.7)	162 (58.1)	65 (52.4)
Male	574 (46.1)	313 (46.4)	85 (50.3)	117 (41.9)	59 (47.6)
Age (years)[a]	63.00 [55.00, 71.00]	62.00 [54.00, 70.00]	64.00 [57.00, 71.00]	65.00 [57.00, 73.00]	65.00 [58.00, 72.00]
Symptoms					
Asymptomatic	312 (25.0)	159 (23.6)	50 (29.6)	57 (20.4)	46 (37.1)
Fever	36 (2.9)	13 (1.9)	4 (2.4)	11 (3.9)	8 (6.5)
Epigastric pain	720 (57.8)	401 (59.5)	84 (49.7)	187 (67.0)	48 (38.7)
Jaundice	71 (5.7)	46 (6.8)	8 (4.7)	10 (3.6)	7 (5.6)
Gastrointestinal symptoms	63 (5.1)	31 (4.6)	12 (7.1)	11 (3.9)	9 (7.3)
Weight loss	4 (0.3)	3 (0.4)	0 (0.0)	0 (0.0)	1 (0.8)
Others	40 (3.2)	21 (3.1)	11 (6.5)	3 (1.1)	5 (4.0)
Complications					
None	1105 (88.7)	599 (88.9)	143 (84.6)	255 (91.4)	108 (87.1)
Abscess	31 (2.5)	15 (2.2)	8 (4.7)	5 (1.8)	3 (2.4)
Severe pancreatitis	38 (3.0)	21 (3.1)	7 (4.1)	8 (2.9)	2 (1.6)
Perforation	28 (2.2)	10 (1.5)	5 (3.0)	9 (3.2)	4 (3.2)
Colon obstruction	25 (2.0)	17 (2.5)	3 (1.8)	1 (0.4)	4 (3.2)
Two or more above	19 (1.5)	12 (1.8)	3 (1.8)	1 (0.4)	3 (2.4)
Smoking					
Absent	1079 (86.6)	559 (82.9)	151 (89.3)	253 (90.7)	116 (93.5)
Present	167 (13.4)	115 (17.1)	18 (10.7)	26 (9.3)	8 (6.5)
Diabetes					
Absent	1087 (87.2)	579 (85.9)	146 (86.4)	249 (89.2)	113 (91.1)
Present	159 (12.8)	95 (14.1)	23 (13.6)	30 (10.8)	11 (8.9)
Gallbladder adenomyomatosis					
Absent	1191 (95.6)	649 (96.3)	156 (92.3)	268 (96.1)	118 (95.2)
Present	55 (4.4)	25 (3.7)	13 (7.7)	11 (3.9)	6 (4.8)
Biliary tract infection					
Absent	1144 (91.8)	648 (96.1)	160 (94.7)	224 (80.3)	112 (90.3)
Present	102 (8.2)	26 (3.9)	9 (5.3)	55 (19.7)	12 (9.7)
Other conditions (schistosomiasis, congenital biliary dilatation/cyst)					
Absent	1197 (96.1)	647 (96.0)	161 (95.3)	269 (96.4)	120 (96.8)
Present	49 (3.9)	27 (4.0)	8 (4.7)	10 (3.6)	4 (3.2)
Imaging variables					
Ultrasound echo					
Hypoechoic	499 (40.0)	265 (39.3)	68 (40.2)	109 (39.1)	57 (46.0)
Mixed echogenicity	42 (3.4)	9 (1.3)	1 (0.6)	32 (11.5)	0 (0.0)
Isoechoic	476 (38.2)	271 (40.2)	67 (39.6)	96 (34.4)	42 (33.9)
Hyperechoic	229 (18.4)	129 (19.1)	33 (19.5)	42 (15.1)	25 (20.2)
Blood flow signal					
None	609 (48.9)	348 (51.6)	89 (52.7)	110 (39.4)	62 (50.0)
Sparse	542 (43.5)	284 (42.1)	74 (43.8)	133 (47.7)	51 (41.1)
Abundant	95 (7.6)	42 (6.2)	6 (3.6)	36 (12.9)	11 (8.9)
Gallbladder stones					
Absent	368 (29.5)	197 (29.2)	49 (29.0)	77 (27.6)	45 (36.3)
Present	878 (70.5)	477 (70.8)	120 (71.0)	202 (72.4)	79 (63.7)
Biliary duct dilation					
Absent	827 (66.4)	438 (65.0)	121 (71.6)	181 (64.9)	87 (70.2)
Present	419 (33.6)	236 (35.0)	48 (28.4)	98 (35.1)	37 (29.8)
Gallbladder morphology					
Irregular	314 (25.2)	150 (22.3)	56 (33.1)	74 (26.5)	34 (27.4)
Regular	932 (74.8)	524 (77.7)	113 (66.9)	205 (73.5)	90 (72.6)

Table 1 (continued) | The baseline characteristics of the training, internal validation, external testing, and real-world cohorts

Variables	Overall (*n* = 1246)	The training cohort (*n* = 674)	The internal validation cohort (*n* = 169)	The external testing cohort (*n* = 279)	Real-world cohort (*n* = 124)
Gallbladder size					
Normal	583 (46.8)	327 (48.5)	74 (43.8)	123 (44.1)	59 (47.6)
Enlarged	494 (39.6)	250 (37.1)	66 (39.1)	128 (45.9)	50 (40.3)
Reduced	169 (13.6)	97 (14.4)	29 (17.2)	28 (10.0)	15 (12.1)
Type of gallbladder wall thickening					
Diffuse	605 (48.6)	356 (52.8)	90 (53.3)	108 (38.7)	51 (41.1)
Focal	537 (43.1)	299 (44.4)	67 (39.6)	104 (37.3)	67 (54.0)
None	104 (8.3)	19 (2.8)	12 (7.1)	67 (24.0)	6 (4.8)
Intramural nodules					
Absent	991 (79.5)	518 (76.9)	141 (83.4)	244 (87.5)	88 (71.0)
Present	255 (20.5)	156 (23.1)	28 (16.6)	35 (12.5)	36 (29.0)
Intraluminal tumor					
Absent	762 (61.2)	406 (60.2)	103 (60.9)	185 (66.3)	68 (54.8)
Present	484 (38.8)	268 (39.8)	66 (39.1)	94 (33.7)	56 (45.2)
Mucosal line					
Discontinuous	801 (64.3)	428 (63.5)	100 (59.2)	185 (66.3)	88 (71.0)
Continuous	445 (35.7)	246 (36.5)	69 (40.8)	94 (33.7)	36 (29.0)
Gallbladder perforation					
Absent	1200 (96.3)	654 (97.0)	162 (95.9)	266 (95.3)	118 (95.2)
Present	46 (3.7)	20 (3.0)	7 (4.1)	13 (4.7)	6 (4.8)
Abscess					
Absent	1208 (97.0)	654 (97.0)	158 (93.5)	276 (98.9)	120 (96.8)
Present	38 (3.0)	20 (3.0)	11 (6.5)	3 (1.1)	4 (3.2)
Invasion of adjacent structures					
Intrahepatic	316 (25.4)	194 (28.8)	43 (25.4)	48 (17.2)	31 (25.0)
Extrahepatic	33 (2.6)	23 (3.4)	3 (1.8)	3 (1.1)	4 (3.2)
Intra- and extrahepatic	77 (6.2)	45 (6.7)	10 (5.9)	12 (4.3)	10 (8.1)
None	820 (65.8)	412 (61.1)	113 (66.9)	216 (77.4)	79 (63.7)
Enlarged peri-tumoral lymph nodes					
Absent	958 (76.9)	517 (76.7)	130 (76.9)	216 (77.4)	95 (76.6)
Present	288 (23.1)	157 (23.3)	39 (23.1)	63 (22.6)	29 (23.4)
Density (plain CT)					
Isodense	366 (29.4)	198 (29.4)	44 (26.0)	92 (33.0)	32 (25.8)
Hyperdense	723 (58.0)	383 (56.8)	111 (65.7)	154 (55.2)	75 (60.5)
Hypodense	157 (12.6)	93 (13.8)	14 (8.3)	33 (11.8)	17 (13.7)
Tumor enhancement pattern (CE-CT)					
Markedly enhanced	759 (60.9)	426 (63.2)	113 (66.9)	142 (50.9)	78 (62.9)
Mildly enhanced	452 (36.3)	242 (35.9)	55 (32.5)	109 (39.1)	46 (37.1)
No enhancement	35 (2.8)	6 (0.9)	1 (0.6)	28 (10.0)	0 (0.0)
T2WI signal					
Isointense	189 (15.2)	110 (16.3)	12 (7.1)	57 (20.4)	10 (8.1)
Hyperintense	1013 (81.3)	543 (80.6)	147 (87.0)	213 (76.3)	110 (88.7)
Hypointense	44 (3.5)	21 (3.1)	10 (5.9)	9 (3.2)	4 (3.2)
DWI signal					
Isointense	1008 (80.9)	562 (83.4)	147 (87.0)	196 (70.3)	103 (83.1)
Hyperintense	238 (19.1)	112 (16.6)	22 (13.0)	83 (29.7)	21 (16.9)
Hypointense	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Tumor enhancement pattern (CE-MRI)					
Markedly enhanced	985 (79.1)	560 (83.1)	145 (85.8)	174 (62.4)	106 (85.5)
Mildly enhanced	241 (19.3)	109 (16.2)	24 (14.2)	90 (32.3)	18 (14.5)
No enhancement	20 (1.6)	5 (0.7)	0 (0.0)	15 (5.4)	0 (0.0)
Laboratory variables					
White Blood Cell ($\times 10^9$/L)[a]	6.30 [5.00, 8.10]	6.10 [5.00, 7.88]	6.40 [5.21, 8.42]	6.30 [4.80, 8.12]	6.83 [5.64, 9.77]
Neutrophil ($\times 10^9$/L)[a]	4.00 [3.00, 5.66]	3.90 [3.00, 5.40]	4.15 [3.16, 6.11]	4.10 [2.89, 5.88]	4.50 [3.46, 7.14]
Lymphocyte ($\times 10^9$/L)[a]	1.46 [1.10, 1.90]	1.53 [1.14, 1.91]	1.51 [1.14, 1.92]	1.30 [0.96, 1.73]	1.44 [1.09, 1.90]
Monocyte ($\times 10^9$/L)[a]	0.41 [0.31, 0.56]	0.40 [0.31, 0.54]	0.43 [0.33, 0.57]	0.40 [0.30, 0.60]	0.43 [0.35, 0.61]
Eosinophil ($\times 10^9$/L)[a]	0.09 [0.05, 0.17]	0.10 [0.06, 0.19]	0.09 [0.04, 0.17]	0.08 [0.03, 0.14]	0.08 [0.03, 0.15]

Table 1 (continued) | The baseline characteristics of the training, internal validation, external testing, and real-world cohorts

Variables	Overall (*n* = 1246)	The training cohort (*n* = 674)	The internal validation cohort (*n* = 169)	The external testing cohort (*n* = 279)	Real-world cohort (*n* = 124)
Basophil (×10^9/L)[a]	0.02 [0.01, 0.04]	0.02 [0.02, 0.04]	0.03 [0.02, 0.04]	0.02 [0.01, 0.03]	0.02 [0.01, 0.03]
Red Blood Cell (×10^{12}/L)[a]	4.29 [3.94, 4.70]	4.33 [4.00, 4.73]	4.27 [3.92, 4.77]	4.21 [3.90, 4.56]	4.31 [3.92, 4.66]
Hemoglobin (g/L)[a]	130.00 [118.00, 140.00]	130.00 [120.00, 141.00]	130.00 [119.00, 142.00]	127.00 [117.00, 138.00]	130.00 [115.00, 138.25]
PLT (×10^9/L)[a]	218.00 [177.00, 275.75]	221.00 [180.00, 284.50]	222.00 [189.00, 281.00]	200.00 [157.00, 242.50]	225.50 [182.50, 292.00]
INR[a]	0.99 [0.94, 1.05]	0.98 [0.94, 1.04]	1.01 [0.96, 1.07]	1.00 [0.95, 1.07]	0.99 [0.96, 1.04]
Fibrinogen (g/L)[a]	3.51 [2.85, 4.61]	3.49 [2.77, 4.63]	3.52 [2.97, 4.64]	3.49 [2.92, 4.50]	3.55 [2.95, 4.75]
Activated Partial Thromboplastin Time (S)[a]	27.90 [25.70, 30.50]	27.20 [25.10, 29.60]	28.10 [26.40, 29.80]	29.60 [25.90, 33.95]	28.55 [27.10, 30.50]
Thrombin Time (S)[a]	17.50 [16.50, 18.40]	17.70 [16.90, 18.50]	17.00 [16.40, 17.70]	17.90 [16.05, 19.65]	16.65 [16.10, 17.40]
Prothrombin Time (S)[a]	11.60 [11.00, 12.40]	11.50 [10.90, 12.10]	11.70 [11.10, 12.50]	11.90 [11.10, 12.80]	11.60 [11.20, 12.10]
D-Dimer[a]	580.00 [281.00, 1279.00]	566.00 [269.25, 1190.00]	677.00 [303.00, 1403.00]	580.00 [290.00, 1385.00]	606.00 [313.50, 1400.50]
Total Protein (g/L)[a]	70.45 [64.60, 74.80]	70.05 [64.82, 74.57]	71.80 [67.50, 74.80]	69.00 [62.30, 74.50]	72.60 [65.90, 76.53]
Albumin (g/L)[a]	41.70 [37.70, 45.00]	42.20 [38.10, 45.48]	43.20 [39.50, 45.80]	39.30 [35.30, 42.60]	42.60 [37.85, 46.92]
Globulin (g/L)[a]	28.00 [24.80, 31.37]	27.40 [24.30, 30.70]	28.00 [24.90, 31.30]	29.20 [26.40, 32.95]	28.65 [25.45, 31.52]
ALT (U/L)[a]	23.00 [14.00, 54.00]	21.00 [14.00, 46.00]	25.00 [14.00, 78.00]	24.30 [14.00, 68.85]	24.50 [14.75, 55.00]
AST (U/L)[a]	22.00 [17.00, 38.00]	21.00 [17.00, 33.00]	23.00 [18.00, 43.00]	22.60 [18.30, 53.55]	24.00 [17.75, 39.00]
ALP (U/L)[a]	99.00 [75.25, 145.00]	97.00 [74.00, 145.00]	102.00 [80.00, 144.00]	101.00 [76.00, 157.10]	95.00 [79.75, 126.25]
Cholinesterase (U/L)[a]	7399.50 [6018.00, 8830.50]	7327.50 [5987.00, 8617.00]	7547.00 [6385.00, 8837.00]	7425.00 [6286.50, 9240.50]	7442.00 [5748.75, 8895.00]
Total Bile Acid (μmol/L)[a]	6.00 [4.00, 10.39]	6.00 [4.00, 10.50]	5.60 [4.00, 8.90]	6.08 [3.40, 12.71]	5.60 [3.90, 8.67]
Total Bilirubin (μmol/L)[a]	12.60 [9.00, 20.90]	12.00 [8.60, 20.00]	12.10 [8.70, 18.30]	15.30 [10.90, 25.35]	12.65 [8.88, 17.92]
Direct Bilirubin (μmol/L)[a]	4.90 [3.50, 8.88]	4.50 [3.30, 8.30]	5.00 [3.50, 8.30]	5.50 [3.80, 10.00]	5.45 [3.68, 8.22]
Indirect Bilirubin (μmol/L)[a]	7.40 [5.00, 10.97]	7.00 [4.90, 10.38]	6.20 [4.60, 9.10]	9.20 [6.70, 14.05]	7.05 [4.70, 9.83]
Gamma-Glutamyl Transferase (U/L)[a]	44.00 [21.70, 148.75]	46.00 [22.00, 149.00]	43.00 [21.00, 130.00]	48.00 [21.25, 206.15]	37.50 [21.00, 104.25]
Glomerular Filtration Rate[a]	91.14 [80.26, 99.92]	92.01 [82.06, 100.74]	90.60 [80.70, 98.20]	89.59 [74.26, 99.00]	89.75 [80.14, 97.80]
Creatinine (μmol/L)[a]	67.00 [57.00, 79.00]	67.00 [57.00, 78.00]	67.00 [56.00, 80.00]	67.70 [58.90, 81.00]	67.50 [56.00, 81.25]
Urea (mmol/L)[a]	4.60 [3.70, 5.84]	4.67 [3.79, 5.74]	4.56 [3.82, 6.00]	4.45 [3.48, 5.97]	4.72 [3.75, 6.02]
Uric Acid (μmol/L)[a]	279.00 [220.00, 344.60]	281.00 [223.00, 349.00]	278.00 [209.00, 337.00]	274.50 [214.60, 339.00]	282.00 [212.00, 346.00]
Triglycerides (mmol/L)[a]	1.26 [0.92, 1.79]	1.31 [0.95, 1.91]	1.22 [0.98, 1.69]	1.21 [0.88, 1.69]	1.22 [0.86, 1.75]
Total Cholesterol (mmol/L)[a]	4.26 [3.61, 4.96]	4.17 [3.53, 4.87]	4.24 [3.66, 5.04]	4.47 [3.80, 5.20]	4.21 [3.57, 4.82]
High-Density Lipoprotein (mmol/L)[a]	1.11 [0.85, 1.43]	1.03 [0.80, 1.27]	1.10 [0.85, 1.38]	1.59 [1.00, 2.93]	1.06 [0.86, 1.38]
Low-Density Lipoprotein (mmol/L)[a]	2.41 [1.86, 3.18]	2.31 [1.81, 2.86]	2.28 [1.78, 2.89]	3.62 [2.37, 6.48]	2.24 [1.71, 2.76]
Fasting Blood Glucose (mmol/L)[a]	5.31 [4.75, 6.21]	5.17 [4.64, 5.98]	5.48 [4.97, 6.37]	5.38 [4.78, 6.85]	5.54 [5.05, 6.39]
AFP (ng/mL)[a]	2.60 [1.90, 3.56]	2.70 [1.90, 3.60]	2.70 [2.10, 3.60]	2.45 [1.85, 3.34]	2.75 [1.87, 3.73]
CEA (ng/mL)[a]	2.50 [1.60, 3.93]	2.40 [1.60, 3.80]	2.50 [1.70, 4.50]	2.70 [1.73, 3.94]	2.35 [1.78, 3.60]
CA125 (ng/mL)[a]	16.90 [10.80, 30.30]	17.05 [10.93, 30.48]	18.10 [11.60, 38.10]	16.00 [10.10, 25.95]	16.95 [11.57, 30.55]
CA199 (ng/mL)[a]	22.04 [6.50, 160.17]	22.30 [6.30, 152.40]	14.90 [6.80, 114.80]	24.76 [5.92, 209.66]	31.85 [7.25, 188.00]
Ferroprotein (ng/mL)[a]	263.30 [134.25, 478.50]	252.70 [121.45, 461.75]	302.00 [150.70, 522.70]	257.70 [140.50, 472.15]	297.55 [143.17, 570.60]
NLR[a]	2.71 [1.82, 4.42]	2.55 [1.74, 3.89]	2.78 [1.90, 4.00]	3.00 [1.92, 5.83]	3.17 [2.04, 6.33]
dNLR[a]	1.90 [1.38, 2.89]	1.82 [1.35, 2.65]	2.02 [1.43, 2.81]	2.00 [1.41, 3.53]	2.07 [1.46, 3.57]
PLR[a]	149.73 [113.34, 205.62]	149.11 [112.37, 200.65]	150.94 [119.88, 211.43]	148.67 [108.82, 205.53]	154.01 [117.47, 234.58]
LMR[a]	3.65 [2.38, 5.03]	3.93 [2.59, 5.39]	3.64 [2.58, 4.88]	3.02 [2.00, 4.63]	3.51 [1.95, 4.80]
AFR[a]	11.72 [8.60, 15.41]	12.04 [8.70, 15.99]	11.66 [8.83, 15.13]	11.05 [8.24, 13.89]	12.42 [8.62, 15.29]
PFR[a]	61.71 [46.64, 79.31]	63.96 [50.00, 81.91]	61.68 [47.56, 78.81]	54.41 [42.15, 70.81]	64.90 [44.82, 83.56]
CA199-to-TB Ratio[a]	1.63 [0.47, 8.77]	1.68 [0.47, 7.97]	1.38 [0.52, 10.04]	1.60 [0.45, 8.78]	1.91 [0.49, 12.87]
SII[a]	582.47 [374.47, 1102.63]	564.83 [361.93, 960.26]	669.20 [402.65, 1100.90]	532.63 [355.37, 1188.44]	721.63 [417.22, 1494.23]
PNI[a]	49.50 [44.30, 53.90]	50.17 [44.91, 54.30]	51.10 [46.60, 54.50]	46.40 [41.12, 50.70]	50.60 [44.45, 55.11]
CNS[a]	1.00 [0.00, 2.00]	1.00 [0.00, 2.00]	1.00 [0.00, 2.00]	1.00 [0.00, 2.00]	1.00 [0.00, 2.00]

Unless otherwise noted, values are numbers of lesions, with percentages in parentheses.

XGC Xanthogranulomatous cholecystitis, *GBC* gallbladder cancer, *CECT* contrast-enhanced computerized tomography, *MRI* magnetic resonance imaging, *T2WI* T2-weighted imaging, *DWI* diffusion-weighted imaging, *PLT* platelet count, *INR* international normalized ratio, *ALT* alanine aminotransferase, *AST* aspartate aminotransferase, *ALP* alkaline phosphatase, *AFP* alpha-fetoprotein, *CEA* carcinoembryonic antigen, *CA125* cancer antigen 125, *CA199* cancer antigen 19-9, *NLR* neutrophil-to-lymphocyte ratio, *dNLR* derived neutrophil-to-lymphocyte ratio, *PLR* platelet-to-lymphocyte ratio, *LMR* lymphocyte-to-monocyte ratio, *AFR* albumin-to-fibrinogen ratio, *PFR* platelet-to-fibrinogen ratio, *CA199-to-TB Ratio* CA199 to total bilirubin ratio, *SII* systemic immune-inflammation index, *PNI* prognostic nutritional index, *CNS* controlling nutritional status score.

[a] Data are median, with interquartile range [P25, P75] in parentheses.

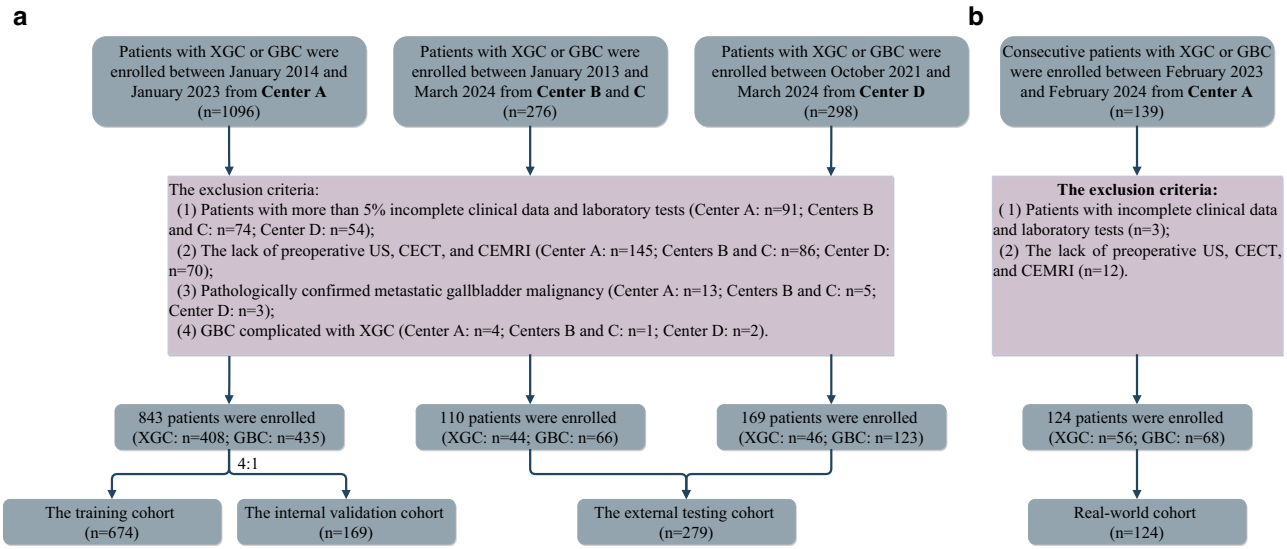


Fig. 2 | Flowchart of the study population. a The training, internal validation, and external testing cohort; **b** The real-world cohort. XGC Xanthogranulomatous cholecystitis, GBC gallbladder cancer, US ultrasound, CECT contrast-enhanced computerized tomography, CEMRI contrast-enhanced magnetic resonance imaging.

continuous mucosal line, elevated fibrinogen level, and higher total bilirubin level. In contrast, independent indicators for GBC included fever, smoking, other conditions (such as schistosomiasis or congenital biliary dilation/cyst), biliary duct dilation, intraluminal tumors, invasion of adjacent structures, enlarged peri-tumoral lymph nodes, hyperdense findings on CT, increased indirect bilirubin, higher CEA levels, and a higher CA199-to-total bilirubin (TB) ratio.

To further refine and determine the optimal number of features, LASSO analysis was employed for an in-depth selection of the 20 independent variables (Supplementary Fig. 1). This analysis ultimately selected 12 key variables for the construction of ML models, which included sex, other conditions, ultrasound echo, gallbladder stones, biliary duct dilation, gallbladder morphology, intramural nodules, intraluminal tumor, mucosal line, enlarged peri-tumoral lymph nodes, fibrinogen level, and indirect bilirubin level. Multicollinearity analysis confirmed that all 12 variables had VIF values below 1.50, indicating no significant collinearity issues among them (Supplementary Table 5).

Diagnostic performance of ML-based models

Using the 12 selected variables, we constructed six ML-based models, including LR, RF, SVM, XGB, LGB, and MLP. In the training, internal validation, and external testing cohorts, the AUCs ranged from 0.98 to 1.00 (95% CI: 0.97–1.00), from 0.92 to 0.94 (95% CI: 0.90–0.98), and from 0.86 to 0.88 (95% CI: 0.81–0.92), respectively (Fig. 3a–c). Notably, the LGB model consistently achieved the highest AUC values in both the internal validation and external testing cohorts, outperforming the other ML-based models. We also compared the differences in AUCs among the six models in the training and internal validation cohorts (Supplementary Table 6), as well as in different external testing cohorts (Supplementary Fig. 2). Additionally, Fig. 3 and Table 2 summarize the diagnostic metrics for each model across the cohorts, including AUC, accuracy, sensitivity, specificity, PPV, NPV, and recall. LIDGAX achieved an AUC of 0.88 (95% CI: 0.84–0.93), accuracy of 0.80 (95% CI: 0.74–0.84), sensitivity of 0.79 (95% CI: 0.73–0.85), and specificity of 0.80 (95% CI: 0.70–0.88) in the external testing cohort. Supplementary Fig. 3 presents the confusion matrices of all models. Calibration curves demonstrated that all models showed good alignment between predicted and observed probabilities for differentiating XGC and GBC in each cohort (Fig. 3d–f). The DCA illustrated the net benefit of clinical utility in the six ML-based models across the three cohorts (Fig. 3g–i). These results strongly suggested that the LGB model outperformed the other five models in various performance parameters. Additionally, the performance of LIDGAX was assessed using time-stratified five-fold cross-validation, demonstrating robust predictive capabilities across all folds (AUC: 0.97–0.98 and 0.94–0.98 in the training and internal validation cohorts; Supplementary Table 7 and Supplementary Fig. 4). This temporal validation strategy effectively simulates real-world clinical deployment scenarios where model performance must remain stable despite temporal shifts in patient characteristics.

Three thresholding strategies demonstrated distinct performance trade-offs across cohorts (Supplementary Table 8). In the external testing cohort, the Youden Index balanced sensitivity (0.79) and specificity (0.80). Maximizing sensitivity achieved near-perfect GBC detection (0.97) but caused significant specificity drops (0.46), increasing false positives. Conversely, maximizing specificity minimized overtreatment risks (0.96) but sacrificed sensitivity (0.42), raising missed diagnosis concerns. Confusion matrices further revealed classification accuracy of three strategies (Supplementary Fig. 5).

Interpretability of LIDGAX model

To enhance the explainability of LIDGAX, the SHAP explainer was utilized to interpret the diagnostic importance of features in the optimal LGB model for distinguishing XGC from GBC. The SHAP beeswarm plot (Fig. 4a) visualizes the 12 key variables, showing each variable's contribution to model predictions. Variables were ranked by importance using average SHAP values and are displayed in descending order (Fig. 4b). SHAP values greater than zero correspond to predictions for the positive class, indicating a higher risk of GBC. For instance, features such as intraluminal tumors or enlarged peri-tumoral lymph nodes were associated with positive SHAP values, which drive predictions toward the "GBC" class. Additionally, Fig. 4c illustrates a case aligned with the "XGC" class, while Fig. 4d represents a case aligned with the "GBC" class according to LIDGAX predictions, with actual variable measurements displayed in each force plot.

Subgroup analyses

We compared the performance of four models built with clinical, imaging, laboratory, and combined variables, respectively (Supplementary Fig. 6). Ultimately, the clinical model was constructed using MLP, the imaging model using MLP, the laboratory model using SVM, and the combined model using LGB (Supplementary Fig. 7). In the external testing cohort, the combined model outperformed the clinical model (0.68 vs. 0.88, adjusted $P < 0.0001$), the imaging model (0.88 vs. 0.88, adjusted $P = 0.699$), and the laboratory model (0.62 vs. 0.88, adjusted $P < 0.0001$). Furthermore, Supplementary Table 9 provides an overview of the AUC, accuracy, sensitivity, specificity, PPV, NPV, and recall of the four models across these cohorts.

Table 2 | Comparison of the diagnostic performance of six ML-based models for differentiating XGC and GBC across the training, internal validation, and external testing cohorts

Cohorts	Models	AUC	Unadjusted *P* (adjusted *P**)	Unadjusted *P* (adjusted *P**)	Unadjusted P (adjusted *P**)	Unadjusted *P* (adjusted *P**)	Unadjusted *P* (adjusted *P**)	Accuracy	Sensitivity	Specificity	PPV	NPV	Recall
The training cohort	LR model	0.99 (0.98–0.99)	Ref.					0.95 (0.93–0.96)	0.97 (0.94–0.98)	0.93 (0.89–0.95)	0.93 (0.90–0.96)	0.96 (0.94–0.98)	0.97 (0.94–0.98)
	RF model	1.00 (1.00–1.00)	<0.0001 (<0.0001)	Ref.				0.99 (0.98–1.00)	0.99 (0.97–1.00)	0.99 (0.98–1.00)	0.99 (0.98–1.00)	0.98 (0.96–1.00)	0.99 (0.97–1.00)
	SVM model	0.98 (0.97–0.99)	0.023 (0.029)	<0.0001 (<0.0001)	Ref.			0.95 (0.93–0.96)	0.96 (0.94–0.98)	0.93 (0.89–0.95)	0.93 (0.90–0.96)	0.96 (0.93–0.98)	0.96 (0.94–0.98)
	XGB model	0.99 (0.98–1.00)	0.004 (0.008)	0.006 (0.009)	<0.0001 (<0.0001)	Ref.		0.97 (0.95–0.98)	0.97 (0.94–0.98)	0.96 (0.94–0.98)	0.97 (0.94–0.98)	0.97 (0.94–0.98)	0.97 (0.94–0.98)
	LGB model	0.98 (0.97–0.99)	0.005 (0.008)	<0.0001 (<0.0001)	0.767 (0.767)	<0.0001 (0.001)	Ref.	0.94 (0.92–0.96)	0.95 (0.92–0.97)	0.93 (0.90–0.95)	0.93 (0.90–0.96)	0.94 (0.91–0.97)	0.95 (0.92–0.97)
	MLP model	0.98 (0.97–0.99)	0.010 (0.014)	<0.0001 (<0.0001)	0.223 (0.257)	0.001 (0.002)	0.280 (0.300)	0.96 (0.94–0.97)	0.96 (0.94–0.98)	0.95 (0.92–0.97)	0.95 (0.92–0.97)	0.96 (0.93–0.98)	0.96 (0.94–0.98)
The internal validation cohort	LR model	0.93 (0.90–0.97)	Ref.					0.86 (0.80–0.91)	0.89 (0.80–0.94)	0.84 (0.74–0.91)	0.86 (0.77–0.92)	0.87 (0.78–0.94)	0.89 (0.80–0.94)
	RF model	0.94 (0.90–0.98)	0.635 (0.902)	Ref.				0.88 (0.82–0.92)	0.86 (0.77–0.93)	0.89 (0.80–0.95)	0.89 (0.81–0.95)	0.86 (0.77–0.92)	0.86 (0.77–0.93)
	SVM model	0.94 (0.91–0.98)	0.206 (0.515)	0.583 (0.902)	Ref.			0.89 (0.83–0.93)	0.90 (0.81–0.95)	0.88 (0.79–0.94)	0.89 (0.80–0.94)	0.89 (0.80–0.95)	0.90 (0.81–0.95)
	XGB model	0.94 (0.91–0.98)	0.264 (0.566)	0.661 (0.902)	0.874 (0.915)	Ref.		0.88 (0.82–0.93)	0.84 (0.74–0.91)	0.93 (0.85–0.97)	0.92 (0.84–0.97)	0.84 (0.75–0.91)	0.84 (0.74–0.91)
	LGB model	0.94 (0.91–0.98)	0.416 (0.781)	0.773 (0.915)	0.830 (0.915)	0.915 (0.915)	Ref.	0.88 (0.82–0.93)	0.86 (0.77–0.93)	0.90 (0.82–0.96)	0.90 (0.82–0.96)	0.86 (0.77–0.93)	0.86 (0.77–0.93)
	MLP model	0.92 (0.87–0.96)	0.151 (0.453)	0.106 (0.399)	0.021 (0.314)	0.042 (0.316)	0.077 (0.386)	0.86 (0.80–0.91)	0.83 (0.73–0.90)	0.90 (0.82–0.96)	0.90 (0.81–0.96)	0.83 (0.74–0.90)	0.83 (0.73–0.90)
The external testing cohort	LR model	0.86 (0.81–0.91)	Ref.					0.81 (0.76–0.85)	0.80 (0.74–0.86)	0.81 (0.71–0.89)	0.90 (0.84–0.94)	0.66 (0.57–0.75)	0.80 (0.74–0.86)
	RF model	0.87 (0.82–0.92)	0.540 (0.824)	Ref.				0.77 (0.71–0.82)	0.74 (0.67–0.80)	0.83 (0.74–0.90)	0.90 (0.84–0.94)	0.60 (0.51–0.69)	0.74 (0.67–0.80)
	SVM model	0.87 (0.82–0.92)	0.716 (0.867)	0.752 (0.867)	Ref.			0.80 (0.74–0.84)	0.78 (0.71–0.83)	0.83 (0.74–0.90)	0.91 (0.85–0.95)	0.64 (0.55–0.73)	0.78 (0.71–0.83)
	XGB model	0.87 (0.83–0.92)	0.306 (0.824)	0.550 (0.824)	0.459 (0.824)	Ref.		0.75 (0.69–0.80)	0.71 (0.64–0.77)	0.83 (0.74–0.90)	0.90 (0.84–0.94)	0.58 (0.49–0.66)	0.71 (0.64–0.77)
	LGB model	0.88 (0.84–0.93)	0.138 (0.792)	0.187 (0.792)	0.049 (0.731)	0.412 (0.824)	Ref.	0.80 (0.74–0.84)	0.79 (0.73–0.85)	0.80 (0.70–0.88)	0.89 (0.84–0.94)	0.65 (0.55–0.74)	0.79 (0.73–0.85)
	MLP model	0.87 (0.82–0.92)	0.625 (0.852)	0.850 (0.910)	0.933 (0.933)	0.537 (0.824)	0.211 (0.792)	0.80 (0.74–0.84)	0.77 (0.70–0.83)	0.86 (0.77–0.92)	0.92 (0.86–0.96)	0.64 (0.54–0.72)	0.77 (0.70–0.83)

Values in parentheses represent the 95% confidence intervals.

ML machine learning, *XGC* Xanthogranulomatous cholecystitis, *GBC* gallbladder cancer, *LR* logistic regression, *RF* random forest, *SVM* support vector machine, *XGB* eXtreme gradient boosting, *LGB* light gradient boosting, *MLP* multilayer perceptron, *AUC* area under the curve, *PPV* positive predictive value, *NPV* negative predictive value, *Ref* reference.

*Indicates statistical significance after Benjamini–Hochberg false discovery rate correction ($q < 0.05$).

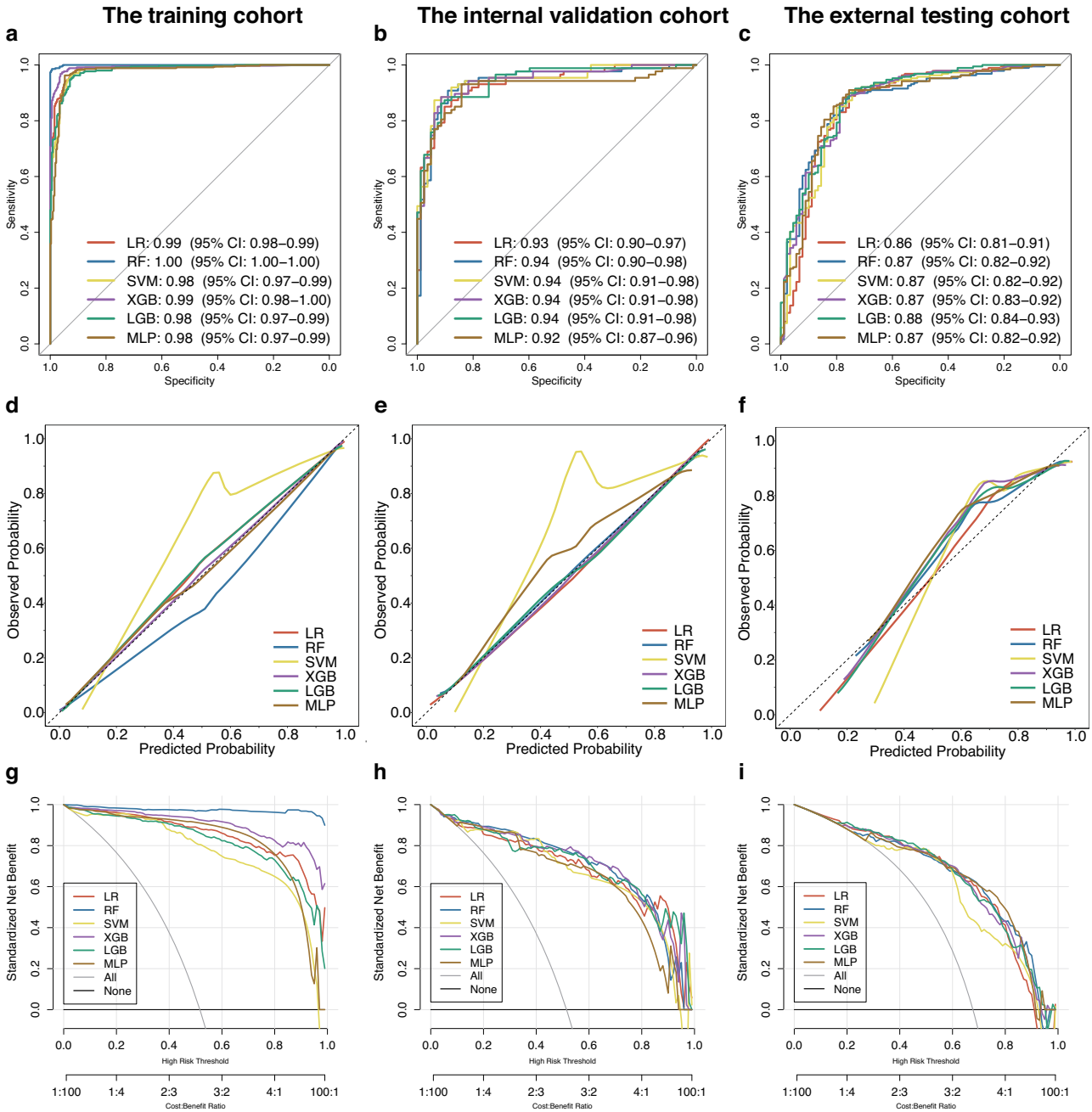


Fig. 3 | Comparison of the diagnostic performance of six ML-based models for differentiating XGC and GBC across the training, internal validation, and external testing cohorts. a–c ROC curves of each model across the three cohorts; **d–f** Calibration curves showing predicted vs. observed probabilities for each model; **g–i** DCA curves indicating net benefit and clinical utility for each model across the three cohorts. ML machine learning, XGC Xanthogranulomatous cholecystitis, GBC gallbladder cancer, LR logistic regression, RF random forest, SVM support vector machine, XGB eXtreme gradient boosting, LGB light gradient boosting, MLP multilayer perceptron, ROC receiver operating characteristic, DCA decision curve analysis, CI confidence interval.

Supplementary Figs. 8–10 present the confusion matrices of six ML models based on clinical variables for differentiating XGC and GBC across all cohorts. The calibration and DCA curves for these models demonstrated that the combined model had a satisfactory alignment and net benefit of clinical utility (Supplementary Figs. 11, 12). All results showed that models utilizing combined variables outperformed those using single-variable groups, with imaging variables making the most substantial contribution. Furthermore, the subgroup analyses demonstrated robust performance of LIDGAX, with AUCs consistently ranging from 0.85 to 0.91 across all subgroups ($P = 0.079–0.682$; Supplementary Fig. 13 and Supplementary Table 10), confirming its generalizability despite demographic, temporal, and institutional variations.

Reader study

To evaluate the performance of LIDGAX compared to six radiologists (including gallbladder specialists, general radiologists, and radiology residents) in distinguishing XGC from GBC, we measured diagnostic accuracy and time efficiency, both with and without LIDGAX assistance (Fig. 5 and Supplementary Table 11). The study involved 169 patients (82 XGC and 87 GBC) from the internal validation cohort. Results showed that all six radiologists performed less accurately than LIDGAX alone in differentiating XGC from GBC (Fig. 5a, d, e), particularly for radiology residents, with significant differences observed in specificity ($P = 0.009–0.010$) and sensitivity ($P = 0.041$). When unassisted, the radiologists demonstrated sensitivity rates

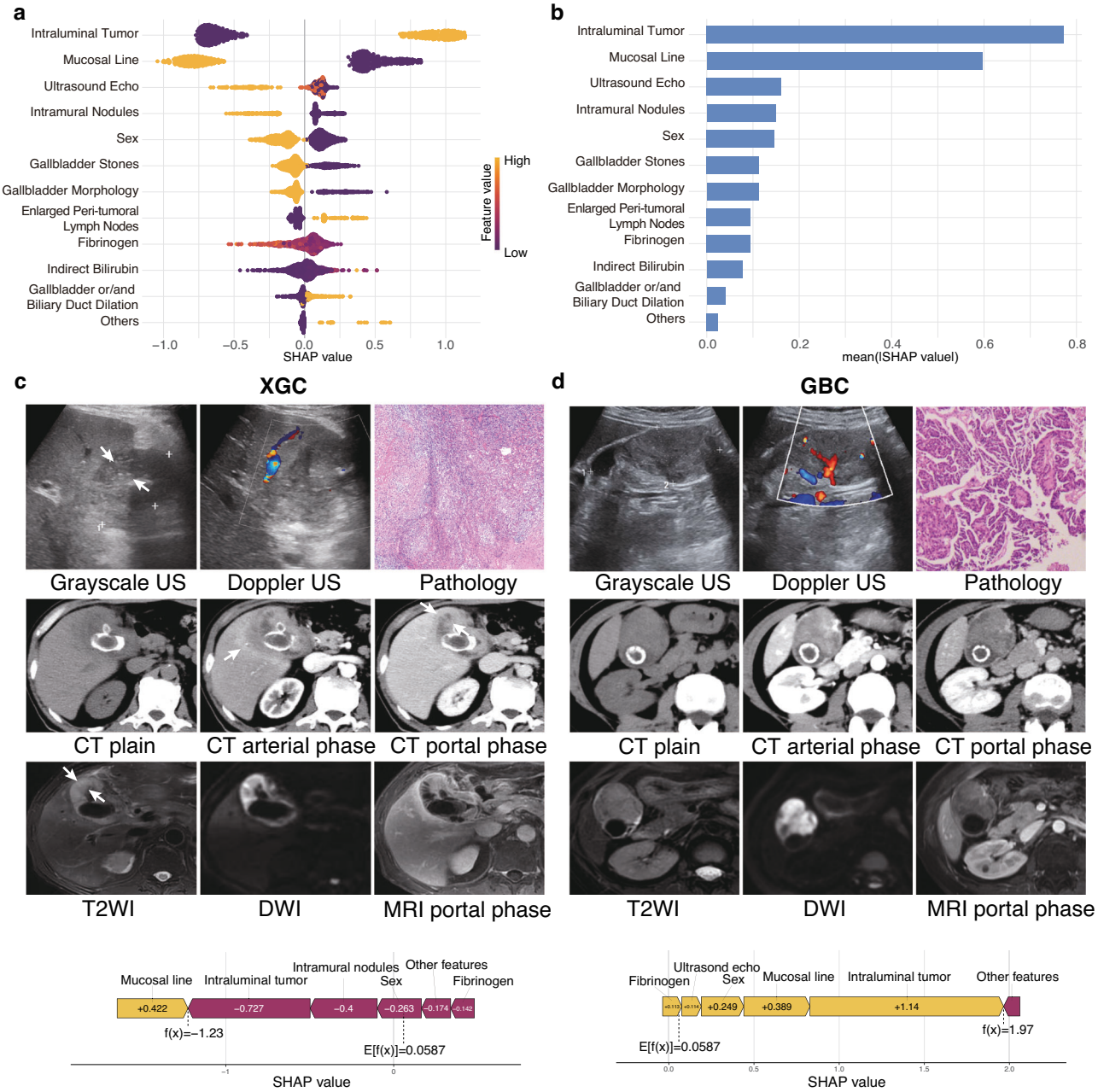


Fig. 4 | Model interpretability using SHAP. a SHAP beeswarm plot illustrates the impact of each variable, where each dot represents a sample with the variable's influence coded by color. **b** SHAP summary bar plot displays variable importance with mean SHAP values on the x-axis, indicating the predictive power of each variable. SHAP force plots with imaging and case details for two individual patients, one with XGC and one with GBC: **c** A 62-year-old male with no additional symptoms. Imaging findings included hypoechoic ultrasound, gallbladder stones, no biliary duct dilation, regular gallbladder morphology, presence of intramural nodules, absence of intraluminal tumor, continuous mucosal line, no enlarged peri-tumoral lymph nodes, a fibrinogen level of 5.27 g/L, and indirect bilirubin at 16.4 μmol/L. LIDGAX predicted XGC, confirmed by post-cholecystectomy pathology. The white arrow highlights the thickened gallbladder wall. **d** A 51-year-old female with no additional symptoms. Imaging showed hypoechoic ultrasound, gallbladder stones, no biliary duct dilation, irregular gallbladder morphology, absence of intramural nodules, presence of an intraluminal tumor, a discontinuous mucosal line, no enlarged peri-tumoral lymph nodes, fibrinogen at 2.34 g/L, and indirect bilirubin at 4.6 μmol/L. LIDGAX classified this case as GBC, also confirmed by pathology. SHAP SHapley Additive exPlanations, XGC Xanthogranulomatous cholecystitis, GBC gallbladder cancer, US ultrasound, CT computerized tomography, MRI magnetic resonance imaging, LIDGAX LightGBM Intelligent Differentiator for XGC and GBC, T2WI T2-weighted imaging, DWI diffusion-weighted imaging.

between 74.4% and 85.4%, which improved to 82.8–89.0% when assisted by LIDGAX (Fig. 5b, f). Similarly, specificity increased from 78.2–86.2% to 82.8–87.4% (Fig. 5b, f). Balanced accuracy rose from 76.3–85.2% unassisted to 82.8–87.6% with LIDGAX assistance, though it remained slightly lower than LIDGAX's performance of 88.2% (Fig. 5c, f). This improvement was most pronounced for radiology residents compared to gallbladder specialists and general radiologists, although not statistically significant ($P = 0.121–0.606$). Furthermore, the average time per assessment decreased significantly from 68.23–89.36 s without LIDGAX to 37.79–53.60 s with its assistance (all P-values < 0.0001, Fig. 5f and Supplementary Table 11), highlighting LIDGAX's potential to improve both diagnostic accuracy and efficiency in differentiating XGC from GBC.

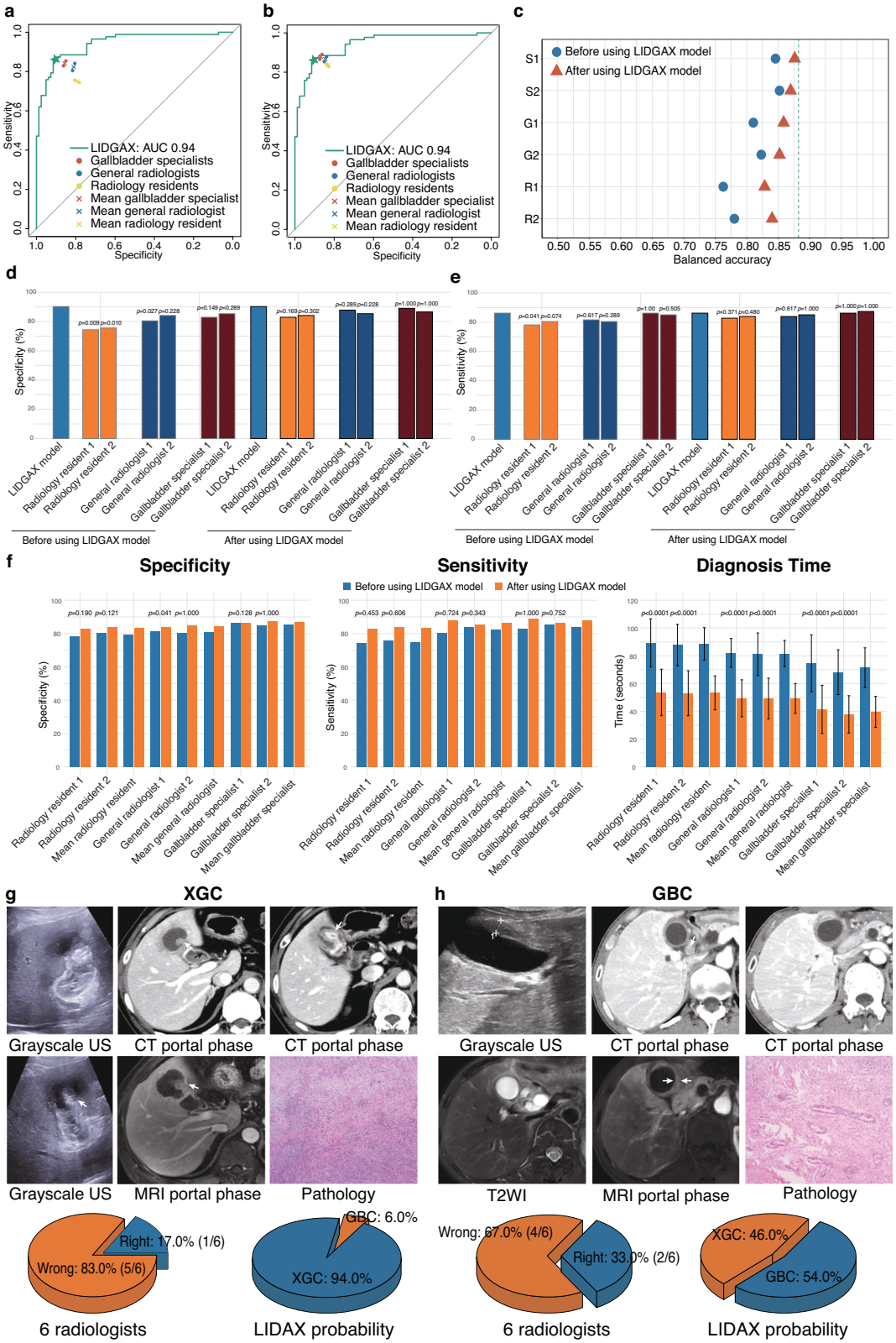


Fig. 5 | Evaluation of LIDGAX versus six radiologists in differentiating XGC and GBC. The evaluation was conducted using 169 patients (82 XGC and 87 GBC) from the internal validation cohort. **a** LIDGAX versus six radiologists without the assistance of LIDGAX. **b** LIDGAX versus six radiologists with the assistance of LIDGAX. **c** Balanced accuracy improvement among radiologists with different levels of expertise for differentiating XGC and GBC. **d** Specificity comparison between LIDGAX and radiologists. **e** Sensitivity comparison between LIDGAX and radiologists. **f** Comprehensive comparison of specificity, sensitivity, and diagnostic time efficiency for radiologists without versus with LIDGAX assistance. **g** A case of XGC misdiagnosed by radiologists but correctly identified as XGC by LIDGAX. **h** A case of GBC misdiagnosed by radiologists but correctly identified as GBC by LIDGAX. XGC Xanthogranulomatous cholecystitis, GBC gallbladder cancer, LIDGAX LightGBM Intelligent Differentiator for XGC and GBC, US ultrasound, CT computerized tomography, MRI magnetic resonance imaging, T2WI T2-weighted imaging.

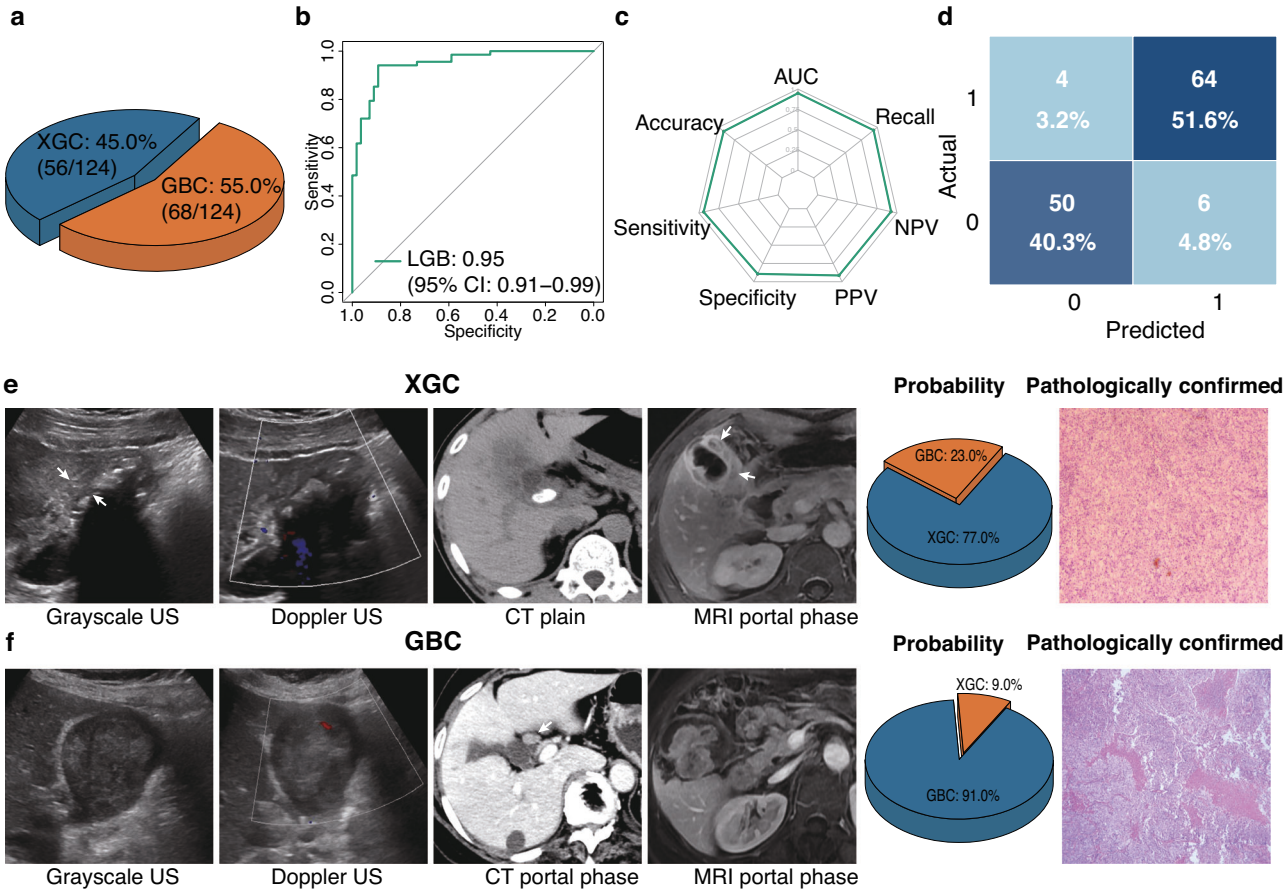


Fig. 6 | Evaluation of the real-world cohort. a Proportions of XGC and GBC cases within the cohort ($n = 124$). **b** ROC curve illustrating LIDGAX's diagnostic performance in differentiating XGC from GBC. **c** Radar chart quantitatively comparing LIDGAX's performance metrics. **d** Confusion matrix of IDGAX in the real-world cohort. **e** An XGC case classified by LIDGAX with a 77.0% probability, confirmed by pathology. **f** A GBC case classified by LIDGAX with a 91.0% probability, also confirmed by pathology. XGC Xanthogranulomatous cholecystitis, GBC gallbladder cancer, LIDGAX LightGBM Intelligent Differentiator for XGC and GBC, AUC area under the curve, PPV positive predictive value, NPV negative predictive value, US ultrasound, CT computerized tomography, MRI magnetic resonance imaging.

Real-world clinical evaluation

To better facilitate clinical translation in real-world settings, we developed an open-source online platform (Supplementary Fig. 14; Version 2.0; https://lidgaxmodel.streamlit.app) based on the LIDGAX model, making it convenient for physicians to use. This retrospective real-world cohort from Center A ultimately comprised 124 individuals, including 56 (45.0%) diagnosed with XGC and 68 (55.0%) with GBC (Fig. 6a). Supplementary Table 12 provides baseline characteristics for XGC and GBC patients in this cohort. LIDGAX achieved an AUC of 0.95 (95% CI: 0.91–0.99), an accuracy of 0.92 (95% CI: 0.86–0.96), a sensitivity of 0.94 (95% CI: 0.86–0.98), a specificity of 0.89 (95% CI: 0.78–0.96), a PPV of 0.91 (95% CI: 0.82–0.97), an NPV of 0.93 (95% CI: 0.82–0.98), and a recall of 0.94 (95% CI: 0.86–0.98) (Fig. 6).

Discussion

In our multicenter real-world study, we present LIDGAX, an advanced ML-based model developed to differentiate between XGC and GBC using clinical, imaging, and laboratory variables. Accurate differentiation between these conditions remains a significant challenge for hepatobiliary surgeons and radiologists, often leading to misdiagnoses and unnecessary healthcare resource usage. By curating a large dataset of pathologically confirmed XGC and GBC cases, we collected relevant clinical, imaging, and laboratory data to construct the LIDGAX model, based on the LGB intelligent differentiator for XGC and GBC, utilizing 12 selected variables. LIDGAX demonstrated high sensitivity (0.86 and 0.79) and specificity (0.90 and 0.80) in distinguishing XGC from GBC in both the internal validation and independent external testing cohorts. The subgroup analyses demonstrated its generalizability despite demographic, temporal, and institutional variations. Its diagnostic accuracy significantly outperformed that of radiologists, particularly enhancing precision and efficiency for residents. Additionally, real-world validation via an online platform further underscored its clinical utility potential.

Artificial intelligence has been utilized to analyze high-throughput data, revealing intricate connections between features and leveraging advanced computational techniques for categorization, prediction, and evidence-based decision-making in novel ways[22]. To our knowledge, only five previous studies have investigated ML- or deep learning (DL)-based approaches for differentiating XGC from GBC. Fujita et al.[23] developed a CT-based DL model that attained high predictive accuracy, achieving an AUC of 0.989 with a dataset of 49 patients. Zhou et al. [24] established an ML-based prediction model that achieved an AUC of 0.888 for the preoperative differentiation of XGC and GBC. Zhang et al.[25] developed a DL nomogram integrating CECT scans, reaching an accuracy of 0.89, a precision of 0.92, and an AUC of 0.92 across two affiliated hospitals, used as an external validation cohort. Gupta et al.[26] employed three DL models to differentiate XGC from GBC on US, demonstrating superior accuracy over radiologists. However, these studies were limited by single-center designs and small sample sizes, with findings unvalidated in independent external cohorts. Another study[27] constructed a predictive nomogram based on 436 patients from two centers, incorporating variables such as sex, Murphy's sign, absolute neutrophil count, glutamyl transpeptidase levels, CEA levels, and imaging findings. Our study included the largest dataset to date—1246

patients from four centers—to differentiate between XGC and GBC. To assess generalizability, reliability, and effectiveness, we validated the LIDGAX model on independent external testing cohorts, achieving AUCs of 0.84 and 0.92 in Centers B and C, and Center D, respectively.

To enhance the clinical applicability of our model, we selected commonly relevant factors for diagnosing gallbladder disease, incorporating general clinical data, imaging features from US, CT, and MRI, and laboratory tests (including routine blood tests, biochemical tests, coagulation tests, and tumor markers). Following multivariate logistic regression for variable selection, 20 variables were identified as independently associated with XGC and GBC. Among these, factors such as sex, symptoms, gallbladder stones, biliary duct dilation, gallbladder morphology, gallbladder size, intramural nodules, intraluminal tumor, mucosal line, invasion of adjacent structures, enlarged peri-tumoral lymph nodes, and CEA were consistent with prior studies[23–25,27,28]. Notably, our analysis revealed that fever was significantly associated with GBC. This association may be mechanistically explained by tumor necrosis and systemic inflammatory response, which triggers elevated pro-inflammatory cytokines (e.g., TNF-α, IL-1) and COX-2 expression[29–31]. Additionally, gallstones or GBC-related biliary obstruction may predispose patients to bacterial infections, thereby further contributing to fever[32,33]. In contrast, XGC—a chronic granulomatous inflammatory condition characterized by lipid-laden macrophage infiltration—typically lacks such pronounced systemic inflammatory responses[34]. CT hyperdensity was another feature associated with GBC, likely reflecting desmoplastic stromal reactions with collagen deposition and fibroblast proliferation, whereas XGC typically exhibits hypodense regions from lipid-laden macrophages. Furthermore, we identified new key risk factors for differentiating XGC and GBC—namely smoking, fibrinogen, total bilirubin, indirect bilirubin, and the CA199-to-TB ratio—factors not previously reported in this context. Prior research has shown that preoperative serum fibrinogen and total bilirubin levels correlate with tumor progression and may independently predict GBC[35–38]. However, their potential role in distinguishing XGC from GBC has not been explored until now. Using LASSO to minimize redundancy, we refined these 20 variables to 12 final input parameters for the six ML-based models.

In our study, the inclusion of patients with complete US, CT, and MRI data may introduce selection bias, as this criterion excluded those typically managed with fewer imaging modalities in routine practice. However, the 12 key variables comprising LIDGAX—particularly key imaging features such as gallbladder stones, biliary duct dilation, gallbladder morphology, intramural nodules, intraluminal tumor, mucosal line, and enlarged peri-tumoral lymph nodes—are not modality-specific. These semantic features can be reliably identified across US, CT, or MRI. Consequently, LIDGAX remains applicable even when MRI is unavailable, provided the essential features are assessable through existing modalities. In clinical practice, US is often the first-line modality, followed by CT or MRI if needed for further characterization. For features requiring MRI confirmation (e.g., occult gallstones undetected by US/CT, intramural nodules demonstrating isoechoic density on US and isodense characteristics on CT), MRI is recommended as a supplementary modality to ensure input accuracy. When MRI is inaccessible, radiologists should flag such cases for multidisciplinary review. This protocol balances diagnostic accuracy with resource constraints. The retrospective requirement for multimodal imaging aimed to minimize feature omission during model development. While potentially introducing selection bias, this strategy ensured comprehensive data collection. In the future, prospective validations will specifically evaluate LIDGAX's performance in settings with restricted imaging protocols. Besides, our study excluded seven GBC complicated with XGC cases, potentially introducing selection bias. Previous studies reported the incidence of XGC-GBC coexistence ranges from 3% to 12.5%[39–41]. For such cases, histopathological examination remains the gold standard for definitive diagnosis. LIDGAX was specifically designed not to replace histopathology but to improve preoperative differentiation between pure XGC and GBC, guiding clinical decision-making and surveillance planning. The exclusion of coexisting cases from model validation ensures alignment with its intended use scenario.

Six ML algorithms we used are widely used in medical diagnostics[16]. Among these, the LGB model (AUC: 0.94 and 0.88) showed superior performance compared to others (AUC: 0.92–0.94 and 0.86–0.87) in both internal validation and external testing cohorts. Multicollinearity can contribute to overfitting; thus, we evaluated it using VIF, with all values under 1.50, indicating no significant multicollinearity among these variables. All six ML-based models exhibited slight overfitting, with significant differences in AUC between the training and internal validation cohorts ($P = 0.002–0.040$). However, the LGB model had the smallest AUC discrepancy and achieved the highest AUC in the independent external testing cohort, highlighting its robustness and generalizability. Calibration and DCA curves further showed that the LGB model offered the best alignment and net clinical benefit. This led us to select it as the optimal model for differentiating XGC from GBC, naming it LIDGAX. Furthermore, the interpretability of LIDGAX is crucial for clinicians and radiologists in decision-making processes. Therefore, we employed SHAP values to enhance interpretability, revealing the underlying relationships between features and outcomes[42]. The SHAP value analysis highlighted that intraluminal tumor and mucosal line contributed the most to distinguishing XGC from GBC.

The choice of thresholding strategy should align with clinical priorities. In high-risk populations, a sensitivity-prioritized threshold is optimal for screening, minimizing missed GBC diagnoses and enabling timely resection. Though this increases false positives, confirmatory biopsies or short-term imaging follow-up can mitigate overdiagnosis risks. Conversely, a specificity-prioritized threshold is critical for surgical decision-making, reducing unnecessary extended hepatectomies, neoadjuvant therapies, and lymph node dissections for benign XGC. For example, misclassifying XGC as GBC could expose patients to toxic chemotherapy or aggressive lymphadenectomy—procedures avoided by LIDGAX's 0.96 specificity. However, its low sensitivity (0.42) necessitates cautious use, particularly in younger patients prioritizing cancer detection. The Youden Index (0.79 sensitivity, 0.80 specificity) balances accuracy and resource allocation, mirroring real-world trade-offs. Future integration of cost-benefit analyses could refine personalized threshold selection.

We also compared LIDGAX with six radiologists of varying experience levels in differentiating XGC from GBC, finding that LIDGAX demonstrated superior diagnostic accuracy. Two main factors contributed to this advantage. First, LIDGAX was trained using a combination of clinical, imaging, and laboratory variables within a supervised learning framework—an approach that provided systematic, digitized data integration across multiple sources, unlike the workflow radiologists typically experience. Second, ML algorithms, such as LIDGAX, are inherently more effective in feature selection and weighting than manual assessments[43], enabling our model to directly learn diagnostic patterns from detailed input data and apply them efficiently. Moreover, LIDGAX's support enhanced radiologists' performance across several metrics: sensitivity improved by 1.2–8.5%, specificity by 0.0–4.6%, and balanced accuracy by 1.8–6.6%. Additionally, the average diagnostic time per patient was reduced by 30.44–35.76 s, indicating that LIDGAX can substantially improve radiologists' accuracy, lower misdiagnosis rates, and optimize time efficiency, particularly in high-demand clinical environments. ML still faces challenges in clinical translation[44]. To support LIDGAX's implementation, we developed an open-source online platform designed for convenient clinical use. This platform enables clinicians to input 12 key factors and instantly receive diagnostic predictions. In a retrospective real-world cohort of 124 patients from Center A, the platform achieved an impressive AUC of 0.95, with an accuracy of 0.92, sensitivity of 0.94, and specificity of 0.89. These results demonstrate that the platform is user-friendly for clinicians and radiologists and achieves robust performance in distinguishing XGC from GBC in real-world clinical settings.

This study has several limitations. First, LIDGAX was developed using data from Chinese populations, so its generalizability to broader, global

populations remains uncertain and requires further validation with additional datasets, despite our dataset being the largest available to date. Second, although LIDGAX was built using the clinical, imaging, and laboratory variables we were able to collect, there may be other relevant variables not considered in this study that could potentially enhance model performance. Future iterations could enhance performance by incorporating emerging biomarkers and genomic data. Third, the retrospective design inherently introduces selection bias, a limitation common to all observational studies. To address this, we are planning prospective randomized controlled trials to validate LIDGAX's clinical efficacy. Lastly, given the rapid advancements in DL for medical imaging, our future goal is to incorporate DL to extract complex, high-dimensional features from multimodal imaging, enabling more accurate and intelligent differentiation between XGC and GBC for clinical applications.

In conclusion, we developed the LIDGAX model, utilizing the LGB algorithm, to accurately differentiate XGC from GBC. The model demonstrated robust diagnostic performance across independent external testing cohorts, surpassing six expert radiologists in diagnostic accuracy. By employing SHAP values, we improved the interpretability of LIDGAX for clinical applications. Additionally, we constructed an open-source online platform to validate the clinical translation potential of LIDGAX. Given its high accuracy and reliability, LIDGAX holds promise as a valuable, non-invasive tool for effectively distinguishing XGC from GBC in clinical settings.

Methods

Patients

This multicenter, retrospective study included patients diagnosed with XGC or GBC between January 2023 and February 2024 from four Chinese hospitals: The First Affiliated Hospital, Zhejiang University School of Medicine (Center A); The Second Affiliated Hospital, Jiaxing University (Center B); Beilun District People's Hospital (Center C); and Huzhou Central Hospital (Center D). Adult participants who underwent either simple or radical cholecystectomy and were pathologically confirmed to have XGC or GBC were included. The exclusion criteria were: (1) patients with more than 5% incomplete clinical data and laboratory tests (Center A: $n = 91$; Centers B and C: $n = 74$; Center D: $n = 54$); (2) the lack of preoperative US, CECT, and CEMRI (Center A: $n = 145$; Centers B and C: $n = 86$; Center D: $n = 70$); (3) pathologically confirmed metastatic gallbladder malignancy (Center A: $n = 13$; Centers B and C: $n = 5$; Center D: $n = 3$); (4) GBC complicated with XGC (Center A: $n = 4$; Centers B and C: $n = 1$; Center D: $n = 2$). Details of the study population are illustrated in Fig. 2a.

In total, 1246 patients were included in the differential diagnosis task, comprising 554 XGC patients and 692 GBC patients. Patients from Center A ($n = 843$) were split chronologically into a training cohort and an internal validation cohort in a 4:1 ratio, while patients from Centers B, C, and D ($n = 279$) were assigned to independent external testing cohorts.

Variable collection

Baseline variables included clinical data, laboratory tests, and imaging features, selected based on consultations with gallbladder specialists and a review of recent literature on risk factors relevant to XGC and GBC. Clinical data were collected from medical records, including sex, age, symptoms, complications, smoking status, diabetes, gallbladder adenomyomatosis, biliary tract infection, as well as other conditions like schistosomiasis and congenital biliary dilatation/cyst. Laboratory tests conducted within two weeks prior to cholecystectomy, including routine blood tests, biochemical analyses, coagulation profiles, and tumor markers, were systematically extracted from electronic medical records. Further details on the clinical data and laboratory tests are provided in Supplementary Table 1. Two abdominal radiologists, each with over 10 years of experience, independently reviewed each US, CECT, and CEMRI scan as a standard reference and assessed all imaging features by consensus. In cases of disagreement, a third radiologist with over 20 years of experience in gallbladder disease performed the final evaluation. All scans conducted within two weeks were used as references. Detailed definitions of imaging features are available in Supplementary Table 2.

Model development

The development of ML-based models followed a two-step approach: (1) selecting robust features related to the differentiation of XGC and GBC from collected clinical, imaging, and laboratory variables; and (2) constructing six ML models using the selected features. For the first step, we implemented a three-stage feature selection approach within the training cohort ($n = 674$) to identify robust clinical, imaging, and laboratory variables. First, a preliminary univariate binary logistic regression analysis was conducted to identify variables with significant differences between XGC and GBC patients. Next, multivariate logistic regression analysis was applied to variables identified in the univariate analysis. Finally, the least absolute shrinkage and selection operator (LASSO) regression method was used to select the most predictive features with non-zero coefficients, with penalty tuning through 10-fold cross-validation (Supplementary Fig. 1).

For the second step, we developed six ML classification algorithms using the features selected by LASSO in the training cohort: logistic regression (LR), random forest (RF), support vector machine (SVM), eXtreme gradient boosting (XGB), light gradient boosting (LGB), and multilayer perceptron (MLP). Each algorithm was fine-tuned through grid search. After identifying optimal hyperparameters, each model was retrained on the full training subset with a set random seed, finalizing the weights and generating a locked model, which was then evaluated on the internal validation cohort.

Model evaluation

To systematically evaluate model performance, we compared the performance of six models across training, internal validation, and external testing cohorts. Evaluation metrics included the area under the curve (AUC), sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), recall, and confusion matrix. Calibration curves and decision curve analysis (DCA) were employed to assess model calibration and clinical utility across all cohorts[45,46]. Based on the integrated performance metrics, the LGB-based algorithm was identified as the optimal model and named LIDGAX (**L**GB **I**ntelligent **D**ifferentiator for **G**BC **a**nd **X**GC). To confirm the robustness of the chronological split strategy, we performed time-stratified five-fold cross-validation on dataset A ($n = 843$) using LIDGAX. The model's decision-making process was visualized using SHapley Additive exPlanations (SHAP)[47,48], which quantified feature importance scores and elucidated the relationships between XGC, GBC, and selected features.

Our study implemented three thresholding strategies to address distinct clinical priorities: (1) Diagnostic balance: Optimized using the Youden Index (sensitivity + specificity − 1) to balance sensitivity and specificity; (2) Screening priority: Maximized sensitivity (>95%) to minimize missed diagnoses in gallbladder cancer (GBC) screening; (3) Treatment precision: Maximized specificity (>95%) to reduce overtreatment risks caused by false positives in therapeutic decision-making.

Subgroup analyses

To evaluate the effectiveness of the combined model, we constructed four separate models using these factors: a clinical model, an imaging model, a laboratory model, and a combined model. Each of these four models was developed using six different ML algorithms. To address potential variability and ensure the robustness of the findings, we conducted subgroup analyses in the external testing cohort, stratified by sex (female and male), age (<60 years and ≥60 years), time periods (2011–2015, 2016–2020, and 2021–2024), and centers (Centers B, C, and D).

Reader study

To assess the performance of radiologists in differentiating XGC from GBC, six radiologists independently diagnosed cases in the internal validation cohort ($n = 169$). Participants included two radiology residents (3–5 years of

experience), two general radiologists (5–10 years of experience in abdominal imaging), and two gallbladder specialists (10–20 years of experience in gallbladder imaging). Prior to the study, a gallbladder specialist with extensive experience (over 3000 case reviews) conducted a training session for each radiologist, covering key imaging features identified in 40 representative cases from the training cohort.

This study included two main steps. In the first step, we compared the diagnostic performance of LIDGAX with that of the radiologists. Each radiologist reviewed anonymized US, CECT, and MRI images in random order using the local picture archiving and communication system, without access to clinical or laboratory data. They were tasked with determining whether each case represented XGC or GBC. In the second step, we assessed LIDGAX's potential to support radiologists in diagnosis. Each radiologist received LIDGAX's probability score for each case and then reanalyzed the same cases from the first step with this additional input. A minimum interval of one month separated the two steps to reduce recall bias.

Real-world study

For the real-world clinical evaluation, we deployed the LIDGAX model on an open-source online computing platform, allowing clinicians and radiologists to easily analyze cases through a user-friendly interface. This retrospective study included consecutive patients diagnosed with XGC or GBC between February 2023 and February 2024 from Center A. Exclusion criteria included: (1) patients with incomplete clinical and laboratory data ($n = 3$), and (2) patients without preoperative US, CECT, and CEMRI scans ($n = 12$). After applying these criteria, 124 patients were included in the real-world evaluation study (Fig. 2b).

Statistical analysis

All statistical analyses were conducted using R software (Version 4.2.2; https://www.rproject.org). To address incomplete clinical and laboratory data, multiple imputation was applied as part of data preprocessing[49]. Categorical variables were compared between groups using either the chi-square test or Fisher's exact test and are presented as numbers and frequencies. For continuous variables, the Kolmogorov–Smirnov test assessed normality. Variables following a normal distribution are expressed as mean ± standard deviation (SD) and were compared using the t-test, while non-normally distributed variables are reported as median (interquartile range, IQR) and analyzed with the Mann–Whitney U test. Non-normally distributed continuous data were normalized before model development. Collinearity among variables was evaluated through the variance inflation factor (VIF), where a VIF > 5 indicated notable collinearity and a VIF > 10 suggested significant collinearity[50]. Univariate and multivariate binary logistic regression analyses identified variables associated with XGC and GBC. Variables found significant in univariate analysis were subsequently included in a stepwise multivariate analysis, using the Akaike information criterion for optimal variable selection[51]. The diagnostic performance of the ML-based models was evaluated by metrics including sensitivity, specificity, accuracy, PPV, NPV, recall, balanced accuracy, and confusion matrix. Model performance comparisons of AUCs between the six algorithms were carried out using the DeLong test. Confidence intervals (95% CIs) were obtained through 1000 bootstrap resampling. To compare the sensitivity and specificity of diagnostic performance before and after LIDGAX assistance, McNemar's test was applied for paired categorical data analysis. Benjamini–Hochberg false discovery rate (BH-FDR, $q < 0.05$) was used for multiple testing correction[52]. Two-tailed P-values < 0.05 were considered statistically significant.

Data availability

The source data for all models, tables, and figures, along with supporting materials, are available from the corresponding author upon reasonable request.

Code availability

The code developed for this study is not publicly accessible to protect proprietary knowledge. However, qualified academic researchers may request access to the code—including preprocessing scripts, model architecture, and training protocols—for noncommercial research purposes. Requests should be submitted to the corresponding author at tiananjiang@zju.edu.cn and must include a detailed research proposal outlining the intended use. Access will be granted following approval by the institutional review committee, which evaluates proposals within 4-6 weeks. A formal code use agreement, prohibiting redistribution or commercial exploitation, will be required prior to release.

Received: 21 February 2025; Accepted: 1 September 2025;
Published online: 01 October 2025

References

1. Rammohan, A., Cherukuri, S. D., Sathyanesan, J., Palaniappan, R. & Govindan, M. Xanthogranulomatous cholecystitis masquerading as gallbladder cancer: can it be diagnosed preoperatively?. *Gastroenterol. Res. Pr.* **2014**, 253645 (2014).
2. Roa, J. C. et al. Gallbladder cancer. *Nat. Rev. Dis. Prim.* **8**, 69 (2022).
3. Baiu, I. & Visser, B. Gallbladder cancer. *Jama* **320**, 1294 (2018).
4. Güneş, Y., Bostancı, Ö, İlbar Tartar, R. & Battal, M. Xanthogranulomatous cholecystitis: is surgery difficult? Is laparoscopic surgery recommended?. *J. Laparoendosc. Adv. Surg. Tech. A* **31**, 36–40 (2021).
5. Benson, A. B. et al. Hepatobiliary Cancers, Version 2.2021, NCCN Clinical Practice Guidelines in Oncology. *J. Natl. Compr. Canc. Netw.* **19**, 541–565 (2021).
6. Feng, L., You, Z., Gou, J., Liao, E. & Chen, L. Xanthogranulomatous cholecystitis: experience in 100 cases. *Ann. Transl. Med.* **8**, 1089 (2020).
7. Truant, S., Chater, C. & Pruvot, F. R. Greatly enlarged thickened gallbladder. Diagnosis: Xanthogranulomatous cholecystitis (XGC). *JAMA Surg.* **150**, 267–268 (2015).
8. Deng, Y. L. et al. Xanthogranulomatous cholecystitis mimicking gallbladder carcinoma: An analysis of 42 cases. *World J. Gastroenterol.* **21**, 12653–12659 (2015).
9. Spinelli, A. et al. Extended surgical resection for xanthogranulomatous cholecystitis mimicking advanced gallbladder carcinoma: a case report and review of literature. *World J. Gastroenterol.* **12**, 2293–2296 (2006).
10. Huang, E. Y. et al. Distinguishing characteristics of xanthogranulomatous cholecystitis and gallbladder adenocarcinoma: a persistent diagnostic dilemma. *Surg. Endosc.* **38**, 348–355 (2024).
11. Xiao, J., Zhou, R., Zhang, B. & Li, B. Noninvasive preoperative differential diagnosis of gallbladder carcinoma and xanthogranulomatous cholecystitis: a retrospective cohort study of 240 patients. *Cancer Med.* **11**, 176–182 (2022).
12. Goshima, S. et al. Xanthogranulomatous cholecystitis: diagnostic performance of CT to differentiate from gallbladder cancer. *Eur. J. Radio.* **74**, e79–e83 (2010).
13. Parra, J. A. et al. Xanthogranulomatous cholecystitis: clinical, sonographic, and CT findings in 26 patients. *AJR Am. J. Roentgenol.* **174**, 979–983 (2000).
14. Ros, P. R. & Goodman, Z. D. Xanthogranulomatous cholecystitis versus gallbladder carcinoma. *Radiology* **203**, 10–12 (1997).
15. Lee, E. S. et al. Xanthogranulomatous cholecystitis: diagnostic performance of US, CT, and MRI for differentiation from gallbladder carcinoma. *Abdom. Imaging* **40**, 2281–2292 (2015).
16. Goecks, J., Jalili, V., Heiser, L. M. & Gray, J. W. How machine learning will transform biomedicine. *Cell* **181**, 92–101 (2020).
17. Zhang, B. et al. Identifying behaviour-related and physiological risk factors for suicide attempts in the UK Biobank. *Nat. Hum. Behav.* **8**, 1784–1797 (2024).
18. Zamanzadeh, D. et al. Data-driven prediction of continuous renal replacement therapy survival. *Nat. Commun.* **15**, 5440 (2024).

19. Liu, R. et al. Development and prospective validation of postoperative pain prediction from preoperative EHR data using attention-based set embeddings. *NPJ Digit Med.* **6**, 209 (2023).
20. Wagner, M. et al. Artificial intelligence for decision support in surgical oncology - a systematic review. *Artif. Intell. Surg.* **2**, 159–172 (2022).
21. Teng, X. et al. Development and validation of an early diagnosis model for bone metastasis in non-small cell lung cancer based on serological characteristics of the bone metastasis mechanism. *EClinicalMedicine* **72**, 102617 (2024).
22. Xu, Y. et al. Artificial intelligence: a powerful paradigm for scientific research. *Innovation* **2**, 100179 (2021).
23. Fujita, H. et al. Differential diagnoses of gallbladder tumors using CT-based deep learning. *Ann. Gastroenterol. Surg.* **6**, 823–832 (2022).
24. Zhou, Q. M. et al. Machine learning-based radiological features and diagnostic predictive model of xanthogranulomatous cholecystitis. *Front. Oncol.* **12**, 792077 (2022).
25. Zhang, W. et al. Deep learning nomogram for preoperative distinction between Xanthogranulomatous cholecystitis and gallbladder carcinoma: a novel approach for surgical decision. *Comput. Biol. Med.* **168**, 107786 (2024).
26. Gupta, P. et al. Deep-learning models for differentiation of xanthogranulomatous cholecystitis and gallbladder cancer on ultrasound. *Indian J. Gastroenterol.* **43**, 805–812 (2024).
27. Fu, T. et al. Machine learning-based diagnostic model for preoperative differentiation between xanthogranulomatous cholecystitis and gallbladder carcinoma: a multicenter retrospective cohort study. *Front. Oncol.* **14**, 1355927 (2024).
28. Ito, R. et al. A scoring system based on computed tomography for the correct diagnosis of xanthogranulomatous cholecystitis. *Acta Radio. Open* **9**, 2058460120918237 (2020).
29. Yang, S. Q. et al. Prognostic significance of tumor necrosis in patients with gallbladder carcinoma undergoing curative-intent resection. *Ann. Surg. Oncol.* **31**, 125–132 (2024).
30. P,rez-Moreno, P., Riquelme, I., Garcja, P., Brebi, P. & Roa, J. C. Environmental and Lifestyle risk factors in the carcinogenesis of gallbladder cancer. *J. Pers. Med.* **12**, 234 (2022).
31. Balkwill, F. Tumour necrosis factor and cancer. *Nat. Rev. Cancer* **9**, 361–371 (2009).
32. Elinav, E. et al. Inflammation-induced cancer: crosstalk between tumours, immune cells and microorganisms. *Nat. Rev. Cancer* **13**, 759–771 (2013).
33. Espinoza, J. A. et al. The inflammatory inception of gallbladder cancer. *Biochim. Biophys. Acta* **1865**, 245–254 (2016).
34. Azari, F. S. et al. Kt. A contemporary analysis of xanthogranulomatous cholecystitis in a Western cohort. *Surgery* **170**, 1317–1324 (2021).
35. Yang, Z. et al. Preoperative serum fibrinogen as a valuable predictor in the nomogram predicting overall survival of postoperative patients with gallbladder cancer. *J. Gastrointest. Oncol.* **12**, 1661–1672 (2021).
36. Zhang, L. et al. Exploring the diagnosis markers for gallbladder cancer based on clinical data. *Front. Med.* **9**, 350–355 (2015).
37. Yang, S. Q. et al. Unraveling early recurrence of risk factors in gallbladder cancer: a systematic review and meta-analysis. *Eur. J. Surg. Oncol.* **50**, 108372 (2024).
38. Liu, F. et al. The prognostic value of combined preoperative PLR and CA19-9 in patients with resectable gallbladder cancer. *Updates Surg.* **76**, 1235–1245 (2024).
39. Bolukbasi, H. & Kara, Y. An important gallbladder pathology mimicking gallbladder carcinoma: xanthogranulomatous cholecystitis: a single tertiary center experience. *Surg. Laparosc. Endosc. Percutan Tech.* **30**, 285–289 (2020).
40. Kwon, A. H. & Sakaida, N. Simultaneous presence of xanthogranulomatous cholecystitis and gallbladder cancer. *J. Gastroenterol.* **42**, 703–704 (2007).
41. Pandey, A., Kumar, D., Masood, S., Chauhan, S. & Kumar, S. Is final histopathological examination the only diagnostic criteria for xanthogranulomatous cholecystitis?. *Niger. J. Surg.* **25**, 177–182 (2019).
42. Nohara, Y., Matsumoto, K., Soejima, H. & Nakashima, N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Comput. Methods Prog. Biomed.***214**, 106584 (2022).
43. Wang, S. & Summers, R. M. Machine learning and radiology. *Med. Image Anal.* **16**, 933–951 (2012).
44. Dinsdale, N. K. et al. Challenges for machine learning in clinical translation of big data imaging studies. *Neuron* **110**, 3866–3881 (2022).
45. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis. Mak.* **26**, 565–574 (2006).
46. Steyerberg, E. W. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 128–138 (2010).
47. Lundberg S. M., Lee S.-I. A unified approach to interpreting model predictions. In *Advances in neural information processing systems* 30, (NeurIPS, 2017).
48. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent individualized feature attribution for tree ensembles. Preprint at https://arxiv.org/abs/1802.03888 (2018).
49. Little, R. J. et al. The prevention and treatment of missing data in clinical trials. *N. Engl. J. Med.* **367**, 1355–1360 (2012).
50. O'brien, R. M. A caution regarding rules of thumb for variance inflation factors. *Qual. Quant.* **41**, 673–690 (2007).
51. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. control* **19**, 716–723 (1974).
52. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B.* **57**, 289–300 (1995).

Acknowledgements

Funding was provided by the Development Project of National Major Scientific Research Instrument (82027803), the National Natural Science Foundation of China (82202151), the Key Research and Development Project of Zhejiang Province (2024C03092), the National Key R&D Program of China (2022YFC2405505), Zhejiang Provincial Natural Science Foundation of China (Y24H180007), and Beilun Health Technology Project (2024BLWSQN002).

Author contributions

Study concept and design: K.Z., J.H., L.X., and T.J.; Acquisition of data: K.Z., J.H., W.J., Q.P., W.X., L.W., and W.S.; Analysis and interpretation of data: K.Z. and J.H.; Drafting of the manuscript: K.Z., J.H., and L.X.; Critical revision of the manuscript: K.Z., J.H., L.X., and T.J.; Statistical analysis: K.Z. and J.H.; Study supervision: K.Z., J.H., W.J., Q.P., W.X., L.W., W.S., L.X., and T.J. All authors have read and approved this manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01991-7.

Correspondence and requests for materials should be addressed to Liting Xie or Tianan Jiang.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025