Received: 13 April 2022 | Accepted: 29 May 2022

DOI: 10.1002/ags3.12589

AGSurg Annals of Gastroenterological Surgery Open Access

WILEY

ORIGINAL ARTICLE

Differential diagnoses of gallbladder tumors using CT-based deep learning

Hiroaki Fujita[1] | Taiichi Wakiya[1] | Keinosuke Ishido[1] | Norihisa Kimura[1] | Hayato Nagase[1] | Taishu Kanda[1] | Masashi Matsuzaka[2] | Yoshihiro Sasaki[2] | Kenichi Hakamada[1]

[1]Department of Gastroenterological Surgery, Hirosaki University Graduate School of Medicine, Hirosaki, Japan

[2]Department of Medical Informatics, Hirosaki University Hospital, Hirosaki, Japan

Correspondence

Taiichi Wakiya, Department of Gastroenterological Surgery, Hirosaki University Graduate School of Medicine, 5 Zaifu-cho, Hirosaki city, Aomori 036-8216, Japan.

Email: wakiya1979@hirosaki-u.ac.jp

Abstract

Background: The differential diagnosis between gallbladder cancer (GBC) and xanthogranulomatous cholecystitis (XGC) remains quite challenging, and can possibly lead to improper surgery. This study aimed to distinguish between XGC and GBC by combining computed tomography (CT) images and deep learning (DL) to maximize the therapeutic success of surgery.

Methods: We collected a dataset, including preoperative CT images, from 28 cases of GBC and 21 XGC patients undergoing surgery at our facility. It was subdivided into training and validation (n = 40), and test (n = 9) datasets. We built a CT patch-based discriminating model using a residual convolutional neural network and employed 5-fold cross-validation. The discriminating performance of the model was analyzed in the test dataset.

Results: Of the 40 patients in the training dataset, GBC and XGC were observed in 21 (52.5%), and 19 (47.5%) patients, respectively. A total of 61 126 patches were extracted from the 40 patients. In the validation dataset, the average sensitivity, specificity, and accuracy were 98.8%, 98.0%, and 98.5%, respectively. Furthermore, the area under the receiver operating characteristic curve (AUC) was 0.9985. In the test dataset, which included 11 738 patches, the discriminating accuracy for GBC patients after neoadjuvant chemotherapy (NAC) (n = 3) was insufficient (61.8%). However, the discriminating model demonstrated high accuracy (98.2%) and AUC (0.9893) for cases other than those receiving NAC.

Conclusion: Our CT-based DL model exhibited high discriminating performance in patients with GBC and XGC. Our study proposes a novel concept for selecting the appropriate procedure and avoiding unnecessary invasive measures.

KEYWORDS

deep learning, gallbladder cancer, neural network, precision medicine, xanthogranulomatous cholecystitis

Hiroaki Fujita and Taiichi Wakiya contributed equally to this study.

1 | INTRODUCTION

Gallbladder cancer (GBC) is the most common malignancy of the bile duct, with a high fatality rate.[1] Radical resection is the best chance for a cure. Various types of surgical resections, from minimally invasive surgery to extended surgery, have been performed in patients with GBC based on the stage of the tumor.[2] Unfortunately, although it should definitely be avoided, unnecessary prolonged surgery has also been performed in patients with benign gallbladder diseases in whom GBC was undeniable. Extended surgery can increase the risk of perioperative complications and in-hospital death.[3] Thus, it is evident that an accurate preoperative diagnosis is needed to avoid unnecessary overinvasive procedures in patients with non-GBC.

Xanthogranulomatous cholecystitis (XGC) mimics GBC. XGC is a benign chronic inflammatory disease of the gallbladder.[4] Characteristically, in addition to being locally aggressive, it can also spread to adjacent organs such as the liver, duodenum, colon, and common bile duct, thereby forming a tumor-like mass around the gallbladder. The clinical and radiological presentations are very similar to GBC, which leads it to be misdiagnosed as GBC.[5] Consequently, surgeons may perform improper surgery owing to a misdiagnosis of GBC. Moreover, due to the high fatality rate of GBC, any preoperative suspicion in diagnosis should lead to wide surgical excisions.[4] To avoid improper and unnecessary surgery for XGC patients, accurately distinguishing between XGC and GBC before surgery is extremely important.

To improve differential diagnostic skills, various approaches have been reported, including using blood biomarkers and imaging analysis. These studies are useful, however there is room for improvement when it comes to distinguishing with accuracy between these two particular diseases. To resolve this issue, we focused on deep learning (DL), which has the potential to revolutionize medical diagnosis and management.[6] Convolutional neural networks, which are a core component of DL, are especially recognized as demonstrating high performance in image recognition. The essence of image recognition by convolutional neural networks entails extracting different levels of features such as low-level edges and color features, as well as more abstract features, through a series of convolutional and pooling layers. This approach has the potential to reveal unforeseen patterns and details that would be hidden without application of advanced data-mining techniques.[7] Indeed, there have been increasing reports of applying DL to the assessment and prediction of radiological images in clinical settings.[8] However, there are no reports on applying the combination of computed tomography (CT) and DL methods for differential diagnosis between GBC and XGC.

Therefore, this study aimed to investigate the potential of DL algorithms to distinguish between XGC and GBC. Here, we have successfully developed a model combining CT images and DL that accurately makes the distinction. We are proposing a novel concept, from a completely different perspective, to maximize the therapeutic success of surgery for both GBC and XGC.

2 | METHODS

2.1 | Patients and study design

This single-center, retrospective, observational study was approved by the Committee of Medical Ethics of Hirosaki University Graduate School of Medicine (Aomori, Japan; reference no. 2021-050). Informed consent was obtained in the form of opt-out on our website (https://www.med.hirosaki-u.ac.jp/hospital/outline/resarch/resarch.html), with the approval of the Committee of Medical Ethics of Hirosaki University Graduate School of Medicine. Our study did not include minors. This study was designed and carried out in accordance with the Declaration of Helsinki.

A total of 49 patients undergoing surgery for GBC and XGC at our facility were included in the study. This group was made up of a

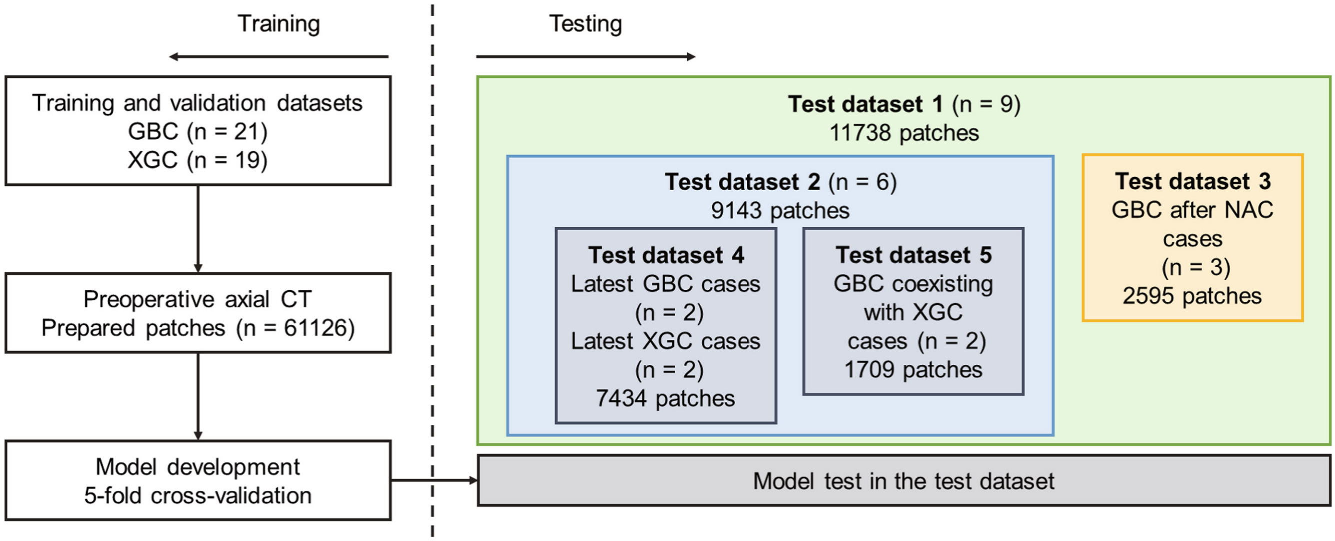

FIGURE 1 The study workflow and methodological process

total of 28 GBC patients who underwent surgery between 2014 and 2020 and 21 XGC patients who underwent surgery between 2008 and 2019. All patients had a pathologically confirmed diagnosis of either GBC or XGC by board-certified pathologists. In total, 40 patients comprised the training and validation sets and nine comprised the test set. Baseline clinicopathologic data were obtained from the medical records.

2.2 | CT acquisition and tumor segmentation

Our workflow is shown in Figure 1. The preoperative axial delayed phase of enhanced CT images for each case were obtained from our facility and were used for this study. The spatial resolution of the CT images was adjusted to 0.0126 mm/pixel. Radiological assessment was performed by board-certified radiologists. Board-certified surgeons performed CT acquisition and tumor segmentation based on the radiological assessment. Using a commercial viewer (ShadeQuest/ViewR; Fujifilm, Tokyo, Japan), CT images showing the tumor area were selected. Images of the obvious tumor region were manually segmented using Adobe Illustrator (Mountain, CA, USA) and saved.

2.3 | Preparation of dataset

We trimmed a square with a size of 128×128 pixels using a 32-pixel stride from the entire segmented tumor area. For the training and validation group, 61 126 patches were obtained from the 40 patients. For the model test cohort, 11 738 patches were obtained from the nine patients. Finally, 72 864 patches were obtained from the 49 patients in the study.

2.4 | Architecture of the convolutional neural network

Residual network (ResNet) 50[9] and Pytorch (a python library) were utilized (available at: https://github.com/pytorch/pytorch). We did not use a pretrained model. The original images acquired of 128×128 pixels were converted into images of 224×224 pixels. We tuned the hyperparameters as follows: number of training epochs, 50; batch size, 128; learning rate, 0.00025 via trial and error; and number of outer layers, two classes (see detail in Appendix S1). We used cross-validation to obtain more accurate results with less bias in the machine-learning studies.[10] In this study the training

TABLE 1 Comparison of the perioperative characteristics

	All cases	GBC	XGC		Logistic regression		
	(n = 49)	(n = 28)	(n = 21)	*P* value	Odds ratio	95% CI	*P* value
Gender							
Male, n	28 (57.1)	12 (42.9)	16 (76.2)	.024	3.680	0.583–23.226	.166
Female, n	21 (42.9)	16 (57.1)	5 (23.8)				
Age, y	70 (50–84)	73 (50–84)	64 (50–77)	.006	1.116	1.007–1.236	.036
Body height, cm	161 (130–173)	155 (130–173)	162 (141–173)	.040	[a]		
Body weight, kg	60 (32–85)	58 (32–83)	66 (46–85)	.036	1.000	0.916–1.091	.992
Body mass index, kg/m^2	23.8 (16.5–30.5)	23.2 (18.2–29.3)	25.2 (16.5–30.5)	.203			
Laboratory values							
WBC, /μL	4650 (3130–8280)	4640 (3130–7880)	4990 (3160–8280)	.768			
NLR	2.1 (0.9–10.1)	2.4 (0.9–8.5)	1.9 (0.9–10.1)	.078			
Hemoglobin, g/dL	13.0 (8.3–16.0)	12.9 (8.3–16.0)	13.8 (10.4–15.7)	.048	0.681	0.415–1.118	.129
AST, U/L	24 (4–139)	22 (4–94)	28 (14–139)	.267			
ALT, U/L	22 (6–207)	18 (6–82)	25 (9–207)	.148			
GTP, U/L	36 (12–518)	24 (12–281)	84 (14–518)	.003	0.989	0.980–0.998	.022
Total bilirubin, mg/dL	0.6 (0.2–19.6)	0.8 (0.3–19.6)	0.5 (0.2–2.1)	.388			
CA19-9, U/mL	23 (2–5109)	26 (6–5109)	20 (2–200)	.380			
CEA, ng/mL	2.4 (0–15.1)	2.5 (1–15.1)	2.2 (0–12.7)	.203			
PET-CT, performed	38 (77.6)	26 (92.9)	12 (57.1)	.005			
SUVmax	6.3 (1.1–19.6)	5.8 (1.1–17.8)	9.1 (4.5–19.6)	.127			

Abbreviations: ALT, alanine aminotransferase; AST, aspartate aminotransferase; CA19-9, carbohydrate antigen 19-9; CEA, carcinoembryonic antigen; CI, confidence interval; GBC, gallbladder cancer; GTP, glutamyl transpeptidase; NLR, neutrophil-to-lymphocyte ratio; PET, positron emission tomography; SUV, standardized uptake value; WBC, white blood cell; XGC, xanthogranulomatous cholecystitis.

[a]Excluded due to multicollinearity with gender.

and validation datasets were split into 5-fold, 1-fold of which is for validation and the other folds are for training. The proportion of patients with GBC vs XGC was equal in each fold. The training and validation processes were repeated five times, using different folds each time. Then we evaluated the discriminating performance using the test datasets. The final results were averaged and the standard deviation was calculated.

2.5 | Evaluation methods

The accuracy, sensitivity, specificity, false-positive rate (FPR), false-negative rate (FNR), positive predictive values (PPV), and negative predictive values (NPV) were evaluated. Furthermore, the model was also evaluated using the area under the receiver operating characteristic (ROC) curve (AUC).

2.6 | Statistical analyses

Continuous variables were expressed as the medians (ranges) and analyzed using nonparametric methods for nonnormally distributed data (Mann–Whitney *U*-test). Categorical variables were reported as numbers (percentages) and analyzed using the chi-squared test or Fisher's exact test, as appropriate. Variables with a significant relationship with GBC in univariate analysis were used in a binary logistic regression analysis. The statistical analyses were performed using IBM SPSS Statistics for Windows, v. 26.0 (IBM, Armonk, NY, USA).

3 | RESULTS

3.1 | Comparison of the perioperative characteristics of the GBC and XGC groups

The clinical characteristics of the 49 enrolled patients are shown in Table 1. In the GBC group, on average, the age was higher and the body weight was lower than the XGC group. The GBC group was significantly associated with a lower level of preoperative hemoglobin (12.9 vs 13.8 g/dL, $P = .048$) and glutamyl transpeptidase (24 vs 84 g/dL, $P = .003$). In contrast, there were no significant differences in tumor biomarkers such as carbohydrate antigen 19-9 and carcinoembryonic antigen between the groups.

Table 2 shows the pathological characteristics of gallbladder cancer cases. These pathological findings are based on the 3rd English edition of the Japanese Society of Hepato-Biliary-Pancreatic Surgery classification of biliary tract cancers.[11] The GBC group included a variety of macroscopic and histological types. In addition, the GBC group included early-stage to advanced stages. These results indicated that the target population of this study was not a particularly biased group. In short, our model was generated based on the images from a diverse population with different tumor stages.

TABLE 2 Pathological characteristics of gallbladder cancer cases

	GBC (n = 28)
Tumor size, mm	30 (10-90)
Macroscopic types	
Papillary-expanding type, n	5 (17.9)
Papillary-infiltrating type, n	7 (25.0)
Nodular-expanding type, n	2 (7.1)
Nodular-infiltrating type, n	7 (25.0)
Flat-expanding type, n	1 (3.6)
Flat-infiltrating type, n	5 (17.9)
Massive type, n	1 (3.6)
Tumor size, mm	30 (10-90)
Histological types	
Papillary adenocarcinoma, n	1 (3.6)
Tubular adenocarcinoma	
Well differentiated, n	15 (53.6)
Moderately differentiated, n	9 (32.1)
Poorly differentiated adenocarcinoma, n	1 (3.6)
Adenosquamous (cell) carcinoma, n	2 (7.1)
Japanese classification T category, n	
is/1a/1b/2/3a/3b/4a/4b	3 (10.7)/2 (7.1)/1 (3.6)/12 (42.9)/7 (25.0)/2 (7.1)/1 (3.6)/0
Japanese classification N category, n	
0/1	17 (60.7)/11 (39.3)
Japanese classification M category, n	
0/1	23 (82.1)/5 (17.9)
Japanese classification stage, n	
0/I/II/IIIA/IIIB/IVA/VB	3 (10.7)/3 (10.7)/10 (35.7)/1 (3.6)/6 (21.4)/0/5 (17.9)
Cancer stromal volume, n	
Medullary/intermediate/scirrhous	8 (28.6)/15 (53.6)/5 (17.9)
Cancer infiltrative (INF) pattern, n	
INFa/INFb/INFc	7 (25.0)/13 (46.4)/8 (28.6)
Lymphatic invasion, n	
0/1/2/3	9 (32.1)/6 (21.4)/10 (35.7)/3 (10.7)
Venous invasion, n	
0/1/2/3	8 (28.6)/1 (3.6)/13 (46.4)/6 (21.4)
Perineural invasion, n	
0/1/2/3	10 (35.7)/4 (14.3)/5 (17.9)/9 (32.1)

Note: These pathological findings are based on the 3rd English edition of the Japanese Society of Hepato-Biliary-Pancreatic Surgery classification of biliary tract cancers.

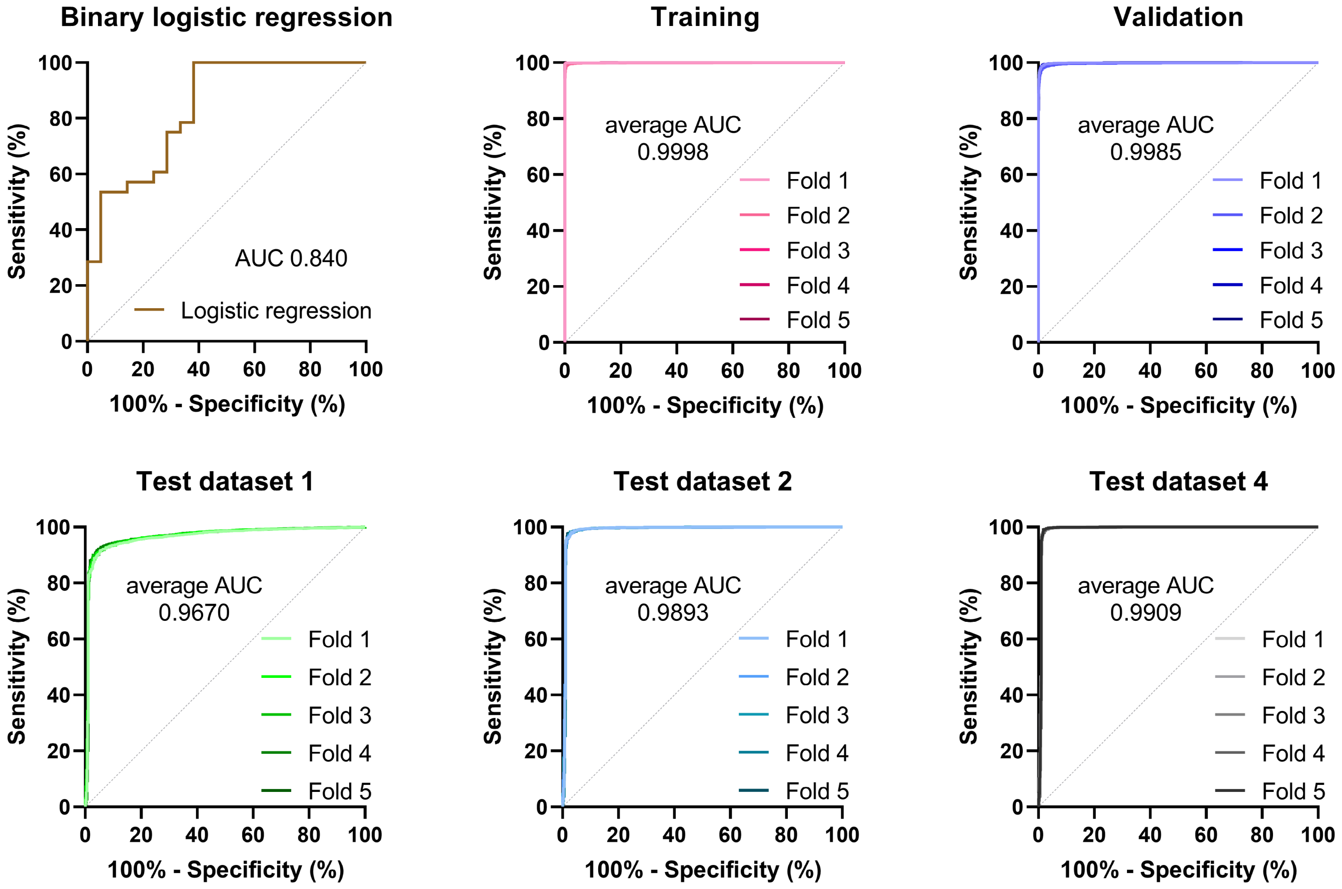

FIGURE 2 The area under the receiver operating characteristic curve (AUC) of logistic regression analysis and the deep-learning (DL) model. The AUC of logistic regression analysis was 0.840. In the DL model, the average AUC in training, validation, and test datasets were 0.9998, 0.9985, 0.9670, respectively

3.2 | Preoperative diagnosis by the radiologist

We retrospectively accessed the actual diagnostic performance of board-certified radiologists with preoperative CT images. Of the 28 patients in the GBC group, 26 patients (92.9%) had a preoperative diagnosis of GBC (n = 4) or suspected GBC (n = 20). In contrast, of the 21 XGC patients, 13 patients (61.9%) had a preoperative diagnosis of GBC (n = 3) or suspected GBC (n = 10). In this analysis, we treated cases with a preoperative diagnosis of both GBC and suspected GBC as malignant. Similarly, we treated cases with a preoperative diagnosis of both XGC and other benign gallbladder diseases as benign. In the setting of differentiating between benign and malignant, the sensitivity, specificity, PPV and NPV, and accuracy were 92.9%, 38.1%, 66.7%, 80.0%, and 69.4%, respectively.

3.3 | Binary logistic regression analysis

To predict GBC, we performed a binary logistic regression analysis, which is one of the traditional methods. We set GBC as the dependent variable. Significant variables linked with GBC, which were found through a univariate analysis ($P<.05$), as listed in Table 1, were entered into a binary regression analysis. Binary logistic regression indicated that patient age and preoperative glutamyl transpeptidase values were significant predictors of GBC (χ^2 = 23.372, and $P<.001$). The result of the Hosmer–Lemeshow test was P = .320. Glutamyl transpeptidase was significant at the 5% level (Wald = 5.214, P = .022). The odds ratio (OR) was 0.989 (95% confidence interval [CI]: 0.980–0.998). Patient age was significant at the 5% level (Wald = 4.415, P = .036). The OR was 1.116 (95% CI: 1.007–1.236). The model correctly predicted 89.3% of GBC cases and 61.9% of XGC cases, giving an overall correct prediction rate of 77.6%. The model achieved an AUC of 0.840 (95% CI: 0.727–0.953) (Figure 2).

3.4 | Performance of the DL model in the training and validation datasets

A total of 22 378 patches were obtained from the 19 patients with XGC. Furthermore, a total of 38 748 patches were obtained from the 21 patients with GBC. Finally, a total of 61 126 patches were extracted from the 40 patients in the current study. In the training and validation datasets, of the 21 patients with GBC, differential diagnosis between GBC and XGC was unable to be determined for 18 patients (85.7%). The other three patients with GBC had a

TABLE 3 Performance of the prediction model in the training and validation datasets

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average	SD
Training dataset							
Sensitivity, %	99.9	99.0	99.6	99.9	99.9	99.7	0.4
Specificity, %	99.9	99.2	99.0	99.9	99.8	99.6	0.5
FNR, %	0.1	1.0	0.4	0.1	0.1	0.3	0.4
FPR, %	0.1	0.8	1.0	0.0	0.2	0.4	0.5
PPV, %	99.9	99.5	99.4	99.9	99.9	99.8	0.3
NPV, %	99.7	98.3	99.4	99.9	99.9	99.4	0.7
Accuracy, %	99.9	99.1	99.4	99.9	99.9	99.6	0.4
AUC	1.0000	0.9994	0.9997	1.0000	1.0000	0.9998	0.0003
Validation dataset							
Sensitivity, %	98.9	98.2	98.9	99.0	99.0	98.8	0.3
Specificity, %	98.5	97.9	97.1	98.6	97.9	98.0	0.6
FNR, %	1.1	1.8	1.1	1.0	1.0	1.2	0.3
FPR, %	1.5	2.1	2.9	1.4	2.1	2.0	0.6
PPV, %	99.1	98.8	98.3	99.2	98.8	98.8	0.3
NPV, %	98.1	96.9	98.1	98.3	98.2	97.9	0.6
Accuracy, %	98.7	98.1	98.2	98.9	98.6	98.5	0.3
AUC	0.9990	0.9979	0.9979	0.9990	0.9988	0.9985	0.0006

Abbreviations: AUC, the area under the receiver operating characteristic curve; FNR, false-negative rate; FPR, false-positive rate; NPV, negative predictive value; PPV, positive predictive value; SD, standard deviation.

TABLE 4 Performance of the prediction model in the test datasets

	Test dataset 1	Test dataset 2	Test dataset 3	Test dataset 4	Test dataset 5
Sensitivity, %	89.9	98.3	61.8	99.5	93.5
Specificity, %	96.0	96.0	N/A	96.0	N/A
FNR, %	10.1	1.7	38.2	0.5	6.5
FPR, %	4.0	4.0	N/A	4.0	N/A
PPV, %	99.8	99.8	100.0	99.7	100.0
NPV, %	29.5	76.5	0	93.2	0
Accuracy, %	90.1	98.2	61.8	99.3	93.5
AUC	0.9670	0.9893	N/A	0.9909	N/A

Note: All values are average value of five tests.

Abbreviations: AUC, the area under the receiver operating characteristic curve; FNR, false-negative rate; FPR, false-positive rate; NPV, negative predictive value; PPV, positive predictive value; SD, standard deviation.

preoperative diagnosis of GBC. Of the 19 XGC patients, 12 patients (63.2%) had a preoperative diagnosis of GBC (n = 3) or suspected GBC (n = 9).

The average accuracy of the discriminating model for predicting GBC was 99.6% for the training dataset. The average sensitivity, specificity, and PPV and NPV were 99.7%, 99.6%, 99.8%, and 99.4%, respectively (Table 3). The model achieved an AUC of 0.9998 (95% CI: 0.9997–1.0000, $P < .0001$) in the training dataset.

Likewise, the model showed high predictive performance in the validation dataset. The average sensitivity, specificity, and PPV and NPV were 98.8%, 98.0%, 98.8%, and 97.91%, respectively. In the validation dataset, the DL model achieved an accuracy of 98.5% (Table 3). The model achieved an AUC of 0.9985 (95% CI: 0.9981–0.9990, $P < .0001$) in the validation dataset (Figure 2).

3.5 | Performance of the DL model in the test datasets

Next, we evaluated the performance of the developing model using the test dataset. In the test dataset, of the seven patients with GBC, the differential diagnosis between GBC and XGC was unable to be determined for five patients (71.4%). Of the two XGC patients, one patient had a preoperative diagnosis of suspected

TABLE 5 Comparison of the diagnostic accuracy in previous reports

Publication	Year	Subject	Modality	Sensitivity, %	Specificity, %	Accuracy, %	AUC
Wang YF, et al[12]	2014	GBC vs cholecystitis, GP, GS, HC	CEA	11.5	97.4	N/A	N/A
			CA19-9	71.7	96.1	N/A	N/A
			CA242	64.1	98.7	N/A	N/A
Chen Z, et al[13]	2020	GBC vs GA, GP	CA19-9	64.1	94.1	N/A	0.84
			CA19-9 + NLR	74.8	89.7	N/A	0.87
			NLR	74.8	64.0	N/A	0.73
Rana S, et al[14]	2012	GBC vs GS, HC	CA242	64.0	83.0	N/A	0.76
Bang SH, et al[15]	2014	GBC vs ADM	US	73.1	96.3	88.8	0.96
Lee ES, et al[16]	2015	GBC vs XGC	US	84.2	91.7	87.5	0.86
Bo X, et al[17]	2019	GBC vs XGC	US	80.0	86.0	N/A	N/A
Chen LD, et al[18]	2017	GBC vs abscess, ADM, cholecystitis	CEUS	92.0	87.0	89.6	0.90
Bo X, et al[17]	2019	GBC vs XGC	CEUS	90.0	93.0	N/A	N/A
Yuan Z, et al[19]	2021	GA, GBC vs ADM, GP	CEUS	87.1	69.0	N/A	N/A
Choi JH, et al[20]	2013	GBC vs ADM, GA, GP, XGC	EUS	90.0	91.1	N/A	N/A
Leem G, et al[21]	2018	GBC vs ADM, cholecystitis, GA, GP	EUS	77.1	82.7	N/A	0.91
Choi JH, et al[20]	2013	GBC vs ADM, GA, GP, XGC	CH-EUS	93.5	93.2	N/A	N/A
Leem G, et al[21]	2018	GBC vs ADM, cholecystitis, GA, GP	CH-EUS	97.1	55.5	N/A	0.94
Kamata K, et al[22]	2018	GBC vs ADM, cholecystitis, GP	CH-EUS	90.0	98.0	96.0	0.97
Bang SH, et al[15]	2014	GBC vs ADM	CT	50.0	98.2	82.5	0.90
Lee ES, et al[16]	2015	GBC vs XGC	CT	88.4	77.5	81.6	0.84
Bo X, et al[17]	2019	GBC vs XGC	CT	71.0	92.0	N/A	N/A
Ito R, et al[23]	2020	GBC vs XGC	CT	77.0	94.0	89.0	0.94
Bang SH, et al[15]	2014	GBC vs ADM	MRI	80.8	98.2	92.5	0.94
Lee NK, et al[24]	2014	GBC vs ADM, cholecystitis, GA, GP, XGC	MRI	97.2	92.2	N/A	0.95
Lee ES, et al[16]	2015	GBC vs XGC	MRI	93.3	84.2	89.3	0.91
Bo X, et al[17]	2019	GBC vs XGC	MRI	75.0	90.0	N/A	N/A
Lee J, et al[25]	2012	GBC vs ADM, GA, GP	PET-CT	85.0	87.0	86.0	0.82
Ramos-Font C, et al[26]	2014	GBC vs ADM, cholecystitis, GA	PET-CT	78.1	88.2	N/A	0.80
Bo X, et al[17]	2019	GBC vs XGC	PET-CT	55.0	90.0	N/A	N/A
Zhou QM, et al[27]	2022	GBC vs XGC	ML-based CT	N/A	N/A	0.84	0.82
			ML-based MRI	N/A	N/A	0.84	0.84
			ML-based CT/MRI	N/A	N/A	0.90	0.90
Jeong Y, et al[28]	2020	GBC etc.[a] vs GP	DL-based US	74.3	92.1	85.7	0.92
Current study	**2022**	**GBC vs XGC**	**DL-based CT**	**99.5**	**96.0**	**99.3**	**0.99**

Abbreviations: ADM, adenomyomatosis; AUC, the area under the receiver operating characteristic curve; CA, carbohydrate antigen; CEA, carcinoembryonic antigen; CEUS, contrast-enhanced ultrasound; CH-EUS, contrast-enhanced harmonic endoscopic ultrasonography; CT, computed tomography; DL, deep learning; EUS, endoscopic ultrasonography; GA, gallbladder adenoma; GBC, gallbladder cancer; GP, gallbladder polyp; GS, gallbladder stone; HC, healthy control; ML, machine learning; MRI, magnetic resonance imaging; NLR, neutrophil-to-lymphocyte ratio; PET, positron emission tomography; US, ultrasound; XGC, xanthogranulomatous cholecystitis.

[a]Including, adenoma, intracystic papillary neoplasm, intracystic tubulopapillary neoplasm, fibroepithelial polyp, adenocarcinoma, intracystic papillary neoplasm with an associated invasive carcinoma, papillary carcinoma, and adenosquamous carcinoma.

GBC, and the other had a preoperative diagnosis of chronic cholecystitis. We made several groupings taking real clinical situations into consideration. In test dataset 1, including all nine test cases, the prediction models showed both an acceptable accuracy (90.1%) and an acceptable AUC (0.9670) (Table 4). To investigate whether postchemotherapy images are also predictive, we evaluated the predictive performance of test dataset 3, consisting only of cases receiving neoadjuvant chemotherapy. As a result, the model demonstrated low performance, suggesting that the developing model was not suitable for postchemotherapy cases. In the clinical setting, we occasionally encounter GBC coexisting with XGC. Thus, we also assessed the model's performance in those cases (test dataset 5), confirming an acceptable performance (Table 4). Finally, in the data from similar patients, used as the training data (test dataset 4), the discriminating model achieved an acceptable accuracy (99.3%) as well as a high AUC (0.9909) (Figure 2). These results provided us with the knowledge that our model could have high versatility.

4 | DISCUSSION

We applied the DL model to distinguish between GBC and XGC. We have successfully shown high discrimination performance using CT images. This report represents the first study in which a DL model based on CT images was used to distinguish GBC and XGC. Our approach may contribute to selecting a proper surgical procedure for these conditions.

The strength of the current study shows extremely high discriminating performance compared to past reports (Table 5). Various approaches have been used to distinguish GBC from benign gallbladder diseases.[12–28] However, there is no gold standard due to insufficient accuracy. These results suggested the need for innovation in the differential diagnosis of gallbladder disease. Machine learning, including DL, has the potential to revolutionize disease diagnosis and management in the medical field.[29] In the field of GBC diagnosis, there are few studies applying machine learning. Zhou et al established a machine-learning-based diagnostic prediction model showing good diagnostic accuracy for the preoperative discrimination of XGC and GBC (AUC 0.888).[27] However, compared with the previous machine-learning-based model, our DL-based model achieved higher predictive accuracy (AUC 0.9893).

The other strength of our study is the unique test dataset setting reflecting real clinical practice. We evaluated the model performance in GBC patients with concomitant XGC, and in GBC patients after NAC, respectively. The frequency of the coexistence of GBC and XGC was reported to be ~10%.[30] A past study using machine learning did not assess these cases of coexistence.[27] Therefore, although the patient number was not large, the setting of our test dataset was significant. Moreover, our model demonstrated insufficient discriminating performance in the GBC patients after NAC. Because NAC can induce tissue changes, this low accuracy seemed reasonable. In short, this model was generated on a training dataset that did not include post-NAC GBC cases, so, as would be expected, it is not suitable for prediction in post-NAC GBC cases.

Intraoperative frozen section diagnosis is a reliable method for differential diagnosis. However, this method is not recommended when considering NAC. NAC for advanced GBC has potential survival advantages, although we do not have strong evidence to recommend it.[31] Generally, confirmed histological diagnosis is necessary before performing NAC. Endoscopic ultrasonography (EUS) fine-needle aspiration (EUS-FNA) has high diagnostic success in gallbladder diseases. In contrast, we cannot ignore the invasiveness of this and the risk of cancer dissemination.[32] To mitigate the unavoidable risk of EUS-FNA, it would be better to use a computer-aided diagnosis proactively. Taken together, our approach could also be advantageous before extended surgery, as well as in NAC settings for GBC.

The present study does have several limitations. As XGC is a relatively rare disease, the patient population was not large. To compensate for this disadvantage, we used a data augmentation technique. We further used the 5-fold cross-validation method and evaluated the generalizability in the additional test dataset. Consequently, our model achieved favorable performance levels even in the test dataset. Generally, models that are generated on huge datasets are able to achieve higher discrimination accuracy with higher generalizability and robustness. To establish clinical applications widely, a future study with huge datasets, such as national or regional datasets, would be attractive to both clinicians and their patients.

In recruiting patients from many facilities, a CT-based approach may be preferred for this topic. In the real world, ultrasound, CT, and magnetic resonance imaging (MRI) are the widely used diagnostic modalities for gallbladder disease. Unlike CT exams, an ultrasound-based approach has no radiation exposure. Among the conventional methods, EUS is the most accurate diagnostic modality for gallbladder disease. However, EUS has limitations, including the lack of standardization and subjective interpretation. Furthermore, particularly with EUS, it may not be easy to reproduce the same image. In short, a CT exam is superior to EUS in objectivity and reproducibility. Indeed, the result of the EUS-based approach using ResNet50, the same architecture as the current study, leaves room for improvement (AUC 0.71).[33] Besides, CT exams are more widely used than MRI exams and are less expensive. Our concept of combing CT images and DL enhances the possibilities for wide availability and good usability in clinical applications worldwide.

In conclusion, our CT-based DL model exhibited high discriminating performance in patients with GBC and XGC. Our study, using DL, proposes a novel concept for avoiding unnecessary invasive procedures and being able to select the appropriate procedure. For a reliable clinical model, conducting a further large-scale study is the hope for the future.

ACKNOWLEDGMENTS

We sincerely thank Shari Joy Berman for professionally editing the English draft of this article.

DISCLOSURE

Conflict of Interest: Author K.H is an editorial member of Annals of Gastroenterological Surgery.

Author contributions: TW contributed to the study conception and design. TW, KI, NK, HN, and KH performed surgical resection and collected the clinical data. HF, TW, and TK performed CT acquisition and tumor segmentation. YS established the prediction model. HF and TW wrote the first draft of the article. MM, YS, and KH contributed to the review, and/or critical revision of the article. All authors approved the final article.

Ethics Statements: This study was approved by the Committee of Medical Ethics of Hirosaki University Graduate School of Medicine (Aomori, Japan; reference no. 2021-050). Informed consent was obtained in the form of opt-out on our website (https://www.med.hirosaki-u.ac.jp/hospital/outline/resarch/resarch.html), with the approval of the Committee of Medical Ethics of Hirosaki University Graduate School of Medicine. Our study did not include minors. This study was designed and carried out in accordance with the Declaration of Helsinki.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author.

ORCID

Hiroaki Fujita https://orcid.org/0000-0002-3259-7054
Taiichi Wakiya https://orcid.org/0000-0003-3681-7736
Keinosuke Ishido https://orcid.org/0000-0002-0342-1199
Norihisa Kimura https://orcid.org/0000-0002-0585-3024
Taishu Kanda https://orcid.org/0000-0003-2839-3347
Masashi Matsuzaka https://orcid.org/0000-0002-8199-3037
Kenichi Hakamada https://orcid.org/0000-0001-6513-1202

REFERENCES

1. Miller KD, Ortiz AP, Pinheiro PS, Bandi P, Minihan A, Fuchs HE, et al. Cancer statistics for the US Hispanic/Latino population, 2021. CA Cancer J Clin. 2021;71:466–87.
2. Matsuyama R, Yabusita Y, Homma Y, Kumamoto T, Endo I. Essential updates 2019/2020: surgical treatment of gallbladder cancer. Ann Gastroenterol Surg. 2021;5:152–61.
3. Mizuno T, Ebata T, Yokoyama Y, Igami T, Yamaguchi J, Onoe S, et al. Major hepatectomy with or without pancreatoduodenectomy for advanced gallbladder cancer. Br J Surg. 2019;106:626–35.
4. Frountzas M, Schizas D, Liatsou E, Economopoulos KP, Nikolaou C, Apostolou KG, et al. Presentation and surgical management of xanthogranulomatous cholecystitis. Hepatobiliary Pancreat Dis Int. 2021;20:117–27.
5. 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. 2022;11:176–82.
6. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172:1122–31.e9.
7. Li Y, Zhou X, Colnaghi T, Wei Y, Marek A, Li H, et al. Convolutional neural network-assisted recognition of nanoscale L12 ordered structures in face-centred cubic alloys. Npj Comput Mater. 2021;7:8.
8. Igarashi S, Sasaki Y, Mikami T, Sakuraba H, Fukuda S. Anatomical classification of upper gastrointestinal organs under various image capture conditions using AlexNet. Comput Biol Med. 2020;124:103950.
9. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016:770–8.
10. Ojala M, Garriga GC. Permutation tests for studying classifier performance. J Mach Learn Res. 2010;11:1833–63.
11. Miyazaki M, Ohtsuka M, Miyakawa S, Nagino M, Yamamoto M, Kokudo N, et al. Classification of biliary tract cancers established by the Japanese Society of Hepato-Biliary-Pancreatic Surgery: 3(rd) English edition. J Hepatobiliary Pancreat Sci. 2015;22:181–96.
12. Wang YF, Feng FL, Zhao XH, Ye ZX, Zeng HP, Li Z, et al. Combined detection tumor markers for diagnosis and prognosis of gallbladder cancer. World J Gastroenterol. 2014;20:4085–92.
13. Chen Z, Liu Z, Zhang Y, Wang P, Gao H. Combination of CA19-9 and the neutrophil-to-lymphocyte ratio for the differential diagnosis of gallbladder carcinoma. Cancer Manag Res. 2020;12:4475–82.
14. Rana S, Dutta U, Kochhar R, Rana SV, Gupta R, Pal R, et al. Evaluation of CA 242 as a tumor marker in gallbladder cancer. J Gastrointest Cancer. 2012;43:267–71.
15. Bang SH, Lee JY, Woo H, Joo I, Lee ES, Han JK, et al. Differentiating between adenomyomatosis and gallbladder cancer: revisiting a comparative study of high-resolution ultrasound, multidetector CT, and MR imaging. Korean J Radiol. 2014;15:226–34.
16. Lee ES, Kim JH, Joo I, Lee JY, Han JK, Choi BI. Xanthogranulomatous cholecystitis: diagnostic performance of US, CT, and MRI for differentiation from gallbladder carcinoma. Abdom Imaging. 2015;40:2281–92.
17. Bo X, Chen E, Wang J, Nan L, Xin Y, Wang C, et al. Diagnostic accuracy of imaging modalities in differentiating xanthogranulomatous cholecystitis from gallbladder cancer. Ann Transl Med. 2019;7:627.
18. Chen LD, Huang Y, Xie XH, Chen W, Shan QY, Xu M, et al. Diagnostic nomogram for gallbladder wall thickening mimicking malignancy: using contrast-enhanced ultrasonography or multi-detector computed tomography? Abdom Radiol. 2017;42:2436–46.
19. Yuan Z, Liu X, Li Q, Zhang Y, Zhao L, Li F, et al. Is contrast-enhanced ultrasound superior to computed tomography for differential diagnosis of gallbladder polyps? A cross-sectional study. Front Oncol. 2021;11:657223.
20. Choi JH, Seo DW, Choi JH, Park DH, Lee SS, Lee SK, et al. Utility of contrast-enhanced harmonic EUS in the diagnosis of malignant gallbladder polyps (with videos). Gastrointest Endosc. 2013;78:484–93.
21. Leem G, Chung MJ, Park JY, Bang S, Song SY, Chung JB, et al. Clinical value of contrast-enhanced harmonic endoscopic ultrasonography in the differential diagnosis of pancreatic and gallbladder masses. Clin Endosc. 2018;51:80–8.
22. Kamata K, Takenaka M, Kitano M, Omoto S, Miyata T, Minaga K, et al. Contrast-enhanced harmonic endoscopic ultrasonography for differential diagnosis of localized gallbladder lesions. Dig Endosc. 2018;30:98–106.
23. Ito R, Kobayashi T, Ogasawara G, Kono Y, Mori K, Kawasaki S. A scoring system based on computed tomography for the correct diagnosis of xanthogranulomatous cholecystitis. Acta Radiol Open. 2020;9:2058460120918237.
24. Lee NK, Kim S, Kim TU, Kim DU, Seo HI, Jeon TY. Diffusion-weighted MRI for differentiation of benign from malignant lesions in the gallbladder. Clin Radiol. 2014;69:e78–85.
25. Lee J, Yun M, Kim KS, Lee JD, Kim CK. Risk stratification of gallbladder polyps (1-2 cm) for surgical intervention with 18F-FDG PET/CT. J Nucl Med. 2012;53:353–8.

26. Ramos-Font C, Gómez-Rio M, Rodríguez-Fernández A, Jiménez-Heffernan A, Sánchez Sánchez R, Llamas-Elvira JM. Ability of FDG-PET/CT in the detection of gallbladder cancer. J Surg Oncol. 2014;109:218–24.
27. Zhou QM, Liu CX, Zhou JP, Yu JN, Wang Y, Wang XJ, et al. Machine learning-based radiological features and diagnostic predictive model of xanthogranulomatous cholecystitis. Front Oncol. 2022;12:792077.
28. Jeong Y, Kim JH, Chae HD, Park SJ, Bae JS, Joo I, et al. Deep learning-based decision support system for the diagnosis of neoplastic gallbladder polyps on ultrasonography: Preliminary results. Sci Rep. 2020;10:7700.
29. Kochanny SE, Pearson AT. Academics as leaders in the cancer artificial intelligence revolution. Cancer. 2021;127:664–71.
30. Yucel O, Uzun MA, Tilki M, Alkan S, Kilicoglu ZG, Goret CC. Xanthogranulomatous cholecystitis: analysis of 108 patients. Indian J Surg. 2017;79:510–4.
31. Naveed S, Qari H, Thau CM, Burasakarn P, Mir AW. Neoadjuvant chemotherapy for advanced gallbladder cancer: do we have enough evidence? A systematic review. Euroasian J Hepatogastroenterol. 2021;11:87–94.
32. Nagino M, Hirano S, Yoshitomi H, Aoki T, Uesaka K, Unno M, et al. Clinical practice guidelines for the management of biliary tract cancers 2019: the 3rd English edition. J Hepatobiliary Pancreat Sci. 2021;28:26–54.
33. Jang SI, Kim YJ, Kim EJ, Kang H, Shon SJ, Seol YJ, et al. Diagnostic performance of endoscopic ultrasound-artificial intelligence using deep learning analysis of gallbladder polypoid lesions. J Gastroenterol Hepatol. 2021;36:3548–55.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Fujita H, Wakiya T, Ishido K, Kimura N, Nagase H, Kanda T, Differential diagnoses of gallbladder tumors using CT-based deep learning. Ann Gastroenterol Surg. 2022;6:823–832. doi:10.1002/ags3.12589